BALTIC-NORDIC-UKRAINIAN NETWORK ON SURVEY STATISTICS

PROCEEDINGS

2023

Proceedings of the 6th Baltic-Nordic Conference on Survey Statistics. – Helsinki, Finland, 2023. – ISBN 978-951-51-9425-1

Programme committee

Maciej Beręewicz, Poland
Andrius Čiginas, Lithuania
Tetiana Ianevych, Ukraine
Danutė Krapavickaitė, Lithuania
Krista Lagus, Finland, Vice chair
Thomas Laitila, Sweden
Risto Lehtonen, Finland
Mārtiņš Liberts, Latvia
Vilma Nekrašaitė-Liegė, Lithuania
Kaja Sõstra, Estonia
Imbi Traat, Estonia
Maria Valaste, Finland, Chair
Olga Vasylyk, Ukraine
Mare Vähi, Estonia
Baiba Zukula, Latvia
Chiara Bocci (University of Florence), Italy
Henri Luomaranta (Statistics Finland), Finland

Organizing committee

Maria Valaste (University of Helsinki), Chair
Kimmo Vehkalahti (University of Helsinki)
Krista Lagus (University of Helsinki)
Maria Litova (University of Helsinki)
Lashini Liyanage (University of Helsinki)
Emilia Carson (University of Helsinki)
Yu Ren (University of Helsinki)
Anastasiia Volkova (University of Helsinki)
Adeline Clarke (University of Helsinki)
Pauli Ollila (Statistics Finland)
Tetiana Ianevych (Ukraine)
Olga Vasylyk (Ukraine)
Centre of Social Data Science staff

Organizing institutions

The Baltic-Nordic-Ukrainian Network on Survey Statistics, in particular:

University of Helsinki, Finland
Statistics Finland
University of Tartu, Estonia
University of Latvia, Latvia
Vilnius University, Lithuania
Vilnius Gediminas Technical University, Lithuania
Poznan University of Economics and Business, Poland
Stockholm University, Sweden
Örebro University, Sweden
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”
Taras Shevchenko National University of Kyiv, Ukraine

https://wiki.helsinki.fi/display/BNU/Home

Compiler: Olga Vasylyk

ISBN 978-951-51-9425-1

SPONSORS

International Association of Survey Statisticians (IASS)

University of Helsinki

Foreword

This proceedings publication includes the abstracts of papers for the 6th Baltic-Nordic Conference on Survey Statistics, BaNoCoSS 2023, taking place on 21–25 August 2023 in Helsinki, Finland and online. Previous conferences were organized in 2002 in Ammarnäs (Sweden), 2007 in Kuusamo (Finland), 2011 in Norrfällsviken (Sweden), 2015 in Helsinki (Finland) and 2019 in Örebro (Sweden).

BaNoCoSS 2023 is a scientific conference presenting new developments in survey statistics theory and methodology in a broad sense. Among other areas, this includes innovations in design-based model-assisted and model-based inferences, small area estimation, new data sources, probability sampling, and integrated probability and nonprobability data. A half-day R workshop on a specific topic complements nicely the programme, presented by Philipp Christian Broniecki of University of Oslo.

We have world-class leading experts in statistical science as keynote speakers. Speakers coming to Helsinki include Jae Kwang Kim of Iowa State University, Jan A. van den Brakel of Statistics Netherlands and Maastricht University, Camelia Goga of University of Franche-Comté and Li-Chun Zhang of University of Southampton; Statistics Norway; University of Oslo. We are pleased to have Andrew Gelman from Columbia University as one of the keynote speakers. He participates online. The program features 16 invited speakers on several interesting topics and more than 30 great presentations from 10 different countries.

The conference provides a platform for discussion and exchange of ideas for a variety of people including statisticians, researchers and other experts from universities and national statistical institutes and other governmental bodies as well as people working at private enterprises. University students in statistics and related disciplines provide an important interest group of the conference. In order to support researchers, university teachers and students in Ukraine and give them an opportunity to take part in the BaNoCoSS 2023, the conference is organized in hybrid mode. In addition to on-site participants in Helsinki, people attending online constitutes a large audience.

The conference is organized by the Baltic–Nordic–Ukrainian Network on Survey Statistics in cooperation with University of Helsinki and Statistics Finland. The scientific programme was developed by the programme committee, and the practical arrangements are made by the organizing committee consisting of the personnel of the Centre of Social Data Science. The proceedings were prepared at National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”. The pdf file is freely available at https://wiki.helsinki.fi/display/BNU/BANOCOSS2023.

We are thankful for the sponsorship of the International Association of Survey Statisticians (IASS), a section of the International Statistical Institute (ISI), the Nordic Council of Ministers, Statistics Finland, SAS Institute and University of Helsinki.

We wish everybody an inspiring conference and enjoyable stay in Helsinki.

On behalf of the organizers,
Risto Lehtonen and Maria Valaste

CONTENTS

1 KEYNOTE PAPERS

New Data Sources and Inference Methods for Official Statistics

Jan van den Brakel[1,2]

[1] Maastricht University, The Netherlands

[2] Statistics Netherlands, The Netherlands
e-mail: jbrl@cbs.nl

Abstract

Official statistics, published by national statistical institutes (NSIs) are traditionally based on repeated probability sampling in combination with design-based inference methods. This is a widely applied approach by NSIs because of its low level of risk. Under this approach NSIs are indeed in control over the availability of the data and the inference methods do not depend on statistical model assumptions of which the validity is often difficult to verify. There is nevertheless a growing interest among NSI's to use registers or other large data sets that are generated as a by-product of processes not directly related to statistical production purposes. The purpose of this, is to reduce data collection costs and response burden, to improve timeliness or to refine the level of detail of official statistics.

Roughly spoken, there are two ways to use these so called non-probability data in the production of official statistics. One approach is to use them as a primary data source for the compilations of official statistics. This generally requires a high risk apatite for the NSI, since there is no control over the availability of the data. On top of that, model-based inference methods are required to correct for selectivity. Most of these methods are based on strong missing ad random assumptions.

A second approach, which requires the acceptance of an intermediate level of risk, is to combine non-probability data sources with sample data in a model-based inference approach. Although inference methods are model-based, the NSI is still in control over the availability of the survey data that are the primary data source under this approach. Since official statistics are predominantly based on repeated surveys, time series methods provide a natural framework.

In this paper it will be illustrated how multivariate state space models and multivariate multilevel time series models can be used as a form of small area estimation by modelling temporal and cross-sectional correlations between previous reference periods and other domains. Extensions to models that include related auxiliary series to further improve the precision of the predictions will be discussed. It will be shown how dynamic factors models can be used in this context to avoid high-dimensionality problems in the case of a large amount of auxiliary series. This easily occurs if data sources like google trends are considered as auxiliary series. A major limitation of standard linear state space models is they assumption that correlations between state variables are constant over time. To relax this assumption a novel non-linear state space model will be proposed. To this end, time varying state correlations are modelled with separate stochastic processes. It will be illustrated how these methods can be used to refine the level of detail and timeliness of official statistics with real life examples at Statistics Netherlands.

Keywords: Small area estimation, multivariate state space models, time varying state correlations, dynamic factor models, multilevel time series models.

Quantifying Discontinuities in Time Series obtained with Repeated Surveys

Jan van den Brakel[1,2]

[1] Maastricht University, The Netherlands

[2] Statistics Netherlands, The Netherlands
e-mail: jbrl@cbs.nl

Abstract

Official statistics are often published repeatedly with the purpose of building consistent time series that describe the evolution of finite population parameters. A significant quality aspect of these surveys is the comparability over time of their estimates. This is a major reason to keep the underlying process of the survey unchanged as long as possible. It is inevitable, however, that adjustment or redesign of the process is needed from time to time, as and when the existing procedures become gradually outdated or more cost-effective methods are required. Recently most European national statistical institutes (NSIs) had to implement significant changes in the survey design of their Labor Force Survey (LFS) to meet the new Eurostat regulation for integrated European social statistics. Partly as a result of the COVID-19 pandemic but also from a cost perspective, many NSIs currently consider to move from uni-mode to mixed-mode data collection strategies.

Implementing such changes in a survey process generally affects measurement and selection bias in the responses of the survey, resulting in a systematic shock in the sample estimates. These shocks or discontinuities disturb comparability with figures published in the past. An important aspect of a survey redesign is to quantify the discontinuities in the main outcomes of the survey. In this way it can be avoided that discontinuities are incorrectly interpreted as real period-to-period changes of the population parameters of interest.

Collecting data under the old and new design in parallel for some period of time, time series modelling or a combination of both are established methods to quantify discontinuities. In this paper different approaches will be discussed and illustrated with real life examples at Statistics Netherlands. One example is a multilevel time series model used to quantify discontinuities due to three different survey redesigns in the Dutch Mobility Survey. Time series are modelled for a breakdown of the population parameter in about 700 domains. Predictions at higher aggregation levels are obtained by aggregation of the predictions of these 700 domains. This result in a numerically consistent set of estimates for all target variables, which are corrected for the different discontinuities. In another application it is shown how discontinuities due to a redesign of the Dutch Crime Victimization Survey are estimated on low regional level, using a small parallel run. With a cross-sectional multivariate Fay-Herriot model prediction for discontinuities at the most detailed regional level are obtained. Numerically consistent predictions for discontinuities at higher output levels are obtained by aggregation. In a third example discontinuities are estimated due to the implementation of the Eurostat 2021 regulations in the Dutch LFS. It is shown how a smooth transition in a rotating panel design is accomplished by integrating data from a parallel run with time series data in a multivariate state space model.

Keywords: multilevel time series models, multivariate Fay-Herriot models, survey redesigns, state space models.

Combining modeling with survey weights

A. Gelman[1]

[1] Department of Statistics, Columbia University, USA
e-mail: gelman@stat.columbia.edu

Abstract

Statistical approaches such as multilevel modeling and poststratification (MRP) can be used for small-area estimation, extrapolation to unsampled groups, and adjustment for differences between sample and population. But challenges arise when applying these methods to real-world surveys that include weights. Weighted regression or weighted likelihood approaches can be statistically inefficient as well as being awkward to incorporate into a model-based framework. Conversely, model-based estimates cannot in general be expressed as weighted averages. We are working on an integrated approach that includes weights as a latent poststratification variable. In this talk we will show the success of this approach or discuss why it does not work as planned, or perhaps both!

MODEL-ASSISTED ESTIMATION IN A HIGH-DIMENSION SETTING FOR SURVEY DATA

C. Goga[1]

[1] University of Franche-Comté, France
e-mail: mailto:camelia.goga@univ-fcomte.fr

Abstract

In sample surveys, model-assisted estimators are commonly used to obtain efficient estimators for interest parameters such as totals or means. Nowadays, it is no longer rare to be confronted with a very large number of auxiliary variables and model- assisted estimators can be less efficient. In this talk, I will discuss the asymptotic efficiency of model-assisted estimators in the presence of a very large number of auxiliary variables and show that they can suffer from additional variability under certain conditions. I will also present two techniques to improve the efficiency of the model-assisted estimator in a high-dimensional context: the first is based on dimension reduction and the second one on ridge-type penalization. The methodology is illustrated on real electricity consumption data for Irish households and companies.

RANDOM-FOREST MODEL-ASSISTED ESTIMATION

C. Goga[1]

[1] University of Franche-Comté, France
e-mail: mailto:camelia.goga@univ-fcomte.fr

Abstract

Abstract: Nowadays, surveys face more and more complex data sets with a large number of variables. These new data sets raise many challenges and traditional parametric methods for estimating interest parameters such as totals, ratios or quantiles may prove inefficient. In this work, we propose a new class of model-assisted estimators based on random forests. Under certain regularity conditions on the study variable, the random forest as well as the sampling design, the proposed model-assisted estimator is shown to be asymptotically design unbiased and consistent for the population total. A consistent variance estimator is proposed and the asymptotic distribution of the random-forest model-assisted estimator is obtained allowing to build confidence intervals. A new variance-estimator based on cross-validation technique is suggested. Simulation illustrate that the proposed estimator is efficient and can outperform state-of-the-art estimators, especially in complex and high-dimension settings.

Semiparametric adaptive estimation under informative sampling

Kosuke Morikawa[1] and J.K. Kim[2] and Yoshikazu Terada[3]

[1] Osaka University, Japan
e-mail: k.morikawa.es@osaka-u.ac.jp

[2] Iowa State University, U.S.A.
e-mail: jkim@iastate.edu

[3] Osaka University, Japan
e-mail: yoshikazu.terada.es@osaka-u.ac.jp

Abstract

In probability sampling, sampling weights are often used to remove the selection bias in the sample. The Horvitz-Thompson estimator is well-known to be consistent and asymptotically normally distributed; however, it is not necessarily efficient. This study derives the semiparametric efficiency bound for various target parameters by considering the survey weights as random variables and consequently proposes two semiparametric estimators with working models on the survey weights. One estimator assumes a reasonable parametric working model, but the other estimator requires no specific working models by using the debiased/double machine learning method. The proposed estimators are consistent, asymptotically normal, and can be efficient in a class of regular and asymptotically linear estimators. A limited simulation study is conducted to investigate the finite sample performance of the proposed method. The proposed method is applied to the 1999 Canadian Workplace and Employee Survey data.

Keywords: Adaptive estimation; Double/debiased machine learning; Semiparametric efficiency

Multiple bias calibration for valid statistical inference under nonignorable nonresponse

Z. Wang[1] and J.K. Kim[2] and S. Yang[3]

[1] Xiamen University, China
e-mail: wangzl@xmu.edu.cn

[2] Iowa State University, U.S.A.
e-mail: jkim@iastate.edu

[3] North Carolina State University, U.S.A.
e-mail: syang24@ncsu.edu

Abstract

Valid statistical inference is challenging when the sample is subject to unknown selection bias. Data integration can be used to correct for selection bias when we have a probability sample from the same population with some common measurements. How to model and estimate the selection probability of a non-probability sample using an independent probability sample is the challenging part of the data integration. We approach this difficult problem by employing multiple candidate models for the propensity score (PS) function combined with empirical likelihood. By incorporating multiple propensity score models into the internal bias calibration constraint in the empirical likelihood setup, the selection bias can be safely eliminated so long as the multiple candidate models contain the true PS model. The bias calibration constraint for the multiple PS models in the empirical likelihood is called the multiple bias calibration. The multiple PS models can include both missing-at-random and missing-not-at-random models. Asymptotic properties are discussed and some limited simulation studies are presented to compare with the existing methods. The proposed method is applied to a real-data-based simulation platform using the Culture & Community in a Time. of Crisis (CCTC) dataset.

Keywords: Empirical likelihood; Propensity score; Selection bias.

Design-based ensemble learning for individual prediction

L.-C. Zhang[1] and D. Lee[2]

[1] University of Southampton, UK and Statistics Norway, Norway
e-mail: L.Zhang@soton.ac.uk

[2] University of Alabama, US
e-mail: dlee84@cba.ua.edu

Abstract

Valid inference of the unobserved individual prediction errors is a fundamental issue to supervised machine learning, no matter how confident one is about the obtained predictor. An independent and identically distributed model of the prediction errors is commonly assumed for algorithm-based learning, such as random forest, support vector machine or neural network, which could be misleading in situations where the available observations are not obtained in a completely random fashion.

Survey sampling has a long tradition for estimating various aggregates of a given population. The inference of the associated uncertainty is based on the known samling design by which the sample of observations are obtained, "irrespectively of the unknown properties of the target population studied" (Neyman, 1934). But there has never been a design-based theory for prediction at the individual level.

We shall define and develop for the first time a general design-based approach to *individual prediction*, given the sampling design and the sample-splitting design for cross-validation, while the outcomes and features are treated as constants associated with the given population. Whether the predictor for the out-of-sample units is selected from an ensemble of models or a weighted average of them, the proposed approach can provide valid inference of the associated risk with respect to the known sampling design.

Keywords: Probability sampling; Ensemble learning; Rao-Blackwellisation.

References

Neyman, J. (1934) On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558-606.

2 INVITED PAPERS

PROCESSING OF COURT DATA

T. Ianevych[1], V. Golomoziy [2] and Yu. Mishura[3]

[1,2,3] Taras Shevchenko National University of Kyiv, Ukraine
e-mails: tetianayanevych@knu.ua, vitaliy.golomoziy@knu.ua, yumishura1@gmail.com

Abstract

Every person intends to protect his/her rights in case of their violation. Even if there is not a violation at the moment, everyone wants to be sure in the availability of such an opportunity, its reliability and efficiency. This has to be provided by the state system of justice. An effective mechanism of equal access to justice for all is the goal that modern states and open societies around the world seek to achieve. At the same time, the effective functioning of the state justice system is a factor that directly affects on its competitiveness and the successful economic development of the state and society.

The transparency of the justice system and the openness of information about the progress of the case and the execution of the court decision are the foundation of public trust, which today is extremely necessary to renew and strengthen. The Ukrainian system of open court decisions and executive proceedings does not ensure real transparency and openness of information about the administration of justice, and, therefore, it is not able to strengthen trust of the system in society. The efficiency indicators of the courts, such as relation of the amount of money has to be collected to those that has been collected, are approximately 0.1%, while the courts are overburdened and almost unable to effectively settle and prevent disputes.

Our goal is to develop the monitoring and data collection system based on the indicators, which will allow the fast and flexible detection of changes. In particular, it is proposed to create a convenient database using the Unified State Register of Court Decisions (https://reyestr.court.gov.ua/) and other court statistics data, for further statistical analysis and producing of recommendations. With the help of a statistical multifactorial analysis of the obtained database on the implementation of civil proceedings and other machine learning algorithms, it is planned to single out the main factors that affect the effectiveness of consideration of private legal cases by the court.

At that moment the register of court decisions includes more than 108 million documents. They are actually text files that have to be transformed into the statistical database. And it is a real challenge that can't be solved without using machine learning algorithms. Transformation of the nonstructural text data into the statistical dataset will allow using it for analysis by scientist, journalists, state officers, politicians and anyone wishing to do it and make the judicial system really transparent.

Keywords: data processing of judicial proceedings, machine learning, effectiveness of justice.

DESIGN-BASED AND MODEL-BASED INFERENCE IN FINNISH FOREST INVENTORY

A. Kangas

Natural Resources Institute Finland, Luke
e-mail: annika.kangas@luke.fi

Abstract

National Forest Inventory (NFI) of Finland was launched in 1921, as a second country after Norway. Currently the NFI is continuous work, with already a 13th round going on. The first NFI was carried out as a line inventory (lines going through the country from south-west to north-east), and from 1960's as a (stratified) cluster sampling.

Due to practical considerations, Finnish NFI, as most NFIs in the world, is based on a systematic sample. This means that only approximate variance estimators are available: the choice is between ignoring the systematic aspect and using design-based estimators and introducing a model and using model-based estimators. In Finnish NFI, mild assumptions on positive autocorrelation and local difference -based estimators have been used for all NFIs, while many other countries apply (conservative) SRS estimators.

During the history of NFI, the need for information to smaller and smaller areas has been increasing. Nowadays, the results are calculated for three regions and 19 counties using purely field plot data. In addition, the results are calculated for 309 municipalities, with a highly variable area. In part of the municipalities, it is possible to calculate useful results with design-based post-stratification with remote sensing material as auxiliary data, but in some cases, there are too few field plots for that. Therefore, we calculate strictly model-based synthetic estimators for all municipalities. However, there is a need for data for even smaller scales, such as single forest stands, for which synthetic estimators are the only possibility. For that purpose, Forestry Centre in Finland carries out a local inventory with a denser plot grid than in NFI. A further complication is that inferences are also needed in pan-European and global scales. Global (model-based) map products are used for policy making in climate change mitigation and biodiversity loss mitigation, to name a few examples. In these maps, the field data used is typically of poor quality and poorly representative leading to high biases.

All in all, complications due to systematic sampling and the multiple scales where information is needed prevent efficiently choosing either design-based or model-based approach, but forces to combine these approaches in many (ad-hoc) ways. Estimators utilizing the best properties of each approach thorough hybrid estimators would be in high demand in forestry.

Keywords: systematic sampling, cluster sampling, hybrid inference.

References

Kangas, A., Räty, M., Korhonen, K.T., Vauhkonen, J. and Packalen, T. (2019) Catering information needs from global to local scales - potential and challenges with national forest inventories. *Forests* 10, 800.

Kangas, A. Myllymäki, M. Mehtätalo L. 2023. Understanding uncertainty in forest resources maps. *Silva Fennica* **57**, 22026.

ACCEPTANCE SAMPLING IN STATISTICAL QUALITY CONTROL

D. Krapavickaitė

Lithuanian Statistical Society, Lithuania
e-mail: danute.krapavickaite@gmail.com

Abstract

A company receives a shipment consisting of the product lots from the producer. The products should be of the settled quality. Nevertheless the receiver has to ascertain this. In order to check the quality of the product lots, the sample of the products is selected and some quality characteristic is measured. Depending on the results of the quality control the lot is accepted or rejected. This stage of the quality control (Montgomery, 2013) is not aimed at the quality improvement, it is used just for the decision making on the lot acceptance.

The aim of a control for the receiver is to ensure that the lot is rejected with high probability if its quality characteristic reaches a non-admissible threshold. Producer's aim is to have an accepted lot with high probability if the amount of the low quality products in a lot is below the fixed threshold. Quality control of the lot items needs time and expenses. Therefore the problem is to choose the lot product sampling plan and sample size in order to satisfy the interests of both, receiver and producer, with the chosen probabilities of their risks (Schilling et al., 2009). The cost of the quality control for acceptance may be taken into account (Kobilinsky, 2005).

Under the traditional assumptions, the number of the insufficient quality items in a lot is distributed by the binomial distribution. Further study of the problem depends on the kind of the quality characteristic used: it may be an attribute (Tešnjak et al., 2014) or variable with some asymmetric distribution (Shahbaz et al., 2018).

A sequence of specific sampling plans are applied for acceptance sampling. Operating characteristic for a sampling plan which quantifies the sampling risk dependency on the probability of defective products in a lot is usually used.

The acceptance sampling will be presented from the point of view of the survey statistician.

Keywords: receiver, producer, lot, attribute, variable, operating characteristic.

References

Kobilinsky, A.; Bertheau, Y. (2005) Minimum cost acceptance sampling plans for grain control, with applications to GMO detection. *Chemometrics and Intelligent Laboratory Systems*, **75**, 189-200.

Montgomery D. C. (2013). *Statistical Quality Control*, John Wiley & Sons, Singapoure.

Schilling, E. G.; Neubauer D. V. (2009). *Acceptance Sampling in Quality Control*, Chapman & Hall/CRC, Boca Raton.

Shahbaz, A. H.; Khan, K.; Shahbaz, M. Q. (2018). Acceptance sampling plans for finite and infinite lot size under power Lindley distribution. *Symmetry*, **10**, 496, doi:10.3390/sym10100496.

Tešnjak, S.; Banovac, I. (2014). Analysis of attribute acceptance sampling properties. *WSEAS Transactions on Systems*, **13**, 720-729.

STATISTICS PRODUCTION SYSTEM 4.0

Thomas Laitila

Örebro University, Sweden
e-mail: thomas.laitila@oru.se

Abstract

Surveys based on the randomization theory have been and are still today a major source for the production of statistics at National Statistical Offices (NSOs). One major problem in the application of the theory is nonresponse. While the theory presupposes full response applications at NSOs turn out with some sample units not responding. The problem of nonresponse has not yet been satisfactorily resolved (e.g. Brick, 2013) and is today a major threat to validity of sample survey statistics. Partly because of this problem and increasing costs of sample surveys, nontraditional data sources are considered by NSOs as an alternative for production of statistics. Such data may also offer interesting features (e.g. Daas et al., 2015; Japec et al., 2015).

There are many challenging problems in using nontraditional data sources in production of official statistics. Issues around accessibility, involving legal and technical aspects, and sustainability are obvious. Validity of statistics considering variable and measurements are also. The most critical problem, the validity, objectivity, and interpretability of produced statistics, are seldom addressed, however. boyd and Crawford (2012) discuss these problems in the context of using Big Data in research.

To my knowledge, there are yet no example of nontraditional data sources replacing traditional ones in an official statistics product. This is remarkable in relation to the world wide interest to do so, and research conducted for at least two decades on how to use nontraditional data sources.

In this paper I suggest the long-time failures of incorporating nontraditional data sources depends on the focus on replacing traditional sample surveys. Instead, the question must be how to integrate nontraditional data sources in the existing production system to improve quality of statistics. This idea is expressed in the conference paper SCB (2017) and is here further elaborated. It summarizes into the Statistics Production system (SP) 4.0.

Keywords: official statistics, selectivity, model based inference, survey design.

References

boyd, D. and Crawford K. (2012). Critical Questions for Big Data, *Information, Communication & Society*, 15:5, 662-679.

Brick, J.M. (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review, *Journal of Official Statistics*, 29:3, 329-353.

Daas, P.J.H., Puts, M.J., Buelens, B. and P.A.M. van den Hurk (2015). Big Data as a Source for Official Statistics, *Journal of Official Statistics*, 31:2, 249-262.

Japec, L., Kreuter, F., Berg, M., Biemer, P., et al. (2015). Big Data in Survey Research: AAPOR Task Force Report, *Public Opinion Quarterly*, 79:4, 839-880.

SCB (2017). Statistics productions system 4.0. Contribution to Conference of European Statisticians, Economic Comission for Europe, Geneva, 19-21 June 2017

The Use of Web Data for EU Official Statistics. Case Studies on Online Job Advertisements and Online Based Enterprise Characteristics.

Jacek Maślankowski[1]

[1] University of Gdańsk, Poland; Statistics Poland
e-mail: jacek.maslankowski@ug.edu.pl

Abstract

The wide use of Big Data and Smart Statistics services in official statistics is a fact. In this paper we are describing two case studies: OJA – Online Job Advertisements and OBEC – Online Based Enterprise Characteristics. The paper relates to the implementation of Big Data as a source for official statistics within the last 6 years and plans for the next 3 years. First project regarding the OJA and OBEC data collection started in 2016 with the Big Data ESSnet 1st Grant. Then it was continued with the Big Data ESSnet II grant. Currently, under Web Intelligence Network initiative, implementation of the production process is on the way for both OJA and OBEC. We are expecting to move these two use cases into production in 2025.

The analysis of OJA data is well known in research papers. Its form varies from qualitative analysis (Hinrichs, Bundtzen, 2021) to quantitative based on text analysis (Cao et al., 2021). Currently Eurostat is analyzing several million job advertisements (Ascheri et al., 2022), based on the system developed in past by the European Centre for the Development of Vocational Training - Cedefop (Descy et al., 2019). For the OBEC data, more data sources are used, i.e. enterprise websites. The challenge is to define legal and methodological frameworks to be replicated by different NSI's, with respect to the business register data and associated URLs of enterprises.

The OJA project has been initiated to collect of job advertisements published online on the web as a new source of data in the field of labour market statistics. For this project several webscraping procedures were developed and tested, including privacy and legal issues of collecting web data. Another document was "webscraping policy" issued in collaboration with the OBEC project (wpc_deliverable_c1_ess_web-scraping_policy_template_2019_07_15.pdf, europa.eu). The development of OJA indicators and analyses of OJA data is now undergoing and data is presented as experimental statistics. Those data can be accessed via Eurostat OJA DataLab where they are regularly improved and updated. Data quality of the OJA data is necessary to decide what data can be disseminated as a part of official statistics. OJA Datalab contains information on many properties of open job advertisements in the market, and the daily dynamics are much richer than quarterly collected survey data. Thus, the data quality assessment also focused on how OJA data refers to official Job Vacancy Statistics (JVS) (Maslankowski et al., 2022).

Regarding OBEC the main goal is to collect the largest number of URLs to be used in web scraping process. The population of OBEC use case includes enterprises employing 10 or more employees. The population was defined as in ICT in Enterprises survey , i.e. "Definition of the enterprise website, From: Methodological manual for data compilers and users of the ICT survey, A6. Does your enterprise have a website? [Scope: enterprises with access to the internet, i.e. A1 > 0], [Type: single answer (i.e. Tick only one); binary (Yes/No); filter question]" (ESSnet WP-C, 2020). There are several ways to acquire URL list of enterprises. One way is to use official registers which in many cases have this information available publicly. For instance, in Poland there are enterprise registers like KRS or CEIDG, which are available online and it is possible to get the website URL if available for the company. There are also commercial databases with the information on enterprises in specific country. One of them is an ORBIS database by Bureau van Dijk which includes detailed information on enterprises, including URLs for numerous enterprises. But the responsibility for the data quality relies on the third-party vendor which

makes the data difficult to be used in official statistics. It is also important to note that fees apply for data acquisition.

Keywords: big data, machine learning,

References

Ascheri, A., Marconi, G., Meszaros, M., Reis, F. (2022) 'Online Job Advertisements for Labour Market Statistics using R', Romanian Statistical Review, (1), pp. 3–26.
Cao, L, Zhang, J., Ge, X., Chen, J., (2021) 'Occupational profiling driven by online job advertisements: Taking the data analysis and processing engineering technicians as an example', PLoS ONE, 16(6), pp. 1–20. doi: 10.1371/journal.pone.0253308.
Descy, P., Kvetan, V., Wirthmann, A., & Reis, F. (2019). Towards a shared infrastructure for online job advertisement data. Statistical Journal of the IAOS, 35(4), 669–675. https://doi.org/10.3233/SJI-190547
ESSnet WP-C Deliverable C6, Deliverable C6: Reference Methodological Framework for processing online based enterprise characteristics (OBECs) data for Official Statistics V .2, 2020, Available at: https://ec.europa.eu/eurostat/cros/sites/default/files/WPC_Deliverable_C6_Reference_Methodological_Framework_v2.0.pdf
Hinrichs, G. and Bundtzen, H. (2021) 'Communicating a Sales Job to Occupational Changers: A Qualitative Content Analysis of Online Job Advertisements', TEM Journal, 10(2), pp. 853–857. doi: 10.18421/TEM102-45.
Maslankowski et al. (2022), WP2: OJA and OBEC Software, Deliverable 2.1: WP2 1st Interim Progress Report.

The European Social Survey in Latvia: Insights from Face-to-Face to Self-Completion Survey Mode Comparison

Jurijs Ņikišins[1]
[1] University of Latvia, Latvia
e-mail: jurijs.nikisins@lu.lv

Abstract

In 2018, Latvia resumed its participation in the European Social Survey, an academic cross-national survey measuring attitudes, beliefs and behaviour patterns. The standard mode of data collection is based on face-to-face interviews at respondents' homes and it was used in Round 9 (2018 – 2020). However, for Round 10, a self-completion approach was specifically designed for use in countries that were particularly severely hit by the pandemic and unable to employ data collection by interviewers via home visits including Latvia. This contribution compares the performance of the national Latvian team and partner survey agencies involved in organizing and carrying out the ESS fieldwork in both rounds (9 and 10), focusing on the composition of the achieved net sample, response rates, population coverage, inconsistencies in answers to survey questions and item non-response. As the ESS is looking forward to adopting the self-completion approach starting from Round 13 in 2027, reflections on the ways to optimize the quality of the ESS data collection in Latvia are presented.

Keywords: European Social Survey, face-to-face, self-completion, response rate, survey questionnaire.

ARTIFICIAL INTELLIGENCE BUSINESS VIABILITY INDEX

Jaan Omblus

Statistics Estonia, Estonia
e-mail: jaan.omblus@stat.ee

Abstract

Businesses are exposed to market and operational risks that can lead to setbacks and potential business failures. In order to correctly identify and manage these risks, huge amounts of data must be analyzed and used as input for the decision-making process. In reality, companies do not have sufficient resources and skills for this process. If management were more data-driven, there would be fewer outages and more stability in companies.

The problem solution is to create an AI (artificial intelligence) solution with the capacity of using considerable amount of real-time market and business data and converting it into actionable management inputs that are available in real-time as basis for the business decision-making process.

The solution would improve the decision-making process in companies by ensuring higher capacity to make better use of the vast amount of data available. The more decisions are data driven, and the more potential market and operational risks are identified (and managed), the more the respective decision making business and the economic ecosystem are stable. The beneficiaries are business customers, employees, investors, business owners, public bodies and society in general. Fewer defaults and better economic performance are preferable to economic and social instability.

The proposed solution leverages the AI concepts to analyze market and business operational data in real-time and derive actionable business management fundamentals in the same timeframe as data is analyzed.

Machine learning, supervised learning and unsupervised learning concepts are used along with deep learning neural network solutions to learn from market and business behavior patterns and use the derived knowledge to support management decision-making process. The essence is to incorporate a huge amount of available data and leverage the interrelationships between different data sources to derive a summary actionable basis for real-time decision making.

The planned AI business support system is designed to achieve the following tactical objectives:
- creation of an analytical report on the respective business activities and market conditions;
- if relevant, generation of risk scenario-based alerts to review and change the course of business action;
- if relevant, production of performance recognition notes to emphasize whether the business is market oriented and successfully run;
- generation of solutions to address the risk scenarios when identified;
- forecasting the outcomes of the solutions offered;
- allowing business representatives to enter business-specific data into the system and to create the forecasts based on these additional inputs.

The proposed solution uses machine learning and artificial neural network tools to ensure the achievement of tactical objectives. From a technical point of view the system is able to perform the following activities:

- it learns the patterns of business practices in the given business sector;
- it learns site-specific issues that affect operations at a particular geographic location;
- it learns the business practices of the particular analyzed business entity;
- it evaluates the business practices of certain companies on industry and location-specific issues;
- it decides if there are specific risks that need to be communicated;
- using the available data inputs and considering potential customer additions, the system forecasts the possible development scenarios for the respective business unit, including corporate finance analytics as well as concepts such as net present value, weighted average cost of capital, internal rate of return, return on investments, customer acquisition costs, breakeven analysis etc.;
- the generated development scenarios could be communicated as solutions to identified risks or as further reinforcement of a positive development scenario.

The solution is made usable without any special competencies needed. Respective menu systems are provided to navigate the application. In case no considerable amendments needed the operational cost are negligible for the system use. As the system uses data for its operation the data needs to be updated and this cost is relevant to the cost of securing updated data.

Customization cost and time of the solution depends on the extent and essence of customization. In case of replacement of data sources from one "similar" data to another, the customization is efficient and the respective code is provided to users. The same applies if some data sources need to be excluded (not all data may be available in similar format in all regions). In case customization relates to adding essential functionalities to the solution the cost and time is dependent on the specific new requirements proposed.

As a result of application of the described system the bankruptcy rate is expected to fall 10% within three years of active system usage while the general insolvency would be expected to improve 25% within the same timeframe.

The Business Viability Index AI support system should be developed in such a way that it is reusable and therefore should have as few dependencies as possible on other specific software solutions. The corresponding documentation should enable reusability and be supplemented.

The system should be made scalable and integrable with other systems. It is essential that the data used in the system described here is available in other systems with which the present solution is intended to be integrated. For some data entries, the input data formats could be partially changed, some entries could be omitted without significant effect on the system result. The corresponding documentation for describing the interoperability must be created and supplemented.

The output of the AI support system, the specific advice, risk warnings, scenario forecasting as well as analysis results should only be available for use by the analyzed company. The results will not be displayed publicly or used as a basis for legal proceedings or categorizations.

The risks of the solution are wrong and misleading evaluations to particular companies' operational statuses, misleading forecasts produced and inappropriate development scenarios communicated. These risks have been mitigated via extensive testing of the solution and evaluation of the results received from testing. Experts and industry professionals are being involved in evaluation process judging AI technology trustworthiness.

Keywords: artificial intelligence, business viability, data integration

On the Use of Auxiliary Variables in Multilevel Regression and Poststratification

Yajuan Si[1]

[1] University of Michigan, USA
e-mail: yajuan@umich.edu

Abstract

Multilevel regression and poststratification (MRP) is a popular method for addressing selection bias in subgroup estimation, with broad applications across fields from social sciences to public health. In this paper, we examine the inferential validity of MRP in finite populations, exploring the impact of poststratification and model specification. The success of MRP relies heavily on the availability of auxiliary information that is strongly related to the outcome. To enhance the fitting performance of the outcome model, we recommend modeling the inclusion mechanisms conditionally on auxiliary variables and incorporating flexible functions of estimated inclusion probabilities as predictors in the mean structure. We present a statistical data integration framework that offers robust inferences for both probability and nonprobability surveys, addressing various challenges that arise in practical applications. Our simulation studies indicate the statistical validity of MRP, which involves a tradeoff between bias and variance, with greater benefits for subgroup estimates with small sample sizes, compared to alternative methods. We have applied our methods to the Adolescent Brain Cognitive Development (ABCD) Study, which collected information on children across 21 geographic locations in the U.S. to provide national representation, but is subject to selection bias as a nonprobability sample. We focus on the cognition measure of diverse groups of children in the ABCD study and show that the use of auxiliary variables affects the findings on cognitive performance.

Keywords: data integration, nonprobability sample, robust inference, model-based, selection/nonresponse bias

References

Y. Si (2023). On the Use of Auxiliary Variables in Multilevel Regression and Poststratification, under review, available at https://arxiv.org/abs/2011.00360.

NATIONAL DATA QUALITY FRAMEWORK FOR PUBLIC SECTOR DATA

Janika Tarkoma[1]

[1] Statistics Finland
e-mail: janika.tarkoma@stat.fi

Abstract

The amount of data available is increasing. However, a large quantity of data does not automatically mean all of it is useful. Data user must understand the content of the data – their limitations and errors. Like other Nordic countries, Finland has numerous public registers. These registers are utilized for their initial purposes, such as taxation or planning services, but also for secondary use. For example, our national census is register-based without fieldwork enumeration. As a general observation, the use of administrative registers could be more efficient in the public sector. This was the key motivation for our 'Opening up and using public sector data' project[1]. Data quality is one of the key elements that enables broader data use. It is important to define data content in a uniform manner so that all users understand the content and its quality in the same way. Uniform terminology and definitions are at the core of data quality criteria.

The data quality criteria are central to the national data quality framework[2]. Each criterion has at least one indicator providing concrete values. Quality criteria and indicators reflect the current situation. They must be adjusted as the world changes. Therefore, the framework includes a management model and tools that support further development.

The data quality criteria and indicators serve as tools for describing and evaluating data quality. They answer the question, 'What does data quality mean?'. Data quality criteria, or quality perspectives, are divided into three groups. Each group addresses a different question about data quality. The objective of these quality criteria groups is to aid in understanding and remembering the key points about data quality. The first group of criteria is called 'How well does information describe reality?' This group contains five criteria, each focusing on describing what should be included in the dataset and assessing how well this goal has been achieved. The second group of criteria, addresses the question, "How has the information been described?". This group focuses on metadata and compliance. The third group addresses the question, "How can I use the information?" It's crucial to note that our project was centered on opening up and reusing data. Therefore, easy access and usability are considered integral parts of data quality.

Our data quality framework acts as a comprehensive starting point for public sector data quality improvement, affecting administrative registers in Finland. For organizations seeking to improve data quality, assessing the current state is crucial. Quality evaluation helps identify areas for quick improvement and areas requiring more collaboration and effort. Regular assessments not only prioritize improvements but also track progress over time. If a change in the production process impacts data quality negatively, this can be corrected.

Keywords: data quality, data quality framework, quality assessment, quality criteria

[1] Opening up and using public data -project: https://vm.fi/en/opening-up-and-using-public-data1

[2] Data Quality Framework: https://stat.fi/dataquality

3 SPECIAL INVITED SESSIONS

Municipalities' size estimation correcting administrative data for coverage errors and misclassification: A Bayesian approach

V. Ballerini[1], M. Di Zio[2], B. Liseo[3] and S. Toti[2]

[1] University of Florence, Italy
e-mail: veronica.ballerini@unifi.it

[2] Istat, Italy
e-mail: dizio@istat.it, toti@istat.it

[3] Sapienza University of Rome, Italy
e-mail: brunero.liseo@uniroma1.it

Abstract

In the process of maintaining the Italian permanent Population Census, the Italian National Statistics Institute (Istat) relies on the *Base Register of Individuals* (BRI, hereafter) to compute population sizes at different levels of aggregation, correcting the administrative data for under- and/or over-coverage. To evaluate the probabilities of under- and over-coverage, Istat conducts two surveys and the *adjusted* population counts estimates are currently obtained by weighting the BRI counts with the ratio of such probabilities. However, this process only produces point estimates for the population sizes, and to obtain an uncertainty quantification of such estimates, a complex bootstrap procedure must be performed, which must consider the complex sampling planes of the two above-mentioned surveys. To overcome such complexities, we approach the problem in a fully Bayesian way by treating the observed BRI counts and the number of under-covered individuals as realizations of random quantities, whose distributions' parameters are functions of the probabilities of over- and under-coverage. The proposed approach makes the model more flexible and able to incorporate uncertainty from different sources. Indeed, we also allow for the concrete possibility that BRI units might be misclassified with respect to some of the individual characteristics, such as citizenship.

We produce an estimate of the posterior distribution of the Italian population size at a municipal level using Markov Chain Monte Carlo methods, overcoming difficulties associated with uncertainty quantification. We illustrate the procedure using the subset of municipalities of less than 18,000 residents (Non-Auto Representative municipalities) in 2018. We also analyze the sensitivity of the results to different prior specifications, demonstrating the robustness of our method.

Keywords: Bayesian inference, permanent census, population size estimation, coverage error, misclassification error.

References

Mancini, L. and Toti, S. (2014) Dalla popolazione residente a quella abitualmente dimorante: modelli di previsione a confronto sui dati del censimento 2011. *Technical report, Istat Working Paper.*

Unit level small area models for business survey data

C. Bocci[1] and P.A. Smith[2]

[1] University of Florence, Italy
e-mail: chiara.bocci@unifi.it

[2] University of Southampton, UK
e-mail: p.a.smith@soton.ac.uk

Abstract

Small area estimation (SAE) encompasses a wide range of methods (Rao and Molina, 2015) that have become a significant component in the official statistician's toolkit, with many applications to a wide range of variables and domains in many different types of surveys. Nevertheless, SAE is still uncommon in business surveys because of challenges arising from the skewness and variability of many size-related variables. SAE methods are generally based on mixed effects models which have assumptions of normal errors, but skewed variables violate this assumption. Moreover, the specific characteristics of sample designs used in business surveys (detailed stratification, non-negligible sampling fractions, large variations in estimation weights) affect the models on which small area estimates are based.

Several approaches have been suggested to deal with such skewed data in different ways: through transformation, by employing robust models to accommodate outlying tail observations, and by directly modelling the skewed distribution. Smith et al. (2021) examined a range of robust approaches, which reduce the impacts of observations in the tails of skewed distributions, in a dataset with known outcomes. Here we replicate this analysis with a second dataset of Italian retail businesses, and compare it with a second group of methods based on transformations of the initial data before modelling. The back-transformed predictions need bias adjustments to produce estimates with acceptable quality. We review the transformation-based methods which have been proposed in the literature and make an assessment of the best approaches to use for business surveys based on our repeated sampling simulation study.

Keywords: robust estimation, outliers, skewed distribution.

References

Rao, J.N.K., Molina, I. (2015) *Small area estimation.* John Wiley & Sons, Hoboken.

Smith, P.A., Bocci, C., Tzavidis, N., Krieg, S., Smeets, M.J.E. (2021), Robust estimation for small domains in business surveys. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **70**, 312–334.

SOME REFLECTIONS ON BAYESIAN INFERENCE IN OFFICIAL STATISTICS

Brunero Liseo[1]
[1] Sapienza University of Rome, Italy
e-mail: brunero.liseo@uniroma1.it

Abstract

Survey sampling and, more generally, Official Statistics are experiencing an important renovation time. On one hand, there is the need to exploit the huge information potentiality that the digital revolution made available in terms of data. On the other hand, this process occurred simultaneously with a progressive deterioration of the quality of classical sample surveys, due to a decreasing willingness to participate and an increasing rate of missing responses.

The switch from survey-based inference to a hybrid system involving register-based information has made more stringent the debate and the possible resolution of the design-based versus model-based approaches controversy. In this new framework, the use of statistical models seems unavoidable and it is today a relevant part of the official statistician toolkit. Models are important in several different contexts, from Small area estimation to non-sampling error adjustment, but they are also crucial for correcting bias due to over and under-coverage of administrative data, in order to prevent potential selection bias, and to deal with different definitions and/or errors in the measurement process of the administrative sources.

The progressive shift from a design-based to a model-based approach in terms of super-population is a matter of fact in the practice of the National Statistical Institutes. However, the introduction of Bayesian ideas in official statistics still encounters difficulties and resistance. In this work, we attempt a non-systematic review of the Bayesian development in this area and try to highlight the extra benefit that a Bayesian approach might provide. Our general conclusion is that, while the general picture is today clear and most of the basic topics of survey sampling can be easily rephrased and tackled from a Bayesian perspective, much work is still necessary for the availability of a ready-to-use platform of Bayesian survey sampling in the presence of complex sampling design, non-ignorable missing data patterns, and large datasets.

Empowering Surveys with Generative AI

Johanna Laiho-Kauranne[1]

[1] DAIN Studios, Finland
e-mail: johanna.laiho-kauranne@dainstudios.com

Abstract

While the AI has been used widely for few years across industries, popularity of generative AI has boomed since the end of 2022 in many daily applications and digital services, utilizing its feature to generate novel content rather than building logic upon existing data, (García-Peñalvo and Váquez-Ingelmo, 2023). Generative AI has the potential to revolutionize many industries, e.g advertising, entertainment, and education (Gozalo-Brizuela and Garrido-Merchan, 2023). This study focuses on its usability for survey industry. Encouraging results by Beck, Dumbert and Feuerhake (2018) show that already in 2018 majority of OECD countries had some applications of machine learning in official statistics, thus it is expected that also new methods such as generative AI will raise interest if ensured the responsible and ethical usage in the field of official statistics and surveys.

The usage of generative AI has potential but also caveats as the facts and fiction can be indistinguishable for service users, and if not used responsibly can deteriorate public trust. In addition, data protection of generative AI solutions cause concerns amongst the public, and experts of data protection professionals. Significant risks related to *bad data* generate by AI are inaccurate decision making, spreading misinformation, privacy violations, legal liabilities, damage to trust (Tang et al., 2023), and may further impact trust to democracy (Arguedas and Simon, 2023). Taking the challenging risks into account, the advantages for improving surveys utilizing the potential of generative AI is the main incentive to explore responsible and ethical use of generative AI for survey research.

There is a global unmet challenge for surveys that are battling with the long prevailing trend of reduction in response rates, reaching levels of severe questioning of the reliability of the estimates. This has been studied for decades, but instead of finding effective strategies for promising initiatives, the survey organizations are increasingly struggling with low or negligible participation to surveys impacting the reliability and accuracy of the survey estimates. Generative AI can be used as a tool survey design in multiple steps of survey process. This is demonstrated in the context of the GSBPM model. Potential use case is defined as using generative AI especially in the design and data collection stages of the GSBPM model, e.g. in improving survey design, respondent approaches, contact strategies and conversion strategies for soft refusals.

Considering the relation to the fundamental aim of surveys providing reliable facts about society or population, the use of generative AI may be argued to be in contradiction with the ultimate purpose of generating reliable information. Thus, it is emphasized that the focus is not on supplementing and creating information arbitrary information content, instead the purpose is to explore usage of generative AI to improve the quality of surveys and the way their processes are managed. In other words, the AI-generative content (AIGC) has been exclude from this study. The data quality components in focus are survey accuracy, and relevance, reviewing also potential impact to survey comparability, timeliness, and accessibility.

The study explores the potential use of generative AI for response conversion strategies. The traditional differentiative motivation strategies base on ad hoc strategies or tailored data analytics can be further enriched using large language models (LLM). These LLMs can be developed so that it is

Indistinguishable from human-generate content (Arguedas & Simon, 2023). The use case is narrowed to reviewing potential applications that require quick adaptation of strategies to build motivation and raise interest of the selected units to be surveyed. Application potential may also reach to support tool for interviewers, or for contact persons to help them to build the best motivation strategies for retrieval of requested information. New technology is fast developing, and promising new solutions are using generative AI text-to-speech diffusion model applications, which features also provide incorporated controllable emotional models (Zhang et al. 2023).

To conclude, AI can be used to explore the validity and relevance of the underlying assumptions of the survey in fast developing societies and survey phenomena. The purpose of this study is to evaluate suitability of generative AI solutions to traditional population surveys. As the usage of generative AI is likely to raise contradiction and successful surveys are based on the trust of data providers, we review the acceptance for information production and survey industry in the AI landscape. As the EU is currently working on the world's first AI law, it is important to pave the way for utilizing future potential and understand the scalability, limitations, and risks of AI in an objective manner. Thus, statistical ethics and code of practice, merely in the European context are reviewed to examine the fit for purpose and preparedness of the legislative grounds.

As the legislative groundings are still evolving, and the generative AI is rather recent, this study is limited to conceptual and theoretical level, and the purpose is to be a starting point for further pilot cases.

Keywords: generative AI, surveys, GSBPM, survey design, respondent conversion.

References

Arguedas, A. & Simon, F. (2023). *Automating democracy: Generative AI, journalism, and the future of democracy.* Institute for Ethics in AI.

Beck, M., Dumbert, F., and Feuerhake, J. (2018). *Machine Learning in Official Statistics*. Preprint.

Gozalo-Brizuela, R. and Garrido-Merchan, E. (2023) *A survey of Generative AI Applications*. Preprint.

García-Peñalvo, F. and Váquez-Ingelmo, A. (2023). What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in Generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence.*

Zhang, C., et al. (2023) *A Survey on Audio Diffusion Models: Text To Speech Synthesis and Enhancement in Generative AI*. Kyung Hee University, South Korea. Manuscript submitted to ACM.

Unleashing the Power of Machine Learning and Artificial Intelligence: Advancing the Production of Official Statistics through MLOps implementation

R. Piela[1] and H. Luomaranta-Helmivuo[2]

[1] Statistical Office, Finland
e-mail: riitta.piela@stat.fi

[2] Statistical Office, Finland
e-mail: henri.luomaranta@stat.fi

Abstract

1. Introduction

1.1. Background and context

Official statistics provide objective and reliable information about various aspects of society. Accurate statistics help identify societal challenges, assess policy effectiveness, and shape evidence-based strategies to address issues such as unemployment, and inequality. Official statistics serve as a critical tool for monitoring economic performance, tracking demographic changes, and assessing the impact of government policies. Reliable data allows governments to identify areas that require intervention, assess the success of existing programs, and allocate resources efficiently to achieve societal goals.

The availability of official statistics should ensure transparency and accountability in decision-making processes. Transparent statistics enhance public trust in government actions and foster open discussions on public policy matters. Implementing ML/AI in statistical processes need the implementation of MLOps (a paradigm for the operationalization of ML efficiently and transparently) principles and the infrastructure and platform to implement these principles. The upcoming AI Act from EU may also set new requirements for statistical production or at least raise people's awareness of the requirements for information produced with the help of artificial intelligence.

1.2. Problem statement

Statistical institutes around the world are increasingly adopting machine learning (ML) techniques to enhance official statistics production. However, globally many statistical institutes encounter difficulties in implementing MLOps principles, such as transparency and reproducibility or they are not yet aware what kind of elements should exist when using AI in statistical production. In the future, decision-makers and the public may lack confidence in the validity and accountability of the ML-driven statistics and its impact on policymaking and governance if the AI driven processes are not transparent enough.

The importance of official statistics in guiding policy formulation, resource allocation, and governance decisions cannot be overstated. In the era of AI and ML, leveraging these advanced technologies has the potential to revolutionize statistical production, improve accuracy, and speed up the ability of statistics to quickly describe the changes taking place in society. Nevertheless, the lack of transparency and reproducibility in ML models raises concerns about their integrity and potential biases.

To gain the trust of decision-makers and the public, statistical institutes must prioritize Responsible AI and MLOps principles, transparency, and reproducibility. Transparency ensures that the decision-making process and ML algorithms are open and explainable, providing stakeholders with insight into how conclusions are reached. Reproducibility guarantees that results can be independently verified, increasing the confidence in the accuracy and reliability of statistical outputs. Concepts such as

Responsible AI and MLOps are not fully established. They contain many common principles, but usually Responsible AI includes also principles related to data protection and data security.

MLOps, an amalgamation of various definitions, represents a paradigm that encompasses best practices, essential concepts, and a development culture aimed at facilitating the end-to-end lifecycle of machine learning products. At its core, MLOps merges principles, tools, and techniques from both machine learning and traditional software engineering to design and construct complex computing systems. However, MLOps is not merely a collection of principles or a stack of technological elements; it extends its reach to encompass data, processes, roles, and methods. Embracing MLOps presents challenges due to its multidimensional nature. One of the prominent challenges lies in navigating the diverse dimensions it encompasses. The implementation of MLOps can be intricate, necessitating the adoption of new roles, such as ML engineers, data scientists, and data engineers, along with the establishment of robust collaboration mechanisms among these roles. Despite the complexities, embracing MLOps is essential for streamlined and accountable AI integration, empowering organizations to leverage the full potential of machine learning and artificial intelligence for advancing official statistics.

By addressing the challenges in implementing MLOps principles, statistical institutes can build a foundation of trust and accountability in their ML driven statistical processes. This trust is crucial for fostering evidence-based policymaking, efficient governance, and the public's acceptance of AI and ML advancements in official statistics. The motivation is to establish a robust and ethical framework that empowers decision-makers to make informed choices and the public to have confidence in the data driving policy and governance decisions.

2. AI in Statistical production

2.1. Current Landscape of ML adoption in Statistical Institutes

As mentioned, machine learning (ML) and artificial intelligence (AI) can bring improvements to the accuracy, efficiency, and timeliness of official statistics production compared to traditional manual processes, or even compared to rules-based systems of validation. Once properly established, ML models can process and analyze data in real-time, allowing statistical offices to produce more up-to-date and accurate data. This is especially valuable in rapidly changing situations of society. ML and AI algorithms can handle large-scale datasets efficiently, enabling statistical offices to process and analyze massive amounts of data quickly and reliably. Manual data validation and editing processes are also prone to human errors. ML-driven automation reduces these errors and is likely to lead to more accurate and reliable official statistics. ML and AI can automate labor-intensive tasks in statistical production processes. Many statistical institutes understand these potential benefits and have explored ML in automating classification and editing tasks, especially when the data are large and difficult. Some effort has been put into exploring new areas such as nowcasting or exploring new data sources such as satellite images. But systematic adoption of ML is not yet achieved. This is likely because these explorations have been done on a case-by-case basis, by few experts dealing simultaneously with many other tasks, and the benefits are not realized at the level of the organization, even less in the statistical system as a whole.

2.2. The Importance of MLOps Principles for Trustworthy Statistical Production

The systematic adoption of transparency, reproducibility and explainability in ML-driven processes play a crucial role in ensuring that decision-makers and the public to understand and trust the outcomes of AI and ML applications in statistical production and ensures that systemwide adoption can be possible. It fosters accountability, promotes ethical decision-making, and facilitates effective communication between stakeholders, leading to better-informed policies and governance decisions. Automated reports and visualization can be great benefit to users of statistics, but only if they are trusted.

In this paper, we underscore the significance of MLOps principles like transparency and reproducibility in the context of machine learning adoption in statistical institutes and aims to inspire statistical institutes to embrace MLOps principles, thereby fostering greater trust among decision-makers and the

public in the credibility and impact of ML-driven statistical outputs. Additionally, the paper seeks to contribute to the responsible and ethical implementation of AI in official statistics, promoting evidence-based policymaking and efficient governance by critically discussing the experiences in Finland, noting that there is much work to be done.

MLOps principles represent a set of best practices that aim to streamline and standardize the entire ML model lifecycle, from development to deployment and maintenance. These principles are fundamental for ensuring that ML models are effectively managed, monitored, and governed, facilitating efficient collaboration among data scientists, decision-makers, and stakeholders.

2.3. Technology and processes implementing MLOps principles

Technology plays a central role in enabling MLOps principles and practices, as it provides the infrastructure, platforms, and frameworks necessary to implement these principles seamlessly. Tools and platforms (or frameworks) like TensorFlow Extended (TFX), MLflow, and Kubeflow offer a suite of tools for automating ML workflows, versioning models, and tracking experiments, promoting transparency and reproducibility. By integrating MLOps principles with appropriate technological solutions, statistical institutes can optimize data collection, validation, analysis, and dissemination processes, leading to more accurate, timely, and trustworthy official statistics.

ML platforms typically provide tools for managing data, model training, serving models, and monitoring, and support features that promote transparency and reproducibility, such as model tracking, artifact management, and version control. In fact, one may consider replacing the entire statistical production process pipelines with the available platforms. Countries such as France, Norway and Lithuania seem to be headed this way.

While choosing an ML platform for implementing MLOps principles, it's essential to consider the specific needs and requirements of your organization. Each platform may have different strengths and may be better suited for certain use cases or deployment environments. Additionally, integration with existing infrastructure and compatibility with the chosen ML frameworks should be considered.

3. Challenges and Ethical Considerations

3.1. Ethical implications of using AI in producing official statistics

Ethics is a critically important issue when using ML in producing data for official statistics due to several key reasons like fairness and equity. ML-driven statistical models have the potential to impact individuals and communities directly through policy decisions. If ML models are not designed with fairness in mind, they may inadvertently perpetuate biases, leading to inequitable outcomes and exacerbating existing societal disparities.

Official statistics influence critical decisions and policies that affect people's lives. Lack of transparency in ML models can lead to "black-box" decision-making, where it's challenging to understand how conclusions are reached. Transparent ML ensures accountability, enabling stakeholders to verify and challenge the basis of statistical outputs.

3.2. Notes regarding to training data

ML-driven official statistics depend on the quality and representativeness of the data used for training. Ethical practices ensure that data used in ML models accurately reflect the population they represent, avoiding skewed or inaccurate results. Establishing governance frameworks for ML use in official statistics ensures that decisions about model deployment and data utilization are made responsibly and with careful consideration of potential impacts on society.

Official statistics often involve sensitive and personal data. ML models must be designed with robust data privacy measures to protect individuals' confidentiality and adhere to data protection regulations. ML models can be susceptible to adversarial attacks, wherein malicious actors manipulate data inputs

to produce inaccurate results. Ensuring the accuracy of ML-driven official statistics is essential for making informed and reliable decisions.

When using data from individuals, obtaining informed consent becomes vital to respecting their autonomy and privacy rights. ML practitioners must ensure that individuals are aware of how their data will be used and have the option to opt-out if desired. ML can make predictions and decisions based on patterns that are not easily understood by humans. This introduces the risk of unintended consequences if models are deployed without careful consideration of potential impacts. ML, as any modeling, can amplify biases present in the data it learns from. Ethical considerations mandate the detection and mitigation of biases to ensure that AI models produce unbiased and impartial results.

3.3. Maintaining public trust

Public trust is one the most important issues in implementing AI systems in official statistics. The adoption of ethical AI practices in official statistics fosters public trust and confidence in the results and decision-making processes. Transparency and fairness lead to greater acceptance of AI-driven insights by policymakers and the public.

The importance of ethics in using AI in producing data for official statistics cannot be overstated. Ethical considerations are crucial for ensuring fairness, transparency, accountability, data privacy, and the accuracy of AI-driven statistical outputs. By upholding ethical principles, statistical institutes can leverage AI to produce data that supports informed policymaking, governance, and decision-making while protecting individuals' rights and promoting societal well-being. Canadian Trust Centre is an excellent effort building trust between public and statistical institute.

3.4. Reproducibility: a cornerstone of scientific research and statistics

Reproducibility is paramount in scientific research and in official statistics. By promoting transparency and fostering collaboration and trust, reproducibility ensures the credibility of results. In official statistics, it serves as a reliable foundation for evidence-based decision-making and contributes to scientific progress.

Both ethical considerations and reproducibility are essential components when deploying AI in official statistics. Maintaining reproducibility ensures that the statistical outputs are verifiable, transparent, and trustworthy, providing a robust basis for effective decision-making. By embracing these principles, statistical institutes can navigate the challenges of AI in official statistics while maintaining public trust and promoting responsible and ethical use of AI technologies.

4.Implementing MLOps, steps towards success at Statistics Finland

4.1. Versioning

In the domain of MLOps, achieving reproducibility is of paramount importance to ensure the credibility and reliability of machine learning models. The main components that play a pivotal role in attaining reproducibility are versioning of the model, data, and code. Model versioning entails the systematic tracking and recording of changes made to the ML model throughout its development lifecycle. By preserving the model's version history in a model registry, it becomes possible to precisely recreate and compare results at different stages of the development process, enabling practitioners to understand the model's evolution and make informed decisions based on its performance. Similarly, data versioning involves meticulously documenting the datasets used for training and evaluation, including data preprocessing steps and any modifications made during the data preparation phase. Proper data versioning ensures that the same dataset can be reliably used in future iterations, allowing for the validation of results, and reducing the risk of introducing data-related discrepancies.

Lastly, code versioning involves the systematic management of the software code used to develop, train, and deploy the ML model. By maintaining a clear record of code changes and dependencies, data scientists and statisticians can reproduce experiments and workflows accurately, leading to consistent results and facilitating collaboration among team members. In combination, model, data, and code

versioning in MLOps contribute to the establishment of a robust and transparent framework, fostering confidence in the reproducibility of machine learning outcomes.

Statistics Finland has implemented a comprehensive model register for ML models, ensuring that each model's version history is well-documented and traceable. Additionally, the use of Model Cards is adopted, which serve as repositories for storing essential metadata related to every model. This includes important information about the model's architecture, training data, hyperparameters, and performance metrics. By maintaining detailed version records and utilizing Model Cards, we enhance transparency, reproducibility, and accountability in our machine learning practices.

4.2. Automation of the workflow – the levels

As the demand for AI and ML solutions continues to soar, organizations face increasing pressure to streamline their ML workflows and enhance model deployment efficiency. Google Cloud's MLOps automation, which has become somewhat of an MLOps standard, has emerged as a transformative approach to address these challenges, presenting three distinct levels of maturity. Each level represents a significant milestone in the automation journey, enabling organizations to achieve higher levels of operational efficiency, model reproducibility, and collaboration among different ML roles.

At the foundational level of MLOps automation, organizations rely on manual processes to manage their ML workflows. Data scientists and ML engineers typically conduct each step of the ML lifecycle manually, from data preprocessing and feature engineering to model training and deployment. Code and data versioning are often managed through traditional version control systems, leading to potential inconsistencies and difficulties in tracking changes. **Level 0** automation serves as a starting point for organizations, providing valuable insights into the intricacies of their ML workflows and paving the way for further advancements in MLOps automation.

As organizations mature in their MLOps journey, they seek to address the challenges encountered at Level 0 by adopting MLOps tools and platforms. **At Level 1**, automation becomes more prevalent in data versioning, model training, and deployment processes. MLOps tools facilitate the automation of ML workflows, offering capabilities for versioning models, tracking experiments, and managing data pipelines. These tools enhance transparency, traceability, and collaboration among ML roles, enabling better reproducibility and efficiency in model development. However, certain elements of the ML process may still require manual intervention, limiting the level of end-to-end automation achieved at this stage.

The pinnacle of MLOps automation is reached at **Level 2,** where organizations achieve full end-to-end automation of their ML workflows. At this stage, advanced MLOps platforms seamlessly orchestrate the entire ML lifecycle, from data ingestion to model deployment and monitoring. AutoML solutions further enhance the efficiency of model development by automating hyperparameter tuning and architecture selection. Fully automated MLOps streamlines collaboration, fosters transparency, and ensures model reproducibility at scale. The robustness of Level 2 automation empowers data scientists and ML engineers to focus on high-value tasks, driving innovation and accelerating the delivery of AI solutions.

At Statistics Finland, we have made significant progress in automation, but the current level falls somewhere between 0 and 1 on the automation scale. While we have successfully implemented many components and have partially addressed versioning requirements, there is still ongoing development work required to achieve seamless automation across process states. Our team is actively working to bridge this gap and enhance our automation capabilities, with the goal of streamlining our processes, improving efficiency, and ensuring the reproducibility of our models and data. As we continue to invest in automation and embrace best practices, we aim to elevate our capabilities to a higher level. It is a difficult management exercise to determine the sweet spot lies where efficiency gains start to show throughout the organization and input costs start to sink in relation to output as compared to the situation without implementing MLOps. Nevertheless, these are necessary sometimes to convince higher leadership. Often, we must rely on demonstrating gains by showing an example and trying to analyze what cost savings are achieved in manual data editing phase which is sometimes offset with increased

investments in technology when a machine learning is adopted, but this does not show the benefits of MLOps, which is a strategy that ensures that the one-off-ML model is maintained, transparent, and available for re-use. What makes it more difficult is that sometimes, short term gains are not favorable for the adoption of a single machine learning model, especially when evaluated case-by-case. However, with system-wide adoption of MLOps and with increased potential output, these gains should be easy to see. MLOps is a strategic investment, more than allowing some statistical team to develop a ML approach to a specific problem.

4.3. Monitoring the performance

Monitoring the performance of machine learning (ML) models is of utmost importance to ensure their continued effectiveness and avoid degradation over time. ML models are typically trained on historical data, and as the underlying patterns in the data change, the model's performance can deteriorate. Monitoring enables us to detect and address these changes promptly, maintaining the model's accuracy and reliability.

The monitoring process involves regularly evaluating the model's performance metrics, such as accuracy, precision, recall, and F1 score, on a representative dataset. By comparing these metrics against predefined thresholds, we can identify deviations and potential performance degradation. Additionally, tracking other relevant statistics, such as data distribution shifts and input-output correlations, can help spot anomalies that might affect the model's performance.

To monitor model performance effectively, a robust and scalable monitoring system is required. This system should be integrated into the MLOps process, enabling automated and continuous monitoring. Regularly retraining the model with fresh data can further help mitigate performance degradation, ensuring that the model adapts to changing patterns in the data.

Moreover, ongoing monitoring facilitates the detection of concept drift, where the relationships between features and target variables change over time. By identifying concept drift early, data scientists can take appropriate actions, such as retraining the model on more recent data or updating the feature set to account for new trends.

At Statistics Finland, our efforts to identify data drift and ensure model monitoring have been ongoing. We have conducted extensive studies using various libraries and off-the-shelf solutions to explore the best approaches for our needs. Notably, we have extensively tested the capabilities of the Alibi Detect library, which offers a collection of state-of-the-art algorithms and methods that can seamlessly integrate into our existing machine learning pipelines. While we have made significant progress, we are still in the process of determining the most suitable implementation strategy for monitoring our ML models effectively. Our commitment to continuous improvement ensures that we will make informed decisions to achieve optimal model performance and maintain the highest standards of data quality and accuracy.

4.4. Explainability

Although explainability and especially explainability of deep learning is not included as an independent principle in MLOps it is still part of transparency. When using deep learning models, achieving explainability is particularly challenging due to the complex and non-linear nature of these models. However, researchers have developed various techniques to provide some level of insight into how deep learning models make decisions.

It's important to note that while methods provide some insights into model behavior, they might not offer a complete understanding of why a deep learning model makes specific decisions. Achieving full explainability in deep learning models remains an ongoing research challenge. The choice of explainability method may depend on the specific use case and the level of interpretability required for the application.

Explainability of deep learning at Statistics Finland is still in its nascent stage. As we delve into the domain of deep learning, we recognize the significance of achieving interpretability for our models. To

tackle this complex challenge, we have actively sought collaborations with academic partners, including those from the FCAI (Finnish Center for Artificial Intelligence) funded by the Research Council of Finland. Explainability in deep learning is a highly theoretical and intricate area, demanding a cautious and methodical approach. We are committed to taking small yet deliberate steps to embark on this journey, working collaboratively to unravel the inner workings of the deep learning models adopted and make them as transparent and interpretable as possible. Through these efforts, we aim to enhance the trustworthiness of our AI-driven statistical insights and pave the way for responsible and ethical implementation of deep learning in official statistics.

4.5. ML Platforms

At this crucial juncture, Statistics Finland is exploring various possibilities to optimize the whole data processing infrastructure (data stack) and leverage the full potential of AI. One option is to replace our in-house ML platform with an off-the-shelf machine learning platform, allowing us to benefit from the advancements and features offered by established solutions. By shifting away from in-house development, we could streamline our processes and reduce maintenance overheads. Additionally, in parallel, we are considering the integration of pre-trained models within the cloud infrastructure of a commercial cloud vendor. While preserving the advantages of the cloud environment, this approach enables us to tap into the power of AI and benefit from the pre-trained models provided by Microsoft/Open AI. However, as we progress towards these potential solutions, we remain mindful of our data's privacy and sensitivity and the general problems with pre-trained foundation models. Also, certain data may require on-premises solutions to ensure the utmost security and compliance with regulatory requirements.

While it may not be classified as a conventional success story, Statistics Finland has shown the foresight and strategic positioning to harness the opportunities presented by AI in official statistics. By being in the right place at the right time, the organization has successfully identified and acted upon the essential elements needed for the effective implementation of AI. This proactive approach has allowed Statistics Finland to make significant strides in adopting AI technology, setting the stage for potential future successes in enhancing the accuracy, efficiency, and impact of official statistics through responsible AI practices.

5. Potential future developments in AI and ML for official statistics

The advent of artificial intelligence (AI) has ushered in transformative opportunities for statistical institutes, promising to revolutionize various stages of the statistical process, from data collection and analysis to validation and dissemination. In this chapter, we delve into the exciting possibilities that AI holds for statistical data production, exploring potential applications, benefits, and challenges on the horizon.

One promising area where AI can make a significant impact is in automating and enhancing data collection processes. By harnessing advanced techniques like natural language processing (NLP) and computer vision, AI facilitates web scraping, social media analysis, and satellite imagery processing, providing real-time and large-scale data streams. This empowers statistical institutes to monitor and respond more effectively to dynamic changes in society. These can improve existing statistics by providing new data sources against which they can be analysed.

Machine learning algorithms, such as deep learning and neural networks, offer powerful tools for data analysis and modeling. AI-driven analytics can identify intricate patterns, correlations, and anomalies in large datasets, enabling statistical institutes to gain deeper insights into complex socio-economic phenomena and forecast trends with improved accuracy. These can lead to new statistical products to complement more established statistical outputs that are based on well-established frameworks.

AI's influence extends to data validation and quality control processes, streamlining the often-time-consuming manual validation and editing tasks. With automated anomaly detection algorithms, potential errors, inconsistencies, and outliers can be swiftly identified, allowing statisticians and subject matter experts to focus on data verification and ensuring the accuracy of statistical outputs.

Furthermore, AI-powered data visualization tools pave the way for real-time and interactive data dissemination. Dynamic visualizations, dashboards, and geospatial mapping enhance data accessibility and understanding for decision-makers and the public, fostering greater engagement with official statistics.

In survey design and sampling methodologies, AI optimization can contribute once scientifically validated. AI-enabled adaptive surveys can dynamically adjust questions based on respondents' previous answers, ensuring personalized and relevant data collection experiences. This leads to more efficient and representative data collection, enhancing the quality of statistical insights.

It is crucial to recognize that AI should not completely replace human expertise but rather complement it in highly specialized field such as official statistics, once the low-hanging fruit of eradicating repetitive tasks is achieved. Statistical institutes can adopt a human-in-the-loop approach, where AI collaborates harmoniously with statisticians, subject matter experts, and decision-makers. This collaborative synergy facilitates more robust analysis, validation, and interpretation of AI-driven statistical outputs.

6. Challenges and Ethical Considerations of the use of foundation models

Foundation models represent a significant shift in the paradigm of AI because they introduce a new approach to language understanding and knowledge representation. Traditionally, AI models were designed with specific tasks in mind and required extensive fine-tuning to perform well on those tasks. However, foundation models, such as large language models like GPT models, are pre-trained on vast amounts of data from the internet, learning the structure of language and general knowledge in an unsupervised manner.

The shift towards foundation models has democratized access to advanced AI capabilities, allowing developers to access state-of-the-art language understanding with minimal effort. It reduces the barriers to entry for AI development and accelerates the pace of innovation in natural language processing tasks.

While leveraging pre-trained models can significantly enhance the efficiency of statistical analysis and insights generation, it's essential to note that pre-trained models may not fully align with the specific needs of official statistics. Customization options might be limited and adapting the models to the specific nuances of official statistical domains could become challenging.

The lack of knowledge about the training data and process raises transparency and explainability concerns. Especially for statistical offices, it is crucial to ensure that the models' underlying data and algorithms align with ethical guidelines and produce interpretable outputs. Hosting and training models on external platforms raise data privacy and security considerations, particularly since official statistics often deal with sensitive data. Sharing such data with external providers requires robust data protection measures.

Dependence on external platforms and models necessitates consideration of the long-term sustainability of such arrangements. It is important to assess potential risks and formulate contingency plans in case the platform or provider becomes unavailable or discontinues services. Delegating model training to external entities requires clear accountability and data governance mechanisms to ensure that the models adhere to official statistical standards and legal frameworks.

Mitigating biases in the usage of foundation models is crucial for ensuring equitable and unbiased official statistics. Efforts to improve the interpretability and fairness of foundation models are essential. Investing in research and development for domain-specific pre-trained models tailored to official statistics can also help address some of the limitations.

To address these challenges effectively, statistical offices should strike a balance between using pre-trained models and maintaining control over the process. Collaborations with external providers should involve transparent agreements on data sharing, data privacy, and explainability. Prioritizing Responsible AI and MLOps principles like transparency, reproducibility, and accountability is essential to ensure the ethical implementation of AI in official statistics. Adopting open-source frameworks and

platforms allows for better scrutiny and adaptation to specific needs while ensuring adherence to best practices.

Conclusions

The implementation of AI in official statistics presents both exciting opportunities and critical challenges. The paradigm shift brought by foundation models has democratized access to advanced AI capabilities, accelerating innovation in natural language processing tasks and other AI applications. However, as we venture into this new frontier of AI, it is crucial to address the ethical implications and challenges that arise. Fairness, transparency, explainability, and data privacy must be prioritized to ensure AI-driven statistical outputs are reliable, equitable, and unbiased. The needs of official statistics can be more specific or differ from what the developers of AI models have in mind. It remains a challenge to adopt what works for official statistics and discard the rest. It is certain however, that high speed of development of AI will continue to shape how we work and what is expected from us.

In the future, the responsible and ethical implementation of AI in official statistics will continue to be a priority. As we navigate the complexities and challenges, we must remain committed to upholding the highest ethical standards, fostering public trust, and contributing to evidence-based decision-making. By leveraging the power of AI responsibly with MLOps, we can shape a brighter and more AI-driven future for official statistics.

Keywords: MLOps, transparency, Artificial Intelligence, Machine Learning

References

Kreuzberger, D., Kühl, N., Hirschl, S. (2023) *Machine Learning Operations (MLOps): Overview, Definition, and architecture.* IEEE Xplore Full-Text PDF:

Google Cloud MLOps. MLOps: Continuous delivery and automation pipelines in machine learning | Cloud Architecture Center | Google Cloud

ml-ops.org ML Ops: Machine Learning Operations (ml-ops.org)

Enhancing natural resource monitoring with Luke NatureWatch

Jonne Pohjankukka[1]

[1] Natural Resources Institute Finland (Luke), Finland
e-mail: jonne.pohjankukka@luke.fi

Abstract

In the era of digital big data, the potential for utilizing artificial intelligence (AI) to automatically extract meaningful insights from visual content has become increasingly evident. The manual processing of vast repositories of digital imagery and video data by researchers often entails an arduous and resource-intensive work. In Luke, researchers conduct periodic manual procedures of collecting and processing of digital data for monitoring the state of Finnish natural resources. These processes consist, for example, the identification of animal species in game data or determining the age of trees by counting of growth rings found in images of trees' cross-sections. There is a growing interest towards AI-based solutions to enhance manual data processing by automating repetitive and monotonic tasks, thus increasing human resources to more demanding tasks. Multitude of various AI-solutions can currently be found on the market, both open-source and commercial, many of which use transfer learning based pretrained models. However, the usage of these solutions can impose limitations depending on the used platform and off-the-self AI-models, while capable, are not necessarily always suitable per se for custom applications. Generally, for an AI-based solution to be implemented and deployed in practice, necessary steps to be taken include building pipelines for data collection and annotation, training and validating the AI-model, and finally deploying and monitoring the model in a real-world application. Also, a common bottleneck in custom AI-applications is the lack of sufficient quantities of training data.

In Luke NatureWatch cloud application (see Figure 1), we provide users easy-to-use pipelines and interfaces for digital image or video data upload, crowdsourcing annotation work via citizen science, creating and monitoring custom computer vision models and AI-generated exportable analytics reports for user provided data. The application's main rationale is to increase efficiency by automating laborious and repetitive data processing tasks and offer a platform for continuous improvement of the underlying AI-models.

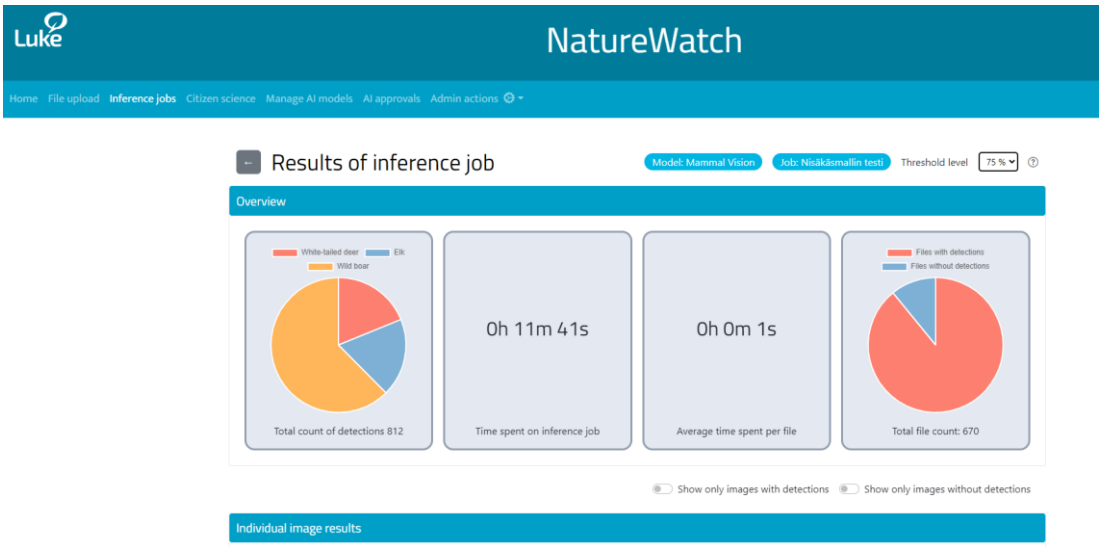


Figure 1: Screenshot of an inference job in Luke NatureWatch application

Keywords: natural resource monitoring, computer vision, automation, cloud computing, citizen science

4 CONTRIBUTED PAPERS

SAMPLING STRATEGIES FOR PROBABILITY SAMPLES USED TOGETHER WITH NON-PROBABILITY DATA

Lyndon Ang[1], Robert Clark[2], Bronwyn Loong[3], and Anders Holmberg[4]

[1] Australian National University, Australia
e-mail: lyndon.ang@anu.edu.au

[2] Australian National University, Australia
e-mail: robert.clark@anu.edu.au

[3] Australian National University, Australia
e-mail: bronwyn.loong@anu.edu.au

[4] Australian Bureau of Statistics, Australia
e-mail: anders.holmberg@abs.gov.au

Abstract

There is a growing trend among statistical agencies to explore alternative data sources for producing more timely and more detailed statistics, while reducing costs and respondent burden. These data sources may include administrative records, big data (or "found data") such as bank transactions or supermarket scanner data, and non-probability surveys such as online web panels. Coverage and measurement error are two issues that may be present in these types of data. These errors may be corrected using available auxiliary information relating to the population of interest, such as from a census or a reference probability sample.

In this paper, we discuss considerations for how a reference probability sample should be designed for the purpose of treating an imperfect data source. We consider in particular the case where the imperfect data relate to businesses. The multiple frame and cut-off sampling frameworks are explored in this context as alternatives to the usual optimal allocation for a single frame sample. A simulation study is conducted to examine the performance of various estimators under these frameworks.

Keywords: Non-probability sampling, Sample design, Multiple Frame, Cut-off Sampling

References

Beaumont, F.-F. (2020) Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46 (1), 1-28.

Lohr, S. L. (2021). Multiple-frame surveys for a multiple-data-source world. *Survey Methodology*, 47 (2), 229–263

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48 (2), 283–311

THE PROBLEM OF DATA COMPARABILITY IN THE POLISH TRANSITION PERIOD OF 1990-2004 AND A PROPOSAL FOR ITS SOLUTION

Lucyna Błażejczyk-Majka[1]

[1] Adam Mickiewicz University in Poznań, Faculty of History, Department of Economic History, Poland
e-mail: majkal@amu.edu.pl

Abstract

The issue of systemic transformation is of great interest to researchers. It is a period in the history of Poland in which a comprehensive reconstruction of the social and economic system was carried out. The Polish society decided to transition from an authoritarian system and centrally controlled economy to a democratic system and market economy (Ratajczak, 2009). The usual assumption is that the systemic transformation in Poland covered the years 1990-2004 (Swadźba, 2005, p. 14; Kaliński, 2009, p. 9).

However, the greatest difficulties in conducting quantitative research for the period of transition arise from the administrative reform in effect since January 1999, which introduced, among other things, a new administrative division of Poland[1]. The country's territory, which previously had 49 provinces, was divided into 16 voivodships. The only administrative units whose boundaries did not change were municipalities. For this reason, most studies referring to the period of transformation concern the entire country or are limited by the time scope of the administrative reforms in force, despite the fact that transformation processes do not usually proceed uniformly over the entire area of a given country (Narkiewicz, 1998; Mync and Komornicki, 2000; Korenik, 2003; Nowińska-Łaźniewska, 2004, chap. Zróżnicowanie rozwoju polskich regionów w okresie transformacji w latach 1990-2001 [Differentiation of the development of Polish regions in the period of transformation in 1990-2001]; Kaliński, 2009, figs 4-20; Sowiński, 2009; Włodarczyk and Nowak, 2011). The direction and pace of the transformation processes are the results of a region's location, its type, the density of residence, different levels of the involvement of human capital and infrastructure resources, as well as cultural or historical conditions of individual provinces or regions (Hryniewicz, 1998; Gorzelak, 1999; Bałtowski and Miszewski, 2015, figs 38-40).

While it is relatively easy to cope with the difficulties arising from inflation or methodological type of changes, the above-described problems arising from administrative reforms significantly limit the research possibilities. Of course, it is possible to use data from the Central Statistical Office and agglomerate the data of a given year at the level of municipalities and, on this basis, construct various types of time series (Stanny, 2013, chap. Zasoby lokalne jako czynniki rowoju obszarów wiejskich w Polsce [Local resources as factors of rural development in Poland] pp. 213-261; Smętkowski and Płoszaj, 2016). However, this is extremely time-consuming and requires, each time, a detailed analysis of the lists of municipalities that make up each voivodeship or region under analysis.

From this perspective, it seemed necessary to develop a quantitative tool that would provide reliable estimates of variables estimated for the entire transition period that is relatively quickly and easily obtainable. This tool would allow the transformation of available quantitative data published by the CSO from a system of 49 provinces into a system of 16 voivodeships. The starting point for its development was the area of the municipalities comprising a given voivodeship. The choice of the area

[1] Journal of Laws 1998 No. 96 pos. 603, https://isap.sejm.gov.pl/isap.nsf/, 26.07.2022.

of municipalities as the basis for the transformation by virtue of the fact that this was the same structure of municipalities that was preserved after the introduction of the administrative reform. Thus, the purpose of this paper is to present the transformation of quantitative data from the transition period at the voivodeship level (49 voivodships to 16 voivodships) and assess its usefulness and reliability in socio-economic and demographic research on the transition period.

Keywords: Data integration, combining data from surveys and registers.

References

Bałtowski, M. and Miszewski, M. (2015) *Transformacja gospodarcza w Polsce*. Wyd. 1, 3. Warszawa: Wydawnictwo Naukowe PWN.

Gorzelak, G. (1999) 'Historia - transformacja - przyszłość. Szkic o strukturze polskiej przestrzeni i jej zmianach', in G. Gorzelak, S. Szczepański, Marek, and T. Zarycki (eds) *Rozwój, region, społeczeństwo. Z okzacji jubileuszu Profesora Bohdana Jałowieckiego*. Warszawa - Katowice: Wydawnictwa Europejskiego Instytutu Rozwoju Regionalnego i Lokalnego, pp. 25–40.

Hryniewicz, J. (1998) 'Wymiary rozwoju gospodarczego gmin', in G. Gorzelak and B. Jałowiecki (eds) *Koniunktura gospodarcza i mobilizacja społeczna w gminach*. Studia Reg. Warszawa: Wydawnictwa Europejskiego Instytutu Rozwoju Regionalnego i Lokalnego, pp. 58–80.

Kaliński, J. (2009) *Transformacja gospodarki polskiej w latach 1989-2004*. Szkoła Główna Handlowa w Warszawie - Oficyna Wydawnicza.

Korenik, S. (2003) *Dysproporcje w rozwoju regionów Polski - wybrane aspekty*. Edited by A.E. im. O.L. (Wrocław). Wydawnictwo. Wrocław: Wrocław : Wydaw. Akademii Ekonomicznej im. Oskara Langego.

Mync, A. and Komornicki, T. (2000) 'Regionalne zróżnicowania procesów rozwoju społeczno-gospodarczego kraju w okresie transformacji', *Ekonomista*, 5, pp. 669–688.

Narkiewicz, J. (1998) 'Regionalne zróżnicowanie rozwoju społeczno-gospodarczego Polski', *Wiadomości Statystyczne*, 10(449), pp. 30–42.

Nowińska-Łaźniewska, E. (2004) *Relacje przestrzenne w Polsce w okresie transformacji w świetle teorii rozwoju regionalnego*. Prace Habi. Poznań: Akademia Ekonomiczna w Poznaniu.

Ratajczak, M. (2009) 'Transformacja ustrojowa w świetle ustaleń i założeń ekonomii instytucjonalnej', *Ruch Prawniczy, Ekonomiczny i Socjologiczny*, LXXI(2), pp. 233–251.

Smętkowski, M. and Płoszaj, A. (2016) 'Zróżnicownia społeczno-gospodarcze gmin według regionów historycznych', in G. Gorzelak (ed.) *Polska gmina* 2015. Warszawa: Wydawnictwo Naukowe Scholar, pp. 116–163.

Sowiński, T. (2009) *Przestrzenne zróżnicowanie jakości kapitału ludzkiego w Polsce w latach 1988-2006*. Opole: Wydawnictwo Uniwersytetu Opolskiego (Studia i Monografie / Uniwersytet Opolski: nr 437).

Stanny, M. (2013) *Przestrzenne zróżnicowanie rozwoju obszarów wiejskich w Polsce*. Warszawa: Instytut Rozwoju Wsi i Rolnictwa Polskiej Akademii Nauk.

Swadźba, S. (2005) 'Transformacja systemowa polskiej gospodarki i jej efekty', in H. Ćwikliński (ed.) *Transformacja polskiej gospodarki: ocena kierunków i dynamiki zmian strukturalnych*. Gdańsk: Fundacja Rozwoju Uniwersytetu Gdańskiego, pp. 13–20.

Włodarczyk, J. and Nowak, A. (2011) 'Zróżnicowanie regionalne poziomu życia ludności w świetle wybranych wskaźników z badań statystyki publicznej', in K. Leszczewska and J. Truszkowska (eds) *Uwarunkowania różnic społeczno-ekonomicznych*. Łomża: Państwowa Wyższa Szkoła Informatyki i Przedsiębiorczości, pp. 11–32.

BAYESIAN MODELLING OF PROFITABLE LANDING PAGE

Yana Bondarenko

Oles Honchar Dnipro National University, Ukraine
e-mail: yana.bondarenko@pm.me

Abstract

Each company applies great efforts on finding ways to make a profit. An effective website is important for the company because a significant part of the target audience receives information about the products and services via the Internet. Company website must be constantly optimized to bring in a profit. A/B testing is an experiment of showing two versions of the landing page to website visitors at the same time and comparing which one performs better for a given conversion aim. Our goal is to compute likelihood that the average income in version B is greater than the average income in version A.

Keywords: Bayesian inference, landing page optimization, likelihood, probability density function.

Let us assume that Bernoulli trials with probabilities of success λ_A, λ_B are conducted in groups of visitors in two versions of landing page. Probabilities of success λ_A, λ_B are unknown random variables. Note that success is a purchase. Suppose $p(\lambda_A)$, $p(\lambda_B)$ are the prior probability density functions for λ_A, λ_B, then $p(\lambda_A \mid x_1,...,x_n)$, $p(\lambda_B \mid y_1,...,y_m)$ are the posterior probability density functions for λ_A, λ_B. Let us remark that sample vectors $x=(x_1,...,x_n)$, $y=(y_1,...,y_m)$ are observed in two versions of landing page during the experiment.

In addition, suppose $f(\theta_A)$, $f(\theta_B)$ are the prior probability density functions for parameters θ_A, θ_B. Note also that these parameters θ_A, θ_B are some functions of the average size of purchase. Then $f(\theta_A \mid z_1,...,z_k)$, $f(\theta_B \mid w_1,...,w_l)$ are the posterior probability density functions for θ_A, θ_B. We stress that sample vectors $z=(z_1,...,z_k)$, $w=(w_1,...,w_l)$ are observed in two versions during the experiment.

Posterior distribution of probability of purchase in version A. Bernoulli trials with two possible outcomes (success is purchase, failure is no purchase) are conducted in version A. The number of successes (that is the number of purchases) in one trial has Bernoulli distribution with parameter λ_A (probability of purchase):

$$P(x,\lambda_A)=\lambda_A^x(1-\lambda_A)^{1-x},\ x=0,1;\ 0<\lambda_A<1. \tag{1}$$

The likelihood function is determined by:

$$p(x_1,\ldots,x_n|\lambda_A)=\lambda_A^{x_1}(1-\lambda_A)^{1-x_1}\ldots\lambda_A^{x_n}(1-\lambda_A)^{1-x_n}=\lambda_A^{\sum_{i=1}^n x_i}(1-\lambda_A)^{n-\sum_{i=1}^n x_i}, \tag{2}$$

where each x_i is equal to 1 or 0. The prior information about probability of purchase λ_A is defined by Beta distribution with parameters $a=1$, $b=1$ (equivalently, Uniform prior distribution as is generally known):

$$p(\lambda_A)=\frac{\lambda_A^{a-1}(1-\lambda_A)^{b-1}}{B(a,b)},\ 0\le\lambda_A\le 1. \tag{3}$$

According to Bayes' theorem, the posterior distribution for probability of purchase λ_A is given by:

$$p(\lambda_A|x_1,\ldots,x_n)=\frac{p(\lambda_A)\,p(x_1,\ldots,x_n|\lambda_A)}{\int_0^1 p(\lambda_A)\,p(x_1,\ldots,x_n|\lambda_A)\,d\lambda_A}. \tag{4}$$

Substituting $p(\lambda_A), p(x_1,\ldots,x_n|\lambda_A)$ in (4), we get:

$$p(\lambda_A|x_1,\ldots,x_n)=\frac{\lambda_A^{a-1}(1-\lambda_A)^{b-1}\lambda_A^{\sum_{i=1}^n x_i}(1-\lambda_A)^{n-\sum_{i=1}^n x_i}}{\int_0^1 \lambda_A^{a-1}(1-\lambda_A)^{b-1}\lambda_A^{\sum_{i=1}^n x_i}(1-\lambda_A)^{n-\sum_{i=1}^n x_i}\,d\lambda_A}. \tag{5}$$

Integrating (5) in λ_A, we obtain:

$$p(\lambda_A|x_1,\ldots,x_n)=\frac{\lambda_A^{a+\sum_{i=1}^n x_i-1}(1-\lambda_A)^{b+n-\sum_{i=1}^n x_i-1}}{B\left(a+\sum_{i=1}^n x_i, b+n-\sum_{i=1}^n x_i\right)}. \tag{6}$$

Thus, we have that the posterior distribution for probability of purchase λ_A is Beta distribution with parameters $(\tilde{a},\tilde{b})$:

$$p(\lambda_A|x_1,\ldots,x_n)=\frac{\lambda_A^{\tilde{a}-1}(1-\lambda_A)^{\tilde{b}-1}}{B(\tilde{a},\tilde{b})},\quad \tilde{a}=a+\sum_{i=1}^n x_i,\quad \tilde{b}=b+n-\sum_{i=1}^n x_i, \tag{7}$$

where n is the number of visitors and $\sum_{i=1}^n x_i$ is the number of complete purchases in version A.

Posterior distribution of probability of purchase in version B. Bernoulli trials with two possible outcomes (success, failure) are conducted in version B. The number of successes (number of purchases) in one trial has Bernoulli distribution with parameter λ_B (probability of purchase):

$$P(x,\lambda_B)=\lambda_B^y(1-\lambda_B)^{1-y},\ y=0,1;\ 0<\lambda_B<1. \tag{8}$$

The likelihood function is determined by:

$$p(y_1,\ldots,y_m|\lambda_B)=\lambda_B^{y_1}(1-\lambda_B)^{1-y_1}\ldots\lambda_B^{y_m}(1-\lambda_B)^{1-y_m}=\lambda_B^{\sum_{i=1}^m y_i}(1-\lambda_B)^{m-\sum_{i=1}^m y_i}, \tag{9}$$

where each y_i is equal to 1 or 0. The prior information about probability of purchase λ_B is defined by Beta distribution with parameters $c=1,\ d=1$ (Uniform prior distribution):

$$p(\lambda_B)=\frac{\lambda_B^{c-1}(1-\lambda_B)^{d-1}}{B(c,d)},\ 0\le\lambda_B\le 1. \tag{10}$$

According to Bayes' theorem, the posterior distribution for probability of purchase λ_B is given by:

$$p(\lambda_B|y_1,\ldots,y_m)=\frac{p(\lambda_B)\,p(y_1,\ldots,y_m|\lambda_B)}{\int_0^1 p(\lambda_B)\,p(y_1,\ldots,y_m|\lambda_B)\,d\lambda_B}. \tag{11}$$

Using (9), (10), we get:

$$p(\lambda_B|y_1,\ldots,y_m)=\frac{\lambda_B^{c-1}(1-\lambda_B)^{d-1}\lambda_B^{\sum_{i=1}^m y_i}(1-\lambda_B)^{m-\sum_{i=1}^m y_i}}{\int_0^1 \lambda_B^{c-1}(1-\lambda_B)^{d-1}\lambda_B^{\sum_{i=1}^m y_i}(1-\lambda_B)^{m-\sum_{i=1}^m y_i}\,d\lambda_B}. \tag{12}$$

Finally, we obtain:

$$p(\lambda_B|y_1,\ldots,y_m)=\frac{\lambda_B^{c+\sum_{i=1}^m y_i-1}(1-\lambda_B)^{d+m-\sum_{i=1}^m y_i-1}}{B\left(c+\sum_{i=1}^m y_i, d+m-\sum_{i=1}^m y_i\right)}. \tag{13}$$

This yields that the posterior distribution for probability of purchase λ_B is Beta distribution with parameters $(\tilde{c},\tilde{d})$

$$p(\lambda_B|y_1,\ldots,y_m)=\frac{\lambda_B^{\tilde{c}-1}(1-\lambda_B)^{\tilde{d}-1}}{B(\tilde{c},\tilde{d})},\quad \tilde{c}=c+\sum_{i=1}^{m}y_i,\quad \tilde{d}=d+m-\sum_{i=1}^{m}y_i, \tag{14}$$

where m is the number of visitors and $\sum_{i=1}^{m}y_i$ is the number of complete purchases in version B.

Posterior distribution of the size of purchase in version A. Now assume that the size of purchase has exponential distribution with parameter θ_A (that is reciprocal of the average size of purchase):

$$p(z,\theta_A)=\theta_A e^{-\theta_A z},\ z\geq 0,\ \theta_A>0. \tag{15}$$

The likelihood function is determined by:

$$p(z_1,\ldots,z_k|\theta_A)=\theta_A e^{-\theta_A z_1}\ldots\theta_A e^{-\theta_A z_k}=(\theta_A)^k e^{-\theta_A\sum_{i=1}^{k}z_i},\ z_i>0 \text{ for all } i. \tag{16}$$

The prior information about parameter θ_A is defined by Gamma distribution with parameters (s,t):

$$f(\theta_A)=\frac{t^s}{\Gamma(s)}\theta_A^{s-1}e^{-t\theta_A},\ \theta_A\geq 0, s>0, t>0. \tag{17}$$

According to Bayes' theorem, the posterior distribution for parameter θ_A is given by:

$$f(\theta_A|z_1,\ldots,z_k)=\frac{f(\theta_A)p(z_1,\ldots,z_k|\theta_A)}{\int_0^{+\infty}f(\theta_A)p(z_1,\ldots,z_k|\theta_A)d\theta_A}. \tag{18}$$

Substituting $f(\theta_A), p(z_1,\ldots,z_k|\theta_A)$ in (18), we get:

$$f(\theta_A|z_1,\ldots,z_k)=\frac{\theta_A^{s-1}e^{-t\theta_A}(\theta_A)^k e^{-\theta_A\sum_{i=1}^{k}z_i}}{\int_0^{+\infty}\theta_A^{s-1}e^{-t\theta_A}(\theta_A)^k e^{-\theta_A\sum_{i=1}^{k}z_i}d\theta_A}. \tag{19}$$

Integrating (19) in θ_A, we obtain:

$$f(\theta_A|z_1,\ldots,z_k)=\frac{\left(t+\sum_{i=1}^{k}z_i\right)^{s+k}}{\Gamma(s+k)}\theta_A^{s+k-1}e^{-\left(t+\sum_{i=1}^{k}z_i\right)\theta_A}. \tag{20}$$

Therefore, the posterior distribution for parameter θ_A is Gamma distribution with parameters $(\tilde{s},\tilde{t})$:

$$f(\theta_A|z_1,\ldots,z_k)=\frac{\tilde{t}^{\tilde{s}}}{\Gamma(\tilde{s})}\theta_A^{\tilde{s}-1}e^{-\tilde{t}\theta_A},\ \tilde{s}=s+k,\ \tilde{t}=t+\sum_{i=1}^{k}z_i, \tag{21}$$

where k is the number of purchases and $\sum_{i=1}^{k}z_i$ is the sum of complete purchases in version A.

Posterior distribution of the size of purchase in version B. Suppose the size of purchase has exponential distribution with parameter θ_B (that is reciprocal of the average size of purchase):

$$p(w,\theta_B)=\theta_B e^{-\theta_B w},\ w\geq 0,\ \theta_B>0. \tag{22}$$

The likelihood function is determined by:

$$p(w_1,\ldots,w_l|\theta_B)=\theta_B e^{-\theta_B w_1}\ldots\theta_B e^{-\theta_B w_k}=(\theta_B)^l e^{-\theta_B\sum_{i=1}^{l}w_i},\ w_i>0 \text{ for all } i. \tag{23}$$

The prior information about parameter θ_B is defined by Gamma distribution with parameters (u,v):

$$f(\theta_b)=\frac{v^u}{\Gamma(u)}\theta_B^{u-1}e^{-v\theta_B},\ \theta_B\geq 0, u>0, v>0. \tag{24}$$

According to Bayes' theorem, the posterior distribution for parameter θ_B is given by:

$$f(\theta_B|w_1,\ldots,w_l)=\frac{f(\theta_B)\,p(w_1,\ldots,w_l|\theta_B)}{\int_0^{+\infty} f(\theta_B)\,p(w_1,\ldots,w_l|\theta_B)\,d\theta_B}. \tag{25}$$

Using (22), (23), we get:

$$f(\theta_B|w_1,\ldots,w_l)=\frac{\theta_B^{u-1}e^{-v\theta_B}(\theta_B)^l e^{-\theta_B\sum_{i=1}^l w_i}}{\int_0^{+\infty}\theta_B^{u-1}e^{-v\theta_B}(\theta_B)^l e^{-\theta_B\sum_{i=1}^l w_i}d\theta_B}. \tag{26}$$

Finally, we obtain:

$$f(\theta_B|w_1,\ldots,w_l)=\frac{\left(v+\sum_{i=1}^l w_i\right)^{u+l}}{\Gamma(u+l)}\theta_B^{u+l-1}e^{-\left(v+\sum_{i=1}^l w_i\right)\theta_B}. \tag{27}$$

Hence, the posterior distribution for parameter θ_B is Gamma distribution with parameters $(\tilde{u},\tilde{v})$:

$$f(\theta_B|w_1,\ldots,w_l)=\frac{\tilde{v}^{\tilde{u}}}{\Gamma(\tilde{u})}\theta_B^{\tilde{u}-1}e^{-\tilde{v}\theta_B},\ \tilde{u}=u+l,\ \tilde{v}=v+\sum_{i=1}^l w_i, \tag{28}$$

where l is number of purchases and $\sum_{i=1}^l w_i$ is sum of accomplished purchases in landing page B.

Likelihood. The foregoing results allows us to describe probability of purchase λ and reciprocal of θ It is clear that λ_A is *the average number of complete purchases* in one trial in version A, λ_B is *the average number of complete purchases* in one trial in version B. Continuing in the same way, we see that $1/\theta_A$ is *the average size of complete purchases* in version A, $1/\theta_B$ is *the average size of complete purchases* in version B. Finally, *the average income* in version A is equal to λ_A/θ_A and *the average income* in version B is equal to λ_B/θ_B. Likelihood that the average income in version B is greater than the average income in version A is defined by:

$$P\left\{\frac{\lambda_B}{\theta_B}>\frac{\lambda_A}{\theta_A}\right\}=P\left\{\frac{\lambda_B}{\theta_B}-\frac{\lambda_A}{\theta_A}>0\right\}=$$

$$=\int_{\frac{\lambda_B}{\theta_B}-\frac{\lambda_A}{\theta_A}>0} p(\lambda_A,\theta_A,\lambda_B,\theta_B\,|\,x_1,\ldots,x_n;z_1,\ldots,z_k;y_1,\ldots,y_m;w_1,\ldots,w_l)\,d\lambda_A d\theta_A d\lambda_B d\theta_B=$$

$$=\int_{\frac{\lambda_B}{\theta_B}-\frac{\lambda_A}{\theta_A}>0} p(\lambda_A\,|\,x_1,\ldots,x_n)f(\theta_A\,|\,z_1,\ldots,z_k)p(\lambda_B\,|\,y_1,\ldots,y_m)f(\theta_B\,|\,w_1,\ldots,w_l)d\lambda_A d\theta_A d\lambda_B d\theta_B=$$

$$=\int_{\frac{\lambda_B}{\theta_B}-\frac{\lambda_A}{\theta_A}>0}\frac{\lambda_A^{\tilde{a}-1}(1-\lambda_A)^{\tilde{b}-1}}{B(\tilde{a},\tilde{b})}\frac{\tilde{t}^{\tilde{s}}}{\Gamma(\tilde{s})}\theta_A^{\tilde{s}-1}e^{-\tilde{t}\theta_A}\frac{\lambda_B^{\tilde{c}-1}(1-\lambda_B)^{\tilde{d}-1}}{B(\tilde{c},\tilde{d})}\frac{\tilde{v}^{\tilde{u}}}{\Gamma(\tilde{u})}\theta_B^{\tilde{u}-1}e^{-\tilde{v}\theta_B}d\lambda_A d\theta_A d\lambda_B d\theta_B.$$

Likelihood can be approximated with Monte Carlo sampling. Many computer languages have procedures for simulating this sampling process.

References

Bondarenko, Ya. (2019) Sequential A/B Testing. In: *Proceedings of the 2019 IEEE International Conference on Advanced Trends in Information Theory*, 488-491. https://ieeexplore.ieee.org/document/9030448

Bondarenko, Ya., Kravchenko, S. (2019) Bayesian approach to landing page testing. *Problems of Applied Mathematics and Mathematical Modelling*, 3-16. https://pm-mm.dp.ua/index.php/pmmm/article/view/240

Bondarenko, Ya. (2021) A sequential probability ratio test for randomized online experiments. *Problems of Applied Mathematics and Mathematical Modelling*, 3-15. https://pm-mm.dp.ua/index.php/pmmm/article/view/304

Stucchio, C. (2015) Bayesian A/B Testing at VWO. Whitepaper, Visual Website Optimizer.

INTEGRATION OF A VOLUNTARY SAMPLE ASSUMING THE NOT MISSING AT RANDOM RESPONSE MECHANISM

I. Burakauskaitė[1,2] **and A. Čiginas**[1,2]

[1] State Data Agency (Statistics Lithuania), Lithuania

[2] Vilnius University, Lithuania

e-mails: ieva.burakauskaite@mif.stud.vu.lt; andrius.ciginas@mif.vu.lt

Abstract

We aim to effectively integrate a voluntary (non-probability) sample for the estimation of population parameters of the Statistical survey on population by ethnicity, native language and religion of the Lithuanian census 2021. We apply the propensity score adjustment to correct the non-probability sample selection bias. We use a parametric model in estimating the propensity scores (probabilities) to participate in the voluntary survey. The modeling of propensity scores is conducted in two ways: assuming the response mechanism to be missing at random, and assuming it to be not missing at random. The maximum likelihood based method is used to estimate the propensity scores in both cases. We compare the obtained estimators in a simulation study based on Lithuanian census data.

Keywords: data integration, non-probability sample, missing at random, not missing at random, propensity score adjustment.

NONPROBSVY – AN R PACKAGE FOR NON-PROBABILITY SAMPLES

Łukasz Chrostowski[1] and Maciej Beręsewicz[2]

[1] Poznan University of Economics and Business, Poland

[2] Poznan University of Economics and Business, Poland
e-mail: maciej.beresewicz@ue.poznan.pl

Abstract

The aim of the nonprobsvy R package is to perform statistical inference on non-probability survey samples (including big data) when auxiliary information from external sources, such as probability samples or population totals or means, is available.

It should be noted that there are several packages that allow correcting for selection bias in nonprobability samples, such as GJRM (Marra et al. 2017), NonProbEst (Rueda et al. 2020), or even sampling (Tillé and Matei 2021). However, these packages do not implement state-of-the-art approaches recently proposed in the literature: Chen et al. (2020), Yang et al. (2020), Wu (2022), nor do they use the survey package (Lumley 2004) for inference.

We have implemented propensity score weighting (e.g. with calibration constraints), mass imputation (e.g. predictive mean matching) and doubly robust estimators that take into account minimisation of the asymptotic bias of the population mean estimators, variable selection or overlap between random and non-random samples. The package uses the functionality of the survey package when a probability sample is available. During the presentation, the functionality of the package and examples will be presented.

The package is under development and can be found on https://github.com/ncn-foreigners/nonprobsvy/

Keywords: Data integration, Doubly robust estimation, Propensity score estimation.

References

Marra et al. (2017). A simultaneous equation approach to estimating HIV prevalence with nonignorable missing responses. JASA, 112(518), 484-496.

Rueda et al. (2020). The R package NonProbEst for estimation in non-probability surveys. The R J, 12(1), 406-418.

Tillé and Matei (2021) sampling: Survey Sampling

Lumley (2004) Analysis of complex survey samples. JSS 9(1): 1-19

Chen et al. (2020). Doubly robust inference with nonprobability survey samples. JASA, 115(532), 2011-2021.

Yang et al. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. JRSS B, 82(2), 445.

Wu (2022). Statistical inference with non-probability survey samples. SM 48(2), 283-311.

OPEN-ENDED QUESTIONS IN SURVEYS

A. Clarke[1], K. Lagus[2] and M. Valaste[3]

[1] University of Helsinki, Finland
e-mail: adeline.clarke@helsinki.fi

[2] University of Helsinki, Finland
e-mail: krista.lagus@helsinki.fi

[3] University of Helsinki, Finland
e-mail: maria.valaste@helsinki.fi

Abstract

The use of open-ended questions in survey research has a very long history and the value of open-ended questions has been rediscovered in survey research (Neuert et al., 2021);(Singer & Couper, 2017). Open-ended questions are an important but challenging way to obtain informative data in surveys. Open-ended question data usually requires extra time investment (Fielding et al., 2013), but open-ended questions are particularly useful if researchers do not want to constrain respondents' answers to pre-specified selections. Open-ended questions allow respondents to provide diverse answers based on their experience, and some answers are probably never thought of by researchers. (He & Schonlau, 2021.) In this paper we look at the use of open-ended responses as part of the analysis of survey data. To accompany our research, we have built an R package which will enable researchers to more easily analyse open-end survey responses in Finnish. The package contributes to the growing suite of tools available for analysing Finnish text data. Our work is part of a national infrastructure project (https://www.dariah.fi/).

Keywords: open-ended questions.

References

Neuert, C. E., Meitinger, K., Behr, D., & Schonlau, M. (2021). Editorial: The Use of Open-ended Questions in Surveys. *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (Mda)*, 15(1), 3–6.

Singer, E., & Couper, M. P. (2017). Some Methodological Uses of Responses to Open Questions and Other Verbatim Comments in Quantitative Surveys. *Methods, Data, Analyses*, 11(2), 115–134. https://doi.org/10.12758/MDA.2017.01.

Fielding, J., Fielding, N., & Hughes, G. (2013). Opening up open-ended survey data using qualitative software. *Quality & Quantity*, 47(6), 3261–3276. https://doi.org/10.1007/s11135-012-9716-1.

He, Z., & Schonlau, M. (2021). Coding Text Answers to Open-ended Questions: Human Coders and Statistical Learning Algorithms Make Similar Mistakes. *Methods, Data, Analyses*, 15(1), Article 1. https://doi.org/10.12758/mda.2020.10.

MULTI-ARMED BANDIT POLICY UNDER DELAYS FOR THE DESIGN OF CLINICAL TRIALS

A. Dzhoha[1] **and I. Rozora**[2,3]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: andrew.djoga@gmail.com

[2] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: irozora@knu.ua

[3] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine

Abstract

Randomized controlled trials are currently considered to be the gold standard method to evaluate the effectiveness of new drugs or medical procedures. Most trials use a fixed randomization method which does not take into account the individual well-being of patients. To conduct clinical trials with the most health benefits for patients, the collected data can be used dynamically to reassign the groups to give more participants a chance for better care during trials. Such an adaptive design is a great example of using the exploration-exploitation trade-off approach. Thompson (1933) introduced the multi-armed bandit problem for this purpose.

The multi-armed bandit problem is well suited to model sequential resource allocation in the face of uncertainty. In setups like clinical trials, the response to an action is not immediate. Thus, the multi-armed bandit policies need adaptation to delays in order to retain their theoretical guarantees in a not strictly sequential environments.

By conducting simulations using the publicly available dataset The International Stroke Trial (Sandercock, Niewada, Członkowska, and the International Stroke Trial Collaborative Group 2011), we show the importance of the adaptation to delayed feedback. We study the impact on the results of experiments and provide asymptotic analysis. Thompson Sampling policy (Bubeck & Cesa-Bianchi 2012, p. 20) with Bernoulli rewards is considered the main baseline.

As another approach to mitigate the issue of delays, we propose to use the estimation of the effect of treatment as the reward feedback. We assume that such evidence of response to a drug can be collected in a relatively short-term period after the procedure and can be used to represent a certainty of successful treatment. For that, we analyze the Upper Confidence Bound policy (Bubeck & Cesa-Bianchi 2012, p. 11) with beta rewards. This policy has the potential to provide lower regret results which give more patients a chance for better care. Additionally, this policy can be corrected for Bernoulli reward or anticipated estimation error.

Keywords: multi-armed bandit, delayed feedback, Upper Confidence Bound policy.

References

Bubeck S., Cesa-Bianchi N. (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, **5**(1), 1–122.

Sandercock PA, Niewada M, Członkowska A, the International Stroke Trial Collaborative Group (2011) The International Stroke Trial database. *Trials*, **12**(1), 101.

Thompson, W. R. (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 285–294.

SMALL AREA PREDICTION FOR EXPONENTIAL DISPERSION FAMILIES UNDER INFORMATIVE SAMPLING

Abdulhakeem Eideh[1] and Emily Berg[2]

[1] Al-Quds University, Palestine
e-mail: msabdul@staff.alquds.edu

[2] Iowa State University, USA
e-mail: emilyb@iastate.edu

Abstract

In complex surveys, the inclusion probability is often correlated with the response variable after conditioning on model covariates, leading to an informative design. Small area estimates are usually constructed from complex survey data. If the design is informative for the model, then procedures that ignore the design can suffer from important biases. The obvious fact that ignoring the design can render erroneous small area predictions is widely noted in the literature (Pfeffermann & Sverchkov 2007, Eideh 2002, Verret et al. 2015, You & Rao 2002, Parker et al. 2019). Past work on small area estimation under informative sampling has focused heavily on linear models or on prediction of means. We propose to generalize existing small area procedures for an informative sample design.
Concomitantly, response variables often have non-normal distributions and require nonlinear models. We develop a small area procedure that addresses both of these issues simultaneously. We develop small area predictors for the broad class of exponential dispersion families under an informative design. This class of models encompasses linear models as well as nonlinear models. We develop predictors of general parameters that may be nonlinear functions of the model response variable. We study the properties of the procedure under two models for the survey weight. We evaluate the procedures through simulation using a logistic mixed model. We then apply the methods to construct small area estimates of several functions of a wetlands indicator using data from a large-scale survey called the National Resources Inventory. Wetlands are crucial for maintaining ecosystem health. We estimate several functions of a wetlands indicator in New Jersey counties. The National Resources Inventory uses a complex design and the association between the weight and the response variable is significant.

Keywords: informative sampling, mean weight model, small area estimation.

References

Cho, Y., Guadarrama, M., Eideh, A., Molina, I. & Berg, E. (2023) Optimal predictors of nonlinear parameters under informative sampling, Manuscript under review by the Journal of Survey Statistics and Methodology .

Eideh, A. H. (2002) Estimation for longitudinal survey data under informative sampling. Unpublished Ph.D., Hebrew University of Jerusalem.

Kim, J. K. & Wang, H. (2023) A note on weight smoothing in survey sampling, Survey Methodology (accepted)

Parker, P. A., Janicki, R. & Holan, S. H. (2019) Unit level modeling of survey data for small area estimation under informative sampling: A comprehensive overview with extensions, arXiv preprint arXiv:1908.10488 .

Pfeffermann, D. & Sverchkov, M. (2007), 'Small-area estimation under informative probability sampling of areas and within the selected areas', Journal of the American Statistical Association 102(480), 1427–1439.

Rao, J. N. K. & Molina, I. (2015) Small area estimation, John Wiley & Sons.

You, Y. & Rao, J. N. K. (2002) A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights, Canadian Journal of Statistics 30(3), 431–439.

Verret, F., Rao, J. N. K. & Hidiroglou, M. A. (2015) Model-based small area estimation under informative sampling', Survey Methodology 41(2), 333–348.

Data collection methods in Latvian Household Finance and Consumption Survey

Andris Fisenko[1]

[1] Bank of Latvia, e-mail: andris.fisenko@bank.lv

Abstract

The Household Finance and Consumption Survey (HFCS) is a statistical survey conducted in the euro area countries by collecting and compiling data on the real assets, financial assets, debt, income, and consumption of households. The HFCS is carried out by the European Central Bank and the national central banks of the European Union Member States. The HFCS is conducted at the national level. To obtain comparable data, the participating countries follow common methodological guidelines (Household Finance and Consumption Network 2019), but do not necessarily use identical questionnaires.

The Latvian HFCS for the third time was conducted in 2020 by the Bank of Latvia, again in a close cooperation with the Central Statistical Bureau of Latvia (CSB). CSB ensured the collection of the HFCS data by two types of interview mode. In 2020, the Computer Assisted Telephone Interview (CATI) method was introduced in the survey.

The recommended survey method by the ECB has been Computer Assisted Personal Interview (CAPI), because there are some questions where the interviewer evaluates the visual appearance of the dwelling being present on site and as well as questions where the answer is selected from the displayed card. However, 2020 brought changes throughout the world and changes in the survey. Most countries were forced to change the method of survey data collection. Thus, Latvia also switched to CATI. CSB already had extensive experience in conducting surveys using the CATI method and even established a special CATI service for this purpose. Therefore, a joint decision was made that for those addresses with available phone numbers, they will be surveyed using the CATI method.

In my work, I will examine whether there have been differences between these data collection methods and what the differences are in specific question groups. Part of the survey consists of administrative data; therefore, I will focus more on those questions that were not affected by the editing of administrative data.

These administrative data, as well as the comments and the paradata provided by interviewers at the conclusion of each interview, are used at the Bank of Latvia during the data editing phase to detect and correct possible mistakes in the survey data. Such quality checks aim to correct various kinds of inconsistencies, such as mistyped or erroneous answers, and it is possible to identify the quality differences between CAPI and CATI methods.

For the current survey wave, all editing has already been done. In my poster, I plan to show some results to demonstrate if there are any differences between these two methods for HFCS data.

Keywords: CAPI, CATI, survey.

References

The Household Finance and Consumption Survey: Methodological report for the 2021 wave. ECB Statistical Paper Series, 28

Selection of demographic variables in post-stratification

Mingmeng Geng[1] **and Roberto Trotta**[1]

[1] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Italy
e-mail: mgeng@sissa.it, rtrotta@sissa.it

Abstract

Post-stratification has been widely used in survey data and performs well in real data. Although it is implemented in different ways, some demographic variables such as gender, age, region, and education level are usually used for post-stratification weighting. This classic setup has been proven very effective in a variety of situations, and our aim in this work is to find some potentially better choice of demographic variables for post-stratification.

Based on multiple public and private data sets containing more demographic variables, we used individual-based machine learning models to predict people's opinions. Under this framework, the effects of different demographic variables on one person's perspective have been explored. Using different feature selection methods and prediction models, we found some possible better combinations of demographic variables for the individual predictions of different questions in the surveys, not only the ones often used before. As a result, more reasonable options for survey design and post-stratification become possible.

We also investigated the treatment and simplification of continuous variables such as age and income. For example, statistics on income are often noisy and we want to know how it affects the prediction model. In practice, these variables are often divided into groups, which facilitates the calculation of post-stratification but obscures the information for individual predictions. Therefore, we also discussed the trade-off between the two parts, for example, the division of age groups.

Keywords: feature selection, machine learning, survey design, continuous variables

VALUE OF INFORMATION IN THE PLANNING OF COST-EFFECTIVE OPERATIONAL FOREST INVENTORIES

Santeri Karppinen[1], Liviu Ene[2] and Juha Karvanen[1]

[1] Department of Mathematics and Statistics
University of Jyvaskyla, Finland
e-mail: juha.t.karvanen@jyu.fi

[2] Skogforsk, Sweden

Abstract

Pre-harvest inventories are used to estimate the volume of timber in a forest stand before the stand is scheduled for harvesting. In the inventory, the volume of timber is measured in randomly selected sample plots and the accuracy of the volume estimates essentially depends on the size and number of these plots. We consider a planning problem with two nested stages. In the inner stage, the posterior volumes are available and the task to schedule the stands for harvesting so that the monthly demand targets for timber can be fulfilled as closely as possible. In the outer stage, the decision on the inventory method is made for each stand. More accurate inventory methods lead to lower variance of the volume estimates but also cost more than less accurate methods. Value of information (Eidsvik et al., 2015) measures the gain from the lower variance. The problem is to assign an inventory method for each stand while making sure that the total cost of inventories is below the given budget limit.

We formulate the problem as a two-stage Bayesian decision problem (Raiffa & Schlaiffer, 1967) where the uncertain timber volumes and their observations are modelled using probability models. We cast the decision problem as a maximisation problem that seeks to maximise the value of information subject to a forest inventory budget constraint. Computing the value of information in our context is analytically intractable, since it requires the solution of an NP-hard binary optimisation problem within a high-dimensional integral. In particular, the binary optimisation problem is a special case of a generalised quadratic assignment problem (cf. Lee and Ma., 2004; Hahn et al., 2008). We present a practical method that solves the problem with an approximation to the value of information which combines Monte Carlo sampling with a greedy, randomised method for the binary optimisation problem.

We apply the developed method with realistic data obtained from Skogforsk, The Forestry Research Institute of Sweden. The data contain 100 forest stands for which some prior knowledge on the timber volume is available . For each stand, the choice of inventory method is made between three alternatives with different costs. We derive optimal inventory decisions for these stands across a range of inventory budgets.

Keywords: Bayesian model, Data collection, Decision making, Forestry

References

Eidsvik J., Mukerji T., Bhattacharjya D. (2015) *Value of information in the earth sciences: Integrating spatial modeling and decision analysis.* Cambridge University Press.

Raiffa H., Schlaiffer R. (1967) *Applied statistical decision theory.* Wiley, New York.

Lee, C.-G. and Ma, Z. (2004) *The generalized quadratic assignment problem.* Technical report, Department of Mechanical and Industrial Engineering, University of Toronto, Canada.

Hahn, P. M. and Kim, B.-J. and Guignard, M. and Smith, J. M. and Zhu, Y.-R. (2008) An algorithm for the generalized quadratic assignment problem. *Computational Optimization and Applications*, (40):351–372.

COMPARING COMBINATIONS OF ESTIMATOR AND SAMPLE ALLOCATION WHEN ESTIMATING POPULATION AND DOMAIN-SPECIFIC PROPORTIONS FOR A BINARY VARIABLE

Mauno Keto[1] and Risto Lehtonen[2]

[1] University of Jyväskylä, Finland
e-mail: maujohketo@gmail.com

[2] University of Helsinki, Finland
e-mail: risto.lehtonen@helsinki.fi

Abstract

In many surveys, both population-level and domain-level statistics are estimated for the target variables which can be continuous or discrete, and the objective is to obtain reliable estimates for each level. The domains may have very diverse sizes and other relevant characteristics. For this reason, it is important to plan stratified sampling and domain estimation carefully, so that the objective is possible to reach. Very small sample sizes are possible for some domains. In this situation, small area estimation, although its basic idea is to utilize information from other domains, does not necessarily produce high-quality estimates for every domain. Sometimes it may be reasonable to set limits to domain-specific sample sizes in order to obtain even moderate estimates for the domains. The concept of optimal allocation for domains depends on the situation. It is a solution of an optimization problem, but all relevant objectives can scarcely be reached. The selected estimator of the target variable may have a strong impact on estimation results, and the combination of sample allocation and estimator is worth studying also. It is possible to develop a model- and estimator-based allocation.

We use planned domains in our study. Our main interest is focused on estimating population and domain-specific proportions for a binary variable by using three model-assisted estimators based on a logistic regression model which use auxiliary data. We compute the domain-specific sample sizes according to six different allocation principles. Five of these allocations are developed by utilizing earlier collected proxy data and the logistic model. But we also test the performances of three other estimators: a direct Horvitz-Thompson estimator, a model-based EBLUP and a model-assisted GREG estimator. The last two estimators are based on the same regression model with random domain-specific effects. We use two allocations with these three estimators. The assessment of the performances of the estimators and allocations at the domain and population levels are based on design-based sample simulations. We measure the performances of the allocations and estimators with quality indicators. We introduce four different R-square measures to assess the suitability of the logistic models in estimation.

The estimators based on the logistic models outperform the design-based and model-related regression estimators, but the performances of different logistic models are close to each other. One allocation can be regarded as slightly more effective than the others. The predictive power of the logistic models can be regarded as moderate.

Keywords: Auxiliary and proxy data, model-assisted logistic regression, direct estimator, model-based and model-assisted regression estimator, performance, optimization, limitation of sample size, trade-off between domains and population, predictive power.

References

Allison, P. (2013). What's the Best R-Squared for Logistic Regression? *Statistical Horizons* (*https://statisticalhorizons.com/r2logistic*).

Cox, D.R. and E.J. Snell (1989) *Analysis of Binary Data*. Second Edition. Chapman & Hall.

Demidenko, E. (2008). Sample size and optimal design for logistic regression with binary interaction. *Statistics in Medicine* 27: 36–46.

Duchesne, P. (2003). Estimation of a Proportion with Survey Data. *Journal of Statistics Education* 11(3).

Gabler, S., Ganninger, M., and Münnich, R. (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika 75*: 15–161.

Keto, M., Hakanen, J., and Pahkinen, E. (2018). Register data in sample allocations for small-area estimation. An International Journal of Mathematical Demography 25, 184-214.

Lehtonen R., Särndal C.E., and Veijanen A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology 29*: 33–44.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition* 7: 649–673.

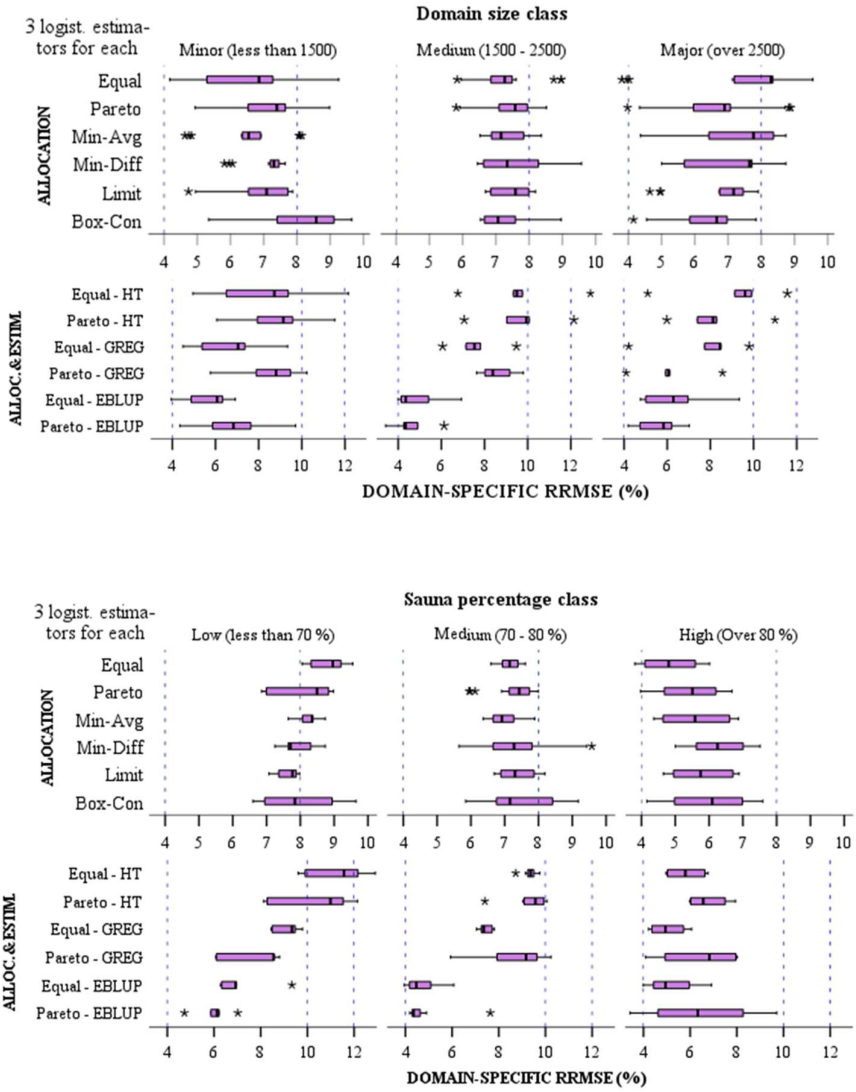
Lehtonen R. and Veijanen A. (2016). Model-assisted methods for Small Area Estimators of Poverty Indicators. In *Analysis of Poverty Data by Small Area Estimation, Pratesi M. (ed.). Wiley and Sons*: 109–127.

Miettinen, K. 1999. Nonlinear Multiobjective Optimization. Kluwer Academic Publishers, Boston.

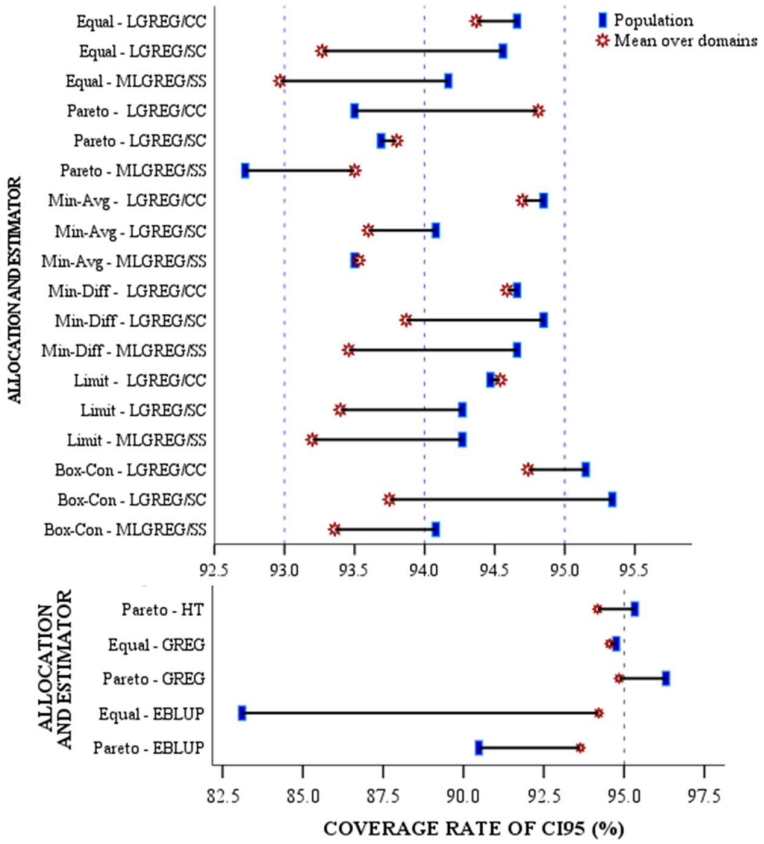
Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* 78: 691–692.

Rao, J. N. K. and Molina, I. 2015. Small Area Estimation (2nd Edition). Hoboken, NJ: John Wiley & Sons, Inc.

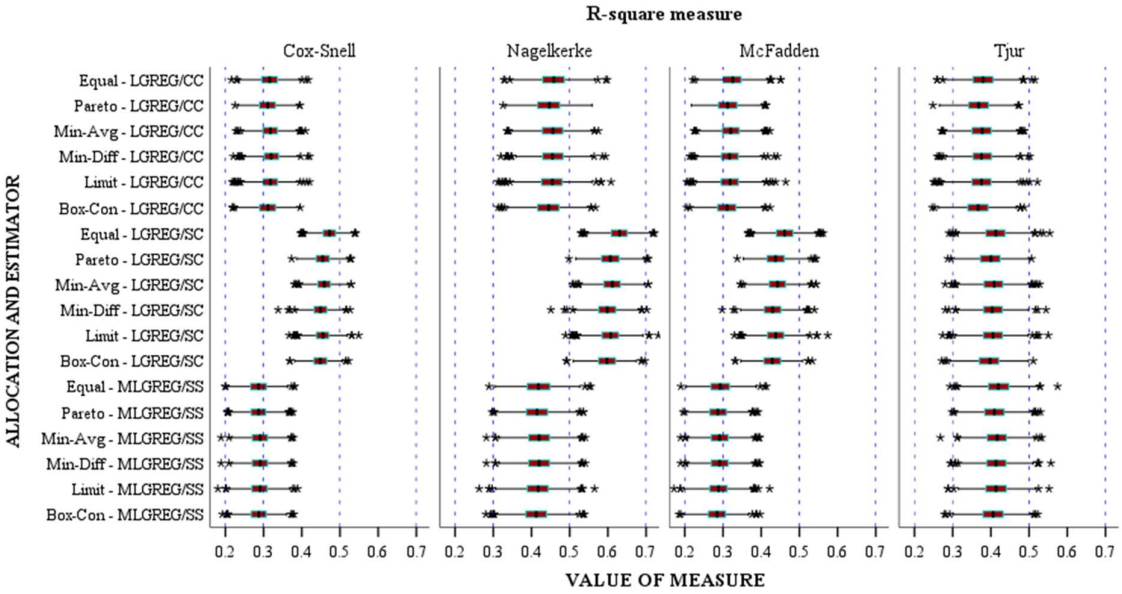
Figures describing results (accuracy RRMSE, CI95 coverage, and R-square measures)



CI95 coverage rates by allocation and estimator



Distributions of different R-square measures in samples by allocation and logistic estimator


R-square measure
Cox-Snell
Nagelkerke
McFadden
Tjur
ALLOCATION AND ESTIMATOR
Equal - LGREG/CC
Pareto - LGREG/CC
Min-Avg - LGREG/CC
Min-Diff - LGREG/CC
Limit - LGREG/CC
Box-Con - LGREG/CC
Equal - LGREG/SC
Pareto - LGREG/SC
Min-Avg - LGREG/SC
Min-Diff - LGREG/SC
Limit - LGREG/SC
Box-Con - LGREG/SC
Equal - MLGREG/SS
Pareto - MLGREG/SS
Min-Avg - MLGREG/SS
Min-Diff - MLGREG/SS
Limit - MLGREG/SS
Box-Con - MLGREG/SS
0.2 0.3 0.4 0.5 0.6 0.7
VALUE OF MEASURE

LIMIT THEOREMS FOR SEQUENCES OF RECORDS

O. Kolesnik
Igor Sikorsky Kyiv Polytechnic Institute, Ukraine
e-mail: lxndr.kolesnik@gmail.com

Abstract

Let $\{X_k, k \geq 1\}$ be a sequence of independent random variables, $\alpha = \{\alpha_k, k \geq 1\}$ be positive real numbers, and F be a continuous distribution function. Assume that the distributions of random variables X_k are such that $P(X_k < x) = (F(x))^{\alpha_k}$. Such a sequence of random variables is called F^{α}-scheme. Define the number of records $\mu(n)$ in the sequence $\{X_k\}$ up to moment n as follows

$$\mu(n) = \sum_{k=1}^{n} \mathbb{I}_k,$$

where $\mathbb{I}_1 = 1$, $\mathbb{I}_k = \mathbb{I}(X_k > \max(X_1, X_2, \ldots, X_{k-1}))$, $k \geq 2$

An important fact of the theory of records is the statistical independence of these indicators. It is also known that $\mathbb{P}(I_n = 1) = \dfrac{\alpha_n}{A_n}$, where $A_n = \sum\limits_{k=1}^{n} \alpha_k$. In some cases, it is possible to express the almost sure asymptotic behavior of $\{\mu(n)\}$ in terms of the sequence $\{A_n\}$. For example,

$$\lim_{n\to\infty} \frac{\mu(n)}{\ln A_n} \to C \quad \text{exists almost surely if} \quad \lim \frac{\alpha_n}{A_n} \quad \text{exists},$$

where C is a nonrandom constant that depends on $\lim \frac{\alpha_n}{A_n}$ (see P. Doukhan, O. I. Klesov, and J. G. Steinebach, 2015).

Some new asymptotic results will be presented in the talk. Below is one of them.

Theorem. Let $0 < p_1 < p_2 < \cdots < p_m < 1, m \geq 1$, be all the partial limits of the sequence $\dfrac{\alpha_n}{A_n}$ and $\Delta_i := \left(\dfrac{p_{i-1} + p_i}{2}, \dfrac{p_i + p_{i+1}}{2}\right)$, $i = \overline{1, m}$, where $p_0 := 0$ and $p_{m+1} := 1$. Assume that

$$\tau_i := \lim_{n\to\infty} \frac{\left|\left\{k \in \mathbb{N} : \ k < n, \dfrac{\alpha_k}{A_k} \in \Delta_i\right\}\right|}{n} \qquad \text{exists for all} \quad i = \overline{1, m}.$$

Then:

$$\frac{\mu(n)}{\ln(A_n)} \to -\frac{\sum\limits_{i=1}^{m} \tau_i p_i}{\sum\limits_{i=1}^{m} \tau_i \ln(1 - p_i)} \ a.s.$$

Keywords: records, F^{α}-scheme, limit theorems.

References

P. Doukhan, O. I. Klesov, and J. G. Steinebach (2015) Strong Laws of Large Numbers in an F^{α}-Scheme. In: *Mathematical Statistics and Limit Theorems, Festschrift in Honour of Paul Deheuvels*, (eds.: M. Hallin, D.M. Mason, D. Pfeifer, J.G. Steinebach), Springer International Publishing, Switzerland, 287–303.

ESTIMATING AVERAGE WAGES IN SMALL POPULATION DOMAINS

Enrika Komarovaitė[1,2] **and Andrius Čiginas**[1,2]

[1] Vilnius University, Lithuania
e-mail: enrika.komarovaite@mif.stud.vu.lt, e-mail: andrius.ciginas@mif.vu.lt

[2] State Data Agency, Lithuania
e-mail: enrika.komarovaite@stat.gov.lt, e-mail: andrius.ciginas@stat.gov.lt

Abstract

In the Statistical survey on the structure of earnings in Lithuania, the average wages are estimated for various domains of the population of employees working in enterprises. The survey sample is designed to ensure sufficient accuracy of estimates in all planned estimation domains. However, users of statistical information tend to ask for estimates in smaller unplanned domains, where domain sample sizes are often too small to obtain reliable results when using direct estimation methods. Therefore, we apply small area estimators based on domain-level models that use average wages derived from administrative data as covariates. Our empirical study shows that these estimators significantly improve the direct ones.

Keywords: small area estimation, auxiliary information, area-level model, composite estimation, average wages.

Multiple hypothesis testing for coronavirus disease in Ukraine

I. Kosareva[1] and R. Yamnenko[2]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: kosarevaivanna777@knu.ua

[2] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: rostyslav.yamnenko@knu.ua

Abstract

In this work, we will consider the data on coronavirus disease in Ukraine by region from the beginning to May 2023[https://index.minfin.com.ua/ua/reference/coronavirus/ukraine/]. The purpose of this study is to find out if the proportion of people who got sick and recovered is equal to 0.5 in each region. The data is arranged in a contingency table and multiple hypothesis testing is planned to be used for its analysis.

Multiple hypothesis testing is a statistical technique used to test multiple hypotheses simultaneously. When it comes to the analysis of contingency tables, multiple hypothesis testing can be used to compare the proportions of different categories across two or more groups.

Multiple hypothesis testing for contingency tables is an important topic in statistics. The need for accurate and efficient analysis of complex data, while avoiding false positives and erroneous conclusions, underscores the relevance of this topic.

The problem with testing multiple hypotheses simultaneously is that the likelihood of making a Type I error (rejecting a true null hypothesis) increases with the number of tests performed. This can lead to spurious or false positive results, which can be misleading and lead to incorrect conclusions. Multiple testing procedures for the contingency table are designed to control the overall error rate while still allowing for the detection of true signals in the data.

In this research, we use the chi-square test to calculate the p-value. The well-known formula for the chi-square statistic used in the chi square test is

$$\chi_c^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \quad (1)$$

where O_i is the observed value, E_i is the expected value, "i" is the "ith" position in the contingency table and c is the degrees of freedom.

The next step is to perform Monte-Carlo simulation on the data and calculate the p-value using the chi-squared statistic for each shuffle. As we test multiple hypotheses false positive rate has to be controlled with the false discovery rate(FDR) method.

After all these procedures we expect to obtain a lower p-value and get the output of the test about the proportion of people who were infected and recovered.

Keywords: multiple hypothesis testing, contingence table, Monte-Carlo simulation, chi-square test, coronavirus disease in Ukraine.

ESTIMATION OF THE SENSITIVE PROPORTION IN ITEM COUNT MODELS UNDER SOME ASSUMPTIONS VIOLATION

Barbara Kowalczyk[1] and Robert Wieczorkowski[2]

[1] SGH Warsaw School of Economics, Poland
e-mail: bkowal@sgh.waw.pl

[2] Statistics Poland, Poland

Abstract

Item count techniques (ICTs) are established and widely applicable methods for surveys with sensitive questions. Estimation of the unconditional probability of possessing the sensitive attribute, i.e. estimation of the sensitive proportion is of main importance. Due to the fact that some control variable (or variables) is used in all item count models the problem of the precision and efficiency of the estimation is especially important. Although in social science practice moment-based estimators are widely used, in the modern methodology of the item count techniques the problem is treated as a problem of incomplete data and therefore ML estimators via either EM or Newton-Raphson algorithm are employed. But the use of a parameter approach to item count methods introduces new problems regarding control variable modelling. To our best knowledge the problem of robustness of various item count models concerning violation of the control variable distribution assumptions has not been studied so far. In the paper we analyze different estimation approaches in various item count techniques, including Poisson and negative binomial ICTs and ICTs with a continuous control variable by taking into account violation of the control variable distribution assumptions. We conduct a comprehensive Monte Carlo simulation study and address the consequences of violations of the theoretical assumptions.

Keywords: surveys with sensitive questions, item count techniques, EM algorithm, robustness.

References

Blair, G., and Imai, K. (2012), Statistical Analysis of List Experiments, Political Analysis, 20, 47–77.

Dempster, A. P., L. N. M. Laird, and D. B. Rubin (1977), Maximum-Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society: Series B, 39, 1–37.

Groves, R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2009), Survey Methodology, Hoboken, NJ: John Wiley & Sons.

Imai, K. (2011), Multivariate Regression Analysis for the Item Count Technique, Journal of the American Statistical Association, 106, 407–416.

Kowalczyk, B., Niemiro, W., Wieczorkowski R., Item count technique with a continuous or count control variable for analyzing sensitive questions in surveys, Journal of Survey Statistics and Methodology, 2021, smab043, https://doi.org/10.1093/jssam/smab043

Krumpal, I., B. Jann, M. Korndorfer, and S. Schmukle (2018), Item Sum Double-List Technique: An Enhanced Design for Asking Quantitative Sensitive Questions, Survey Research Methods, 12, 91–102.

Kuha, J., and J. Jackson (2014), The Item Count Method for Sensitive Survey Questions:
Modeling Criminal Behavior, Journal of the Royal Statistical Society: Series C, 63, 321–341.

Liu, Y., Tian, G.-L., Wu, Q., and Tang, M.-L.. 2019. Poisson–Poisson item count techniques for surveys with sensitive discrete quantitative data, Statistical Papers, 60, 1763-1791.

Miller, J. D. (1984), "A New Survey Technique for Studying Deviant Behavior," PhD thesis, The George Washington University, USA.

Tian, G.-L., M.-L. Tang, Q. Wu, and Y. Liu (2017), Poisson and Negative Binomial Item Count Techniques for Surveys with Sensitive Question, Statistical Methods in Medical Research, 26, 931–947.

Tourangeau, R., and T. Yan (2007), Sensitive Questions in Surveys, Psychological Bulletin,
133, 859–883.

Trappman, M., I. Krumpal, A. Kirchner, and B. Jann (2014), Item Sum: A New Technique for Asking Quantitative Sensitive Questions, Journal of Survey Statistics and Methodology, 2, 58–77.

TECHNOLOGIES FOR CREATING AND ANALYZING TESTS IN ADVANCED MATHEMATICS

N. Kruglova[1], O. Dykhovychnyi[2] and M. Poprozhuk[3]
[1] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
e-mail: natahak@ukr.net

[2] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
e-mail: a.dyx@ukr.net

[3] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
e-mail: poprozhuk19@gmail.com

Abstract

Over the past three years, distance education has become one of the primary formats of education for Ukrainian students, initially due to the COVID-19 pandemic and currently because of ongoing Russian missile strikes on civilian targets in Ukraine. It is highly likely that distance education will remain the only viable option for a considerable period of time. Therefore, creating high-quality content to test students' knowledge under these conditions has become an exceedingly critical task.

A team of instructors at National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" are currently working on mathematics tests design in the two main areas:

A. Test item design. To facilitate this process, we employ Wolfram Mathematica as a practical tool that allows us to perform problems' solutions verification, designing numerical problems with integer answers, modeling high-quality distractors for multiple-choice questions, creating images for problem illustration, and identifying problems that are likely to be solved by students using online resources.

B. Statistical analysis of test quality. The quality of the designed tests is analyzed by testing them on a control group of students, with the aim of improving the tests by removing or reformulating items that are either too easy or too difficult.

Classical Test Theory (CTT) and Item Response Theory (IRT), Crocker (2006) are the primary conventional tools used for this purpose. However, we apply a more advanced framework called Multidimensional Item Response Theory (MIRT), Reckase (2009). This framework permits a more nuanced differentiation of the individual characteristics and abilities of examinees. A key challenge when applying MIRT is to select an appropriate model for analysis, particularly when it comes to choosing the model's dimensionality. The model selection methodology we employ is as follows:

1) Exploratory Factor Analysis (EFA). We conduct EFA to initially select the model's dimensionality based on Auerswald (2019) approach. At this stage, we use methods like Parallel Analysis, Empirical Kaiser Criterion, etc. to determine the number of competencies being tested as well as the number of latent numerical parameters that characterize a student.

2) Estimation of model parameters. For both compensatory and non-compensatory MIRT models, we utilize Confirmatory Factor Analysis (CFA), Chalmers (2012) approach for model parameters estimation employing EM or NH-RM algorithms.

3) Assessing model adequacy. Since different EFA algorithms can yield different dimensionalities, the most adequate model is selected based on the following criteria: M2, RMSEA, TLI, CFI.

We chose the R programming language for statistical analysis of test quality.

Keywords: IRT, MIRT, EFA, CFA

References

Auerswald, A., and Moshagen, M. (2019) How to Determine the Number of Factors to Retain in Exploratory Factor Analysis: A Comparison of Extraction Methods Under Realistic Conditions. *Psychological Methods*. Advance online publication. 1-24.

Chalmers, R. (2012) Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, Volume 48, Issue 6, 1-25.

Crocker, L., and Algina, J. (2006) *Introduction to Classical and Modern Test Theor*y, Belmont, CA:Wadsworth.

Reckase, M. D. (2009) *Multidimensional item response theory*, Springer, New York.

The Self-Organizing Map for the Analysis of Survey Data

M. Litova[1] and K. Lagus[2]

[1] University of Helsinki, Finland
e-mail: maria.litova@helsinki.fi

[2] University of Helsinki, Finland
e-mail: krista.lagus@helsinki.fi

Abstract

The self-organizing map (SOM) algorithm is implemented to analyze the survey data obtained from the faculty well-being project conducted in spring 2022 at the Faculty of Social Sciences in the University of Helsinki. The self-organizing map is an unsupervised neural network and data analysis method that enables dimensionality reduction, exploration of variable variation and dependencies and visualization of similarity relations (Kohonen, 2001). With its strong clustering capabilities and visualization potential, the self-organizing map can be effectively applied to the analysis of survey data, particularly data collected through questionnaires.

The survey conducted as part of the faculty well-being project comprised three question groups. The first group focused on gathering background information, while the second group consisted of closed Likert-scale questions aimed at capturing individuals' subjective experiences of well-being within the faculty. Lastly, the survey included open-ended text questions that explored various topics, including social interaction, the role of the faculty in promoting well-being, maintaining equality, and other related subjects (Laine et al., 2022). Closed background and Likert-scale questions were chosen to be analyzed with the SOM method.

The utilization of the self-organizing map algorithm for the analysis of closed questions facilitated the identification of seven profiles (clusters) among survey participants. These profiles were obtained based on the varying experiences concerning well-being and the accompanying background information, including gender, position within the faculty and proficiency in the Finnish language. The implementation of the SOM method can be described as an experimental undertaking that typically involves several key steps. These steps include the identification of the optimal set of parameters for the SOM training, the selection of an appropriate approach for encoding variables, and the handling of missing values within the dataset.

Keywords: self-organizing map, survey data, clustering, well-being.

References

Kohonen, T. (2001). Self-Organizing Maps. Springer-Verlag; Berlin, Heidelberg.

Laine, K., Litova, M., and Oikarinen T. (2022). Crowdsourcing social wellbeing non-probability survey. In*: Baltic-Nordic-Ukrainian workshop on survey statistics 2022,* Tartto, 46.

ANALYSIS OF RESPONSE REPRESENTATIVENESS IN CASE OF ADAPTIVE SURVEY DESIGN

A. Meļņičuka[1] and J. Voronova[2]

[1] Central Statistical Bureau, Latvia
e-mail: Anastasija.Melnicuka@csp.gov.lv

[2] Central Statistical Bureau, Latvia
e-mail: Jelena.Voronova@csp.gov.lv

Abstract

In practice, researchers still focus mainly on reducing survey non-response through data collection, thus sometimes sacrificing representative response. To achieve representativeness, balancing of response levels across different groups of respondents has major importance. If the sample is not representative, the estimates of the variables of interest may deviate significantly from the population values, even those having high response rates. To address this issue, Schouten, Cobben, and Bethlehem (2009) introduced Representativeness Indicators (R-indicators). Representativeness is measured by degree of difference between respondents and non-respondents based on response propensities. Logistic regression model is a typical tool used to estimate response propensities. Consequently, R-indicators used in adaptive data collection to ensure representativeness of a sample can additionally help in non-response adjustment through weighting procedures aimed at reducing non-response bias by considering individual response propensities.

This study was conducted at the Central Statistical Bureau of Latvia and is focused on analyzing data from the Adult Education Survey 2022 (AES). The AES is aimed at gathering internationally comparable data on adult participation in lifelong learning activities – formal education, non-formal education and training, and informal learning. During AES data collection, an adaptive survey design (ASD) was implemented to improve response rates among underrepresented groups. The collected response datasets were fixed on assigned dates – before and after the adaptive design was implemented, assuming the fieldwork was over. It allows us to review possible results and examinate changes in R-indicators and partial R-indicators as well as changes in estimators used for target variables. We will assess ASD effectiveness by estimating accuracy of variable estimates and considering details of the weighting procedures at different stages of data collection. Target variables were estimated by using multiple weights with non-response correction, including homogeneity groups correction and calculations dependents on individual response propensities. The results were compared based on calculation of the variable of interest and precision estimates. The research was aimed at analyzing effects of adaptive design, testing different weighting schemes, evaluation of estimate quality and results produced. The goal of the study is to improve weighting and data collection process in the AES.

Keywords: non-response analysis; adaptive design; R-indicators; response propensity; weighting.

References

Schouten B., Cobben F., and Bethlehem J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35(1), 101-113.

The Bahadur Representation of Sample Quantiles in General Unequal Probability Sampling Designs

Hitoshi MOTOYAMA[1]

[1] Delft University of Technology, Netherlands;
Aoyama Gakuin University, Japan;
The Institute of Statistical Mathematics, Japan
e-mail: H.M.Motoyama@tudelft.nl; hitoshi@aoyamagakuin.jp.

Abstract

This presentation establishes the Bahadur representation of sample quantiles in general unequal probability sampling designs.

The Bahadur representation, proposed by and named after Bahadur(1966), is a useful linear representation of quantiles that has been extended to M-estimators and time-series data. Moreover, in a sample survey framework, the Bahadur representation of quantiles plays an important role. Francisco and Francisco and Fuller(1991) and Shao(1994) developed the Bahadur representation for stratified cluster sampling and stratified multistage sampling, respectively. However, it has been considered difficult to develop Bahadur representations in the general sampling framework. In fact, Wu and Thompson(2020), citing Chen and Wu(2002), state that "With complex survey data, however, Bahadur representations are difficult to establish even under very restrictive regularity conditions."

In this presentation, we establish the Bahadur representation of sample quantiles in general unequal probability sampling designs under the fairly general and mild conditions similar to Boistard *et al.*(2017).

Consider a sequence of finite population $\mathcal{U}^N$, of size $N = 1, 2, \ldots$. With each population, we associate a set of indices $U_N = \{1, 2, \ldots, N\}$. Furthermore, for each index $i \in U_N$, we have a number $x_i \in \mathbb{R}$ which is the values of the variable of interest. For all $N = 1, 2, \ldots$, let a sequence of sample $\mathcal{S}_N = \{s : s \subset U_N\}$ be the collection of subsets of U_N. We define the sample size $n = n(N)$ for the sample $\mathcal{S}_N$ as the cardinal number of $\mathcal{S}_N$. Moreover, we define the first-order inclusion probability π_i of the unit i as $\pi_i = P(i \in s)$.

Define F_N and F_n be the population distribution function and the Hájek estimator for F_N, respectively,

$$F_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}(x_i \le x) \quad \text{and} \quad F_n(x) = \frac{\sum_{i\in s} I(x_i \le x)/\pi_i}{\sum_{i\in s} 1/\pi_i}, \tag{1}$$

where $\mathrm{I}(A)$ is the indicator function defined by 1 if A is true and 0 otherwise.

Under some regularity conditions, we establish the Bahadur representation of sample quantiles:

$$\hat{\theta}_n = \theta_N + \frac{p - F_n(\theta_N)}{f_N(\theta_N)} + o_p(n^{-1/2}), \tag{2}$$

where $\hat{\theta}_n = \inf\{x : F_n(x) \ge p\}$ are the p-th sample quantiles and $\theta_N = \inf\{x : F_N(x) \ge p\}$ are the p-th population quantiles.

Keywords: Bahadur representation; Quantiles; Finite population; Survey Sampling.

References

Boistard, Hélène and Lopuhaä, Hendrik P. and Ruiz-Gazen, Anne. (2017). Functional central limit theorems for single-stage sampling designs. *The Annals of Statistics*, **45**, 1728–1758.

Chen, Jiahua and Wu, Changbao. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, **12**, 1223–1239.

Francisco, Carol A. and Fuller, Wayne A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, **19**, 454–469.

Shao, Jun. (1994). L-statistics in complex survey problems. *The Annals of Statistics*, **22**, 946–967.

Wu, Changbao and Thompson, Mary E. (2020). *Sampling theory and practice*. Springer.

Statistical analysis of the impact of the blackout caused by the Russian attack on the infrastructure of Ukraine on the educational process at NTUU «KPI»

Mulyk Olena [1], Tetiana Pryhalinska [2] and Linieana Svystun-Zolotarenko [3]
[1] NTUU Igor Sikorsky Kyiv Polytechnic Institute, Ukraine
e-mail: mulyk.olena@gmail.com

[2] NTUU Igor Sikorsky Kyiv Polytechnic Institute, Ukraine
e-mail: tania.krivorot@gmail.com

3 Medical Statistics Office of Kyiv Clinical Hospital of Ukrzaliznytsia No. 1, Ukraine
e-mail: lineana.sv@gmail.com

Abstract

For three years, humanity has been living under the influence of the Covid-19 pandemic. Now, in addition to the problems related to the Covid-19 pandemic, our state has faced another extraordinary hard challenge the decision by President Vladimir Putin and all of Russia to launched a full-scale re-invasion of Ukraine a year and a half ago. It has tragically been changing of the lives of children and young people in Ukraine. Everyone knows that last year, starting from October, Russia has been targeting Ukraine's energy infrastructure with missiles and drones. Since October 10, 30% of Ukraine's power stations have been destroyed, causing massive blackouts across the country. Therefore, students had to search connection to the Internet in "points of unbreakability (Invincibility points).

The goal of the work is to discuss the peculiarities of the university educational process in the conditions of war. Peculiarities and problems of teaching higher mathematics are considered in the example of our own experience in the under-bombing period in Ukraine and causing massive blackouts across the country. To consider the results of tests provided for medical engineering faculty NTUU "KPI" students about investigating the level of stress caused by a power outage affects. Find out the ability and problems of students to continue their studies under this kind of stress

Content and methods. Last three years distance education was continued as the main form of education, but in October 2022 it was interrupted. The teachers and students had significant tests in the use of distance education when there was a power outage. We want to investigate the level of stress caused by a power outage affects. Find out the ability and problems of students to continue their studies under this kind of stress

From the middle of October we began to observe a decline in the learning results of our students. We had seen their passivity in online classes or great anxiety and worry in behavior, they handed in their papers with great delays. They had written in personal messages that they cannot do tasks by the deadline and do not understand the material. Also, they are unable to perform practical and control tasks because they do not have access to the Internet and the electricity is turned off in their region. Because of this, we have conducted a questionnaire in order to identify the needs and problems of students during the power outage period. We meant had the goal of finding and implementing available forms of presentation and control of educational information.

As the results of the survey showed, among the interviewed students of NTUU "KPI" 20.6% were forced to go abroad. Almost 80% remained in Ukraine and had to continue their studies in the difficult conditions of the war. 64.8% of respondents admitted that the events related to Russia's

military invasion of Ukraine became a significant stressful experience for them. For 27.3% of students, the war had a significant impact on their further education plans; 65.6% of the surveyed respondents noted that in connection with the Russian invasion, the educational process became "very" and "tangibly" stressful for them. Almost 60% of students felt negative emotions (fear, anger); 25% experienced problems in experiencing positive emotions (joy and love); 33% of students began to feel strong physical reactions, such as heart palpitations, difficulty breathing, sweating at the mention of the Russian invasion. 78% of students were stressed even if the power outage occurred on schedule; more than 75% of respondents had obsessive thoughts that the lights would switch off, and they wouldn't have time to complete their homework or test, or wouldn't have time to send the assignment to the teacher for review by the deadline. The lack of mobile and Internet connection became a tangible stress for 68% of surveyed students. The test results are shown in Fig. 1 and Fig. 2.

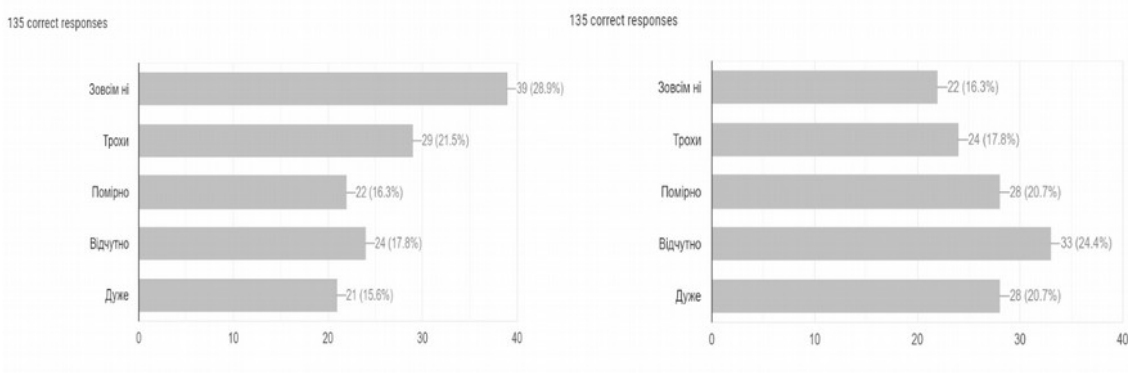


Fig.1. Results of answers to questions:

Left: Have you had strong physical reactions (palpitations, difficulty breathing, sweating) when something reminds you of the events of the Russian invasion?

Right: During the Russian invasion, did you experience problems in experiencing positive emotions (for example, inability to feel joy or love)?

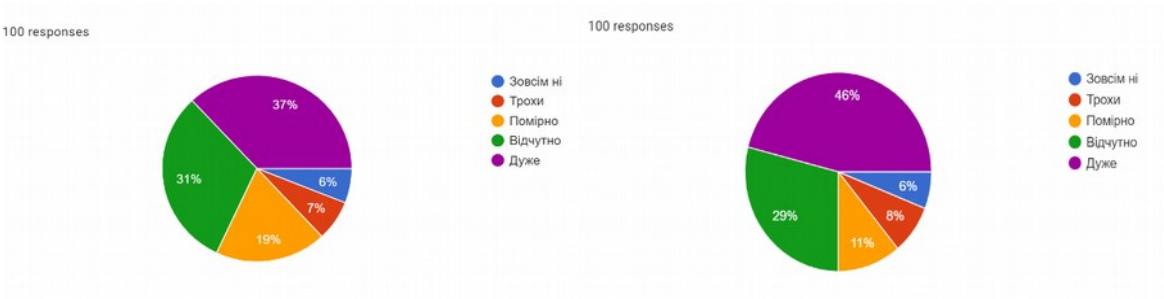


Fig.2. Results of answers to questions:

Left: How stressful was it for you to lose contact during your studies?

Right: How much did the stress of the sudden blackout affect your ability to concentrate on your studies?

The remote form in the pre-war period allowed testing students of technical specialties on the Moodle platform, but it became unusable during a blackout. Therefore, we began to use Google Forms

more actively (95% of students were inclined to this form of monitoring the learning of the lesson tasks, compared to the Moodle platform). The Google form developed by the teacher can be opened at any time and, by personal agreement between the teacher and students, executed several times. This adds flexibility to the system of self-monitoring of the level of knowledge both for the student himself and for control by the teacher while maintaining the objectivity of the assessment. Google Forms has proven to be extremely effective when combining theoretical questions with practical ones. In addition, the advantage of a Google form compiled in this way is that it can be given to students several times if the result obtained was low. Repeated tasks contribute to a better understanding of the content because students are forced to review the educational material more carefully and not much stress on the student.

Also, we have been videotaping the all lessons (and continue doing) and using the YouTube network and the Telegrams channel to contact students.

In conclusion, we can say that the system of higher education should be based on the study of the state, conditions, and development of the main indicators of the field of education, the application of statistical methods allows solving this problem. The training of specialists in Ukraine must be based on a systematic approach, which is why the organization of the process of managing the training of specialists must take into account all the conditions and factors that shape this process. Therefore, to study the state and development of the higher education system in Ukraine, it is suggested to use such statistical methods as distribution analysis, expert evaluations, and factor analysis.

Since the results of student testing have shown a high level of stress, we will continue the research in more detail and adapt conclusions to the educational process in the future.

References

Nikolaev E., Riy G., Shemelinets I. (2023) Higher education in Ukraine: changes due to the war: analytical report. *Borys Grinchenko Kyiv University*.

Кудзіновська І. П., Трофименко В. І. Особливості викладання математичних дисциплін у закладах вищої освіти в умовах воєнного стану. *Інтернаука. Сер.* : педагогічні науки. URL: https://www.inter-nauka.com/uploads/public/16606360462734.pdf.

Програма великої трансформації *«Освіта 4.0: український світанок»*. URL: https://mon.gov.ua/storage/app/media/news/2022/12/10/Osvita-4.0.ukrayinskyy.svitanok.pdf

Про затвердження плану дій органів виконавчої влади з відновлення деокупованих територій територіальних громад*: розпорядження Кабінету Міністрів України від 30.12.2022 № 1219-р.* URL: https://www.kmu.gov.ua/npas

Demographic Patterns and Prevalence of Mental Health Disorders in Europe

S. Myrvoda[1] and H. Livinska[2]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: sofiia.myrvoda@knu.ua

[2] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: hanna.livinska@knu.ua

Abstract

Mental health disorders are a major public health concern globally, and according to the World Health Organization (WHO) in 2019 one in every eight people in the world suffers from a certain mental problem. Mental health disorders are widespread throughout various demographic groups in Europe, although there are noticeable variations. Our work aims to explore the occurrence and patterns of mental health disorders, with a particular focus on demographic factors such as gender and region.

With this objective, we conducted a systematic review of the existing literature and analyzed data over the period of 1990-2019 from the Global Burden of Disease (2019). The data included measures of mental health disorders and demographic variables such as location, sex, age, etc. It is revealed that mental health disorders are prevalent across all demographic groups, highlighting that mental health is an issue that impacts everyone regardless of their background.

However, our analysis also showed that there are some notable differences in the rate of mental health disorders across different demographic groups. For example, women had a higher prevalence of depression and anxiety compared to men. This is consistent with previous research that has found that women are more likely to experience mental health disorders than men. Furthermore, people in Western Europe had a higher rate of anxiety than Central or Eastern Europe, indicating that there may be regional differences in the prevalence of mental health disorders.

Our study highlights the need of targeted interventions to address the significant burden of mental health disorders. Given the variations in their prevalence across different demographic groups, the interventions need to be tailored to the specific needs of each group. For example, actions that address the higher prevalence of depression and anxiety among women could include targeted counseling and support services. Similarly, measures that address the higher prevalence of anxiety in Western region could focus on reducing stress and improving access to mental health services in those regions.

Keywords: Demography, Mental Health Disorders, Machine Learning.

References

Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2019 (GBD 2019) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2020. Available from https://vizhub.healthdata.org/gbd-results/.

World Health Organization (2022) Fact sheets - Mental Disorders. Available from https://www.who.int/news-room/fact-sheets/detail/mental-disorders

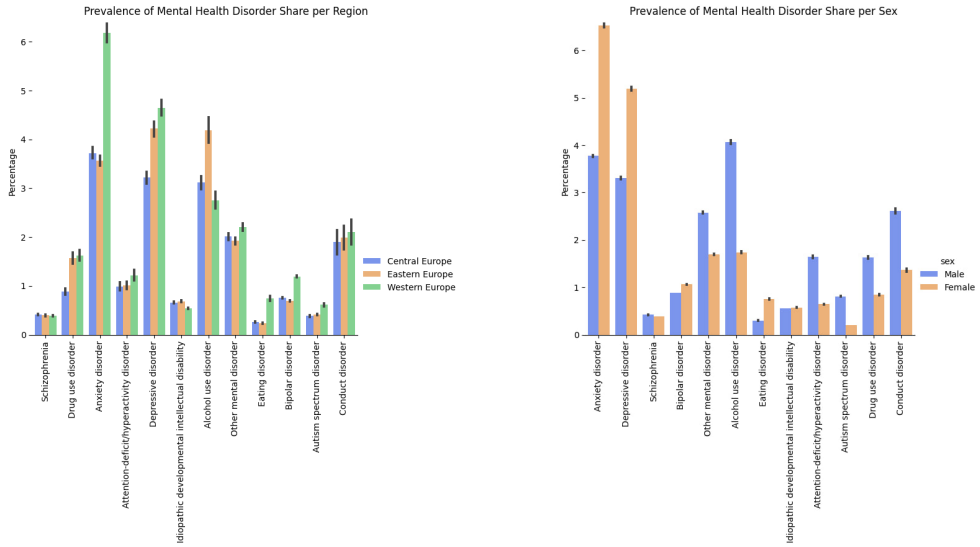


Figure 1: Prevalence of Mental Health Disorders per Region and per Sex.

Estimating Proportions from Integrated Probability and Non-Probability Samples

V. Nekrašaitė-Liegė[1,2], A. Čiginas[1,3] and D. Krapavickaitė[4]

[1] State Data Agency, Statistics Lithuania, Lithuania

[2] Vilnius Gediminas Technical University, Lithuania
e-mail: Vilma.Nekrasaite-Liege@vilniustech.lt

[3] Vilnius University, Lithuania
e-mail: Andrius.Ciginas@mif.vu.lt

[4] Lithuanian Statistical Society, Lithuania
e-mail: Danute.Krapavickaite@gmail.com

Abstract

The estimation of finite population parameters by combining data from probability and non-probability samples is a prevalent research topic in survey sampling today. Various scenarios for data availability have been investigated in the literature. In some cases, the study variable is observed only in a non-probability sample and there is no possibility to combine samples at the unit level (Chen et al., 2020 and Yang et al., 2020). In other cases, both samples can be combined at the unit level and the study variable is observed either in non-probability sample (Kim and Haziza, 2014 and Kim and Wang, 2018) or both (Tam and Kim, 2018).

In this study, we focus on the situation where the study variable is available in both samples, and we aim to estimate the proportion using post-stratification and composite estimation methods.

We pay attention to the assessment of the variance for the estimator, taking into account not only the randomness of the probability sample but also the randomness of the non-probability sample. The influence of the non-probability sampling on the variance estimator is evaluated with respect to the distribution of estimated propensity scores.

Keywords: non-probability sample, post-stratification, propensity score, composite estimator.

References

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association* **115:532** 2011-2021, DOI: 10.1080/01621459.2019.1677241.

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Stat. Sin.* **24**, 375-394.

Kim, J. K. and Wang, Z. (2019). Sampling Techniques for Big Data Analysis. *International Statistical Review* **87**.

Tam, S.-M. and Kim J.-K. (2018). Big data, selection bias and ethics – an official statistician's perspective. *Statistical Journal of the IAOS* **34**, 577–588.

Yang, S., Kim, J. K. and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high-dimensional data. *J R Stat Soc Series B Stat Methodol.* **82**(2), 445-465. doi: 10.1111/rssb.12354. Epub 2020 Jan 7. PMID: 33162780; PMCID: PMC7644042.

Evaluation of the Efficiency in Healthcare using queueing modelling: A Case Study of Intensive Care Units in Kyiv

A. Pererva[1] **and H. Livinska**[2]

[1] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: tont0n@knu.ua

[2] Taras Shevchenko National University of Kyiv, Ukraine
e-mail: hanna.livinska@knu.ua

Abstract

Queueing theory gives suitable tools for studying special systems that model repeated execution of the same type of tasks that appear in many fields of production, household services, economy, finance, etc. Therefore such models are widely used for designing different types of stochastic objects. One of the goals of queueing theory is to calculate performance measures of the systems, in particular queue length, number of customers in the system, probability of refusals caused by a mismatch between the demand for a service and the capacity to satisfy the demand.

The healthcare industry around the world suffers from queuing refusals. Most commonly, this is a problem of estimating the level of service provided to patients, the cost of the medical service, the average waiting time, the number of patients in the queue, the capacity used and the probability that a patient needs to wait, or the probability of a patient being turned away when all service servers are occupied in the case of a non-queuing system. Hospitals should constantly improve their work, because poor planning and a faulty logistics system lead to excessively long stays.

Our work examines an intensive care unit, where certain urgent mechanical or pharmacological support is required. We consider an intensive care unit as a queueing system without waiting places (without a queue), where beds are model servers and patients are customers. The flow of customers, that is a flow of emergency patients, is supposed to be a Poisson one, such flows naturally appear in our case (patients arrive independently, usually one at a time, with stable intensity during a day). The rate of patient arrivals assumed to be constant. The time spent in the intensive care unit is determined by an exponential distribution.

By modelling the unit as a queueing system, we aim to study and evaluate the performance measures of the system such as the probability of patient rejection, the optimal number of beds in the unit, the rate of the arrival of patients per day, and the average time of patient service. We estimate parameters of the system based on the data for emergency unit in Oleksandriv Hospital of Kyiv. Having the parameters, we obtain the probability of refusal. Different optimization problems for the cost of the system service can be formulated and solved in order to find a balance between the average workload and the probability of failure.

Queuing and failure analysis can significantly improve medical productivity, patient satisfaction, and cost-effectiveness of health care. This determines the relevance of the study of the efficiency of the Oleksandriv Hospital and its comparison with the average Kyiv indicators. One of the purposes of the work is to identify the impact on the efficiency of intensive care units in Kyiv under changing the basic model parameters such as the number of beds, service time, and the arriving rate of patients.

Keywords: queuing model, failure probability, emergency department.

References

Komenda I. (2013) *Modelling Critical Care Unit Activities Through Queueing Theory.* School of Mathematics, Cardiff University, Cardiff.

Yan Z., Jie Z., Min T.,Jian S.,Degang Z. (2022) *Forecasting patient arrivals at emergency department using calendar and meteorological information.* Hangkong University, Nanchang.

Creemers S., Lambrecht M. (2008) *Healthcare queueing models.* Catholic University Leuven, Belgium.

MODEL-ASSISTED SMALL-AREA ESTIMATION WITH AUTOMATED MODEL BUILDING FOR SWISS NATIONAL FOREST INVENTORY USING TWO-PHASE SAMPLING

M. Pulkkinen[1], J. Zell[2] and A. Lanz[2]

[1] Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Switzerland &
Laboratory of Forest Inventory, National Institute of Geographic and Forest Information IGN, France
e-mail: minna.pulkkinen@ign.fr

[2] Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Switzerland

Abstract

National Forest Inventories (NFIs) provide estimates of parameters related to the state and change of forest resources within a country. In addition to national and regional estimates, information is increasingly being demanded also for smaller areas and domains. In the Swiss NFI, the estimates are based on the measurements of permanent field plots placed on a rectangular grid over the country. In small areas, however, the current design-based post-stratified estimators are imprecise or impossible to apply due to the insufficient number of field plots. With reasonable auxiliary data available (that covary with target variable data and cover the entire country on a denser grid than target variable data), estimates for small areas can be obtained with design-based model-assisted estimators, which exploit auxiliary data via models between target variables and auxiliary variables.

We employed the model-assisted estimators introduced by Mandallaz (2008, 2013) for two-phase sampling of continuous populations in order to estimate the forest area and growing-stock volume in the Swiss municipalities (n = 2932), forest districts (n = 101) and cantons (n = 25). The first-phase auxiliary data on a 100 m × 100 m grid came from several sources (vegetation height model derived from digital aerial images; Copernicus satellite image products; digital elevation model based on a LiDAR campaign) and comprised several dozen variables (with transformations and interactions included, altogether 80–120 potential explanatory variables). The second-phase target variable data on a √2 km × √2 km sub-grid were obtained from the field plot measurements of the 4th Swiss NFI. For forest area estimation, we used a common external logistic model in all the small areas. For volume estimation, we built internal linear models individually for each small area (i) by fitting a model of a common pre-determined form (established in a separate pre-study), and (ii) by building a linear model in an automated procedure, where the explanatory variables included in the model were selected with lasso. Individual model-building data for each small area consisted of a fixed number of field plots (fulfilling certain quality criteria) closest to the area centroid and were obtained by adding plots from outside the area; the required number of plots varied according to the type of the area (municipality < forest district < canton).

The resulting model-assisted estimates were compared to the standard Swiss NFI post-stratified estimates in forest districts and cantons (where such estimates could be computed). The model-assisted point estimates were found to be close to the NFI estimates and generally more precise than the NFI estimates. Consequently, the model-assisted estimates in municipalities (for which no NFI reference estimates exist) can be considered fairly plausible. The models constructed with the automated procedure generally resulted in more precise estimates than the models of the pre-determined form. Further work is needed, however, to study the effect of the model-building settings (size and quality of model-building data, transformations and interactions of auxiliary variables, elastic net instead of just lasso for variable selection) on estimation precision. In conclusion, design-based model-assisted small-area estimation with automated model building could be developed into an operational system, to which new target parameters and updated auxiliary data sets can easily be

added, as no prior modelling effort is needed. Model-assisted estimation also exhibits a clear potential for improving precision even in areas and domains where standard estimators can be applied.

Keywords: continuous population, design-based estimation, difference estimator, lasso, regression estimator.

References

Mandallaz, D. (2008) *Sampling Techniques for Forest Inventories*. Chapman & Hall/CRC, Boca Raton, USA.

Mandallaz, D. (2013) Design-based properties of some small-area estimators in forest inventory with two-phase sampling. *Canadian Journal of Forest Research*, **43**, 441–449.

DATA COLLECTION METHODS FOR RESEARCH ON EDUCATION

B. Sloka[1] and K. Liepina[2]

[1] University of Latvia, Latvia
e-mail: Biruta.Sloka@lu.lv

[2] University of Latvia, Latvia
e-mail: Kristine.Liepina@lu.lv

Abstract

Recent data and data file availability for research in education make the research process available deeper and wider using data from EU-SILC surveys, from Labour Force Surveys and other data available in Official Statistics portal which are representative with ell described meta data and surveys they very useful, but often there are needed data which are available only using additional surveys, expert surveys and other sources there huge work and contribution by researchers is needed. For additional organized surveys and expert surveys for survey questionnaire preparation, pilot surveys, survey organization and data collection theoretical findings are on great importance where classical findings (Sarndal, 2022) and recent findings in the field (Valente, Salavisa, Lagoa, 2016) have to be taken into account.

Aim of the research: find best possible scientific and practical suggestions for survey organization and expert survey organization for data collection for education management research.

Research methods applied: scientific publications and previous conducted research analysis and practical findings evaluations for obtaining as good as possible data for research.

Research results indicated that the experience on international institutions (for example, EUROSTAT, OECD, etc.) experience and publications are very good examples for practical survey development in education management research

Keywords: research in education management; questionnaire design; pilot survey; survey organisation.

References

Sarndal C.-E. (2022). Progress in survey science and practice yesterday - today – tomorrow. Presenation on Workshop on Survey Statistics 2022, Tartu, 2022, available at file:///C:/Users/EVF/Downloads/resumetartu202206010.pdf

Valente, A.C., Salavisa, I., Lagoa, S. (2016). Work-based cognitive skills and economic performance in Europe. *European Journal of Innovation Management*, 19(3), 383-405.

Yousef, D.A. (2016). The use of the learning styles questionnaire (LSQ) in the United Arab Emirates. *Quality Assurance in Education*, 24(4), 490-506.

COMBINING POPULATION STATISTICS, CONJOINT DATA AND PURCHASE HISTORY IN PRICE OPTIMIZATION

L. Valkonen[1], S. Tikka[1], J. Helske[1] and J. Karvanen[1]

[1] Department of Mathematics and Statistics
University of Jyvaskyla, Finland
e-mail: juha.t.karvanen@jyu.fi

Abstract

With multiple data sources, it is possible to address estimation problems that would otherwise remain unsolvable. The inherent differences between populations or the sampling methods provide a major challenge for data-fusion. We demonstrate how causal modeling and Bayesian estimation can overcome these challenges in a business scenario.

The pricing of products and services is one of the most important decisions that almost every company faces. Our scenario is motivated by a real business case and focuses on the pricing of a subscription-based service, such as video or music streaming, an audiobook service, or a digital newspaper. The subscription is automatically renewed every month unless canceled by the consumer. In price optimization (Phillips, 2021), a company has to estimate the impact of a price change on the behavior of both current customers and potential new customers. Thus, price optimization is essentially a causal problem.

The data sources for the price optimization scenario are the following:

Population statistics are available from Statistics Finland and provide information on the joint distribution of age, gender, and location in the target population.

Conjoint data originates from a study where the price is varied and a group of customers is asked to make imaginary purchases in an artificial setup (Rao, 2014).

Purchase history contains customer-level data on the subscription periods and is collected by the company as a part of daily operations.

The proposed approach (Valkonen et al. 2023) consists of four steps: 1) The causal relations of the purchase process are described in the form of a directed acyclic graph (DAG). 2) The causal effect of the price on purchases is identified from the data sources presented in a symbolic form (Tikka et al., 2021). 3) A hierarchical Bayesian model is fitted to estimate the causal effect based on the obtained identifying functionals. 4) The posterior distribution of the optimal price is found by maximizing the expected gross profit defined as a function of the price and the purchase probabilities estimated in step 3. The approach is demonstrated with simulated data resembling the features of real-world data.

Keywords: Bayesian model, Causal inference, Data-fusion, Demand estimation, Transportability

References

Phillips R. L. (2021) *Pricing and Revenue Optimization*. Stanford University Press.

Rao V. R. (2014) *Applied Conjoint Analysis*. Springer.

Tikka S., Hyttinen A., Karvanen J. (2021) Causal effect identification from multiple incomplete data sources: A general search-based approach. *Journal of Statistical Software*, 99(5):140.

Valkonen L., Tikka S., Helske J., Karvanen J. (2023) Price optimization combining conjoint data and purchase history: A causal modeling approach, *arXiv:2303.16660*, https://arxiv.org/abs/2303.16660.

STATISTICS FOR BIODIVERSITY MONITORING

J. Vanhatalo[1], E. Numminen[2] and J. Siren[3]

[1] University of Helsinki, Finland
e-mail: jarno.vahnhatalo@helsinki.fi

[2] University of Helsinki, Finland
e-mail: elina.numminen@helsinki.fi

[3] University of Helsinki, Finland
e-mail: jukka.p.siren@helsinki.fi

Abstract

Key ecosystem services ultimately depend on biodiversity (Cardinale *et al.*, 2012). There is comprehensive qualitative evidence of a global biodiversity change (e.g., Cardinale *et al.* 2012; Chase *et al.* 2020), which has catalysed a general demand for biodiversity preserving policies and management (MEA, 2005, UN, 2021). Yet, without reliable quantitative information on biodiversity patterns and trends, rational biodiversity management is logically impossible. What we do not know and do not measure, we cannot effectively manage, either. To be directly applicable, such information should be generated at a high spatiotemporal resolution, thus matching that of other national assets, such as infrastructure, land use, and industry (Dasgupta, 2021).

Despite the obvious information need, for most taxonomic groups, and for biodiversity as a whole, we still lack critically validated methods for producing such information, and for validating the uncertainties associated with it. In most countries, long-term biodiversity monitoring programs have been initiated for different reasons, are currently implemented by multiple actors, and remain focused on a few selected taxonomic groups (for example, birds, butterflies and game species, each monitored using a separate design). Individual programs lack coordination both within and between countries, and taxon-specific assessments are rarely combined into holistic analyses of the general state of biodiversity (Roslin & Laine, 2022). The current situation is a major obstacle to achieving sustainable development.

In this talk, we will present an overview of the state of Finnish nature monitoring programs and analyse their usefulness for biodiversity monitoring. Our results highlight that most of the monitoring programs do not provide statistically representative data on Finnish biodiversity – implying that classical design-based estimates are inadequate for analysing these data. As an alternative, we will consider modern model-based approaches and show how they can alleviate the challenge, and what are their current limits (e.g., Foster *et al.*, 2021). Furthermore, we will show, how ecological processes themselves might cause bias to population and biodiversity estimates even when statistically perfect sampling design is applied. Changes in habitat availability often change species behavior so that design-based methods give biased estimates for population change (Numminen *et al.*, 2023).

To tackle the modern challenges posed by the biodiversity change, we need holistic planning for future biodiversity monitoring programs. We will present preliminary results from our on-going work, where we analyze, how future monitoring programs should be arranged in Finland to achieve good cost-efficiency. Our approach is based on Bayesian approach where we calculate the expected utility of alternative monitoring design (Liu & Vanhatalo, 2020).

Keywords: biodiversity monitoring, ecology, species distribution modeling, monitoring design.

References

Cardinale *et al.* (2012). Biodiversity loss and its impact on humanity. Nature 486, 59–67.

Chase *et al.* (2020). Ecosystem decay exacerbates biodiversity loss with habitat loss. Nature 584, 238–243

Dasgupta, P. (2021), The Economics of Biodiversity: The Dasgupta Review. (London: HM Treasury)

Foster, S.D *et al.* (2021). Effects of Ignoring Survey Design Information for Data Reuse. Ecological Applications, 31(6):e02360

Liu, J. & Vanhatalo, J. (2020). Bayesian model based spatio-temporal sampling designs and partially observed log Gaussian Cox process. Spatial Statistics, 35:100392.

MEA (2005). Millennium Ecosystem Assessment. Ecosystems and Human Well-being: Synthesis. Island Press, Washington, DC. Millennium Ecosystem Assessment. Available from: www.millenniumassessment.org

Numminen, E. *et al.* (2023). Species ecology can bias population estimates. Biological Conservation, in press.

Roslin, T., & Laine, A.-L. (2022). The changing fauna and flora of Finland – discovering the bigger picture through long-term data. Memoranda Societatis Pro Fauna Et Flora Fennica, 98 (Supplement 2), 40–53.

UN (2021). United Nations et al. System of Environmental-Economic Accounting –Ecosystem Accounting (SEEA EA). https://seea.un.org/ecosystem-accounting.

OUTLIERS IN LOSS RESERVING

O. Vasylyk[1] and V. Shunder[2]

[1] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
e-mail: vasylyk@matan.kpi.ua

[2] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
e-mail: 1618422a@gmail.com

Abstract

Impact of outliers in loss reserving is a very serious problem in insurance business. An insurance company needs to estimate a reserve as accurately as possible to be able to meet its future obligations arising from incurred but not reported claims. This task is often impaired by outlier-contaminated datasets. Outliers in insurance typically are not data errors but large financial claims that are an important pricing component.

Common reserving techniques include deterministic chain-ladder method, stochastic chain-ladder method, Mack's model, Bornhuetter-Ferguson technique. They are applied on data at a certain level of aggregation, often presented in triangles. The chain-ladder method is the most popular one. The outstanding claims reserve can be obtained as a result of applying the chain-ladder method for the development triangle of a univariate business line or a multivariate chain-ladder method for several development triangles of a company with multiple business lines. But the chain-ladder method is very sensitive towards outliers, and reserve estimates can be significantly shifted in the presence of even one outlier. Widely used methods to eliminate outliers are either limiting the number of outliers by robust statistical methods or by change of development factors. But these methods have several disadvantages.

We shall discuss detection of outliers in datasets, the impact of outliers on reserve estimates, and alternative robust techniques to treat outliers, suggested in resent studies.

Keywords: chain-ladder method, loss reserving, outliers, robust estimation.

References

Avanzi, B., Lavender, M., Taylor, G., Wong, B. (2016) On the impact, detection and treatment of outliers in robust loss reserving.
https://www.actuaries.asn.au/Library/Events/GIS/2016/PaperAvanziLavenderTaylorWong.pdf.

Avanzi, B., Lavender, M., Taylor, G., Wong, B. (2022) On the impact of outliers in loss reserving. *https://arxiv.org/abs/2203.00184.*

Barlak J., Bakon M., Rovnak M., Mokrisova M. (2022) Heat Equation as a Tool for Outliers Mitigation in Run-Off Triangles for Valuing the Technical Provisions in Non-Life Insurance Business. *Risks*, 10(9):171. https://doi.org/10.3390/risks10090171

Jeng, H. (2010). On Small Samples and the Use of Robust Estimators in Loss Reserving. *Casualty Actuarial Society E-Forum, Fall 2010.*

Verdonck, T., Wouwe, M., Dhaene, J. (2009). A Robustification of the Chain-Ladder Method. *North American Actuarial Journal,* **13 (2)**. DOI: 10.1080/10920277.2009.10597555.

Wouwe, M., Phewchean, N. (2016). Robustifying the multivariate chain-ladder method: A comparison of two methods. *Journal of Governance and Regulation*, **5 (1)**. DOI: 10.22495/jgr_v5_i1_p9.

SEPARATING CROSS-CULTURAL AND CROSS-NATIONAL: INSIGHTS FROM THE EUROPEAN VALUES STUDY

Anastasiia Volkova[1]

[1] University of Helsinki, Helsinki, Finland
e-mail: anastasiia.volkova@helsinki.fi

Abstract

Previous quantitative research on morality often uses data from European Values Study (EVS), as this longitudinal survey have a scale on morally debatable behaviors (MDBS). This work investigates the differences in moral values across 28 European countries, using data from the Morally Debatable Behaviors Scale (MDBS) in the European Values Study (EVS). The MDBS measures moral values by asking for justifications for different actions and events, from claiming social benefits to homosexuality and suicide. Results of multi-group confirmatory factor analysis followed by validity tests prove that the MDBS successfully measures the leniency of moral judgments. However, for a deeper understanding, the basis for intergroup analysis in the case of personal-sexual morality may be not countries but other cultural clusters. This paper explores alternative groupings that may explain differences in personal-sexual moral values and similarities between countries, such as religious constructs (based on denominations) and cultural zones (based on historical context, both religious and political). The preliminary findings of this paper support the results of previous research, arguing that while the legal-fair dimension is universal, the personal-sexual dimension depends on cultural context. It is demonstrated how researchers can use the MDBS to analyze and visualize value differences across Europe, drawing on insights from survey methodology and comparative social research.

Keywords: Comparative research, Values and Attitudes Measurement, Morally Debatable Behaviors Scale, European Values Study, Longitudinal and panel surveys, Questionnaire design and testing.

ADAPTIVE SAMPLE SURVEY DESIGN IN DATA COLLECTION

J.Voronova[1]

[1] Central statistical bureau of Latvia, Latvia
e-mail: jelena.voronova@csp.gov.lv

Abstract

The responsive adaptive survey design (ASD), utilizing R-indicators as measures of representativeness, is tested in Central statistical bureau of Latvia (CSB) as a flexible approach for organizing social surveys. R-indicators help to identify potential bias by measuring the degree of difference between responding and non-responding sample groups. Based on the monitoring and analyses of the R-indicators, active interventions are implemented during data collection process to increase the chances of obtaining a representative set of final response unit, thereby reducing variance in the weights of the final survey data.

Using the notation and definition of response propensities as set out in Schouten, Cobben and Bethlehem (2009) and Shlomo, Skinner and Schouten (2012), denote U the set of units in the population U=1,2,..,i,..N and s the set of units in the sample s=1,2,..,i,..n . Denote a response indicator variable R_i which takes the value 1 if unit i in the population responds and the value 0 otherwise. The response propensity is defined as the conditional expectation of R_i given the vector of values x_i of the vector X of auxiliary variables:

$$\rho_x\ (x_i)=E(R_i=1|X=x_i)\ =P(R_i=1|X=x_i) \quad (1)$$

and also denote this response propensity by ρ_x .

Define the R-indicator as:

$$R(\rho_x) = 1 - 2S(\rho_x) \quad (2)$$

Estimation of the response propensity is based on logistic regression model and estimator of the variance of the response propensities:

$$\hat{S}^2(\hat{\rho}_x) = \frac{1}{N-1}\sum_s d_i(\hat{\rho}_x(x_i) - \overline{\hat{\rho}_x})^2 \quad (3)$$

where $d_i = \pi_i^{-1}$ is the design weight or inverse inclusion probabilities and $\overline{\hat{\rho}_x} = \frac{1}{N}\sum_s d_i\ \hat{\rho}_x(x_i)$

. Thereby, estimation of the R-indicator $\hat{R}(\hat{\rho}_x) = 1 - 2\hat{S}(\hat{\rho}_x)$.

As in variance analysis, R-indicator has the same characteristics and could be split into unconditional partial indicators, which measures the distance to representative response for single auxiliary variables and are based on the between variance given a stratification with categories of Z and conditional partial R-indicators measure the remaining variance due to variable Z within sub-groups formed by all other remaining variables as in Schouten, Shlomo and Skinner (2011).

Survey responsive data collection design concept was piloted in CSB on three person surveys all of them was conducted by using systematic stratified simple random sample:

- Objective of the survey “Mobility of Latvian population in 2021” (MOBS) is to find out the mobility habits of the population. 8 978 persons aged 15 to 84 years living in private households in Latvia selected into sample, the response rate in the survey accounted for 60.4 %.

The survey took place at time when the spread of COVID-19 had particularly intensified and stricter restrictions were introduced in Latvia in order to reduce this spread.

- First "Survey on Gender-Based Violence" (SGBV) 2021 is aimed at collecting information on prevalence of various types of violence in Latvia based on common methodology developed by the Eurostat. SGBV covers personal safety and experience with unwanted behaviour at work, in society, partnership, family, and childhood. The target population of the survey covers people aged 18–74 living in private households in Latvia. Within the framework of the survey, 6 300 people were interviewed. The survey was conducted during rapid spread of COVID-19 and strict restrictions imposed to fight it.
- Adult Education Survey (AES) 2022 is aimed at acquiring internationally comparable data on adult participation in lifelong learning activities – formal education, non-formal education and training, informal learning. The questions covered participation in education activities within the last 12 months. Target population of the survey starting from 2022 cover people aged 18–69 living in private households - 8764 usually residents of Latvia. Answers to the questionnaire questions developed by the CSB were given by 5 492 persons.

The focus of the ASD approach was set on ensuring the quality of fieldwork, with a particular emphasis on the representativeness of sample's response unit set. Several steps were taken during fieldwork to achieve the goal. At the first part of the data collection, R-indicators were used for monitor needs, afterward the groups of imbalance were identified and resources of interviewers were redirected to data collection of those groups.

Response propensity model was developed for each monitoring date during data collection period. The response propensities were estimated a generalized linear model (GLM), a generalization of the classical linear model, with the binomial family logistic-regression model (logistic link function). The set of auxiliary variables were built from social-demographic variables and paradata. Various approaches were used for variable selection, including correlation analysis, evaluation of the amount of available data, level of explanation of the propensity to respond. Individual final set was evaluated for each survey. Selection of the final model specification was evaluated by the automatic *stepAIC* procedure from the *MASS* package (Venables, W. N. & Ripley, B. D. 2002), thus iteratively were reviewed all possible models from the initially passed parameters and left only those variables where the AIC criteria was the smallest.

There were no pre-defined methods for ASD, CSB usually uses multi-mode research method, but the impact of COVID-19 was still significant in 2021. The cancellation of face-to-face interviews led to shift to CATI in 2020. It the situation of a defined fieldwork period, a limited resources as number of interviewers were available, at least one contact for every sample unit were allowed. All resources were planned to be redirected to imbalanced groups.

An important implication of the study was individuality of the survey aim and scope, its influence on the results of the tests. Although more active intervention was made in MOBS, the representativeness of the response set increased in both (MOBS and SGBV) at the end of data collection period. In association with assessed results, possible assumption is the survey aim and type of questions affects representativity - MOBS questionnaire is about habits in specific period, while SGBV survey questions is more about whole life experience.

One of the aims of data processing was to assess variance, and the analysis showed better results in MOBS than SGBV, mostly because of different goals of the surveys. MOBS data, before and after re-directing interviewer resources, showed a reduction of variance. Additionally, an overestimation of the variable of interest was observed in the imbalanced response set. SGBV showed an imbalance in the final response set by sex, and the impact of COVID-19 restrictions on survey results were also observed.

Some valuable lessons were learned in organization and managing ASD in CSB during the 2021 surveys. An ASD dashboard for monitoring needs, which was evaluated by the survey manager, was introduced in AES 2022. Process of results analysis of ASD is ongoing, and the results will be available later this summer.

Keywords: Nonresponse, representativeness, R-indicator, adaptive design

References

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schouten B., Cobben F., Bethlehem J. (2009) *Indicators for the representativeness of survey response,* Computer Science, Chemistry Dalton Transactions.

Schouten B., Peytchev A., Wagner J. (2018) *Adaptive survey design*, Chapman & Hall/CRC Statistics in the Social; Behavioral Sciences.

Schouten B., Shlomo N. (2015*) Selecting adaptive survey design strata with partial R-indicators,* CBS, https://www.cbs.nl/-/media/imported/documents/2015/51/2015-selecting-adaptive-survey-design-strata-with-partial-r-indicators.pdf?la=nl-nl

Schouten B., Shlomo N., Skinner C.J. (2011) *Indicators for monitoring and improving representativeness of response*, Journal of Official Statistics, Vol. 27, No. 2, 231-253.

Shlomo N., Schouten B., de Heij V. (2013) *Designing adaptive survey designs with R-indicators*, NTTS 2013, https://ec.europa.eu/eurostat/cros/system/files/NTTS2013fullPaper_63.pdf

Shlomo N., Skinner C., Schouten B. (2012) Estimation of an indicator of the representativeness of survey response, Volume 142, Issue 1, January 2012, 201-211.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

AUTOREGRESSIVE MODELS FOR AIR QUALITY INVESTIGATION

O. Zalieska[1] and H. Yailymova[2]

[1] Taras Shevchenko National University of Kyiv, Ukraine
[2] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
e-mail: zalieskaolena@knu.ua, yailymova.hanna@lll.kpi.ua

Abstract

The aim of the work is to build a forecast of air quality in Kyiv for some period of time. For this purpose we preprocessed and analized data, selected and fitted a model.

Keywords: machine learning, autoregression, air quality, time series

1. Introduction

The development of technology, increased production of certain products contribute to higher emissions of harmful substances into the air. Air pollution causes climate change, increases the number of people suffering from heart and respiratory diseases etc. Therefore, air quality in Kyiv is monitored in order to minimize possible negative consequences.

2. Prediction using the SARIMA model

2.1. Problem Statement

PM (particulate matter) - small particles that are air pollutants (dust, dirt, smoke, etc.) distinguished by diameter (PM1, PM10, PM2.5). The Air Quality Index (AQI) provides information about air pollution [1]. Usually the AQI is calculated for indicator PM2.5, because, as stated in [3], it is the most dangerous pollutant. Therefore, PM2.5 can be considered a target.

Table 1: 3 rows of the data

logged at	pm25
2020-12-01 00:08:10	0.518428
2020-12-01 00:13:04	0.729244
2020-12-01 00:16:25	0.770710

The data is taken from [2]. The dataset contains values: phenomenon - the measured indicator; value (of the indicator); logged_at - the exact time when the measurements were taken. Let's try to identify any dependencies between PM2.5 feature and time. Table 1 shows the data after pre-processing.

Let's look on the average value of the PM2.5 feature by hour (Figure 1).The average at about 2-3 pm is the lowest, the highest value is at 7 am, another peak occurs at 10 pm. This is probably due to the increase in the number of cars during the hours when people go to or from work.

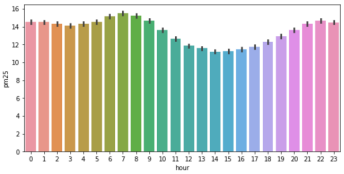


Figure 1: Average PM2.5 by hour

2.2. Time Series Research, Choosing and Fitting Model

We will consider the averaged values of the PM2.5 column for every 2 hours as a time series. We can find outliers by decomposing the series into trend, seasonality and error as those values that deviate significantly from the combined seasonal component and trend, i.e., there is a large error (with sigma rule).

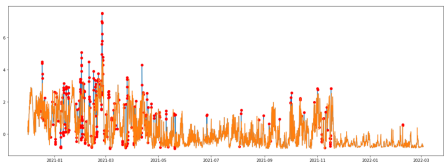


Figure 2: time series and outliers

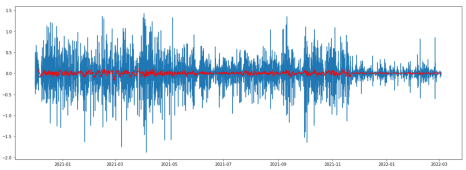


Figure 3: time series after the first differentiation

The figure 2 shows the points that were identified as outliers, the series before removing these points is colored blue, after removing - orange. Since there is a nonlinear trend, the series is not stationary. After trying to make it stationary by taking differences we get the series shown on Figure 3.

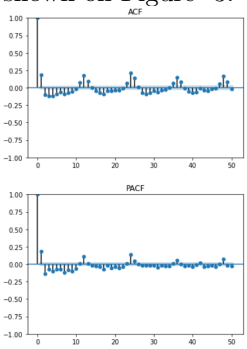


Figure 4: Values of ACF and PACF for the series after the first differentiation

Using ACF and PACF (Figure 4) we can see that there's a seasonality (a large correlation at the value of 12). This means that the data is correlated with what happened 24 hours ago. At 48, 72, 96 hours, the correlation decreases, but still remains quite high. So we need to get rid of seasonality.

For this series, the hypothesis in the Dickey-Fuller test is not rejected, so the series is indeed stationary.

One approach to time series forecasting is to use autoregressive and moving average models, as well as their modifications. The SARIMA (Seasonal autoregressive integrated moving average) model is used if there are both seasonality and trend in a time series.

So, we will use SARIMA ((p, 1, q)(P, 1, Q), 12) to predict the values of the series.

We will build the model on PM2.5 values until 2021-07-01 4pm and try to build a forecast for 30 hours ahead. We will search through the possible model parameters and determine the best ones using the Python and R built-in functions. The model defined in this way is SARIMA ((5,1,0), (2,1,0), 12). Let's build a forecast and display the predicted and real data (Figure 5).

The mean square error (MSE) is about 0.0967 on the test sample, and 0.1038 on the training sample. That is, the model made a fairly good prediction.

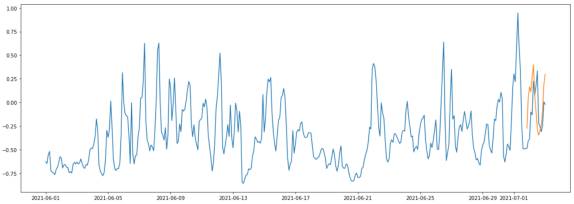


Figure 5: Prediction of PM2.5 for 30 hours

Conclusions

The data from one of the air quality monitoring stations in Kyiv was chosen. For this dataset the time series with the values of the concentration of PM2.5 pollutants in the atmosphere was analyzed. Based on this analysis, an autoregressive model was chosen to predict the values of this indicator in the future.

The SARIMA model was used to make a 30-hour forecast of PM2.5 in the atmosphere. The model made a good prediction, but there are other approaches and ways to improve the model (adaptive methods for building autoregressive models, neural networks, etc.)

References

1. *Beijing Air Pollution: Real-time Air Quality Index (AQI)*- `https://aqicn.org`,

3. *SaveEcoBot* - `https://www.saveecobot.com`,

3. *Undark - The Weight of Numbers: Air Pollution and PM2.5* -`https://undark.org/breathtaking`,