ABSTRACTS

Session 1 -- Keynote Address

(E) Quality and Official Statistics: Present and Future

Paul P. Biemer, RTI International, USA

The word “quality” is as ambiguous as it is ubiquitous. I will attempt to reduce some of the ambiguity and answer age-old questions such as: what quality it is; how is it evaluated; what are some proven strategies for improving it; and how can it be effectively managed in a statistical organization. My talk will focus primarily on “official statistics;” i.e., statistics that are produced by national statistical offices (NSOs) for purposes related to public policy. I will touch on the concepts of organizational, process and product quality; quality dimensions and their uses; total survey error vs. total statistical uncertainty; intrinsic vs. residual quality risks and practical quality assessment methodologies. I will also discuss some of the quality issues associated with integrating survey and non-survey data sources to produce hybrid data products, an activity that is becoming more common. The implications of these concepts for NSOs and other statistical institutions will also be discussed with an eye toward the future of official statistics. Examples from recent quality initiatives at a several statistical organizations will be used to illustrate the ideas.

Session 2A -- Small area estimation with Big data

(E) Big Data, Big Promise, Big Challenge: Can Small Area Estimation Play a Role in the Big Data Centric World?

Partha Lahiri, Lijuan Cao, Ying Han, Kartik Kaushik and Cinzia Cirillo, University of Maryland, USA

The demand for various socio-economic, transportation, and health statistics for small geographical areas is steadily increasing at a time when survey agencies are desperately looking for ways to reduce costs to meet fixed budgetary requirements. In the current survey environment, the application of standard sample survey methods for small areas, which require a large sample, is generally not feasible when considering the costs. The accessibility of Big Data from different sources is now bringing new opportunities for statisticians to develop innovative SAE methods. I will first explain how relevant information can be extracted from social media Big Data to solve a SAE problem. I will then describe how SAE can help solve a seemingly different problem of predicting in real-time traffic by exploiting rich vehicle probe Big Data.

(E) Small Area Estimation to correct for measurement errors in big population registries

Danny Pfeffermann and Dano Ben-hur, Central Bureau of Statistics, Israel

As in many countries, Israel has a fairly accurate population register at the national level, which consists of about 8.5 million persons. However, the register is much less accurate for small domains, with an average domain enumeration error of about 13%. The main reason for the inaccuracy at the domain level is that people moving in or out an area often report late their change of address. In order to correct the errors at the domain level, we investigate the following three-step procedure.

A- Draw a sample from the register to obtain initial sample estimates for the number of persons residing in each domain on “census day”,

B- Apply the Fay-Herriot model to the initial estimates so as to improve their accuracy,

C- Compute a final estimate for each domain as a linear combination of the estimates obtained in Step B and the register figure.

The presentation will discuss the considerations motivating this approach and show empirical results based on the last census in Israel. A procedure to deal with not missing at random nonresponse in Step A will be considered and illustrated.

(E) Small Area Model-Based Estimation using Big Data

Stefano Marchetti, University of Pisa, Italy

National statistical offices have the mission to produce statistics for citizen and policy-makers. Survey sampling has been recognized to be an effective method to obtain timely and reliable estimates for a specific area in socio-economic fields. Usually, it is important to infer population parameters at a finer area level, where the sample size is usually small and does not allow for reliable direct estimates. Small area estimation (SAE) methods – which uses linking models based on auxiliary variables that “borrow strength” across small areas – try to obtain reliable estimates when the directs are unreliable. SAE methods can be classified into unit-level and area-level models: Unit-level models require a common set of auxiliary variables between survey and Census/registers known for all the population units; area-level models are based on direct estimates and aggregated auxiliary variables. Privacy policies and high census costs make difficult the use of unit-level data. Aggregated auxiliary variables from different sources are easily available and can be used in are-level models. Big Data – a collection of data that contains greater variety arriving in increasing volumes and with ever-higher velocity – adequately processed can be used as auxiliary variables in area-level models. Here, I present two applications of SAE: first, I use mobility data to estimates poverty incidence at local level in Tuscany, Italy; latter, I use twitter data to estimate the share of food consumption expenditure at the province level in Italy. Moreover, the use of Big Data sources in official statistics is discussed focusing on future challenges.

Session 2B -- Data integration I

(E) Estimating annual wealth distributions within Canada’s System of National Accounts

Margaret Wu and Cilanne Boulet, Statistics Canada, Canada

In recent years, there has been growing interest in distributional measures of economic well-being. To address this need, Statistics Canada is building a series of annual tables integrating macro-level national accounts data with micro-level survey data on wealth, income and consumption. This product, the Distributions of Household Economic Accounts, adds a distributional component to Canada’s macroeconomic accounts, thereby giving a more complete picture of the economic well-being of Canadian households within the national accounts framework.

This talk will describe the methodology used to build the wealth tables. In the past, Statistics Canada has conducted a survey of wealth only on an occasional basis. One of the major challenges of this project is to find a way to fill the relatively long gaps in the data between survey years. Modelling, calibration, benchmarking, and raking are combined to address the challenges of filling these gaps and ensuring consistency with macroeconomic accounts totals

and the survey data. The incorporation of newly obtained administrative data sources on liabilities into the wealth tables is also being investigated.

(E) Transactional Data Processing System

Agnes Waye, Serge Godbout and Nathalie Hamel, Statistics Canada, Canada

Transactional data is becoming more commonly used as a source of administrative data or in surveys. The richness and volume of the data allows the user to gain valuable insight and to conduct more thorough analysis of trends. However, such large datasets with complex structures pose unique challenges in terms of data processing and estimation, and classic data processing methods require adapted solutions. At Statistics Canada, there is a gap in the statistical infrastructure to process transactional data. We have identified the need to develop a system that is more robust to process transactional data since a high level of flexibility is required. A transactional data processing system has been developed for transportation surveys, which include many surveys with transactional data. One survey has been integrated into this system so far (Fare Basis Survey), and gradually other surveys from aviation, rail and trucking statistics programs will be integrated as well. This system implements steps from the process phase as identified in the Generic Statistical Business Process Model (GSBPM), including features such as data import, edit and imputation, data integration, balancing data, and estimation. This paper will describe the definition and the specific characteristics of transactional data, how they are processed, lessons learned, challenges we faced as well as future issues to resolve in the transactional data system.

(E) Measuring mobility by combining big data and administrative data sources at the Centre for Big data Statistics of Statistics Netherlands

Marko Roos, Statistics Netherlands, The Netherlands

At the Centre for Big data statistics of Statistics Netherlands work is done on measuring mobility patterns , among others themes. For this purpose several administrative data sources are combined with big data sources. Specifically, big data based on traffic loop data and data from mobile phone providers is compared to data from administrative sources on wages. The administrative data on wages is indicative of commuting patterns, because it holds addresses of companies for employees. Combining those with living locations, and projecting commuting routes on the resulting origin-destination matrices, yield possible commuting patterns. Big data sources originating from traffic loops and mobile phone providers are used to corroborate and refine these patterns. This approach may enable timely and detailed insight into mobility within the Netherlands.

In this paper, the first results of this approach are presented within the context of a resume of other themes within the Centre for Big data statistics of Statistics Netherlands.

(E) A look inside the box: Combining aggregate and marginal distributions to identify joint distributions

Marie-Hélène Felt, Bank of Canada, Canada

This paper proposes a method for estimating the joint distribution of two or more variables when only their marginal distributions and the distribution of their aggregate are observed. Nonparametric identification is achieved by modeling dependence using a latent common-factor structure. Multiple examples are given of data settings where multivariate samples from the joint distribution of interest are not readily available, but some aggregate measures are observed.

In the application, intra-household distributions are recovered by combining individual-level and household-level survey data. The Bank of Canada monitors Canadian's payment behaviour via two surveys: the Method-of-Payments (MOP) Survey and the Canadian Financial Monitor (CFM). Both surveys collect information on payment choices at the point of sale as well as cash management habits; they differ with respect to their unit of observation. In the MOP, the unit of observation is the individual respondent; all the questions relate to the respondent's own individual characteristics and behaviours. In the CFM, the main unit of observation is the household; demographic characteristics are observed for the female and male heads of the household, but cash and alternative methods of payment quantities are collected at the aggregated household level - the respondent is asked to report the monthly family total.

By applying my methodology I can investigate intra-household influences, with respect to payment and cash management practices, in the absence of intra-household data. I find that, for individual living in couple relationships, personal cash management practices in terms of withdrawals and holdings are significantly influenced by the partner's cash and single-purpose stored-value cards utilization.

(E) Transport Survey Estimate Adjustment by Permanently Installed Highway-sensors Using Capture-recapture Techniques

Jonas Klingwort, Bart Buelens and Rainer Schnell, University of Duisburg-Essen, Germany and Statistics Netherlands, the Netherlands

Recently, the integration of sensor-based data has become increasingly relevant in official statistics. The integration of sensor data is in particular highly valuable if it can be linked with survey and administrative data. The application demonstrated here combines such datasets to quantify and adjust underreporting in survey point estimates.

We use the Dutch Road Freight Transport Survey (RFTS) and road sensor data produced by automated weighing stations (WIM) installed on Dutch highways. The RFTS is a probability sample of registered truck owners who report the trips and the weight of the cargo for the sampled truck in a specified week. The nine WIM stations measure continuously every passing truck and use a camera system scanning the license plates to identify trucks. The datasets can be linked one-to-one using the license plate as a unique identifier. Since administrative registers provide information on the empty weight of each truck and trailer, the weight of the load can be computed. Hence, the survey and the sensors independently measure the same target variable: the transported cargo weight.

Capture-recapture methods are used to estimate underreporting in the RFTS. Heterogeneity of the vehicles with respect to capture and recapture probabilities is modeled through logistic regression and log-linear models. Different estimators will be compared and discussed. Results show the approach being promising in terms of validating and adjusting survey data using external sensor data. This application is a new example of multi-source statistics to improve the benefits of sensor data into the field of official statistics.

Session 3A -- Prioritizing the use of administrative data for official statistics

(F) Development of a census methodology combining administrative data and data obtained through traditional collection

Jean-François Simard, Roxanne Gagnon, Georgina House, Francis Demers and Christian Nadeau, Statistics Canada, Canada

In June 2016, Statistics Canada launched the Census Program Transformation Project to develop a census methodology that maximizes the use of available administrative data and maintains the quality and relevance of census products. The preferred methodology is a combined census that relies heavily on statistical registers built largely from administrative data. In a combined census, statistics are taken from registers or other sources of administrative data, to which information is added from a sample for certain variables or partial enumeration for others.

Simulations are done to measure the reliability of population and dwelling counts produced from administrative data supplemented with data obtained through traditional collection. The objective of the simulations is to evaluate different methodological options. These simulations are the first in a series that is expected to culminate in the complete simulation of a combined census in 2021.

The simulations combine data from the Canadian Statistical Demographic Database with 2016 Census data. The population counts produced are compared with equivalent counts from the traditional census at different levels of geography and for demographic characteristics such as age, sex and household composition. Various possibilities for improving and measuring the coverage of a combined census are included in the simulations.

This presentation will discuss the proposed methods for the possible implementation of a combined census in Canada as well as the results of the first simulations.

(E) Admin-First as a Statistical Paradigm for Canadian Official Statistics: Meaning, Challenges and Opportunities

Eric Rancourt, Statistics Canada, Canada

For decades, National Statistical Offices have claimed they intend to use more and more administrative data; and they have actually use them to various degrees from one program to another. But with the advent of the data revolution, it is no longer a wish, a side issue, a marginal method, or an increasing trend; it has become the central focus of attention for the future of programs. Whether the objective is to enhance relevance, reduce response burden, increase efficiency, or produce faster with more details, the use of administrative data (in the broadest sense) is proliferating within and without statistical systems at a sky rocket pace. Statistics Canada is facing the new data world by modernizing itself and embracing an admin-first paradigm. This paper attempts to explain what this means, to highlight some of the practical and theoretical challenges and to point out possible opportunities.

(E) Use of administrative data in New Zealand's 2018 Census – the future has arrived

Nathaniel Matheson-Dunning, Anna Lin, Christine Bycroft, Statistics New Zealand, New Zealand

The NZ 2018 Census is a modernised full-enumeration census, with a digital-first response mode and increased use of administrative data to supplement survey responses. For the first time, administrative sources (including the previous census) will form an integral part of census data collection.

The 2018 Census is a step towards a long-term future where the census would be based on administrative sources, supported by surveys, and reflects our goal of being an 'administrative data first' organisation.

We are fortunate to have a rich array of linked administrative and survey data sources available in Stats NZ's Integrated Data Infrastructure (IDI). Quality assessments have determined which admin-derived variables will produce high quality information for the census.

For the 2018 Census, administrative sources, as well as information provided in the previous 2013 Census, is being used to impute responses that are missing for a given individual.

As well as item non-response, in some cases, whole households do not respond to the census. In our quality assessments we found that it is more difficult to place the correct people in households constructed from the administrative addresses. Building on work by the US Census Bureau, we have developed a model to assess which IDI-derived households are likely to have the most reliable household membership. The administrative data can then be used to derive responses for each household member.

We will present the methods used for both forms of imputation. We summarise the findings so far, and discuss the likely implications for the final outputs from the 2018 Census.

Session 3B -- Innovative methods

(E) The Imitation Game: An Overview of a Machine Learning Approach to Code the Industrial Classification

Javier Oyarzun, Statistics Canada, Canada

Statistics Canada's Business Register (BR) plays a fundamental role in the mandate of Statistics Canada. The BR is a database that includes all businesses operating in Canada. Close to a hundred business surveys use the BR in various ways, but, mainly it is used for establishing survey frames, sampling, collecting and processing data, and producing estimates.

The Business Register has a direct impact on the efficiency of the business survey process, the reliability of data produced by business statistics programs and the coherence of the national accounting system. In early 2018, Statistics Canada started developing a new methodology to probabilistically code the industrial classification of businesses. This methodology which uses data mining, text mining and machine learning will provide Statistics Canada with a tool to code missing industrial classifications and improve the overall quality of the industrial classifications on the BR.

This paper will discuss the North American Industrial Classification System (NAICS) and its usage in statistical programs at Statistics Canada. It will also present current and new approaches to coding the NAICS. Finally, the paper will discuss the challenges related to the portion of the BR which is not coded and present complex cases of NAICS coding.

(E) Modelling measurement errors to enable consistency between monthly and quarterly turnover growth rates

Arnout van Delden, Sander Scholtus, and Nicole Ostlund, Statistics Netherlands, The Netherlands

For a number of economic sectors, Statistics Netherlands (SN) produces turnover growth rates of businesses: monthly figures based on a sample survey and quarterly figures based on administrative tax data. SN aims to benchmark the monthly growth rates on the quarterly ones in order to produce consistent output, especially for the National Accounts which uses both series as input.

Preliminary results of benchmarking showed that monthly growth rates were adjusted differently between quarters. In fact, the quarterly administrative turnover turned out to be relatively large in the four quarter of the year compared to the survey data whereas the

opposite was true in the first quarter. This effect is probably caused by quarterly patterns in measurement errors, for instance due to administrative processes within businesses. These patterns may also occur in other countries that use or aim to use short-term administrative data and compare results with survey data.

We present a methodology that aims to automatically detect and correct for such measurement errors. Starting with a large set of background variables, found through discussions with administration offices, we test which of those are associated with the quarterly patterns in the measurement errors by using decision trees. From this we construct composite variables that are best associated with the seasonal measurement errors. These composite variables are used in a mixture regression model that describes the relation between quarterly administrative and survey turnover. The model consists of several groups of units, each capturing a different measurement error structure, with one group using the composite variable to explain quarterly measurement effects. The measurement errors, that would otherwise be an obstruction to sensible benchmarking, could be corrected for using the parameter estimates of this mixture model.

(E) Bringing survey sampling techniques into big data

Antoine Rebecq, Ubisoft Montréal, Canada

Big data tools are advanced and complex, and as a result, computer science culture is more widely spread than statistical culture among data science teams. Nevertheless, analysis of big datasets presents statistical challenges that data science and machine learning professionals have to be aware of. We present a few examples of how adding some sampling tools in big data helped Ubisoft's data analysts better understand what players enjoy in games.

At Ubisoft, the primary source of player data from games are events that are sent by the game client to servers, and then go through a big data pipeline to make it available to analysts. The volume of the data is so huge that the events still sometimes need to be sampled. As a result, some events are observed with intricate inclusion probabilities. Data scientists are often unaware of the biases, typically when they train machine learning models. We will show that we can improve the performance of algorithms by using calibration on margins.

Some other data is also collected using conventional marketing opportunity samples. We'll discuss how we can create efficient sampling designs and reweighting for such data by leveraging big data collected from in-game events. Text mining techniques have also become increasingly popular in the last few years. Companies look at what users say on social media to target the areas of their products that need improvement. We show that sampling techniques can also help treat biases (e.g. auto-selection bias) that are inherent to these datasets.

Finally, sampling tools are put to use in more exploratory projects. Network data is incredibly valuable, but also very hard to handle. We show how sampling can be used to make network data analyses cheaper and more reliable.

(E) Developing an exploratory open database of buildings

Alessandro Alasia, Jean Le Moullec, Marina Smailes, Statistics Canada, Canada

TBA

(E) The challenges of producing national estimates of child maltreatment using administrative data from different jurisdictions

David Laferrière and Catherine Deshaies-Moreault, Statistics Canada, Canada

STC was approached to conduct a feasibility study on how to develop a surveillance system for child maltreatment. The Canadian Reported Child Maltreatment Surveillance System (CRCMSS) would integrate data from child welfare agencies in each province and territory to calculate annual estimates of child maltreatment in five categories: physical abuse, emotional abuse, sexual abuse, neglect, and exposure to inter-partner violence. To reduce burden on child welfare workers, the primary source of data would be a census of administrative data from the provinces and territories (P/Ts).

There are several challenges to overcome in order to implement CRCMSS. Each P/T has its own legislation on what constitutes child maltreatment and how to categorize it. The P/Ts also have different systems to identify and track cases of maltreatment. The content and comprehensiveness of the administrative data, which can include both microdata and text narratives, varies substantially.

Traditionally, identifying cases of maltreatment from narratives would require coders. However, for CRCMSS natural language processing techniques from machine learning will be explored to determine if cases of maltreatment could automatically be identified and classified from narrative reports.

Another challenge is that administrative data might not be available from every P/T. When administrative data is not available, one alternative is to have child welfare workers complete surveys.

We discuss the CRCMSS feasibility study, a project that explores the practical and technical challenges of using traditional approaches as well as more modern techniques to create coherent national estimates from the administrative data and survey data of 13 P/Ts.

Session 4 -- Waksberg Award Winner Address

(E) The Sage Statistician and Conditional Calibration
Donald Rubin, Harvard University, USA

TBA

Session 5A -- Geospatial data

(E) Monitoring spatial sustainable development: (Semi-)Automated analysis of satellite and aerial images for energy transition and sustainable indicators.
R.L. Curier, T.J.A. de Jong, D. Iren, S. Bromuri, Statistics Netherlands and Business Intelligence and Smart Services Institute, The Netherlands

Europe aims to replace by 2050 at least 30% of the demand for fossil fuels by renewable resources, requiring thus urban energy systems that emit less carbon and use less energy. Nowadays, solar energy plays an important role in the energy transition, and policy makers and net operators are strongly interested in mapping the solar panels. Current statistics on solar energy are based on surveys from solar panels importation and solely provides national figures on a yearly basis while the energy transition create a demand for information at regional to local scale with shorter time scales.

The current study aims to produce a map of solar panels along with statistics on the number of solar panels by automating the detection of solar panels. To this end, the information content from high resolution satellite and aerial images is analysed by means of artificial intelligence to allow for the automatic detection and classification of solar panels. Two machine learning approaches, support vector machine and convolutional deep neural networks, will be used to identify solar panels in images of various resolutions. Further, the project will also make use of existing registers such as information on the VAT returns from solar panels owners and information acquired from the energy providers to the machine learning algorithm.

In this presentation, the preliminary results for the province of Limburg (NL), Flanders (BE) and North Rhine Westfalia (DE) will be discussed and the added value of the use of remote sensing data to infer information will be addressed.

(E) Do fences really make good neighbors? A side-by-side comparison of RDD and geofencing methods using risk factor surveys

James Dayton and Matt Jans, ICF, USA

TBA

TBA

Session 5B -- Non-probability samples

(E) Calibrating Non-Probability Samples With Probability Samples Using LASSO

Jack Chen, Michael R. Elliott and Rick Valliant, University of Michigan, USA

One of the most prominent applications of survey research is election polling. Due to declining land-line phone coverage and improved phone screening technology, it has become a significant challenge for election pollsters to capture voting intentions in a timely way. This has lead to the expanded use heap and easily accessed samples of individuals obtained from non-probability web surveys. But non-probability samples are at risk for selection bias due to differential access, degrees of interest, and other factors. Calibration is a standard method used in official statistics and other settings that uses weights to adjust total estimates from a sample to known totals in a population. Because non-probability samples do not have robust inferential properties, we consider use of model-assisted calibration methods that allow robust estimation of population totals. In particular, we consider calibration to estimated population totals using adaptive LASSO regression – estimated-controlled LASSO (ECLASSO). Adaptive LASSO can yield a consistent estimator of a population total as long as a subset of the true predictors is included in the prediction model, thus allowing large numbers of possible covariates to be included without risk of overfitting. This allows to the possibility of calibration to estimates from higher-quality probability samples with modest sample sizes. We apply ECLASSO to predict the voting result for 11 gubernatorial elections and 8 senate elections in the U.S. 2014 midterm election. Since the actual election results are published, we can compare the bias and root-mean square error of ECLASSO with traditional weighting adjustment methods.

(E) Statistical Inference with Non-probability Survey Samples

Changbao Wu, Yilin Chen and Pengfei Li, University of Waterloo, Canada

We establish a general framework for statistical inferences with non-probability survey samples when relevant auxiliary information is available from a probability survey sample. We develop a rigorous procedure for estimating the propensity scores for units in the non-probability sample, and construct a doubly robust estimator for the finite population mean. Variance estimation is discussed under the proposed framework. Results from simulation studies show the robustness and the efficiency of our proposed estimators as compared to existing methods. The proposed method is used to analyze a non-probability survey sample collected by the Pew Research Center with auxiliary information from the Behavioral Risk Factor Surveillance System and the Current Population Survey. Our results illustrate a general approach to inference with non-probability samples and highlight the importance and usefulness of auxiliary information from probability survey samples.

(E) Understanding the Effects of Record Linkage on Estimation of Total when Combining a Big Data Source with a Probability Sample

Benjamin Williams, Lynne Stokes, Southern Methodist University, USA

Record linkage is a useful tool which links records from two lists that refer to the same unit but lack a unique identifier. The effect of matching error from record linkage has not been considered for the case of estimating the total of a population from a capture-recapture model.

The National Marine Fisheries Service (NMFS) estimates the total number of fish caught by recreational anglers. NMFS arrives at this by estimating the total effort (the number of fishing trips) and the catch per unit effort or CPUE (the number of fish caught per species per trip), and then multiplying them together. Effort data are collected via a mail survey of potential anglers. CPUE data are collected via face-to-face intercepts of fishing trips. The effort survey has a high non-response rate and is retrospective, causing a lengthy estimation process, precluding in-season management.

Due to these limitations, NMFS is experimenting with replacing the effort survey with electronic self-reporting. The anglers report trip details via an electronic device and remain eligible to be sampled in the dockside intercept.

In this scenario, proposed estimators of total use the self-reports (a large non-probability sample) alongside the dockside intercept (a probability sample), using capture-recapture methodology (Liu et al 2017). For the estimators to be valid, data from trips that both self-reported and were sampled in the intercept must be linked. Current estimators assume perfect matching, however this is difficult in practice due to device and measurement error.

In this paper, we propose several additional estimators and develop a record linkage algorithm to match trips. We examine the effect of matching errors on the estimators and illustrate these effects using data from one of the electronic reporting experiments.

(E) Web Scraping as an alternative data source to predict E-Commerce indicators

José Márcio Martins Júnior, Marcelo Trindade Pitta, Joao Victor Pacheco Dias and Pedro Luis do Nascimento Silva, Brazilian Network Information Center and National School of Statistical Science from IBGE, Brazil

To estimate E-Commerce indicators, like the proportion of Webpages which sell products and/or services, through traditional surveys and considering the need for providing both disaggregated and timely data, is expensive and time consuming. To estimate such indicators, this paper proposes an approach that combines survey data with the information scrapped from the source code (HTML) of the Webpages of the companies as an additional data source. The sample of companies surveyed includes all enterprises selected and responding to the 2017 Brazilian Survey on Information and Communications Technologies in Enterprises (TIC-

Empresas). As the survey information includes the addresses of the Webpages and also the answers to the questions regarding e-commerce practices of the companies, using the survey database and accessing the corresponding companies Webpages enabled the building of a model to evaluate the HTML in order to establish automated responses to some of the variables required to estimate required e-commerce indicators. To get the source codes of the Webpages, a custom Java crawler was used. The model attempted to use the information in the first page (homepage) as the input for a logistic regression model regarding the answers to selected indicators. The explanatory variables in the model correspond to the words in Webpages, these variables derived from a dictionary of the most relevant words created to predict the outcome of these indicators. Using this methodology, the sample can be enlarged by crawling over larger samples of companies Webpages, and the resulting data used to estimate the required indicators more accurately, faster and more disaggregated, thus, requiring less resource than the usual way. Then, it is possible to compute more up-todate estimates of the required indicators whenever necessary.

(F) Indirect sampling applied to capture-recapture models with dependency between the sources

Herménégilde Nkurunziza and Ayi Ajavon, Statistics Canada, Canada

Capture-recapture is a widely used method for estimating the unknown size of a population. The method consists of drawing two independent samples from the population of interest. The often-used Petersen estimator of population size depends on sample size and overlap. Lavallée and Rivest (2012) examined samples from indirect sampling and introduced a generalization of the Petersen estimator based on the generalized weight share method. In practice, the independence assumption on which the estimator is based is not often verified (Brenner 1995). In this article, we will examine capture-recapture models with dependency between samples and propose an extension of the estimator put forward by Lavallée and Rivest (2012). We analyze the properties of the resulting estimator and illustrate the method using simulated data.

Session 6A -- Using alternative data sources for social statistics programs

(E) Herding and exploring combinations of electronic transaction data for alternative HBS-design purposes

Anders Holmberg, Statistics Norway, Norway

When Statistics Norway cancelled its 2018 Household Budget Survey it was because of cost/quality trade-off concerns and hesitation whether an essentially traditional survey-based diary approach would be satisfactory. The decision to cancel, kicked-off intensified investigations to acquire alternative data sources for household consumption data. The national transaction data landscape can therefore now be well described, and three different sources of electronic transaction data as well as how they can be combined are being examined on experimental basis. One source is transaction data from a key payment provider in Norway (card transactions and other electronic transactions, also B2B transactions). The coverage rate is quite substantial and covers most card transactions done in Norway. We also investigate cash register data of from retail chains and samples of retail chain membership data. All these data sources are interesting by themselves but finding ways to combine the different sources is what really adds value, at least from an HBS-perspective. E.g. methods of record linking card transactions with cash register data to combine both the demographic spending dimension and the detailed consumption dimension down to detailed COIOCOP. The paper discusses the possibilities and the methodological and technical experiences made from this work.

(F) Modernizing the Household Expenditure Program

Christiane Laperrière, Denis Malo and Johanne Tremblay, Statistics Canada, Canada

The Survey of Household Spending collects information that is an essential input into the Consumer Price Index and the System of National Accounts. These data are also used by a broad community of users who are generally interested in analyzing expenditures by the socioeconomic characteristics of households. The detailed content and the traditional data collection practices using personal interviews and an expenditure diary impose a heavy burden on respondents and lead to high collection costs. As part of Statistics Canada's modernization initiative, the Household Expenditure Program is exploring potential new sources of data. Using such a large volume of data creates new challenges, including the use of automated learning algorithms to classify data into expenditure categories relevant to the program. The framework for integrating these multiple data sources must also be established and will require the development of new methods. Studies have begun to evaluate how to adapt the integration methods to the needs of the different types of users. In this presentation, we will discuss the innovative ideas being considered for classification and integration, the challenges encountered in exploring these new data sources, and the results of ongoing evaluations.

(E) Combining Multiple Data Sources to Enhance U.S. Federal Statistics

Brian Harris-Kojetin and Robert M. Groves, US National Academies of Sciences, Engineering, and Medicine and Georgetown University, USA

Large-scale probability sample surveys have long been the foundation for producing many U.S. national statistics, but the costs of conducting such surveys have been increasing while response rates have been declining, and many surveys are not keeping up with growing demands for more timely and detailed local information. The Committee on National Statistics at the National Academies of Sciences, Engineering, and Medicine convened a committee of experts in social science research, sociology, survey methodology, economics, statistics, privacy, public policy, and computer science to explore a possible shift to an approach combining data sources to give users richer and more reliable datasets. The panel conducted a two-year study and produced two reports with conclusions and recommendations for federal statistical agencies. The first report, Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy, reviews the current approach for producing federal statistics, examines other data sources such as government administrative data and private sector data sources, including Internet and other big data sources, which could also be used for federal statistics. The second report, Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps, examines more in-depth the necessary infrastructure, methods, and skills to implement a multiple data sources paradigm for official statistics. Together, the reports provide an overview of statistical methods and quality frameworks that have been used for combining information and outline research that is needed for the development of methods for combining data sources and protecting privacy. The presentation will give an overview of the major conclusions and recommendations from these reports.

Session 6B -- Record linkage

(E) Evaluation of the Accuracy of Links or Candidate Pairs within Record Linkage Analyses

Dean Resnick and Lisa Mirel, National Opinion Research Center at the University of Chicago and National Center for Health Statistics, USA

Any complete record linkage strategy requires an approach to evaluating the accuracy of candidate pairs or links on a case-by-case basis or in aggregate. The seminal Fellegi and Sunter record linkage paper suggest that pairs between the rejection cutoff score and acceptance cutoff score be reviewed manually to set link status. Nevertheless, it is an open question whether clerical reviewers can accurately infer match status, to say nothing of the expense of conducting this review for a large number of pairs. We surmise that where the human reviewer may add value by being able to see a name transposition or substitution or be able to factor name rarity into the analysis, it would often (but not always) be the case that these types of techniques should have already been coded into the programmed linkage evaluation routine.

Alternatively, where a high quality test deck is available, this can be used to measure link quality, but this will usually not exist. In its absence, a highly accurate ID field (such as Social Security Number) can be used instead and the valid linkage rate would be approximately the level of agreement on this field within the linked pairs. But even if this were not available, we can use an approach that applies record linkage theory to first estimate pair match validity and use that in-turn to estimate linkage accuracy. This talk will present different methods to evaluate record linkage quality that go beyond clerical review.

(E) Decision Rules and Error-Rate Estimation for Record Linkage Using a Probability Model

Clayton Block, Elections Canada, Canada

Since 1997, Elections Canada has maintained the National Register of Electors, a database of Canadians aged 18 and over, used to administer federal elections. This database is updated from several federal and provincial administrative sources, linked to electors in the database using personal information such as names, date of birth, gender, and address. Initially, commercial linkage software based on Fellegi-Sunter theory was used for these linkage activities. Gradually, the methodology and software used have shifted towards custom-built solutions, providing more flexibility over how potential pairs get processed, and reducing the classification error rates associated with the linkage process. One key improvement to the methodology is a reformulation of the familiar Fellegi-Sunter decision rule, now put in terms of a probability of interest and compared to actual error tolerances. For matching on personal information, the required probabilities are calculated from the observed pairs with the aid of a simple probability model for chance agreement on date of birth. The model assumptions should be quite realistic. The probabilities calculated for each pair can also be simply added up to produce estimates of the two types of matching error, requiring no specialized software and no complex mathematical procedures. The methods described have been used for various linkage processes at Elections Canada, each with different expected match rates. In all cases, the error rates produced appear to be believable. In the future, these results could be compared and contrasted with those obtained from competing, more complicated error-rate estimation methods.

(E) Combining commercial company data with administrative employer-employee data – methodological challenges of linkage, preparation and representativity

Manfred Antoni, Marie-Christine Laible, Institute for Employment Research (IAB) Research Data Center, Germany

We describe the linkage of commercial company data from Bureau van Dijk (BvD) with administrative employment data of the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB). BvD is a commercial provider of company data and its databases have mainly been used for analyzing business intelligence. Meanwhile, the FDZ has been providing cost-free access to administrative and survey data to researchers for over a decade.

To combine the research potentials of both data sources, the FDZ has performed a record linkage of the companies (independent unit) given in BvD's database Orbis with the establishments (dependent subunits) given in the Establishment History Panel (BHP) of the FDZ. First, we outline the linkage process and the methods the FDZ applied. Thus far, no large-scale linkage between BvD data and administrative data had been successful, as the major obstacle is that the two data sources do not contain a common identifier that would allow a direct linkage. The FDZ thus performed the linkage by comparing, amongst others, the names of companies and establishments given in the original databases. Second, we present the steps of creating a research dataset from the company-establishment correspondence table generated by the record linkage. We describe the challenges encountered, such as multiple assignments, and the methods applied to overcome them. The resulting dataset contains longitudinal information on companies, their dependent establishments and all of their employees. Third, we present representativity analyses that examine the selectivity of the research dataset.

(E) Quality assessment of linked Canadian clinical administrative hospital and vital statistics death data

Nancy Rodrigues, Canadian Institute for Health Information, Canada

Three Canadian clinical-administrative hospital databases were linked to the Canadian Vital Statistics Death Database (CVSD) to provide information about patients who died following discharge from hospital as well as supplementary information about patients that died in-hospital.

The linked datasets were created to develop and validate health care indicators and performance measures and perform outcome analyses. It was therefore imperative to evaluate the data's fitness for use. Quality was assessed by calculating coverage of deaths for all linked contributors, creating a profile of the linked dataset and analyzing issues that were identified by users. These analyses were guided by an existing Data Source Assessment Tool, which provides a set of criteria that allow for assessment across five dimensions of quality.

Variables had good data availability with rates of 95% or higher. 1.4% of acute hospital deaths had discrepancies in the death date captured in the two linked sources; the vast majority had a difference of only one day. A user group and issue tracking process were created to share information about the linked data and ensure that issues are triaged to the appropriate party and allow for timely follow up with the data supplier.

A guided data assessment ensured that strengths and limitations were identified and shared to support appropriate use. Feedback to the data supplier is supporting ongoing improvements to the linkage methodology.

(E) Inverse Probability Weighting to Correct for Outcome Misclassification Using Linked Administrative Datasets

Christopher A. Gravel, Kristian B. Filion, Pauline M. Reynier and Robert W. Platt, McGill University and Lady Davis Institute, Canada

Large observational real-world heath care data, such as claims data, can be used for post market drug safety and effectiveness studies. The use of inverse probability of treatment (propensity score) weights can be used to address measured confounding in studies of this nature, under the assumption of accurate measurement of the outcome variable.

Many of these datasets suffer from systematic outcome misclassification due to coding and/or recording errors. As an example, studies have demonstrated the potential for incomplete ascertainment of cardiovascular outcomes in the Clinical Practice Research Datalink (CPRD) - a repository of general practice information from the United Kingdom. We introduce a novel set of weights that can be used in conjunction with propensity score weights to produce consistent estimation of the marginal causal odds-ratio in the presence of binary outcome misclassification (diagnostic error) using internal validation information.

To acquire a source of internal validation for the outcomes housed in the CPRD, we use linked hospital records taken from the Hospital Episode Statistics (HES), a database containing hospital admission records for patients in the UK. We then exemplify the proposed weighted approach using an example studying post myocardial infarction statin use and the one-year risk of stroke. We compare our findings to the results of a meta-analysis of randomized clinical trials. Finally, we present simulation studies investigating the properties of the proposed weighted estimator, including the impact of model selection on bias reduction and variability.

Session 7A -- Data fusion and record linkage

(E) Simple linkage-data linear regression methods for secondary analysis
Li-Chun Zhang and Tiziana Tuoto, University of Southampton and Istat, UK and Italy

Unless a unique identifier is available, linkage of two separate datasets will generate errors that can cause bias of the subsequent analysis, if the linked data are treated as if they were truly observed. We consider linear regression from the perspective of secondary analysts, who are only given the linked dataset, but not the unlinked records in the two datasets. Moreover, we assume that the analyst has neither access to all the linkage key variables nor the details or tools of the actual linkage procedure, but is instead provided with some non-confidential information about the record linkage precision.

We develop some simple linkage-data linear regression methods, and compare them to the prevalent existing frequentist methods. In particular, these existing methods are constructed under the assumption that the two separate datasets form a complete-match space, so that each record in either dataset is a true match to one unique record in the other dataset. Our approach relaxes this assumption, allowing for the more realistic situation where either dataset may contain some non-matched records.

Given the lack of information about the underlying measurement error mechanisms that have caused the linkage errors, all secondary analysis methods must make an assumption of non-informative linkage errors. We propose a diagnostic test of non-informative linkage errors, which can be performed by the secondary analysts. The test can be useful in practice, for deciding whether the actual regression analysis is acceptable.

All the proposed methods will be illustrated using publically available linkage datasets.

(E) Statistical inference from multiple data files linked by probabilistic record linkage
Partha Lahiri and Ying Han, University of Maryland, USA

Probabilistic record linkage (PRL) methods are frequently used by government statistical agencies to quickly and accurately link two or more large files that contain information on the same individuals or entities using available information, which typically does not include unique, error free identification codes. Because PRL utilizes already existing databases, it enables new statistical analysis without the substantial time and resources needed to collect new data. Linkage errors are inevitable when combining multiple files using PRL because of

the unavailability of error-free unique identifiers. Even a small amount of linkage errors can lead to substantial bias and increased variability in estimating parameters of a statistical model. The importance of incorporating uncertainty of record linkage into statistical analysis cannot be overemphasized. We develop a theoretical framework for statistical inference using a general integrated model that includes a linkage model to incorporate uncertainty due to the probabilistic record linkage process and a mixture model to estimate parameters of the linkage model. In order to reduce the computational burden, simplified version of the proposed methodology and numerical algorithm are provided. I will end the talk by raising several challenging problems in this growing important research area.

(F) Pairwise Estimating Equations for the primary analysis of linked data

Abel Dasylva, Statistics Canada, Canada

A new estimating equation methodology is proposed for the primary analysis of linked data, i.e. an analysis by someone having an unfettered access to the related microdata and project information. It is described when the data come from the linkage of two registers with an exhaustive coverage of the same population, or from the linkage of two overlapping probability samples, as when the said registers have some undercoverage. This methodology accounts for the uncertainty about the match status of record pairs, with a mixture model for the marginal distribution of the agreements vector in a pair. It relies on the conditional independence assumption between the agreements vectors and the responses given the covariates.

Session 7B -- Use of administrative or alternative data sources I

(E) Practical issues in linking data: Experiences with the Linked File Environment (LFE) and its future infrastructure.

Peter Timusk, Julio Rosa, Statistics Canada, Canada

TBA

(F) Moving from a census to tax sources: changing the survey frame to better coordinate INSEE samples

Thomas Merly-Alpa, Institut national de la statistique et des études économiques, France

As a result of greater availability and better quality of administrative sources from taxation authorities, the Institut National de la Statistique et des Études Économiques (INSEE) has developed an alternative survey frame: the Fichier Démographique sur les Logements et les Individus (Fidéli), which comprises tax information, adjusted for duplicates and administrative management rules.

Using Fidéli instead of INSEE's census of population, the survey frame traditionally used at INSEE, will reduce the size of survey zones and improve their sampling design. However, as a result of this change, information disappears, concepts are modified and new variables emerge, in particular to identify surveys that can be based on more precise geolocation.

The renewal of the master sample, planned for 2020 under the Nouvelle Application Utilisée pour le Tirage des Individus et des Logements des Enquêtes (Nautile) project, as well as the sample of the Emploi en Continu (EEC) survey, are based on this file, which has the characteristics of a good survey frame, such as completeness.

Coordinating these two samples has been studied with a view of limiting interviewer travel. The primary units in the master sample are drawn by spatially balanced sampling; the EEC sample is a set of compact clusters of about 20 dwellings also selected by spatially balanced sampling on proxy variables based on the labour market situation. They are coordinated through the introduction of coordination units (CUs) selected by indirect sampling via primary units, with the EEC clusters selected in the CUs in a final phase.

(E) A cross-border traveller's tale: Do I stay or do I go?

Tanja Armenski and Zdenek Patak, Statistics Canada, Canada

Inclement weather can affect cross-border vehicular traffic flows in a number of ways. Weather events such as rain, freezing rain, snow and temperature extremes may lead to large fluctuations in the volume of vehicles entering and exiting the country. This article tries to develop a better understanding of the impact of weather on cross-border traffic flows by integrating weather and traffic data.

The research is conducted using traffic data collected by Canadian Border Service Agency (CBSA) and used by Statistics Canada as an administrative source of Frontier Counts. Weather data were obtained from Environment and Climate Change Canada. Weather parameters such as mean temperatures, rainfall and snowfall are obtained from individual weather stations throughout Canada.

To explain the variations in cross-border traffic series, ARIMA models with weather and holiday related regressors were used. Outbound traffic, Canadian vehicles returning from the USA, and inbound traffic, the USA vehicles entering Canada, are analyzed separately. Practical implications are discussed, and recommendations for further research are provided.

(E) Using multiple sources of data to create and refine geographic aggregations for sub-county surveillance

Angela K Werner and Heather Strosnider, Division of Environmental Health Science and Practice, National Center for Environmental Health, Centers for Disease Control and Prevention, USA

The Center for Disease Control and Prevention's National Environmental Public Health Tracking Program's (Tracking Program) mission is to provide information from a nationwide network of integrated health and environmental data, driving actions to improve the health of communities. The Tracking Program plans on regularly disseminating data at a higher geographic resolution to improve environmental health surveillance and help drive more local-level changes. When displaying higher resolution data, several considerations, such as stability and suppression of those data, must be taken into account. This requires temporal aggregation and/or new aggregated geographies to minimize suppression and instability in displays while remaining classified as sub-county. The method to create these geographies must be standardized so the geographic units will be comparable across states for use in a national surveillance system.

Using multiple sources of data, including census tract boundaries, health data, and population data, optimal aggregations were created for two aggregation schemes (i.e., a rare outcome aggregation scheme and a more common outcome aggregation scheme) for a set of pilot states. An initial review of the new aggregations and consultations with states revealed several issues such as cross-county merging, variations in merges, and geographic units with larger populations than needed. After establishing suitable population thresholds for the two aggregation schemes, an alternative method of merging using population-weighted centroids was explored. Future work includes further refinement of the aggregated geographies by

addressing some of the challenges that were encountered and exploring the use of additional factors in the aggregation process.

(E) Next Generation Data Governance Technologies for Official Statistics

Ryan White, Statistics Canada, Canada

This paper will present proof-of-concept outcomes of key technologies for computing, analysis and data models which are used in an administrative data preprocessing and data validation framework. The technologies and preprocessing framework facilitate sound data governance and reproducible data science based on open-source, cloud native principles. Benchmark data pipelines for record linkage using synthetic data sources are presented which are based on container technology and elastic cloud computing clusters. Modern and innovative data science tools are used for data and processing management that provide reproducibility and provenance of data and processes. The benchmark data pipeline is an automated, reproducible data process that serves as a process profiling and research tool for in-memory data analytics and scalable, distributed data analysis. Comparisons of data processing using traditional row-oriented datasets (csv) to industry standard columnar data format organized for efficient analytic operations on modern computing hardware is presented. The prototype data preprocessing and validation framework serves as a data modeling, profiling and conversion engine of administrative data that demonstrates a data strategy for the next generation of reproducible data science and data analytics at Statistics Canada.

Session 8A -- Methods in Demography

(F) Projecting the level of literacy using a microsimulation model

Samuel Vézina and Alain Bélanger, Institut national de la recherche scientifique, Canada

The purpose of this article is to present a module to project the literacy level of adults (aged 25 to 64 years) using the LSD-C microsimulation model. This model projects the Canadian population by demographic variables (age, sex, place of residence, place of birth, generation status and immigration status), ethnocultural variables (mother tongue, language spoken most often at home, knowledge of official languages, visible minority group, religion) and socioeconomic variables (education, labour force status).

An in-depth analysis of data from three cross-sectional surveys on the skills of adults in Canada was conducted before this modelling and projection exercise. This analysis, stratified by immigration status, helped to identify the factors that determine the literacy level of the population and to judge the comparability of the data over time (pseudo-longitudinal measure of the cohort effect).

The method chosen is simple because the available data cannot be used to dynamically model the trajectory of the population's literacy level. The literacy score is imputed for the simulated cases based on several characteristics: age, sex, region of residence, education, language skills, labour force status, age at immigration, duration of residence in Canada, country of birth and country of graduation. The score of the individual is recalculated each time there is a change of status in one of the above characteristics. It is possible to calculate the average score of the total adult population and to measure the impact of sociodemographic changes projected on this average score.

This method has been applied to another microsimulation model, PÖB, which projects Austria's population.

(E) Putting admin data at the core of the Population Statistics System – the experience in England and Wales

Rebecca Tinsley, Office for National Statistics, United Kingdom

The Office for National Statistics (ONS) has been using integrated data to research into the Government ambition that "censuses after 2021 will be based on alternative sources of data." The Administrative Data Census project has made progress in: producing a range of Research Outputs that are typically produced by the ten-yearly census; comparing these outputs with official statistics; and seeking feedback from users.

Research so far has covered a range of outputs including the size of the population, household statistics, and a range of population characteristics. These outputs have used a range of integrated methods and data sources including administrative data, commercial (aggregate mobile phone data) and survey data.

Recently, we have shifted the focus of our research to understand how admin data can be used to produce not only the stock count of the population, but also components of population change, including estimates of international migration.

This is a key milestone in an ambitious programme of work to transform to a new administrative data-led system for population and migration statistics for England and Wales by Spring 2020.

This work builds and expands upon our research on an Administrative Data Census to move into a routine understanding of population and migration using all available sources.

This presentation will cover:

Progress made so far

Challenges in estimating coherent stocks and flows of population from admin data, including outcomes from collaboration at an international workshop on the same topic

Insights into user feedback

(E) The Demographic Characteristics File (DCF): Assigning Race and Hispanic Origin to Domestic Migrants using Census and Federal Administrative Data Sources

Amel Toukabri, U.S. Census Bureau, USA

Internal migration between states and counties is a large driver of annual change in the official subnational population estimates published by the U.S. Census Bureau. The Census Bureau uses a rich constellation of federal administrative and decennial census data to produce domestic migration estimates by demographic characteristics. Data from annual federal tax returns and the Social Security Administration are linked by person to census data to provide a demographic portrait of state and county movers by age, sex, race, and Hispanic origin. The result is a master file of individuals in each migration period referred to as the Demographic Characteristics File (DCF). In this paper, we present the method behind the DCF with a particular focus on a multi-staged imputation process for cases with missing race and Hispanic origin. Compared to the previous method, the DCF improves the age distribution by race for children born after Census 2010. We will present distributional differences by age and race in county population between the DCF and the previous method and highlight a few case studies. Finally, we close with limitations, next steps, and future research directions.

(E) The Utility of Using Web Surveys to Measure and Estimate Major Health Outcomes, A Pilot Study.

Yulei He, Hee-Choon Shin, Bill Cai, Van Parsons, Peter Meyers and Jennifer Parker, Division of Research and Methodology National Center for Health Statistics U.S. Centers for Disease Control and Prevention, USA

Given the increasing cost and resource constraints of traditional surveys, web surveys have been frequently used as a flexible alternative in the field. However, past literature has well documented the general limitations of using web surveys. There also exists vibrant research on improving their uses from both applied and methodological perspectives. One major question is whether the web surveys can produce (either directly or being calibrated) national, official estimates of major health outcomes (e.g. diabetes prevalence, insurance coverage). The National Center for Health Statistics at US Centers for Disease Control and Prevention has conducted a pilot study of assessing the utility of using web surveys to accurately measure and estimate important health outcomes, in conjunction with the National Health Interview Survey (NHIS). NHIS can be viewed as a reference data source for producing national estimates of major health outcomes in the US. This talk will present the background and some initial study results. Specifically, we compare the estimates from the two sources in general and across key subgroups. We also explore advanced statistical methods for calibrating the web survey estimates using NHIS as the yardstick.

(E) Data integration method: a consolidation of semantic heterogeneity and data sources with the England and Wales custodial policy evaluation project

Marie-Eve Bedard, Statistics Canada, Canada

This research will highlight the methodological issues that arose from the use of multiple data source for the research project done on the 2014 England and Wales evaluation of safety policies performances in custody. The project used several data sources, such as administrative data, survey data, key performance indicators, and Prison Quality Model data gathered by researchers. Since the concept measurements surrounding the data were taken from different sources and some data were not collected on both time points, there is a lack of certitude on cause and effect of the policy. Using administrative data has also brought other issues to the surface. One of the main issues was the limitations of the sample, since the sample is predefined by the administration, subsequently the researcher is working with a pool of individuals that is predetermined and a different one at two time points. The same issue was observed with the survey data. A method was developed to consolidate these data sources and their semantic heterogeneity for evaluation purpose, by means of residual change score analysis, principle component factor analysis, robust standard error regression and scales. These methods have been combined with other methods, such as Cook's distance, y computation and variance inflation factor test in order to unify these data and make them comparable for this kind of evaluation. With this type of data integration method, although successful, the data analysis remained with several limitations, which leaves room for further research into finding a way to close the gap between these data sources.

(E) Combining Unit and Area Level Data for Small Area Estimation via Penalized Multi Level Models

Jan Pablo Burgard, Joscha Krause and Ralf Münnich, Trier University, Germany

Small Area Estimation (SAE) is widely used to obtain area quantity estimates in the presence of small samples. Modern SAE applications are based on unit or area level models, depending on data availability on the respective topic. A unit level model considers data below the area

level for model parameter estimation, while an area level model uses data on the area level. If data on both levels is available, it should be combined in order to maximize the explanatory power of the underlying model and improve area quantity estimates relative to the consideration of only one data level.

However, the combination of unit and area level data raises methodological issues. Linking the information demands model parameter estimation on both levels. Thus, the number of estimands increases, which may destabilize area quantity estimates due to a lack in degrees of freedom. Further, unit and area level data have different distributional characteristics, for example in terms of dispersion patterns and covariance structures within the covariates. Therefore, the different data sources should not be treated as equal within the estimation process.

We investigate penalized multi level models to solve these problems and combine unit and area level data to improve area quantity estimates relative to standard SAE methods. Multivariate l1-norm, l2-norm and elastic net penalties are used for level-specific regularization to balance the different data sources and produce optimal model predictions. An empirical application in social medicine is provided by combining German survey and micro census data.

(E) Quality Measures for the Monthly Crude Oil and Natural Gas (MCONG) Report

Evona Jamroz, Lihua An, and Sanping Chen, Statistics Canada, Canada

The Monthly Crude Oil and Natural Gas (MCONG) program is a critical component of Canada's monthly gross domestic products. It brings together three categories of input data: data reported by multiple "feeder" surveys, administration data from government agencies, and historical allocation data based on "expert opinions." A new system environment is being built for the MCONG program by integrating the data from the above three sources. In this paper, we summarize our ongoing work and remaining challenges for developing quality measures for the estimates in the new MCONG program.

For the three data sources, the government administrative data are provided in macro format, for which we assume no error. For the survey data, the variance due to sampling and/or imputation can be estimated using conventional methods. A particular challenge is to estimate the error associated with a parameter that is based on expert opinion. We propose a Bayesian approach for such a parameter.

We then propose a process largely based on Taylor linearization to integrate these variance components into a single coefficient of variation (CV) for the final MCONG estimates. The situations in which CV is not an adequate quality measure will also be discussed.

(E) Findings from the Integrated Data Workshops hosted by the Federal Committee on Statistical Methodology and Washington Statistical Society

Alexandra Brown, Andrew Caporaso, Katharine G. Abraham and Frauke Kreuter (Presenter: Linda Young), University of Maryland, USA

Across Federal Statistical agencies, there has been growing interest in integrating traditional survey and census data with auxiliary data (both structured and unstructured) in order to increase the quality and the timeliness of the data and statistics that are produced. To address the gap that exists in understanding the quality of such datasets (as compared to sample surveys) and improve communication between data producers and data users, the Federal Committee of Statistical Methodology (FCSM) and the Washington Statistical Society (WSS) co-hosted three workshops that explored the current practices of transparent reporting on the

quality of integrated data. This report summarizes the three workshops and pulls together the key themes.

Session 9A -- Data Access Issues - Privacy and Confidentiality in the era of multiple sources

(E) The Framework behind the Confidentiality Classification Tool and its Role in Modernization

Joseph Duggan, Michelle Marquis, Claude Girard and Jack Gambino, Statistics Canada, Canada

With the Confidentiality Classification Tool, Statistics Canada is implementing a small, but key supporting component of its recent Modernization Initiative. Sensitive statistical information (SSI) is now being classified along a continuum of risk, replacing the traditional binary classification that underpinned its two separate working environments: Network A for internal use and processing of protected information, and Network B for external communication and the dissemination of our statistical products. The combination of this change along with other initiatives - relating to the use of alternative and combined data sources, modernized access to microdata, and bringing together more partners in collaborative efforts - will align our Disclosure Control practices with current and future IT Infrastructures. The tool seeks to facilitate all of this by increasing awareness of confidentiality issues and practices, while helping data custodians determine a level of confidentiality for any selected data holding in Statistics Canada. This paper describes the methodology behind the intentionally-simple Confidentiality Classification Tool and the lessons that were learned in its development.

(E) Why the U.S. Census Bureau Adopted Differential Privacy for the 2020 Census of Population and Housing

John M. Abowd, U.S. Census Bureau, USA

The U.S. Census Bureau acknowledges that database reconstruction attacks as defined by Dinur and Nissim (2003) are a real vulnerability of the disclosure avoidance systems used to protect all statistical publications from previous decennial censuses. Our own research confirms the vulnerability, which is now designated as an enterprise issue. Differential privacy was invented in 2006 (Dwork et al. 2006) to address this vulnerability and to insure that statistics produced from confidential data sources could be protected in a principled way. These formal privacy systems have two properties that traditional disclosure limitation systems lack: (1) they are closed under composition and (2) the protection does not degrade with post-processing. The composition property means that the privacy-loss from a sequence of differentially private algorithms applied to the same data is no greater than the sum of the privacy losses from each component. These properties imply that statistics protected with differential privacy satisfy a known, computable privacy-loss budget, which is the correct global disclosure risk measure. They also imply that the disclosure protection is “future proof”—its strength does not depend on assumptions about current or future information sets or computational abilities of the data user. The next United States decennial census, to be taken in 2020, will feature published data products protected by differential privacy. This protection is being engineered subject to a global privacy-loss budget to be determined by the Census Bureau’s Data Stewardship Executive Policy Committee. Each statistic’s fitness-for-use will be directly measured, inclusive of the uncertainty due to statistical disclosure limitation.

(E) Implementing Privacy-preserving National Health Registries

Rainer Schnell and Christian Borgs, University of Duisburg-Essen Lotharstr, Germany

Most developed nations operate health registers such as neonatal birth registries. These kind of registries are important for medical research applications, for example, follow-up studies of cancer treatments. Linking such registers with administrative data or surveys offers research opportunities, but may rise privacy concerns. Due to the recent harmonization of data protection rules in Europe with the General Data Protection Regulation (GDPR), criteria for operating such registers in a privacy-preserving way can be derived.

A health data register used for linking needs to be secured against re-identification attacks, while retaining high linkage quality. We will demonstrate solutions providing strong resilience against re-identification attacks while preserving linkage quality for research purposes. Several state of the art privacy-preserving record linkage (PPRL) techniques were compared during the development. For real world testing, we matched mortality data from a local administrative registry (n = 14, 003) with health records of a university hospital (n = 2, 466). Scaling of the proposed solutions was tested by matching 1 million simulated records from a national database of names with a corrupted subset(n = 205, 000).

Re-Identification risk of different implementations will be discussed, considering recently developed new attack methods. Finally, we will give detailed recommendations for operating privacy preserving health registries for linkage, including operational guidelines and best-practice suggestions.

Session 9B -- Use of administrative or alternative data sources II

(E) Combining Census and Manitoba Hydro data to understand residential electricity usage

Chris Duddek, Statistics Canada, Canada

Statistics Canada has received monthly Manitoba Hydro files from 2015 onwards. Since Manitoba Hydro is the electricity provider for most Manitobans, it would seem easy to obtain total annual residential electricity usage. When compared to published estimates, however, there are discrepancies. To better understand why, we modify geographic variables on the file in order to compare the Manitoba Hydro data to 2016 Census counts. The comparison allows us to pinpoint issues arising from the way the Hydro file data is structured. A brief cross sectional analysis of the data is followed by a look at longitudinal elements of the data. The paper ends by outlining the steps necessary to rectify the problems in order to estimate total annual electricity usage.

(E) Valuing Wellbeing Impacts with the General Social Survey

Xavier Lemyre, Department of Canadian Heritage, Canada

The Three-Stage Wellbeing Valuation approach is used to value the impacts of participation in activities such as going to the theater on life satisfaction. The method measures the amount of money that would be necessary to make an activity participant as well off in the absence of his participation as when participation occurs. Due to issues of endogeneity, the approach relies on an instrumental variable framework to measure income's effects on wellbeing. Moreover, the approach estimates activity and income effects on life satisfaction in separate stages, which permits the use of different datasets for separate stages.

In this presentation, we will examine how simultaneous equations are used to estimate monetary values that express the wellbeing effects of the participation in arts, culture and sports activities. New applications that are made possible due to linked administrative data and detailed microdata access will be presented, as well as opportunities for further research.

(E) Towards a Register-centric Statistical System: Recent Developments at Statistics Canada

Jean Pignal, Philippe Gagné, Christian Wolfe and Tanvir Quadir, Statistics Canada, Canada

The Statistical Register Transformation and Integration Project aims to build and maintain a statistical registers infrastructure (SRI) that comprises core interconnected registers (population, building, business and activity) based on administrative data. The goals of the SRI are to maximize the use of actionable information from the data already collected, improve the timeliness of statistics, and reduce response burden while protecting the privacy of Canadians. This presentation will cover the following: the methodology and framework for the ongoing development of de-identified population and building registers, the redesign of the existing Business Register, and the conceptual framework of an activity register. The presentation will also include an outline of the Canadian System of Integrated Statistical Registers (CSISR), which assembles the underlying core registers and lays the foundation for the register infrastructure.

(E) Replacing Census questions using linked administrative data

Scott McLeish and Athanase Barayandema, Statistics Canada, Canada

Beginning in 2011, administrative data from Immigration, Refugees and Citizenship Canada's (IRCC) has been linked to census data for the purposes of informing edit and imputation processing, as well as certification of census results. This administrative immigration data includes characteristics at landing for all immigrants who have landed in Canada since 1980, with some data going back to 1952. As a result of linking to this administrative data, two variables, immigrant admission category and applicant type, were added to the 2016 Census for the first time.

These linkages to administrative data sources have highlighted the quality of current census questions related to immigration, including a better understanding of errors due to non-response and measurement. For example, the relationship between the number of years immigrants have resided in Canada and the accuracy of their responses has been established.

This presentation will outline a study examining the feasibility of replacing the immigration questions on the 2021 Census using linked administrative records. It will cover the strengths and limitations of both the status quo and the use of the linked administrative data, including issues associated with data integration, such as inconsistencies between data sources.

(E) Automated Methods of Data Collection: Retail locations and shopping centres in Vancouver

Paul Holness, Statistics Canada, Canada

Retail sales estimates are one of the leading economic indicators used by the Bank of Canada and the business community to generate strategic policy, guide investment decisions (strategies) and assess economic performance. The latest figures show that the retail industry (NAICS 44-45) contribution to GDP was approximately $90.5b in 2014 representing 4.9% of total GDP in Canada. The retail sector is the largest industry in Canada, employing close to two million people. Retail big box stores and retail chains are among Canada's top importers and in recent years, more and more of these fix point-of-sales retailers have begun to scale up their e-commerce offerings.

The Monthly Retail Trade Survey (MRTS) collects sales, e-commerce sales, and the number of retail locations by province, territory, and selected Census Metropolitan Areas (CMA) from a sample of retailers. This paper uses web-based collection methods to investigate whether

the Monthly Retail Trade Survey (RTMS) survey frame can be automatically generated using data from Google Web Services Application Programming Interface (API) and the Canadian Directory of Shopping Centres Directory (CDSC) Register.

The new application uses reference data collected manually from the Vancouver CMA from June 2017 and recently extracted web-based data to develop a prototype which features an executive dashboard design as the presentation layer of the system. The proposed system leverages automated business services from Google Web Services, Nearby Search and Text Search, and the Web-scraping tools Scrapy and Selenium to extract information from the Canadian Directory of Shopping Centres (CDSC). Together, these technologies provide direct access to source data and reduce the instances of manual intervention in data acquisition. The extracted JSON data is easily parsed into comma separated vector (CSV) datasets that are easily read into SAS or Excel. Finally, results showed that the web-based methods produced a comparable population of retail locations and required significantly less resources compared to the manual process.

Our study provides detailed quantitative information to allow managers to: (1) evaluate the use of automated methods in the preparation, maintenance and evaluation of MRTS survey frame; (2) Asses the impacts of such methods on the quality of retail location attributes on Statistics Canada's Business Register (BR); and (3) Estimate the overall cost and efficiency of these methods on the RSID program.

Session 11 -- Plenary Session

(E) Measuring Uncertainty with Multiple Sources of Data

Sharon Lohr, Arizona State University, USA

In a probability sample with full response, the margin of error provides a reliable, theoretically justified measure of uncertainty. When combining estimates from multiple samples or administrative data sources, however, traditional margins of error underestimate the uncertainty---differences among statistics from various sources often exceed the estimated sampling variability. I examine methods that have been proposed for measuring uncertainty from combined data sources, with application to estimating smoking prevalence and sexual assault rates, and outline some possible directions for research.