Catalogue no. 12-206-X
ISSN 1705-0820

Statistical Methodology Research and Development Program Achievements, 2020/2021

Release date: October 6, 2021

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistical Information Service 1-800-263-1136

- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Statistical Methodology Research and Development Program

Achievements, 2020/2021

This report summarizes the 2020/2021 achievements of the Methodology Research and Development Program (MRDP) sponsored by the Modern Statistical Methods and Data Science Branch at Statistics Canada. This program covers research and development activities in statistical methods with potentially broad application in the agency's statistical programs; these activities would otherwise be less likely to be carried out during the provision of regular methodology services to those programs. The MRDP also includes activities that provide client support in the application of past successful developments in order to promote the use of the results of research and development work. Selected prospective research activities are also presented. Contact names are provided for obtaining more information on any of the projects described. For more information on the MRDP as a whole, contact:

Susie Fortier
(613-220-1948, susie.fortier@statcan.gc.ca)

Jean-François Beaumont
(613-863-9024, jean-francois.beaumont@statcan.gc.ca)

Statistical Methodology Research and Development Program

Achievements, 2020/2021

Table of Contents

1 Modeling, data integration and data science

1.1 Small Area Estimation

Standard design-based estimates of population parameters, called direct estimates, are generally reliable provided that the sample size in the domains of interest is not too small. Indirect estimates, that borrow strength over areas or over time, often yield substantial gains of efficiency for small domains at the expense of introducing model assumptions. In recent years, there has been a renewed interest at Statistics Canada in investigating and using indirect model-based estimation methods for small domains. The main goals of this project are:

- to develop new estimation methods for small domains that address issues found in the production of small area estimates;
- to study properties of existing methods under different scenarios to better understand how and when to use them;
- to determine suitable small area estimation methodology for some candidate surveys;
- to develop and test prototypes implementing new or existing methods that could be beneficial to statistical programs.

So far, progress has been made in the following sub-projects.

SUB-PROJECT: Local diagnostics for the Fay-Herriot model

Model validation tools such as graphs of residuals are often used to assess the plausibility of the Fay-Herriot model. Model-based Mean Square Error (MSE) estimates are then used to evaluate the efficiency gains of small area estimators over direct estimators. All these techniques are useful to assess the overall performance of small area estimates. However, users are often interested in their specific domain only and a quality indicator for their specific domain estimate is more relevant to them. The model-based MSE achieves partially this goal but integrates out the local random effect (linking model error) that is of interest to users of a specific domain. The design MSE would be more relevant to these users but design-unbiased estimates of the design MSE are known to be very unstable (e.g., Rao, Rubin-Bleuer and Estevao, 2018). In this project, we developed and investigated two new local diagnostics for the evaluation of small area estimates.

Progress:

A paper was previously written and submitted to *Survey Methodology*. We received reviewers' comments this year and revised the paper accordingly. The paper has been accepted and will be published in a 2021 issue of *Survey Methodology* (Lesage, Beaumont and Bocci, 2021).

SUB-PROJECT: Robust small area models and estimation with application using Labour Force Survey data

The Fay-Herriot model with normal sampling errors and random effects is widely used in small area estimation to improve direct survey estimates. Ghosh, Myung and Moura (2018) proposed t priors for the random effects and a specifically modified Jeffrey's prior using a Hierarchical Bayes (HB) approach. In this project, we will evaluate the HB models with normal and t distributions based on the You and Chapman

(2006) model, compare the various HB small area models and study the effects of robust Small Area Estimation (SAE) modelling with t distribution for sampling errors and random effects. The proposed models and methods will be programmed in R and compared using Labour Force Survey (LFS) data.

Progress:

We considered the You and Chapman (2006) model and studied two types of robust SAE models based on normal and t distributions for sampling errors and random effects. Model specification and programs have been completed and programmed. The models and methods have been applied to LFS data for evaluation and comparison. A research report is completed (You, 2021a). The normal-t distribution model may be extended to the sampling variance model of Sugasawa, Tamae and Kubokawa (2017) and You (2021b).

SUB-PROJECT: Small area estimation with sampling variance smoothing and modeling

We consider the Fay-Herriot model and study Empirical Best Linear Unbiased Prediction (EBLUP) and HB approaches for small area estimation with sampling variance smoothing and modeling. We study and propose sampling variance smoothing and modeling methods and compare the models of You and Chapman (2006), You (2016) and Sugasawa et al. (2017) for the estimation of the LFS unemployment rate and other small area estimation problems.

Progress:

We studied EBLUP and HB approaches with sampling variance smoothing and modeling and applied our methods to LFS data. Our results indicate that sampling variance smoothing can improve the efficiency and accuracy of the model-based estimator. Sampling variance modeling can also be useful and performs much better than simply using direct sampling variance estimates. A research paper on this project has been written and will appear in the December issue of *Survey Methodology* (You, 2021b).

SUB-PROJECT: Small Area Estimation Prototype

A SAS prototype has been developed to produce small area estimates. This prototype forms the basis for the implementation of new SAE methods within the generalized estimation system, G-Est. The SAE prototype is used as a production tool to respond to the increasing demand for estimates at disaggregated levels.

Progress:

We implemented and tested a new model selection method to assist users in the identification of key auxiliary variables in the model. It is based on a backward selection approach using a Chi-square criterion to identify significant regressor variables. This improvement provides a considerable reduction in the time and effort to build a proper model.

SUB-PROJECT: Small area estimation of mean liquid assets by census division

Small area estimation of mean liquid assets at the census division level is investigated. The survey estimates of mean liquid assets at the census division level are calculated using data from the 2016 Survey

of Financial Security. These survey estimates are then modeled as a function of socio-economic characteristics of families taken from the Census of Population of 2016 at the same level to produce small area estimates of mean liquid assets through the Fay-Herriot model.

Progress:

The Fay-Herriot model was carefully validated and small area estimates of mean liquid assets at the census division level for 2016 were produced. A report was written (Bocci, Morissette and Beaumont, 2020), which includes a description of the small area estimation models.

SUB-PROJECT: Estimation of the design Mean Square Error in Small Area Estimation

The use of the Fay-Herriot model to produce small area estimates has increased at Statistics Canada in the last years. These estimates are typically accompanied with estimates of their model Mean Square Error (MSE). However, users are typically accustomed with estimates of the design MSE. The design MSE has the advantage over the model MSE of not integrating out the specificity of a particular domain, and may be more relevant to users as an indicator of the quality of the estimates. Design-based estimates of the design MSE are known to be unstable (e.g., Rao, Rubin-Bleuer and Estevao, 2018). We plan to investigate the use of a conditional approach to obtain a more efficient estimator of the design MSE.

Progress:

In previous research, an estimator of the design MSE based on a conditional approach was developed and a draft internal report was written. In the latter part of this fiscal year, a simulation study was started to evaluate the proposed conditional approach.

For more information, please contact:
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

REFERENCES

Ghosh, M., Myung, J. and Moura, F.A.S. (2018). Robust Bayesian small area estimation. *Survey Methodology*, 44, 101-115.

Rao, J.N.K., Rubin-Bleuer, S. and Estevao, V.M. (2018). Measuring uncertainty associated with model-based small area estimators. *Survey Methodology*, 44, 151-166.

Sugasawa, S., Tamae, H. and Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44, 150-167.

You, Y. (2016). Hierarchical Bayes sampling variance modeling for small area estimation based on area level models with applications. Methodology Branch working paper, ICCSMD-2016-03-E.

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97-103.

1.2 Real-time estimation via time series methods

SUB-PROJECT: Nowcasting Economic Indicators

We aim to develop an early indicator of Canadian economic activity in collaboration between various teams within Statistics Canada. The goal is to leverage traditional and new machine learning methods combined with alternative data to nowcast monthly values of economic activities.

Progress:

A cloud-based modeling environment on Statistics Canada's Advanced Analytics Workspace has been established. An extensive time-series database with ElasticSearch was build consisting of economic and alternative data of different frequencies. Processing pipelines were built in Python and R to allow time series specialists to access the database for the modeling. Modeling results indicate that some of the alternative data sources used can improve the nowcasting performance and might be relevant and very timely data for future macroeconomic modeling. Time series models of type ARIMAX, PROPHET and different machine learning models including XGBoost were compared in performance. The ARIMAX model performed only slightly better than PROPHET and the tuned XGBoost model. Early results indicate that a hybrid (meta) model combining the three models can improve the performance of ARIMAX further.

For more information about the Advance Analytics Workspace, you can visit:
https://www.statcan.gc.ca/eng/data-science/network/cloud-platform.

SUB-PROJECT: Modelling and Forecasting in the context of real-time estimation

Increasing timeliness of statistical indicators is an important priority for Statistics Canada and one option for doing so is through time series modelling to nowcast economic indicators much earlier than the point in time where the first traditional estimator is produced.

Progress:

Evaluation of statistical models in this context was continued, identifying ARIMA models and state space models as the most appealing candidates. A presentation was made to Statistics Canada's Advisory Committee on Statistical Methods to gather feedback on the progress and future plans for this work (Matthews, Patak and Picard, 2020).

Two case studies were conducted in collaboration with the Data Science Division, to evaluate predictive methods for nowcasting Building Permits, and monthly Gross Domestic Product. These studies compared more traditional models such as RegARIMA with machine learning predictive models, indicating a similar performance and highlighting not only differences between the methods but also many similarities (Matthews and Ritter, 2020). A presentation of this work has been developed to present at Statistics Canada's 2021 Methodology Symposium.

An in-depth analysis of Dynamic Factor Models was conducted, to understand the underlying theory and the potential for application in the context of nowcasting Canadian economic indicators. This method is widely used in other statistical organizations for nowcasting. It combines dimensionality reduction techniques to generate auxiliary variables, coupled with linear models in a state space framework to

generate the predictions. This evaluation was conducted in partnership with Dr. Rafal Kulik from the University of Ottawa, along with a graduate student, Ismael Zie Diamoutene who is also collaborating on this analysis. This investigation concluded that the predictive power of dynamic factor models relies very heavily on the availability of several highly correlated auxiliary variables, which are still being identified for Canadian economic indicators.

A draft guidelines document was prepared and discussed at the Data to Information: Modern Methods Committee, to promote standardization of terminology for producing advance indicators (including model-based nowcasts), to outline criteria to support the decision to publish new advance indicators, and provide guidance on appropriate methods for nowcasting along with advantages and disadvantages (Matthews, 2020b, 2021b). Progress has been made towards finalizing the guidelines document for inclusion in Statistics Canada's policy suite. This is expected to be completed within the first half of the coming fiscal year.

For more information, please contact:
Steve Matthews (613-854-3174, Steve.Matthews@statcan.gc.ca).

1.3 Machine Learning and other selected activities in data science

SUB-PROJECT: Identifying Covid-19 pandemic hubs at the health region level

In the early stage of the Covid-19 pandemic, Statistics Canada undertook a collaboration with Public Health Agency of Canada to identify and predict Covid-19 pandemic hotspots at the health-region level by using Machine Learning models. The main goal of the project was to develop a pseudo-spatiotemporal forecast model by socio-economic and socio-demographic, health-related, epidemiological and publicly available mobility data to divert public health resources from lower risk regions to higher risk regions.

Progress:

Supervised (deep learning based) risk-level prediction models were successfully developed (Arim, Hennessey and Molladavoudi, 2021). These are highly customizable and would allow risk prediction with various forecasting ranges/sliding windows (up to several days) at the health region level. An interactive dashboard was also developed that would allow federal/provincial health authorities to monitor trends in the covid-19 cases and deaths at the health-region level and divert public health resources (e.g. personal protective equipment, medical frontline workers …) from lower risk regions to higher risk regions and contain those cases in higher risk areas more quickly through isolation, contact tracing and quarantine of contacts. The proof-of-concept was successfully completed in fall 2020. The next step would be for the models and dashboard to be used in the context of Mobile Health Units in collaboration with Health Canada.

SUB-PROJECT: Event Detection - Detecting events from news articles.

The objective of this project is to identify the economic events related to specified companies from news articles (coming from different sources such as Factiva and Newsdesk) and represent the events in a timeline. The project also deals with providing a user-friendly way to interact with the data by highlighting identified events and companies over a dashboard.

Progress:

To detect relevant companies in news articles we have used BERT based Named-Entity-Recognition (NER) algorithms which are available in the library SpaCy. Since these NER models are language specific, deep learning translation models were considered. Translating the French articles to English allowed us to apply the NER algorithm on these as well. A multi-label classification approach was then used to detect specific events related to economic activities. For the front end, we have used Dash and Kibana technologies to design dashboards. Results are promising with next steps focusing on conducting more user testing.

For more information, you can visit: https://www.statcan.gc.ca/eng/data-science/network/cloud-tools.

SUB-PROJECT: Census Comment Classifier

At the end of the Census of Population questionnaires, respondents are provided with a text box to make comments which are classified by subject matter area - such as education, labour or demography - and shared with the corresponding expert analysts. The information is used to support decision making regarding content determination for the next census and to monitor factors such as respondent burden. During a proof of concept project, we evaluated whether machine learning techniques could be used to classify census respondents' comment into 16 different categories.

Progress:

We used labelled data from 2019 Census test to train a Bilingual Semi-Supervised Text Classification algorithm. We evaluated the performance of 4 different models: Support Vector Machines, single headed and multi headed Long Short Term Memory, Multi headed Convolutional Neural Network. The proof of concept project was successfully completed and the classifier will be used in production starting May to classify thousands of respondent comments from Canada's 2021 Census.

For more information, you can visit: https://www.statcan.gc.ca/eng/data-science/network/census-comment-2021.

SUB-PROJECT: Accessible Canada Act Indicator

Highlighting the regulations to be implemented by a given company involves reading thousands of long accessibility plans of no uniform format and can be very subjective. The idea behind this proof of concept is to explore how natural language processing can be used to search more than 200 regulation requirements amongst 10 to more than 150 pages of accessibility plans to quickly highlight the regulations to be implemented by a given company.

Progress:

At the end of the proof of concept, the team delivered a natural language processing model that used pre-trained word embedding to match accessibility plan items back to its regulation. At the same time, the pilot project identified nudges to increase the correctly extracted meaningful information from published plans. It should be noted that this pilot project is not an exhaustive evaluation of all plans and regulations; to apply the findings of the pilot project to an evaluation of Accessible Canada Act plans and regulations will require significant effort to refine the chosen model. Still, this work shows it is feasible.

SUB-PROJECT: Dynamic Topic Modelling

This project examined the potential application of topic modelling methods on the Canadian Coroner and Medical Examiner Database (CCMED) to detect changes in death patterns. The CCMED contains unstructured data in the form of free-text variables, called narratives, that provides detailed information on the circumstances surrounding these reported deaths. The goal of this project is to apply machine learning techniques to the narrative variable to uncover hidden semantic structures within the CCMED and analyze these structures dynamically (over time) to detect emerging death narratives.

Progress:

A first proof of concept to develop a dynamic topic modelling system has been designed, implemented and deployed using the British Columbia CCMED data. The system includes a model, data ingestion and processing pipeline and data visualization tools. The model groups death narratives into clusters with similar semantic (topics) structures while using a novel Bayesian dynamic topic modelling approach in which previously learned topics are used as Bayesian priors for current topics. To interpret, visualize and analyze the model outputs, a dashboard user interface has also been deployed. The next steps are: (1) to adapt this model for other Canadian provinces; and (2) to construct an indicator metrics in order to detect significant changes in different topics.

SUB-PROJECT: Learning Optimal COVID-19 Mitigation Strategies using Reinforcement learning

An agent based simulation environment was built to represent a Canadian population in Canada during COVID-19. Reinforcement learning was used to learn agent behaviours that minimize the spread of COVID within the simulation environment. The simulation environment was built using freely available data (from Statistics Canada, Canadian Institute of Health Information, and Public Health Agency of Canada). The goal is to optimize non-pharmaceutical intervention strategies implemented and determine the optimal set of population behaviours that minimize the spread of an infection within simulations.

Progress:

A population of 50,000 agents was built and 100 simulations were run while applying reinforcement learning. Agent behaviours were optimized in order to minimize the number of infections within the simulation environment. Behaviours were analyzed and interesting insights were discovered. Additionally, scenarios including “Covid fatigue” and non-compliance were studied. The project was successfully completed and led to a publication as a chapter in a forthcoming book on mathematical modelling for Covid-19 (Denis, El-Hajj Drummond, Abiza and Gopaluni, 2021).

For more information, you can visit: https://www.statcan.gc.ca/eng/data-science/network/npi.

SUB-PROJECT: Evaluating Office Building Re-opening Strategies during COVID-19 using Multi-Armed Bandits

Return to work and “re-opening” following Covid-related business and industry closures was considered. This project was done in collaboration with the Public Health Agency of Canada, and involved building an agent-based simulation environment for differently sized office buildings with various numbers of floors and elevators, to represent a work day under various scenarios. Over 200,000 different workplace scenarios were explored, comprising of the presence of screening mechanisms, employee capacity

control, the distribution of employees over floors, meeting sizes, elevator usage as well as work hour scheduling. We leverage sequential decision theory to control sampling of simulation scenarios using Multi-Armed Bandits. Through this approach, each scenario was evaluated based on the ability to minimize the number of infection events that occur within the workplace. This is the first modelling approach to consider return to work scenarios for an office building workplace setting, and resulted in recommendations shared across the government regarding to decisions around returning to work to office buildings.

Progress:

This project was successfully completed. Specifically, an agent-based simulation environment representing different sized office buildings was created. Over 200,000 distinct office building scenarios were simulated, comprising of a wide range of factors and variables unique to the office building workplace setting. Multi-armed bandits were used to control the sampling process. As well, key findings and recommendations from this work was circulated to agencies and departments across Canada. The Public Health Agency of Canada and Statistics Canada aim to prepare and publish a manuscript.

SUB-PROJECT: Crop Area Estimates on Data Analytics as a Service (DAaaS) Platform using Satellite Imagery

Building on the success of two previous proof-of-concepts, this project aims to build a novel machine learning model that takes as input a variable number of satellite images of a particular crop field in Canada, and perform regression on the image, estimating the acreage of 46 different crop types. The satellite imagery is paired with crop insurance data for specific Canadian provinces: Alberta, Manitoba and Saskatchewan. This project will allow for Statistics Canada to perform in-season crop area estimation as early as June or July. The goal is to perform a parallel trial for the 2021 crop season and eventually reduce or even replace costly surveys.

Progress:

Made possible by Statistics Canada's Advanced Analytics Workspace, we have done the costly pre-processing of 7 years of historical data, using Landsat-8 Imagery. The pre-processed data is being made available to train a neural network, and the massive amount of data led to the most accurate model that can be achieved. In parallel, the prototype "production system" is being developed, which will ultimately use the machine learning model to produce rolling estimates for the agricultural analysts.

For more information, you can visit: https://www.statcan.gc.ca/eng/data-science/network/satellite-imaging.

For more information about the Advance Analytics Workspace, visit:
https://www.statcan.gc.ca/eng/data-science/network/cloud-platform.

SUB-PROJECT: Spatial Layout based Information and Content Extraction

Portable Document Format (PDF) is most commonly used by companies for financial reporting purposes. The absence of effective means to extract data from these highly unstructured PDF files in a layout-aware manner presents a significant challenge for statistical organizations to efficiently analyze and process information in this rich admin data source in a timely manner. This research introduces 'Spatial Layout

based Information and Content Extraction' (SLICE) - a novel computer vision algorithm designed and developed by Statistics Canada that simultaneously uses textual, visual, and layout information to segment several data points into a tabular structure. This proposed modular solution significantly reduces the manual hours and efforts spent on identifying and capturing required information by automating the financial variable extraction for a variety of PDF documents.

Progress:

Over the four phases of this project, a number of methods where tested and applied. In its final state, SLICE uses pixels to determine ideal page divisions and map the entire page into several rectangular subsections. Once mapped, these subsections are leveraged using graph traversal techniques to determine position of each token within a tabular matrix. The product was showcase within Statistics Canada and to other departments as well and is now ready to move to production.

SUB-PROJECT: Detection of Construction sites using deep learning.

The goal of the project is to demonstrate the potential for reducing the cost of identification and categorization of construction sites and their properties by completely automatizing the detection process. Through extraction from satellite images using Deep Learning, this would enable faster (monthly) detection of construction sites. To detect the construction sites, we adopted a supervised machine learning based approach to classify the object class for each pixel within an image (semantic segmentation).

Progress:

We developed a data and meta-data structure for collaborative workflow, a parallelized image processing pipeline which can process 100GBs of images to create training sets on Advanced Analytics Workspace, accessible to collaborators.

We developed multiple machine learning models using semantic segmentation approach (U-net) of Ronneberger, Fischer et Brox (2015) for construction detection based on the image data and evaluated on unseen validation and test areas. We developed a post-processing pipeline to process the pixel-level model predictions and create polygons. On a test set, the construction starts model obtained results of significant enough quality to warrant continuation of this proof of concept further.

SUB-PROJECT: Ottawa COVID-19 Short-term hospital occupancy forecasts

A hierarchical Bayesian model was constructed to generate local COVID-19 hospital occupancy forecasts, based on COVID-19 daily hospital new admission counts and daily hospital midnight census counts. The two key phenomena being modelled are the random delay between (unobserved) infection and (observed) hospital admission, and that between hospital admission and (observed) discharge/death.

This proof-of-concept project was commissioned by Health Canada in September 2020. Sub-provincial COVID-19 hospital occupancy forecast could inform their decision making on resource deployment in Canada's pandemic response. The objective of this project was to determine the feasibility of generating such forecasts with sufficient accuracy based only on daily hospital admission and occupancy data, using Ottawa as a test case. This model was adapted from Flaxman *et al.* (2020).

Progress:

A model was developed to generate short-term forecasts of Ottawa COVID-19 hospital occupancy, based on Ottawa COVID-19 daily hospital new admission counts and midnight census counts. A modeling/forecast pipeline was implemented and deployed on Statistics Canada's Advanced Analytics Workspace (collaborative cloud computing infrastructure). Using open data reported for February 10, 2020 up to March 9, 2021, an efficacy test was carried out in the form of a series of mock forecasts, by truncating training data at 15 consecutive Mondays (November 23, 2020 – March 1, 2021). A model was fitted to each of the 15 truncated data sets, forecasts were generated accordingly for a period (up to 21 days) immediately following the training data truncation date, and the forecasts were compared with reported data. The model yielded satisfactory results, despite certain limitations, given the simplicity and limited scope of the observed data (Bosa and Chu, 2020). Potential subsequent work includes expansion to other jurisdictions, pending availability of local data.

SUB-PROJECT: Epidemiological modelling of COVID-19 for Personal Protective Equipment demand estimation

The SARS-CoV-2 (COVID-19) pandemic has put unprecedented demands on the Government of Canada to provide timely, accurate and relevant information to inform policy-making around a host of issues, including personal protective equipment (PPE) procurement and PPE deployment to the provinces and territories. Epidemiological models can be used to project the trajectory of epidemics, under different future assumptions, allowing policy-makers to consider a range of scenarios. One of the aims for this project is to improve the age-stratified dynamic compartmental model developed by the Public Health Agency of Canada (PHAC) in collaboration with Statistics Canada (Ludwig et al, 2020) by implementing additional pandemic scenarios. In this model, population is assigned to various compartments and flows through the model at a defined rate.

Progress:

A vaccine model was developed where vaccination phases were divided into 4 compartments: 1) first shot and did not develop immunity yet; 2) first shot and developed partial immunity; 3) second shot and still have partial immunity; 4) second shot and have developed maximum immunity. Flow from one vaccine compartment to the next is controlled by the number of available doses each province can distribute that day. In accordance with the National Advisory Committee on Immunization guideline (PHAC, 2021), the model simulates vaccination process of prioritizing seniors then moving to younger population, while allowing small number of doses to be distributed in middle age group to account for the effect of vaccinating the healthcare workers. Different flows from one compartment to the others were simulated taking efficacy of the vaccine, vaccination rate and level of immunity into account.

A model for variant-of-concern have been generated and tested to model the effect of a specific variant (B117) can pose in Canadian provinces. First, findings by Davies et al (2021) which reported that infectivity of B117 can be higher than the wildtype variant were added to the model. In addition, Challen et al (2021) reported that B117 increases the risk of hospital admittance (i.e. severe cases) and this information was also incorporated in the model.

For more information, please contact:
Saeid Molladavoudi (613-294-7418, saeid.molladavoudi@statcan.gc.ca**)**.

REFERENCES

Challen, R., Brooks-Pollock, E., Read, J.M., Dyson, L., Tsaneva-Atanasova, K. and Danon, L. (2021). Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. BMJ, 372.

Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., Pearson, C.A., Russell, T.W., Tully, D.C., Washburne, A.D. and Wenseleers, T. (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England. Science, 372(6538).

Flaxman, S., Mishra, S., Gandy, A. *et al.* (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature **584**, 257–261. https://doi.org/10.1038/s41586-020-2405-7

Ludwig A., Berthiaume P., Orpana H., Nadeau C., Diasparra M., Barnes J., Hennessy D., Otten A., and Ogden N. (2020). Assessing the impact of varying levels of case detection and contact tracing on COVID-19 transmission in Canada during lifting of restrictive closures using a dynamic compartmental model. Canada Communicable Disease Report, 46(11/12):409-421.

Public Health Agency of Canada (PHAC). (2021). Recommendations on the use of COVID-19 vaccines. National Advisory Committee on Immunization (NACI): Statements and publications. https://www.canada.ca/content/dam/phac-aspc/documents/services/immunization/national-advisory-committee-on-immunization-naci/recommendations-use-covid-19-vaccines/recommendations-use-covid-19-vaccines-en.pdf

Ronneberger, O., Fischer, P. and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

1.4 Record linkage

Record linkage plays an important role in the production of official statistics. However, it is susceptible to errors because it is often based on quasi-identifiers that are not unique and recorded with variations and typographical errors. This project looks at the production and use of linked data, including the accurate estimation of linkage errors.

SUB-PROJECT: Estimation of linkage errors

The accurate estimation of linkage errors is critical for applications, including the estimation of the number of false negatives due to blocking. This project estimates these errors without any clerical reviews with a model extending that by Blakely and Salmond (2002), to account for the record heterogeneity (see the details in Dasylva and Goussanou, 2020). This model involves a finite mixture, where each component is the sum of a Bernoulli variable with an independent Poisson variable. The estimation is based on the numerical maximization of a composite likelihood.

Progress:

The model has been applied to estimate the number of false negatives due to blocking when linking two duplicate-free sources including a file and a register or census with nearly complete coverage. The proposed methodology was evaluated successfully in an empirical study, which is described in a paper to appear in *Survey Methodology* (Dasylva and Goussanou, 2021). It is of interest when linking tax or mortality data to the census (Blakely and Salmond, 2002; Statistics Canada, 2017).

A new estimation procedure has been developed to address some limitations of the previous one. Indeed, that procedure used expectation-maximization and was slow to converge. Besides, the number of mixture components had to be selected manually. Instead, the new procedure uses a faster Newton-Raphson procedure and selects the number of components automatically based on Akaike's information criterion. It also provides likelihood ratio confidence intervals based on a subset of the observations. This procedure may be used to estimate the error rates at the end of a linkage, regardless of the selected method (e.g. probabilistic, deterministic, hierarchical, etc.), provided that the decision to link two records depends on no other record. With the probabilistic method, it may also be used during the development to automatically select the linkage weights and the thresholds (given the target error rates), without having to specify how the variables are correlated. Thus, it greatly simplifies the error estimation process and dispenses with conditional independence assumptions that are a potential source of correlation bias (Blakely and Salmond, 2002; Newcombe, 1988, p. 149).

For more information, please contact:
Abel Dasylva (613-408-4850, abel.dasylva@statcan.gc.ca).

REFERENCES

Blakely, T., and Salmond, C. (2002). "Probabilistic record linkage and a method to calculate the positive predicted value", *Journal of Epidemiology*, 31, 1246-1252.

Newcombe, H. (1988). *Handbook of Record Linkage*, Oxford University Press.

Statistics Canada (2017). *2016 Census of Population Income Reference Guide*. Catalog no. 98-500-X2016004.

1.5 Qualitative research

SUB-PROJECT: Perceptions of data sensitivity

The Questionnaire Design Resource Centre (QDRC) has an ongoing connection to Canadians when conducting cognitive interviews. This connection is a great vehicle to collect information about the perceived sensitivity of various survey and non-survey data sources and to contribute in part to the sensitivity scale project within the Necessity and Proportionality Framework.

Progress:

A first draft of a short questionnaire was developed and administered to people taking part in the QDRC's usual cognitive interviews. Following this first phase, some early survey results were obtained, an internal

report was written (Reicker, 2020) and changes to the questionnaire were recommended. The updated questionnaire was then used in collection for a brief second phase. Results of that second phase should be reported in 2021.

For more information, please contact:
Paul Kelly (613-371-1489, paul.kelly2@statcan.gc.ca).

2 Confidentiality / Access

Confidentiality research at Statistics Canada continues to focus on developing new methods and ideas that offer alternative forms of access while continuing to ensure that personal individual and business information is not disclosed in any way. The Centre for Confidentiality and Access group at Statistics Canada also continues to offer consultation services to internal and external partners as a way to help develop capacity in disclosure risk identification and treatment.

2.1 Perturbation-based methods

SUB-PROJECT: Random Tabular Adjustment

Random Tabular Adjustment (RTA) is a disclosure control method that relies on adding random noise to estimates rather than suppressing them. The primary focus is to avoid suppression for continuous variables collected under economic surveys.

Progress:

The data of the Annual Survey of Research and Development in Canadian Industry (RDCI) was released in December of 2020, using the RTA method. This release built upon the first successful use of the RTA method, in 2019, for the release of the data of the Survey of Innovation and Business Strategy (SIBS). For 2020, the main focus continued to be on developing ideas to allow RTA to be applied on a wider variety of products at Statistics Canada as the agency moved towards decreasing its dependence on suppression strategies.

The main challenge being investigated was adding a correlated noise function. This would have several benefits including: maintaining correlation structures between related variables; minimizing the impact that noise addition has on aggregate cells by adding negative noise to related cells; and preserving trends found in repeated surveys. A SAS prototype has been developed, but there are optimization issues with this strategy as the mathematical problems can quickly explode to a point where it is not practical. Future work will look at some compromises that may allow the problems to be solved more easily, and application for small programs and studies using higher powered computing environment such as cloud technology and high-powered servers.

SUB-PROJECT: Use of perturbation for Census Public Use Microdata Files (PUMF)

The potential application of the post-randomization method (PRAM) (Gouweleeuw et al. 1998) within the disclosure control protection strategy of the Canadian Census PUMFs was investigated in an attempt to reduce the number of suppressions for certain key variables and increase the overall utility of the files. PRAM uses perturbative methods, rather than suppression techniques to protect the data, via various swapping of variable values.

Progress:

Through extensive investigations and analysis, several options of implementing PRAM were explored. One preferred approach was further developed. It could allow for improved utility of the PUMF files using

PRAM. Given the confidential nature of the protection methods applied to the PUMFs, further details will not be discussed, such as how specifically it could be applied or whether in fact PRAM will be applied for the Canadian Census PUMFs.

For more information on confidentiality, please contact:
Steven Thomas (613-882-0851, Steven.Thomas@statcan.gc.ca).

REFERENCE

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and de Wolf P.-P. (1998) Post *Randomisation for Statistical Disclosure Control: Theory and Implementation,* Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 463-478

2.2 Expanding access to business data

SUB-PROJECT: Public-Use Business Files

Public Use Microdata Files (PUMFs) are one of Statistics Canada's many data access solutions. The original microdata is transformed through confidentiality methods in order to reduce the risk of disclosure enough for it to be safe to release to the general public. Statistics Canada has never released a PUMF for a business survey. There are factors pertaining to business-level data that afford it an extra level of confidentiality protection in comparison to person-level data. A paper describing the techniques along with the risk and utility challenges was developed (Gilchrist, 2020) with an application to an economic database.

SUB-PROJECT: Research Data Centre Access to Business Files

Statistics Canada is researching and developing access solutions to allow researchers access to real microdata. This is a useful scenario for researchers while it poses a certain access risk if not implemented strategically. The Research Data Centre (RDC) program (provides a safe setting for vetted researchers where outputs are vetted carefully. Methods are being developed to allow access to business data through the RDC program and a general output vetting strategy is being developed. This process was previously handled by the Canadian Centre for Data Development and Economic Research (CDER) program.

For more information on confidentiality, please contact:
Steven Thomas (613-882-0851, Steven.Thomas@statcan.gc.ca).

2.3 Synthetic data

The Government of Canada's Directive on Open Government aims to ensure Canadian's get access to the most government information and data as possible. One solution for open data is synthetic data. A synthetic version of a database would address confidentiality issues with personal data while retaining as much analytical value as possible. The methods used by Statistics Canada in creating synthetic data has been documented by Kenza Sallier as the winning submission to the 2020 Young Statisticians' Prize (Sallier, 2020).

One of the challenges with synthetic data is the terminology and nomenclature used for describing the various types of synthetic data along with the various methods and tools available. A guide was last developed in 2002 and is being revised to include many of the modern methods that are focused on synthetic files preserving analytical content rather than creating simple dummy files. This guide may contribute or rely on some of the work being developed by the UNECE and the High-level Group for the Modernisation of Official Statistics. Statistics Canada has also contributed to this group by sharing its experiences (Sallier, 2021).

Finally, efforts were dedicated to the study of other tools and methods that could complement or replace those used until now. A study comparing the R packages Synthpop with SimPop was carried out (Zhao, 2021).

As Statistics Canada acquires more expertise in producing synthetic data files of high analytic value, we begin to tackle new challenges with the synthesis of data that preserves hierarchical structures in the form of families where records are linked and share common traits that must be preserved. These challenges also arise when synthesizing other structured data such as business data. A paper is in development that will concentrate on a real-world example of applying this strategy to income data for families.

For more information, please contact:
Steven Thomas (613-882-0851, Steven.Thomas@statcan.gc.ca).

2.4 Cryptography

SUB-PROJECT: Private text classification on scanner data using Homomorphic Encryption

With the imminent closing of data centres at Statistics Canada, the pressure is on to figure out how to migrate projects involving sensitive data to the cloud. One solution is to use privacy-preserving techniques to protect the private data cryptographically. One such technology is Homomorphic Encryption, which allows one to perform computations on encrypted data without first decrypting it. One of StatCan's sensitive data sources is the so-called scanner data, which comes from major retailers and is used to calculate the Consumer Price Index, among others. The machine-learning-assisted pre-processing step of classifying product descriptions into North American Product Classification System codes needs to be performed on the cloud, and to protect the sensitive descriptions they need to be encrypted.

Progress:

Two proof of concept projects were started and completed successfully this year. Research was performed on the feasibility of implementing both statistical analysis (compute mean, totals and variance) and neural network modeling (training a model to classify products and generate products prediction) using scanner data. A compute model was designed and a program implementing the model has been produced (Zanussi, Santos and Molladavoudi, 2021). In the future, we want to evaluate the feasibility of encrypted processing for record linkage using private set intersection. When all partners are ready, the project can be extended and brought into production.

For more information on this sub-project, please contact:
Zachary Zanussi (613 298-1808, zachary.zanussi@statcan.gc.ca).

3 Theory and framework

3.1 Inference from non-probability samples

There is a growing interest in National Statistical Offices to produce Official Statistics using non-probability sample data, such as big data or data from volunteer web surveys. Indeed, Statistics Canada has recently conducted several online volunteer surveys, called crowdsourcing surveys, to evaluate the impacts of the COVID-19 pandemic on different aspects of the life of the Canadian population. The main motivation for using non-probability samples is their low cost and respondent burden, and quick turnaround since they allow for producing estimates shortly after the information needs have been identified. However, non-probability samples are well known to produce estimates that may be fraught with significant selection bias. Beaumont and Rao (2021) discuss this important limitation, along with an illustration, and describe some remedies that involve the integration of data from the non-probability sample with data from a probability sample. Renaud and Beaumont (2020) describe four recent research initiatives to leverage non-probability sample data.

How to obtain meaningful estimates and make valid inferences from non-probability samples is an important question that still requires research and experimentations. The following four sub-projects attempted to address this question.

SUB-PROJECT: Bias reduction of non-probability sample estimators through propensity score weighting methods with application to crowdsourcing data

The goal of this project is to study and develop propensity score weighting methods that combine data from a non-probability sample that contains variables of interest and auxiliary variables with data from a probability sample that contains the same auxiliary variables (or a subset of them). Propensity score weighting involves modelling the probability of participation in the non-probability sample.

Progress:

We first considered a logistic model (see Chen, Li and Wu, 2019) and developed a variable selection procedure based on a modified Akaike Information Criterion (AIC). Our modified AIC properly accounts for the data structure and the possibly complex probability sampling design. We also developed a simple method of forming homogeneous post-strata. Moreover, we extended the Classification and Regression Trees (CART) algorithm (Breiman, Friedman, Stone and Olshen, 1984) to this data structure, and developed a pruning procedure that again properly accounts for the probability sampling design. Our new algorithm is called nppCART. Some details about the pruning procedure are given in Beaumont and Chu (2020). We also developed a bootstrap variance estimator that reflects two sources of variability: the probability sampling design and the participation model. Our methods are currently being evaluated using crowdsourcing data and Labour Force Survey (LFS) data. A draft paper has been written, which is close to being completed.

SUB-PROJECT: An approximate Bayesian approach to improving probability sample estimators using a supplementary non-probability sample

A Bayesian method of combining data from a probability and non-probability sample was recently published in the *Journal of Official Statistics* (Sakshaug, Wisniowski, Ruiz, and Blom, 2019). These authors dealt with the estimation of model parameters when the dependent variable *y* and the vector of explanatory variables **x** are observed in both samples. They used the non-probability sample as a means of determining the prior mean for the model parameters under the assumption that the probability sampling design is ignorable. The goal of this project is to extend their method to the estimation of finite population parameters, under a possibly non-ignorable probability sampling design, and see if we can obtain model-based estimates that are more efficient than standard survey-weighted estimates.

Progress:

We have made the extension to the estimation of a finite population mean under a non-ignorable probability sampling design. We have also conducted preliminary simulation experiments that have been summarized in You (2021c). We plan to complete the simulation study and write a paper that will be presented at the 2021 Statistics Canada's Symposium.

SUB-PROJECT: Mean square error estimation for non-probability sample estimates using small area estimation techniques

Administrative data and other non-probability sources of data are being increasingly considered by National Statistical Offices as a means of obtaining directly the information sought on a population. However, estimates from these non-probability sources are often subject to a number of errors, and there is a need to develop indicators of their accuracy.

Progress:

We developed an estimator of the conditional Mean Square Error (MSE) of non-probability sample estimates applicable when a probability sample with the same variables is also available. Our conditional MSE estimator is obtained by making use of Small Area Estimation techniques. We have evaluated the method using data from the Longitudinal Social Development Data program (LSDDP). The LSDDP file is constructed from administrative files and allows for the estimation of labour force characteristics. The Labour Force Survey (LFS) is used as the probability sample. We have started writing an article that summarizes the findings.

SUB-PROJECT: Statistical data integration using a prediction approach

We consider the problem where a non-probability sample is available that contains a vector of auxiliary variables, *x*, for each sample unit. We assume that this non-probability sample covers a significant portion of the population. A probability sample is also available that contains *x* as well as the variable of interest *y* for each sample unit. The indicator of participation in the non-probability sample is available in the probability sample. This scenario is relevant to a survey on postal traffic conducted by La Poste in France. Alain Dessertaine proposed a predictor for that scenario. We developed variance estimators, including a bootstrap variance estimator, for evaluating the quality of the proposed predictor. The details are given in an internal draft report.

Progress:

The collaboration with La Poste, Toulouse School of Economics and the university of Besançon continued and the objective is to write a joint paper. The results of this project will first be presented in an invited session at the Colloque francophone sur les sondages in the fall 2021.

For more information, please contact:
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

REFERENCES

Breiman, L., Friedman, J.H., Stone, C.J. and Olshen, R.A. (1984). *Classification and regression trees*. CRC Press.

Chen, Y., Li, P. and Wu, C. (2019). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association* (published online).

Sakshaug, J.W., Wisniowski, A., Ruiz, D.A.P. and Blom, A.G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statsitics*, 35, 653-681.

3.2 Machine Learning framework

Statistics Canada uses machine learning techniques to solve large-scale data problems. At the same time, national statistical institutes are facing unprecedented pressure to demonstrate to citizens, businesses and users that they are trustworthy and transparent institutions. Statistics Canada has developed a framework for the responsible use of machine learning techniques, including guidelines for constructing and implementing ethically and methodically sound processes. This framework is built on four themes—respect for people, respect for data, sound application and sound methods —and a number of attributes for each theme. It is aligned with the Directive on Automated Decision-Making and its Algorithmic Impact Assessment Tool, developed by the Treasury Board Secretariat (2020).

Progress:

Following testing and a review by management, the framework for responsible machine learning processes was formally adopted by Statistics Canada in July 2020. A checklist accompanying the guidelines has been finalized and is used to help assess responsible machine learning processes. A process for evaluating machine learning applications that will go into production has been developed, which includes an independent review by experts, adequate documentation and completion of the project-related checklist. The project methodology as well as the review carried out by the experts are presented to a scientific review committee which will issue recommendations on the proposed methodology. In the past year, seven projects were evaluated using this framework. Next steps for this project include: developing a dashboard for recording reviews, developing a template of documentation that should be provided to reviewers, and a new literature review regarding the concepts of explainability and interpretability to be up to date in this area.

For more information, please contact:
Keven Bosa (613-863-8964, keven.bosa@statcan.gc.ca).

3.3 Quality indicator research

In order to provide users with quality indicators for programs that combine administrative data sources, the Quality Secretariat has initiated work to develop a composite indicator that combines quality indicators related to different stages of data processing (record linkage, imputation, geocoding, etc.) into a single indicator. The objective is to give a global view of the quality of an estimate by taking into account several factors that can introduce errors. A first program will publish these indicators along with the estimates in the summer of 2021. This project will be presented at the European Establishment Statistics Workshop in September 2021 (Beaulieu and Lebrasseur, 2021).

For more information, please contact :
Martin Beaulieu (613-854-2406, martin-j.beaulieu@statcan.gc.ca).

4 Support (Resource Centres)

4.1 Record Linkage Resource Centre

The objectives of the Record Linkage Resource Centre (RLRC) are to provide consulting services to both internal and external users of record linkage methods, including recommendations on software and methodology and collaborative work on record linkage applications, to evaluate alternative record linkage methods and develop improved methods. We evaluate software packages for record linkage and, where necessary, develop prototype versions of software incorporating methods not available in existing packages and assist in the dissemination of information concerning record linkage methods, software and applications to interested persons both within and outside Statistics Canada.

Progress:

We continued to provide the development team of G-Link with support and follow up on any raised issues. The RLRC also provided internal and external G-Link users with support when help/comments/suggestions regarding G-Link were sought through requests at G-Link_info.

During the year, much of methodology's work revolved around the development and the support of users of the new G-Link version (Version 3.5), which included the addition of profile-based linkage, identification and treatment of orphan records and integrated pseudokeys. Additionally, work has been done to explore record linkage in the cloud using the platform Databrics and to integrate a clerical review tool (quality assessment) as well as correct and improve some threshold estimators.

The RLRC also worked on a variety of other record linkage-related projects during the year, including holding more instances of the Record Linkage Forum.

Our record linkages helped us document performance and issues pertaining to management and developers and was used as an opportunity to field test new G-Link 3.5 features and develop more systematic and theoretically coherent approaches of defining and adjusting record linkages under servers and SAS Grid. The RLRC updated the tutorial and the user guide of G-Link 3.5.

For more information, please contact:
Abdelnasser Saïdi (613-863-7863, abdelnasser.saidi@statcan.gc.ca).

4.2 Generalized Systems

The Economic Generalized Systems team is responsible for the support and development of four Generalized Systems, namely G-SAM – the generalized sampling system, BANFF – the generalized system for edit and imputation, G-EST – the generalized system for estimation and G-SERIES- the generalized system for time series techniques.

Progress:

A large volume of support cases were processed by the project team, mainly on G-EST, BANFF and G-SAM. Most of these were resolved with suggestions on how to apply the systems properly. However, several required more involvement. Two new versions of the generalized systems were released this year. G-EST 2.03.002 was released including performance improvements for calibration and parallel processing for sampling variance (Statistics Canada, 2020). BANFF version 2.08 was also released including improvements to the BANFF processor and bug fixes to improve error localization (Statistics Canada, 2021).

Two support cases were identified that involved processing issues with large complex datasets in applying SEVANI (part of G-EST) to estimate variance due to imputation. Consequently, development work was done to build a prototype to allow for unnecessary calculations to be omitted for specific cases, and for parallel processing. These prototypes are being tested and will be part of a future release of G-EST.

The development of ImpACT, a system to visualize imputation was continued, and the system was used to evaluate imputation strategies from two Statistics Canada projects. This work was presented to the Scientific Review Committee of the Modern Statistical Methods Branch (Gray, 2020a). Expansion of this tool to include further visualisations, and application to other programs is planned for the future.

Members of the team participated in training through formal courses with Statistics Canada's training institute, as well as seminars for recently recruited statisticians and other ad hoc presentations to analysts and other organisations. The team also contributed to the organisation of the UNECE data editing workshop, and presented work on the ImpACT tool (Gray, 2020b). This presentation generated interest from international colleagues and is expected to lead to collaborations.

A major initiative for the generalized systems is a long-term planning exercise to consider the upcoming evolution of the systems, including the possibility of using open-source tools, inclusion of modern methods, and revisiting the approach to manage future developments. As part of this work, a number of over-arching requirements for Generalized Systems were outlined (Matthews, 2020a) and work was begun to define the business requirements for each individual system in the short, medium and long term. These requirements will feed directly into the long-term evolution plan for each system that will be elaborated over the next year in partnership with the Informatics Technology team (Matthews, 2021a).

For more information, please contact:
Steve Matthews (613-854-3174, Steve.Matthews@statcan.gc.ca).

4.3 Questionnaire Design Resource Centre

The Questionnaire Design Resource Centre (QDRC) is a focal point of expertise at Statistics Canada for questionnaire design and evaluation. The QDRC provides consultation and support services, and carries out projects and research related to the development, testing and evaluation of survey questionnaires. The QDRC plays a very important role in quality management and responds to program requirements throughout Statistics Canada by consulting with clients, respondents and data users and by pre-testing survey questionnaires.

While much of the QDRC's work is carried out on a cost-recovery basis, the Center is frequently approached on an ad hoc basis for expert reviews and consultation services on a wide variety of surveys. The group also offers courses on questionnaire design.

Progress:

The QDRC conducted many reviews of survey questionnaires. While most of these involved Statistics Canada questionnaires, several were conducted for surveys being done by other government organizations such as the Bank of Canada, the Office of the Superintendent of Bankruptcy, the Canadian Office of the Ombudsman for Responsible Enterprise, Global Affairs Canada and others.

In response to the very quick questionnaire development schedules for new COVID-19 related surveys, the QDRC developed and implemented a new qualitative research panel that allowed for rapid testing of these new data collection tools.

The QDRC continued to experiment with mixed method research. As well some research and experimentation began with asynchronous qualitative methods. The group also contributed to various corporate consultation initiatives.

For further information, please contact:
Paul Kelly (613-371-1489, paul.kelly2@statcan.gc.ca).

4.4 Quality Secretariat

The Quality Secretariat's mandate includes designing and managing quality management studies and responding to requests for quality management information or assistance from Statistics Canada's various programs or other organizations.

SUB-PROJECT: Capacity building with internal, national and international partners

The Quality Secretariat's objective is to provide advice and undertake capacity-building measures internally, with national partners (other departments or others) and international partners, primarily by giving a general overview of Statistics Canada's quality management practices and official quality-related documents (the Quality Assurance Framework and the Quality Guidelines) and by providing quality management support services.

Progress:

The Quality Secretariat undertook capacity building for many partners during the reporting period. Internally, training workshops were offered through various courses for staff. At the national partner level, formal presentations on quality management practices were made to three organizations, in addition to holding a number of workshops and seminars. Material on data quality and good quality management practices was provided to Statistics Canada's Data Literacy Training Initiative. Discussions occurred within the Data Governance Standardization Collaborative and the Data Quality Working Group. The latter group, co-chaired by Statistics Canada, aims to define a data quality framework applicable to all Government of Canada organizations as part of the implementation of the Data Strategy. The validation of the quality of

a statistical process carried out as well as the validation of the quality of a data source used by another federal agency has also been completed. At the international level, involvement as the United Nations Expert Group on National Quality Assurance Frameworks continued in preparation for the implementation of the United Nations National Quality Assurance Framework Manual for Official Statistics (United Nations, 2019).

SUB-PROJECT: Quality indicators for statistics from integrated data**.**

Details on this sub-project can be found in Section 3.3 (Quality indicator research).

For more information, please contact:
Martin Beaulieu (613-854-2406, martin-j.beaulieu@statcan.gc.ca).

REFERENCE

United Nations (2019). United Nations National Quality Assurance Frameworks Manual for Official Statistics. https://unstats.un.org/unsd/methodology/dataquality/un-nqaf-manual/.

4.5 Quality Assurance Resource Centre

The objectives of the Quality Assurance Resource Centre (QARC) are to conduct research and development activities on statistical methods of quality assurance and control, aimed at improving the outgoing quality of survey data collection and processing operations within the bureau. This includes offering methodological services for G-Code which is used at Statistics Canada to create coding databases for data processing. Research on quality assurance and control is often generic in nature and involves issues of efficiencies and automation that are frequently applied to many steps of survey operations.

Progress:

The methodological support team helped the development team and tracked user inputs to help identify ideas for potential improvements for G-Code. The QARC also provided internal and external (international users) G-Code users with support when help/comments/suggestions regarding G-Code was needed.

During the year, work revolved around the implementation of a new version of G-Code (Version 3.2), which included the addition of machine learning capabilities (XgBoost and FastText). The QARC team has been involved in a coding and classification proof of concept looking at the integration of the FastText algorithm into our Generalized Coding tool (G-CODE). The new algorithms have been widely used to code industry and occupation for the Business Register, the Labour Force Survey and many other programs/surveys (including Census of Population). Additionally, these new functionalities have been presented to external agencies (Australian Bureau of Statistics and Statistics New Zealand). Lately, the QARC team has been helping with the integration of Pytorch into G-Code. PyTorch is an open-source machine learning library based on the Torch library, used for applications such as computer vision and natural language processing, primarily developed by Facebook's AI Research lab.

The QARC also worked on implementing a variety of quality controls on the coding processes for the Labour Force Survey, Canadian Community Health Survey, Job Vacancy and Wage Survey and the Business Register.

For further information, please contact:
Javier Oyarzun (613-302-8454, javier.oyarzun@statcan.gc.ca).

4.6 Data Analysis Resource Centre and consultation

The main goal of Data Analysis Resource Centre (DARC) is to give advice on the appropriate use of data analysis tools and methods, and to promote best practices in this area. DARC's services - which focus mainly on survey, census or administrative data - are available to the employees of the Agency or other departments, as well as to analysts and researchers from academia or the Research Data Centres (RDCs).

Progress:

Consultations

Consultation services were provided as requested by internal and external clients. The questions varied in complexity and included use of survey bootstrap weights, *p*-values, constructing confidence intervals and testing hypotheses with survey data, estimating *effect size* with survey data, analysis with linked data, fitting logistic regression, etc. We also helped our clients with implementation of methods in SUDAAN, SAS, STATA and R software.

Provision of Training and Training Material

DARC developed and presented Session 1 of *Statistical Modelling Course*, entitled "Statistical Modelling with Complex Survey Data – Linear Regression" (Mach and Michaud, 2020). This new course was offered virtually by the Statistical Talent Development Working Group to employees of Modern Statistical Methods and Data Science Branch of Statistics Canada.

The document "Data Visualization - Best practices" (DARC, 2020a) was finalized and is available internally in both official languages. It is intended to be a tool for Statistics Canada analysts and dissemination teams.

Collaboration

We collaborated in developing measurement strategies for three projects: i) Workplace Mental Health Performance Measurement Project, ii) Beyond2020 Public Service Renewal Index, and iii) Accessibility Strategy for the Public Service of Canada.

All three projects use data collected by the Public Service Employee Survey (PSES) to measure latent variables like psychological risk factors, behaviors, etc. The measurement models were developed using factor analysis and structural equation modelling as discussed by DARC (2020b) and Blais et al. (2020 and 2021).

For further information, please contact:
Harold Mantel (613-863-9135, harold.mantel@statcan.gc.ca).

4.7 Time Series Research and Analysis Centre

The objective of the time series research is to maintain high-level expertise and offer needed consultation in the area, to develop and maintain tools to apply solutions to real-life time series problems as well as to explore current problems without known or acceptable solutions.
The projects can be split into various sub-topics with emphasis on the following:

- Consultation and training in Time Series (including course development and delivery)
- Support and Enhancement of Time Series Processing System
- Development and Support for Seasonal Adjustment and Trend-Cycle Estimation
- Modelling and Forecasting, particularly in the context of real-time estimation

SUB-PROJECT: Consultation and training in Time Series Methods

The Time Series Research and Analysis Centre is responsible for developing and delivering training on time series methods including seasonal adjustment, reconciliation and time series modelling to participants internal to Statistics Canada as well as those from other agencies. In addition, the Centre provides guidance and consultation on time series projects in general for Statistics Canada, including requests coming from within and outside of the agency.

Progress:

With the sudden change to remote working, a reduced number of courses were delivered. However, those that were delivered required modifications to the content and organisation of the courses to accommodate shorter but more frequent sessions to deliver the material. Specifically, courses on benchmarking (H-0436), raking (H-0437) and seasonal adjustment (H0434) were delivered during the year via the Statistics Canada training centre (Statistics Canada, 2021). In addition, a course on time series modelling and forecasting (H-0433) was delivered to participants from Immigration, Refugees and Citizenship Canada via a remote format. Members of the Centre also participated in outreach and training to other groups in Statistics Canada on time series topics as part of training for recent recruits and financial officers.

The Centre also participated in review of papers for refereed journals related to the COVID-19 pandemic, in particular on forecasting of case counts for selected countries, and informal discussions with statisticians involved in statistical modelling related to the pandemic. Members of the Centre also participated in the United Nations Economic Commission for Europe - High Level Group on Machine Learning, contributing to a comparison of machine learning and traditional forecasting methods (Picard, 2020). The Centre also reviewed the methodology behind several new statistical products released by Statistics Canada, namely the provincial economic indicator (Statistics Canada, 2021b), and excess mortality estimates (Statistics Canada, 2020) to validate time series features in the development of these indicators.

SUB-PROJECT: Support and Enhancement of Time Series Processing System

The Time Series Processing System is a customizable SAS-based application to apply time series techniques including file validations, application of revision strategies as well as seasonal adjustment and reconciliation techniques used extensively in the production of seasonally-adjusted estimates for mission critical programs within Statistics Canada. Developed over 10 years ago, the system is in a fairly mature and stable state. However, it requires updating on an ongoing basis to broaden functionality and address new needs of programs in the agency. In the longer term, a new version of the system may need to be developed to support processing and allow flexibility to use new techniques available from open-source software.

Progress:

Minor fixes were applied to version 3.08 of the Time Series Processing System (TSPS) described in Ferland (2019) as well as to a number of supporting tools for analysis and system development. Notably, a tool that is used as a key component in quality assurance was expanded to include improved diagnostics on residual seasonality and relative roughness statistics at different stages in the seasonal adjustment process to allow for more efficient trouble-shooting and development of solutions.

Investigations were made to evaluate tools available to apply benchmarking through open-source tools, including those available in R. A comparison was made between the current functionality of G-Series, and an R package called tempdisagg, in terms of available methods and their parametrization (Picard, 2021). Statistics Canada joined the Seasonal Adjustment Centre of Excellence of Eurostat as a partner organization to participate in discussions and development of related tools.

SUB-PROJECT: Development and Support for Seasonal Adjustment and Trend-Cycle Estimation

Analysis and evaluation of new methods and techniques for seasonal adjustment as well as consultation and centralization of expertise in applying seasonal adjustment.

Progress:

The Time Series Research and Analysis Centre provided extensive support for seasonal adjustment in order to ensure the quality of results during the unprecedented economic shocks due the COVID 19 pandemic. Many exchanges were held with representatives from national statistical offices, either via email exchanges, one-on-one conversations, or videoconferences in small groups. The exchanges included members from Eurostat, the Office for National Statistics, the United States Census Bureau and the Bureau of Labor Statistics, Statistics Norway, INSEE (Institut national de la statistique et des études économiques), Statistics Israel, the Australian Bureau of Statistics, Statistics New Zealand and others. These exchanges were extremely valuable to compare and contrast approaches for the short and long term, and were almost universally in line with the approach suggested by Eurostat in Eurostat (2020). The Centre shared findings and recommendations from these consultations with relevant contacts within the agency through periodic updates to ensure that the information was widely available.

For the seasonal adjustment of programs that are directly supported by the Centre, a quality assurance strategy for seasonal adjustment was developed and implemented including increased communication with the subject-matter analysts to ensure that important information was shared with subject-matter experts to provide guidance on seasonal adjustment. In addition, the strategy for annual review of

seasonal adjustment parameters was developed for each program, taking into account the timing of the review, as well as the extent of shocks in the time series. This overall strategy was documented and shared with other seasonal adjustment practitioners during a round-table discussion organized by the government statistics section of the American Statistical Association, see American Statistical Association (2021).

The Time Series Research and Analysis Centre also provided extensive consultations on seasonal adjustment within the agency for programs not formally supported by the team. Consultations were provided to groups producing seasonally adjusted statistics on trade in services, tourism, trade and exporter characteristics, the business register entry and exits counts, and various components within the system of national accounts. Explorations were done to seasonally adjust statistics from other programs. Notably, statistics on railcar loadings were analysed and scenarios to produce seasonally adjusted estimates for some high-level time series were suggested. A similar analysis was done for electricity generation. In both cases, many seasonal series were identified and a decision is pending on whether the seasonally adjusted estimates will be produced for dissemination.

Continued progress was made on making the Seasonal Adjustment Dashboard available to analysts for individual programs, with two mission critical surveys now able to access the tool to understand and explain seasonally adjusted results, and a plan to increase this to four more mission critical surveys in the first quarter of the coming fiscal year. The dashboard is presented in Matthews (2019).

As well, trend-cycle estimation methods were evaluated in the context of the COVID-19 pandemic. In particular, given the sharp nature of the economic shocks, the trend-cycle may present an overly smooth impression of the economy so a number of alternative measures were identified including breaking the series at a particular point, and introducing outlier effects to model shocks more directly. These methods will be further evaluated taking into account their advantages and disadvantages for potential application in some programs.

SUB-PROJECT: Modelling and Forecasting, particularly in the context of real-time estimation

Details on this sub-project can be found in Section 1.2 (Real-time estimation via time series methods).

For more information, please contact:
Steve Matthews (613-854-3174, Steve.Matthews@statcan.gc.ca).

REFERENCES

Matthews, S. (2019). De-Mystifying Seasonal Adjustment: A visual tool to understand the process. Presentation to the Seasonal Adjustment Practitioners Workshop, United States Census Bureau.

Eurostat (2020). Guidance On Time Series Treatment In The Context Of The Covid-19 Crisis. https://ec.europa.eu/eurostat/documents/10186/10693286/Time_series_treatment_guidance.pdf.

Ferland, M. (2019). What's new in the Time Series Processing System – v3.08. Internal Document, Statistics Canada.

Statistics Canada, (2020). Excess mortality in Canada during the COVID-19 pandemic. https://www150.statcan.gc.ca/n1/pub/45-28-0001/2020001/article/00076-eng.htm.

Statistics Canada (2021). Workshops, training and references. https://www.statcan.gc.ca/eng/wtc/training.

Statistics Canada (2021b). Experimental indexes of economic activity in the provinces and territories, December 2020. https://www150.statcan.gc.ca/n1/daily-quotidien/210413/dq210413d-eng.htm.

4.8 Confidentiality

Part of Statistics Canada's role and responsibility continues to be outreach and support for confidentiality strategies. Some of the activities carried out throughout the year include consultation with Employment and Social Development Canada (ESDC) on Pay Transparency Statistics, presentation to the G20 DGI-2 workshop on access strategies, and Confidentiality workshops for Health Canada and ESDC.

Other research and development activities related to this topic are described in section 2.

For more information, please contact:
Steven Thomas (613-882-0851, Steven.Thomas@statcan.gc.ca).

4.9 Data Science Communities of Practice

SUB-PROJECT: Machine Learning Community of Practice

The Statistics Canada Machine Learning Community of Practice has the goal of facilitating collaboration and knowledge transfer as well as improving our machine learning operations at Statistics Canada.

Through various activities pertaining to machine learning bringing together 50 to 80 people, such as lunch-and-learns, presentations, reading groups, viewing groups and information-sharing on a site developed and updated by the members, the Community is, through its active presence, still collaborating in the development of the machine learning capacities of Statistics Canada's employees.

Progress:

Despite remote work, the Community organized a number of presentations on several areas of machine learning, such as a series of four presentations on machine learning applications developed by Statistics Canada to help frontline agencies assess and prepare for COVID-19 spread scenarios. The Community also continued the activity of viewing free online machine learning courses, enabling participants to discuss the topics covered after the viewing. The Community has updated the list of existing methodology projects that explore or use machine learning and continues to collaborate with the other communities of practice, including a new collaboration with the Citizen Development Working Group. Finally, the Community continues to provide information to its members on a variety of relevant external activities related to data science.

SUB-PROJECT: Machine Learning Text Analysis Community of Practice (CoP)

Machine Learning Text Analysis CoP is a centralized inter-departmental place for practitioners of various expertise to discuss practical applications of Natural Language Processing (NLP). Various practitioners across GoC come together to learn, discuss and adopt ethical applications of NLP. Monthly meetings bring together about 50-60 attendees to share each other's solutions and problems. About half of the participants are from federal departments outside Statistics Canada.

Progress:

Throughout 2020-2021, practitioners from different fields of Statistics Canada or from other department such as the Office of the Superintendent of Financial Institutions, Canada Border Services Agency, Immigration, Refugees and Citizenship Canada or Canada Revenue Agency presented their high-quality NLP solutions. Each presenter illustrated his or her modern methodology to quickly process their data source whether that be survey data, administrative data and public reports. Discussions after the presentation broke down the complex concepts for attendees to comprehend. The attendees came from over 20 different departments.

COVID-19 and the move to teleworking was an unexpected setback but we capitalized this as an opportunity to strengthen our teleconferencing framework. This invited a more inclusive inter-departmental community.

For more information, please contact:
Yanick Beaucage (613-854-2397, Yanick.Beaucage@statcan.gc.ca).

5 Divisional research and other activities

5.1 Economic Statistics Methods Division

SUB-PROJECT: Longitudinal Analysis for the project "Toking 9 to 5? Examining the Impact of Cannabis Legalization on Workplace Cannabis Use and Perceptions among Canadian Workers"

Toking 9 to 5? Examining the Impact of Cannabis Legalization on Workplace Cannabis Use and Perceptions among Canadian Workers is a four-wave multiple-cohort longitudinal study from the Canadian Institutes of Health Research. One of the main objectives of that study is to examine the impact of cannabis legalization on longitudinal patterns of workplace cannabis consumption, perceptions of risk, norms, and availability; and whether those trends in patterns differ by age, sex, labour market gender roles, occupational groups, and geography. The idea of Carrillo & Karr (2013) is appealing for tackling this issue since it proposed a method for marginal analysis of multiple-cohort surveys. However, the work of Carrillo & Karr was limited to *probabilistic* surveys; whereas the Toking 9 to 5 survey is *non-probabilistic*. Therefore, an evaluation/modification of said method for non-probabilistic surveys is required before it can be applied to the Toking 9 to 5 data.

Progress:

As there are no survey weights for the *Toking 9 to 5* survey, the first step of the project is to create pseudo weights that can be used with the method of Carrillo & Karr (2013). We reviewed some of the literature available for the creation of pseudo-weights for non-probabilistic surveys (Valliant & Dever, 2011, Wang et al., 2020a, b). All these methods require the use of a probabilistic survey as reference. We examined different possibilities for surveys to be used as reference. We determined that the Canadian Community Health Survey (CCHS) and the Canadian Tobacco, Alcohol and Drugs Survey (CTADS) are the best options; these surveys have many variables in common with the *Toking 9 to 5* panel and the target populations can be reconciled. We harmonized the variables between the *Toking 9 to 5* panel and CCHS variables for the first two waves of the panel.

SUB-PROJECT: Data Integration through outcome adaptive LASSO

Administrative data, or non-probability sample data, are increasingly being used to obtain official statistics due to their many benefits over survey methods. In particular, they are less costly, provide a larger sample size, and are not reliant on the response rate. However, it is difficult to obtain an unbiased estimate of the population mean from such data due to the absence of design weights. Due to selection bias, the sample mean of an outcome using a non-probability sample would not necessarily be an accurate estimate of the population mean of the outcome. Several estimation approaches have been proposed recently using an auxiliary probability sample which provides covariate information on the target population. In this research, we focus on the Chen, Li and Wu (2019) method. These authors developed a doubly robust inference estimation approach to inference with non-probability samples. The main objective of this project is to assess the possibility of extending Chen, Li and Wu (2019) approach to variable selection using the LASSO technique.

Chen, Li and Wu (2019) considered the situation where the auxiliary variables are given. However, when the covariate information is high-dimensional, variable selection is not a straightforward task even for a subject-matter expert. The goal of this project was to extend Chen, Li & Wu (2019) approach by applying

an outcome adaptive LASSO (Shortreed & Ertefaie, 2017) based penalty in order to perform variable selection. This work is relevant for several reasons. Studies have shown that including the variables in the propensity score model that influence the selection into the non-probability sample and also cause the outcome (Van der weele and Shiptser, 2011) leads to unbiased estimates. However, including variables in the propensity score model that affect the selection into the non-probability sample but not the outcome will increase the variance of the estimators relative to estimators that exclude such variables (Schistermann et al, 2009; Schneeweiss et al, 2009; van der Laan & Gruber, 2010). In the same sense, including variables in the propensity score model that are only related to the outcome will increase the precision of the estimator without affecting the bias (Brookhart et al, 2006; Shortreed & Ertefaie, 2017).

Progress:

We developed an extension of outcome adaptive LASSO in the context of combining non-probability and probability samples (Bahamyirou, 2021). We also developed an R program that implements the proposed method along with the Chen, Li and Wu (2019) estimator. The results of this project were submitted to the *Canadian Journal of Statistics*.

For more information, please contact:
Wesley Yung (613-951-4699, wesley.yung@statcan.gc.ca).

REFERENCES

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. and Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163, 1149–-1156.

Carrillo I.A., and Karr A.F. (2013). Combining cohorts in longitudinal surveys. *Survey methodology*, 39(1), 149-182.

Chen, Y., Li, P. and Wu, C. (2019). Doubly robust inference with Non-probability Survey samples. *Journal of the American Statistical Association* (published online).

Gruber, S., and van der Laan, M.J. (2010). An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *International Journal of Biostatistics*, 6(1), 18.

Schisterman, E.F, Cole, S. and Platt, R.W (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20, 488.

Schneeweiss, S., Rassen, J.A., Glynn, R.J., Avorn, J., Mogun, H. and Brookhart, M.A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20, 512.

Shortreed, S.M., and Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73(4), 1111–1122.

Valliant, R., and Dever, J.A. (2011). Estimating propensity adjustments for survey volunteer web surveys. *Sociological Methods & Research*, 40(1), 105-137.

Van der Weele, T.J., and Shipitser, I. (2011). A new criterion for confounder selection, Biometrics. 2011 Dec; 67(4): 1406–1413.

Wang L., Graubard B.I., Katki H.A. and Li Y. (2020). Improving external validity of epidemiological cohort analyses: a kernel weighting approach. *JRSS, A*, 183(3), 1293-1311.

Wang L., Graubard B.I., Katki H.A. and Li Y. (2020). Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts. https://arxiv.org/abs/2011.14850v1

5.2 Social Statistics Methods Division

SUB-PROJECT: Confidence Intervals for Estimated Counts

The goal of the research was to develop a confidence interval method for estimated counts with good coverage properties for the 2021 Census long form release. Most Census long form estimates are estimated counts, i.e., estimates of the total number of units that have a characteristic of interest. The usual method of constructing confidence intervals has poor coverage for estimated counts based on small sample sizes, and when nearly all sampled units have a value of 0 for the variable of interest or when nearly all units have a value of 1.

Progress:

A Wilson confidence interval for estimated counts was derived by mimicking the approach used by Wilson (1927) for proportions, and adapted for complex survey data using the method proposed by Kott and Carr (1997). Extensive simulation studies were undertaken to determine how the proposed interval should be implemented for domain estimation and calibration. The proposed method was also tested through simulations designed to reproduce the Census long form environment. An approximation of the modified Wilson interval was developed to circumvent constraints imposed by the Generalized Tabulation System (GTAB), which will be used for Census long form production. Simulation results show that the modified Wilson interval for estimated counts has good coverage properties for nearly all scenarios considered. The method and results are documented in Hidiroglou (2020), Neusy (2021) and Savard (2020,2021).

SUB-PROJECT: Combining survey data with crowdsourced data using small area estimation techniques

Crowdsourcing is a cost-effective and timely collection method where information is collected through a questionnaire that is open to any willing participant. However, crowdsourced data cannot be directly used to draw statistical inferences as the process is completely non-probabilistic. The use of Small Area Estimation (SAE) methods for combining crowdsourced data with probabilistic survey data was tested with the goal of variance reduction.

Progress:

The Fay-Herriot model was used to integrate categorical COVID-19 crowdsourced data with COVID-19 survey data sources from the first and third cycles of the Canadian Perspectives Survey Series (CPSS). The Fay-Herriot model was tested for its simplicity and interpretability. Overall, the Fay-Herriot model showed moderate success in integrating categorical crowdsourced data with survey data. When proper model fit

was achieved, significant improvements in the precision of estimates was observed. This held true even for the third crowdsourcing cycle which featured a dramatic reduction in the number of participants. However, proper model fit was not always achievable, potentially indicating the need to evaluate the use of other SAE methods.

A research report and a Social Statistics Methods Division (SSMD) working paper on this project was written (Chatrchi and Ding, 2021).

SUB-PROJECT: Integrating survey data and crowdsourced data: Sample Matching method

Sample matching is a relatively new methodology proposed by Rivers (2007) which consist of using non-probabilistic data as a pool of donor, and impute all answers for a probabilistic sample using a set of matching variables. It was tested in the context of COVID-19 surveys, where the imputed estimates were compared to the estimates from the probabilistic surveys.

Progress:

As a first step, the sample matching method was tested as Rivers proposed it: all units of a probabilistic sample were imputed using a donor from a non-probabilistic data source. The sample consisted of respondents of a survey using the same questions than the non-probabilistic data. Thus the estimates from both the probabilistic survey and the sample matching experiment were compared. As a second step, the sample matching method was used to impute the non-respondents of a non-probabilistic survey using a donor from a non-probabilistic data source. The estimates from both this combined method and from the probabilistic survey were compared.

Both steps were performed using the first wave of the Canadian Perspectives Survey Series (CPSS) as the probabilistic survey, and the first wave of the crowdsourcing projects as the non-probabilistic data source. Then they were performed again using the third wave of CPSS as the probabilistic survey, and the crowdsourcing component of the third wave of CPSS as the non-probabilistic data source.

The research report was completed and an SSMD Working Paper on this project is being reviewed (Poirier, Chatrchi and Wang, 2021).

For more information, please contact:
François Brisebois (613-222-8310, francois.brisebois@statcan.gc.ca).

REFERENCES

Kott, P.S., and Carr, D.A. (1997). Developing an Estimation Strategy for a Pesticide Data Program. *Journal of Official Statistics*, 13, 4, 367-383

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

Wilson, E.B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22, 209-212.

5.3 Statistical Integration Methods Division

SUB-PROJECT: Override errors in the data used in record linkage with the development and study of a new linkage method through the random generation of pseudo keys and decision processes.

The use of pseudo keys is common in linkage projects. A pseudo key is defined from portions of variables from both tables to be linked so that it forms a new variable that will serve as a linkage key. It simulates the role of a real key (such as the Social Insurance Number) by requiring uniqueness in the same table and forms pairs from records that have the same pseudo key value. In addition to being used in deterministic linkages, it is also important in probabilistic ones, where they are applied to help form a set of pairs related to the "blocking" step.

This project develops a linkage method based on the generation of random pseudo keys through sampling variables and aggregating the results of classifying pairs obtained by the application on the tables of the generated pseudo keys.

The primary goal was to implement new developments or functionalities of the method in Python. Another objective was to test the effectiveness of the method developed in the context of a two-table matching, based on different scenarios. The final objective was to begin designing a theoretical framework for pseudo keys and their uniqueness indicators.

Progress:

The method was developed to produce a range of indicators, referenced by Christen (2007), and tools for visualizing the quality of the linkage using a truth table. These functionalities helped achieve the second objective by studying the impact of the variation in the degree of inclusion and exclusion of the two input tables on the linkage results. Similarly, an analysis of the influence of the minimum threshold parameter for a vote between classifiers was produced mainly by the interactive tool of the "Receiver Operating Characteristic" (ROC) Curve, which presents these results in comparison with individual pseudo keys. Finally, theoretical development began with the formalization of concepts enabling a mathematical definition of "pseudo keys" in linkage, similar to those presented in Atencia, David and Euzenat (2014) and Pernelle, Saïs, Symeonidou (2013), as well as the randomness used in this method. This definition leads to a naive predicted likelihood of uniqueness, assuming character independence and i.i.d distribution of records, which was experimentally compared with the actual uniqueness rates of pseudo keys.

For more information, please contact:
Gautier Gissler (613-294-2574, gautier.gissler@statcan.gc.ca).

REFERENCES

Atencia, M., David, J. and Euzenat, J. (2014). Data interlinking through linkkey extraction. ECAI 2014: 15-20.

Christen, P., and Goiser, K. (2007). Quality and Complexity Measures for Data Linkage and Deduplication. Quality Measures in Data Mining, Studies in Computational Intelligence, vol. 43, Springer.

Pernelle, N., Saïs, F. and Symeonidou, D. (2013). An automatic key discovery approach for data linking. *Journal of Web Semantics*, Elsevier, 23, 16-30.

5.4 International Cooperation and Methodology Innovation Centre

SUB-PROJECT: Non-response follow-up for business surveys

In the last two decades, survey response rates have been steadily falling. In that context, it has become increasingly important for statistical agencies to develop and use methods that reduce the adverse effects of non-response on the accuracy of survey estimates. Follow-up of non-respondents may be an effective, albeit time and resource-intensive, remedy for non-response bias. The goal of this research project is to shed some light on a number of practical questions about non-response follow-up. For instance, assuming a fixed non-response follow-up budget, what is the best way to select units to be followed up? Should all non-respondents be followed up or just a sample of them? If a sample is followed up, how should it be selected?

Progress:

A simulation study was previously conducted to respond to the above questions. In the current year, we have added some theoretical results to better justify our empirical results, and we completed an article that was submitted to a peer-reviewed statistical journal (Neusy, Beaumont, Yung, Hidiroglou and Haziza, 2021).

SUB-PROJECT: Bootstrap estimation of the conditional bias for measuring influence in complex surveys

In sample surveys that collect information on skewed variables, it is often desirable to assess the influence of sample units on the sampling error of survey-weighted estimators of finite population parameters. The conditional bias is an attractive measure of influence that accounts for the sampling design and the estimation method. It is defined as the design expectation of the sampling error conditional on a given unit being selected in the sample. The estimation of the conditional bias is relatively straightforward for simple sampling designs and estimators. However, for complex designs or complex estimators, it may be tedious to derive an explicit expression for the conditional bias. In such cases, variance estimation is often achieved through replication methods such as the bootstrap. Bootstrap methods are typically implemented by producing a set of bootstrap weights that is made available to users along with the survey data. We study how to use bootstrap weights to obtain an estimator of the conditional bias. This estimator can be used to construct robust estimators of finite population parameters that are less negatively affected by influential units than standard survey-weighted estimators. We plan to evaluate our bootstrap estimator of the conditional bias in a simulation study.

Progress:

The theory was developed in previous research. This year, we conducted a simulation study and wrote an article (Beaumont, Bocci and St-Louis, 2021), which was submitted to a peer-reviewed statistical journal.

SUB-PROJECT: Development of a prototype system for robust estimation

In many economic and a few social surveys, variables with skewed distributions are collected, which may result in the presence of outliers and influential units. Traditional estimation methods may produce highly-inefficient estimators in that scenario. The idea of robust estimation is to diminish the effect of influential sample units on the estimates. The conditional bias is used as a measure of influence. The traditional sample estimate is decreased by a function of the conditional bias of the sample units. The conditional bias was first proposed by Moreno-Rebollo, Munoz-Reyez and Munoz-Pichardo (1999) and later used to develop a robust estimator by Beaumont, Haziza and Ruiz-Gazen (2013). This work is relevant to the many economic and social surveys at Statistics Canada that collect skewed variables.

Progress:

A SAS prototype has been developed and tested to include many of the specifications for robust estimation presented in Estevao (2021). It consists of a series of macros for the various functions related to the production of domain estimates. It includes a macro to calculate the traditional domain estimates as well as the domain robust estimates. The domain robust estimates generally do not have the additivity property of the domain estimates, so a macro exists to meet this requirement by creating domain coherent estimates from the domain robust estimates through minimal change of their values. There is also another macro to produce recalibrated weights for the sample units to ensure that they reproduce the domain coherent estimates and any known totals of auxiliary variables.

In the last year, a bootstrap method for variance estimation has been included to produce variance estimates for the domain robust estimates. A user guide with examples (Estevao, 2021) is available describing the first 6 of the 9 main functions. This will be updated soon to include a description of all the functions.

The compiled macros and the user guide for their application are available through the support of members of the International Cooperation and Methodology Innovation Centre.

SUB-PROJECT: Bootstrap variance estimation for multistage sampling with application to nonresponse

The bootstrap is often used for variance estimation in surveys with a stratified multistage sampling design. It is typically implemented by producing a set of bootstrap weights that is made available to users and that accounts for the complexity of the sampling design. The method of Rao, Wu and Yue (1992) is often used to produce the required bootstrap weights. It is valid under stratified with-replacement sampling at the first stage or when the first-stage sampling fractions are negligible. Some surveys do not satisfy these conditions. The goal of this project is to propose a bootstrap methodology for multistage designs that would be applicable when the conditions for the validity of the Rao, Wu and Yue (1992) bootstrap are not satisfied.

Progress:

In the previous year, we developed a simple bootstrap method that is valid even with non-negligible sampling fractions. It is applicable to any multistage sampling design as long as a valid bootstrap method is available for each individual stage of sampling. Our method is also applicable to two-phase sampling designs with Poisson sampling at the second phase. We use this design to derive bootstrap weights that account for nonresponse weighting. In the current year, we have started simulation studies to evaluate

the proposed bootstrap method and wrote the theoretical part of a paper, which we plan to submit to a peer-reviewed statistical journal.

For more information, please contact:
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

REFERENCES

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 555-569.

Moreno-Rebollo, J.L., Munoz-Reyez, A.M. and Munoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: conditional bias. *Biometrika*, 86, 923-928.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, 209-217.

5.5 Data Science Division

SUB-PROJECT: Data Science Division work

In September 2019, Statistics Canada created the Data Science Division in order to provide data science expertise and act as the front door for advanced data processing and analytics of big and unstructured data. Its multi-disciplinary team (data scientists, methodologists, managers), experts in the latest open source, coding, hardware and Cloud services techniques was created to tackle projects on any structured and unstructured data (text, images, etc…). The team promotes as well ethical science practices and methods that produce valid statistical inference. Its mission statement is to build data science capacity within the organisation by solving concrete problems.

Progress:

In the last year, the team conducted many data science projects using different data science techniques: reading and extracting information from PDF, using satellite imagery to derive agricultural information, exploring privacy preserving techniques, extracting information from news sources, comments or narratives, modeling COVID-19 related scenarios to help first line agency, etc…We also conducted horizontal activities such as launching in October 2020, the Data Science Network for the Federal Public Service which has now close to 2 000 members and published monthly newsletter. The Network is open to anyone interested in data science in Canada, whether they are data scientists or managers of data scientists. The five features of the Network are: talent management, training, information sharing, collaboration and joint services.

The division has also worked on a first version of Statistics Canada's upcoming data science strategy, based on the following six pillars: governance, operationalisation, infrastructure and tools, integration, people and, culture and leadership. The vision is: the agency is integrating machine learning and artificial intelligence methods, new processes, technologies and standards, with long standing analytical data stewardship and privacy preserving expertise, to provide better social and economic insights to Canadians

and policy makers. Lastly, the Data Science Division is contributing to raise data science capacity within the organization by leading or supporting related communities of practice, contributing to the Statistics Canada data literacy initiative or by proposing appropriate data science training to all levels of employees through on-line and formal courses, seminars and forum.

For more information, you can visit:
Data Science Network for the Federal Public Service web page: https://www.statcan.gc.ca/eng/data-science/network
Data literacy training on machine learning: https://www.statcan.gc.ca/eng/wtc/data-literacy/catalogue/892000062021004

Or contact:
Sevgui Erman (613-716-5906, sevgui.erman@statcan.gc.ca).

5.6 Survey Methodology Journal

Survey Methodology is an international journal available at www.statcan.gc.ca/surveymethodology that publishes articles in both official languages on various aspects of statistical development that are relevant to a statistical agency. Its editorial board includes world-renowned leaders in survey methods from the government, academic and private sectors. The journal is released in fully accessible HTML format and as a PDF.

The work related to the editorial and production processes include correspondence with authors, reviewers, associate editors, and referees; review of reviewer comments and author revisions; the layout and editing of texts; copy editing of manuscripts; liaison with translation and dissemination; and maintenance of a database of submitted papers.

Progress:

The June and December 2020 issues (46-1 and 46-2) were released in PDF and HTML versions. These issues include six and five regular papers, respectively. From April 2020 to March 2021, the Survey Methodology pages were viewed 47,000 times and nearly 18,000 copies of papers were downloaded. A total of 57 papers were submitted for publication.

In January 2021, Jean-François Beaumont, Senior Statistical Advisor at Statistics Canada, was appointed as the new editor of Survey Methodology. Beaumont has been associated with the journal for 20 years: He served as assistant editor from 2000 to 2010, and then as associate editor from 2010 to 2020. He succeeds Wesley Yung, who has been the editor since 2015. Meeting authors' heightened expectations in terms of the timeliness of the editorial process, while maintaining the same fundamental value of scientific rigour, is a priority for the new editor.

For more information, please contact:
Susie Fortier (613-220-1948, susie.fortier@statcan.gc.ca).

5.7 Knowledge Transfer – Statistical Training

The Statistical Talent Development Working Group, whose primary mandate continues to be the modernization of statistical training within the agency, had a busy and productive year. Although a well-established curriculum exists with courses that cover a range of statistical themes, the Group continues to develop and prioritize new learning activities that can be developed in a timely manner with an emphasis on active learning.

Early in the year, as the pandemic hit and many of Statistics Canada's programs were temporarily put on hold, employees had to quickly adjust to working from home full-time. This proved to be an ideal time for many employees to focus on training. As a result, a large inventory of available online self-learning tools was quickly developed and shared with employees of the Branch. Topics covered were varied, including sampling techniques, variance estimation, data science and machine learning, data integration, modelling and various statistical software (R, SAS, Python) to name but a few. This inventory continues to be a great training resource for employees to this day.

With data science and data modelling continuing to be areas of priority within the Agency, the focus of new initiatives remained mainly on these topics. A couple of data science courses were offered by notable university professors. The first was a repeat of the course piloted the previous year, which is now part of our curriculum. The second one targeted managers within the agency, focusing on what the methods can and cannot do rather than on the mathematical theory behind the methods. It generated great interest and should become part of our regular program for years to come as well. Also, a new learning activity on statistical modelling was piloted for the first time this past year. Three topics were covered, linear regression, logistic regression and cross- validation. Participants were first instructed to watch a video explaining the modelling technique. This was followed by an in-depth discussion with the group and a moderator, and finally by a demonstration of an application of the method in an existing Statistics Canada program. This combination proved very successful and really allowed for an active participation of all the participants. The experience will be repeated next year. Furthermore, as the use of R becomes more widespread within the Agency, additional offerings of the in-house introductory course developed last year were given and will continue to be offered.

Another new successful initiative in the past year was the launch of the Methodological Tidbits. These are short videos, from 30 to 40 minutes, developed by Statistics Canada employees covering a wide range of topics. More are in development for next year. Finally, it is worth noting that most of the in-class existing courses in our curriculum had a seamless transition to a virtual format.

For more information, please contact:
Pierre Caron (613-612-6910, pierre.caron@statcan.gc.ca).

5.8 Statistics Canada's 2021 International Methodology Symposium

Statistics Canada's 2021 International Methodology Symposium "Adopting Data Science in Official Statistics to Meet Society's Emerging Needs" will, for the first time, take place virtually over several weeks in the fall of 2021. Usually held in the National Capital Region of Canada, Symposium 2021 will be accessible online every Friday between October 15 and November 5, 2021, inclusively. The Symposium

will be free, will include plenary, parallel and poster sessions that cover a wide variety of topics and will be preceded by three workshops on Thursday, October 14th.

Progress:

The organizing committee has been very active in developing the different activities surrounding the symposium. Being held virtually, a new format was developed, potential platforms were assessed and an on-line abstract submission form was implemented. The committee has been busy with the usual activities such as advertising the symposium, developing the themes and the call for papers announcement, organizing invited sessions, and securing keynote and special presenters. The committee has also started the organization of the workshops, as well as the panel – determining topics and potential participants. Following the closing of the submissions, we will develop the full program, finalize the platform selection, obtain slides from presenters, translate material, develop a testing environment and hold the Symposium. After that, the last steps will be to collect papers for the proceedings, then coordinate and review their translation for electronic publication.

For more information, you can visit: https://www.statcan.gc.ca/eng/conferences/symposium2021/index.

6 Research papers sponsored by the Methodology Research and Development Program

American Statistical Association (2021). Time Series and Seasonal Adjustment Estimation During the COVID-19 Pandemic. Round-table discussion hosted by the Government Statistics Section, https://community.amstat.org/governmentstatisticssection/professionaldevelopmentmentoring/virtualworkshoppracticumfall2020349.

Arim, R., Hennessey, D. and Molladavoudi, S. (2021). Data, data analytics and data science at Statistics Canada during Covid-19 pandemic. Presented at Nexus 2021 Data Science Conference, University of Manitoba.

Bahamyirou, A. (2021). Data integration through outcome adaptive LASSO and a collaborative propensity score. Internal report, Statistics Canada.

Beaulieu, M., and Lebrasseur, D. (2021). « Measuring and Communicating Quality for Programs Using Administrative Data Sources Exclusively ». Presented at the forthcoming European Establishment Statistics Workshop.

Beaumont, J.-F., Bocci, C. and St-Louis, M. (2021). Bootstrap estimation of the conditional bias for measuring influence in complex surveys. Submitted for possible publication in a peer-reviewed statistical journal.

Beaumont, J.-F., and Chu, K. (2020). Statistical data integration through classification trees. Paper presented at the Advisory Committee on Statistical Methods, June 2020, Statistics Canada.

Beaumont, J.-F., and Rao, J.N.K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, 83, 11-22.

Blais, A.-R., Mach, L., Michaud, I. and Simard, J.-F. (2020). Analysis of the Public Service Employee Survey Items as Measures of the Psychosocial Risk Factors. Presentation to the Workplace Mental Health Performance Measurement Steering Committee, October 7, 2020.

Blais, A.-R., Michaud, I., Simard, J.-F., Mach, L. and Houle, S. (2021). Measuring Workplace Psychosocial Factors in the Federal Government. Submitted to *Health Reports*.

Bocci, C., Morissette, R. and Beaumont, J.-F. (2020). Overview of small area estimation methods used for the estimation of mean liquid assets. Internal report, Statistics Canada.

Bosa, K., and Chu, K. (2020). Ottawa COVID-19 Hospital Occupancy Forecasts – Final Report. Internal report (shared with Health Canada collaborators), Statistics Canada.

Chatrchi, G., and Ding, A. (2021). Combining survey data with crowdsourced data using small area estimation techniques, Social Statistics Methods Division working paper, SSMD-2021-02E, Statistics Canada.

Dasylva, A., and Goussanou, A. (2020). Estimating linkage errors under regularity conditions. In *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Dasylva, A., and Goussanou, A. (2021). Estimating the false negatives due to blocking. *Survey Methodology* (to appear in the December 2021 issue).

Data Analysis Resource Centre (2020a). Data Visualization - Best practices. Internal document, Modern Statistical Methods and Data Science, Statistics Canada.

Data Analysis Resource Centre (2020b). Measuring Culture of Accessibility, A Brief Introduction to Factor Analysis. Presentation to the Accessibility Culture Goal Working Group, May 1, 2020.

Denis, N., El-Hajj, A., Drummond, B., Abiza, Y. and Gopaluni, K.C. (2021). Learning COVID-19 Mitigation Strategies using Reinforcement learning. Chapter to appear in an upcoming Springer book.

Estevao, V.M. (2021). Robust Estimation - Parameter Description and User Guide, Methodology Specifications. Statistics Canada internal document.

Gilchrist, P. (2020). Public Use Microdata Files and Business Data: A Summary of Challenges for Confidentiality. Internal Paper, Statistics Canada.

Gray, D. (2020a). Evaluation of some imputation methods for monthly industry surveys using the Evaluation Framework and Visualisation. Internal document presented to the Modern Statistical Methods Branch Scientific Review Committee, Statistics Canada.

Gray, D. (2020b). Evaluating Imputation Methods using ImpACT: First Case Study. Presented at the *United Nations Statistical Commission and Economic Commission for Europe – Workshop on Statistical Data Editing.* https://unece.org/statistics/events/SDE2020.

Hidiroglou, M. A. (2020). Wilson Intervals for Proportions and Counts. Internal document, Statistics Canada.

Lesage, É., Beaumont, J.-F. and Bocci, C. (2021). Two local diagnostics to evaluate the efficiency of the empirical best predictor under the Fay-Herriot model. *Survey Methodology*, 47 (to appear).

Mach, L., and Michaud, I. (2020). Statistical Modelling with Complex Survey Data – Linear Regression. Session 1 of *Statistical Modelling Course*, Statistical Talent Development Working Group, Modern Statistical Methods and Data Science, Statistics Canada.

Matthews, S. (2020a). Over-arching principals for the Statistical Generalized Systems. Internal document presented to the Statistical Generalized Systems Steering Committee, Statistics Canada.

Matthews, S. (2020b). Strategy for Development of Advance Indicator. Internal document presented to the Data to Information: Modern Methods Committee, Statistics Canada.

Matthews, S. (2021a). Updated Generalized Systems Evolution Plan. Internal document presented to the Statistical Generalized Systems Steering Committee, Statistics Canada.

Matthews, S. (2021b). Guidelines for Advance Indicators of Official Statistics. Internal document presented to the Data to Information: Modern Methods Committee.

Matthews, S., Patak, Z. and Picard, F. (2020). Time Series Modelling to Produce Economic Indicators in (near) Real-time. Presented to Statistics Canada's Advisory Committee on Statistical Methods.

Matthews, S., and Ritter, C. (2020). Building a better nowcast – towards a real-time modelling environment. Presentation to Statistics Canada's Corporate Research and Development board.

Neusy, E., Beaumont, J.-F., Yung, W., Hidiroglou, M. and Haziza, D. (2021). Non-response follow-up for business surveys. Submitted for possible publication in a peer-reviewed journal.

Neusy, E. (2021). Wilson Confidence Intervals for Proportions and Totals of Binary Variables Using Complex Survey Data. Internal document, Statistics Canada.

Picard, F. (2020). SARIMA models for early estimates of energy balance statistics. Article presented to the United Nations Economic Commission for Europe – High level group on Modernizing Official Statistics.

Picard, F. (2021). A comparison between PROC BENCHMARKING and the R package TEMPDISAGG. Internal document, Statistics Canada.

Poirier, G., Chatrchi, G. and Wang, Y. (2021) Integrating survey data and crowdsourced data: Sample Matching method, Social Statistics Methods Division working paper (review in progress), Statistics Canada.

Reicker, K. (2020). Perceptions of data sensitivity, pilot study with questionnaire testing participants. Internal report, Statistics Canada.

Renaud, M., and Beaumont, J.-F. (2020). Crowdsourcing during a pandemic: the Statistics Canada experience. Paper to be presented at the Advisory Committee on Statistical Methods, October 2020, Statistics Canada.

Sallier, K. (2020). Toward More User-Centric Data Access Solutions: Producing Synthetic Data of High Analytical Value by Data Synthesis. *Statistical Journal of the International Association for Official Statistics*, 36, 1059-1066.

Sallier, K. (2021). Statistics Canada's experience creating public synthetic datasets using the FCS and the Synthpop Package. Presentation to the HLG-MOS Synthetic Data Guide working group.

Savard, S.-A. (2020). Intervalles de confiance pour les comptes estimés au questionnaire long du recensement. Présentation au Comité d'examen scientifique, 2 octobre 2020.

Savard, S.-A. (2021). Intervalle de confiance pour les comptes estimés. Rapport interne, Statistique Canada.

Statistics Canada (2020). G-EST 2.03.002 User Guide. Internal document, Statistics Canada.

Statistics Canada (2021). BANFF version 2.08 User Guide. Internal document, Statistics Canada.

You, Y. (2021a). Evaluation of robust small area estimation using normal and t distribution models. ICMIC internal research report, Statistics Canada.

You, Y. (2021b). Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling. *Survey Methodology*, 47 (to appear in the December issue).

You, Y. (2021c). A Bayesian approach to regression estimation and prediction inference for survey data analysis. ICMIC internal research report, Statistics Canada.

Zanussi, Z., Santos, B. and Molladavoudi, S. (2021), Supervised text classification with leveled homomorphic encryption. To appear in: Proceedings of the 63rd ISI World Statistics Congress.

Zhao, Z. (2021). Exploring available tools to generate synthetic data with high analytical value: R packages synthpop and simPop. Co-op report, directed by H. Gauvin. Internal document, Statistics Canada.