Conference in honour of Fred Smith and Chris Skinner

Online, 7-9 July 2021

Conference abstracts

Wednesday 7 July

Invited Session 1

Danny Pfeffermann University of Southampton, CBS Israel and Hebrew University of Jerusalem
Time series modeling of repeated survey data for estimation of finite population parameters

In the first part of the article, I review and discuss the pioneering contributions of the late Alastair Scott and T.M.F Smith to time series analysis of repeated survey data. In the second part, I review and discuss the extensive theoretical and applied developments in this area, emerging from their work over the ensuing 40 years or so.

Contributed Session 1

Roger Sugden
IID case under Poisson sampling

Fred Smith and I last worked on inference under informative sampling designs in 2012 and he presented some results at a conference in Hong Kong. Here I show again in the IID case under Poisson Sampling that the 'full likelihood' reduces to the 'sample likelihood' of Pfeffermann *et al*. but only when we condition on *s*, the observed labels or equivalently the observed sample size *n*. Their 'sample complement distribution' for unobserved responses also emerges from this conditioning. I briefly discuss whether this conditioning is justified and also consider the extension of this analysis to the more realistic Conditional Poisson Sampling.

Jan Pablo Burgard Trier University and **Stefan Zins** Institute for Employment Research (IAB) of the German Federal Employment Agency (BA)
Planning domain sizes in cluster sampling

Multi-stage cluster sampling is a common sampling design of social surveys because populations of interest are often structured by, or partitioned into, disjoint organizational and administrative units. The need to use cluster sampling can conflict with survey planners' goal to select a sample that contains a specific number of elements from certain domains of interest. This can be a complex problem if sampling units, i.e. clusters, cut across the domains of interest, as it is often the case. For example, an analysis require sufficient observations from certain age and gender categories. But the population is clustered within schools, hospitals, establishments, or municipalities and hence age-gender categories cannot be used for stratification.

We propose a quadratic optimization approach to define inclusion probabilities that can be used for drawing balanced cluster samples that comply with predefined sample sizes from domains of interest. Henceforth the clusters may cut across domains. We also provide an application of the proposed solution to the domain size problem for an existing social survey on migration and emigration in Germany.

Guillaume Chauvet ENSAI and **Audrey-Anne Vallée** Université Laval

Inference for two-stage sampling designs

Two-stage sampling designs are commonly used for household and health surveys. To produce reliable estimators with associated confidence intervals, some basic statistical properties like consistency and asymptotic normality of the Horvitz-Thompson estimator are desirable, along with the consistency of associated variance estimators. These properties have been mainly studied for single-stage sampling designs.

In this work, we prove the consistency of the Horvitz-Thompson estimator and of associated variance estimators for a general class of two-stage sampling designs, under mild assumptions. We also study two-stage sampling with a large entropy sampling design at the first stage, and prove that the Horvitz-Thompson estimator is asymptotically normally distributed through a coupling argument. When the first-stage sampling fraction is negligible, simplified variance estimators which do not require estimating the variance within the Primary Sampling Units are proposed, and shown to be consistent.

An application to a panel for urban policy, which is the initial motivation for this work, is also presented.

Marcel D.T. Vieira Universidade Federal de Juiz de Fora, **Loveness Dzitiki** and **Brendan Girdler-Brown** University of Pretoria

Approaches for combining data collected from multiple simple random samples

It is sometimes desirable to combine data from different surveys either to increase the sample size and precision of estimates for sub-populations of interest or to improve coverage, when compared to individual surveys. Large surveys may not have enough sampling units for subgroups of interest to allow any meaningful inference. Even though there is substantial literature on studies that pool survey data, it is still not clear which are the most efficient methodologies and sampling designs for pooling data from different surveys. It is important to know whether the estimates from the surveys involved should be given equal weights in the calculation of the combined estimate or not. If they are not given equal importance, then it should be clear how they should be weighted and why.

There are two main approaches which may be used when combining surveys, which are the separate and the pooled approaches (Wannell and Thomas, 2009; Roberts and Binder, 2009). In the pooled approach, individual records from the surveys are combined. Original survey weights may be modified, and estimates may then be calculated based on the new weights and the pooled sample. In the separate approach, estimates are obtained from each of the surveys separately.

In this paper we consider the separate approach and explore alternative sampling designs and importance measures for the different surveys involved in the estimation of the combined estimate, including the sample size, the variance, and the coefficient of variation (cv) of the estimator, and the misspecification effect (Skinner, 1986).

Current and proposed methods are evaluated through simulation, in the context of simple random sampling without replacement, stratified random sampling and two stage cluster random sampling from finite populations generated from alternative super-population models.

Simulation results suggest superpopulation variance does not influence the choice of weighting method. However, the population size appears to influence this choice. Combining samples improved the precision of estimates regardless of weighting method used for all sampling techniques. Moreover, combining stratified samples led to more efficient estimates than combining simple random samples and 2 stage cluster samples, which presented the poorer results. Our illustrative examples considering data from the South African Community Survey and from the South African General Household survey confirmed combining samples yield better estimates compared to using estimates obtained from one sample survey.

Contributed Session 2

James Jackson Lancaster University, **Robin Mitra** Cardiff University, **Brian Francis** Lancaster University and **Iain Dove** Office for National Statistics

Developing synthetic data methods for large, sparse administrative databases

When releasing data, respondents' privacy is protected through statistical disclosure control (SDC). Professor Chris Skinner made several key contributions to this field, notably his work on disclosure risk in microdata (Skinner and Elliot, 2002).

Over the past three decades, the use of synthetic data (Rubin, 1993; Little, 1993) for SDC has continually developed. Methods have adapted to account for different data types, but mainly within the domain of survey data sets.

Administrative databases are being increasingly considered as a way to fuel researchers' demand for data. Such databases are inherently confidential: respondents provide their information for administrative purposes, not for statistical analyses. There are characteristics of administrative databases which present challenges for synthetic data generation. First, these databases tend to be comprised of categorical variables, some with many categories which, when converted from microdata into frequency tables, gives rise to large, sparse tables. Also, these tables can have structural zeros, that is, unobservable combinations of levels.

The categorical nature of administrative databases allows the synthesis to be undertaken at

the tabular level rather than at the individual level. We will show that the fitting of saturated models allows administrative databases to not only be synthesized quickly, but also allows risk and utility to be formalised in a manner inherently unfeasible in other techniques. The flexibility afforded by multi-parameter count distributions, such as the negative binomial, can be utilised to protect respondents' privacy. The synthesis distribution's parameters can be adjusted to achieve, analytically, certain criteria post-synthesis, for example, the probability that a count of one is synthesized to one, which is equivalent to saying that a unique in the original data is also a unique in the synthetic data.

Finally, we give an empricial example, synthesising a database that can be viewed as a surrogate to the English School Census. (This data set was constructed using information from various sources, primarily 2011 census tables.)

Li-Chun Zhang University of Southampton and **Gustav Haraldsen** Statistics Norway

Secure data collection and processing: means and opportunities

“Survey respondents are usually provided with an assurance that their responses will be treated confidentially. These assurances may relate to the way their responses will be handled within the agency conducting the survey or they may relate to the nature of the statistical outputs of the survey…” (Skinner, 2009). Statistical output disclosure limitation with respect to confidentiality constraints has long been an important topic, when it comes to the dissemination of statistical tables and, in particular, microdata. Many techniques for assessing various forms of identification risk and for masking of the released statistical outputs have been developed, including, it seems, a surge of interest in differential privacy in the last decade.

Meanwhile, there is an increasing need for greater protection of confidentiality as the National Statistical Institute (NSI) collects and processes data, en route to the statistical outputs. One can easily identify at least two main drivers behind the pressure. On the one hand, the ever-more digitalised life form has created a multitude of big non-survey data sources, offering potentially many opportunities for better, quicker or richer statistical outputs, which would have been either extremely demanding or simply infeasible based on surveys in the traditional sense. On the other hand, the General Data Protection Regulation (GDPR) requires the businesses and agencies that handle personal data must implement measures to safeguard the data, which at once raises barriers for the NSI who would like to collect and use such data for statistical purposes.

To address this need we develop a compartmented data collection and processing system, which can greatly reduce the identification risk throughout the statistical value chain. The design is inspired by the relevant ideas and concepts underlying the so-called trusted execution environment (TEE) which, strictly speaking, refers to a secure area (enclave) of the main processor of a device, such as smart phone, whose memory or execution state is invisible to any other process, including the device's operating system. Similarly, on entry to our compartmented system, any dataset is subdivided into separate packages of unit IDs and fragmented attributes, and critical processes are isolated from one another and inaccessible to the process owner. The details will be described in the paper.

The proposed system aims to gain acceptance from data owners, stake holders and the public, thereby smooth access to the many useful big-data sources, including various transactions and remote sensing signals. As we shall discuss, such data of digital traces differ to survey data fundamentally in several respects. Indeed, in many contexts, they can be both more accurate and richer than survey data. Thus, instead of considering the undertakings required of the secure system merely as extra troubles, we prefer to welcome them as means to better and richer data that can both improve and enlarge the official statistics outputs. Taking transaction data as examples, we shall outline a number of topics, where transaction data can either replace or enhance the existing sample surveys, or provide new statistics that are currently infeasible via surveys.

Mark Bun Boston University, **Jörg Drechsler** Institute for Employment Research, **Marco Gaboardi** University of Maryland and **Audra McMillan** Northeastern University

Controlling privacy loss in survey sampling

In social science research, surveying an entire population is typically infeasible due to time and budgetary constraints. As a result, much of the data collected by statistical agencies are based on surveys performed on a random sample of the target population. While the main motivation for sampling is often financial, a commonly held belief is that sampling provides additional privacy. Intuitively, data subjects are afforded some plausible deniability by the fact that they may, or may not, have been sampled. This intuition of increased privacy from sampling has been formalized for some types of sampling schemes (such as simple random sampling with and without replacement or Poisson sampling) in a series of papers in the differential privacy literature. These types of privacy amplification by sampling results are useful for a variety of reasons including producing more accurate results at a target privacy level, and for providing additional incentive for participation in a survey study.

However, simple sampling schemes like simple random sampling with or without replacement and Poisson sampling are rarely used in practice. Statistical agencies typically use more complex sampling designs to increase the accuracy of the results and/or to reduce the costs of data collection. Unfortunately, these more complicated sampling schemes can result in privacy degradation. That is, sampling can actually make privacy guarantees worse.

It is often the case that the design of a survey is based on sensitive historic or auxiliary data. Thus, the survey design itself can leak additional information about the population. We find that this degradation of the privacy guarantee occurs even for reasonably simple sampling designs when the sampling design is data dependent.

We focus on two common sampling schemes: cluster sampling and stratified sampling. We will discuss the implication of these sample designs on the effective privacy loss, in terms of the parameter from differential privacy. For stratified sampling, we will discuss how the data dependent nature of this sampling design results in privacy degradation. We will then discuss a potential method for controlling the privacy loss for proportional allocation, a common sampling design for stratified sampling. For cluster sampling, we will discuss how the lack of independence between the inclusion of different data subjects affects the privacy guarantee. We will show that even for relaxed notions of privacy and "ideal" data, while the privacy does not degrade, privacy amplification is negligible.

Shijie Guo Dartmouth College and **Jingchen (Monika) Hu** Vassar College
Data privacy protection and utility preservation through Bayesian data synthesis

When releasing record-level datasets containing sensitive information to the public, the data disseminator is responsible for protecting the confidentiality of each record in the dataset, while preserving important features in the data for data analysis. This goal can be achieved by data synthesis, where sensitive data are replaced with synthetic data, which is simulated based on statistical models estimated on the confidential data. In this paper, we synthesize price and number of available days in a sample of the New York Airbnb Open Data. We evaluate the data utility and the disclosure risks of our synthetic data. Our main focus is to investigate how intruder's knowledge would influence the identification disclosure risk of the synthetic data. In particular, we explore several realistic scenarios of uncertainties in intruder's knowledge of available information and evaluate their impacts on the identification disclosure risk.

Invited Session 2

Natalie Shlomo University of Manchester and **Chris Skinner** London School of Economics and Political Science
Measuring risk of re-identification in microdata: state of the art and new directions

We review the influential research carried out by Chris Skinner and his co-authors in the area of statistical disclosure control and more specifically quantifying the risk of re-identification in sample microdata. The disclosure risk scenario for the release of sample microdata where the sample is drawn from a finite population is based on the following assumptions: (1) there is an 'intruder' (someone with malicious attempt to discredit the statistical office) who has access to the microdata and other auxiliary information from the population that allows him/her to link data sources in order to identify individuals in the sample microdata; (2) there is no 'response knowledge' meaning that the intruder does not know who was drawn into the sample. The risk of re-identification is conceptualized with respect to population uniqueness – given an observed sample unique on a set of cross-classified (categorical) quasi-indicators, what is the probability that the sample unique is also a population unique. In this framework, the survey design can be accounted for. If the sample is drawn from a population that is known, for example a sample drawn from a census, then the risk of re-identification is straight-forward to calculate. However, in the types of government surveys that we focus on here, samples are generally drawn from address-based sampling frames and the population is not known. Therefore, a statistical modelling framework is needed to estimate the disclosure risk measures. The framework proposed by Chris Skinner and his co-authors is based on log-linear modelling to estimate unknown parameters of the population which are then used to estimate disclosure risk measures assuming a Poisson distribution. We also review extensions to this framework as well as related work by other authors.

We also introduce a new extension to this approach for measuring the risk of re-identification in microdata, but here we focus on a dataset containing a subpopulation in a register that is not representative of the general population, for example a register of cancer patients. In this framework, we use a random reference sample to estimate both the population parameters and the parameters to estimate the propensity to be included in the register. These parameters are then included in an extended probabilistic framework to estimate the disclosure risk measures. This research was initiated during the Data Linkage and Anonymization Programme at the Isaac Newton Institute, Cambridge UK from July-December 2016, but never published until now. We demonstrate this approach in an application study based on UK census data to compare the estimated disclosure risk measures to the known truth. Further work will focus on measuring the risk of re-identification for a sample drawn from the register or more generally for non-probability samples, where the selection mechanism needs to be modelled separately and included in the statistical framework for quantifying disclosure risk.

Jae Kwang Kim Iowa State University, **Jon Rao** Carleton University and **Yonghyun Kwon** Iowa State University

Analysis of clustered survey data based on two-stage informative sampling and associated two-level models

In a pioneering paper, Scott and Smith (1969), proposed a Bayesian model-based or prediction approach to estimating a finite population mean under two-stage cluster sampling. WE provide a brief account of their pioneering work. We also review two methods for the analysis of two-level models based on matching two-stage samples. Those methods are based on pseudo maximum likelihood and pseudo composite likelihood taking account of survey design weights. We then propose a method for the analysis of two-level models based on a normal approximation to the estimated cluster effects and taking account of design weights. This method does not require cluster sizes to be constants or unrelated to cluster effects. We evaluate the relative performance of the three method in a simulation study. Finally, we apply the methods to real data obtained from 2011 Prive Education Expenditure Survey in Korea.

Invited Session 3

David Steel University of Wollongong and **Robert Clark** Australian National University

Sample design for analysis using high influence probability sampling

Sample designs are typically developed to estimate summary statistics such as means, proportions and prevalences. Analytical outputs may also be a priority but there are fewer methods and results on how to efficiently design samples for the fitting and estimation of statistical models. This paper develops a general approach for determining efficient sampling designs for probability-weighted maximum likelihood estimators and considers application to generalized linear models. We allow for non-ignorable sampling, including outcome-dependent sampling. The new designs have probabilities of selection closely related to influence statistics such as dfbeta and Cook's distance. The new approach is shown to perform well in a simulation based on data from the New Zealand Health Survey.

Ray Chambers University of Wollongong, **Setareh Ranjbar** University of Lausanne, **Nicola Salvati** and **Barbara Pacini** University of Pisa

Weighting, informativeness and causal inference, with an application to rainfall enhancement

Sampling is informative when probabilities of sample inclusion depend on unknown variables that are correlated with a response variable of interest. This can be a problem when the sample data analyst only has access to secondary data sources for controlling the impact of the sampling method. When sample inclusion probabilities are available, inverse probability weighting can be used to account for informative sampling in a secondary analysis situation, though usually at the cost of less precise inference. This paper reviews two important research contributions by Chris Skinner that modify these weights to reduce their variability while at the same time retaining consistency of the weighted estimators. In some cases, however, sample inclusion probabilities are not known, and are estimated based on the observed sample. This can be an issue in causal analysis, and double robust methods that protect against misspecification of the sampling process have been the focus of much recent research.

In this paper we propose a simple model-based modification to the popular inverse probability weighted estimator of an average treatment effect, and then illustrate its use in a causal analysis of a rainfall enhancement experiment that was carried out in Oman between 2013 and 2018.

Student Prize Session

(**Loveness Dzikiti** and **Caio Gonçalves** were also winners, but appear elsewhere on the programme)

Dehua Tao Australian National University

Linkage bias adjustment for integrated data sets

By linking the data from one source with another, the resultant integrated data set provides a rich data source for analysis by researchers. However, unless direct and accurate identifiers are available for data linking, integrated data sets may be subject to linking errors, and the resultant analysis subject to linkage bias. In this paper, we outline a method based on the parametric bootstrap to determine if such bias indeed exists, and construct bias-corrected estimators. Because only finite bootstrap samples can be drawn and used to provide Monte Carlo estimates of the linkage bias, the single bootstrap may not be sufficient to remove the linkage bias completely. This paper also provides a method to test if a higher-order bootstrap, e.g. a double bootstrap or triple bootstrap etc, is needed for complete linkage bias removal. Finally, we illustrate the methods outlined in this paper using a simulation study and a real-life example.

Estelle Medous Université Toulouse 1 Capitole and La Poste

Pros and cons of the double indirect sampling for "many to one" links

In probabilistic surveys, when there is no sampling frame for the target population, a solution is to find a frame population linked in some way to the target population and use indirect sampling. The sampling weights can be determined using the generalized weight share method (GWSM) as detailed in [1] and [2]. However, this method requires the observation of all links for the indirectly sampled units. As a result, it cannot be applied when some of the links between the frame population and the sample in the target population are missing or difficult to retrieve exhaustively. This is especially true when one unit of the frame population is linked to only one unit of the target population, but one unit of the target population can be linked to many units of the frame population. This type of links is called "Many-to-One" and appears in particular situations such as the French postal traffic survey where the frame populationis made of addresses and the target population of the postman tours.

A solution to avoid this issue of missing links is to consider an intermediate population linked in some way to both the frame and target populations and use a double indirect sampling. Then the GWSM can be used twice, first between the frame and intermediate populations and then between the intermediate and target populations. If the links between each population are of type "Many-to-One", this method needs less information on links compared to the simple indirect sampling. Indeed the sample in the target population only requires the observation of all links to the intermediate population and only the sampled units in the intermediate population requires the observation of all links to the frame population. The links between the frame population and non (indirectly) sampled units of the intermediate population don't have to be observed, even for units linked to a sampled unit of the target population. The gain in the number of links to observe can be important which gives a definite advantage to the double GWSM. In the case of the French postal survey, the intermediate population is made of outgoing mail sorting boxes. While there is a need to observe 500 addresses per round in average for a simple GWSM, there are only 50 addresses per box and 10 boxes per round in average, which lowers to 60 the number of links to observe for a double GWSM.

However, as illustrated with the French postal traffic survey results, double indirect sampling with two GWSM may deteriorate the precision of estimators in some situations. Using mathematical derivations, it is possible to derive explicitly the magnitude of the loss of precision in situations similar to the French postal context. These derivations allow a better understanding of the loss of precision. Monte-Carlo simulations illustrate how significant the deterioration can be.

When the double GWSM method leads to an excessive loss of precision compared to the simple one, it can be of interest to use a double indirect sampling together with an estimator close to a simple GWSM. It can be noted that a simple GWSM estimator can be derived using a double indirect sampling, as long as the links between the frame population and the non sampled intermediate units related to the target sample are known. If these links are unobserved but auxiliary information is available, it can be used to fit a model and estimate the number of links needed to compute a simple GWSM estimator. A similar solution is considered in [3] for treatment of links nonresponse.

Fernanda Lang Schumacher University of Campinas

Canonical fundamental skew-t linear mixed models

In clinical trials, studies often present longitudinal data or clustered data, which are frequently affected by missing data. These studies are commonly analyzed using linear mixed models (LMMs), usually considering Gaussian assumptions for both random effect and error term. Recently, several proposals extended the restrictive assumptions from traditional LMM by more flexible ones, that can accommodate skewness and heavy-tails, and consequently are more robust to outliers. This work proposes a canonical fundamental skew-t (ST) LMM, that allows for asymmetric and heavy-tailed random effects and heavy-tailed errors, and includes several important cases as special cases, which are presented and considered for model selection. For this robust and flexible model, we present an efficient EM-type algorithm for parameters estimation via maximum likelihood, which is implemented in a closed form by exploring the hierarchical representation of the ST-LMM. In addition, the estimation of standard errors and random effects is discussed. The methodology is illustrated through an application to schizophrenia data and some simulation studies.

Luiz Eduardo da Silva Gomes Federal University of Rio de Janeiro

A Bayesian network approach to food security modeling in Brazil

This work proposes a probabilistic decision tool that integrates the main factors influencing food insecurity in Brazil. Food security exists when individuals have access to sufficient food to meet their dietary needs in quantity and quality. Evidence indicates that food insecurity is associated with malnutrition, diabetes, cardiovascular diseases, some cancers, deterioration of mental health, inability to manage chronic disease, worse children's academic performance, and social skills. Government policies on welfare, farming, the environment, employment, health, and others impact food security at various levels. Each of these is a dynamic sub-system that considers factors influencing food insecurity such as weather, economy, food availability, social and demographic aspects. In the context of policies for complex systems, it is difficult for decision-makers to account for all the variables within the system. The usual approach to relate factors and outcomes is based on linear regression models that do not allow for cause-effect inference. Our proposal is based on Bayesian networks that can capture both non-linearities and complex cause-effect relationships. This model also considers a time-varying Dirichlet process for smooth changes in the economic, demographic, and weather series affecting food security. Thus, the effectiveness and sustainability of interventionist policies such as the Fome Zero program {the larger initiative to combat hunger in Brazil, can be measured. Food security was measured using a national psychometric scale consisting of questions related to the direct experience of food insecurity available on the Brazilian National Household Sample Survey (PNAD). The outcome of this project is a decision support system that integrates the main factors influencing food insecurity in a probabilistic model. With this tool, the decision-makers will optimize the cost-effectiveness of future interventions, simulate scenarios, and compare several candidate policies.

Contributed Session 3

Caio Gonçalves National School of Statistical Sciences - ENCE/IBGE, **Luna Hidalgo** Brazilian Institute of Geography and Statistics, **Denise Britz do Nascimento Silva** National School of Statistical Sciences - ENCE/IBGE and **Jan van den Brakel** Statistics Netherlands and Maastricht University

Model-based unemployment rate estimates for the Brazilian Labour Force Survey

The Brazilian Labour Force Survey (BLFS) is a quarterly rotating panel survey with 80% sample overlap between two successive quarters. Monthly unemployment rate estimates are regularly produced based on a three-month average of direct estimates. Due to the unforeseen situation of COVID19 pandemic and its effects in the economy and labour market, there was a need to investigate model-based estimation procedures to obtain unemployment rate single-month estimates. We present structural time series models developed to produce model-based single month estimates at national level as well as small area (state-level) estimates at a higher frequency than those currently being published. Using the state-space framework, the models account for the autocorrelation due to sample overlap and the increased dynamics in the labour force series in 2020. In addition, bivariate models that combine claimant count and survey data as well as multivariate models integrating the analysis for different states are investigated. The models not only yield estimates with better precision than direct estimates, since the latter were affected by a rise in non-response, but they can deliver reliable state-level public statistics at a monthly frequency that are presently required. The new improved model-based estimates were proposed as experimental statistics for the Brazilian national statistical office (IBGE).

Jan van den Brakel Statistics Netherlands and Maastricht University, **Martijn Souren** and **Sabine Krieg** Statistics Netherlands

Estimating monthly labour force figures during the COVID-19 pandemic in the Netherlands

Official monthly statistics about the Dutch labour force are based on the Dutch Labour Force Survey (LFS). This survey is based on a rotating panel design. Responding households are interviewed five times at quarterly intervals. In 2010, Statistics Netherlands implemented a model-based estimation procedure based on multivariate structural time series model to produce monthly figures about the labour force. This time series model is used as a form of small area estimation to produce sufficiently reliable monthly estimates, despite the relative small monthly sample sizes. The model also accounts for systematic differences between the outcomes obtained in the subsequent waves (rotation group bias) and for systematic difference in the outcomes due to two major redesigns that took place in 2010 and 2012 using intervention variables.

Data collection of the Dutch LFS is based on a sequential mixed mode design that starts with Web Interviewing (WI). Non-respondents are followed up with computer assisted telephone interviewing (CATI) if a listed telephone number is available or by computer assisted personal interviewing (CAPI) if no listed telephone number is available.

Due to the COVID-19 pandemic, the Netherlands went in a lockdown on March 16, 2020. Due to this lockdown, CAPI stopped. It can be anticipated that this has a systematic effect on the outcomes of the LFS. At the same time, it can be expected that the lockdown effects the real monthly labour force figures. The unemployed labour force time series show a steady decrease since 2014, while the evolution of the employed labour force shows a steady increase during these last six years. The lockdown indeed marked a sharp turning point in the evolution of these series. Moreover, these sharp turning points induced by the lockdown strongly increased the dynamics in the labour force series, which was, according to standardized innovations, not sufficiently picked up by the time series model.

In this paper two problems related to the lockdown are addressed. The first problem is how to separate a sudden change in the mode effects because CAPI stopped, from real period-to-period changes in the labour force figures. Extending the time series model with an intervention variable, is not a good solution, since the lockdown also resulted in a turning point at exactly the same time. In that case it can be expected that a major part of the real period-to-period change is incorrectly absorbed in the discontinuity estimate. In this paper three different approaches to correct for the loss of CAPI responses are proposed and compared. The second problem is how to adapt the time series model to the increased dynamics in the labour force figures.

It is discussed how the different approaches to account for both problems are evaluated during the first month of the lock down in order to choose the most promising method for the publication of official monthly labour force figures. Consequences of these choices are discussed by comparing the different methods in an in real time analysis, up to the most recent available data.

Andres Gutierrez United Nations Economic Commission for Latin America and the Caribbean, **Hanwen Zhang** Universidad Autónoma de Chile, **Pedro Luis do Nascimento Silva** Escola Nacional de Ciências Estatísticas and **Leonardo Trujillo** Universidad Nacional de Colombia

Estimation of gross labour-force flows for some non-response models: a case study during the coronavirus outbreak

One of the several contributions of Chris Skinner to the statistical science was his work related to the analysis of complex surveys (Skinner, Holt & Smith 1989). From all of his valuable contributions, it is worth mentioning his work on estimating gross flows (Skinner & Torelli 1993, Pfeffermann, Skinner & Humphreys 1998). In this paper, we present and apply a pseudo-likelihood based procedure for the estimation of gross labour-force flows from complex survey data in the presence of nonrandom missing data. Most existing approaches rely on assuming independent and identically distributed observations or considering that non-response is random; however, these assumptions do not hold when dealing with complex survey data. Extending the ideas of Fienberg & Stasny (1983) to the complex sampling approach, we propose non-response models that yielded empirically design-consistent estimates of the gross flows parameters under different assumptions about the non-response mechanism. The parameters and their corresponding standard errors were estimated by using the pseudo-likelihood approach (Binder 1983).

These models were applied to the Chilean Labour Force Survey to estimate the gross labour-force flows between the first and second quarters of 2020, just in the middle of the coronavirus outbreak that directly affected the population's occupational status. Furthermore, in this period, the non-responses rates of the Survey severely increased. Each respondent was classified in one of four states: employed (formal and informal), unemployed and inactive. We chose the best fitting model by using Pearson's χ2 with Rao - Scott adjustment from these plausible models. The chosen model estimated the gross-flows between the first two quarters of 2020, providing valuable information to understand the impact of COVID-19 on Chile's labour-force situation. One of the paper's main findings is that 41% of people with informal jobs and 47% of unemployed people during the first quarter of 2020 changed their occupational status to inactive during the second quarter because they stopped searching for a job. These transitions reveal the tremendous impact of the ongoing pandemic on the labour market in Chile.

Guilherme Jacob and **Pedro Luis do Nascimento Silva** Escola Nacional de Ciências Estatísticas

Gross labour flows estimation with non-response using the Brazilian National Household Sample Surveys

In many countries, labour force surveys use a rotating panel strategy, which allows for the estimation of gross labour flows. The gross labour flows enable a more detailed analysis of the labour market dynamics. However, issues like non-response and misclassification errors might introduce bias in the estimates. Humphreys and Skinner (1997) and Pfeffemann, Skinner, and Humphreys (1998) developed methods to account for both misclassification and complex survey design in the estimation. Stasny (1987) developed methods to account for non-response bias under simple random sampling, with Gutiérrez (2014) and Gutiérrez, Trujillo, and Silva (2014) developing methods for complex sample surveys.

Stasny (1987) proposed models to separate the non-response mechanism from the transitions in the observed flows. The observed flows are seen as a result of a two-stage process: (1) a Markov Chain that allocates individuals in an unobservable gross flows matrix; and (2) a process in which individuals in each of the gross flows table cells lose their classification with probabilities that vary according to their status. Gutiérrez, Trujillo, and Silva (2014) show how the parameters of these models can be estimated using complex samples, also describing a method to calculate variance estimates of both gross flows and model parameter estimates.

While these methods are known to produce reliable results, their usage in official statistics production requires well documented, reliable, and accessible statistical software. This work presents a software that implements the methodology developed by Gutiérrez, Trujillo, and Silva (2014), producing estimates for both the gross flows, initial and transition probabilities as well as for the non-response mechanism. By using a syntax based on that of Lumley (2004), this software reduces the cost of learning a new syntax and allows for easy usage of the broad range of methods already implemented. Simulations are used to assess the properties of the estimators and their software implementation.

Using the Brazilian National Household Sample Survey, the method and software are applied to estimate quarter-to-quarter labour flows from 2013 to 2019. The results confirm that gross flows can reveal important facts about the dynamics of the labour markets, usually not revealed by the analysis of net flows. The results highlight the impact of the flows between employment, unemployment and non-participation in the labour force, enhancing the official labour statistics products. While this approach is useful for labour flows, it can also be applied to estimate gross flows for other variables.

Invited Session 4

Graham Kalton University of Maryland

Probability vs. nonprobability sampling: from the birth of survey sampling to the present day

At the beginning of the 20th century, there was an active debate about random selection of units versus purposive selection of groups of units for survey samples. Neyman's (1934) paper tilted the balance strongly towards varieties of probability sampling combined with design-based inference, and most national statistical offices have adopted this method for their major surveys. However, nonprobability sampling has remained in widespread use in many areas of application, and over time there have been challenges to the Neyman paradigm. In recent years, the balance has tilted towards greater use of nonprobability sampling for several reasons, including: the growing imperfections and costs in applying probability sample designs; the emergence of the internet and other sources for obtaining survey data from very large samples at low cost and at high speed; and the current ability to apply advanced methods for calibrating nonprobability samples to conform to external population controls. This paper presents an overview of the history of the use of probability and nonprobability sampling from the birth of survey sampling at the time of A. N. Kiær (1895) to the present day.

Wayne Fuller Iowa State University

Post strata based on sample quantiles

The standard method of creating post strata is to define the boundaries of the strata on the basis of population characteristics of auxiliary variables. Such samples often contain empty cells requiring adjustment to the estimation procedure. To avoid empty post strata, we propose using the sample distribution function of the auxiliary variable to define the post strata. We show that the large-sample efficiency of the sample-based-post-stratification procedure is the same as that of the equivalently-defined population-based procedure. In the simulation, the sample-based procedure was slightly more efficient than the classical procedure. The Monte Carlo coverage of a nominal 95 percent interval was approximately 95 percent for the sample-based procedure and approximately 94 percent for the classical procedure.

Panel Session (Chair: Nikos Tzavidis)

Dennis Trewin
An integrated statistical approach to managing pandemics

Panelists: **Len Cook**, **Pedro Luis do Nascimento Silva** Escola Nacional de Ciências Estatísticas, **David Steel** University of Wollongong

As can be seen from COVID-19 pandemic, good quality statistical information is crucial to managing pandemics and understanding the consequences on households and businesses. Many countries implemented a range of data collection activities although, in most countries, the involvement of national statistical offices was limited. Epidemiologists drove the data collection activity but, among other things, most ignored the possibilities from probability surveys and, as a consequence, there was a limited knowledge of asymptomatic and mildly symptomatic cases that were not tested. The focus of the session is on the design of a pandemic information plan for future pandemics which might include future waves of COVID-19.

The presentation describes the shortcomings of the statistical approach in Australia and then proposes a more integrated approach underpinned by a commonly used SEIR epidemiological model. It uses a mixture of epidemiological data (eg positive cases), other data extracted from administrative systems and data derived from probability surveys. This data is used to estimate the all-important reproduction number. However, the dispersion among different socio-economic groups is also Important to the management of the pandemic but this aspect is often ignored in data collection.

We also consider waste-water surveillance. This is potentially especially important for the early detection of the virus. Statisticians can assist with the design of efficient systems that use group testing. They can also assist with designs that address any lack of sensitivity and specificity in tests.

We describe the range of other ways in which statisticians can assist. These include modelling to estimate reproduction and dispersion numbers with challenging data sets, the design of an information model to enable extraction of data from test and trace data, design of impact surveys, and assessing the validity of big data sets (eg mobility data) to inform pandemic management.

Contributed Session 4

Mihaela-Catalina Anastasiade-Guinand Swiss Federal Statistical Office, **Alina Matei** and **Yves Tillé** Université de Neuchatel
Estimating a counterfactual wage heavy-tailed distribution using survey data

We work in the framework of the gender wage modelisation using survey data. The wage of an employee is hypothetically a reflection of their characteristics, such as the education level or the work experience. It is possible that a man and a woman with the same characteristics get different salaries. To measure the difference in the gender wages we use the concept of counterfactual distribution. This is done in order to estimate what the former group would earn, if they had the characteristics of the latter group. The usual regression approach of Blinder-Oaxaca consists of modeling the mean of the wage of each individual conditionally on their characteristics. The aim is to isolate the part attributable to gender at the mean level by estimating the part of the wage difference that is explained by the differing characteristics.

Conditional to some characteristics, we assume that the conditional wage distribution of each woman follows a given theoretical distribution with unknown parameters. First, we estimate the parameters of the distribution of each woman given their characteristics. Next, the marginal women wages distribution is fitted based on the individual woman wage distributions. A counterfactual distribution is constructed by reweighting the women characteristics. We provide two parametric methods to estimate the gender wage quantiles and counterfactual wage quantiles, respectively, and estimate their differences. The goal is to capture the shape of the wage distributions and to go beyond the simple mean differences, by determining the estimator of 'gender wage discrimination' at different quantiles.

Since, in general, wage distributions are heavy-tailed, the main interest is to model wages by using heavy-tailed distributions like the GB2 distribution. We illustrate the two proposed methods using the GB2 distribution and real data from the Swiss Federal Statistical Office.

Nora Würz Freie Universität Berlin, **Timo Schmid** Universität Bamberg and **Nikos Tzavidis** University of Southampton

Estimating regional income indicators with access to limited auxiliary information

Reliable knowledge of the spatial distribution of income and wealth is essential for evidence-based policymaking. High spatial resolution direct estimates of income that use household surveys are likely to be unreliable because of the small sample sizes at the spatial scale of interest. A possible way to overcome this problem is by using small area estimation methods (Rao and Molina, 2015; Tzavidis et al., 2018) and in particular model-based methods under the nested error regression (unit-level) model (Battese et al., 1988). The estimation of income indicators, including non-linear ones, for example the poverty gap and poverty severity, has been researched extensively (Molina and Rao, 2010; Tzavidis et al., 2018; Molina and Martín, 2018). These indicators are typically estimated by using model-based methods that assume access to auxiliary information from population micro-data. In countries like Germany and the UK population micro-data are not publicly available and access to such data is even challenging within gatekeeper organizations. Instead, population-level auxiliary data is often only available at some aggregate level.

In this work we propose small area methodology for estimating small area means based on the transformed nested error regression model when only aggregate population-level auxiliary information is available. Especially skewed variables, like income and consumption, can often not be adequately described by the available auxiliary variables and therefore lead to error terms that are not normally distributed. Thus, the models we consider in this work use fixed logarithmic (Molina and Martín, 2018) or data-driven (Sugasawa and Kubokawa, 2019; Rojas-Perilla et al., 2020) transformations for the dependent variable. In the absence of population micro-data, appropriate bias-corrections for small area prediction are presented. Under the proposed approach we do not make any parametric assumptions about the auxiliary variables and instead use aggregate statistics (means and covariances) and kernel density estimation to resolve the issue of not having access to population micro-data. Regarding the estimation of the mean squared error, we propose a parametric bootstrap that captures the uncertainty due to the use of transformations and kernel density estimation. For estimating means, Li et al. (2019) propose an alternative approach that uses a smearing estimator assuming access only to aggregate covariates. This is one of the alternative approaches we compare our method against.

Extensive model-based and design-based simulations are used to compare the proposed method to alternative methods for example, the EBP under transformations (assuming the availability of unit-level Census data) and the estimator of Li et al. (2019). Results show that, compared to alternative methods, the proposed methodology leads to comparable results. The proposed methodology is also applied to the 2011 Socio-Economic Panel and aggregate census information from the same year to estimate the average individual income for 96 regional planning regions in Germany.

Antônio Teixeira Senac – Departamento Nacional and **Pedro Luis do Nascimento Silva** Escola Nacional de Ciências Estatísticas

Cross-classified sampling and regression estimation for price index estimation

Cross-classified or two-dimensional sampling (Ohlsson, 1996) was one of the topics of interest to Chris Skinner in his long and prominent research career. His 2015 paper on "Cross-classified sampling: Some estimation theory" was a key reference to the first author's PhD thesis – Teixeira Júnior (2020). This article presented expressions for estimating population totals under cross-classified sampling designs, namely designs that sample independently in the two dimensions of a population to select the units to be observed / measured. Expressions considering simple random, stratified, and PPS with replacement for sampling in each of the dimensions were provided.

Teixeira Júnior (2020) extended these expressions for estimating population totals under cross-classified sampling using regression estimators, and also considered an application that Skinner (2015) used as an example, namely the selection of trading places and products for a price index survey.

This article compares the efficiency of regression and Horvitz-Thompson (HT) estimators of a price index under cross-classified sampling, considering both simple random and stratified simple random sampling in the two dimensions. Design effects are used to measure the relative efficiency. Two-dimensional sampling is also compared with two-stage cluster sampling taking the trading places as clusters in the first stage of sampling.

The initial results show that the coefficients of variation (CVs) obtained with the regression estimator are smaller than the CVs obtained with the HT estimator. The efficiency gains achieved by regression estimation are slightly bigger for cross-classified sampling than for cluster sampling.

Elizabeth Belo Hypolito and **Denise Britz do Nascimento Silva** National School of Statistical Sciences - ENCE/IBGE

Impacts of proxy response in the Brazilian labor force survey

Data quality, under the approach of the total survey error, aims to minimize the accumulated error in the estimates. Studies on the non-sampling components of the total survey error, even when partial, are important for the evaluation and improvement of the survey process, as well as for the efficient allocation of resources. One of the possible sources of non-sampling errors in household surveys is the use of proxy respondents to get information about absent or disabled persons who cannot provide information about themselves. In general, labor force surveys use this resource widely once they are collected and disseminated in a short period of time. Although it is an important resource for reducing nonresponse, the use of proxy response raises some questions about the reliability of the data collected. It is reasonable that, even the proxy with extensive knowledge about the selected person, does not know all the information requested by the survey. Thereby, they can contribute to an increase in item nonresponse and measurement error and, eventually, in the total survey error. This study aims to investigate possible impacts of proxy response in the National Continuous Household Sample Survey (Continuous PNAD), the main source of labor force data in Brazil, carried out by the Brazilian Institute of Geography and Statistics (IBGE). The results show that, in 2017, about 55% of the survey questionnaires were answered by proxy respondents, mainly household reference persons or spouses. Persons who had their information answered by proxy were primarily children or men in any position in the household. The imputation rate for income data was considerably low in the survey (less than 2%). However, this rate was higher for questionnaires answered by proxies than for those answered by self-respondents. The divergence rate for some basic characteristics that should have remained unchanged for the same person in two consecutive waves (second and third quarters of 2017) were noticeably high, especially in the presence of proxy response or replacement of interviewers. In order to better understand the effect of respondents and interviewers in this process, logistic regression were employed to model the probability of divergence for selected variables. The results point out that the replacement of respondents (proxy with self-respondent, selfrespondent with proxy or proxy with another proxy respondent) or interviewers between waves led to higher probabilities of divergence. The work provides evidence that the use of proxy respondent increases the item nonresponse and the measurement error in the survey. In addition, they draw attention to the need for further studies related to nonresponse errors in the Brazilian household surveys.

Contributed Session 5

Phillip S. Kott RTI International

The role of weights in regression modeling and Imputation

When fitting observations from a complex survey, the standard regression model assumes that the expected value of the difference between the dependent variable and its model-based prediction is zero no matter what the values of the explanatory variables. A rarely-failing extended regression model assumes only that the model error is uncorrelated with the model's explanatory variables. When the standard model holds, it is possible to create alternative analysis weights that retain the consistency of the model-parameter estimates while increasing their efficiency by scaling the inverse-probability weights by an appropriately chosen function of the explanatory variables.

When a regression model is used to impute for missing item values in a complex survey, and item missingness is a function of the explanatory variables of the regression model and not the item value itself, near unbiasedness of an estimated item mean requires that either the standard regression model for the item in the population holds or the analysis weights incorporate a correctly-specified and consistently estimated probability of item response. By estimating the parameters of the probability of item response with a calibration equation, one can sometimes account for item missingness that is (partially) a function of the item value itself.

Ton de Waal Statistics Netherlands

Calibrated imputation for multivariate numerical data

Non-response is a major problem for anyone collecting and processing data, such as national statistical institutes. When left untreated, non-response can lead to biased estimates or results from statistical analyses. Non-response can be subdivided into item non-response, where some values from otherwise observed units are missing, and unit non-response, where entire units are not observed. A commonly used technique to deal with missing data, especially due to item-nonresponse, is imputation. In imputation, missing values are estimated and filled in into the dataset.

Imputation, just like any other technique for dealing with missing data, can become challenging when data from several sources are combined. When all data sources would be available at the same time, imputation can simply be carried out by first linking the data sources as well as possible and then treating the linked dataset as a single data source. Imputation itself is then basically the same as imputation for a single data source with missing data. Unfortunately, data sources often are not all available at the same moment as waiting for the last data source to become available may not be a viable option due to the time delay. The data are then often used for multiple purposes: very timely but less detailed results are published when the first data sources become available, less timely but more detailed results are published later when other data sources have become available as well. It is here where the use of imputation becomes especially challenging: unless special precautionary measures are taken, the later results will deviate from the earlier published results. This is deemed undesirable by many national statistical institutes.

Especially at national statistical institutes, the imputation problem is often complicated further owing to the existence of constraints in the form of edit restrictions that have to be satisfied by the data.

Examples of such edit restrictions are that young children cannot have an academic degree and that the turnover of an enterprise should be non-negative.

In our paper, we develop imputation methods for multivariate data such that previously published estimated population totals are preserved while edit restrictions on the data are satisfied. To this end, we adopt a fully conditional specification imputation approach and extend this with a weighted calibration step that takes the previously published estimated population totals into account and a step that ensures that ensures that edit restrictions are satisfied.

Mehdi Dagdoug and **Camelia Goga** Université de Bourgogne-Franche-Comté and **David Haziza** University of Ottawa

Model-assisted estimation through random forests in finite population sampling

Estimation of finite population totals is of primary interest in survey sampling. Often, additional auxiliary information is available at the population level. Model-assisted estimators use this supplementary source of information to construct estimators built upon predictors. In this work, new classes of model-assisted estimators based on random forests are suggested.

Generally speaking, random forest is an ensemble method which consists of creating a large number of regression trees and combining them to produce more accurate predictions than a single regression tree would.

Under mild conditions, the proposed estimators are shown to be asymptotically design unbiased and consistent. Their asymptotic variance is derived, and a consistent variance estimator is suggested. The asymptotic distribution of the estimators is obtained, allowing for the use of normal-based confidence intervals.

Simulations illustrate that the proposed estimator is particularly efficient and can outperform state of-the-art estimators, especially in complex settings such as small sample sizes, high-dimensional regressor space or complex superpopulation models.

Savano Pereira, Luiz P. Calôba and **Pedro Luis do Nascimento Silva** Escola Nacional de Ciências Estatísticas

Artificial neural networks with complex survey data

Artificial Neural Network (ANN) models are often fitted to sample data as if they were obtained by simple random sampling with replacement, i.e. as if the sample observations are indeendent and identically distributed. However, in practice, sample surveys rarely provide data where such assumptions are tenable because they use complex sampling schemes, involving stratification, clustering and unequal probabilities of selection. Such sampling strategies are required to accommodate existing population structure and/or survey administration requirements. The statistical literature provides several approaches for modelling data from complex surveys. However, the ANN literature does not provide approaches for when the data comes from complex sample surveys. We developed a superpopulation approach for modelling data from complex surveys using ANN. This approach can incorporate typical aspects of complex sampling such as stratification, clustering and unequal weighting, both when estimating parameters of the ANN as well as their corresponding precision. An empirical evaluation of the proposed approach was carried out through simulation, in a scenario where stratification and unequal weighting lead to informative sampling. The results suggest that our approach can reduce estimation bias for some of the ANN parameters, although not for all. Further work is ongoing to improve understanding of the empirical performance of our approach.

Invited Session 5 (Chair: Paul Smith)

Denise Britz do Nascimento Silva Escola Nacional de Ciências Estatísticas, **Eduardo Santiago Rosseti** FGV Projetos and **Antônio Etevaldo Teixeira Júnior** Senac - Departamento Nacional

Compositional time series analysis of labour force status in the Brazilian National Household Survey

Most surveys are multivariate and multipurpose, often containing several variables with multinomial response. Frequently, the interest lies in survey estimates of the proportion of units classified in each response category. In this case, the survey estimates form a composition, proportions of a whole subject to a unity sum constraint. The vector of proportions of people classified by labour force status (employed, unemployed and not in the labour force) is a well-known example of compositional survey data. As many surveys are repeated on several occasions, compositional time series from repeated surveys are also produced.

The problems of modelling and analysing compositional data are discussed comprehensively in Aitchison (1986) who demonstrated the difficulties of applying standard statistical methods due to the sum constraint. He considered logratio transformations to allow the use of standard unconstrained multivariate statistics applied to transformed data. Brunsdon (1987), supervised by Fred Smith, and Brunsdon and Smith (1998) developed models for time series analysis of compositional data from repeated surveys with direct application to UK monthly public opinion polls of voting intentions. In addition, Silva (1996), also supervised by Fred Smith, and Silva and Smith (2001) proposed state-space models for compositional time series from rotating panel surveys considering the autocovariance structure of the sampling errors. Later, Van den Brakel and Roels (2010) used structural time series models with intervention variables to estimate the effect of a survey redesign on compositional time series from repeated surveys.

The Brazilian National Household Survey (BNHS), carried out by the Brazilian Institute of Geography and Statistics (IBGE), is a quarterly rotating panel survey with 80% sample overlap between two successive quarters and is the main data source for the country's labour market indicators since 2012. The paper presents state-space models to produce model-based estimates of labour force status composition, the unemployment rate series, and corresponding unobserved structural components, taking into account the autocorrelation of sampling errors. In addition, estimates of labour force status flows are used to highlight changes in composition due to economic and political events as well as the unforeseen pandemic situation.

Pedro Luis do Nascimento Silva Escola Nacional de Ciências Estatísticas and **Fernando Moura** Universidade Federal do Rio de Janeiro

Fitting multivariate multilevel models under informative sampling

A model-dependent approach for multivariate multilevel normal modelling that accounts for informative sampling of group and unit level population elements is developed. The approach involves extracting the multilevel model holding for the sample data given the selected sample as a function of the corresponding population model and the sample selection probabilities, and then fitting the resulting sample model using Bayesian methods. An illustration is provided using the approach for modelling jointly Portuguese Language and Mathematics proficiency scores obtained from a Brazilian evaluation study of basic education conducted by the Brazilian National Institute of Education Research (INEP). The scores stem from applying Item Response Theory models to test results from the 'Prova Brasil 2009' study. A two-level multivariate hierarchical normal model is fitted, where the students are the first level units and schools are the groups (second level units). The analysis is restricted to the students from the year in elementary schools from the municipality of Rio de Janeiro. Joint modelling of proficiency scores is the added value of our approach, since typical multilevel modelling of such scores is carried out separately using univariate models.