BALTIC-NORDIC-UKRAINIAN WORKSHOP ON SURVEY STATISTICS 2024

Poznań, 26 – 30 August 2024

BOOK OF ABSTRACTS

POZNAŃ UNIVERSITY OF ECONOMICS AND BUSINESS

Contents

Session 7 - Contributed presentations
Thursday 29 August, 11:15 – 12:45, Chair: Kristi Lehto

Session 8 - Invited lecture
Thursday 29 August, 14:15 – 15:00, Chair: Tomasz Żądło

Session 9 - Contributed presentations
Thursday 29 August, 15:30 – 17:00, Chair: Adeline Clarke

Session 10 - Keynote lecture
Friday 30 August, 09:00 – 09:55, Chair: Jelena Voronova

Session 11 - Invited lecture
Friday 30 August, 10:00 – 10:45, Chair: Dankmar Böhning

Session 12 - Contributed presentations
Friday 30 August, 11:15 – 12:15, Chair: Biruta Sloka

Voting-based predictor's selection in survey sampling

Wed. 28 Aug,
09:15 – 10:10

T. Żądło[1], A. Wolny-Dominiak[1]

[1]University of Economics in Katowice

We propose a simulation-based procedure to select the best predictor in survey sampling using various models, including machine learning. The proposed method is applicable to both cross-sectional and longitudinal data, enabling the incorporation of multiple user-selected parametric and nonparametric models to predict a vector of any population or subpopulation characteristics. It is presented in a real-world application, where the best predictor is chosen based on a simulation study considering six parametric and non-parametric models and estimating prediction accuracy based on three measures. Moreover, the procedure can be applied beyond survey sampling and small area estimation to any prediction problem, including time series analysis. An R package, allowing for easy implementation of our proposal, is currently under development.

Keywords: model-based approach, parametric and non-parametric models, prediction accuracy, predictor's selection, voting algorithms

Population estimation in Ukraine after Russia's full-scale invasion: Methods, data sources, and challenges

Wed. 28 Aug, 10:15 – 11:00

V. Sarioglo[1]

[1]Institute for Demography and Life Quality Problems of the National Academy of Science of Ukraine, Ukraine
e-mail: sarioglo@idss.org.ua

The crisis brought on by the February 2022 Russian full-scale invasion of Ukraine has led to large-scale displacement, both internally and across borders, of the Ukrainian population. In such conditions estimating the size and place of residence of the population has become a big problem. This problem was aggravated by the fact that official statistics did not have reliable data on population statistics before the war due to the lack of census data (the last census was conducted in 2001) and the absence of high-quality population registers in Ukraine. This has necessitated the development of relevant approaches for obtaining actual and timely population statistics to inform stakeholders, first of all the Government authorities, international organizations, and other users for humanitarian, social, economic, and community reconstruction efforts.

Together with the State Statistics Service of Ukraine (SSSU), Institute for Demography and Social Studies of the NAS of Ukraine, UNFPA (leading developer), and other international organizations an approach has been developed to estimate the number, structure and location of the population as of mid-2023 (Silva, R., and others 2023). In particular, it was established that there were about 31.7 million people in the Government Controlled Area of Ukraine.

The report examines sources of population data that exist in Ukraine and can be used to estimate the size, distribution and dynamics of the population. In particular, the following main data sources were used: the Ministry of Justice of Ukraine - on the number of registered deaths and births; the Public Health Center of the Ministry of Health of Ukraine - on the number of newborns; the Ministry of Social Policy of Ukraine - on the number and placement of registered IDPs; the International Organization for Migration - results of humanitarian surveys, first of all the General Population Survey, in Ukraine; the UN Refugee Agency (UNHCR) - on the number, demographic characteristics, and regions of origin of forced external migrants from Ukraine; the Pension Fund of Ukraine, the Ministry of Education and Science of Ukraine etc.

Special attention is paid to the use of data from mobile operators on the number of subscribers, and the results of a special population sample survey on the use of mobile communication for the purpose of estimating the population (Sarioglo, V., and Ogay, M. 2023). It is shown that population estimates based on mobile operator data for large cities in the pre-war period are, in general, more credible than official population statistics. Approaches to statistical estimation of the reliability of population indicators, and used verification procedures by administrative units are discussed. Examples of estimates of the size, distribution and dynamics of the population in 2022 and 2023 for administrative-territorial units of different levels are considered.

The report also discusses the main strengths, and limitations of the developed approach. At the same time, it is shown that the approach can be used to estimate and monitor the number and location of the population during the significant population movements due to military conflicts or environmental disasters.

Keywords: population estimation, administrative data, mobile operators' data, sample survey

References

Silva, R., Snyder, M., Han, M. D., Sariogolo, V., Libanova, E. (2023). Subnational population projections for humanitarian response in Ukraine: integration and cross-validation of traditional and non-traditional sources: Quetelet 2023 Seminar, (November 9–10, 2023). Leuven. Retrieved from `https://uclouvain.odoo.com/en_US/event/quetelet-2023-seminar-403/page/introduction-quetelet-2023-seminar`.

Sarioglo, V., Ogay, M. (2023). Approach to Population Estimation in Ukraine Using Mobile Operators' Data. *Statistics in Transition new series*, 24(1), 131-144.

Performance of Rao-Blackwellization-based estimators

Wed. 28 Aug,
11:30 – 13:00

T. Ianevych[1], O. Zalieska[1]

[1]Taras Shevchenko National University of Kyiv, Ukraine
e-mail: tetianayanevych@knu.ua, e.zalesskaya35@gmail.com

The accuracy of estimates is an important issue that determines whether they can be used in practice. Improving the quality of estimates can be achieved by reducing their bias. In this paper, we applied the leave-one-out Rao-Blackwellization (LOO-RB) estimation method (1) proposed by Sande & Zhang, (2021), combined with various machine learning (ML) models, to estimate the mean for different populations.

$$\hat{Y}_{RB} = \frac{1}{n}\sum_{j\in s}\left(\sum_{i\in s_1} y_i + \sum_{i\in U\setminus s_1}\mu(\mathbf{x}_i, s_1) + \frac{y_j - \mu(\mathbf{x}_j, s_1)}{\pi_{2j}}\right), \tag{1}$$

where U is a population, s is a sample, $|s| = n$, $s_1 = s \setminus \{j\}$, $\mu(\mathbf{x}_i, s_1)$ is the prediction of the model trained on s_1, $\pi_{2j} = P(j \in s_2|s_1)$, $s_2 = s\setminus s_1$, y_i is the value of the target characteristics for the i-th observation, $\mathbf{x}_i$ is the vector of auxiliary values for the i-th observation.

We analyze how essential is the bias reduction for LOO-RB estimators utilizing different ML-models. In particular, we generated population ($|U| = 1000$), in which observations form two clusters of different sizes. The Table 1 shows the results of the mean estimation for 500 Monte Carlo simulations based on samples ($|s| = 100$) obtained by the simple random sampling.

Table 1: Comparison of estimators with and without RB

Estimators	Bias	RMSE	RRMSE
HT	-0.0078	0.3328	3.1461
LREG	-0.0352	0.2754	2.6034
LOO-RB-LREG	-0.0159	0.2750	2.5997
SVM	-0.0758	0.2543	2.4040
LOO-RB-SVM	-0.0152	0.2471	2.3359
KNN	-0.0580	0.1538	1.4539
LOO-RB-KNN	-0.0101	0.1625	1.5362

Keywords: Rao-Blackwellization, bias, machine learning methods

References

L. S. Sande, L.-Ch. Zhang, (2021) Design-Unbiased Statistical Learning in Survey Sampling. *Sankhya A*, 83(2), 714-744.

Testing sampling method for validated transition between population estimation models

Wed. 28 Aug, 11:30 – 13:00

J. Voronova[1]

[1]Central statistical Bureau of Latvia

Since the last traditional census in Latvia held in 2011, the use of administrative data has been meeting the needs of the population census, estimating the population size of Latvian inhabitants. A logistic regression model, trained on 2011 census data, is currently employed, integrating various administrative data sources. Recognizing the finite lifecycle of any model and its potential degradation over time, as well as the extended availability and quality of administrative data, a methodology replacement mechanism has been developed. The Sollogit model, belonging to the class of unsupervised methods, has been tested and used for comparisons and analysis. To ensure the accuracy and reliability of this transition between methodologies, an audit pilot survey has been introduced. One of the aims of this survey is to measure and validate the precision of the applied methods, providing an approval-based transition from one model to another, and explaining model estimations. The presentation will cover the fundamental stages of developing the audit pilot survey plan and the applied approach, highlighting data collection issues and sharing obtained results along with lessons learned.

Keywords: model-based approach, logistic regression, population, census, pilot survey

Blocking records for probabilistic record linkage using approximate nearest neighbours algorithms

Wed. 28 Aug, 11:30 – 13:00

M. Beręsewicz[1,2]

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

[2]Statistical Office in Poznań, Poland

Blocking is a crucial aspect of probabilistic record linkage studies in official statistics. The objective of this procedure is to reduce the number of possible comparison pairs (cf. Steorts et al. 2014). This procedure assumes that blocking variables, such as sex, birth dates, or country of origin, are free of errors. However, in practice, such variables are often observed with typos or missing data.

The goal of this paper is to present a new approach to block records using approximate nearest neighbors (ANN) algorithms and graphs. The proposed blocking procedure has three primary objectives: (1) to significantly reduce the number of comparison pairs, (2) to account for the possibility that blocking variables may be measured with errors, and (3) to speed up the blocking procedure by employing advanced ANN algorithms.

The proposed method is based on the `rnndescent` (Melville 2024a), `RcppHNSW` (Melville 2024b), `RcppAnnoy` (Eddelbuettel 2024), and `mlpack` (Curtin et al. 2023) packages, which implement state-of-the-art ANN algorithms. The `igraph` (Csárdi & Nepusz, 2006, Csárdi et al. 2024) package is used for the creation of blocks. Moreover, the package facilitates straightforward integration with the `reclin2` (van der Laan, J. 2024) package and allows for the assessment of the quality of the blocking procedure.

Keywords: data integration, administrative data, official statistics

References

Curtin, R., Edel, M., Shrit, O., Agrawal, S., Basak, S., Balamuta, J., Birmingham, R., Dutt, K., Eddelbuettel, D., Garg, R., Jaiswal, S., Kaushik, A., Kim, S., Mukherjee, A., Sai, N., Sharma, N., Parihar, Y., Swain, R., Sanderson, C. (2023). mlpack 4: a fast, header-only C++ machine learning library. *Journal of Open Source Software*, 8(82).

Csárdi, G., Nepusz, T. (2006). The igraph software package for complex network research. *Complex Systems*, 1695, 1–9.

Csárdi, G., Nepusz, T., Traag, V., Horvát Sz, Zanini, F., Noom, D., Müller, K. (2024). igraph: Network Analysis and Visualization in R. `https://doi.org/10.5281/zenodo.7682609`. Retrieved from `https://CRAN.R-project.org/package=igraph`.

Eddelbuettel, D. (2024). RcppAnnoy: 'Rcpp' Bindings for 'Annoy', a Library for Approximate Nearest Neighbors [R package version 0.0.22]. Retrieved from `https://CRAN.R-project.org/package=RcppAnnoy`.

van der Laan, J. (2024). reclin2: Record Linkage Toolkit [R package version 0.5.0]. Retrieved from `https://CRAN.R-project.org/package=reclin2`.

Melville, J. (2024a). rnndescent: Nearest Neighbor Descent Method for Approximate Nearest Neighbors [R package version 0.1.6]. Retrieved from `https://CRAN.R-project.org/package=rnndescent`.

Melville, J. (2024b). RcppHNSW: 'Rcpp' Bindings for 'hnswlib', a Library for Approximate Nearest Neighbors [R package version 0.6.0]. Retrieved from `https://CRAN.R-project.org/package=RcppHNSW`.

Parihar, Y. S., Curtin, R., Eddelbuettel, D., Balamuta, J. (2024). mlpack: 'Rcpp' Integration for the 'mlpack' Library [R package version 4.3.0.1]. Retrieved from `https://CRAN.R-project.org/package=mlpack`.

Steorts, R. C., Ventura, S. L., Sadinle, M., Fienberg, S. E. (2014). A comparison of blocking methods for record linkage. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2014, Ibiza, Spain, September 17-19, 2014. Proceedings* (pp. 253-268). Springer International Publishing.

Analysing open-ended survey responses in Finnish

Wed. 28 Aug,
14:15 – 15:00

A. Clarke[1], M. Valaste[1]

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

A survey is a method of gathering information using relevant questions from a sample of entities with the aim of understanding populations as a whole. Surveys may use various methods to collect information from the respondents. Perhaps the most common is the use of the questionnaire, a standardized set of questions administered to the respondents in a survey. The questions are typically administered in a fixed order and often with fixed answer options. (Groves et al., 2009.) In addition to these closed questions, the survey has the option to include open-ended questions. Open-ended questions in survey research have a very long history and the value of open-ended questions have been rediscovered in survey research (Neuert et al., 2021, Singer & Couper, 2017).

Open-ended responses provide a unique opportunity to gather more in-depth and diverse information about respondents' views, experiences, and opinions. They allow respondents to express their thoughts in their own words, which can reveal new perspectives and themes that the researcher might not have anticipated (He & Schonlau, 2021). This can lead to a richer and more nuanced understanding of the phenomenon under study. Additionally, open-ended responses can highlight individual differences and diversity that might go unnoticed with closed-ended questions. This is particularly important in the study of complex and multidimensional phenomena such as attitudes, beliefs, and experiences.

Statistical software analyzes data as if the data were collected using simple random sampling (SRS). However, this is not always the case. When sampling method is not SRS, then we need to take into account the design that was used. The sampling design affects the calculation of the points estimates and standard errors.

There's limited support for conducting analysis on Finnish open-ended survey responses, so open-ended survey responses tend not to be utilized properly. Our aim is to build tools for text data that work with Finnish language with sufficient ease and to support explorative analysis of open- ended survey responses. For this purpose, we have created the R package `finnsurveytext`. In the next update of the package, we will include additional functionality which enables finnsurveytext to integrate with the popular `survey` package in R through the `svydesign` object furthering our package's useability in survey analysis and enabling analysis of open-ended questions to be better integrated with analysis of closed questions.

Keywords: open-ended questions, survey, text as data.

References

Clarke, A. P., Lagus, K., Laine, K. H., Litova, M., Nelimarkka, M., Oksanen, J., Peltonen, J., Oikarinen, T. S., Tirkkonen, J.-M., Toivanen, I., Valaste, M. (2024). finnsurveytext:

Analyse Open-Ended Survey Responses in Finnish. https://cran.r-project.org/web/packages/finnsurveytext/.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., Tourangeau, R. (2009). Survey methodology (Vol. 561). John Wiley & Sons.

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: a journal for quantitative methods and survey methodology (mda)*, 11(2), 115–134. https://doi.org/10.5281/zenodo.7682609.

He, Z., Schonlau, M. (2021). Coding Text Answers to Open-ended Questions: Human Coders and Statistical Learning Algorithms Make Similar Mistakes. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, 15(1), 103-120. https://doi.org/10.12758/mda.2020.10.

Capture-recapture methods and their applications: The case of one-inflation in zero-truncated count data

Thu. 29 Aug, 09:00 – 09:55

D. Böhning[1]

[1]University of Southampton

Estimating the size of a hard-to-count population is a challenging matter. We consider unilist approaches in which the count of identifications per unit is the basis of analysis. Unseen units have a zero count and do not occur in the sample leading to a zero-truncated setting. Because of various mechanisms, one-inflation is often an occurring phenomena that can lead to seriously biased estimates of population size. The talk will review some recent advances on one-inflation and zero-truncation modelling, and furthermore focuses here on the impact it has on population size estimation. The zero-truncated one-inflated and the one-inflated zero-truncated model is compared (also with the model ignoring one-inflation) in terms of Horvitz–Thompson estimation of population size. Simulation work shows clearly the biasing effect of ignoring one-inflation. Both models, the zero-truncated one-inflated and the one-inflated zero-truncated one, are suitable to model ongoing one-inflation. It is also important to choose an appropriate base-line distributional model. Considerable emphasis is allocated to a number of case studies which illustrate the issues and the impact of the work.

Evaluating the impact of a non-probability sample-based estimator in a linear combination with an estimator from a probability sample

Thu. 29 Aug,
10:00 – 10:45

V. Nekrašaitė-Liegė[1,2], A. Čiginas [1,3], 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

The proliferation of diverse data sources presents an opportunity to enhance the accuracy of indicators estimated in official statistics. This study explores the estimation of finite population parameters by combining data from both probability and non-probability samples, a method increasingly discussed in survey sampling literature (Beaumont, 2020; Kim and Tam, 2021; Rao, 2021). We focus on scenarios where the study variable is available in both samples or just in the non-probability sample.

When the study variable is available in both samples, we suggest using a composite estimator. 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.

When the study variable is available only in the non-probability sample, we explore several estimators and their variances. Our findings suggest that integrating non-probability samples with probability sample can lead to improved estimator efficiency, providing a pathway to more accurate and reliable official statistics.

Keywords: composite estimator, non-probability sample, variance estimation

References

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

Kim, J.-K. and Tam S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89(2), 382–401.

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 83(1), 242–272.

Combining online job advertisements with probability sample data for enhanced small area estimation of job vacancies

Thu. 29 Aug, 11:15 – 12:45

D. Ševinskas[1,2], A. Čiginas[1,2], I. Burakauskaitė[1,2]

[1]State Data Agency, Statistics Lithuania, Lithuania
e-mail: donatas.slevinskas@stat.gov.lt, andrius.ciginas@stat.gov.lt, ieva.burakauskaite@stat.gov.lt

[2]Vilnius University, Lithuania
email: donatas.slevinskas@mif.stud.vu.lt, andrius.ciginas@mif.vu.lt, ieva.burakauskaite@mif.vu.lt

Auxiliary information available in probability sample surveys is important in obtaining as accurate parameter estimates in the finite population and its domains as possible. Having auxiliary data related to the study variables at the unit or domain (area) level provides a range of models to choose from that can improve the direct design-based estimates. In this application, we combine the probability sample data on job vacancies with online job advertisements (OJA) information and administrative data to improve the estimates of job vacancy totals in small population domains.

One of the ways of integrating big data samples such as OJA is to stratify the population into big data stratum and a missing data stratum (Kim and Tam, 2021). Then, apply the calibration method with different imposed conditions on artificial strata to exploit big data sample as complete auxiliary information. However, typical calibration methods, though explicitly not stated, rely on unit-level linear relationship between the variables.

A model-calibration (MC) approach of (Wu and Sitter, 2001) allows more general underlying unit-level models. The MC approach to improving the direct probability sample-based estimates in small areas is based on the predictions of the study variable in the big data stratum through a measurement error model Predictions of study variable in big data stratum are further used in data integration calibration constraints. A variety of models can be utilised: linear regression, nonlinear parametric models or non-parametric models such as k-nearest neighbors or others.

The estimated totals through model-calibration and data integration approach are subsequently used in Fay-Herriot area-level model (Fay and Herriot, 1979) to obtain the empirical linear unbiased predictions (EBPLUPs) of domain totals.

The methodology proposed based on this application can be used in other estimation problems where incomplete additional information is available from administrative or alternative data sources.

Keywords: non-probability sample, nearest neighbor imputation, model-calibration, small area estimation, area-level

References

Deville, J. C., Särndal, C. E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418), 376–382.

Fay, R.E., Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a), 269–277.

Kim, J.-K., Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89(2), 382–401.

Rao, J.N.K., Molina, I. (2015). *Small Area Estimation.* 2nd edition, John Wiley & Sons, Inc., Hoboken, New Jersey.

Wu, C., Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453), 185–193.

Small area estimation of tourism indicators using online booking platform data

Thu. 29 Aug, 11:15 – 12:45

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

[1]State Data Agency, Statistics Lithuania, Lithuania
e-mail: ieva.burakauskaite@stat.gov.lt, andrius.ciginas@stat.gov.lt

[2]Vilnius University, Lithuania

State Data Agency (Statistics Lithuania) has been carrying out the quarterly census survey on accommodation services in order to collect statistical information for the short-stay accommodation tourism indicators, including the number of tourists and nights spent. However, the availability of data on listings and bookings from four major booking platforms ($Airbnb$, $Booking.com$, $Tripadvisor$, and $ExpediaGroup$) enables a shift to a sample-based survey and the reduction of the burden on respondents. We demonstrate how domain-level small area estimation modeling using auxiliary data from online booking platforms might lead to more accurate estimates in small domains, such as municipalities, compared to direct estimates. Therefore, the enhanced estimation process could help refine tourism statistics further.

Keywords: non-probability sample, small area estimation, tourism indicators, official statistics

Nowcasting Consumer Confidence Indicators Using Social Media and Google Trends Data

Thu. 29 Aug, 11:15 – 12:45

A. Vitkauskaitė[1,2], A. Čiginas[1,2]

[1]State Data Agency, Statistics Lithuania, Lithuania
e-mail: Akvile.Vitkauskaite@stat.gov.lt, Andrius.Ciginas@mif.vu.lt

[2]Vilnius University, Lithuania

The Consumer Confidence Index (CCI) measures public sentiment about the economy through a probability sample survey considering four questions regarding household finances, economic outlook, and spending plans over the past and coming years. It is calculated as a mean of these responses. The main objective of the study is to nowcast the CCI, estimating the current month's values faster than traditional survey methods, which usually provide results at the end of the month. We examine the relationship between traditional survey-based indicators and consumer sentiment expressed on social media. Social media expressions are collected from X (Twitter). The sentiment analysis of tweets enables us to create a Social Media Indicator (SMI), offering a distinct advantage in our predictive models. To improve forecast accuracy, we include Google Trends data, providing additional insights into consumer search behaviour and related trends in economic confidence. The study also explores integrating key economic indicators such as inflation rate, income statistics, and unemployment. In essence, by combining traditional economic indicators with advanced sentiment analysis from X and Google Trends data, the study seeks to deliver prompt CCI predictions ahead of standard survey timelines.

Keywords: consumer confidence, X, Google Trends, sentiment analysis, nowcasting

Clustered-based demographic typology of rural municipalities: The case of three Baltic States

Thu. 29 Aug, 14:15 – 15:00

A. Dahs[1]

[1]University of Latvia, Latvia
e-mail: aleksandrs.dahs@lu.lv

Implementation of an effective measurement and evaluation methodology is crucial for achieving regional development goals and monitoring their performance. Previous research from other European countries (e.g. see Stonawska and Vaishar, 2018 or Hasek, 2020) outlines multiple problems linked with the classification and assessment of territorial units according to their demographic, social or economic characteristics. Standard regional demographic classifications based on perceived level of urbanization does not reflect the important population development aspects. With this in mind, it is necessary to establish a data-driven technique for categorization of territorial units based on their demographic characteristics that can be applied in drafting policy measures (Malinen et al, 1994). Proposed methodology relies on unsupervised non-hierarchical partitioning clustering algorithm. The study focuses on predominantly rural municipalities of the three Baltic countries – Estonia, Latvia and Lithuania.

This study aims to provide an update to the demographic typology of predominantly rural municipalities of the three Baltic States and to discuss main differences between the identified municipality groups. Case of Latvia is reviewed in greater detail, using latest population census and survey data. Detailed review of the municipalities included in specific clusters helps to better understand internal regional processes and factors shaping local demographic trends. Provided clustering examples confirm that such analytical approach can be beneficial for monitoring regional population developments and evaluating the efficiency of policy responses.

References

Hasek, O., (2020). Regionální diferenciace plodnosti podle typologie venkova. (The Regional Differentiation of Fertility by Rural Typology in Czechia). *Demografie*, 62, 3–13.

Malinen, P., Keränen, R., Keränen, H., (1994). *Rural area typology in Finland - a tool for rural policy*. University of Oulu, Research Institute of Northern Finland Research Reports 123. Available at: `https://jyu.finna.fi/Record/jykdok.487435?lng=en-gb`.

Stonawska, K., Vaishar, A., (2018). Differentiation and Typology of the Moravian Countryside. *European Countryside*, 10(1), 127–140.

Mixed-mode census 2021 survey with voluntary part in Estonia

Thu. 29 Aug, 15:30 – 17:00

K. Lehto[1]

[1]Statistics Estonia
e-mail: kristi.lehto@stat.ee

The 2021 census in Estonia was mostly based on administrative data. All EU-mandatory characteristics were collected from registers (Statistics Estonia, 2022). However, the purpose of the sample survey was to collect information on persons living in Estonia that is not available in the registers (religious affiliation, knowledge of languages and dialects, existence of a long-term illness or health problem and health-related limitations on daily activities).

In 2011 census, when Estonia used the first time of self-enumeration (CAWI - Computer Assisted Web Interview) the response rate of CAWI was 67%. This gave an idea to use the voluntary CAWI part in 2021 census survey. The population and housing census survey was mandatory for sample persons according to the law. In CAWI mode all those who wished could respond voluntarily even outside the sample. CAWI response rate was 43,1% which is very high concerning that answering was voluntary (Statistics Estonia, 2022 November).

For weighting, it was important to keep in mind that CAWI respondents are different from CATI/CAPI respondents. They are younger, healthier, less religious and know more foreign languages based on Census 2011. In order to obtain the unbiased estimates, it was necessary to combine the data of different modes. The calculation of the design weights and the calibration of the weights were done separately in two parts of the population. All households with at least one CAWI respondent belongs to the CAWI population, the rest of the people belong to the CATI/CAPI population. Households that were originally included in the random sample, but responded via CAWI, are part of the general population of CAWI. The population of the CATI/CAPI is described by people who were randomly sampled and answered to the interviewers (CATI/CAPI). To mix probability and non-probability (voluntary) sample helped to improve quality of the census estimates and publish more detailed breakdowns.

Keywords: census survey, mixed-mode, survey with voluntary part

References

Statistics Estonia (2022). *Population and Housing Census. Methodology*. `https://rahvaloendus.ee/en/census-2021/methodology`.

Statistics Estonia (2022, November). *Description of the sample survey methodology for the 2021 census*. `https://www.stat.ee/sites/default/files/2022-11/Loenduse%20valikuuringu%20metoodika%20raport.pdf` (only in Estonia).

Teaching of survey statistics - Needs and challenges of students in business and economic studies

Thu. 29 Aug, 15:30 – 17:00

B. Sloka[1]

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

All students at Economics and Business study programs have to use knowledge and skills on survey creation, data collection and survey data analysis as they need to include results of their empirical research on their final papers: bachelor theses, master theses and doctoral dissertations. For those empirical studies knowledge and skills on creation of questionnaire for the survey, testing the questionnaire, pilot survey, final preparation of the questionnaire, organization of questionnaire on survey platform (mostly used *QuestionPro* as University of Latvia provide individual licenses for students for *QuestionPro* use for their survey). Every student has to apply individually to Study Department, than the license can be used only for research at University of Latvia and not for commercial needs is provided on individual basis for each student. The individual applications are very good as every student has to think about the survey organization on a timely manner and license cannot be obtained on the last moment. It is good to recognize that students use this opportunity and obtained survey data export to SPSS where wide data analysis tools are available and practically used by students: analysis of descriptive statistics, cross-tabulations, comparing means using t-tests, analysis of variance – ANOVA, correlation analysis, regression analysis, factor analysis and other statistical methods. Usually during the preparation of final graduation paper many additional consultations are requested and additional methodological materials and literature sources are needed and used (Association of Statisticians of Latvia, 2024; Bryman, 2012; Sapsford, 2007, Lohr, 2019; Greenlaw and Brown-Welty, 2009) as well as recommended literature on survey organization reflected in *Sage Research Methods* platform (Sage Research Methods, 2024) is available for each student with University of Latvia login and password. Very useful information and data are available on Household and Finance Survey realized by National Banks under supervision of European Central Bank (Bank of Latvia, 2024) as well as different surveys organized by Central Statistical Bureau: Labor Force Survey, EU-SILC, etc. (OSP, 2024) where information about the surveys and also anonymized data are available for researchers and student use.

Keywords: questionnaire design, testing of questionnaire, survey platform QuestionPro, survey data analysis

References

Association of Statisticians of Latvia (2024). Readings (Lasījumi). Aailable at `https://www.statistikuasociacija.lv/category/lasijumi/` [accessed 17.06.2024].

Bank of Latvia (2024). *Household and Finance Consumption Survey*. Available `https://www.bank.lv/en/statistics/stat-data/hfcs` [accessed 20.06.2024].

Bryman, A. (2012). *Social Research Methods*, 4th edit, Oxford University Press, 7p. 66.

Greenlaw C. and Brown-Welty S. (2009). A Comparison of Web-Based and Paper-Based Survey Methods: Testing Assumptions of Survey Mode and Response Cost. *Evaluation Review*, 33, 464–480.

Lohr, S.L. (2019). *Sampling: Design and Analysis*. 2nd edit., Boca Raton, CRC Press, p. 596.

Official Statistics Portal of Republic of Latvia (2024). *Data for research*. Available at `https://stat.gov.lv/lv/petniecibai` [accessed 20.06.2024].

Sage Research Methods (2024). Database – platform available for students and staff of University of Latvia (with registration indicating student/staff ID).

Sapsford, R. (2007). *Survey Research*, 2nd edit., Sage Publications, p. 276.

Methods of estimating loss reserves based on data with outliers

Thu. 29 Aug, 15:30 – 17:00

O. Vasylyk[1], V. Shunder[1]

[1]National Technical University of Ukraine, Ukraine
e-mail: vasylyk@matan.kpi.ua, 1618422a@gmail.com

Actuarial calculations for various insurance products are the basis for ensuring the solvency of the insurance company. The impact of outliers on loss reserving estimates is a very serious problem. Outliers in insurance are not data errors but large financial claims. Even if such events occur very rarely, the actuary cannot ignore outliers in the data on which a forecast is based, as this may lead to a false estimate of the required insurance reserves and subsequent insolvency of the insurance company. In particular, this applies to the estimation of reserves for incurred but not reported claims (IBNR). Another problem occurs when for some periods there are no claims, that is, we face missing data.

We consider various calculation methods for loss reserves such as the chain ladder method, the Bornhuetter-Ferguson method, the Cape Cod method and others, analyse the advantages and disadvantages of these methods, and give examples of their use in the case of data with outliers. The problem of robustification of the chain ladder method is solved both in the case of too large payments and in the case of absent insurance payments (missing data). Also we consider application of the bootstrap method in loss reserving.

Keywords: bootstrap method, Bornhuetter-Ferguson method, Cape Cod method, chain-ladder method, loss reserving, outliers, robust estimation

References

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.

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.

Jeng, H. (2010). On Small Samples and the Use of Robust Estimators in Loss Reserving. In *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). https://doi.org/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). https://doi.org/10.22495/jgr_v5_i1_p9.

Data integration and population size estimation: The risk of misidentification

Fri. 30 Aug, 09:00 – 09:55

T. Tuoto[1]

[1]ISTAT

Data integration and data linkage are valuable statistical tools for combining potentially noisy data from different sources in the absence of a unique identifier, both to identify duplicate information and to increase the informative content of each single data source. They are widely used by survey statisticians in many fields and are a prerequisite for estimating population sizes, both in the case of single and multiple lists. However, when unique identifiers are not available the integration procedures are prone to errors, and uncertainty due to linkage errors propagates in the downstream analyses. Compared with other statistical models, the estimation of population size is extremely sensitive to linkage errors, and the usual approach of minimising false links is not useful here, since missed matches introduce serious bias into population size estimates. The talk will examine the effect of linkage errors in the context of population size estimation, considering both the cases of multiple list and single lists. Linkage error adjusted estimators, from the frequentist and Bayesian perspective, will be discussed. The advantages and limitations of the proposals will be reviewed, highlighting potential uses in official statistics.

Recognizing duplicate entities in multimodal data

Fri. 30 Aug,
10:00 – 10:45

P. Kałużny[1]

[1]Allegro sp. z o.o.
e-mail: piotr.kaluzny@allegro.com, kaluzny.piotr.it@gmail.com

Recognizing duplicate entities, particularly in the context of e-commerce, is essential for accurately comparing products, detecting duplicates and crucial for maintaining data quality. This task involves comparing data representations of two objects based on probabilistic measures to identify duplicate entities. The topic has been long present in the scientific literature from databases and financial systems to managing CRMs. In e-commerce, it ensures accurate product listings, preventing duplicate entries[1] and ultimately improving customer experience. In statistics, accurate entity matching is vital for linking datasets and drawing reliable conclusions from multiple sources. Other domains, such as healthcare and finance, also benefit from precise entity matching for unify patient records or detecting fraudulent transactions.

Entity matching is inherently complex due to the diverse nature of data types involved, but essential for maintaining data integrity and driving informed decision-making in various domains. Records describing entities can include text, images, and tabular data, each of which presents unique challenges. While there are available tools (e.g. PRLT[2]) their use is limited to only the simplest cases of record linkage in databases. Text data, for instance, can contain variations such as typos, synonyms, and different formatting styles, making exact matches difficult to identify. Addition of images introduces the need for sophisticated algorithms to detect visual similarities, which can be complicated by variations in lighting, angles, and resolutions. Even tabular data adds another layer of complexity, particularly when dealing with missing values, inconsistent formats, and varying units of measurement. There exists a multitude of research in the field exemplified by ready to use frameworks (Köpcke and Rahm, 2010) but the advancements of AI systems have enabled significant extensions in matching multi-modal data consisting of structured and unstructured text and images with both supervised and unsupervised approaches, which differ whether human-annotated training data is needed to find an entity matching strategy. First problem of entity matching size of the data space for possible comparisons. With databases spanning millions of entities pre-ranking or filtering potential duplicate candidates is needed (Li et al., 2021). The multi-modality and volume of data makes pairing every entity computationally costly. This can be solved by utilizing a combination of classification models, unsupervised clustering techniques like DBScan, and various distance-based algorithms to optimize product matching to a subset smaller than the dot product of entities.

Entity matching across different data modalities requires tailored methods to effectively handle the unique characteristics of each type. For **tabular and short text data**, distance metrics such as Jaro-Winkler, Levenshtein, cosine distance, and affine gap can be used to measure similarity, especially when dealing with typos or slight variations. For more unstructured text preprocessing

[1]See e.g. `https://www.kaggle.com/c/shopee-product-matching/`
[2]`https://recordlinkage.readthedocs.io/`

steps like splitting text into words, removing duplicates and stopwords, and applying TF-IDF weighting are essential for enhancing the accuracy of these comparisons. Tools utilizing fuzzy matching like HMNI[3] can further aid in handling misspellings.

For unstructured **long text**, more sophisticated methods are required. Keyword extraction and Named Entity Recognition (NER) can be used to identify and compare key attributes within unstructured descriptions. Techniques such as TF-IDF with optimized sparse matrix multiplication, followed by top-n result selection, are effective for handling large feature vectors. While manual extraction of parameters for comparison is possible, alternatives utilising neural networks can also be applied (Barlaug and Gulla, 2021). Embeddings like word2vec or the use of BERT (Li et al., 2021) can capture the semantic meaning of text, enabling accurate entity matching even when the text is lengthy and complex. Often data augmentation is used to improve the process.

When dealing with **image data**, methods such as optical character recognition (OCR) can extract image-embedded text for further analysis. One of the simplest ways however to compare images is perceptual hashing with the chosen underlying resolution at a level needed for the specific use case (Hadmi et al., 2012; Monga and Evans, 2006). Perceptual hashing is useful for identifying visually similar images by converting them into comparable hash values and allows distance measures to be used. For handling multiple angles techniques like SIFT (Scale-Invariant Feature Transform) help in identifying similarities based on distinctive features that might provide more robustness to the object rotation and position in the image at the cost of computational complexity (Wu et al., 2013). Advanced techniques like object detection and image captioning (Ghandi et al., 2023; Xu et al., 2023) (image-to-text) algorithms can be employed to extract meaningful content from images while approaches like Local Sensitive Hashing (LSH) provide faster but less precise results that might be used in the pairs preprocessing and both can help in processing multimodal data (Nguyen et al., 2024). Moreover, employing new deep learning models such as detectron2 (Wu et al., 2019) for object detection can further refine the final result. Similarly, utilizing models like recently published Florence-2 (Xiao et al., 2024) might aid at providing a unified interface for usefulness from a multitude of vision tasks.

Important topic of entity recognition is deciding two objects are the same. With the wealth of metrics it might be useful to use optimization to select only the ones that best describe the relevant similarities in the data, akin to feature selection (Wang et al., 2011). Similarly, cross-domain findings on entity similarity calculation used in recommendation systems[4] can be considered. Designing approaches aimed at particular modalities is possible but with multiple metrics might result in a very complex data model. To avoid that, **multi-modal embedding algorithms like contrastive language image pretraining (CLIP) encoders** (Radford et al., 2021) can be used to reduce complexity of entity linkage systems. Those embedding bridge the gap between text and image modalities by creating joint embeddings, especially helpful in dealing with missing records for datasets with significant sparsity. As they allow for the comparison of images and textual descriptions within the same semantic space, they make it easier to identify matches across different types of data. Due to single representation, they can efficiently be

[3] `github.com/Christopher-Thornton/hmni`

[4] `ghostday.pl/wp-content/uploads/2024/04/Chrabrowa-Dense-Retrieval-for-Allegro-Search-Engine.pdf`

compared using methods like k-nearest neighbors, enabling highly accurate identification of duplicate or similar entities across various data types. Different architectures can also play a role in design of these systems. Siamese networks, particularly those utilizing triplet-loss functions or approaches like "Efficient Learning based Record Deduplication" (Ravikanth et al., 2024) offer approaches to learn the rules for entity matching in an supervised but more robust manner. Given labeled data in training on pairs of similar and dissimilar products, a siamese network learns to produce embeddings that closely align for items within the same class while diverging for items from different classes. These varied approaches ensure **that entity matching can be effectively applied across different data types in an unified manner**.

Keywords: multi-modal data, product matching, record matching, entity matching, duplicate detection, entity linkage, entity resolution, product similarity, product variants, e-commerce, LLM, computer vision, perceptual hashing, NLP, machine learning

References

Barlaug, N., Gulla, J. A. (2021). Neural networks for entity matching: A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15 (3), 1–37

Ghandi, T., Pourreza, H., Mahyar, H. (2023). Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56 (3), 1–39.

Hadmi, A., Ouahman, A. A., Said, B. A. E., Puech, W. (2012). Perceptual image hashing.

Köpcke, H., Rahm, E. (2010). Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, 69 (2), 197–210.

Li, Y., Li, J., Suhara, Y., Wang, J., Hirota, W., Tan, W.-C. (2021). Deep entity matching: Challenges and opportunities. *J. Data and Information Quality*, 13 (1). `https://doi.org/10.1145/3431816`.

Monga, V., Evans, B. L. (2006). Perceptual image hashing via feature points: Performance evaluation and tradeoffs. *IEEE transactions on Image Processing*, 15 (11), 3452–3465.

Nguyen, T., Gadre, S. Y., Ilharco, G., Oh, S., Schmidt, L. (2024). Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763.

Ravikanth, M., Korra, S., Mamidisetti, G., Goutham, M., Bhaskar, T. (2024). An efficient learning based approach for automatic record deduplication with benchmark datasets. *Scientific Reports*, 14 (1), 16254.

Wang, J., Li, G., Yu, J. X., Feng, J. (2011). Entity matching: How similar is similar. *Proceedings of the VLDB Endowment*, 4 (10), 622–633.

Wu, J., Cui, Z., Sheng, V. S., Zhao, P., Su, D., Gong, S. (2013). A comparative study of sift and its variants. *Measurement science review*, 13 (3), 122–131.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., Girshick, R. (2019). Detectron2. `https://github.com/facebookresearch/detectron2`.

Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L. (2024). Florence-2: Advancing a unified representation for a variety of vision tasks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4818–4829.

Xu, L., Tang, Q., Lv, J., Zheng, B., Zeng, X., Li, W. (2023). Deep image captioning: A review of methods, trends and future challenges. *Neurocomputing*, 546, 126287.

Coherence coefficient and its applications

Fri. 30 Aug,
11:15 – 12:15

D. Krapavickaitė

One of the quality requirements in official statistics is coherence of statistical information across domains, in time, in national accounts, and internally. However, no measure of its strength is used. The concept of coherence is also met in signal processing, wave physics, and time series. The definition of the coherence coefficient for a weakly stationary time series will be recalled and discussed in the presentation. The coherence coefficient is a correlation coefficient between two indicators in time indexed by the same frequency components of their Fourier transforms and shows a degree of synchronicity between the time series for each frequency. The application of this coefficient is illustrated through the coherence and Granger causality analysis of a collection of numerical economic and social statistical indicators. The coherence coefficient matrix-based non-metric multidimensional scaling for visualization of the time series in the frequency domain will be the second example. The aim of the talk is to propose the use of this coherence coefficient and its applications in official statistics.

Keywords: time series, periodogram, cross-covariance, Granger causality, multidimensional scaling

References

Borg, I., Groenen, P. J. F. (2005). *M*odern Multidimensional Scaling. Springer: Berlin/ Heidelberg, Germany.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 424–438.

Krapavickaitė, D. (2022). Coherence Coefficient for Official Statistics. *Mathematics* 10(7): 1159. `https://doi.org/10.3390/math10071159`.

Wei, W. W. S. (2006). *Time Series Analysis: Univariate and Multivariate Methods*. Pearson Education: London, UK.

Evaluation of some optimal characteristics of a metro station by queueing modelling

Fri. 30 Aug, 11:15 – 12:15

M. Tkach [1], H. Livinska[1]

[1]Taras Shevchenko National University of Kyiv, Ukraine
e-mail: mashatkach1@gmail.com, hanna.livinska@knu.ua

It is impossible to overestimate the work of the metro in a big city. First of all, the safety of passenger transportation is important, a comfortable and safe stay of passengers at the metro station is also an important factor, which in turn depends on the waiting time for the train. The main goal of the work is to study passenger flow at a metro station in Kyiv and to find the optimal interval between trains by using the tools of queueing theory. This problem has a practical application in real life and greatly affects the efficiency of passenger movement.

Our approach is based on a model of tandem queues for the metro station and a model for estimating trains loading. One of the queueing systems in our tandem model is the checkpoints in the station, and another is the station itself. Providing such a model unable us to solve a lot of problems. This gave us the opportunity to point out the condition of the overloaded regime of system nodes, number of service devices, calculate the number of passengers who can be at the station at the same time (maximum and safe-comfortable) and the optimal (safe and/or comfortable) interval between trains. Due to the flexibility of our model, all the performance measures of the system can be easily recalculated if the rates of the passenger flow or other system parameters are changed. Simulation approach using the R programming language enable us to estimate the performance measures under different intensity rate of the passenger flow and station load. This can contribute to improving the safe capacity of the metro station, reducing the waiting time of passengers and, accordingly, reducing traffic jams.

So, first, given estimated passengers flow into the station, we calculate the optimal number of devices at the subway checkpoints, so that with this minimum number, the queue of passengers would not accumulate. That is, to make passenger service comfortable. Since the metro station under consideration already has a set number of validators, we can compare how the overall experience of staying at the station would change by simulating the flow of passengers with different numbers of devices.

Next important performance measure, that can be calculated, is the interval between trains. It means an interval such that the number of passengers, who will arrive at the platform within it, does not exceed the average number of people that can comfortably fit on the train. To model train loading, we introduce two additional functions. The first is a load factor of the train. This coefficient depends on the time of day and on the intensity of passenger traffic corresponding to this time of day, as well as on the train waiting time tau. The second factor is the percentage of passengers who get off the carriages at the station. This value depends on the time of day and on a specific station, and in our case, it is a constant value. The simulation of passenger flow on a platform and the function of carriage load under input rate at the station equal 36 is shown on Fig. 1. Under given values of the parameters of the system the optimal interval between trains is for $\delta = 36$ passengers/min: $\tau <= 2.51294$.

It should be noted that the proposed model can be complicated by the introduction of additional transitions between stations, the introduction of different types of passengers into the system, the introduction of the time dependence of the input flow intensity function, etc.

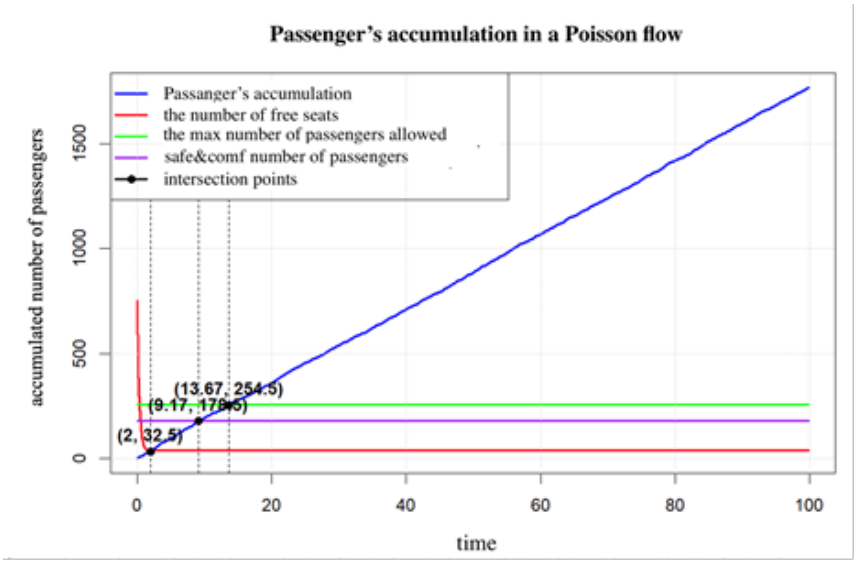

Figure 1: Passenger's accumulation in a Poisson flow

Keywords: queueing system, system in tandem, passengers flow

References

Ng, C. H., & Boon-Hee, S. (2008). *Queueing modelling fundamentals: With applications in communication networks*. John Wiley & Sons, p. 271.

Chen, S., Di, Y., Liu, S., & Wang, B. (2017). Modelling and analysis on emergency evacuation from metro stations. *Mathematical Problems in Engineering*, 2017(1), 2623684.

Xu, X. Y., Liu, J., Li, H. Y., & Hu, J. Q. (2014). Analysis of subway station capacity with the use of queueing theory. *Transportation research part C: emerging technologies*, p. 28–43.