Volume 26, Number 2, June 2025 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

ISSN 1234-7655

Address for correspondence
Statistics Poland, al. Niepodległości 208, 00-925 Warsaw, Poland, tel./fax: +48 22 – 825 03 95

CONTENTS

STATISTICS IN TRANSITION new series, *June 2025*
Vol. 26, No. 2, pp. III

Submission information for Authors

Statistics in Transition new series (SiTns) is an international journal published jointly by the Polish Statistical Association (PTS) and Statistics Poland, on a quarterly basis (during 1993–2006 it was issued twice and since 2006 three times a year). Also, it has extended its scope of interest beyond its originally primary focus on statistical issues pertinent to transition from centrally planned to a market-oriented economy through embracing questions related to systemic transformations of and within the national statistical systems, world-wide.

The SiT*ns* seeks contributors that address the full range of problems involved in data production, data dissemination and utilization, providing international community of statisticians and users – including researchers, teachers, policy makers and the general public – with a platform for exchange of ideas and for sharing best practices in all areas of the development of statistics.

Accordingly, articles dealing with any topics of statistics and its advancement – as either a scientific domain (new research and data analysis methods) or as a domain of informational infrastructure of the economy, society and the state – are appropriate for *Statistics in Transition new series*.

Demonstration of the role played by statistical research and data in economic growth and social progress (both locally and globally), including better-informed decisions and greater participation of citizens, are of particular interest.

Each paper submitted by prospective authors are peer reviewed by internationally recognized experts, who are guided in their decisions about the publication by criteria of originality and overall quality, including its content and form, and of potential interest to readers (esp. professionals).

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl,
GUS/Statistics Poland,
Al. Niepodległości 208, R. 296, 00-925 Warsaw, Poland

It is assumed, that the submitted manuscript has not been published previously and that it is not under review elsewhere. It should include an abstract (of not more than 1600 characters, including spaces). Inquiries concerning the submitted manuscript, its current status etc., should be directed to the Editor by email, address above, or w.okrasa@stat.gov.pl.

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: https://sit.stat.gov.pl/ForAuthors.

STATISTICS IN TRANSITION new series, *June 2025*
Vol. 26, No. 2, pp. V–VI

Policy Statement

The broad objective of *Statistics in Transition new series* is to advance the statistical and associated methods used primarily by statistical agencies and other research institutions. To meet that objective, the journal encompasses a wide range of topics in statistical design and analysis, including survey methodology and survey sampling, census methodology, statistical uses of administrative data sources, estimation methods, economic and demographic studies, and novel methods of analysis of socio-economic and population data. With its focus on innovative methods that address practical problems, the journal favours papers that report new methods accompanied by real-life applications. Authoritative review papers on important problems faced by statisticians in agencies and academia also fall within the journal's scope.

***

Abstracting and Indexing Databases

Statistics in Transition new series is currently covered in:

BASE – Bielefeld Academic Search Engine

CEEOL – Central and Eastern European Online Library

CEJSH (The Central European Journal of Social Sciences and Humanities)

CNKI Scholar (China National Knowledge Infrastructure)

CNPIEC – cnpLINKer

CORE

Current Index to Statistics

Dimensions

DOAJ (Directory of Open Access Journals)

EconPapers

EconStore

Electronic Journals Library

Elsevier – Scopus

Genamics JournalSeek

Google Scholar

Index Copernicus

J-Gate

JournalGuide

JournalTOCs

Keepers Registry

MIAR

Microsoft Academic

OpenAIRE

ProQuest – Summon

Publons

QOAM (Quality Open Access Market)

ReadCube

RePec

SCImago Journal & Country Rank

TDNet

Technische Informationsbibliothek (TIB) – German National Library of Science and Technology

Ulrichsweb & Ulrich's Periodicals Directory

WanFang Data

WorldCat (OCLC)

Zenodo

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. VII–X

From the Editor

The June issue of *Statistics in Transition new series* contains 11 articles by 32 authors from seven countries, namely (in order of appearance): the USA, Poland, India, Egypt, Libya, Indonesia and Nigeria. I am convinced that they will be excellent reading for both theoretically- and practically-oriented experts and enthusiasts of several fields of statistics – just right for the summer holidays.

Invited papers

In the first paper, *Thresholding nonprobability units in combined data for efficient domain estimation,* **Terrance D. Savitsky, Matthew R. Williams, Vladislav Beresovsky** and **Julie Gershunskaya** propose a new method for thresholding or excluding convenience units to minimise the variance of the resulting survey-weighted domain estimator. The authors compare their thresholding method with other such constructions in a simulation study for two classes of datasets based on the degree of the overlap between survey and convenience samples on covariate support. The research results show that excluding convenience units that each express a low probability of appearing in both reference and convenience samples reduces the risk of the occurrence of an estimation error.

Original Research papers

Andrzej Młodak and **Tomasz Józefowski's** article ***Application of Statistical Disclosure Control methods to protect the confidentiality of the 2020 agricultural census microdata*** presents the attempt to develop an efficient disclosure control algorithm for microdata in a statistical portal used for releasing detailed statistical information at various levels of spatial aggregation. The proposed algorithm is based on perturbative methods, such as microaggregation with Gower's distance for categorical variables and the addition of correlated noise for continuous variables. The proposed algorithm can be used to assess the loss of information by measuring distribution disturbances (based on a complex distance that accounts for all measurement scales) and the impact of the Statistical Disclosure Control (SDC) on the strength of correlations between variables (for continuous variables).

Formulation of estimator for population mean in stratified successive sampling using memory-based information by **Sanjoy Majumder**, **Arnab Bandyopadhyay** and **Arindam Gupta** discusses the development of a memory-type estimator for a popula-

tion mean in stratified successive sampling. The authors use the past sample information together with the current sample information through hybrid, exponentially-weighted moving-averages statistics, and employ the information available on auxiliary variable to construct the proposed estimator. The performance of this estimator is compared with a conventional estimator of a population mean. The results are obtained using both the simulated as well as population survey data. In the light of the findings of the study, the proposed estimator is recommended for survey statisticians for solving real-life research problems.

In ***Inference of dynamic weighted cumulative residual (DWCR) entropy for Burr XII distribution based on progressive censoring*** by **Amal S. Hassan**, **E. A. Elsherpieny** and **Wesal E. Aghel,** the authors introduce the DWCR entropy performing as an additional measure of uncertainty related to the residual lifetime function in several disciplines, including survival analysis and reliability. In this work, the DWCR formula is based on Havarda and Charvat and uses progressive Type II censoring to investigate the implications of the DWCR Tsallis entropy (DWCRTE), DWCR Rényi entropy (DWCRRE), and DWCRHCE for the Burr XII distribution. Both classical and Bayesian methods are used to derive the estimators of these entropy metrics. Assuming independent gamma priors, the authors obtain the Bayes estimator of the suggested measures. Due to the lack of explicit forms, the Metropolis-Hastings approach is used to determine the Bayes estimates for symmetric and asymmetric loss functions. To assess the efficacy of the suggested estimating techniques, several simulations are run for different censoring schemes.

The paper ***Synchronization and similarity between regional and sectoral output gaps in the Polish manufacturing industry*** by **Mirosław Błażej, Mariusz Górajski** and **Magdalena Ulrichs** presents a new micro-econometric model to calculate regional and sectoral output gaps in Poland's manufacturing industry in the years 2009–2020. The model was estimated using official data sources, which included business tendency surveys and annual enterprise activity reports. The authors assessed the synchronization and similarities between the sectoral and regional output gaps in Poland's manufacturing industry. They also analyzed the impact of the COVID-19 pandemic on the coherence of the regional and sectoral output gaps. The study contributes to the unification of the method for estimating and assessing the levels of unobservable potential production.

Sapriadi Rasyid, **Siswanto Siswanto** and **Sitti Sahriman's** article, ***Clustering based on poverty indicator data using K-Means cluster with Density-Based Spatial Clustering of Application with Noise*** describes a statistical method to cluster people affected by poverty on the basis of error indicators for each region, serving as a reference for providing assistance. The cluster analysis shown to be appropriate due to minimizing object differences within one cluster and maximizing them between clusters. The

study employs two methods, namely K-Means and Density-Based Spatial Clustering of Application with Noise (DBSCAN), to compare their effectiveness based on the Silhouette Coefficient. The data used for the analysis comes from eight poverty indicators for the South Sulawesi province in 2022. The study demonstrates that the K-Means method is more effective than the DBSCAN in helping the government to group the poverty characteristics of each region.

In **Mbanefo S. Madukaife**, **Uchenna C. Nduka** and **Everestus O. Ossai's** paper ***Testing for multinormality with goodness-of-fit tests based on phi divergence measures***, the authors use the phi divergence measure, DΦ(F,G), between F and G distributions to obtain a goodness-of-fit test to multivariate normality (MVN) based on the theoretical density function of the beta transformed random variable and a window-size spacing-sample density function. Three versions of the statistic are derived from the three known phi divergence measures that are based on the sum of squares. The empirical critical values of the statistics are obtained and the empirical type-one-error rates as well as powers of the statistics in comparison with those of other well-known competing statistics are computed through extensive simulation study. The study shows that the new statistics have good control over type-one-error and are highly competitive with the existing well-known ones in terms of power performance.

Marta Marszałek's article ***Households' invisible input to the economy: a review of its measurement methods and results*** discusses the unpaid domestic work as the main part of the non-market household production, which is not covered by official statistics (GDP). The monetary value of unpaid work is identified within the gross value added (GVA), which is 60-80% of (the invisible) non-market household production. This paper shows that different wages used in input methods do not change the final proportion of the gross value added GVA of unpaid work to total household production. The analysis confirms that a regular implementation of the Household Satellite Accounts (HHSA) jointly with the core system – the European System of Accounts (ESA) – is a valuable tool for assessing the total output of household production.

In the paper ***Ratio regression type estimators of the population mean for missing data in sample surveys*** by **Prachi Garg, Namita Srivastava** and **Manoj Kumar Srivastava,** new ratio regression-type estimators with imputation are proposed as a means to overcome the problem of missing data for a studied variable in a survey. It is shown that the suggusted estimators are more efficient than the mean method of imputation, the ratio method of imputation, the regression method of imputation, or several other estimators. The authors derive the bias and the mean square errors of the suggested estimators and conduct a comprarative study using real and simulated data. The results seem tobe improvement as compared to all the methods discussed.

Conference papers

XV SCIENTIFIC CONFERENCE

MASEP 2024 -Measurement and Assessment of Social and Economic Phenomena, Warsaw, Poland

In the paper ***Harnessing AI for business transformation: strategies for effective implementation and market advantage***, **Jakub Kubiczek, Elżbieta Roszko-Wójtowicz, Julianna Koczy, Izabela Waszkiewicz** and **Klaudia Woś** check in what ways AI-driven tools contribute to business transformation, focusing on their impact on the operational efficiency, customer engagement and market competitiveness. The research employs a multi-method approach, including literature reviews, secondary data analyses and case studies of the AI use in selected enterprises in Poland. The authors' findings highlight both the positive and the negative results of using AI. While it enhances productivity, accuracy and sustainability, businesses must also navigate risks related to data security, compliance, and financial feasibility. The study underscores the importance of dynamic capabilities in leveraging AI for strategic growth while mitigating associated challenges. The results contribute to the discourse on AI's role in shaping modern commerce, offering practical insights for companies seeking to integrate AI-driven solutions effectively.

Research Communicates and Letters

Mateusz Borkowski and **Ewa Gruszewska's** study presented in their paper ***The integrity of the innovation process on the example of EU countries: a PLS-SEM approach*** assesses the integrity of the elements of innovation process and their efficiency with regard to EU countries on the basis of the Schumpeter concept. The study applies the partial least squares structural equation modeling (PLS-SEM), which allows the analysis of the latent variables (LVs). On the basis of the PLS-SEM models for 2010 and 2020, it is concluded that innovation proceed in an integrated manner in EU countries. Not only did the modeling results indicate a positive and moderate effect of the invention inputs on the innovation efficiency LV, but they also showed a positive and strong influence of innovation efficiency LV and the innovation diffusion LV in the analyzed countries. The technological process integrity of EU economies was lower in 2020 than in 2010. In order to improve the functioning of innovation activities, it is necessary to increase technology inputs and the efficiency of their use in R&D.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 1–19, https://doi.org/10.59139/stattrans-2025-013
Received – 04.09.2024; accepted – 13.02.2025

Thresholding nonprobability units in combined data for efficient domain estimation

Terrance D. Savitsky[1], **Matthew R. Williams**[2], **Vladislav Beresovsky**[3], **Julie Gershunskaya**[4]

Abstract

Quasi-randomization approaches estimate latent participation probabilities for units from a nonprobability / convenience sample. Estimation of participation probabilities for convenience units allows their combination with units from the randomized survey sample to form a survey-weighted domain estimate. One leverages convenience units for domain estimation under the expectation that estimation precision and bias will improve relative to solely using the survey sample; however, convenience sample units that are very different in their covariate support from the survey sample units may inflate estimation bias or variance. This paper develops a method to threshold or exclude convenience units to minimize the variance of the resulting survey-weighted domain estimator. We compare our thresholding method with other thresholding constructions in a simulation study for two classes of datasets based on the degree of overlap between survey and convenience samples on covariate support. We reveal that excluding convenience units that each express a low probability of appearing in *both* reference and convenience samples reduces estimation error.

Key words: survey sampling, nonprobability sampling, data combining, quasi randomization, thresholding units, bayesian hierarchical modeling

1. Introduction

Declining response rates for randomized survey instruments administered by government statistical agencies (Williams & Brick, 2017) have encouraged the development of quasi-randomization processes such as those of Elliott (2009); Elliott & Valliant (2017); Wang et al. (2021); Savitsky et al. (2023) to allow the inclusion of responses derived from a nonrandom convenience sample that includes responses for covariates that overlap those measured by the randomized survey or reference sample. Directly combining responses for units participating in the convenience sample with those selected into the randomized or reference sample may be expected to induce bias for inference about an underlying latent population, however, because the convenience sample is not generally representative of that population (Bethlehem, 2010; Meng, 2018; VanderWeele & Shpitser, 2011).

[1]Office of Survey Methods Research, U.S. Bureau of Labor Statistics, USA.
E-mail: Savitsky.Terrance@bls.gov. ORCID: https://orcid.org/0000-0003-1843-3106.
[2]RTI International, USA. E-mail: mrwilliams@rti.org. ORCID: https://orcid.org/0000-0001-8894-1240.
[3]Office of Survey Methods Research, U.S. Bureau of Labor Statistics, USA.
E-mail: Beresovsky.Vladislav@bls.gov. ORCID: https://orcid.org/0009-0002-8375-5195.
[4]OEUS Statistical Methods Division, U.S. Bureau of Labor Statistics, USA.
E-mail: Gershunskaya.Julie@bls.gov. ORCID: https://orcid.org/0000-0002-0096-186X.

Quasi-randomization methods propose model formulations to estimate unit participation probabilities as if the convenience sample is realized from a *latent* or unknown selection process. Quasi-randomization uses the reference sample and associated known inclusion probabilities to provide information about the underlying sampling frame that is, in turn, used to estimate convenience sample inclusion probabilities. The goal in using a statistical model to estimate the convenience sample inclusion probabilities is to allow inclusion of the convenience sample units in a combined (reference and convenience sample) data estimator for a domain mean (e.g., employment for computer services in New York city) with minimal bias.

Beresovsky et al. (2024) provides a comprehensive overview of quasi-randomization methods and compares the variance performances of a collection of methods for domain estimation that are mostly differentiated by assumptions about the degree of overlap in memberships in the convenience and reference samples, on the one hand, and the form of approximating inference on the non-sampled portion of the population, on the other hand. Elliott (2009) and Elliott & Valliant (2017) assume that the reference sample size is sufficiently small that there is a negligible overlap in unit inclusions with the convenience sample. This negligible overlap assumption is increasingly untenable under ever larger convenience samples. Later methods dispense with this assumption; in particular, Savitsky et al. (2023) and Wang et al. (2021) make no assumption about the degree of overlaps in units to allow more robust inference. Similarly, recent methods differ on how to estimate likelihoods specified for the population on realized (convenience and reference) samples. Wu (2022) and Wang et al. (2021) use a pseudo likelihood approach by approximating unknown population units with the weighted reference sample units. The use of reference sample-weighted units may inflate estimation variance for small-sized reference samples. Savitsky et al. (2023) directly specify a likelihood for the realized samples that avoids using reference sample weights.

To motivate the focus of our paper, we highlight a key covariate balance requirement of these methods to produce combined reference and convenience sample domain estimators with reduced bias (as compared to domain estimators obtained from solely using the reference sample).

Quasi-randomization methods require availability of the covariates used to determine the sampling design (governing the reference sample) for convenience sample units. This requirement is generally readily satisfied for sampling designs parameterized by demographic variables; for example, in the case of surveys conducted from business establishments by the U.S. Bureau of Labor Statistics, these covariates might include a discretized employment size class, industry classification and metropolitan statistical area designation.

Valliant (2020) further notes that the target population units are assumed to have positive probabilities to be included into both samples conditional on the shared set of covariates among both reference and convenience samples. They refer to this condition of positive participation probabilties for all units in both samples as a requirement for "common support ". Satisfying common support requires that the support of covariate values expressed by units in the population is also expressed by units included in *both* the reference and convenience samples. This paper addresses estimation bias that arises when common support is satisfied but where a subset of population units selected into the reference sample with relatively moderate-to-large inclusion probabilities may express vanishingly low convenience sample

participation probabilities. Heuristically, there are often subsets of the population purposefully emphasized in the reference sample that are poorly represented in the convenience sample.

Since a convenience sample derives from an opt-in or self-initiated participation process there will typically be some units in the realized convenience sample that are very different from those represented in the randomized reference sample. To be precise, there may be some units in the convenience samples whose covariate values don't well overlap those for the reference sample. Gelman & Hill (2007) discuss degrees of "partial overlap" in the space of covariate values that may occur between treatment and control sample arms in the causal inference experimental set-up and the increase in bias and variance in the resulting propensity scores. The low overlap of covariate values for those convenience units with the reference sample provides less information to estimate associated participation probabilities for them, which produces estimates with large errors. Including these low overlap convenience units along with reference units to formulate a domain estimator would be expected to inflate bias and variance rather than reduce it. The error inflating effect of these low overlap convenience units on the domain estimator would partially offset the variance reduction benefit of incorporating high overlap convenience units along with the reference units discussed in Savitsky et al. (2023).

This paper introduces an approach to identify and exclude a subset of convenience sample units whose covariate values poorly overlap the reference sample in order to further reduce the error in domain estimators that incorporate convenience units (and their estimated participation probabilities). Our approach for excluding or thresholding units uses estimated reference and convenience sample inclusion and participation probabilities for the *convenience* units as a uni-dimensional summary of the overlap of multivariate covariate values. In the sequel, we develop a set of alternative statistics used for thresholding where each statistic represents distinct functional combinations of the estimated reference and convenience sample inclusion and participation probabilities for the convenience units. We note that Savitsky et al. (2023) specify a Bayesian modeling approach that provides estimates for both convenience *and* reference sample participation and inclusion probabilities for the convenience units. The most simple example of using these estimated probabilities to threshold units would be to exclude convenience units with low reference sample inclusion probabilities below some threshold quantile. The logic for such a thresholding statistic is that convenience units with low values for estimated reference sample inclusion probabilities may be expected to express a low degree of overlap in covariate values with the reference sample.

We introduce a thresholding statistic for excluding convenience sample units that arises by minimizing of the variance of a domain mean estimator that is a function of the estimated reference and convenience sample inclusion and participation probabilities for the convenience sample units in Section 2. We begin by deriving the variance optimal thresholding statistic under the simpler set-up that composes the domain mean estimator using solely estimated convenience sample inclusion probabilities for convenience units (and excludes estimated reference sample inclusion probabilities for the convenience units). We then derive our main result under a set-up that constructs a threshold statistic composed of both estimated reference and convenience sample marginal probabilities for the convenience

units. Section 2.3 introduces an additional thresholding statistic motivated by Beresovsky et al. (2024). We compare the reductions in bias and means squared error offered by the alternative thresholding statistics with a Monte Carlo simulation study in Section 3 and conclude with a discussion in Section 4.

2. Optimal Variance Thresholding

2.1. Thresholding based solely on convenience sample probabilities

We begin this section using only convenience sample participation probabilities (obtained from co-modeling with the reference sample) for convenience units to construct our estimator to introduce our notation under a simpler thresholding construction. This set-up contrasts with the use of *both* estimated convenience and reference participation and inclusion probabilities for the convenience units to compose our domain mean estimator. We label the set-up that utilizes solely convenience sample participation probabilities (for convenience sample units) to define our thresholding statistic and set as "one-arm". By contrast, our main result will use the more general set-up that defines the thresholding statistic from both estimated convenience and reference sample probabilities, which we label as "two-arm".

Our main result defines a subset of $x \in \mathbb{X}$, where units in the convenience sample whose threshold statistic percentile (as a function of x) is less than a some small value (α) will be excluded from the subset. Only convenience sample units that are members of the subset will be used to render our weighted domain mean estimator, $\hat{\mu}$.

Let $\delta_c \in \{0,1\}$ denote unit participation in the convenience sample, where $\delta_c = 1$ denotes participation in the sample and $\delta_c = 0$ denotes a non-participating unit from the population frame, U, where $|U| = N$. Define marginal participation probability $\pi_c(x) = \Pr[\delta_c = 1 \mid X = x]$, where $X \in \mathbb{X}$ is a random variable. This construction for $\pi_c(x)$ defines a marginal participation probability (rather than a propensity score). We proceed to extend and adapt a result of Crump et al. (2009) from the literature on causal inference that defines a threshold statistic and acceptance set for units constructed from a subset of $x \in \mathbb{X}$, where the value of the threshold statistic is exceeded. The acceptance set formed by excluding units whose value lies below some percentile of the threshold statistic constructed by Crump et al. (2009) is guaranteed to produce a minimizing variance for the domain mean estimator after excluding those x not in the acceptance set. We repurpose and extend their result from treatment and control arms under their causal inference set-up to reference and convenience sampling arms under our survey sampling set-up. We begin our extension of their result with a simpler result that defines an acceptance set and formulation for a thresholding statistic for units in a convenience sample that produces a minimum variance for the domain mean estimator constructed solely from convenience sample participation probabilities.

Our population quantity of inferential interest is $\mu = \mathbb{E}(Y)$, where Y denotes a univariate response variable of interest. Define our domain mean estimator as,

$$\hat{\mu} = \mu + \frac{1}{N} \sum_{i=1}^{N} \frac{z_i \delta_i}{\hat{\pi}_c(x_i)}, \tag{1}$$

where we are assuming N is known and $z = y - \mu$. Treating N as known may be relaxed, in practice. Let

$$\phi(Y, \delta, X, \mu, e) = \frac{z\delta}{\pi_c(X)} \tag{2}$$

$$\hat{\mu} = \mu + \frac{1}{N}\sum_{i=1}^{N} \phi(y_i, \delta_i, x_i, \mu, e_i). \tag{3}$$

Then $\phi(Y, \delta, X, \mu, e)$ has 0 expectation and variance (Hirano et al., 2003, p. 1182),

$$\mathbb{E}\left[\phi(Y, \delta, X, \mu, e)^2\right] = \frac{1}{N}\mathbb{E}\left[\frac{\sigma_1^2(X)}{\pi_c(X)}\right], \tag{4}$$

where $\sigma_1^2 = \mathbb{V}(Y \mid \delta = 1, X = x)$. The expectation on the LHS of Equation 4 is taken with respect to the joint distribution for X and the taking of a sample from the underlying frame on which $\mathbb{X}$ is defined. The expectation on the RHS is taken with respect to the distribution for X.

Equation 4 may be used in combination with Corollary 1 of Crump et al. (2009) to produce the following result for the optimal threshold level, α.

Theorem 1 (One-arm extension of Crump et al. (2009))**.** *Assume $\pi_c(x) > 0 \; \forall x \in \mathscr{X}$. Then set $\mathbb{A} = \{x \in \mathbb{X} : \pi_c(x) > \alpha\}$ denotes the variance optimal subset of $\mathbb{X}$ after thresholding units where $\mathbb{A}$ is defined on the basis of thresholding conditional inclusion probability, $\pi_c(X)$. The minimum variance quantile α is constructed by,*

$$\frac{1}{\alpha} = 2\mathbb{E}\left[\frac{1}{\pi_c(X)} \;\; \frac{1}{\pi_c(X)} < \frac{1}{\alpha}\right]. \tag{5}$$

For computation of α we approximate the expectation with sums over units $i \in S_c$, where S_c denotes the observed convenience sample,

$$\frac{1}{\alpha} = 2\frac{\sum_{i \in S_c} \mathbf{1}(\hat{\pi}_c(x_i) > \alpha)\frac{1}{\hat{\pi}_c(x_i)}}{\sum_{i \in S_c} \mathbf{1}(\hat{\pi}_c(x_i) > \alpha)}. \tag{6}$$

Proof. Plugging in $\pi_c(X)$ for $e(X)$ into Theorem 1 of Crump et al. (2009) and using the result of Equation 4 for the case of where we utilize solely the convenience sample participation probabilities (for the convenience units) produces the result. □

Remark 1. The result of Theorem 1 utilizes a one-arm set-up that composes the mean estimator from solely the convenience sample. A companion, separate reference sample is required in order to estimate the convenience sample inclusion probabilities, $\hat{\pi}_c(x_i),\ i \in (1, \ldots, N)$. In the sequel, we will further extend Theorem 1 by additionally estimating the reference sample inclusion probabilities for the same convenience units, $\hat{\pi}_r(x_i),\ i \in (1, \ldots, N)$, also using the reference sample inclusion probabilities estimated on the convenience units. See Savitsky et al. (2023) for more details on estimating $(\hat{\pi}_c(x_i), \pi_r(x_i))$ (where subscript "r" denotes reference sample) for convenience sample units. They specify

a model for the observed membership indicator in the pooled sample, $\mathbf{1}_{z_i}$, which is set to 1 if unit i is included in the convenience sample and 0 if the unit belongs to the reference sample. Units in the convenience and reference samples are "stacked", which allows for a unit included in the convenience sample to also be included in the reference sample without the requirement to *know* the identity of that unit. They utilize a Bayesian hierarchical modeling approach that specifies a Bernoulli likelihood for indicator $\mathbf{1}_{z_i}$ for all units in the pooled sample. A likelihood term is also included for $\pi_r(X_i)$, only for units in the observed reference sample (where $\pi_r(X_i)$ is known) to borrow further modeling strength. This modeling set-up of Savitsky et al. (2023) may also be performed in the frequentist paradigm. The main advantage of the Bayesian approach is that it treats values $\pi_r(X_i)$ for the *convenience* sample as *unknown* and allows their estimation in the model. By contrast, in the frequentist set-up (see Beresovsky et al. (2024)), $\pi_r(X_i)$ are assumed known for *all* convenience and reference sample units.

Remark 2. In this one-arm case where the domain estimator is constructed solely from the estimated convenience sample inclusion probabilities, the resulting thresholding is performed on the convenience sample inclusion probabilities, $\pi_c(x_i),\ i \in S_c \subset U$ (where S_c denotes units in frame U that participate in the convenience sample), without accounting for the estimation quality of $\pi_c(X)$. So, this is a traditional regularization approach used to stabilize the variance of a survey domain estimator by excluding units with extreme weight values. This approach trades some small increase in bias for a large decrease in variance.

Remark 3. We include an alternative, direct derivation for the result of Theorem 1 in an Appendix A assuming Equation 4 is everywhere differentiable (on $x \in \mathbb{X}$). We also include an illustration to show that the result of the Theorem does, indeed, produce a minimum variance estimator for $\hat{\mu}$.

Equation 4 can now be generalized in the manner of Section 3.1 of Crump et al. (2009) to develop an alternative to their Theorem 1 and Corollary 1 under a composite estimator that includes both reference and convenience sample inclusion and participation probabilities.

2.2. Thresholding using both reference and convenience sample probabilities

Let δ_c and δ_r denote random inclusion indicators (governed by a survey design distribution) for convenience and reference samples, respectively, and let $\pi_c(x) = \Pr[\delta_c = 1 \mid X = x]$ and similarly for π_r. Define our estimator as,

$$\hat{\mu} = \mu + \frac{1}{N}\sum_{i=1}^{N} \frac{z_i \delta_{ci}}{\hat{\pi}_c(x_i)} + \frac{z_i \delta_{ri}}{\pi_r(x_i)}. \tag{7}$$

Although the above estimator is defined disjointly on the reference sample using $\pi_r(X)$ and the convenience sample using $\hat{\pi}_c(X)$, the resulting optimal variance thresholding rule of Equation 11 applies to *only* units in the convenience sample. So, as mentioned in Remark 4, below, we may use the estimated $\hat{\pi}_c(x_i)$ *and* $\hat{\pi}_r(x_i)$ for *each* unit $i \in S_c$ to apply the thresholding rule of Equation 11. To demonstrate that this trick works, we may generate an estimator identical to Equation 7 that includes both convenience and reference sample

probabilities defined *solely* for convenience units. Use $\{\pi_c(x_i)\}_{i\in S_c}$ to generate a pseudo population of size N (from units $i \in S_c$, allowing for replicates). Next take a random / probability sample from this pseudo population using $\{\pi_r(x_i)\}$ of the same size as the reference sample. Now form the same estimator as Equation 7, but the universe of units is actually confined to $i \in S_c$.

Let

$$\phi(Y,\delta_c,\delta_r,X,\mu,e_c,e_r) = \frac{z\delta_c}{\pi_c(X)} + \frac{z\delta_r}{\pi_r(X)} \tag{8}$$

$$\hat{\mu} = \mu + \frac{1}{N}\sum_{i=1}^{N} \phi\left(y_i,\delta_{ci},\delta_{ri},x_i,\mu,\pi_c(x_i),\pi_r(x_i)\right). \tag{9}$$

Then, from Hirano et al. (2003), the variance of our estimator is

$$\mathbb{E}\left[\phi(Y,\delta,X,\mu,e)^2\right] = \frac{1}{N}\mathbb{E}\left[\frac{\sigma_c^2(X)}{\pi_c(X)} + \frac{\sigma_r^2(X)}{\pi_r(X)}\right], \tag{10}$$

where $\sigma_c^2 = \mathbb{V}(Y \mid \delta_c = 1, X = x)$ and similarly for σ_r^2. The expectation on the LHS of Equation 4 is taken with respect to the joint distribution for X and the taking of a sample from the underlying frame on which $\mathbb{X}$ is defined. The expectation on the RHS is taken with respect to the distribution for X. We have used the assumption of independence between the sampling arms with respect to the design distribution.

We may now use Equation 10 to generalize Corollary 1 of Crump et al. (2009) under the assumption of $\sigma_c^2 = \sigma_r^2 = \sigma^2$.

Theorem 2 (Two-arm extension of Crump et al. (2009))**.**
Assume $(\pi_c(x) > 0, \pi_r(x) > 0),\ \forall x \in \mathbb{X}$.
Then $\mathbb{A} = \left\{x \in \mathbb{X} : \sqrt{\pi_r(X)\pi_c(X)/(\pi_r(X)+\pi_c(X))} > \alpha\right\}$ *defines the optimal subset of* $\mathbb{X}$*, where threshold* α *is obtained as a solution to,*

$$\frac{1}{\alpha^2} = 2\mathbb{E}\left[\frac{1}{\pi_c(X)} + \frac{1}{\pi_r(X)}\ \frac{1}{\pi_c(X)} + \frac{1}{\pi_r(X)} \le \frac{1}{\alpha^2}\right]. \tag{11}$$

Proof. Plugging in $\pi_c(x)$ for $e(X)$ and $\pi_r(X)$ for $1-e(X)$ into Theorem 1 of Crump et al. (2009) and using the result of Equation 10 for the case of where we utilize both the convenience sample and reference sample participation and inclusion probabilities (for the convenience units) produces the result. □

Remark 4. Defining variance optimal subset, $\mathbb{A}$, by thresholding $\sqrt{\pi_r(x_i)\pi_c(x_i)/(\pi_r(x_i)+\pi_c(x_i))} > \alpha$, is a harmonic mean that tends to exclude units i where $\pi_r(x_i)$ is a very different value from $\pi_c(x_i)$. We may even better understand the behavior of this thresholding statistic by noting the result from Beresovsky et al. (2024) that $\Pr[i \in S_c, i \in S_r \mid i \in S] = \pi_{ri}\pi_{ci}/(\pi_{ri}+\pi_{ci})$, where $S = S_c \bigotimes S_r$ denotes the pooled convenience and reference sample. This result reveals that convenience units with low probabilities of being in *both* the convenience and reference samples tend to be excluded. This thresholding behavior matches intuition because units with low probabilities to appear in

both samples will tend to have low overlaps in their covariate supports. We further note that our derivation of this variance minimizing threshold statistic was done without explicit reference to this joint probability, which makes the concordance of the two expressions (for the thresholding statistic, on the one hand, and the joint probability of inclusion in both samples, on the other hand) to be quite fortuitous. We label this thresholding statistic as "balanced" because it favors inclusion of records for estimating the domain mean that have relatively high probabilities of participating in *both* samples.

Remark 5. This thresholding method can be used in practice solely directed to units $i \in S_c$, because we have both estimated $(\hat{\pi}_c(x_i), \hat{\pi}_r(x_i))$ available.

Remark 6. Theorem 2 assumes both $(\pi_r(x), \pi_c(x))$ are *known* for the convenience units when, in fact, they are estimated. We explore the sensitivity to the performance of the variance minimizing thresholding statistic (for the domain mean) of this theorem to estimation uncertainty for $(\hat{\pi}_r(x), \hat{\pi}_c(x))$ in the simulation study to follow.

2.3. Thresholding statistic motivated by variance structure of model score function

Our derivation of the thresholding statistic of Section 2.2 treats $\pi_c(\mathbf{x})$ as known. By contrast, Beresovsky et al. (2024) suppose a generalized linear model, $\text{logit}(\pi_{ci}(\boldsymbol{\beta})) = \boldsymbol{\beta}^{\boldsymbol{T}}\mathbf{x}_i$, with a linear form under a logit link function for logistic regression. They derive the variance of the domain mean, $\hat{\mu}$, that includes an additive term for variance of the score function, $S(\boldsymbol{\beta})$, which has two parts:

$$\text{Var}[S(\boldsymbol{\beta})] = \text{Var}[S_c(\boldsymbol{\beta})] + \text{Var}[S_r(\boldsymbol{\beta})] =: \mathbf{A} + \mathbf{D}$$
$$\boldsymbol{D} = \text{Var}_d\left[\sum_{S_r} \frac{g_i}{1+g_i}(1-\pi_{ci})\mathbf{x}_i\right],$$

where $g_i = \pi_c(\mathbf{x}_i)/\pi_r(\mathbf{x}_i)$ and Var_d denote the design variance. Motivated by the dependence of $\boldsymbol{D}$ on g_i, we propose to use this statistic as another thresholding option.

We propose the following acceptance set that uses g:

$$\mathbb{A} = \{x \in \mathbb{X} : \pi_r(x)/\pi_c(x) > \alpha\}.$$

Remark 7. The use of $\pi_r(x)/\pi_c(x)$ as a thresholding statistic may be intuitively motivated by noting that it will tend to threshold or exclude units $i \in S_c$, where $\pi_r(\mathbf{x}_i)$ is relatively small for each unit and $\pi_c(\mathbf{x}_i)$ is relatively large, which may occur if the value for $\mathbf{x}_i$ for some $i \in S_c$ is not well covered by or represented in the reference sample, S_r.

3. Simulation study

3.1. Simulation design

We conduct a Monte Carlo simulation study that generates a finite population on each iteration to include covariates $\mathbf{x}$ that govern both the convenience and reference sample designs. The sample designs are size-based as a linear function of $\mathbf{x}$ where we vary the

coefficients of the linear function to draw two categories of reference and convenience samples: 1. Where the covariate spaces of resulting reference and convenience samples express a *high* degree of overlap; 2. Where the two samples express a *low* degree of overlap. We also generate a response variable of interest, y, for the finite population. A domain mean, μ, is constructed for the population and *estimated* by a combined weighted estimator over the reference and convenience samples. Finally, we compare the three thresholding methods we developed in Section 2 in terms of their bias, error and coverage performances. We expect that conducting thresholding of sampled convenience units using one or more of our thresholding statistics will reduce the estimation error.

We utilize the simulation data generation process of Savitsky et al. (2023). We briefly summarize the procedure and refer the reader for a more detailed exposition. We generate $M = 30$ distinct populations, each of size $N = 4000$. Design covariates, X, of dimension $K = 5$ are generated (all binary, with one continuous). Outcome variable, y_i, is generated as $\log(y_i) \sim \mathcal{N}(\mathbf{x}_i\beta, 2)$ for $i = 1, \ldots, N$.

A randomized reference sample of size $n_r = 400$ is taken from the finite population under a proportion-to-size (PPS) design with size variable, $s_{r_i} = \log(\exp(\mathbf{x}_i \times \beta) + 1)$.

For the convenience sample, we set $n_c \approx 800$, which is a relatively larger sampling fraction that we choose to explore the full range of $\pi_c \in [0, 1]$ that we would expect to see for business establishment data in the U.S. Bureau of Labor Statistics. We use a size-based Poisson sample with $\pi_{c_i} = \text{logit}^{-1}(\mathbf{x}_i \times \beta_c + \text{offset})$. We control 'high' and 'low' overlap by varying β_c compared to the reference sample.

Figure 1 presents a violin (rotated and reflected density) plot for the percentage overlap of *units* in both the convenience and reference samples over the Monte Carlo iterations. The left-hand plot represents the high overlap samples and the right-hand plot represents the low-overlap samples. We see that the number of shared units in both samples is notably higher for the high overlap samples than for the low overlap samples. We expect fewer units to be thresholded for a high overlap sample since their covariate supports express relatively more overlap such that units in the convenience sample are more similar to those in the reference sample. Since our modeling obtains information about the population from the reference sample (and reference sample inclusion probabilities), we are able to better estimate participation probabilities for convenience units that are similar in covariate values to the reference units.

3.2. Thresholding of convenience units

In this paper, we employ the Bayesian model formulation of Savitsky et al. (2023) that estimates both $(\pi_r(\mathbf{x}_i), \pi_c(\mathbf{x}_i)),\ i \in S_c$. In the sequel we use $(\pi_{ri} = \pi(\mathbf{x}_i), \pi_{ci} = \pi_c(\mathbf{x}_i))$ for ease-of-reading and to emphasize the dependence on $i \in S_c$.

Within each Monte Carlo iteration, $m \in 1, \ldots, M$, we conduct thresholding of convenience units and computation of the domain mean for each posterior/MCMC sample in the following procedure:

1. For *each* posterior/MCMC draw $s \in 1, \ldots S$, compute the thresholding statistic (e.g., balanced thresholding statistic) for each unit $i \in S_c$ as a function of $(\hat{\pi}_{rsi}, \hat{\pi}_{csi})$. Denote the focus thresholding statistic as, $T(\hat{\pi}_{rsi}, \hat{\pi}_{csi})$, which allows us to provide a general

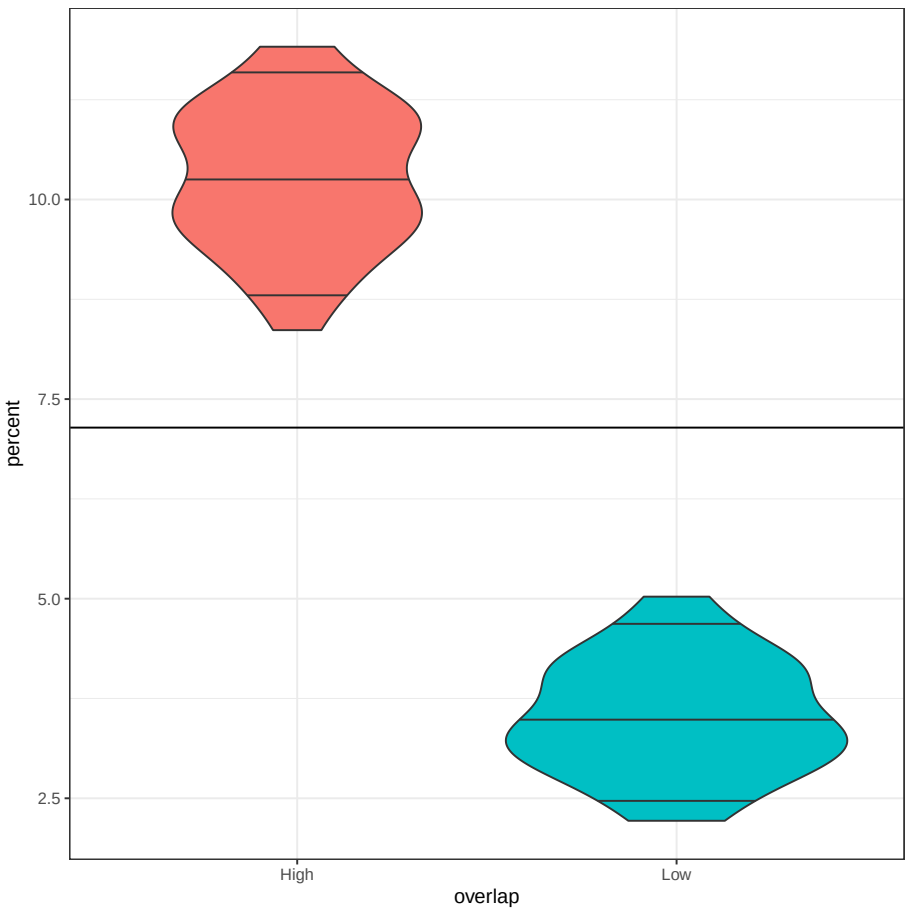


Figure 1: Distribution over $M = 30$ Monte Carlo iterations of the percentage of units overlapping between realized reference and convenience samples (taken on each Monte Carlo iteration)

exposition of how we conduct thresholding of convenience units; for example, we may set $T(\hat{\pi}_{rsi}, \hat{\pi}_{csi}) = \sqrt{\hat{\pi}_{rsi}\hat{\pi}_{csi}/(\hat{\pi}_{rsi} + \hat{\pi}_{csi})}$ for $i \in S_c$.

2. For MCMC iteration s: evaluate the distribution of the thresholding statistic $T(\hat{\pi}_{rsi}, \hat{\pi}_{csi})$ over the convenience units, $i \in S_c$, and compute threshold quantile, α_s, associated with target percentile, γ, below which convenience units are excluded / thresholded.

3. *Retain/accept* those *convenience* units where $\mathbb{A}_s = \{i \in S_c : T(\hat{\pi}_{rsi}, \hat{\pi}_{csi}) > \alpha_s\}$.

4. Use the retained units in draw s to construct the domain mean, $\mu_s = (\sum_{i\in\mathbb{A}_s} y_i/\hat{\pi}_{csi} + \sum_{i\in S_r} y_i/\hat{\pi}_{rsi})/(\sum_{i\in\mathbb{A}_s} 1/\hat{\pi}_{csi} + \sum_{i\in S_r} 1/\hat{\pi}_{rsi})$.

5. One now has the induced posterior distribution over the S MCMC samples for μ from which one may estimate the mean (e.g., $\mu = 1/S\sum_{s=1}^{S} \mu_s$).

Remark 8. The above procedure is a form of "soft" thresholding, because a unit $i \in S_c$ may be excluded on posterior sampling draw s in forming domain mean estimator μ_s, but then *included* in posterior draw $s^{'}$ to construct $\mu_{s'}$. So each μ_s may be constructed from a differing set of convenience units. This occurs because $(\hat{\pi}_{rsi}, \hat{\pi}_{csi})$ are parameters estimated from our model, so the distribution of $T(\hat{\pi}_{rsi}, \hat{\pi}_{csi})$ over convenience units $i \in S_c$ will vary from over the posterior draws, $s \in 1, \ldots, S$.

We formulate a variation to this procedure that produces a "hard" threshold to compare the performance to our main soft thresholding procedure. For the hard thresholding alternative, we construct the acceptance sets $\mathbb{A}_s,\ s = 1, \ldots, S$ as described in the first 3 steps of the above procedure. We then count the percentage of the S posterior draws where unit i is in each acceptance set $\mathbb{A}_s$. If the percentage is less than 50%, we *exclude* or threshold unit i. In other words, we form a single acceptance set over the S MCMC draws with $\mathbb{A} = \{i \in S_c : i \in \mathbb{A}_s$ for a total of $S^i = \sum_s^S (1 : i \in \mathbb{A}_s) > 0.5S\}$. So, our first addtional steps formulates $\mathbb{A}$, the set of non-thresholded convenience units. We then use this *same* set of units to compute μ_s for each MCMC draw. So, either unit i is included to construct all the μ_s or it is excluded. We use the label "two-step" for this hard thresholding alternative since we first threshold the units over all MCMC draws and then compute the domain mean estimator.

Remark 9. Although our thresholding procedure is constructed under the Bayesian model formulation of Savitsky et al. (2023) for developing a thresholded posterior distribution for domain mean, μ, steps $1-4$ of our thresholding procedure may be applied under the frequentist generalized linear formulation of Beresovsky et al. (2024), to obtain a thresholded estimator of μ with no loss of generality or applicability. Instead of thresholding each MCMC draw, s, one would threshold the statistic formed from the maximum likelihood estimators of the convenience sample participation probabilities under frequentist model estimation.

3.3. Results

Figure 2 presents plot panels for bias, root mean squared error (RMSE), median absolute deviation (MAD) and coverage results over the M Monte Carlo iterations. The left side of

each horizontal bar in the plot panels represents a result for "L" or the low overlap sample, while the right side of each horizontal bar represents a result for "H" or the high overlap sample. The top row of bars in blue presents results using the unknown *true* values for both the reference sample inclusion probabilities for the reference sample units and the convenience sample inclusion probabilities for the convenience units as if they were known. The next row of bars down from the top in red presents the result from the model of Savitsky et al. (2023) that smooths or co-models the inclusion probabilities for the reference sample units. No thresholding is conducted for the results in these first two rows. The next two rows of bars present results for our variance-optimal balanced threshold statistic: the orange bar uses our main soft thresholding procedure, while the yellow bar uses the alternative hard thresholding procedure that we label as "two-step". The next row of light green bars presents the results for thresholding π_{ri} while the last row of green bars presents results for the ratio (π_{ri}/π_{ci}) thresholding statistic. We remind the reader that the statistics and thresholding are performed over $i \in S_c$ (the convenience sample) and that our Bayesian model estimates both (π_{ri}, π_{ci}) for each unit in the convenience sample. The vertical black dashed line in each plot panel represents the result using *only* the reference sample (and excluding the convenience sample). We use the $\gamma = 5\%$ of the distribution over the convenience for each threshold statistic to compute the thresholding quantile, α.

One notes that the estimation errors (RMSE, MAD) are little different both with and without thresholding and among the thresholding statistics for the high (H) overlap samples, which is expected because there is less need for thresholding due to the high degree of overlap in covariate spaces between the reference and convenience samples such that most convenience sample inclusion probabilities are well-estimated. By contrast, we observe that the estimation errors for the balanced statistic perform best among the different thresholding statistics and even better than the case where we use the true convenience sample participation probabilities (blue bars) as if they were known. The slight increase in bias relative to the blue bar is more than offset by a decrease in variance, producing lower estimation error. There is little difference between the soft and hard thresholding alternatives under the balanced statistic, though the soft thresholding produces a slightly higher amount of bias but also a slightly lower amount of estimation error as compared to hard thresholding. Perhaps we are not surprised that the balanced threshold statistic performed best because it was derived as a minimum variance estimator for the domain mean, though it is surprising that this thresholding option performed better for low overlap (L) samples than did the domain mean estimator constructed from the true (rather than estimated) convenience sample inclusion probabilities (as if they were known).

Lastly, while the balanced threshold statistic produces only a slight improvement in the error for high overlap (H) samples, the notion of whether a convenience sample is high or low overlap is relative such that the practitioner may not know whether their realized reference and convenience samples represent H or L. Nevertheless, since thresholding with the balanced statistic never produces worse errors than not thresholding, and sometimes much better, there is little risk to use thresholding.

We chose a reasonably small (5%) percentile for thresholding, so we next experiment with 10% and 1% under our best performing balanced thresholding statistic (under soft thresholding). Figure 3 presents the results. While the estimation errors are similar for

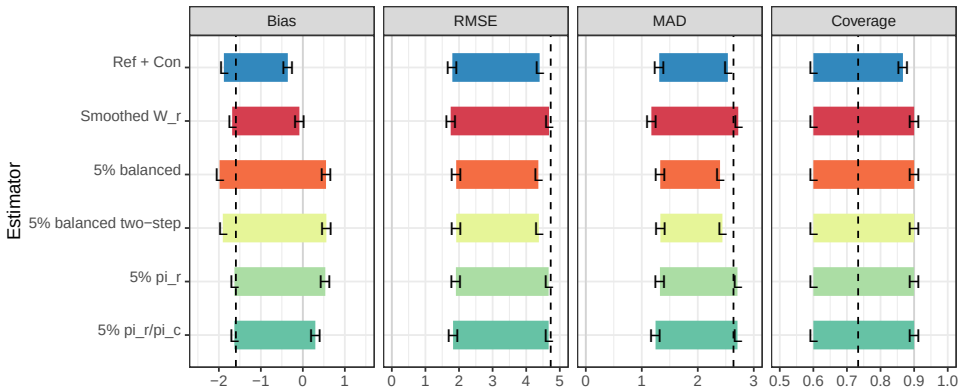


Figure 2: Performance of the weighted mean estimator between high (H) and low (L) overlapping samples using variations of the two-arm method across Monte Carlo Simulations for (top to bottom): True weights for both samples (Blue), Smoothed weights for reference sample (Red), minimum variance or balanced $\sqrt{\pi_r(x)\pi_c(x)/(\pi_r(x)+\pi_c(x))}$ (Orange), balanced based on posterior mean (Yellow), π_r only (Light Green), π_r/π_c (Dark Green). Left to right: Bias, root mean square error, mean absolute deviation, coverage of 90% intervals. Vertical reference line corresponds to using the reference sample only

the 3 different percentiles for low overlap (L) samples, we nevertheless note that the error performance is notably better for 1% balanced thresholding under high overlap (H) samples than the other two higher thresholding percentiles, and even performs slightly better than the blue bar that uses true convenience sample participation probabilities. Thresholding fewer units for high overlap samples intuitively makes sense since convenience units are relatively more similar to reference units. The low overlap sample MAD is, however, worst for the 1% threshold and best for the 10% threshold, which also accords with intuition since the convenience units in low overlap samples are less similar (in their covariate values) to reference sample units. Yet, the worsening of estimation error in the low overlap is a much smaller magnitude than the improvement in estimation error for high overlap. Our results suggest that the practitioner may generally favor a relatively lower value for the thresholding percentile.

While thresholding does notably reduce estimation errors (RMSE/MAD) on low overlap samples, as expected, uncertainty quantification is little improved (and continues to express undercoverage) even after thresholding due to the limited estimation improvement offered for a low overlap convenience sample. The fidelity of uncertainty quantification is driven by the underlying degree of overlap in the covariate supports of the reference and convenience sampling arms and is not much affected by thresholding relatively few convenience units. As a result of the low quality of uncertainty quantification under the low overlap samples, the coverage performances for all methods express little differentiation. By contrast, for high overlap the coverage results are more robust and nominal coverage is achieved when thresholding relatively fewer units, as expected. Thresholding is most important for low

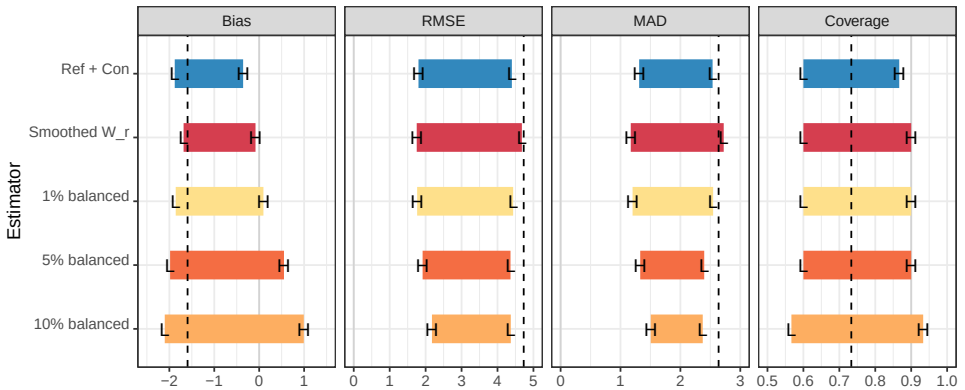


Figure 3: Comparison of the variance for the balanced threshold, $\sqrt{\pi_r(x)\pi_c(x)/(\pi_r(x)+\pi_c(x))}$ between high (H) and low (L) overlapping samples for (top to bottom): True weights for both samples (Blue), Smoothed weights for reference sample (Red), 1% (Yellow) vs. 5% (Orange) and 10% (Light Orange). Left to right: Bias, root mean square error, mean absolute deviation, coverage of 90% intervals. Vertical reference line corresponds to using the reference sample only

overlap samples to prevent non-representative outliers from inducing large errors (due to biased estimation of their convenience sample inclusion probabilities). Our results show that thresholding for low overlap samples provides a notable improvement in error control over repeated sampling.

We recall that the balanced threshold statistic was derived to produce a minimum variance domain mean estimator. Yet, the result in Section 2 assumes that the reference sample inclusion and convenience sample participation probabilities for convenience sample units, (π_{ri}, π_{ci}), $i \in S_c$, are *known* when, in fact, they are estimated. We seek to assess the sensitivity of the thresholding statistics to uncertainty in the estimation of these inclusion and participation probabilities for convenience units.

Each curve in a each plot panel of Figure 4 presents a sequence of 90% credibility intervals of percentiles for the fit statistic estimated on each MCMC iteration. More specifically, if we fix an MCMC iteration, we next compute the estimated thresholding statistic from the probabilities for each unit and compute its percentile of the distribution of the statistic over the convenience sample units. We repeat this process for each MCMC draw, which gives us a range of percentiles of the thresholding statistic for each convenience sample unit. Each horizontal line in the curve represents the 90% credibility interval of the percentiles for a convenience sample unit. These lines are ordered along the horizontal axis by the posterior mean of estimated thresholding statistic for each unit. The longer the horizontal lines, the greater the estimation uncertainty for the thresholding statistic. The blue-colored horizontal lines represent those units that have switched from being on one side of threshold to the other (meaning, they were sometimes included and sometimes excluded) more than 10% of the MCMC samples. The horizontal dashed lines in each panel represent $1\%, 5\%$and10%

thresholds (from bottom-to-top).

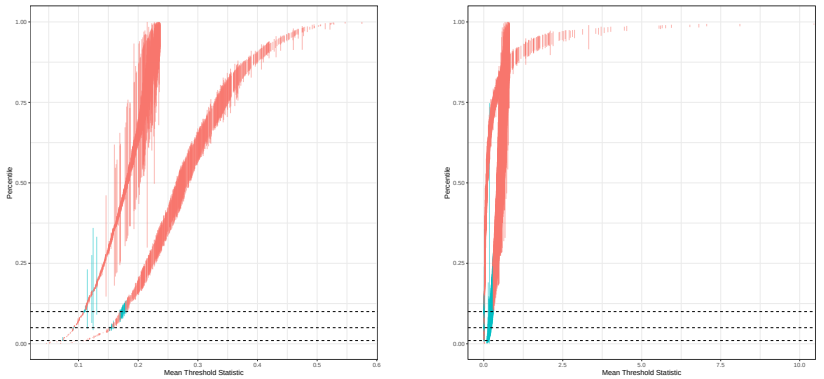
The left-hand curve in each plot panel represents estimations under low overlap samples and the right-hand represents high overlap samples. The left plot panel represents the the balanced thresholding statistic, while the right panel represents the ratio thresholding statistic.

Focusing on the left-hand panel for the balanced thresholding statistic, we see that the relatively wider horizontal lines for the low overlap sample express more estimation uncertainty than do those for the high overlap sample. That accords with our expectation, because the reference sample provides less information about convenience units whose covariate values are different from those of the reference sample. Yet, we see relatively few units (colored in blue) that switch between being excluded/thresholded and included for estimating the domain mean. So, the uncertainty does not impact the thresholding set and that explains why the balanced thresholding statistic turned out to be variance-optimal as compared to the other thresholding statistics despite the uncertainties in estimating inclusion and participation probabilities. By contrast, we observe a relatively higher number of units that switch between inclusion and exclusion under the ratio thresholding statistic in the right-hand plot panel. So, the performance of this thresholding statistic is less robust under uncertainty about the probabilties than is the balanced thresholding statistic.

4. Conclusion

The quasi-randomization method of Savitsky et al. (2023) that treats the non-randomized convenience sample as if it arose from a latent survey design process with an unknown sampling distribution provides a start-of-art method for producing survey-weighted domain estimates. Yet, the estimation quality of inclusion and participation probabilities for convenience units depends on the degree of overlap in the design covariate spaces between the randomized reference and convenience samples. It is typically the case that the estimated convenience sample participation probabilities for some convenience units whose design covariate values are very different from the reference sample are not well-estimated. Incorporating these units can partially defeat the purpose of leveraging the convenience sample by actually increasing bias and variance as compared to excluding them.

We devised a soft thresholding procedure for excluding convenience sample units that are very different from reference sample units and achieved a notable reduction in estimation error for low overlap (in their design covariate spaces) samples. We began by developing a new formulation for a balanced threshold statistic that minimized the resulting variance of the domain estimator. Our balanced thresholding statistic proposes to exclude some convenience sample units and is constructed from inclusion and participation probabilities for convenience units that effectively serve as one-dimensional summaries of the design covariates. It was particularly interesting to discover that the balanced threshold statistic derived from a theoretical exposition turns out to be a function of the joint probability that a unit is in *both* the reference and convenience samples. This formulation makes intuitive sense because our procedure proposes to exclude those convenience units that express low probabilities of being in both samples.



(a) Balanced Thresholding **(b)** Ratio Thresholding

Figure 4: Vertical lines are percentiles of Threshold statistic distribution over 700 MCMC draws of π_{ci}, π_{ri} for convenience units. Left is L and Right is H. Blue denotes unit that jumped threshold $> 10\%$, 5%, and 1% of draws

We motivated an additional thresholding statistic that we labeled “ratio" as the ratio of reference and convenience sample inclusion probabilities based on the variance formulation of the domain mean estimator derived in Beresovsky et al. (2024).

We designed a soft thresholding procedure that constructed an acceptance set for convenience units to be included in domain mean estimator on each MCMC iteration such that a unit might be included in some iterations but not others.

Our result revealed that the balanced threshold statistic produced the greatest reduction in the variance of the domain estimator, particularly for relatively lower overlap samples. We also showed that this reduction is relatively insensitive to the percentile cutoff for the estimated distribution of the balanced threshold statistic over the convenience sample units. Finally, we showed that this variance reduction result is robust against estimation uncertainty because the units that are thresholded are minimally impacted under our soft thresholding procedure.

References

Beresovsky, V., Gershunskaya, J. and Savitsky, T. D., (2024). Review of quasi-randomization approaches for estimation from non-probability samples.

Bethlehem, J., (2010). Selection bias in web surveys. *International Statistical Review*, 78(2), pp. 161 – 188.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A., (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1), pp. 187–199.

Elliott, M. R., (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2, pp. 813–845.

Elliott, M. R. and Valliant, R., (2017). Inference for Nonprobability Samples. *Statistical Science*, 32(2), pp. 249 – 264.

Gelman, A. and Hill, J., (2007). *Data analysis using regression and multilevel/hierarchical models*, volume Analytical methods for social research. New York: Cambridge University Press.

Hirano, K., Imbens, G. and Ridder, G., (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), pp. 1161–1189.

Meng, X.-L., (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *Annals of Applied Statistics*, 12(2), pp. 685 – 726.

Savitsky, T. D., Williams, M. R., Gershunskaya, J. and Beresovsky, V., (2023). Methods for combining probability and nonprobability samples under unknown overlaps. *Statistics in Transition*, 24(5), pp. 1–34.

Valliant, R., (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2), pp. 231–263.

VanderWeele, T. J. and Shpitser, I., (2011). A new criterion for confounder selection. *Biometrics*, 67(4), pp. 1406 – 1413.

Wang, L., Valliant, R. and Li, Y., (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Stat Med.*, 40(4), pp. 5237–5250.

Williams, D. and Brick, J. M., (2017). Trends in U.S. Face-To-Face Household Survey Nonresponse and Level of Effort. *Journal of Survey Statistics and Methodology*, 6(2), pp. 186–211.

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

Appendix

A. Direct derivation of variance minimizing threshold for one-arm sample

The Hajek mean estimator from the convenience sample S_c is:

$$\hat{\bar{y}} = \frac{\sum_{S_c} \frac{y(x)}{\hat{e}(x)}}{\sum_{S_c} \frac{1}{\hat{e}(x)}}$$

where $\hat{e}(x)$ is an estimated propensity score.

The associated model-based variance of this estimator is:

$$\text{var}\left(\hat{\bar{y}}\right) = \frac{\sum_{S_c} \frac{\sigma_y^2(x)}{\hat{e}^2(x)}}{\left[\sum_{S_c} \frac{1}{\hat{e}(x)}\right]^2}.$$

Assume that all variance $\sigma_y^2(x) = \sigma_y^2$ are equal. Order convenience sample units by response propensity $\widehat{e}(x)$. Units can be listed by $\widehat{e}(x)$ with density $w(\widehat{e}(x)) = \widehat{e}(x)$. Variance estimated from full convenience sample S_c without cut-off may be expressed as the integral over the distribution of response propensity $\widehat{e}(x)$

$$\text{var}\left(\hat{\bar{y}}\right) = \frac{\int_0^1 \frac{\sigma_y^2(x)}{\hat{e}^2(x)} w\left(\hat{e}(x)\right) d\left(\hat{e}(x)\right)}{\left[\int_0^1 \frac{1}{\hat{e}(x)} w\left(\hat{e}(x)\right) d\left(\hat{e}(x)\right)\right]^2} = \frac{\sigma_y^2 \int_0^1 \frac{1}{\hat{e}(x)} d\left(\hat{e}(x)\right)}{\left[\int_0^1 d\left(\hat{e}(x)\right)\right]^2}.$$

If sample units are trimmed by response propensity at level ε, then variance depending on ε is

$$\text{var}\left(\hat{\bar{y}}, \varepsilon\right) = \frac{\sigma_y^2 \int_\varepsilon^1 \frac{1}{\hat{e}(x)} d\left(\hat{e}(x)\right)}{\left[\int_\varepsilon^1 d\left(\hat{e}(x)\right)\right]^2} = \frac{\sigma_y^2 F(\varepsilon)}{G^2(\varepsilon)},$$

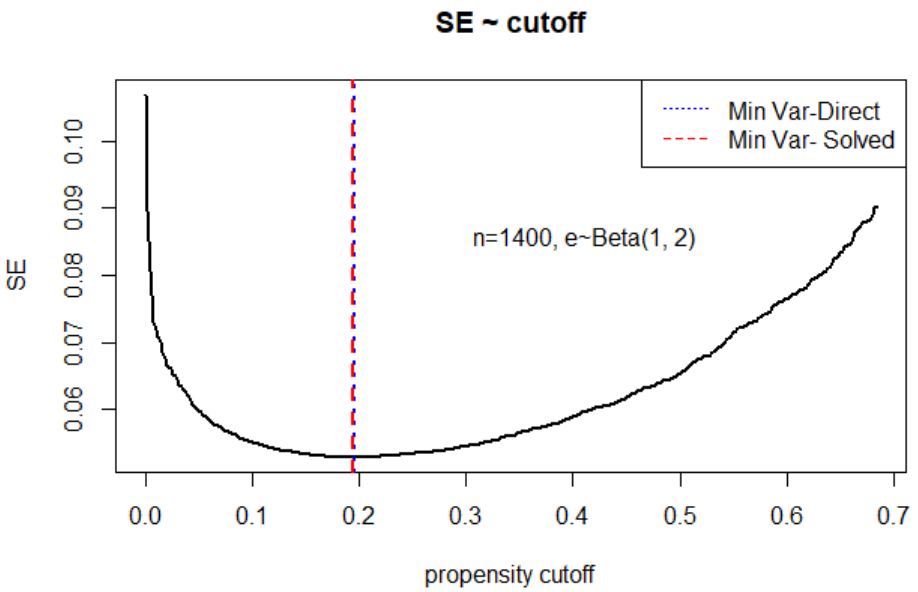
where $F(\hat{e}(x))$ is a primitive of $f(\hat{e}(x)) = 1/\hat{e}(x)$ and $G(\hat{e}(x))$ is a primitive of 1.

Minimize the trimmed variance by ε

$$\frac{d\,\text{var}\left(\hat{\bar{y}}, \varepsilon\right)}{d\varepsilon} = \frac{\sigma_y^2 F'(\varepsilon) G^2(\varepsilon) - 2G'(\varepsilon) G(\varepsilon) \sigma_y^2 F(\varepsilon)}{G^4(\varepsilon)} = 0.$$

Here we have:

$$F'(\varepsilon) = \frac{d}{d\varepsilon}\left(F(1) - F(\varepsilon)\right) = 0 - \frac{1}{\varepsilon} \times 1$$
$$G'(\varepsilon) = \frac{d}{d\varepsilon}\left(G(1) - G(\varepsilon)\right) = G'(1) - G'(\varepsilon) = -1.$$



Optimal propensity cut-off point ε can be estimated from the numerator null condition

$$\frac{1}{\varepsilon_c} G(\varepsilon_c) - 2F(\varepsilon_c) = 0$$

$$\frac{1}{\varepsilon_c} = \frac{2F(\varepsilon_c)}{G(\varepsilon_c)} = \frac{2\sum_{S_c} \frac{1}{\widehat{e}(x)} |\widehat{e}(x) > \varepsilon_c}{\sum_{S_c} 1 |\widehat{e}(x) > \varepsilon_c}.$$

Results of simulations:

- Sample size $n = 1,400$
- Propensity score $\widehat{e} \sim Beta(1,2)$

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 21–38, https://doi.org/10.59139/stattrans-2025-014
Received – 29.07.2024; accepted – 30.01.2025

Application of Statistical Disclosure Control methods to protect the confidentiality of the 2020 agricultural census microdata[1]

Andrzej Młodak[2], **Tomasz Józefowski**[3]

Abstract

In this paper, we describe an attempt made to develop an efficient disclosure control algorithm for microdata in a statistical portal used for releasing detailed statistical information at various levels of spatial aggregation. The proposed algorithm is based on perturbative methods, such as microaggregation with Gower's distance for categorical variables and the addition of correlated noise for continuous variables, but it also offers several alternative options in this regard. Moreover, the algorithm can be used to assess the loss of information by measuring distribution disturbances (based on a complex distance that accounts for all measurement scales) and the impact of the Statistical Disclosure Control (SDC) on the strength of correlations between variables (for continuous variables). Through the application of the tools offered by the `sdcMicro`R package, the algorithm was tested using microdata about agricultural farms and farm animals collected in the 2020 Polish Agricultural Census. We present the results of the tests and discuss the main problems and challenges connected with the use of such tools.

Key words: Statistical Disclosure Control, perturbative methods, disclosure risk, information loss, agricultural census.

1. Introduction

Censuses are the biggest and most informative statistical data collection undertakings. They provide key data about the population, households and farms. This is why they are of particular interest to all groups of users, including government agencies and units of local government administration, policy makers, and various organizations. In other words, the demand for detailed and comprehensive census data is especially high.

Before census data can be safely released, they have to undergo a meticulous process of statistical disclosure control (SDC) to ensure that sensitive information remains confidential. The primary task of every national statistical institute consists in striking an optimal balance between minimizing the risk of disclosure and maximizing the utility of disclosed data

[1]The paper was presented at the International Conference "Privacy in Statistical Databases 2024" (PSD 2024) in Antibes Juan-les-Pins, France, September 25–27, 2024.

[2]Statistical Office in Poznań, Centre for Small Area Estimation; address: Statistical Office in Poznań, Branch in Kalisz, ul. Piwonicka 7-9, 62-800 Kalisz, Poland; e-mail: a.mlodak@stat.gov.pl & University of Kalisz, Interfaculty Department of Mathematics and Statistics, ul. Nowy Świat 4a, 62-800 Kalisz, Poland. ORCID: https://orcid.org/0000-0002-6853-9163.

[3]Poznań University of Economics and Business, Department of Statistics, al. Niepodległości 10, 61-875 Poznań, Poland; e-mail: tomasz.jozefowski@ue.poznan.pl & Statistical Office in Poznań, Centre for Small Area Estimation, ul. Wojska Polskiego 27/29, 60-624 Poznań, Poland. ORCID: https://orcid.org/0000-0001-9485-1946.

(i.e. minimizing the loss of information resulting from SDC measures). Which SDC methods and to what extent they should be used depends, of course, on the format of data publication and their specific characteristics.

In this paper we present and discuss solutions used to test the process of releasing microdata from the Agricultural Census of 2020. This work was part of a broader research project concerning applications of SDC to protect census data (including microdata from the 2021 National Population and Housing Census). The project focused on three different forms of releasing statistical information:

- microdata
- data at 1 km $\times$ 1 km resolution
- hypercubes.

Different methods and tools are required in order to protect data confidentiality in each of these cases. Sets of microdata are collections of information about individual units (e.g. persons); hypercubes are multidimensional tables, while gridded data can be treated as a kind of table (a special type of hypercube), but given very small counts in the majority of the cells, it is possible to apply some methods dedicated to microdata.

The use of SDC methods for census data has been broadly investigated in the literature (cf. e.g. Zayatz (2002), Shlomo et al. (2010), Jansson (2012), Calian (2020), Kraus (2021), SNI et al. (2022) or Muralidhar and Domingo-Ferrer (2023)). The authors of these papers propose various approaches for specific cases with a specific set of parameters. A number of methods have been developed to protect census data, such as Targeted Record Swapping or the Cell Key Method. Moreover, the US Census Bureau adopted a differential privacy approach based on the assumption that a change to one entry in a database only creates a small change in the probability distribution of the outputs (cf. e.g. Abowd (2018)). Since then, this approach has been modified in various ways. For instance, Tran et al. (2024) propose using quantile regression to improve the utility of data protected by differential privacy. Jackson et al. (2024) demonstrate how to apply differential privacy to efficiently protect tabular data using the Poisson synthesis mechanism. It is also becoming increasingly common to rely on neural networks – such as Generative Adversarial Networks, GAN – to generate synthetic data (cf. Yoon et al. (2020)). Innovative approaches have also been proposed to measure the risk of disclosure and information loss. For example, Shlomo (2022) develops distance metrics to compare the overall distributions in the original data versus synthetic data for a particular variable and, more specifically, within equivalence classes based on the Kullback–Leibler distance, the Total Variation and Hellinger's Distance. Shlomo and Skinner (2022) use microdata from a sample survey to infer population parameters when the population is unknown, and estimate the risk of re-identification based on the notion of population uniqueness using probabilistic modelling. A synthetic review of modern concepts in this field can also be found in Templ (2017). Młodak (2020) proposes a new method of assessing information loss in terms of distribution disturbance based on the idea of the Gower distance, where the cyclometric function is used in the partial distance for continuous variables, and a method of computing information loss on regarding relationships between continuous variables using an inverse correlation matrix. An improved

version of this method is described in Młodak et al. (2022), which also presents a method of assessing external (ex post) disclosure risk – when it is assumed that the end user has access to an alternative data set containing information that can be linked with statistical data in order to identify a unit.

We show the particular nature of the data from Polish censuses and propose optimum SDC methods to protect microdata from the 2020 Agricultural Census. The purpose of the work conducted during this study was to prepare an algorithm for protecting microdata to be released in the Geostatistics Portal maintained by Statistics Poland. The microdata were used primarily to test the algorithm's efficiency, but, ultimately, they will be also uploaded to the portal. The work was conducted between 1st July 2018 to 31st October 2022 as part of the project "Spatial statistical data in the state information system", implemented under the Operational Programme Digital Poland within Priority axis II – "E-administration and open government" Measure 2.1. "High availability and quality of public e-services". All actions were financed from the European Regional Development Fund (ERDF). The main contractor was a consortium of companies with a documented track record of data processing and statistical analysis. The underlying assumptions of the SDC process were specified by Statistics Poland.

The final algorithm was based on the perturbative methods, such as microaggregation and noise addition. The efficiency of the methods we applied was assessed with measures of disclosure risk (based on k-anonymity and the concept of individual and global risk) and information loss (above all, these proposed by Młodak et al. (2022)). The methods we chose and their parameters as well as the most important problems to be solved in the future are presented and discussed below.

The paper is organised as follows. Section 2 presents basic assumptions of the Geostatistics Portal and the project "Spatial statistical data in the state information system", especially with respect to microdata. Section 3 describes the set of microdata from the 2020 Agricultural Census and their basic characteristics. Section 4 contains a description of various methods of statistical disclosure control and dedicated software. Section 5 contains the most important results and addresses problems encountered during the study. Section 6 includes the key conclusions.

2. The release of microdata via the Geostatistics Portal

The Geostatistics Portal of Statistics Poland (https://portal.geo.stat.gov.pl/en/home/)[4] was created to satisfy the demand for detailed and high-quality data at various levels of spatial aggregation, enabling users to conduct their own studies and analyses and present results in their preferred form (tabular or graphical). The Portal was developed as part of the Spatial Statistical Data project, with a goal of expanding the scope and availability of statistical information and geostatistical analysis methods that rely on publicly available statistical data.

Before any innovations were implemented, user needs and current limitations of the Geostatistics Portal were analyzed. One of the expectations was that the portal should

[4] Some information given in the following paragraphs is based on the description included on the webpage.

provide a wide range of possible tools for analyzing the spatial distribution of various socio-economic phenomena in specific areas and with a high level of detail and precision. Therefore, an option of analyzing gridded microdata (1 km $\times$ 1 km grid) was added and tools were provided to enable users to independently conduct advanced statistical analyses, especially at various levels of spatial aggregation.

The improved functionalities of the Portal include tools for statistical analyses at any level of spatial aggregation, possibility of combining statistical data with users' own data, geocoding user objects used for geostatistical analyses, using exploratory analyses of spatial data based on statistical information, performing geostatistical modelling and the option of enriching users' own content with geostatistical information and analyses.

In summary, the main outcome of the project are a number of publicly available e-services:

- the ability to access statistical information collected in the Portal from a remote computer and perform advanced spatial analyses using available data and metadata (users can select data, area, visualization method and method parameters). The user can generate an analysis of a given spatial area (also at 1km x 1km resolution) and present its results in a choropleth map or various types of editable cartodiagrams, which can show the variability of statistical data over time,

- the ability to access the portal from a mobile device (using an Android and iOS application). The aforementioned functionalities are adapted to being displayed on smaller screens,

- the ability to perform exploratory analyses of spatial data using statistical information stored in the Portal; using the available tools users can examine the spatial distribution of selected variables and determine spatial connections, interdependencies and identify clusters. A variety of descriptive statistics and statistical methods are available (e.g. central tendency statistics, dispersion statistics, measures of asymmetry and concentration, variable correlation analyses, etc.). It is also possible to perform cluster analysis and check spatial autocorrelation and similarity of objects,

- the ability to conduct analyses involving geostatistical modelling. e.g. to generalize/ estimate results based on a random sample to other surveyed units or the population of these units. Users can create models and apply a probabilistic model for statistical inference (estimation) concerning values of the response variable based on results of a random sample survey and the assumed probability distribution. There are statistics and tests that can be used to verify the quality of these models as well as some spatial interpolation and imputation methods,

- the ability to enhance user's own content with geostatistical information and analyses provided by the Portal (semantic access to documents related to the analytical work and the ability to supplement user's own text-based content with graphical elements). Additionally, advanced users can use programming languages to access the Portal's databases via the API.

All of these functionalities can support users in decision-making processes related to statistical and spatial information and enable them to benefit from spatial and data mining analyses, either in the context of business activity, or in policy-making by government and local government administration, or in scientific research.

However, before any such detailed statistical data can be released to enable advanced analyses, they have to undergo statistical disclosure control to protect data confidentiality. Apart from satisfying legal requirements, it is necessary to apply additional tools to minimize the risk of potential identification of individual units and unauthorized disclosure of sensitive information about them. Given the complexity of information provided to Portal users, the kind data from other sources they may have access to and the sophistication of their analyses, the SDC process should be conducted thoroughly by competent staff.

Outputs of any analyses conducted in the Portal using data designated as protected (i.e. from internal statistical databases), including map visualizations, have to be checked in terms of primary confidentiality. According to this requirement, values of aggregates can only be displayed (visualized) if they contain a sufficiently large number of units (data records) – at least 3 (it is the fundamental rule established in the Polish Act on Official Statistics), and, in some cases, at least 10. However, in some situations aggregate values suppressed to protect statistical confidentiality could be recalculated by the user on the basis of correlations between results of various analyses (queries). For example, if a higher-order aggregate consists of several lower-order groupings and the value of only one of them is hidden (because it would violate the statistical confidentiality), the hidden value can be determined by subtracting the sum of the displayed components (lower-order groupings) from the value of the higher-order aggregate. Such situations require secondary confidentiality. Normally, this is achieved by additionally suppressing aggregates which apparently (from the point of view of primary confidentiality) do not violate the protection rules, but allow the protected values to be recalculated through the use of indirect dependencies. Because the tools available in the Portal system are flexible and diverse, they enable users to analyze and aggregate data in any way and not be limited to pre-defined formats, it is not possible to create algorithms that will reliably control secondary confidentiality at the stage of presenting analytical results/ data summaries by hiding appropriate aggregates.

For this reason, any analyses based on a protected set of data that are to be released to external users are not performed on the original set of microdata, but on a set of data subjected to data distortion techniques designed to protect statistical confidentiality. Therefore, although users do not get direct access to unit-level data in the system (they can only see aggregates containing at least the minimum number of units), because of the flexibility offered by the aggregation tools which are associated with a a high risk of recalculating information about individual units based on the dependencies between the data, they can work (i.e. perform self-defined aggregations) only on datasets that have been disturbed by appropriate SDC methods. In this way, even if the user is able to recalculate values pertaining to individual units (records) in the disturbed set, this should not result in the disclosure of actually protected information, unless the user relies on individual data from other studies (e.g. registers of labor offices and the Labour Force Survey). Then, the risk of revealing sensitive information by linking relevant records from different sources may increase. In such situations, it may be necessary to carry out an additional - joint -

verification of the provided files in terms of statistical confidentiality. This will also be necessary, if, in the future, users of the system are able to use their own, external data sources.

In summary, under the adopted approach, only disturbed sets of publicly available unit-level data can made available for analysis. To implement this form of protection, the project team created a parameterized script to perform perturbations involving SDC methods.

Perturbation cannot be performed automatically. Each set of statistical data to be released to external users through the system must be perturbed separately. The disruption process (which may have to be repeated in the event of data update) must be performed by an analyst with a knowledge of the specific dataset and SDC methods. In each case the operation involves creating an appropriate script based on the template provided by the contractor, who should define the role of individual variables of the input set in the disturbance process and the method parameters.

Therefore, the SDC process in the system is enabled by an R script, which relies on functions implemented in the `sdcMicro` package (Templ et al. (2015)). The functions are used to apply specific perturbative methods and control the disclosure risk and information loss. Although the script relies on two main families of perturbations, it can be adapted to include other methods, if necessary. The following sections describe the data used for testing and details of the script.

3. Microdata from the 2020 Agricultural Census

Our analysis was based on a dataset containing microdata collected during the Agricultural Census conducted in Poland between 1st September to 30th November 2020, with reference to 1st June 2020. The data are to be made available through the Geostatistics Portal, and, in other forms, to all interested persons, especially for scientific purposes. So they will have to be perturbed to prevent potential unit identification and disclosure of its sensitive information. The set in question contained 1,317,400 records and 81 variables. The following 12 variables describe the main features of farms:

- NR_GOS – farm ID,
- SP – legal status,
- Wo_SG – the province where the farm is located;
- Pow_SG – the district (LAU 1 unit) where the farm is located;
- Gm_SG – the commune (LAU 2 unit) where the farm is located;
- KTS1_SG – the macroregion (NUTS 1) where the farm is located;
- KTS3_SG – the region (NUTS 2) where the farm is located[5],
- KTS4_SG – subregion (NUTS 3) where the farm is located,

[5] In Poland regions coincide with the provinces except for the Mazowieckie Province, which is divided into two regions: the City of Warsaw and the rest of the province.

- UG2w – total land area,
- UG2a – area of agricultural land,
- UG_W1 – area of arable land,
- UG_W2 – area of permanent grassland.

The remaining 69 variables describe various aspects of the livestock population. They are presented in Table 1.

Table 1. Variables describing the livestock population in the analyzed dataset

Symbol	Description	Symbol	Description
ZW1	Breeding of farm animals (yes/no)	ZW45b	Number of laying hens for the production of table eggs
ZW2	Cattle breeding (yes/no)	ZW45c	Number of laying hens for the production of hatching eggs
ZW3w	Total cattle population	ZW45d	Number of turkeys
ZW3a	Number of bulls under 1 year of age	ZW45e	Number of geese
ZW3b	Number of heifers under 1 year of age	ZW45f	Number of ducks
ZW3c	Number of bulls aged 1 to 2 years (except for exactly 2-year-old bulls)	ZW45g	Number of remaining poultry
ZW3d	Number of heifers aged 1 to 2 years (except for exactly 2-year-old heifers)	ZW45h	Number of ostriches
ZW3e	Number of male cattle aged 2 years and over	ZW47	Number of horses
ZW3f	Number of heifers aged 2 years and over	ZW47a	Number of horses three years old and over
ZW3g	Number of dairy cows	ZW48	Total number of rabbits kept for meat
ZW3h	Number of other cows	ZW48a	Number of female rabbits capable of breeding

Symbol	Description	Symbol	Description
ZW34	Farm breeding pigs (yes/no)	ZW49	Number of other fur animals (including fur rabbits)
ZW35w	Total pig population	ZW49a	Number of remaining female fur animals
ZW35a	Number of piglets weighing up to 20 kg	ZW50	Number of bee trunks
ZW35b	Number of weaners weighing 20-50 kg	ZW51	Number of remaining animals
ZW35c	Number of breeding boars	ZW51a	Number of deer animals
ZW35d	Number of pregnant sows	ZW_W1_3	Number of calves under 1 year of age
ZW35e	Number of sows pregnant for the first time	ZW_W2_3	Number of cattle aged 1-2 years
ZW35f	Number of remaining sows (loose, not pregnant)	ZW_W3_3	Number of cattle aged 2 years and over
ZW35g	Number of gilts has never been bred	ZW_W4_3	Total number of cows
ZW35h	Number of pigs for fattening	ZW_W5_3	Number of female cattle aged 2 years and over
ZW40	Sheep breeding (yes/no)	ZW_W1_35	Number of pigs for breeding weighing 50 kg and more
ZW41w	Total number of sheep	ZW_W2_35	Total number of breeding sows
ZW41a	Number of sheep lambs	ZW_W1_41	Total number of sheep ewes
ZW41b	Number of sheep ewes used for milk production	ZW_W2_41	Total number of adult sheep
ZW41c	Number of sheep ewes used in other directions	ZW_W1_45	Total number of chicken poultry
ZW41d	Number of remaining adult sheep	ZW_W2_45	Total number of laying hens
ZW42	Goat breeding (yes/no)	ZW_W_SD	Animal population in LSUs[a]
ZW43w	Total goat population	ZW_W1_SD	Number of cattle in LSUs
ZW43a	Number of female goats one year old and older	ZW_W2_SD	Number of pigs in LSUs

Symbol	Description	Symbol	Description
ZW43b	Number of female goats used for milk production	ZW_W3_SD	Number of sheep in LSUs
ZW43c	Number of remaining goats	ZW_W4_SD	Number of goats in LSUs
ZW44	Poultry breeding (yes/no)	ZW_W5_SD	Number of poultry in LSUs
ZW45w	Total poultry population	ZW_W6_SD	Number of rabbits in LSUs
ZW45a	Number of broiler chickens		

[a] The livestock unit, abbreviated as LU (or sometimes as LSU - Livestock Standard Unit), means a standard measurement unit that allows the aggregation of various categories of livestock (various species, sex and age) in order to enable them to be compared. Data on animals are converted into livestock units using the following coefficients: equidae – 0.80, young cattle aged less than 1 year old (calves) – 0.40, male bovines aged between 1 and 2 years – 0.70, female bovines aged between 1 and 2 years – 0.70, male bovines aged 2 years and over – 1.00, heifers of bovines aged 2 years and over – 0.80, dairy cows – 1.00, other cows (sucklers) – 0.80, sheep – 0.10, goats – 0.10, piglets with a live weight of less than 20 kg – 0.027, breeding sows with a live weight of 50 kg or more – 0.50, other pigs (young pigs with a live weight of 20 kg or more but less than 50 kg, breeding boars and fattening pigs with a live weight of 50 kg and more) – 0.30, broilers of chickens – 0.007, laying hens – 0.014, other poultry (ducks, turkeys, geese, domestic quails, guinea-fowls, and other poultry but apart from ostriches) – 0.030, ostriches – 0.35 and female of rabbits – 0.020. The reference unit used for the calculation of livestock units (=1 LSU) is the grazing equivalent of one adult dairy cow producing 3 000 kg of milk annually, without additional concentrated foodstuffs.

Source: Based on the metadata for the dataset and information provided by Statistics Poland (https://stat.gov.pl/en/metainformation/glossary/terms-used-in-official-statistics/1394,term.html).

14 of the above variables are categorical (NR_GOS, SP, Wo_SG, Pow_SG, Gm_SG, KTS1_SG, KTS3_SG, KTS4_SG, ZW1, ZW2, ZW34, ZW40, ZW42 and ZW44). The remaining 67 variables are numerical.

The following 26 variables are derived from 55 primary ones: (KTS1_SG, KTS3_SG, KTS4_SG, ZW3w, ZW35w, ZW41w, ZW43w, ZW45w, ZW_W1_3, ZW_W2_3, ZW_W3_3, ZW_W4_3, ZW_W5_3, ZW_W1_35, ZW_W2_35, ZW_W1_41, ZW_W2_41, ZW_W1_45, ZW_W2_45, ZW_W_SD, ZW_W1_SD, ZW_W2_SD, ZW_W3_SD, ZW_W4_SD, ZW_W5_SD, ZW_W6_SD).

These facts were taken into account when planning the SDC process. The next section contains a description of how this information was used to determine the set of quasi-identifiers to be protected and choose appropriate SDC methods.

4. Methods and tools of statistical disclosure control

The main problem in defining successive steps of the SDC process was to identify a set of quasi-identifiers that need to be protected. First, the 26 derived variables were excluded from further analysis because any perturbations applied to these variables could cause significant deviations from their original dependencies on primary variables. For instance, the value of ZW43w is the sum of the values of ZW43a, ZW43b and ZW43c. Hence, additivity of these cells should be retained in the safe dataset. Therefore, values of

these derived variables should be re-calculated *ex post*, i.e. after the whole SDC process has been completed. Of course, deviations on particular values of the variables on the basis of which a given derived variable is obtained can accumulate in this way. However, variables are derived at the level of units (not aggregated data, as in tables), so the final accumulation of deviations should be rather low.

There is a group of key quasi-identifiers that describe a given unit's geographical location. These are: Wo_SG, Pow_SG and Gm_SG. Since each of these variables contains only unit codes for a given level (the codes do not contain symbols denoting higher level units) the exact location can only be obtained only by concatenating codes in Wo_SG, Pow_SG and Gm_SG. However, since we allow record swapping between communes (LAU2), we have replaced Wo_SG and Pow_SG by their concatenation, denoted as GEO_ID.

Thus, the variables under analysis are: NR_GOS, GEO_ID, Gm_SG, UG2w, UG2a, ZW1, ZW2, ZW3a, ZW3b, ZW3c, ZW3d, ZW3e, ZW3f, ZW3g, ZW3h, ZW34, ZW35a, ZW35b, ZW35c, ZW35d, ZW35e, ZW35f, ZW35g, ZW35h, ZW40, ZW41a, ZW41b, ZW41c, ZW41d, ZW42, ZW43a, ZW43b, ZW43c, ZW44, ZW45a, ZW45b, ZW45c, ZW45d, ZW45e, ZW45f, ZW45g, ZW45h, ZW47, ZW47a, ZW48, ZW48a, ZW49, ZW49a, ZW50, ZW51 and ZW51a.

Categorical variables were perturbed using microaggregation based on Gower's distance (first described in a PhD thesis by Kowarik (2015) and later also by Templ (2017)). In this method records are combined to form a number of groups. Then, the true value of each sensitive attribute is replaced by a value representing a certain measure of central tendency of this attribute (e.g. mode or mean) for the group a given record belongs to. Groups are formed using a criterion of maximum similarity. Gower's distance is used to compute the distance between any two records, taking into account all measurement scales of variables. Clusters for which microaggregation was to be conducted were established using the variable GEO_ID. Therefore, microaggregation was performed within districts (LAU 2). The Gower's distance was computed using the following variables: UG2w, UG2a, ZW1, ZW2, ZW34, ZW40, ZW42 and ZW44. The mechanism of microaggregation was defined by the `maxCat` function, i.e. the level with the most occurrences is normally chosen or the selection is random if the maximum is not unique. The aggregation level was adjusted for the properties of the analyzed dataset. It is an efficient method of perturbing variables whose values are expressed on various measurement scales, since it offers several possibilities of choosing the form of perturbation and its parameters and its results are easy to interpret. These features give it an advantage over other methods[6]. On the other hand, the method of perturbing continuous variables was chosen because it ensures that relationships between them are retained as much as possible and it reduces the impact of outliers better than many other approaches, which is consistent with basic expectations of users of disclosed data. Of course, the method is slightly sophisticated (but its results are easy to interpret) and might flatten the original distributions (which can be controlled to some extent).

Continuous variables were perturbed using correlated noise addition. The approach involves adding random values selected from a continuous distribution while preserving

[6]In the general script, an alternative use of post-randomization (PRAM) for perturbing categorical variables is available. However, the efficient setting of necessary entries in the transition matrix is more difficult.

the structure of covariances of the original variables and assuring, by way of additional transformations, that the sample covariance matrix of the suppressed variables is an unbiased estimator for the covariance matrix of the original variables (cf. e.g. Kim (1986) or Brand (2002)). The basic parameter δ and the amount of noise were optimized in a series of trials.

However, the algorithm offers the possibility of using other perturbation methods that are better suited to data with different properties or different user expectations. These other options include post-randomization for categorical variables and microaggregation for continuous variables. The algorithm is an R script and relies on functions from the `sdcMicro` package. To be more precise, the function `microaggrGower` was used to apply microaggregation based on the Gower distance to categorical variables. Noise was added to continuous variables using the `addNoise` function. The risk of disclosure was computed by setting relevant parameters of the `sdcMicroObj` object. The `IL_variables` function was used to assess information loss regarding the distribution and the `IL_correl` function was applied to estimate information loss.

5. Results and problems encountered during the exercise

The algorithm took almost 35 hours to complete its run, which mainly resulted from the large number of variables, the complexity of the script and the limitations of the computational environment.

The dataset under analysis contains a few categorical variables. Categories with smallest frequency appear in more than 6 thousands records. Therefore, it is not surprising that the k-anonymity rules for $k = 2,3$ and 5 are practically not violated (it is, of course, not a rule; however, the higher the frequency of the "smallest" categories, the lower the probability that the rare combinations occur). Therefore, the risk associated with categorical variables is negligible.

The situation looks very different for the continuous variables. In this case, the risk of disclosure is assessed using the basic function implemented in the sdcMicro package and described by Templ (2017). The function reports the percentage of observations falling within an interval centered on its masked value, whereas the upper bound of such an interval corresponds to the worst case scenario in which an intruder is sure that each nearest neighbor is indeed the true link. The function compares data before and after the SDC process. For raw data the risk – by definition expressed as a percentage – is always in the range between 0% and 100%. The computation showed that after the SDC process the risk interval ranged from [0.00%,100.00%] to [0.00%,0.00%]. Thus, the protection is ideal.

To get a full picture of the efficiency of the SDC process, it is necessary to measure the loss of information. In this experiment, it was assessed in two ways:

- by measuring the distribution disturbance,
- by measuring the disturbance of correlations between the variables.

The measure of distribution disturbance was proposed by Młodak (2020), improved by Młodak et al. (2022) and implemented in the `sdcMicro` package as the `IL_variables` function. It is based on Gower's distance between original and perturbed values and is

defined as the sum of partial distances. In the case of nominal variables, these partial distances amount to 0 if they are the same and 1 otherwise; in the case of ordinal variables, they are equal to the normalized number of categories by which the compared values differ, and in the case of continuous variables, they are computed using the cyclometric function. This measure takes values from [0,1]. The larger the value of the measure, the bigger the loss of information. Information loss can be measured both at the global level and for particular variables. Table 2 shows information loss computed for particular variables.

Table 2. Information loss for particular variables (in %)

Variable	Information loss	Variable	Information loss
NR_GOS	0.0	ZW41c	56.3
Gm_SG	0.0	ZW41d	39.0
UG2w	92.5	ZW42	0.1
UG2a	92.1	ZW43a	32.2
ZW1	0.0	ZW43b	24.7
ZW2	0.0	ZW43c	8.1
ZW3a	64.7	ZW44	0.0
ZW3b	66.0	ZW45a	99.8
ZW3c	59.6	ZW45b	99.9
ZW3d	62.1	ZW45c	99.2
ZW3e	27.3	ZW45d	99.4
ZW3f	49.3	ZW45e	98.1
ZW3g	77.2	ZW45f	98.6
ZW3h	47.8	ZW45g	94.9
ZW34	0.0	ZW45h	3.4
ZW35a	95.9	ZW47	32.8
ZW35b	96.1	ZW47a	21.6
ZW35c	9.1	ZW48	89.2
ZW35d	88.5	ZW48a	64.2
ZW35e	68.6	ZW49	99.5
ZW35f	82.0	ZW49a	97.4

Variable	Information loss	Variable	Information loss
ZW35g	74.4	ZW50	75.7
ZW35h	96.5	ZW51	82.3
ZW40	0.1	ZW51a	39.6
ZW41a	49.0	GEO_ID	0.0
ZW41b	42.7	GEO_ID_G	0.0

Source: Results obtained by applying the `IL_variables()` function from the `sdcMicro` package.

The variable GEO_ID_G was created for technical reasons by concatenating symbols for province, district and commune. The overall information loss amounts to 53.8%. As can be seen, the level of information loss for particular variables varies greatly. Of course, some variables (e.g. ID_GOS or GEO_ID) could not be changed because of the underlying assumptions of the SDC process. Nevertheless, information loss for the remaining ones varies significantly – from 0.0 to as much as 99.9%. This may be the result of adjustments in the amount of correlated noise in the case of the continuous variables and the fact that some perturbed values may go beyond the range defined for a given variable (and hence some *ex post* corrections will be necessary). On the other hand, however, the distance component for continuous variables (based on the cyclometric function – arcus tangent) tends to take values close to 1 (100%) for larger differences between original and perturbed values. As a result, information loss can be overestimated. On the other hand, information loss can also be overestimated when the original range of values is exceeded as a result of perturbations. As we have noted in Section 6, these inconveniences can be corrected *ex post*, which should reduce this problem. But such overestimation could be helpful when identifying problem areas in the SDC process.

Information loss has some impact on the descriptive statistics of the analyzed variables. Table 3 shows the mean, median and third quartile of primary continuous variables before and after the SDC process.

Table 3. Basic descriptive statistics for primary continuous variables before and after the SDC process

Variable	Original			After SDC		
	mean	median	3rd quartile	mean	median	3rd quartile
UG2w	12.6530	5.6300	11.8600	20.5326	10.4000	32.1400
UG2a	11.3503	4.6900	10.2800	18.8962	9.2500	29.5400
ZW3a	0.6535	0.0000	0.0000	1.8686	0.0000	3.0000
ZW3b	0.6582	0.0000	0.0000	1.9537	0.0000	3.0000

Variable	Original			After SDC		
	mean	median	3rd quartile	mean	median	3rd quartile
ZW3c	0.6833	0.0000	0.0000	1.6146	0.0000	2.0000
ZW3d	0.6492	0.0000	0.0000	1.7089	0.0000	3.0000
ZW3e	0.1009	0.0000	0.0000	0.3713	0.0000	1.0000
ZW3f	0.1628	0.0000	0.0000	0.8382	0.0000	1.0000
ZW3g	1.6839	0.0000	0.0000	4.0118	1.0000	6.0000
ZW3h	0.1972	0.0000	0.0000	0.8302	0.0000	1.0000
ZW35a	1.7826	0.0000	0.0000	27.8934	1.0000	46.0000
ZW35b	2.5555	0.0000	0.0000	30.1932	1.0000	49.0000
ZW35c	0.0107	0.0000	0.0000	0.0993	0.0000	0.0000
ZW35d	0.4303	0.0000	0.0000	7.4268	0.0000	12.0000
ZW35e	0.0784	0.0000	0.0000	1.7361	0.0000	3.0000
ZW35f	0.1875	0.0000	0.0000	3.9925	0.0000	7.0000
ZW35g	0.0396	0.0000	0.0000	2.3313	0.0000	4.0000
ZW35h	3.4979	0.0000	0.0000	34.7676	2.0000	56.0000
ZW41a	0.0531	0.0000	0.0000	0.7486	0.0000	1.0000
ZW41b	0.0394	0.0000	0.0000	0.5812	0.0000	1.0000
ZW41c	0.0910	0.0000	0.0000	1.0270	0.0000	2.0000
ZW41d	0.0433	0.0000	0.0000	0.5094	0.0000	1.0000
ZW43a	0.0291	0.0000	0.0000	0.3814	0.0000	1.0000
ZW43b	0.0158	0.0000	0.0000	0.2702	0.0000	0.0000
ZW43c	0.0119	0.0000	0.0000	0.0911	0.0000	0.0000
ZW45a	106.7432	0.0000	0.0000	1258.1000	28.0000	2003.0000
ZW45b	34.6117	0.0000	7.0000	2200.8400	19.0000	3693.0000
ZW45c	7.3861	0.0000	0.0000	203.2564	1.0000	334.0000
ZW45d	13.3664	0.0000	0.0000	248.3547	4.0000	403.0000
ZW45e	4.2672	0.0000	0.0000	69.5504	1.0000	113.0000
ZW45f	4.4782	0.0000	0.0000	96.2474	1.0000	158.0000

Variable	Original			After SDC		
	mean	median	3rd quartile	mean	median	3rd quartile
ZW45g	0.6233	0.0000	0.0000	20.7730	0.0000	35.0000
ZW45h	0.0016	0.0000	0.0000	0.0358	0.0000	0.0000
ZW47	0.1188	0.0000	0.0000	0.4614	0.0000	1.0000
ZW47a	0.0694	0.0000	0.0000	0.2764	0.0000	0.0000
ZW48	0.5543	0.0000	0.0000	8.0797	0.0000	14.0000
ZW48a	0.1079	0.0000	0.0000	1.4254	0.0000	2.0000
ZW49	3.3246	0.0000	0.0000	297.2698	1.0000	498.0000
ZW49a	0.6726	0.0000	0.0000	46.6540	0.0000	78.0000
ZW50	0.4925	0.0000	0.0000	2.8839	0.0000	4.0000
ZW51	0.0873	0.0000	0.0000	3.9979	0.0000	7.0000
ZW51a	0.0184	0.0000	0.0000	0.5001	0.0000	1.0000

Source: Results obtained using the SAS Studio software.

As one can see, in most cases the SDC process did not significantly change the presented statistics. Moreover, the original first quartile was 0 except for UG2w (2.8600) and UG2a (2.320), whereas after perturbation this quartile for all variables amounted to 0. However, for some variables – e.g. ZW35a, ZW35b, ZW45a, ZW45b and ZW45c – the differences are more significant. This situation can be due to a large degree of variation in relevant data across various farms and spatial areas, which can have an impact on the noise distribution adjusted to preserve the correlation, according to our assumptions.

The loss of information resulting from the disturbance of correlations between variables, which reflects the degree to which relationships between variables have been preserved, is measured using the approach developed by Młodak (2020), improved by Młodak et al.(2022) and implemented in the `sdcMicro` package as the `IL_correl` function. It is based on distances of normalized sums of diagonal entries of an inverse correlation matrix and takes values from [0,1] (again, the larger the value, the bigger the loss). In the analyzed situation the measure amounts to 7.9%. Therefore, the loss of information about relationships is small. This is largely the result of using correlated noise. Thus, in this respect, SDC seems to be fully efficient.

6. Final conclusions

Statistical disclosure control is necessary to ensure the safe and efficient disclosure of statistical information. It is worth emphasizing that without the use of these methods, much

of statistical data would either have to remain unavailable to end users or would largely be useless.

The above exercise indicates that perturbative SDC methods can be very useful, especially if most variables in a given dataset are numerical. As a result, the risk of disclosure associated with these variables is significantly reduced. If there are few categorical variables, then the risk of disclosure associated with them tends to be low (or even negligible).

However, the risk of disclosure is reduced at the cost of some information loss. Perturbations introduced in some variables result in large differences between their original distributions and those after the SDC process. This happens because perturbations cannot preserve some features of the original variables resulting from their definitions, e.g. the range of permitted values. Therefore, additional corrections may be required. On the other hand, in the finally disclosed dataset secondary variables have to be determined (to avoid violations of additivity or related rules, it is reasonable to omit them in the SDC process and to compute them again after it is over), which can have some (rather moderate in the case of microdata) impact on the final information loss.

The use of correlated noise in relation to continuous variables with appropriately chosen parameters results in a small loss of information about relationships between them. So, it is a very important aspect of disclosed data. When appropriate summations of continuous variables are performed to derive secondary variables, these interactions should not be violated.

The algorithm can be a good tool for performing SDC on microdata. However, its application reveals the whole complexity of the process, especially as regards steps that have to be taken before and after the perturbation procedure in order to obtain an output that is efficiently protected and simultaneously sufficiently useful for its users. Thus, each stage of this procedure should be treated with equal care. Of course, it is possible to consider some dynamic methods of protecting data confidentiality. Data in the geostatistics portal will be available as microdata, so SDC methods for tabular data, such as cell-key adjustment, will not be appropriate. Nonetheless, the use of other dynamic SDC tools can be an interesting challenge for future research, which can focus, e.g. on reducing the computational overload, which is unavoidable in the case of such large files.

References

Abowd, J. M., (2018). The US Census Bureau adopts differential privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, p. 2867.

Brand, R., (2002). Microdata protection through noise addition. *Inference Control in Statistical Databases: From Theory to Practice*, pp. 97–116.

Calian, V., (2020). Methods of statistical disclosure control for aggregate data with a case study on the new Icelandic geospatial system of statistical output areas. *Working Papers of Statistics Iceland*, 105(6).

Jackson, J., Mitra, R., Francis, B., and Dove, I., (2024). Obtaining (ε, δ) - Differential Privacy Guarantees When Using a Poisson Mechanism to Synthetize Contingency Tables. *Privacy in Statistical Databases: International Conference, PSD 2024, Antibes Juan-les-Pins, France, September 25–27, 2024. Proceedings*, pp. 102–112.

Jansson, I., (2012). Issues and plans for the disclosure control of the Swedish Census 2011. *En Workshop on Statistical Disclosure Control of Census Data*, Luxembourg.

Kim, J. J., (1986). A method for limiting disclosure in microdata based on random noise and transformation. *Proceedings of the Section on Survey Research Methods*, pp. 303–308.

Kowarik, A., (2015). *New computational tools and methods for official statistics* [Doctoral dissertation, Technische Universitat Wien].

Kraus, J., (2021). Statistical Disclosure Control methods for Harmonised Protection of Census Data: A Grid Case. *Demografie*, 63(4), pp. 199–215.

Młodak, A., Pietrzak, M., and Jozefowski, T., (2022). The trade-off between the risk of disclosure and data utility in SDC: A case of data from a survey of accidents at work. *Statistical Journal of the IAOS*, 38(4), pp. 1503–1511.

Młodak, A., (2020). Information loss resulting from Statistical Disclosure Control of output data [(in Polish)]. *Wiadomości Statystyczne. The Polish Statistician*, 65(09), pp. 7–27.

Muralidhar, K., Domingo-Ferrer, J., (2023). A Rejoinder to Garfinkel (2023) – Legacy Statistical Disclosure Limitation Techniques for Protecting 2020 Decennial US Census: Still a Viable Option. *Journal of Official Statistics*, 39(3), pp. 411–420.

Shlomo, N., (2022). How to Measure Disclosure Risk in Microdata? *The Survey Statistician*, 86, pp. 13–21.

Shlomo, N., Skinner, C., (2022). Measuring risk of re-identification in microdata: State-of-the art and new directions. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 185(4), pp. 1644–1662.

Shlomo, N., Tudor, C., and Groom, P., (2010). Data swapping for protecting census tables. *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2010, Corfu, Greece, September 22–24, Proceedings*, pp. 41–51.

SNI et al., (2022). *Census 2021 Statistical Disclosure Control Methodology*. Northern Ireland Statistics & Research Agency.

Templ, M., (2017). *Statistical Disclosure Control for Microdata. Methods and Applications in R*. Springer International Publishing AG, Cham, Switzerland.

Templ, M., Kowarik, A., and Meindl, B., (2015). Statistical Disclosure Control for Micro-Data Using the R Package `sdcMicro`. *Journal of Statistical Software*, 67(4), pp. 1–36.

Tran, T., Reimherr, M., and Slavkovic, A., (2024). Differentially private quantile regression. *Privacy in Statistical Databases: International Conference, PSD 2024, Antibes Juan-les-Pins, France, September 25–27, Proceedings*, pp. 18–34.

Yoon, J., Drumright, L. N., and Van Der Schaar, M., (2020). Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8), pp. 2378–2388.

Zayatz, L., (2002). SDC in the 2000 US Decennial Census. In *Inference Control in Statistical Databases: From Theory to Practice*, pp. 193–202. Springer.

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 39–56, https://doi.org/10.59139/stattrans-2025-015
Received – 20.08.2023; accepted – 30.10.2024

Formulation of estimator for population mean in stratified successive sampling using memory-based information

Sanjoy Majumder[1], Arnab Bandyopadhyay[2], Arindam Gupta[3]

Abstract

In study described in this article, we developed a memory type estimator for the population mean in stratified successive sampling. We used the past sample information together with the current sample information through hybrid exponentially weighted moving averages statistics. We have also used the information available on auxiliary variable to construct the proposed estimator. We studied the properties of the proposed estimator. Further, we examined the performance of the proposed estimator in comparison with conventional estimator of the population mean and the results are demonstrated by using the data set of simulation as well as natural population. After observing the auspicious findings, we suggest that the proposed estimator can be applied to solve real-life problems.

Key words: successive sampling, HEWMA, regression estimator, variance, minimum variance, efficiency.

1. Introduction

In the context of socio-economic surveys, our society is composed of various classes of individuals like business owners, salaried persons, daily wages labourers, etc., and these distinct classes have various levels of income and expenditure. Therefore, when assessing these socio-economic parameters, it is needed to categorize these heterogeneous units of the population into different homogenous strata based on their socio-economic status, which necessitates the application of stratified random sampling techniques. It may also be noticed that socio-economic factors such as income and expenditure are changing over a period of time. When the characteristics are liable to change over time, successive sampling is the most preferred method for estimating

[1] Department of Mathematics and Statistics, Aliah University, Kolkata-700160, India. E-mail: sanjoybiostat@gmail.com. ORCID: https://orcid.org/0000-0001-8320-003x

[2] Department of Basic Science and Humanities (Mathematics), Dr. B. C. Roy Engineering College, Durgapur 713206, India. E-mail: arnab.bandyopadhyay4@gmail.com. ORCID: https://orcid.org/0000-0002-0769-7491.

[3] Department of Statistics, The University of Burdwan, Burdwan-713104, India. E-mail: guptaarin@gmail.com. ORCID: https://orcid.org/0000-0001-9381-5150

the population parameters at different points of time along with measuring the change over a period of time. Consequently, estimating these socio-economic parameters for various classes of individuals, stratified successive sampling may be the effective technique. Jessan (1942) was first introduced successive sampling under simple random sampling and later it was further developed by Patterson (1950), Rao and Graham (1964), Sen (1971), Sen (1972, 1973), Das (1982), Chaturvedi and Tripathi (1983) and many others. This work was further developed by, Singh and Singh (2001), Singh (2003) among others.

In many circumstances, information regarding auxiliary variable is easily available on both the first and second occasions. Using the information on auxiliary variable (first and second occasions) Feng and Zou (1997) suggested an estimator for population mean on current occasion. It was further extended by Birader and Singh (2001), Singh and Karna (2009), Singh and Vishwakarma (2009), Singh *et al.* (2011) and many others, who suggested estimation procedures for population mean on current (second) occasions in successive sampling (two occasions).

It is common practice in successive sampling that we treat the variables y and x as the character under study at current occasion and previous occasion respectively. Therefore, we estimate the parameters on y and x based on information gathered on single occasion i.e., current occasion and previous occasion respectively. It is obvious that the use of information gathered from past samples at different points of time (i.e., occasions) along with the information on the current occasion improves the performance of the estimator. It is noted that hybrid exponential weighted moving average statistics (HEWMA) help us in developing the estimator for the parameter on the variable y at the current occasion (i.e., y_t at time t) based on the information from the previous occasion such as y_{t-1} from t-1 and y_{t-2} from time t-2 and so on. It may be observed that several authors including Noor-ul-Amin (2020, 2021) and Aslam *et al.* (2020, 2023), Bhushan *et al.* (2023) developed effective estimation technique using HEWMA statistics for the population parameters in sample surveys. The utilization of HEWMA statistics may be proven to be commendable for estimating population parameter in successive sampling where collection of information at different points of time is necessary.

However, it may be noted that almost no attempt has been done for estimating population mean in stratified successive sampling using memory-based information. Motivated with these arguments, we have formulated a memory type estimator for population mean in stratified successive sampling and we examined detailed properties of the proposed estimator through empirical investigation carried over the data set of simulation as well as natural population.

2. Sample structure

Let $U = \{U_1, U_2, U_3, \ldots, U_N\}$ be finite population of size N which splits into K non overlapping strata or homogeneous strata, with each stratum containing N_h (h = 1, 2, . . ., K) units such that $N = \sum_{h=1}^{K} N_h$. The population U has been sampled over two occasions. The characters under study are denoted by x and y at the first and second (current) occasion respectively. A simple random sample S_{n_h} (WOR) of n_h units is drawn from each of the h^{th} stratum having N_h (h = 1, 2, . . ., K) units of the first occasion. A random subsample S_{m_h} (h = 1, 2, …, K) of m_h units is retained (matched) from each of the n_h units of the sample S_{n_h} (h = 1, 2, . . ., K) for its use on the second occasion. Again, a fresh simple random sample S_{u_h} (WOR) of u_h ($n_h - m_h$) units is drawn on the second occasion from each of the N_h unit of the population so that the sample size on the second occasion is also n_h ($n_h = u_h + m_h$; h = 1, 2, . . ., K).

Let y, x and z be the study variable at current occasion, study variable at previous occasion and auxiliary variable which is stable over occasions respectively taking values y_{hi}, x_{hi} and z_{hi} for the i^{th} unit (i = 1, 2, . . ., N_h) of the h^{th} stratum (h = 1, 2, . . ., K).

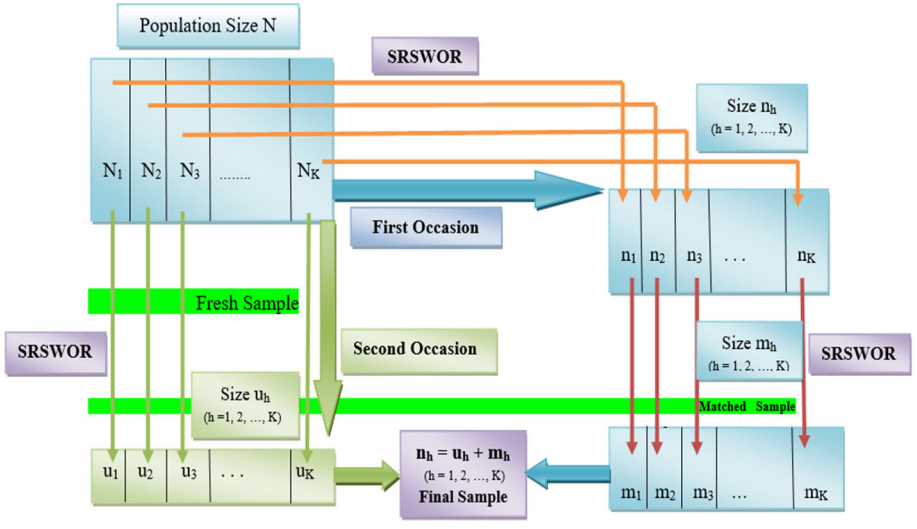


Figure 1: Sample structure for two-occasion stratified successive sampling

Henceforth, we use the following notations:

$\bar{Y}_h = \sum_{i=1}^{N_h} \frac{y_{hi}}{N_h}, \bar{X}_h = \sum_{i=1}^{N_h} \frac{x_{hi}}{N_h}, \bar{Z}_h = \sum_{i=1}^{N_h} \frac{z_{hi}}{N_h}$: Population means of the respective variables on the h^{th} stratum (h = 1, 2, . . ., K).

$W_h = \frac{N_h}{N}$: The original weight of the h^{th} stratum (h = 1, 2, . . ., K).

$\bar{Y} = \sum_{h=1}^{K} \bar{Y}_h W_h, \bar{X} = \sum_{h=1}^{K} \bar{X}_h W_h, \bar{Z} = \sum_{h=1}^{K} \bar{Z}_h W_h$: Overall population means of the respective variables.

$\bar{y}_{m_h} = \frac{1}{m_h} \sum_{i=1}^{m_h} y_{hi}, \bar{x}_{m_h} = \frac{1}{m_h} \sum_{i=1}^{m_h} x_{hi}, \bar{z}_{m_h} = \frac{1}{m_h} \sum_{i=1}^{m_h} z_{hi}$: Sample means of the respective variables based on the sample S_{m_h} of size m_h on the h^{th} stratum (h = 1, 2, . . ., K).

$\bar{x}_{n_h} = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}, \bar{z}_{n_h} = \frac{1}{n_h} \sum_{i=1}^{n_h} z_{hi}$: Sample means of the respective variables based on the S_{n_h} of size n_h on the h^{th} stratum (h = 1, 2, . . ., K).

$\bar{z}_{u_h} = \frac{1}{u_h} \sum_{i=1}^{u_h} z_{hi}$: Sample mean of the variable z based on the sample S_{u_h} of size u_h on the h^{th} stratum (h = 1, 2, . . ., K).

$C_{y_h} = \frac{S_{y_h}}{\bar{Y}_h}, C_{x_h} = \frac{S_{x_h}}{\bar{X}_h}, C_{z_h} = \frac{S_{z_h}}{\bar{Z}_h}$: Coefficient of variations of the respective variables based on the h^{th} stratum (h = 1, 2, . . ., K).

$\rho_{yx_h}, \rho_{yz_h}, \rho_{zx_h}$: Correlation coefficients between the respective variables based on the h^{th}stratum (h = 1, 2, . . ., K).

$\beta_{yx_h}, \beta_{xz_h}$: Sample regression coefficients between the respective variables based on the h^{th} stratum $(h = 1, 2, \ldots, K)$.

$S_{y_h}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(y_{hi} - \bar{Y}_h \right)^2$: Population mean square of the variable y based on the h^{th} stratum (h = 1, 2, ..., K).

$S_{x_h}^2, S_{z_h}^2$: Population mean squares of the variables x and z based on the h^{th} stratum (h = 1, 2, …, K).

3. Conventional hybrid exponentially weighted moving average statistics (HEWMA)

Let us assume $X_1, X_2, \ldots, X_n$ are the independently and identically distributed random variables. From this we build a sequence $HE_1, HE_2, \ldots, HE_n$ such as

$$E_t = \gamma_2 \overline{X}_t + (1-\gamma_2)E_{t-1},\ 0 < \gamma_2 \leq 1 \tag{1}$$

$$HE_t = (1-\gamma_1)HE_{t-1} + \gamma_1 E_t,\ 0 < \gamma_1 \leq 1 \tag{2}$$

where γ_1 and γ_2 are real scalars and $\overline{X}_t$ is the mean of the variable X at the time t (i.e., current occasion). It is also noticed from the above-mentioned statistic that the efficiency of the estimator can be enhanced by incorporating the information from the current sample together with the information available from the previous samples such as from time t-1, from time t-2 and so on, and HE_t denotes the hybrid exponentially weighted moving average statistic which is based upon E_t (exponentially weighted moving average). It was first introduced by Roberts (1959) in control charting. The initial values of E_t and HE_t are taken as usual mean, which may be estimated from a pilot survey; it is considered as zero i.e., $HE_0 = Et_0 = 0$. This statistic was proposed by Haq (2013). The respective expected value and variance of HEWMA statistic evaluated by Haq (2016) and and Noor-ul-Amin (2020) is given by

$$E(HE_t) = \mu \tag{3}$$

$$\mathrm{Var}(HE_t) = \frac{(\gamma_1\gamma_2)^2}{(\gamma_1-\gamma_2)^2}\left[\frac{(1-\gamma_1)^2\{1-(1-\gamma_1)^{2t}\}}{1-(1-\gamma_1)^2} + \frac{(1-\gamma_2)^2\{1-(1-\gamma_2)^{2t}\}}{1-(1-\gamma_2)^2} - \frac{2(1-\gamma_1)(1-\gamma_2)\{1-(1-\gamma_1)^t(1-\gamma_2)^t\}}{1-(1-\gamma_1)(1-\gamma_2)}\right]\frac{\sigma^2}{n} \tag{4}$$

where $t \geq 1$, μ and σ^2 are the mean and variance of the variable of interest.

The limiting form of the variance is described as

$$Var(HE_t) = \frac{(\ _2\ _1)^2}{(\ _{2-}\ _1)^2}\left[\frac{(1-\gamma_1)^2}{1-(1-\gamma_1)^2} + \frac{(1-\gamma_2)^2}{1-(1-\gamma_2)^2} - \frac{2(1-\gamma_1)(1-\gamma_2)}{1-(1-\gamma_1)(1-\gamma_2)}\right]\frac{\sigma^2}{n} = \frac{(\ _2\ _1)^2}{(\ _{2-}\ _1)^2}R\frac{\sigma^2}{n} \tag{5}$$

where $R = \frac{(1-\gamma_1)^2}{1-(1-\gamma_1)^2} + \frac{(1-\gamma_2)^2}{1-(1-\gamma_2)^2} - \frac{2(1-\gamma_1)(1-\gamma_2)}{1-(1-\gamma_1)(1-\gamma_2)}$.

4. Proposed memory type estimator

Using the concept of HEWMA statistics for the variables x, y, z under stratified successive sampling, we have developed an estimation strategy in the following way:

a. HEWMA statistics for the variable y, based on the sample S_{m_h} of size m_h (h = 1, 2, …., K) has been defined as

$$E_{ty_{m_h}} = \gamma_2 \overline{y}_{tm_h} + (1-\gamma_2)E_{(t-1)y_{m_h}}$$
$$A'_{tm_h} = \gamma_1 E_{ty_{m_h}} + (1-\gamma_1)A'_{(t-1)m_h} \tag{6}$$

where $\overline{y}_{tm_h}$ indicates sample mean of the variable y, which is based on the sample S_{m_h} at time t.

b. HEWMA statistics for the variable x, which is based on the sample S_{m_h} of size m_h (h = 1, 2, …., K) has been defined as

$$E_{(t-1)x_{m_h}} = \gamma_2 \overline{x}_{(t-1)m_h} + (1-\gamma_2) E_{(t-2)x_{m_h}}$$
$$B'_{(t-1)m_h} = \gamma_1 E_{(t-1)x_{m_h}} + (1-\gamma_1) B'_{(t-2)m_h} \quad (7)$$

where $\overline{x}_{(t-1)m_h}$ indicates sample mean of the variable x, which is based on the sample S_{m_h} at time t-1.

c. HEWMA statistics for the variable x, which is based on the sample S_{n_h} of size n_h (h = 1, 2, …., K) has been defined as

$$E_{(t-1)x_{n_h}} = \gamma_2 \overline{x}_{(t-1)n_h} + (1-\gamma_2) E_{(t-2)x_{n_h}}$$
$$B''_{(t-1)n_h} = \gamma_1 E_{(t-1)x_{n_h}} + (1-\gamma_1) B''_{(t-2)n_h} \quad (8)$$

where $\overline{x}_{(t-1)n_h}$ indicates sample mean of the variable x, which is based on the sample S_{n_h} at time t-1.

d. HEWMA statistics for the variable y, which is based on the sample S_{u_h} of size u_h (h = 1, 2, …., K) has been defined as

$$E_{ty_{u_h}} = \gamma_2 \overline{y}_{tu_h} + (1-\gamma_2) E_{(t-1)y_{u_h}}$$
$$A''_{tu_h} = \gamma_1 E_{ty_{u_h}} + (1-\gamma_1) A''_{(t-1)u_h} \quad (9)$$

where $\overline{y}_{tu_h}$ indicates sample mean of the variable y, which is based on the sample S_{u_h} at time t.

The statistics draw up in equations (6), (7), (8) and (9) are unbiased estimators for the population mean $\overline{Y}_h$, $\overline{X}_h$, $\overline{X}_h$ and $\overline{Y}_h$ respectively.

Using the above HEWMA statistics defined in equation (9), we have constructed a memory type estimator for population mean based on the sample of size u_h taken afresh on the second occasion, which is presented below:

$$T'_{tu_h} = A''_{tu_h} \left[\frac{\overline{Z}_h}{\overline{z}_{u_h}} \right] \quad (10)$$

Therefore, the estimator of the overall population mean, i.e., $\overline{Y}$, which is based on the sample S_{u_h} of size u_h at time t, is described as

$$T_{tu} = \sum_{h=1}^{K} W_h T'_{tu_h} \quad (11)$$

where W_h is the stratum's weight.

Motivated by Kiregyra (1984) and using the above HEWMA statistics defined in equations (6), (7) and (8), we have constructed another memory type estimator for population mean based on the sample size m_h common for both the occasions, which is defined as

$$T''_{tm_h} = A'_{tm_h} + \beta_{yx_h}\left[\left(B''_{tn_h} - B'_{tm_h}\right) + \beta_{xz_h}\left(\overline{z}_{n_h} - \overline{Z}_h\right)\right] \tag{12}$$

Therefore, the estimator of the overall population mean, i.e., $\overline{Y}$, which is based on the sample S_{m_h} of size m_h at time t is described as

$$T_{tm} = \sum_{h=1}^{K} W_h T''_{tm_h} \tag{13}$$

where W_h is the stratum's weight.

Combining the estimator T_{tu} and T_{tm}, we get the final estimator of the population mean as follows:

$$T_{pe} = \emptyset T_{tu} + (1-\emptyset)T_{tm} \tag{14}$$

5. Variance of the proposed estimator T_{pe}

The variance of the proposed estimator T_{pe} is defined as

$$V\left(T_{pe}\right) = E\left(T_{pe} - \overline{Y}\right)^2 = E\left[\emptyset\left(T_{tu} - \overline{Y}\right) + (1-\emptyset)\left(T_{tm} - \overline{Y}\right)\right]^2$$

$$= \emptyset^2 \mathrm{V}(\mathrm{T_{tu}}) + (1-\emptyset)^2 \mathrm{V}(\mathrm{T_{tm}}) + 2\emptyset(1-\emptyset)\mathrm{Cov}(\mathrm{T_{tu}}, \mathrm{T_{tm}})$$

where $V(T_{tu}) = E\left(T_{tu} - \overline{Y}\right)^2, V(T_{tm}) = E\left(T_{tm} - \overline{Y}\right)^2$

and $Cov(T_{tu}, T_{tm}) = E\left(T_{tu} - \overline{Y}\right)\left(T_{tm} - \overline{Y}\right)$

where, T_{tu} and T_{tm} are based on two independent samples u and m, so their covariance term is zero., i.e., $Cov(T_{tu}, T_{tm}) = 0$

Therefore, $V\left(T_{pe}\right) = \emptyset^2 V(T_{tu}) + (1-\emptyset)^2 V(T_{tm})$ (15)

Variance of the estimator T_{tu} is defined as

$$\begin{aligned} V(T_{tu}) &= E\left(T_{tu} - \overline{Y}\right)^2 \\ &= \sum_{h=1}^{K} W_h^2 V\left(T'_{tu_h}\right) \\ &= \sum_{h=1}^{K} W_h^2 E\left(T'_{tu_h} - \overline{Y}_h\right)^2 \end{aligned} \tag{16}$$

Similarly, the variance of the estimator T_{tm} is defined as

$$\begin{aligned} V(T_{tm}) &= E\left(T_{tm} - \overline{Y}\right)^2 \\ &= \sum_{h=1}^{K} W_h^2 V\left(T''_{tm_h}\right) \\ &= \sum_{h=1}^{K} W_h^2 E\left(T''_{tm_h} - \overline{Y}_h\right)^2 \end{aligned} \tag{17}$$

To obtain the variance, we use the following transformation:

$$A'_{tm_h} = \overline{Y}_h(1+e_0),\ B'_{tm_h} = \overline{X}_h(1+e_1),\ A''_{tu_h} = \overline{Y}_h(1+e_2),$$
$$\overline{z}_{u_h} = \overline{Z}_h(1+e_3),\ B''_{tn_h} = \overline{X}_h(1+e_4),\ \overline{z}_{n_h} = \overline{Z}_h(1+e_5)$$

Such that

$$E(e_i) = 0 \text{ and } |e_i| < 1\ \forall\ i = 0,1,2,3,4,5.$$

The expected values of the parameters under the above transformation are obtained as

$$E(e_0^2) = f_{1h}C_{y_h}^2\frac{(\ _2\ _1)^2}{(\ _2-\gamma_1)^2}R, \qquad E(e_1^2) = f_{1h}C_{x_h}^2\frac{(\ _2\ _1)^2}{(\ _2-\gamma_1)^2}R, \qquad E(e_2^2) = f_{2h}C_{y_h}^2\frac{(\ _2\ _1)^2}{(\ _2-\gamma_1)^2}R,$$

$$E(e_3^2) = f_{2h}C_{z_h}^2\frac{(\ _2\ _1)^2}{(\ _2-\gamma_1)^2}R, \qquad E(e_4^2) = f_{3h}C_{x_h}^2\frac{(\ _2\ _1)^2}{(\ _2-\gamma_1)^2}R, \qquad E(e_5^2) = f_{3h}C_{z_h}^2\frac{(\ _2\ _1)^2}{(\ _2-\gamma_1)^2}R,$$

$$E(e_0e_1) = f_{1h}\rho_{yx_h}C_{y_h}C_{x_h}\frac{(\ _2\ _1)^2}{(\ _2-\gamma_1)^2}R, \qquad E(e_0e_4) = f_{3h}\rho_{yx_h}C_{y_h}C_{x_h}\frac{(\ _2\ _1)^2}{(\ _2-\gamma_1)^2}R, \tag{18}$$

$$E(e_0e_5) = f_{3h}\rho_{yz_h}C_{y_h}C_{z_h}\frac{(\ _2\ _1)^2}{(\ _2-\gamma_1)^2}R, \qquad E(e_1e_4) = E(e_4^2) = f_{3h}C_{x_h}^2\frac{(\ _2\ _1)^2}{(\ _2-\gamma_1)^2}R,$$

$$E(e_1e_5) = E(e_4e_5) = f_{3h}\rho_{xz_h}C_{x_h}C_{z_h}\frac{(\ _2\ _1)^2}{(\ _2-\gamma_1)^2}R,$$

$$E(e_2e_3) = f_{2h}\rho_{yz_h}C_{y_h}C_{z_h}\frac{(\ _2\ _1)^2}{(\ _2-\gamma_1)^2}R,$$

where,

$$f_{1h} = \frac{1}{m_h} - \frac{1}{N_h},\ f_{2h} = \frac{1}{u_h} - \frac{1}{N_h},\ f_{3h} = \frac{1}{n_h} - \frac{1}{N_h}$$

Expressing equation (10) in terms of e's we have

$$T'_{tu_h} = \overline{Y}_h(1+e_2)\left[\frac{\overline{Z}_h}{\overline{Z}_h(1+e_3)}\right]$$

Neglecting terms of e's having power greater than two, we have

$$T'_{tu_h} = \overline{Y}_h(1+e_2)\,(1-e_3+{e_3}^2)$$

It is found that

$$\left(T'_{tu_h} - \overline{Y}_h\right) \cong \left[\left\{\overline{Y}_h(1+e_2)\,(1-e_3+{e_3}^2)\right\} - \overline{Y}_h\right] \tag{19}$$

Squaring both sides of equation (19) and neglecting terms of e's having power greater than two and taking expectation using the result form equation (18), we have

$$\mathrm{E}\left(\mathrm{T}'_{\mathrm{tu_h}} - \overline{\mathrm{Y}}_{\mathrm{h}}\right)^2 = \left\{\mathrm{C}^2_{\mathrm{z_h}} - 2\rho_{\mathrm{yz_h}}\mathrm{C}_{\mathrm{y_h}}\mathrm{C}_{\mathrm{z_h}} + \mathrm{C}^2_{\mathrm{y_h}}\right\}\overline{\mathrm{Y}}_{\mathrm{h}}{}^2\mathrm{f}_{2\mathrm{h}}\frac{(\gamma_2\gamma_1)^2}{(\gamma_2-\gamma_1)^2}\mathrm{R} \tag{20}$$

Substituting the expression of $\mathrm{E}\left(\mathrm{T}'_{\mathrm{tu_h}} - \overline{\mathrm{Y}}_{\mathrm{h}}\right)^2$ into the equation (16), we get

$$\mathrm{V}(\mathrm{T_{tu}}) = \sum_{\mathrm{h}=1}^{\mathrm{K}} \mathrm{W}_{\mathrm{h}}^2\left\{\mathrm{C}^2_{\mathrm{z_h}} - 2\rho_{\mathrm{yz_h}}\mathrm{C}_{\mathrm{y_h}}\mathrm{C}_{\mathrm{z_h}} + \mathrm{C}^2_{\mathrm{y_h}}\right\}\overline{\mathrm{Y}}_{\mathrm{h}}{}^2\mathrm{f}_{2\mathrm{h}}\frac{(\gamma_2\gamma_1)^2}{(\gamma_2-\gamma_1)^2}\mathrm{R} \tag{21}$$

Similarly, expressing equation (12) in terms of e's, we have

$$\begin{aligned} T''_{tm_h} &= \overline{Y}_h(1+e_0) + \beta_{yx_h}\left[\overline{X}_h(1+e_4) - \overline{X}_h(1+e_1) + \beta_{xz_h}\left\{\overline{Z}_h(1+e_5) - \overline{Z}_h\right\}\right] \\ &= \overline{Y}_h(1+e_0) + \beta_{yx_h}\left[\overline{X}_h(e_4-e_1) + \beta_{xz_h}\overline{Z}_h e_5\right] \end{aligned}$$

It is found that

$$T''_{tm_h} - \overline{Y}_h = \left\{\overline{Y}_h(1+e_0) + \beta_{yx_h}\left[\overline{X}_h(e_4-e_1) + \beta_{xz_h}\overline{Z}_h e_5\right]\right\} - \overline{Y}_h \tag{22}$$

Squaring both sides of equation (22) and neglecting terms of e's having power greater than two and taking expectation using the result form equation (18), we have

$$E\left(T''_{tm_h} - \overline{Y}_h\right)^2 = \left\{(1-\rho^2_{yx_h})f_{1h} + \left(\rho_{yx_h} + 2\rho_{xz_h}\rho_{yz_h} + \rho_{yx_h}\rho^2_{xz_h}\right)\rho_{yx_h}f_{3h}\right\}S^2_{y_h}\frac{(\quad_2\ _1)^2}{(\quad_2-\gamma_1)^2}\mathrm{R} \tag{23}$$

Substituting the expression of $\mathrm{E}\left(\mathrm{T}''_{\mathrm{tm_h}} - \overline{\mathrm{Y}}_{\mathrm{h}}\right)^2$ into the equation (17), we get

$$V(T_{tm}) = \sum_{h=1}^{K}\left[\begin{matrix} W_h^2\left\{(1-\rho^2_{yx_h})f_{1h} + \left(\rho_{yx_h} + 2\rho_{xz_h}\rho_{yz_h} + \rho_{yx_h}\rho^2_{xz_h}\right)\rho_{yx_h}f_{3h}\right\} \\ S^2_{y_h}\frac{(\quad_2\ _1)^2}{(\quad_2-\gamma_1)^2}R \end{matrix}\right] \tag{24}$$

Substituting the expression of $\mathrm{V}(\mathrm{T_{tu}})$ and $\mathrm{V}(\mathrm{T_{tm}})$ in the equation (15), we get

$$\begin{aligned} V(T_{pe}) = \emptyset^2 &\left[\sum_{h=1}^{K} W_h^2\left\{C^2_{z_h} - 2\rho_{yz_h}C_{y_h}C_{z_h} + C^2_{y_h}\right\}\overline{Y}_h{}^2 f_{2h}\frac{(\gamma_2\gamma_1)^2}{(\gamma_2-\gamma_1)^2}R\right] \\ &+ (1 \\ &- \emptyset)^2\sum_{h=1}^{K}\left[\begin{matrix} W_h^2\left\{(1-\rho^2_{yx_h})f_{1h} + \left(\rho_{yx_h} + 2\rho_{xz_h}\rho_{yz_h} + \rho_{yx_h}\rho^2_{xz_h}\right)\rho_{yx_h}f_{3h}\right\} \\ S^2_{y_h}\frac{(\gamma_2\gamma_1)^2}{(\gamma_2-\gamma_1)^2}R \end{matrix}\right] \end{aligned} \tag{25}$$

6. Minimum variance the proposed estimator T_{pe}

We see that the variance of the estimator T_{pe} derived in equation (25) is a function of unknown constant ∅. So, to get the minimum (optimum) value of ∅, we differentiated equation (25) with respect to ∅ and equated the outcome to zero, which gives us the minimum (optimum) value of ∅ as

$$\emptyset_{opt} = \frac{V(T_{tm})}{V(T_{tu})+V(T_{tm})} \tag{26}$$

Substituting the optimum value of ∅ in the equation (15), we obtain the minimum (optimum) variance of the estimator T_{pe} as

$$V(T_{pe})_{opt} = \frac{V(T_{tu}).V(T_{tm})}{V(T_{tu})+V(T_{tm})} \tag{27}$$

Substitute the expression of $V(T_{tu})$ and $V(T_{tm})$ in the equation (27), we get the expression of minimum variance of the estimator T_{pe}.

7. Empirical study

To test the performance of the proposed estimator, we have compared our estimator with the conventional sample mean estimator of population mean $\overline{Y}$ based on sample of size n_h $(h = 1,2, \ldots, K)$given by

$$\text{i.e.}, T_1 = \sum_{h=1}^{K} W_h^2 \, \overline{y}_{n_h}$$

and its variance is obtained as

$$Var(T_1) = \sum_{h=1}^{K} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) W_h^2 \, \overline{Y}_h{}^2 C_{y_h}^2 \tag{28}$$

We have examined the merits of the proposed estimator against the conventional one through artificially generated population data as well as natural population data set. For this purpose, we have calculated the PRE of the proposed estimator with respect to T_1 as

$$PRE = \frac{Var(T_1)}{V(T_{pe})_{opt}} \times 100 \tag{29}$$

7.1. Numerical illustration using artificially generated population

The main perspective of simulation study is that it replicates the actual system. Simulation helps us to compare the efficiency through artificial population generation technique and conclude whether a newly developed method is superior than the existing ones. Inspired by the work of Singh and Deo (2003), Sing *et al.* (20017), Maji

et al. (2019) of artificially population generation techniques, we have generated three sets of independent random numbers (x, y, z) of size N (N=100) i.e., x[k], y[k] and z[k] (k = 1, 2, 3, …, N) from a standard normal distribution by using statistical software R.

For generating the population artificially, we use the following algorithm:

1. Generate random variables x1, y1, z1 and a (temporary variables) which are normally distributed with mean 0, S.D. =1 and are of size 100.
2. Define

   N =100.
3. Define

   x = 0, y = 0, z = 0, ry1x1 = 0.5 (correlation coefficient between y1 and x1), rx1z1=0.75 (correlation coefficient between x1 and z1), Sx1 = $\sqrt{50}$,

   Sy1 = $\sqrt{50}$, Sa = $\sqrt{40}$

   (Sx1, Sy1, Sa are S. D. (standard deviations) of x1, y1, z1 and variable a respectively)

   mx1= 20 (mean of x1), mz1 = 25 (mean of z1).
4. Define

   a1=Sy1*Sy*(1-(ry1x1^2))

   a2=Sz*Sz*(1-(rx1z1^2))
5. for (j in 1 to N)

   {

   y[j]=20.0+(sqrt(a1)*y1[j])+(ry1x1*sy1*x1[j])

   x[j]=25.0+(sx1*x1[j])

   z[j]=15+(sqrt(a2)*z1[j])+(rx1z1*sz1*x1[j])

   }
6. Take output of the variables x, y and z.

In such a way artificial population data set 1 has been generated.

Repeating the above algorithm, artificial population data set 2 has been generated with the following changed ry1x1 = 0.75 and rx1z1=0.5 in step 3.

Similarly, artificial population data set 3 has been generated by changing the value of ry1x1 = 0.75 and rx1z1=0.75 in step 3.

By the above algorithm we have generated artificial population of size 100 for each population data set and each data set further divided into five strata sequentially with sizes 15, 18, 21, 22 and 24 respectively. Details are given below in Table 1.

We have calculated the PREs of the proposed estimator T_{pe} for the different values of γ_1 and γ_2 from the Artificial Data Set-I, Artificial Data Set-II, Artificial Data Set-III.

For generating the table, we have taken ρ_{yx} = ry1x1 and ρ_{xz}= rx1z1.

Table 1: Parameters of the stratum 1-5

Size	Simulated Data				
	Stratum 1	**Stratum 2**	**Stratum 3**	**Stratum 4**	**Stratum 5**
N_h	15	18	21	22	24
n_h	8	7	9	10	10
m_h	5	4	5	6	6
u_h	3	3	4	4	4

Table 2: PRE of the proposed estimator with respect to T_1 for simulated data sets

Constant		PRE		
γ_1	γ_2	**Artificial Data Set-I** $(\rho_{yx} = 0.50, \rho_{xz} = 0.75)$	**Artificial Data Set-II** $(\rho_{yx} = 0.75, \rho_{xz} = 0.50)$	**Artificial Data Set-III** $(\rho_{yx} = 0.75, \rho_{xz} = 0.75)$
0.06	0.05	1136.458	1003.131	1464.382
	0.25	624.383	551.132	804.549
	0.50	557.616	492.198	718.516
	0.75	532.230	469.789	685.804
	1.00	515.532	455.051	664.288
0.10	0.05	923.523	815.176	1190.004
	0.25	411.458	363.186	530.184
	0.50	344.727	304.284	444.197
	0.75	319.428	281.953	411.599
	1.00	302.942	267.401	390.355
0.14	0.05	832.053	734.438	1072.141
	0.25	320.000	282.458	412.336
	0.50	253.309	223.591	326.401
	0.75	228.104	201.343	293.923
	1.00	211.831	186.980	272.955

7.2. Numerical Illustration using Natural Population data set

We have considered two natural population data sets to examine the merits of the proposed estimator T_{pe}. The sources of populations and details of the variables y, x and z and the values of various parameters are mentioned bellow.

7.2.1. Population Data Set-I [Literacy Rate by Sex of India, Census India (2011)]

y: Number of literates (male and female) in the year 2011;
x: Number of literates (male and female) in the year 2001;
z: Female literacy rate (2011).

We have divided 32 states of India into 6 different strata (zone wise) as shown in Table 3.

Table 3: Values of different parameters of the respective variables

Strata	States	Statistical Parameters
Stratum I	Andhra Pradesh	$N_h = 4, n_h = 3, m_h = 2, \overline{X}_h = 79.13,$ $\overline{Y}_h = 72.86, \overline{Z}_h = 73.18, S^2_{y_h} = 127.7,$ $S^2_{x_h} = 172.18, S^2_{z_h} = 194.02, \rho_{yx_h} =$ $0.9929, \rho_{yz_h} = 0.9980, \rho_{xz_h} = 0.9970$
	Karnataka	
	Kerala	
	Tamil Nadu	
Stratum II	Goa	$N_h = 5, n_h = 4, m_h = 2, \overline{X}_h = 78.18,$ $\overline{Y}_h = 71.62, \overline{Z}_h = 70.62, S^2_{y_h} = 69.82,$ $S^2_{x_h} = 67.83, S^2_{z_h} = 142.49, \rho_{yx_h} = 0.9864,$ $\rho_{yz_h} = 0.9864, \rho_{xz_h} = 0.9759$
	Gujarat	
	Maharashtra	
	Punjab	
	Rajasthan	
Stratum III	West Bengal	$N_h = 4, n_h = 3, m_h = 2, \overline{X}_h = 69.95,$ $\overline{Y}_h = 58.07, \overline{Z}_h = 60.35, S^2_{y_h} = 40.20,$ $S^2_{x_h} = 93.24, S^2_{z_h} = 73.06, \rho_{yx_h} = 0.9997,$ $\rho_{yz_h} = 0.9900, \rho_{xz_h} = 0.9930$
	Odisha	
	Jharkhand	
	Bihar	
Stratum IV	Manipur	$N_h = 7, n_h = 4, m_h = 2, \overline{X}_h = 76.73,$ $\overline{Y}_h = 65.61, \overline{Z}_h = 71.66, S^2_{y_h} = 49,$ $S^2_{x_h} = 34.95, S^2_{z_h} = 64.02, \rho_{yx_h} = 0.9266,$ $\rho_{yz_h} = 0.9702, \rho_{xz_h} = 0.8663$
	Meghalaya	
	Nagaland	
	Arunachal Pradesh	
	Assam	
	Sikkim	
	Tripura	
Stratum V	Uttar Pradesh	$N_h = 5, n_h = 4, m_h = 2, \overline{X}_h = 74.42,$ $\overline{Y}_h = 65.56, \overline{Z}_h = 65.08, S^2_{y_h} = 47.02,$ $S^2_{x_h} = 87.15, S^2_{z_h} = 69.85, \rho_{yx_h} = 0.9987,$ $\rho_{yz_h} = 0.9981, \rho_{xz_h} = 0.9941$
	Haryana	
	Himachal Pradesh	
	Uttarakhand	
	Jammu & Kashmir	
Stratum VI	A & N Island	$N_h = 7, n_h = 4, m_h = 2, \overline{X}_h = 85.67,$ $\overline{Y}_h = 78.37, \overline{Z}_h = 79.54, S^2_{y_h} = 21.76,$ $S^2_{x_h} = 89.89, S^2_{z_h} = 52.67, \rho_{yx_h} = 0.9532,$ $\rho_{yz_h} = 0.9822, \rho_{xz_h} = 0.9862$
	Chandigarh	
	Daman & Diu	
	D & N Haveli	
	Delhi	
	Lakshadweep	
	Pondicherry	

7.2.2. Population Data Set-II [Abortion Rate, Statistical Abstract of the United States (2011)]

y: Number of abortions reported in the year 2008.
x: Number of abortions reported in the year 2007.
z: Number of abortions reported in the year 2005.

We have divided 51 states of the United States into 4 different strata (zone wise) as shown in Table 4.

Table 4: Values of different parameters of the respective variables

Strata	States	Statistical Parameters
Stratum I	Wyoming	$N_h = 14, n_h = 8, m_h = 5, u_h = 3, \overline{X}_h = 6.551, \overline{Y}_h = 6.59, \overline{Z}_h = 6.720, S^2_{y_h} = 4.56, S^2_{x_h} = 4.51, S^2_{z_h} = 5.21, \rho_{yx_h} = 0.9784, \rho_{yz_h} = 0.9725, \rho_{xz_h} = 0.9484$
	Missouri	
	Mississippi	
	Kentucky	
	Oklahoma	
	Arkansas	
	Indiana	
	Nebraska	
	South Carolina	
	Wisconsin	
	Utah	
	South Dakota	
	Idaho	
	West Virginia	
Stratum II	Alaska	$N_h = 25, n_h = 12, m_h = 7, u_h = 5, \overline{X}_h = 15.031, \overline{Y}_h = 15.11, \overline{Z}_h = 14.851, S^2_{y_h} = 6.91, S^2_{x_h} = 6.93, S^2_{z_h} = 8.87, \rho_{yx_h} = 0.9413, \rho_{yz_h} = 0.878, \rho_{xz_h} = 0.8988$
	Montana	
	New Hampshire	
	Minnesota	
	Vermont	
	Ohio	
	Arizona	
	New Mexico	
	North Dakota	
	Maine	
	Michigan	
	Massachusetts	
	Washington	
	Kansas	
	Virginia	
	North Carolina	
	Oregon	
	Pennsylvania	
	Texas	
	Louisiana	
	Colorado	
	Tennessee	
	Iowa	
	Alabama	
	Georgia	

Table 4: Values of different parameters of the respective variables (cont.)

Strata	States	Statistical Parameters
Stratum III	Hawaii	$N_h = 7, n_h = 4, m_h = 2, u_h = 2, \overline{X}_h = 24.562, \overline{Y}_h = 24.48, \overline{Z}_h = 24.095, S^2_{y_h} = 31.42, S^2_{x_h} = 37.22, S^2_{z_h} = 65.29, \rho_{yx_h} = 0.9885, \rho_{yz_h} = 0.9442, \rho_{xz_h} = 0.95454$
	Rhode Island	
	Connecticut	
	Nevada	
	Florida	
	California	
	Illinois	
Stratum IV	Maryland	$N_h = 5, n_h = 4, m_h = 2, u_h = 2, \overline{X}_h = 33.528, \overline{Y}_h = 33.55, \overline{Z}_h = 36.533, S^2_{y_h} = 19.31, S^2_{x_h} = 29.03, S^2_{z_h} = 53.75, \rho_{yx_h} = 0.9751, \rho_{yz_h} = 0.39256, \rho_{xz_h} = 0.53965$
	District of Columbia	
	New Jersey	
	New York	
	Delaware	

We have calculated the PREs of the proposed estimator T_{pe} for the different values of γ_1 and γ_2 from the Population Data Set-I and Population Data Set-II.

Table 5. PRE of the proposed estimator with respect to T_1 for Natural Population data sets

Constant		PRE	
γ_1	γ_2	Population Data Set-I	Population Data Set-II
0.06	0.05	11392.639	4077.631
	0.25	6259.249	2240.298
	0.50	5589.930	2000.737
	0.75	5335.434	1909.648
	1.00	5168.048	1849.738
0.10	0.05	9258.025	3313.614
	0.25	4124.737	1476.318
	0.50	3455.776	1236.885
	0.75	3202.167	1146.114
	1.00	3036.894	1086.959
0.14	0.05	8341.070	2985.419
	0.25	3207.905	1148.167
	0.50	2539.341	908.876
	0.75	2286.668	818.440
	1.00	2123.543	760.054

8. Conclusion

We have examined the performances of our proposed strategy against the conventional ones for different data sets as presented. The results of such a comparison are discussed in Table 2 and Table 5. Thus, the following interpretations may be read out from the respective Tables.

a) It is noted that the development of the estimation procedure in successive sampling is still in progress under various designs but no one has incorporated the past sample information in the form of HEWMA statistics to construct an effective estimation strategy in the successive sampling scheme. The proposed estimator uses the current sample information along with the information available from past samples in the form of HEWMA statistics. The results in Tables 2 and 5 demonstrated that the proposed estimator is more efficient than the conventional estimator. It may be concluded that the suggested strategy is more efficient in estimation of population mean.

b) From Table 2, it is observed that for fixed values of ρ_{yx}, the values of PREs of our proposed estimator are increasing with the increasing values of ρ_{xz}. A similar pattern may be noted for fixed values of ρ_{xz} with increasing values of ρ_{yx}. Thus, it is clear that for the presence of higher values of correlation coefficient between study and auxiliary variable, our proposed strategy produces more precise estimates.

   From Table 5, it is also noted that the correlation coefficient between variables (y and x) and (x and z) is high in most of the strata and consequently the PREs are very high. Therefore, the findings of the simulation studies are also justified with the natural population studies.

c) It may also be noted that in artificially generated population, values of the various statistical parameter such as means, variances, correlation coefficients etc. are almost strata wise similar while in case of natural populations, their parametric values are different from strata to strata. Our proposed estimator performs profoundly for both the types of population which enhance their recommendation in practice.

Thus, it is observed from the interpretation of the result that the use of memory-based information of HEWMA statistics for the estimation of population mean under stratified successive sampling is highly encouraging. Moreover, the calibration technique utilized in the formulation of the estimator helps us in enhancing the efficiency of the proposed strategy. Therefore, looking at the pleasing findings, we are recommended our proposed strategy to the survey practitioners for their application in real life.

Acknowledgement

Authors are thankful to the reviewers for their constructive suggestions, which enhanced the quality of the manuscript.

References

Aslam, I., Noor-ul-Amin, M., Yasmeen, U. and Hanif, M., (2020). Memory type ratio and product estimators in stratified sampling. *Journal of Reliability and Statistical Studies*, 13(1), pp. 1–20.

Aslam, I., Noor-ul-Amin, M., Hanif, M. and Sharma, P., (2023). Memory type ratio and product estimators under ranked-based sampling schemes. *Communications in Statistics-Theory and Methods*, 1–23. *Communications in Statistics - Theory and Methods,* 52, pp. 1–23.

Bhushan, S.; Kumar, A., Al-Omari, A. I. and Alomani, G. A., (2023). Mean Estimation for Time-Based Surveys Using Memory-Type Logarithmic Estimators. *Mathematics*, 1(9), pp. 21–25. https://doi.org/10.3390/math11092125.

Biradar, R. S., Singh, H. P., (2001). Successive sampling using auxiliary information on both occasions. *Calcutta Statistical Association Bulletin,* 51, pp. 243–251.

Chaturvedi, D. K., Tripathi, T. P., (1983). Estimation of population ratio on two occasions using multivariate auxiliary information. *Journal of Indian Statistical Association*, 21, pp. 113–120.

Das, A. K., (1982). Estimation of population ratio on two occasions. *Journal of the Indian Society of Agricultural Statistics*, 34, pp. 1–9.

Feng, S., Zou, G., (1997). Sample rotation method with auxiliary variable. *Communications in Statistics-Theory and Methods*, 26, 6, pp. 1497–1509.

Haq, A., (2013). A new hybrid exponentially weighted moving average control chart for monitoring process mean. *Quality and Reliability Engineering International,* 29(7), pp. 1015–1025.

Haq, A., (2016). A new hybrid exponentially weighted moving average control chart for monitoring process mean Discussion. *Quality and Reliability Engineering International*, 33(7), pp. 629–1631.

Jessen, R. J., (1942). Statistical Investigation of a Sample Survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, No. 304, Ames, Iowa, USA, pp. 1–104.

Kiregyera, B., (1984). Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations. *Metrika,* 31, pp. 215–226.

Maji, R., Singh, G. N. and Bandyopadhyay, A., (2019). Estimation of Population Mean in Presence of Random Non-Response in Two-Stage Cluster Sampling. *Communications in Statistics - Theory and Methods,* 48 (14), pp. 3586–3608.

Noor-ul-Amin, M., (2020) Memory type ratio and product estimators for population mean for time-based surveys. *Journal of Statistical Computation and Simulation*, 90(17), pp. 3080–3092.

Noor-ul-Amin, M., (2021). Memory type estimators of population mean using exponentially weighted moving averages for time scaled surveys. *Communications in Statistics-Theory and Methods*, 50(12), pp. 2747–2758.

Patterson, H. D., (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, 12, pp. 241–255.

Rao, J. N. K., Graham, J. E., (1964). Rotation design for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, pp. 492–509.

Roberts, S., (1959). Control chart tests based on geometric moving averages. *Techno metrics*, 1(3), pp. 239-250.

Sen, A. R., (1971). Successive sampling with two auxiliary variables. *Sankhya*, 33, Series B, pp. 371–378.

Sen, A. R., (1972). Successive sampling with p (p ≥1) auxiliary variables. *Annals Mathematical Statistics*, 43, pp. 2031–2034.

Sen, A. R., (1973). Theory and application of sampling on repeated occasions with several auxiliary variables. *Biometrics*, 29, pp. 381–385.

Singh, G. N., Singh, V. K., (2001). On the use of auxiliary information in successive sampling. *Journal of the Indian Society of Agricultural Statistics*, 54(1), pp. 1–12.

Singh, G. N., (2003). Estimation of population mean using auxiliary information on recent occasion in h-occasion successive sampling. *Statistics in Transition*, pp. 523-532.

Singh, S., Deo, B., (2003). Imputation by power transformation. *Statistical Papers*, 4, pp. 555–579.

Singh, G. N., Karna, J. P., (2009). Estimation of population mean on current occasion in two-occasion successive sampling. *Metron - International Journal of Statistics*, vol. LXVII, no. 1, pp. 87–103.

Singh, H. P., Vishwakarma, G. K., (2009). A general procedure for estimating population mean in successive sampling. *Communications in Statistics-Theory and Methods,* 38(2), pp. 293–308.

Singh, H. P., Tailor, R., Singh, S. and Kim, J. M., (2011). Estimation of population variance in successive sampling. *Quality and Quantity*, 45, pp. 477–494.

Singh, G. N., Sharma, A. K. and Bandyopadhyay, A., (2017). Effectual Variance Estimation Strategy in Two Occasions Successive Sampling in Presence of Random Non-Response. *Communications in Statistics-Theory and Methods*, 46(14), pp. 7201–7224.

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 57–84, https://doi.org/10.59139/stattrans-2025-016
Received – 05.03.2024; accepted – 20.02.2025

Inference of dynamic weighted cumulative residual entropy for Burr XII distribution based on progressive censoring

Amal S. Hassan[1], **E. A. Elsherpieny**[2], **Wesal E. Aghel**[3]

Abstract

The dynamic weighted cumulative residual (DWCR) entropy is regarded as an additional measure of uncertainty related to the residual lifetime function in several disciplines, including survival analysis and reliability. This article presents the DWCR formula based on Havarda and Charvat. This measurement is called the DWCR Havarda and Charvat entropy (DWCRHCE). This work uses progressive Type II censoring to investigate the implications of DWCR Tsallis entropy (DWCRTE), DWCR Rényi entropy (DWCRRE), and DWCRHCE for the Burr XII distribution. Both classical and Bayesian methods are used to derive the estimators of these entropy metrics. Assuming independent gamma priors, we get the Bayes estimator of the suggested measures. Due to the lack of explicit forms, the Metropolis-Hastings approach was offered to determine the Bayes estimates for symmetric and asymmetric loss functions. To determine the efficacy of the suggested estimating techniques, several simulations were run for different censoring schemes. The simulation analysis leads us to the conclusion that, under a precautionary loss function followed by a linear exponential loss function, the Bayesian estimates of DWCRTE are generally more effective than the DWCRHCE or DWCRRE. Compared to maximum likelihood estimates, Bayesian estimates are preferred for different metrics. After that, a detailed explanation of the process is provided by looking at real data. The analysis of real-world data, specifically the Shasta reservoir water capacity data, aligns with the findings from simulated data. Notably, these findings have crucial implications for effective water resource management decisions.

Key words: Burr XII distribution, dynamic weighted cumulative residual entropy, Bayesian estimators, precautionary loss function.
Mathematical Subject Classification: 62F10.

[1] Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt.
E-mail: amal52_soliman@cu.edu.eg. ORCID: https://orcid.org/0000-0003-4442-8458.
[2] Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt.
E-mail: Elsayed.Elsherpieny@cu.edu.eg. ORCID: https://orcid.org/0000-0002-2635-8081.
[3] Corresponding Author. Faculty Sciences, Zawia University, Libya. E-mail: Wesalagil@yahoo.com, 12422022546163@pg.cu.edu.eg. ORCID: https://orcid.org/0009-0000-7246-7660.

1. Introduction

1.1. Progressive Type II censoring

Censorship is common in many fields, including pharmacology, social economics, and engineering, particularly in reliability and survival analysis (Wang & Gui, 2021). Due to time and cost constraints, it is difficult to completely observe the sample data in actual production. Even though they have been thoroughly examined, conventional censoring systems such as Type-I, Type-II, and hybrid schemes are inflexible, in that units cannot be removed arbitrarily. Cohen (1963) proposed the progressive censoring scheme (PCS) in order to overcome this restriction. In the PCS, units are removed from the experiment at different time points, with the number of units eliminated at each time point specified in advance.

The progressive type-II censoring (PT-IIC) method is one of the most popular censoring schemes. Here, we provide its description. Assume that m failures will be noticed when n identical units are put through a test. At the moment of the initial failure ($y_{(1)}$), the number r_1 of the surviving units ($n-1$) is randomly selected and removed from the experiment. At the second failure ($y_{(2)}$), the number r_2 of the surviving units ($n-r_1-2$) is randomly selected and removed from the experiment, and so on, until the m^{th} failure ($y_{(m)}$) occurs, at which point all the remaining $n-m-r_1-r_2-\cdots-r_{m-1}$ units are removed. Thus, the PCS includes m observed samples of failure $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(m)}$, and survival items $\boldsymbol{r} = (r_1, r_2, \cdots, r_m)$ such that $n = m + r_1 + \cdots + r_m$. Note that in the PCS, $\boldsymbol{r} = (r_1, r_2, \cdots, r_m)$ is prefixed. Balakrishnan and Aggrawala (2000) offered historical context and a comprehensive overview of PCS. Notably, the following special cases can be noticed:

- Classical Type-II censoring**:** This occurs when $r_1 = r_2 = \cdots = r_{m-1} = 0$ and $r_m = n - m$.
- Complete sample**:** This occurs when $m = n$ and $r_i = 0$, $i = 1, 2, \cdots, n$.

1.2. Entropy measures

Shannon (1948) proposed the concept of entropy as a metric for quantifying uncertainty. Nowadays, the fields of economics, physics, telecommunications, communication theory, and reliability have given this criterion significant consideration. One parameter generalization of the Shannon entropy that may be applied as a randomness metric is the Rényi entropy. Numerous disciplines, including biology, genetics, electrical engineering, computer science, economics, chemistry and physics use Rényi entropy in their work. The entropy function as an extension of the Shannon entropy was first presented by Rényi (1961), followed by Havrda and Charvat (1967). The Rényi and Havrda and Charvat entropy of order c are defined respectively by the following expressions:

$$\boldsymbol{R(c) = \frac{1}{1-c} log\left(\int_{-\infty}^{\infty} (f(y))^c \, dy\right); c > 0, c \neq 1,}$$

$$\boldsymbol{H(c) = \left(2^{1-c} - 1\right)^{-1}\left(\int_{-\infty}^{\infty} (f(y))^c - 1\right); \quad c > 0, c \neq 1.}$$

Tsallis (1988) also presented the ideas of Tsallis entropy as a measure of quantifying randomness. The entropy can be used to determine the level of uncertainty associated with a random observation. Tsallis later utilized its distinctive characteristics and situated it within a physical framework. This metric is designated for any continuous random variable of order c, where and it is defined as follows:

$$\boldsymbol{T(c) = \frac{1}{(c-1)}\left(1 - \int_{-\infty}^{\infty} f^{c}(y)dy\right); \quad c > 0, c \neq 1.}$$

Recently, entropy measure estimation utilizing different statistical distributions and sampling strategies has been studied by several writers (see, for example, Baratpour et al., 2007; Abo-Eleneen, 2011; Cho et al., 2015; Lee, 2017; Hassan and Zaky, 2019, 2021; Helmy et al., 2021; Hassan et al., 2022; Helmy et al., 2023; and Hassan at al., 2024b).

Different measures of uncertainty for probability distributions have attracted many writers, especially in works related to reliability analysis and survival. In light of Rényi's entropy's utility, Sunoj and Linu (2012) presented cumulative residual Rényi entropy (CRRE) of order c as below:

$$\boldsymbol{\bar{R}(c) = \frac{1}{(1-c)} log\left(\int_{t}^{\infty} (\bar{F}(y))^{c} dy\right); c > 0, c \neq 1,}$$

where $\bar{F}(y) = 1 - F(y)$ is the survival function (SF). Additionally, Sunoj and Linu (2012) examined the primary characteristics of the dynamic version of the CRRE (DCRRE), which is extended to the residual lifetime $Y_t = (Y - t \mid Y > t)$ based on survival function, rather than using probability density function (PDF), and found it to be beneficial for reliability modeling. The DCRRE of order c is given by:

$$\boldsymbol{\bar{R}^{*}(c) = (1-c)^{-1} log\left(\frac{1}{(\bar{F}(t))^{c}} \int_{t}^{\infty} (\bar{F}(y))^{c} dy\right); \quad c > 0, c \neq 1.} \tag{1}$$

Sati and Gupta (2015) established the DCR Tsallis entropy (DCRTE) as follows:

$$\boldsymbol{\bar{T}^{*}(c) = (c-1)^{-1}\left(1 - (\bar{F}(t))^{-c} \int_{t}^{\infty} (\bar{F}(y))^{c} dy\right); \quad c > 0, c \neq 1,}$$

In the literature, some statistical inferences based on the above dynamic entropy measures and their related works have been considered by several authors. For the Pareto distribution, Bayesian estimates (BEs) of the DCR entropy under various sampling conditions have been examined by several researchers (see Renjini et al., 2016a, 2016b, and 2018; and Ahmadini et al., 2020). The Lindley distribution's BE of DCRRE was examined by Almarashi et al. (2021). Al-Babtain et al. (2021) supplied the maximum likelihood estimates (MLEs) and BEs of the DCRRE for the Lomax distribution. Mohamed (2022) used generalized order statistics to study DCRTE and cumulative residual Tsallis entropy. The BEs of the DCRTE for the moment exponential distribution were recently studied by Alyami et al. (2023). For more recent studies, refer to Kayal and Balakrishnan (2023), Nair and Sathar (2024) and Smitha et al. (2024).

Weighted distributions provide a valuable tool for modeling statistical data in situations where standard distributions may not accurately capture the underlying characteristics of the data. Guiasu (1986) applied weighted entropy in order to balance the degree of homogeneity and information contained in a data partition into classes. The idea of weighted distributions was used in a number of domains, such as biostatistics (Wang, 1996), reliability modeling (Navarro et al., 2001) and renewal theory (Sunoj and Mayi, 2006). In the context of theoretical neurobiology, uncertainty measures based on the concept of weighted entropy were explored by Belis and Guiasu (1968). Di Crescenzo and Longobardi (2006) extended the concept of weighted entropy to residual and past lifetimes, introducing weighted residual and past entropies. This work builds upon previous research by Belzunce et al. (2004) and Nanda and Paul (2006), which characterized distribution functions using weighted dynamic measures. Misagh and Yari (2011) further investigated this concept by studying the weighted differential information measure for two-sided truncated random variables. Sunoj and Linu (2012) introduced a new measure of uncertainty based on the length-biased weighted function and called it dynamic weighted CRRE (DWCRRE). Based on the SF given in Equation (1), the DWCRRE is as follows:

$$\boldsymbol{R^{*}(c) = (1-c)^{-1} log\left(\left(\overline{F}(t)\right)^{-c}\int_{t}^{\infty} y\left(\overline{F}(y)\right)^{c} dy\right);\quad c>0, c\neq 1,} \tag{2}$$

where y is the length-biased weighted function. The dynamic weighted CRTE (DWCRTE), introduced by Khammar and Jahanshahi (2018), is another significant weighted metric. It is defined as:

$$\boldsymbol{T^{*}(c) = (c-1)^{-1}\left(1-\left(\overline{F}(t)\right)^{-c}\int_{t}^{\infty} y\left(\overline{F}(y)\right)^{c} dy\right);\quad c>0, c\neq 1.} \tag{3}$$

Using Type II right-censored data, a weighted version of CRTE and DCRTE was created by Khammar and Jahanshahi (2018), and many of its reliability properties.

1.3. Work Motivation

As far as we are aware, no research so far has taken into account the PT-IIC for dynamic weighted cumulative residual in entropy measurements. Therefore, our research question is: "How to find the estimate of dynamic weighted cumulative residual entropy measures under PT-IIC for the Burr XII distribution (BXIID)?". We decided to investigate this topic because of the significance of the BXIID and its widespread application in numerous sectors (as shown in Section 2). Our work involves the following steps:

- The DWCRE based on the Havrda and Charvat measure is defined following the idea of Sunoj and Linu (2012). This measure is called the dynamic weighted cumulative residual Havrda and Charvat entropy (DWCRHCE).
- The estimators for the DWCRRE, DWCRTE, and DWCRHCE are derived using both Bayesian and non-Bayesian techniques. Both symmetric and asymmetric loss

functions yield Bayesian estimators for the suggested measures. Also, because there are no explicit forms for the BEs of various measures, we use the Markov chain Monte Carlo (MCMC) approach to approximate the estimates.

- Simulation studies are employed to evaluate and contrast the precision of various approximations about their mean squared error (MSE) and average of estimates. For illustration reasons, application to real data is shown.

The rest of the paper is organized in the following way: Section 2 presents a model description. Dynamic weighted cumulative residual entropy expressions are determined in Section 3. The PT-IIC and maximum likelihood estimation are used in Section 4. The BE of the DWCRE for the BXIID under symmetric and asymmetric loss functions is presented in Section 5. Section 6 provides an example of a real-data application. A simulation study is provided in Section 7. Based on the outcomes of the numerical studies, the paper draws a few conclusions in the last section.

2. Model Description

The Burr distribution is a versatile family that covers several widely used distributions as limiting asymptotic approximations, and it comprises a broad range of distribution shapes. A significant amount of the curve shape properties in the Pearson family are covered by correctly selecting the parameters of the Burr distribution, as Burr (1942) showed. Because its shape parameter generates a variety of forms that are excellent fits for different data, the BXIID has been used in research related to medical, business, chemical engineering, quality control and reliability. Evans and Simons (1975), Wingo (1993) and Gupta et al. (1996) are a few references to consult for a detailed explanation of such circumstances. The PDF and SF of the BXIID have the following specifications:

$$f(y) = \delta\lambda y^{\delta-1}\left(1 + y\ \right)^{-(\lambda+1)}, y > 0, \tag{4}$$

$$\overline{\boldsymbol{F}}(\boldsymbol{y}) = \left(\mathbf{1} + \boldsymbol{y}^{\boldsymbol{\delta}}\right)^{-\boldsymbol{\lambda}}, \boldsymbol{y} > \mathbf{0}, \tag{5}$$

where$\delta >$ and$\lambda > 0$ are shape parameters. Recently, numerous scholars have conducted extensive research on the estimation utilizing the BXIID. Estimation in step-stress partially accelerated life tests for the BXIID using Type I censoring was covered by Abd-Elfattah et al. (2008). Works on BXIID inferences under the PCS were discussed by, for example, Mousa and Jaheen (2002), Wu and Yu (2005), Soliman (2005), Li et al. (2007) and Hassan et al. (2024a). According to Panahi and Sayyareh (2014), statistical inference and prediction about BXIID parameters based on a Type II censored sample were covered. On the basis of the competing risk model, Qin and Gui (2020) derived the MLE and BE of BXIID parameters.

3. Dynamic Weighted Cumulative Residual Entropy Expressions

Inspired by the DWCRRE entropy developed by Sunoj and Linu (2012) and the DWCRTE presented by Khammar and Jahanshahi (2018), we introduce two novel information measures: the DCR Havrda and Charvat entropy (DCRHCE) and the second measure is the DWCRHCE. This measure is based on the cumulative residual entropy originally proposed by Rao et al. (2004), and the dynamic cumulative residual entropy proposed by Asadi and Zohrevand (2007) and Sati and Gupta (2015).

Definition: The DCRHCE and DWCRHCE of a random variable Y of order c are defined by:

$$\boldsymbol{H(c) = \left(2^{1-c}-1\right)^{-1}\left((\overline{F}(t))^{-c}\int_t^{\infty}(\overline{F}(y))^{c}-1\right); c>0, c\neq 1,}$$
$$\boldsymbol{H^{*}(c) = \left(2^{1-c}-1\right)^{-1}\left((\overline{F}(t))^{-c}\int_t^{\infty} y(\overline{F}(y))^{c}-1\right); c>0, c\neq 1,} \quad (6)$$

where y, is the length-biased weighted function. Now, the expressions of DWCRRE, DWCRTE, and DWCRHCE for the BXIID are obtained. To do so, we first obtain the following integral by using the SF given in Equation (5)

$$I=\int_t^{\infty} y(\bar{F}(y))^{c}dy=\int_t^{\infty} y\left(1+y\ \right)^{-\lambda c}dy. \quad (7)$$

Using the transformation $z = y\ , y = (z)^{\frac{1}{\delta}}, dy = \delta^{-1}z^{\frac{1}{\delta}-1}dz$in Equation (7) yields:

$$I=\int_{t^{\delta}}^{\infty} z^{\frac{1}{\delta}}(1+z)^{-\lambda c}\frac{1}{\delta}z^{\frac{1}{\delta}-1}dz=\frac{1}{\delta}\int_{t^{\delta}}^{\infty}(1+z)^{-\lambda c}z^{\frac{2}{\delta}-1}dz.$$

Use the transformation $x=(1+z)^{-1}$, then $z=\frac{(1-x)}{x}$, and$dz=\frac{-dx}{x^2}$, thus the integral I is as follows:

$$I=\tfrac{1}{\delta}\int_0^{\left(1+t^{\delta}\right)^{-1}} x^{\lambda c}\left(\frac{1-x}{x}\right)^{\frac{2}{\delta}-1}\frac{dx}{x^2}=\tfrac{1}{\delta}\int_0^{\left(1+t^{\delta}\right)^{-1}} x^{\lambda c-\frac{2}{\delta}-1}(1-x)^{\frac{2}{\delta}-1}dx=\tfrac{1}{\delta}B\left(\tfrac{2}{\delta},\lambda c-\tfrac{2}{\delta},\left(1+t\ \right)^{-1}\right), \quad (8)$$

where $B(a,b,x)=\int_0^x y^{a-1}(1-y)^{b-1}\,dy$ is the incomplete beta function. Now the formula of the DWCRRE is obtained by inserting Equation (8) in Equation (2), in the following way:

$$R^{*}(c)=\frac{1}{(1-c)}log\left(\frac{1}{\left(1+t^{\delta}\right)^{-\lambda c}}B\left(\tfrac{2}{\delta},\lambda c-\tfrac{2}{\delta},\left(1+t\ \right)^{-1}\right)\right). \quad (9)$$

Also, expressions of DWCRTE and DWCRHCE measures are obtained by inserting Equation (8) in Equations (3) and (6), respectively, as below:

$$T^{*}(c)=\frac{1}{(c-1)}\left[1-\frac{1}{\left(1+t^{\delta}\right)^{-\lambda c}}\left(B\left(\tfrac{2}{\delta},\lambda c-\tfrac{2}{\delta},\left(1+t\ \right)^{-1}\right)\right)\right], \quad (10)$$

$$H^{*}(c)=\frac{1}{(2^{1-c}-1)}\left[\frac{1}{\left(1+t^{\delta}\right)^{-\lambda c}}\left(B\left(\tfrac{2}{\delta},\lambda c-\tfrac{2}{\delta},\left(1+t\ \right)^{-1}\right)\right)-1\right]. \quad (11)$$

Note that expressions (9), (10) and (11) are a function of parameters δ, λ, and c.

4. Maximum Likelihood Estimators

Here, the MLEs of DWCRRE, DWCRTE, and DWCRHCE of the BXIID, based on PT-IIC samples, are obtained. Assume that $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(m)}$be the PT-IIC of

size m from a sample of size n taken from density given in Equation (4) and SF given in Equation (5) with censoring scheme $r_1, r_2, \ldots, r_m$. Based on the PT-IIC sample, the likelihood function reads:

$$L(\delta,\lambda) = D\prod_{i=1}^{m} f(y_{(i)})\left[1-F(y_{(i)})\right]^{r_i} == D\delta^m\lambda^m\prod_{i=1}^{m} y_{(i)}^{\delta-1}\left(1+y_{(i)}\right)^{-\lambda(r_i+1)-1},$$

where $\boldsymbol{D = n(n-r_1-1)(n-r_1-r_2-2)\ldots n-m+1-\sum_{i=1}^{m-1} r_i}$. Additionally, the corresponding log-likelihood function, say ℓ^*, is

$$\boldsymbol{\ell^* \propto m\, ln(\delta) + m\, ln(\lambda) + (\delta-1)\sum_{i=1}^{m} ln(y_{(i)}) - \sum_{i=1}^{m}[\lambda(r_i+1)+1]\, ln\left(1+y_{(i)}^{\delta}\right)}.$$

Then, the first derivative of the log- likelihood function, with respect to the parameters δ and λ, are as follows:

$$\frac{\partial \ell^*}{\partial \delta} = \frac{m}{} + \sum_{i=1}^{m} lny_{(i)} \quad - \sum_{i=1}^{m}\frac{[\lambda(r_i+1)+1]\, lny_{(i)}}{\left(1+y_{(i)}^{-\delta}\right)}, \tag{12}$$

and

$$\frac{\partial \ell^*}{\partial \lambda} = \frac{m}{\lambda} - \sum_{i=1}^{m}(r_i+1)\, ln\left(1+y_{(i)}^{\delta}\right). \tag{13}$$

The MLEs of δ and $\lambda$are determined by solving Equations (12) and (13) after simultaneously setting them to zero by using a numerical technique such as the Newton-Raphson method, to get $\hat{\delta}_{ML}$and $\hat{\lambda}_{ML}$. Based on the invariance property, the MLEs of DWCRRE, DWCRTE, and DWCRHCE are obtained by inserting$\hat{\delta}_{ML}$and $\hat{\lambda}_{ML}$ in Equations (9), (10), and (11).

5. Bayesian Estimator

This part covers the BE of $T^*(c)$, $R^*(c)$ and $H^*(c)$ under both symmetric and asymmetric loss functions for the two-parameter BXIID. It is assumed that the prior of parameters δ and λ have independent gamma priors. The joint prior distribution can be written as:

$$\boldsymbol{\pi_0(\delta,\lambda) \propto \delta^{b_1-1}\lambda^{b_2-1}e^{-(\delta a_1+\lambda a_2)}},$$

where the hyper-parameters a_1, a_2, b_1, and b_2 are known and non-negative. The gamma prior was chosen for its flexibility in modeling a wide range of prior beliefs. They may take on a wide range of forms based on hyper-parameters. The joint posterior for parameters, denoted by $\pi(\delta,\lambda)$, is

$$\boldsymbol{\pi(\delta,\lambda) = W^{-1}\delta^{m+b_1-1}\lambda^{m+b_2-1}e^{-(\delta a_1+\lambda a_2)}\prod_{i=1}^{m} y_{(i)}^{\delta-1}\left(1+y_{(i)}^{\delta}\right)^{-[\lambda(r_i+1)+1]}},$$

where $\boldsymbol{W = \int_0^\infty\int_0^\infty \delta^{m+b_1-1}\lambda^{m+b_2-1}e^{-(\delta a_1+\lambda a_2)}\prod_{i=1}^{m} y_{(i)}^{\delta-1}\left(1+y_{(i)}^{\delta}\right)^{-[\lambda(r_i+1)+1]} d\delta d\lambda}$.

So, the conditional posterior distribution of the unknown parameters δ and λ, is given, respectively, by:

$$\boldsymbol{\pi^*\left(\delta\left|\lambda,\underline{y}\right.\right) \propto \delta^{m+b_1-1}e^{-\delta\left(a_1+\sum_{i=1}^{m} lny_{(i)}\right)-\sum_{i=1}^{m}[\lambda(r_i+1)+1]\, ln\left(1+y_{(i)}^{\delta}\right)}}, \tag{14}$$

and

$$\boldsymbol{\pi^{**}\left(\lambda\left|\delta,\underline{y}\right.\right)} \propto \text{Gamma}\ (m+b_2, a_2+\textstyle\sum_{i=1}^{m}(r_i+1)\, ln(1+y_{(i)})). \tag{15}$$

The Bayesian estimators of $g^{\bullet}(c)$, under squared error loss function (SLF), linear exponential loss function (LLF), and precautionary loss function (PRLF) are obtained in the following way:

$$\widehat{\boldsymbol{g}}^{\bullet}_{\boldsymbol{SLF}}(\boldsymbol{c}) = \boldsymbol{E}\left[\boldsymbol{g}^{\bullet}(\boldsymbol{c})\left|\underline{\boldsymbol{y}}\right.\right] = \boldsymbol{W}^{-1}\int_0^{\infty}\int_0^{\infty} \boldsymbol{g}^{\bullet}(\boldsymbol{c})\boldsymbol{\delta}^{\boldsymbol{m}+\boldsymbol{b}_1-1}\boldsymbol{\lambda}^{\boldsymbol{m}+\boldsymbol{b}_2-1}\boldsymbol{e}^{-(\delta\boldsymbol{a}_1+\lambda\boldsymbol{a}_2)}\prod_{i=1}^{m}\boldsymbol{y}_{(i)}^{\delta-1}\left(\mathbf{1}+\boldsymbol{y}_{(i)}^{\delta}\right)^{-[\lambda(r_i+1)+1]}\boldsymbol{d\delta d\lambda}, \quad (16)$$

$$\widehat{\boldsymbol{g}}^{\bullet}_{\boldsymbol{LLF}}(\boldsymbol{c}) = \frac{-1}{\tau}\boldsymbol{ln}\left[\boldsymbol{E}\left(\boldsymbol{e}^{-\tau\boldsymbol{g}^{\bullet}(\boldsymbol{c})}\left|\underline{\boldsymbol{y}}\right.\right)\right] = \boldsymbol{W}^{-1}\int_0^{\infty}\int_0^{\infty}\boldsymbol{\delta}^{\boldsymbol{m}+\boldsymbol{b}_1-1}\boldsymbol{\lambda}^{\boldsymbol{m}+\boldsymbol{b}_2-1}\boldsymbol{e}^{-(\tau\boldsymbol{g}^{\bullet}(\boldsymbol{c})+\delta\boldsymbol{a}_1+\lambda\boldsymbol{a}_2)}\prod_{i=1}^{m}\boldsymbol{y}_{(i)}^{\delta-1}\left(\mathbf{1}+\boldsymbol{y}_{(i)}^{\delta}\right)^{-[\lambda(r_i+1)+1]}\boldsymbol{d\delta d\lambda}, \quad (17)$$

and

$$\widehat{\boldsymbol{g}}^{\bullet}_{\boldsymbol{PRLF}}(\boldsymbol{c}) = \left[\boldsymbol{E}\left((\boldsymbol{g}^{\bullet}(\boldsymbol{c}))^2\left|\underline{\boldsymbol{y}}\right.\right)\right]^{0.5} = \boldsymbol{W}^{-1}\int_0^{\infty}\int_0^{\infty}(\boldsymbol{g}^{\bullet}(\boldsymbol{c}))^2\boldsymbol{\delta}^{\boldsymbol{m}+\boldsymbol{b}_1-1}\boldsymbol{\lambda}^{\boldsymbol{m}+\boldsymbol{b}_2-1}\boldsymbol{e}^{-(\delta\boldsymbol{a}_1+\lambda\boldsymbol{a}_2)}\prod_{i=1}^{m}\boldsymbol{y}_{(i)}^{\delta-1}\left(\mathbf{1}+\boldsymbol{y}_{(i)}^{\delta}\right)^{-[\lambda(r_i+1)+1]}\boldsymbol{d\delta d\lambda}, \quad (18)$$

where $g^{\bullet}(c) = R^{*}(c)$ to obtain the DWCRRE, $g^{\bullet}(c) = T^{*}(c)$ to calculate the DWCRTE, and $g^{\bullet}(c) = H^{*}(c)$ to produce the DWCRHCE. The Gibbes sampler, Metropolis-Hastings (M-H), and random walk Metropolis algorithms are used to generate the MCMC samples from the posterior density functions (14) and (15), respectively, because integrals (16)–(18) do not take a closed form. As a result, the BEs of δ and λ under the SLF, LLF, and PRLF are calculated from their posteriors as the mean of the simulated samples. The M-H algorithm is one of the most famous subclasses of the MCMC method in Bayesian literature. It is used to simulate the deviates from the posterior density and produce good approximate results. The M-H algorithm uses an acceptance/rejection rule to converge to the target distribution. The initial values of the unknown parameters (δ, λ) must be specified, along with a suggested distribution, in order to implement the M-H algorithm for the DWCRTE, DWCRRE, and the DWCRHCE of the BXIID. A normal distribution will be used to calculate the proposal distribution, i.e.,$h(\delta'|\delta)$ and $h(\lambda'|\lambda) \equiv N(\theta, \sigma_{\theta}^2)$, where$\theta \equiv (\delta, \lambda)$ and σ_{θ}^2 is the variance-covariance matrix (Va-CoM) for the MLEs of (δ, λ). The MLE for θ is taken into account, that is,$\theta^{(0)} = \hat{\theta}_{MLE}$. Asymptotic Va-CoM, say $I^{-1}(\hat{\theta}_{MLE})$, where $I(.)$is the Fisher information matrix, and is assumed to represent the choice of σ_{θ}^2. First, the M-H algorithm employs the steps mentioned below to extract a sample from the posterior density given by Equations (14) and (15).

Step 1: Set the initial value ofθ as $\theta^{(0)} = \hat{\theta}_{MLE}$.

Step 2: For i =1, 2, 3,…, M repeat the following steps:

2.1: Set $\theta = \theta^{(i-1)}$.

2.2: Generateλ_1^i from Gamma$(\boldsymbol{m} + \boldsymbol{b_2}, \boldsymbol{a_2} + \sum_{i=1}^{m}\boldsymbol{\lambda}(\boldsymbol{r_i}+\mathbf{1})\,\boldsymbol{ln}(\mathbf{1}+\boldsymbol{y}_{(i)}^{\delta}))$.

2.3: Create a new candidate parameter value using $N(\theta, S_{\theta})$.

2.4: Compute the formula $\phi = \frac{\pi(\theta'|y)}{\pi(\theta|y),}$ where $\pi(.)$is the posterior density of Equations (14) and (15).

2.5: Create a sample u from the uniform U (0,1) distribution.

2.6: Accept or reject the new candidate θ'. $\begin{cases} If u \le \phi put\theta^{(i)} = \theta' \\ elsewhereput\theta^{(i)} = \theta. \end{cases}$

Step 3: Obtain the Bayesian estimator of θ and compute the DWCRTE, DWCRRE, and DWCRHCE functions of $T^*(c)$, $R^*(c)$ and $H^*(c)$ with respect to the loss functions as follows:

$$\hat{H}^*(c) = \frac{1}{M-Q}\sum_{i=Q+1}^{M} H^*\left(c,\theta^{(i)}\right), \hat{R}^*(c) = \frac{1}{M-Q}\sum_{i=Q+1}^{M} R^*\left(c,\theta^{(i)}\right), \hat{T}^*(c) = \frac{1}{M-Q}\sum_{i=Q+1}^{M} T^*\left(c,\theta^{(i)}\right),$$

where Q is the number of samples that have been burned. Ultimately, the estimates of DWCRTE, DWCRRE, and DWCRHCE are obtained by subtracting 2,000 burn-in samples from the 10,000 samples that the posterior density produced.

6. Real Data Analysis

The data set represents the monthly water capacity data from the Shasta reservoir in California, USA, and was taken for the month of February from 1991 to 2010 (http://cdec.water.ca.gov/reservoir_map.html). The maximum capacity of the reservoir is 4552000 AF, and the data set was transformed to the interval [0, 1] (for more details, see Nadar et al., 2013). In the first step, it should be checked if the BXIID is well fitted to this data. By fitting BXIID, the Kolmogorov-Smirnov distance and associated p-value are 7.7852, 7.8451, 0.22479 and 0.2274, respectively. The p-values show that BXIID yields suitable fits for the given dataset. In Figure 1, the empirical distribution functions and histogram are provided.

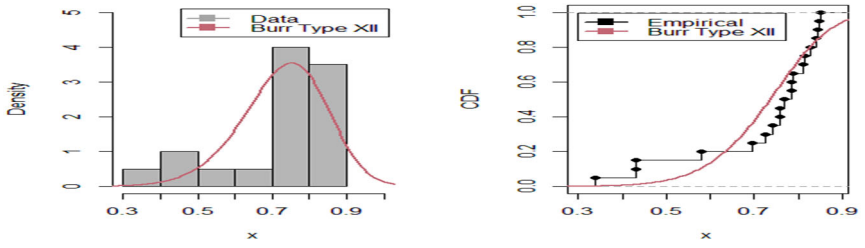


Figure 1. The histogram (left) and the empirical distribution function (right) for a given dataset.

Source: created by researchers utilizing the R programming language.

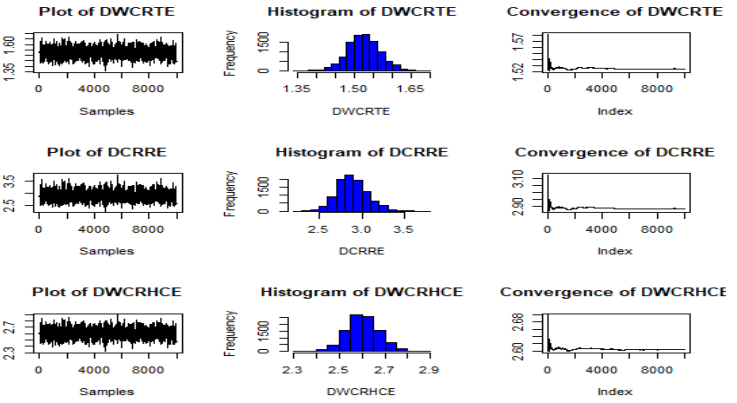
The number of stages in the PT-IIC scheme is assumed to be m = 12 and the removed items r_i are assumed as in the following scheme (Sch.):

$$S_1: (8,0^{*11}), \quad S_2: (4,0^{*10}, 4), \; S_3: (0^{*2}, 1^{*8}, 0^{*2}), \text{ and } S_4: (0^{*11}, 8).$$

The complete case also considers instances where $n = m = 2$ and $r_i = 0$, $i = 1, 2, \ldots, n$. Table 1 uses the PT-IIC to compute MLEs based on the produced data for each censoring strategy. Then employed to compute $T^*(c)$, $R^*(c)$ and $H^*(c)$ given t and c, where t = 0.1, 0.2, and c = 1.5, 2.5. For the BEs, the M-H algorithm will be used under

different loss functions in the case of a uniform prior, where $a_1 = a_2 = b_1 = b_2 = 0.001$. Different loss functions, including SLF, LLF-1($\tau = 0.5$), LLF-2($\tau = -0.5$) and PRLF are assumed. After that, the estimated values are calculated using the previous values. The estimates of DWCRTE, DWCRRE, and DWCRHCE are then obtained after 2,000 burn-in samples are subtracted from the 10,000 samples that the posterior density produced.

Convergence of the MCMC estimates for the DWCRTE, DWCRRE, and the DWCRHCE using M-H algorithms is shown in two figures in the case of complete sampling. Each figure shows the plot, histogram and cumulative mean where $t = 0.1$



and $c = 1.5$ in Figure 2 while Figure 3 represents the case where $t = 0.2$ and $c = 2.5$.

Figure 2. MCMC convergence of DWCRTE, DWCRRE, and DWCRHCE estimates at $t = 0.1$ and $c = 1.5$

Source: created by researchers utilizing the R programming language.

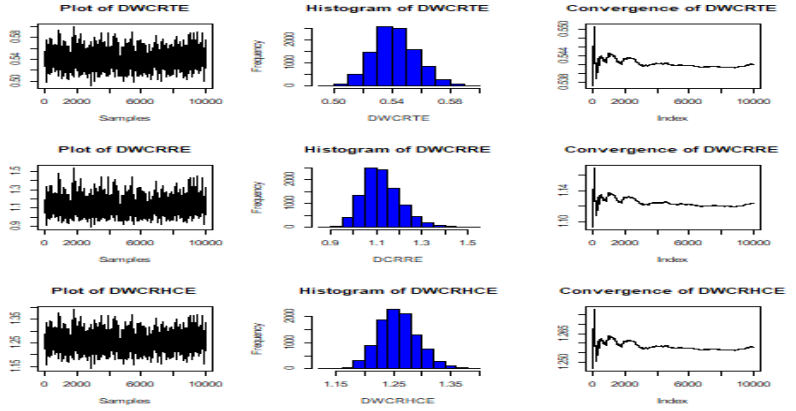


Figure 3. MCMC convergence of DWCRTE, DWCRRE, and DWCRHCE estimates at $t = 0.2$ and $c=2.5$

Source: created by researchers utilizing the R programming language.

Table 1. Estimates of different entropy measures under PT-IIC schemes given real dataset where m=15

Sch.	Method	$t = 0.1, c = 1.5$			$t = 0.2, c = 2.5$		
		DWCRTE	DWCRRE	DWCRHCE	DWCRTE	DWCRRE	DWCRHCE
S_1	MLE	1.5100	2.8129	2.5777	0.5322	1.0673	1.2349
	SLF	1.5249	2.8747	2.6032	0.5406	1.1102	1.2544
	LLF-1	1.5147	2.8321	2.5857	0.5401	1.1074	1.2531
	LLF-2	1.5612	3.0335	2.6651	0.5488	1.1550	1.2733
	PRLF	1.5309	2.9001	2.6134	0.5417	1.1159	1.2568
S_2	MLE	1.4986	2.7671	2.5583	0.5427	1.1217	1.2594
	SLF	1.4860	2.7175	2.5368	0.5414	1.1147	1.2563
	LLF-1	1.4828	2.7049	2.5313	0.5417	1.1162	1.2569
	LLF-2	1.4915	2.7388	2.5461	0.5415	1.1151	1.2565
	PRLF	1.4939	2.7484	2.5503	0.5431	1.1235	1.2601
S_3	MLE	1.4282	2.5043	2.4381	0.5105	0.9674	1.1845
	SLF	1.4396	2.5446	2.4576	0.5171	0.9964	1.1999
	LLF-1	1.4350	2.5281	2.4497	0.5185	1.0025	1.2031
	LLF-2	1.5249	2.8747	2.6031	0.5375	1.0939	1.2471
	PRLF	1.4561	2.6043	2.4858	0.5209	1.0136	1.2087
S_4	MLE	1.3820	2.3487	2.3591	0.4951	0.9047	1.1487
	SLF	1.3866	2.3638	2.3671	0.4998	0.9234	1.1597
	LLF-1	1.3971	2.3981	2.3849	0.5059	0.9484	1.1740
	LLF-2	1.3761	2.3298	2.3491	0.4932	0.8976	1.1445
	PRLF	1.3922	2.3822	2.3767	0.5006	0.9267	1.1616
com-plete	MLE	1.5228	2.8660	2.5996	0.5384	1.0986	1.2492
	SLF	1.5324	2.9066	2.6160	0.5434	1.1256	1.2610
	LLF-1	1.5272	2.8844	2.6071	0.5436	1.1265	1.2614
	LLF-2	1.5560	3.0104	2.6563	0.5485	1.1536	1.2728
	PRLF	1.5354	2.9195	2.6211	0.5438	1.1277	1.2619

Source: created by researchers utilizing the R programming language.

7. Simulation Study

This section evaluates the performance of estimate methodologies for BXIID under the PT-IIC scheme using a Monte Carlo simulation analysis. Specifically, maximum likelihood and Bayesian processes employing MCMC are considered. From the BXIID, we generate 1,000 random samples using the following guidelines:

1. Two cases of parameters of BXIID are assumed, namely: $(\delta, \lambda) = (1.5, 2.5)$ and$(\delta, \lambda) = (2.5, 1.5)$
2. Two cases of parameters of weighted entropy measures are assumed, namely: $(t, c) = (0.5, 1.5)$ and $(t, c) = (1.5, 2.5)$.
3. The true values of different entropy measures are listed in Table 2.

Table 2. True values of different entropy measures

δ	λ	t	c	$T^*(c)$	$R^*(c)$	$H^*(c)$
1.5	2.5	0.5	1.5	1.3621	2.2853	2.3252
		1.5	2.5	0.3434	0.4825	0.7968
2.5	1.5	0.5	1.5	1.2567	1.9796	2.1453
		1.5	2.5	0.3943	0.5964	0.9115

Source: created by researchers utilizing the R programming language.

4. The sample size is assumed to be $n = 40$ and $n = 60$.
5. The number of stages in the PT-IIC scheme is $m = 20, 30$ at $n = 40$, $m = 40$ and 50 at $n = 60$.
6. Removed items r_i are assumed to have n and m values as shown in Table 3, where$r_m = n - m - \sum_{i=1}^{m-1} r_i$ and $r_i$is the number of failure items.

Based on the produced data and the previously-made assumptions, the MLEs are calculated using PT-IIC. After that, the MLEs are used to calculate $T^*(c)$, $R^*(c)$ and $H^*(c)$ given the values of t and c. For the Bayesian method, BEs using the M-H algorithm under different loss functions in the case of gamma prior are computed, where the following hyper-parameters are assumed: $(a_1, b_1, a_2, b_2) = (1.5, 2.5, 1.75, 2.75)$.

Different loss functions, including SLF, LLF-1($\tau = 0.5$), LLF-2($\tau = -0.5$) and PRLF, are given. These values are then employed to determine the estimated values. The average (Avg.) of all entropy estimates and the Avg. of the MSE are presented in Tables 4 (a) to 5 (d), which incorporate all of the Monte Carlo simulation's inputs.

Table 3. Numerous patterns for removing items from life test at different number of stages

(n, m)	Censoring Schemes			
	S_1	S_2	S_3	S_4
(40,20)	$(20, 0^{*19})$	$(10, 0^{*18}, 10)$	(1^{*20})	$(0^{*19}, 20)$
(40,30)	$(10, 0^{*29})$	$(5, 0^{*28}, 5)$	$(0^{*10}, 1^{*10}, 0^{*10})$	$(0^{*29}, 10)$
(60,40)	$(20, 0^{*39})$	$(10, 0^{*38}, 10)$	$(0^{*10}, 1^{*20}, 0^{*10})$	$(0^{*39}, 20)$
(60,50)	$(10, 0^{*49})$	$(5, 0^{*48}, 5)$	$(0^{*20}, 1^{*10}, 0^{*20})$	$(0^{*49}, 10)$

Here, $(2^{*4}, 0)$, for example, means that the censoring scheme employed is (2,2,2,2,0).
Source: created by researchers.

Table 4 (a). The Avg. and MSE of different weighted entropy estimates for BXIID under PT-IIC schemes at$(\delta, \lambda) = (1.5, 2.5)$, and $(n, m) = (40, 20)$

(t, c)	Sch.	Estimate	DWCRTE		DWCRRE		DWCRHCE	
			Avg.	MSE	Avg.	MSE	Avg.	MSE
(0.5,1.5)	S_1	MLE	1.4108	0.0794	2.7101	2.6286	2.4084	0.2314
		SLF	0.9446	0.4437	1.4681	1.3845	1.6126	1.2929
		LLF-1	0.8564	0.5946	1.3158	1.6815	1.4620	1.7328
		LLF-2	1.0257	0.3293	1.6214	1.1383	1.7510	0.9596

(t, c)	Sch.	Estimate	DWCRTE		DWCRRE		DWCRHCE	
			Avg.	MSE	Avg.	MSE	Avg.	MSE
		PRLF	1.0293	0.3130	1.6195	1.1060	1.7571	0.9121
	S_2	**MLE**	1.3937	0.2089	3.0689	12.2305	2.3792	0.6087
		SLF	0.4167	3.2758	0.8684	3.3052	0.7114	9.5464
		LLF-1	0.2216	4.7021	0.6644	3.9941	0.3783	13.7029
		LLF-2	0.5888	2.2816	1.0751	2.7086	1.0051	6.6492
		PRLF	0.6137	2.0044	1.0844	2.6074	1.0477	5.8414
	S_3	**MLE**	1.4055	0.1730	3.0630	10.8346	2.3994	0.5042
		SLF	0.4537	2.7111	0.8827	3.1863	0.7746	7.9007
		LLF-1	0.2675	3.8704	0.6797	3.8568	0.4566	11.2793
		LLF-2	0.6184	1.9003	1.0883	2.6073	1.0557	5.5378
		PRLF	0.6402	1.6882	1.0959	2.5141	1.0929	4.9198
	S_4	**MLE**	1.4494	0.1353	3.6653	12.6546	2.4743	0.3944
		SLF	0.7864	0.4795	1.1076	1.8434	1.3424	1.3972
		LLF-1	0.7149	0.5654	0.9811	2.1068	1.2205	1.6476
		LLF-2	0.8485	0.4207	1.2309	1.6438	1.4485	1.2260
		PRLF	0.8377	0.4300	1.2074	1.6717	1.4300	1.2532
(1.5,2.5)	S_1	**MLE**	0.3752	0.0147	0.6654	0.6670	0.8707	0.0792
		SLF	0.2018	0.0438	0.2752	0.0897	0.4682	0.2358
		LLF-1	0.1756	0.0545	0.2384	0.1057	0.4076	0.2936
		LLF-2	0.2271	0.0349	0.3130	0.0763	0.5269	0.1880
		PRLF	0.2281	0.0337	0.3130	0.0742	0.5292	0.1816
	S_2	**MLE**	0.3793	0.0274	0.8829	2.5356	0.8800	0.1473
		SLF	0.0796	0.1314	0.1351	0.1853	0.1846	0.7077
		LLF-1	0.0394	0.1632	0.0905	0.2174	0.0913	0.8786
		LLF-2	0.1184	0.1049	0.1815	0.1565	0.2747	0.5647
		PRLF	0.1223	0.0994	0.1836	0.1518	0.2839	0.5350
	S_3	**MLE**	0.3818	0.0250	0.8531	2.2068	0.8859	0.1347
		SLF	0.0848	0.1228	0.1380	0.1797	0.1968	0.6611
		LLF-1	0.0453	0.1526	0.0937	0.2113	0.1051	0.8215
		LLF-2	0.1230	0.0979	0.1843	0.1516	0.2855	0.5271
		PRLF	0.1265	0.0930	0.1860	0.1471	0.2936	0.5008
	S_4	**MLE**	0.4105	0.0254	1.1969	4.9892	0.9526	0.1366
		SLF	0.1754	0.0390	0.2198	0.0918	0.4071	0.2098
		LLF-1	0.1592	0.0436	0.1955	0.1015	0.3694	0.2348
		LLF-2	0.1901	0.0358	0.2436	0.0849	0.4412	0.1926
		PRLF	0.1878	0.0360	0.2394	0.0855	0.4358	0.1941

Source: created by researchers utilizing the R programming language.

Table 4 (b). Avg. and MSE of different weighted entropy estimates for BXIID under PT-IIC schemes at $(\delta, \lambda) = (1.5, 2.5)$ and $(n, m) = (40, 30)$

(t, c)	Sch.	Estimate	DWCRTE		DWCRRE		DWCRHCE	
			Avg.	MSE	Avg.	MSE	Avg.	MSE
(0.5,1.5)	S_1	MLE	1.4217	0.0503	2.6143	0.6428	2.4270	0.1465
		SLF	1.1547	0.1414	1.8470	0.6809	1.9712	0.4122
		LLF-1	1.1107	0.1751	1.7480	0.7857	1.8961	0.5102
		LLF-2	1.1966	0.1143	1.9466	0.5968	2.0427	0.3330
		PRLF	1.1964	0.1113	1.9420	0.5834	2.0424	0.3243
	S_2	MLE	1.4242	0.0626	2.6550	0.7982	2.4313	0.1825
		SLF	1.0862	0.2244	1.7193	0.9094	1.8543	0.6540
		LLF-1	1.0335	0.2789	1.6101	1.0545	1.7643	0.8128
		LLF-2	1.1362	0.1802	1.8293	0.7898	1.9397	0.5252
		PRLF	1.1370	0.1744	1.8256	0.7714	1.9411	0.5082
	S_3	MLE	1.4242	0.0601	2.6487	0.7695	2.4313	0.1752
		SLF	1.0926	0.2134	1.7294	0.8854	1.8653	0.6219
		LLF-1	1.0415	0.2643	1.6224	1.0243	1.7779	0.7703
		LLF-2	1.1412	0.1719	1.8373	0.7709	1.9482	0.5011
		PRLF	1.1416	0.1668	1.8327	0.7538	1.9488	0.4861
	S_4	MLE	1.4242	0.0785	2.7197	1.5945	2.4313	0.2287
		SLF	1.0046	0.3581	1.5816	1.1984	1.7149	1.0435
		LLF-1	0.9404	0.4489	1.4612	1.3958	1.6054	1.3081
		LLF-2	1.0651	0.2848	1.7036	1.0321	1.8182	0.8300
		PRLF	1.0673	0.2734	1.7008	1.0069	1.8221	0.7967
(1.5,2.5)	S_1	MLE	0.3770	0.0103	0.6030	0.1413	0.8747	0.0554
		SLF	0.2679	0.0189	0.3700	0.0496	0.6217	0.1016
		LLF-1	0.2525	0.0224	0.3444	0.0556	0.5859	0.1204
		LLF-2	0.2830	0.0160	0.3960	0.0449	0.6567	0.0859
		PRLF	0.2831	0.0156	0.3954	0.0439	0.6570	0.0839
	S_2	MLE	0.3799	0.0127	0.6378	0.3206	0.8815	0.0685
		SLF	0.2445	0.0271	0.3361	0.0637	0.5674	0.1457
		LLF-1	0.2270	0.0321	0.3085	0.0719	0.5266	0.1728
		LLF-2	0.2617	0.0228	0.3643	0.0571	0.6072	0.1227
		PRLF	0.2621	0.0222	0.3639	0.0558	0.6081	0.1195
	S_3	MLE	0.3792	0.0121	0.6334	0.3127	0.8799	0.0651
		SLF	0.2468	0.0259	0.3388	0.0619	0.5726	0.1392
		LLF-1	0.2297	0.0306	0.3118	0.0697	0.5330	0.1649
		LLF-2	0.2635	0.0218	0.3664	0.0555	0.6114	0.1175
		PRLF	0.2637	0.0213	0.3657	0.0543	0.6118	0.1147
	S_4	MLE	0.3821	0.0155	0.7030	0.7839	0.8865	0.0836
		SLF	0.2189	0.0380	0.3008	0.0804	0.5079	0.2043
		LLF-1	0.1988	0.0451	0.2711	0.0913	0.4614	0.2429
		LLF-2	0.2385	0.0318	0.3314	0.0713	0.5534	0.1714
		PRLF	0.2392	0.0309	0.3312	0.0696	0.5550	0.1664

Source: created by researchers utilizing the R programming language.

Table 4(c). The Avg. and MSE of different weighted entropy estimates for BXIID under PT-IIC schemes at $(n, (\delta, \lambda) = (2.5, 1.5)$ and $(n, m) = (60, 40)$

(t, c)	Sch.	Estimate	DWCRTE		DWCRRE		DWCRHCE	
			Avg.	MSE	Avg.	MSE	Avg.	MSE
(0.5,1.5)	S_1	MLE	1.4210	0.0362	2.5731	0.4596	2.4258	0.1055
		SLF	1.2338	0.0707	2.0056	0.4220	2.1062	0.2060
		LLF-1	1.2049	0.0839	1.9326	0.4710	2.0569	0.2445
		LLF-2	1.2617	0.0598	2.0792	0.3841	2.1539	0.1744
		PRLF	1.2606	0.0589	2.0740	0.3773	2.1520	0.1715
	S_2	MLE	1.4226	0.0486	2.6113	0.6132	2.4285	0.1415
		SLF	1.1639	0.1287	1.8595	0.6310	1.9870	0.3751
		LLF-1	1.1274	0.1540	1.7750	0.7132	1.9246	0.4487
		LLF-2	1.1992	0.1076	1.9450	0.5639	2.0472	0.3137
		PRLF	1.1986	0.1054	1.9400	0.5537	2.0461	0.3072
	S_3	MLE	1.4234	0.0457	2.6074	0.5854	2.4300	0.1331
		SLF	1.1741	0.1171	1.8784	0.5962	2.0043	0.3413
		LLF-1	1.1392	0.1396	1.7965	0.6718	1.9448	0.4069
		LLF-2	1.2078	0.0983	1.9612	0.5346	2.0619	0.2865
		PRLF	1.2068	0.0966	1.9554	0.5257	2.0601	0.2814
	S_4	MLE	1.4209	0.0642	2.6426	0.7779	2.4256	0.1871
		SLF	1.0802	0.2275	1.7009	0.9062	1.8440	0.6629
		LLF-1	1.0334	0.2748	1.6034	1.0341	1.7642	0.8007
		LLF-2	1.1251	0.1879	1.8001	0.7982	1.9207	0.5476
		PRLF	1.1253	0.1829	1.7953	0.7834	1.9209	0.5329
(1.5,2.5)	S_1	MLE	0.3734	0.0073	0.5727	0.0428	0.8663	0.0395
		SLF	0.2944	0.0107	0.4083	0.0323	0.6830	0.0576
		LLF-1	0.2836	0.0123	0.3890	0.0353	0.6579	0.0662
		LLF-2	0.3050	0.0094	0.4279	0.0301	0.7078	0.0505
		PRLF	0.3048	0.0092	0.4270	0.0296	0.7072	0.0497
	S_2	MLE	0.3757	0.0098	0.5874	0.0588	0.8717	0.0525
		SLF	0.2686	0.0176	0.3687	0.0463	0.6234	0.0948
		LLF-1	0.2557	0.0204	0.3469	0.0512	0.5932	0.1096
		LLF-2	0.2815	0.0153	0.3909	0.0423	0.6531	0.0822
		PRLF	0.2814	0.0150	0.3901	0.0415	0.6529	0.0807
	S_3	MLE	0.3754	0.0091	0.5843	0.0550	0.8710	0.0489
		SLF	0.2725	0.0161	0.3740	0.0436	0.6323	0.0869
		LLF-1	0.2601	0.0186	0.3529	0.0481	0.6035	0.1002
		LLF-2	0.2848	0.0140	0.3954	0.0400	0.6608	0.0755
		PRLF	0.2845	0.0138	0.3944	0.0393	0.6602	0.0743
	S_4	MLE	0.3771	0.0125	0.6227	0.2506	0.8749	0.0673
		SLF	0.2405	0.0272	0.3275	0.0631	0.5580	0.1466
		LLF-1	0.2249	0.0316	0.3031	0.0705	0.5219	0.1704
		LLF-2	0.2558	0.0234	0.3527	0.0568	0.5937	0.1260
		PRLF	0.2559	0.0229	0.3519	0.0558	0.5939	0.1234

Source: created by researchers utilizing the R programming language.

Table 4(d). The Avg. and MSE of different weighted entropy estimates for BXIID under PT-IIC schemes at $(\delta, \lambda) = (2.5, 1.5)$ and $(n, m) = (60, 50)$

(t, c)	Sch.	Estimate	DWCRTE		DWCRRE		DWCRHCE	
			Avg.	MSE	Avg.	MSE	Avg.	MSE
(0.5,1.5)	S_1	MLE	1.4218	0.0295	2.5565	0.3722	2.4272	0.0861
		SLF	1.2762	0.0462	2.1027	0.3103	2.1787	0.1347
		LLF-1	1.2546	0.0531	2.0441	0.3366	2.1417	0.1549
		LLF-2	1.2973	0.0405	2.1615	0.2912	2.2147	0.1180
		PRLF	1.2962	0.0400	2.1569	0.2864	2.2128	0.1165
	S_2	MLE	1.4231	0.0332	2.5708	0.4189	2.4295	0.0966
		SLF	1.2558	0.0580	2.0556	0.3618	2.1439	0.1690
		LLF-1	1.2323	0.0670	1.9938	0.3952	2.1037	0.1953
		LLF-2	1.2787	0.0505	2.1177	0.3365	2.1829	0.1472
		PRLF	1.2777	0.0498	2.1131	0.3310	2.1812	0.1451
	S_3	MLE	1.4231	0.0326	2.5692	0.4127	2.4293	0.0951
		SLF	1.2570	0.0570	2.0581	0.3586	2.1459	0.1661
		LLF-1	1.2339	0.0657	1.9972	0.3912	2.1064	0.1916
		LLF-2	1.2796	0.0497	2.1194	0.3338	2.1844	0.1449
		PRLF	1.2784	0.0491	2.1145	0.3285	2.1824	0.1429
	S_4	MLE	1.4237	0.0373	2.5833	0.4680	2.4303	0.1087
		SLF	1.2340	0.0726	2.0070	0.4215	2.1066	0.2116
		LLF-1	1.2083	0.0843	1.9414	0.4635	2.0627	0.2457
		LLF-2	1.2591	0.0628	2.0732	0.3886	2.1494	0.1830
		PRLF	1.2581	0.0618	2.0686	0.3822	2.1477	0.1802
(1.5,2.5)	S_1	MLE	0.3726	0.0060	0.5657	0.0341	0.8646	0.0323
		SLF	0.3098	0.0074	0.4331	0.0246	0.7189	0.0401
		LLF-1	0.3014	0.0083	0.4174	0.0262	0.6993	0.0449
		LLF-2	0.3181	0.0067	0.4490	0.0235	0.7382	0.0362
		PRLF	0.3179	0.0066	0.4482	0.0231	0.7377	0.0357
	S_2	MLE	0.3737	0.0068	0.5710	0.0391	0.8671	0.0364
		SLF	0.3017	0.0091	0.4198	0.0284	0.7001	0.0489
		LLF-1	0.2927	0.0102	0.4034	0.0304	0.6792	0.0549
		LLF-2	0.3106	0.0081	0.4365	0.0269	0.7208	0.0438
		PRLF	0.3104	0.0080	0.4358	0.0265	0.7203	0.0432
	S_3	MLE	0.3734	0.0066	0.5698	0.0380	0.8665	0.0355
		SLF	0.3023	0.0089	0.4207	0.0279	0.7015	0.0478
		LLF-1	0.2935	0.0100	0.4045	0.0299	0.6810	0.0536
		LLF-2	0.3110	0.0080	0.4370	0.0265	0.7217	0.0429
		PRLF	0.3108	0.0079	0.4362	0.0260	0.7211	0.0423
	S_4	MLE	0.3744	0.0076	0.5757	0.0441	0.8688	0.0407
		SLF	0.2934	0.0110	0.4065	0.0325	0.6807	0.0591
		LLF-1	0.2837	0.0124	0.3892	0.0351	0.6582	0.0667
		LLF-2	0.3029	0.0098	0.4241	0.0306	0.7029	0.0526
		PRLF	0.3028	0.0096	0.4234	0.0301	0.7025	0.0518

Source: created by researchers utilizing the R programming language.

Table 5(a). The Avg. and MSE, of different weighted entropy estimates for BXIID under PT-IIC schemes at $(\delta, \lambda) = (1.5, 2.5)$ and $(n, m) = (40, 20)$

(t, c)	Sch.	Estimate	DWCRTE		DWCRRE		DWCRHCE	
			Avg.	MSE	Avg.	MSE	Avg.	MSE
(0.5,1.5)	S_1	MLE	1.3067	0.0516	2.2314	0.7609	2.2307	0.1504
		SLF	1.0486	0.1430	1.5868	0.5515	1.7901	0.4168
		LLF-1	0.9962	0.1856	1.4846	0.6601	1.7007	0.5410
		LLF-2	1.0976	0.1101	1.6882	0.4660	1.8738	0.3209
		PRLF	1.1056	0.1024	1.7018	0.4431	1.8873	0.2983
	S_2	MLE	1.3135	0.0780	2.3729	3.1072	2.2422	0.2274
		SLF	0.8639	0.4090	1.2871	1.0673	1.4748	1.1918
		LLF-1	0.7868	0.5359	1.1649	1.2812	1.3432	1.5618
		LLF-2	0.9345	0.3116	1.4074	0.8892	1.5953	0.9080
		PRLF	0.9491	0.2842	1.4277	0.8407	1.6202	0.8282
	S_3	MLE	1.3174	0.0793	2.4249	4.2087	2.2489	0.2311
		SLF	0.8543	0.4305	1.2749	1.1017	1.4584	1.2547
		LLF-1	0.7775	0.5607	1.1538	1.3151	1.3272	1.6339
		LLF-2	0.9247	0.3304	1.3942	0.9238	1.5786	0.9628
		PRLF	0.9402	0.3007	1.4157	0.8724	1.6050	0.8763
	S_4	MLE	1.3107	0.1344	2.5900	9.0307	2.2375	0.3917
		SLF	0.5819	1.4653	0.9524	1.9576	0.9934	4.2703
		LLF-1	0.4553	2.0057	0.8024	2.3463	0.7773	5.8452
		LLF-2	0.6946	1.0729	1.0997	1.6268	1.1857	3.1266
		PRLF	0.7249	0.9285	1.1309	1.5257	1.2374	2.7059
(1.5,2.5)	S_1	MLE	0.4189	0.0078	0.7511	0.5409	0.9721	0.0420
		SLF	0.3141	0.0171	0.4534	0.0597	0.7288	0.0921
		LLF-1	0.2966	0.0213	0.4210	0.0698	0.6883	0.1145
		LLF-2	0.3310	0.0137	0.4864	0.0517	0.7681	0.0738
		PRLF	0.3328	0.0130	0.4889	0.0497	0.7722	0.0702
	S_2	MLE	0.4250	0.0121	0.9170	1.9017	0.9862	0.0654
		SLF	0.2601	0.0374	0.3672	0.1030	0.6036	0.2012
		LLF-1	0.2378	0.0459	0.3312	0.1202	0.5518	0.2473
		LLF-2	0.2816	0.0302	0.4039	0.0883	0.6534	0.1628
		PRLF	0.2847	0.0286	0.4079	0.0849	0.6607	0.1539
	S_3	MLE	0.4259	0.0120	0.9255	2.0453	0.9882	0.0648
		SLF	0.2591	0.0379	0.3661	0.1043	0.6013	0.2041
		LLF-1	0.2371	0.0464	0.3305	0.1212	0.5502	0.2496
		LLF-2	0.2803	0.0308	0.4022	0.0899	0.6503	0.1661
		PRLF	0.2836	0.0291	0.4065	0.0863	0.6581	0.1568
	S_4	MLE	0.4289	0.0179	1.1547	4.2872	0.9952	0.0965
		SLF	0.1958	0.0752	0.2797	0.1632	0.4543	0.4047
		LLF-1	0.1666	0.0921	0.2391	0.1898	0.3865	0.4958
		LLF-2	0.2237	0.0610	0.3211	0.1398	0.5190	0.3283
		PRLF	0.2291	0.0569	0.3272	0.1338	0.5316	0.3066

Source: created by researchers utilizing the R programming language.

Table 5(b). The Avg. and MSE of different weighted entropy estimates for BXIID under PT-IIC schemes at $(\delta, \lambda) = (1.5, 2.5)$ and $(n, m) = (40, 30)$

(t, c)	Sch.	Estimate	DWCRTE		DWCRRE		DWCRHCE	
			Avg.	MSE	Avg.	MSE	Avg.	MSE
(0.5,1.5)	S_1	MLE	1.3110	0.0342	2.1955	0.3041	2.2381	0.0995
		SLF	1.1581	0.0580	1.7958	0.2938	1.9770	0.1692
		LLF-1	1.1276	0.0706	1.7270	0.3326	1.9250	0.2057
		LLF-2	1.1872	0.0482	1.8640	0.2651	2.0267	0.1405
		PRLF	1.1907	0.0461	1.8710	0.2569	2.0327	0.1344
	S_2	MLE	1.3152	0.0380	2.2154	0.3439	2.2453	0.1108
		SLF	1.1294	0.0749	1.7371	0.3508	1.9281	0.2183
		LLF-1	1.0968	0.0911	1.6660	0.3996	1.8724	0.2656
		LLF-2	1.1605	0.0619	1.8075	0.3128	1.9811	0.1805
		PRLF	1.1645	0.0592	1.8150	0.3028	1.9879	0.1724
	S_3	MLE	1.3155	0.0391	2.2181	0.3516	2.2457	0.1140
		SLF	1.1232	0.0810	1.7264	0.3695	1.9174	0.2359
		LLF-1	1.0909	0.0979	1.6567	0.4192	1.8623	0.2853
		LLF-2	1.1539	0.0673	1.7955	0.3305	1.9699	0.1961
		PRLF	1.1581	0.0643	1.8034	0.3199	1.9770	0.1873
	S_4	MLE	1.3186	0.0432	2.2352	0.3925	2.2510	0.1258
		SLF	1.0939	0.1014	1.6684	0.4316	1.8674	0.2954
		LLF-1	1.0580	0.1235	1.5936	0.4944	1.8061	0.3600
		LLF-2	1.1280	0.0834	1.7424	0.3812	1.9256	0.2431
		PRLF	1.1326	0.0795	1.7508	0.3684	1.9335	0.2317
(1.5,2.5)	S_1	MLE	0.4198	0.0054	0.6888	0.0451	0.9740	0.0292
		SLF	0.3519	0.0082	0.5220	0.0349	0.8166	0.0439
		LLF-1	0.3406	0.0097	0.4983	0.0389	0.7903	0.0522
		LLF-2	0.3630	0.0069	0.5460	0.0321	0.8424	0.0373
		PRLF	0.3637	0.0067	0.5469	0.0312	0.8439	0.0361
	S_2	MLE	0.4219	0.0062	0.7046	0.0948	0.9791	0.0335
		SLF	0.3421	0.0102	0.5037	0.0411	0.7937	0.0551
		LLF-1	0.3301	0.0121	0.4795	0.0460	0.7660	0.0653
		LLF-2	0.3537	0.0087	0.5282	0.0375	0.8207	0.0467
		PRLF	0.3545	0.0084	0.5293	0.0365	0.8225	0.0452
	S_3	MLE	0.4217	0.0063	0.7094	0.1337	0.9784	0.0341
		SLF	0.3406	0.0107	0.5014	0.0425	0.7903	0.0577
		LLF-1	0.3289	0.0126	0.4778	0.0474	0.7632	0.0681
		LLF-2	0.3520	0.0091	0.5252	0.0388	0.8167	0.0491
		PRLF	0.3528	0.0088	0.5265	0.0378	0.8187	0.0476
	S_4	MLE	0.4239	0.0073	0.7324	0.2407	0.9837	0.0390
		SLF	0.3301	0.0132	0.4824	0.0493	0.7660	0.0709
		LLF-1	0.3173	0.0156	0.4573	0.0553	0.7363	0.0841
		LLF-2	0.3426	0.0111	0.5078	0.0445	0.7949	0.0600
		PRLF	0.3435	0.0108	0.5091	0.0434	0.7971	0.0580

Source: created by researchers utilizing the R programming language.

Table 5(c). The Avg. and MSE, of different weighted entropy estimates for BXIID under PT-IIC schemes at $(\delta, \lambda) = (2.5,1.5)$ and $(n, m) = (60,40)$

(t, c)	Sch.	Estimate	DWCRTE		DWCRRE		DWCRHCE	
			Avg.	MSE	Avg.	MSE	Avg.	MSE
(0.5,1.5)	S_1	MLE	1.3122	0.0258	2.1817	0.2285	2.2401	0.0751
		SLF	1.2038	0.0334	1.8889	0.1946	2.0550	0.0974
		LLF-1	1.1833	0.0385	1.8390	0.2104	2.0199	0.1121
		LLF-2	1.2237	0.0294	1.9385	0.1839	2.0891	0.0858
		PRLF	1.2259	0.0286	1.9432	0.1800	2.0928	0.0832
	S_2	MLE	1.3162	0.0300	2.2017	0.2704	2.2469	0.0874
		SLF	1.1725	0.0464	1.8201	0.2450	2.0016	0.1353
		LLF-1	1.1502	0.0538	1.7681	0.2688	1.9635	0.1569
		LLF-2	1.1941	0.0404	1.8718	0.2268	2.0384	0.1176
		PRLF	1.1966	0.0391	1.8771	0.2215	2.0427	0.1138
	S_3	MLE	1.3166	0.0308	2.2042	0.2768	2.2476	0.0896
		SLF	1.1687	0.0494	1.8136	0.2570	1.9951	0.1440
		LLF-1	1.1469	0.0570	1.7631	0.2808	1.9580	0.1660
		LLF-2	1.1898	0.0432	1.8638	0.2387	2.0311	0.1259
		PRLF	1.1925	0.0418	1.8695	0.2331	2.0358	0.1218
	S_4	MLE	1.3190	0.0358	2.2210	0.3236	2.2516	0.1042
		SLF	1.1347	0.0678	1.7424	0.3206	1.9370	0.1977
		LLF-1	1.1093	0.0792	1.6861	0.3558	1.8937	0.2308
		LLF-2	1.1591	0.0584	1.7983	0.2923	1.9787	0.1701
		PRLF	1.1622	0.0563	1.8045	0.2848	1.9841	0.1641
(1.5,2.5)	S_1	MLE	0.4174	0.0039	0.6745	0.0313	0.9686	0.0213
		SLF	0.3685	0.0048	0.5522	0.0233	0.8551	0.0261
		LLF-1	0.3605	0.0055	0.5345	0.0251	0.8365	0.0298
		LLF-2	0.3764	0.0043	0.5701	0.0221	0.8733	0.0231
		PRLF	0.3767	0.0042	0.5706	0.0217	0.8742	0.0226
	S_2	MLE	0.4195	0.0048	0.6845	0.0393	0.9734	0.0258
		SLF	0.3573	0.0066	0.5304	0.0294	0.8291	0.0358
		LLF-1	0.3488	0.0076	0.5122	0.0320	0.8094	0.0409
		LLF-2	0.3657	0.0058	0.5488	0.0275	0.8485	0.0315
		PRLF	0.3662	0.0057	0.5495	0.0270	0.8496	0.0307
	S_3	MLE	0.4192	0.0048	0.6839	0.0393	0.9728	0.0259
		SLF	0.3568	0.0069	0.5297	0.0302	0.8279	0.0369
		LLF-1	0.3486	0.0078	0.5122	0.0327	0.8089	0.0420
		LLF-2	0.3648	0.0061	0.5474	0.0284	0.8465	0.0327
		PRLF	0.3654	0.0059	0.5483	0.0278	0.8478	0.0319
	S_4	MLE	0.4211	0.0058	0.6993	0.0829	0.9772	0.0313
		SLF	0.3442	0.0093	0.5060	0.0376	0.7988	0.0499
		LLF-1	0.3349	0.0106	0.4866	0.0411	0.7770	0.0573
		LLF-2	0.3534	0.0081	0.5255	0.0347	0.8201	0.0437
		PRLF	0.3541	0.0079	0.5264	0.0340	0.8216	0.0426

Source: created by researchers utilizing the R programming language.

Table 5(d). The Avg. and MSE of different weighted entropy estimates for BXIID under PT-IIC schemes at$(\delta, \lambda) = (2.5,1.5)$ and $(n, m) = (60, 50)$

(t, c)	Sch.	Estimate	DWCRTE		DWCRRE		DWCRHCE	
			Avg.	MSE	Avg.	MSE	Avg.	MSE
(0.5,1.5)	S_1	MLE	1.3126	0.0213	2.1735	0.1882	2.2407	0.0620
		SLF	1.2278	0.0238	1.9409	0.1514	2.0960	0.0694
		LLF-1	1.2117	0.0265	1.9003	0.1590	2.0685	0.0773
		LLF-2	1.2435	0.0218	1.9812	0.1470	2.1228	0.0635
		PRLF	1.2450	0.0213	1.9846	0.1448	2.1254	0.0621
	S_2	MLE	1.3142	0.0224	2.1803	0.1996	2.2434	0.0653
		SLF	1.2190	0.0263	1.9201	0.1614	2.0809	0.0766
		LLF-1	1.2026	0.0294	1.8794	0.1707	2.0530	0.0856
		LLF-2	1.2349	0.0239	1.9606	0.1555	2.1081	0.0696
		PRLF	1.2365	0.0233	1.9641	0.1530	2.1108	0.0680
	S_3	MLE	1.3145	0.0229	2.1820	0.2035	2.2440	0.0666
		SLF	1.2161	0.0276	1.9141	0.1672	2.0760	0.0803
		LLF-1	1.1999	0.0308	1.8740	0.1769	2.0484	0.0897
		LLF-2	1.2319	0.0250	1.9540	0.1608	2.1029	0.0729
		PRLF	1.2335	0.0245	1.9577	0.1582	2.1057	0.0713
	S_4	MLE	1.3156	0.0238	2.1871	0.2132	2.2458	0.0695
		SLF	1.2088	0.0297	1.8970	0.1757	2.0635	0.0866
		LLF-1	1.1919	0.0334	1.8555	0.1872	2.0347	0.0973
		LLF-2	1.2252	0.0268	1.9383	0.1678	2.0916	0.0781
		PRLF	1.2269	0.0262	1.9419	0.1649	2.0944	0.0762
(1.5,2.5)	S_1	MLE	0.4170	0.0032	0.6694	0.0252	0.9675	0.0174
		SLF	0.3775	0.0035	0.5698	0.0181	0.8760	0.0188
		LLF-1	0.3711	0.0039	0.5551	0.0191	0.8611	0.0210
		LLF-2	0.3839	0.0032	0.5847	0.0176	0.8908	0.0172
		PRLF	0.3841	0.0031	0.5850	0.0173	0.8913	0.0169
	S_2	MLE	0.4177	0.0035	0.6726	0.0274	0.9693	0.0187
		SLF	0.3743	0.0039	0.5632	0.0196	0.8685	0.0209
		LLF-1	0.3678	0.0043	0.5485	0.0207	0.8534	0.0233
		LLF-2	0.3807	0.0035	0.5781	0.0189	0.8835	0.0190
		PRLF	0.3810	0.0035	0.5784	0.0186	0.8840	0.0187
	S_3	MLE	0.4177	0.0035	0.6727	0.0278	0.9692	0.0190
		SLF	0.3735	0.0040	0.5617	0.0201	0.8667	0.0217
		LLF-1	0.3671	0.0045	0.5473	0.0213	0.8518	0.0242
		LLF-2	0.3798	0.0037	0.5763	0.0194	0.8813	0.0197
		PRLF	0.3801	0.0036	0.5767	0.0191	0.8820	0.0194
	S_4	MLE	0.4184	0.0038	0.6759	0.0300	0.9709	0.0202
		SLF	0.3705	0.0044	0.5557	0.0215	0.8598	0.0237
		LLF-1	0.3639	0.0049	0.5407	0.0228	0.8443	0.0265
		LLF-2	0.3771	0.0040	0.5708	0.0206	0.8751	0.0215
		PRLF	0.3774	0.0039	0.5712	0.0203	0.8757	0.0211

Source: created by researchers utilizing the R programming language.

It is evident from the tabulated result values that:

1- As n and m increase, the MSE of the Bayes estimate gradually decreases.

2- In comparison to other estimates, the MSE of all estimates based on DWCRTE frequently yields the smallest values.

3- As can be seen in Tables 4 (a) to 4 (d), in the majority of the cases, the precision measures of DWCRTE, DWCRRE and DWCRHCE under PRLF and LLF-2 are preferable to the corresponding estimates under LLF-1 and SLF for all schemes.

4- At true values $T^*(c) = 0.3434$ and 1.3621, the BEs of $T^*(c)$ under all loss functions are preferred over the other entropy measures for all schemes (see Tables 4(a) and 4(d)).

5- The MLEs and BEs of $T^*(c)$, $R^*(c)$ and $H^*(c)$ under different loss functions are decreasing as n and m increase from (40, 20) to (60, 50) (see Tables 4(a) and 4(d)).

6- The MSE of all BEs of $T^*(c)$ gets the smallest values compared to the others for all schemes at $(n\ ,\ m) = (60, 40)$ and $(t,\ c) = (0.5, 1.5)$ and $(1.5, 2.5)$ (see Table 5(c)).

7- In most cases, the MSEs of all estimates have the largest values in the case of S_3 and S_4 compared to other schemes, at $(n,\ m) = (60, 50)$, $(t,\ c) = (0.5, 1.5)$ and $(1.5, 2.5)$ (see Table 5(d)).

8- The MSE of $T^*(c)$ gets the smallest values under all loss functions in the case of S_1 at $m = 20$, $n = 40$ (see Figure 4).

9- At true values of $H^*(c) = 0.7968$ and $T^*(c) = 0.3434$, it can be observed that the BE of DWCRHCE gets the largest MSE, while the BE of DWCRTE gets the smallest value for all loss functions in Sch.1 (see Figure 5).

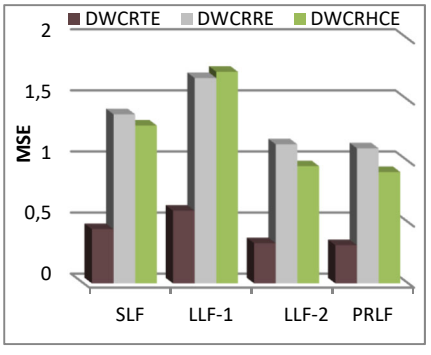


Figure 4. The MSE of different entropy estimates at $t = 0.5$

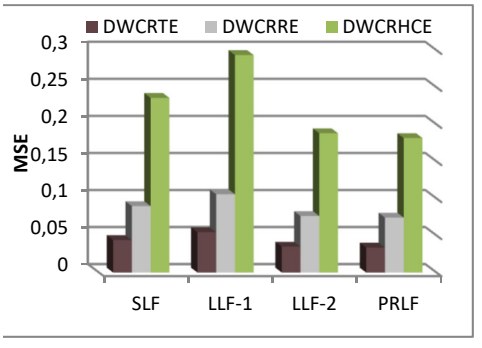


Figure 5. The MSE of different entropy estimates at $t = 1.5$

Source: created by researchers utilizing Microsoft Excel.

10- The MSE of the BE for $R^*(c)$ based on S_1 has the biggest values at $m = 50$, $n = 60$ and $t = 0.5$ (see Figure 6).

11- The MSE of the DWCRTE estimate under different loss functions based on S_1 typically produces the smallest values when compared to other estimates at $m = 50$, $n = 60$ and t=1.5 (see Figure 7).

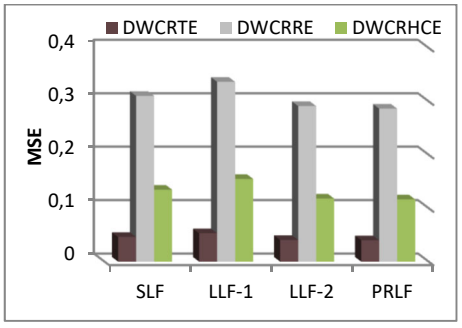


Figure 6. MSE of different Entropy estimates at $t = 0.5$, $m = 50$, and $n = 60$

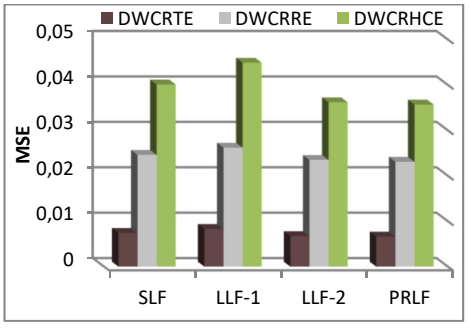


Figure 7. MSE of different Entropy estimates at $t = 1.5$

Source: created by researchers utilizing Microsoft Excel.

12- It can be observed from Figure 8 that the MSE of $\hat{T}^*(c)$ under PRLF takes the smallest values compared to the others in Sch. 4 at $m = 20$, $n = 40$, and $t = 0.5$.

13- At $m = 20$, $n = 40$, where the true value of $\hat{T}^*(c) = 0.3434$, the MSE of $\hat{T}^*(c)$ under LLF-2 takes the smallest values compared to the others under S_4 (Figure 9).

14- When compared to the other estimates from S_2, S_3, and S_4, the MSE of all entropy estimates based on S_1 often has the smallest values. The majority of entropy estimates (MLE, SLF, LLF-1, LLF-2, and PRLF) show a slight decrease as t increases.

15- It should be highlighted that, in comparison to the other estimates based on S_4, S_3 often produces the shortest MSE outcomes.

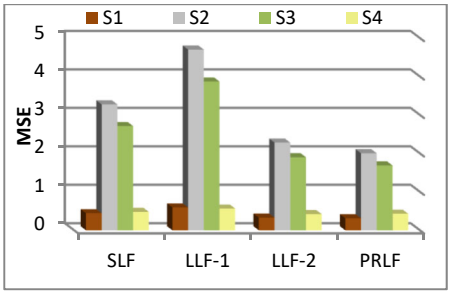


Figure 8. MSE of DWCRTE for different loss functions when $t = 0.5$

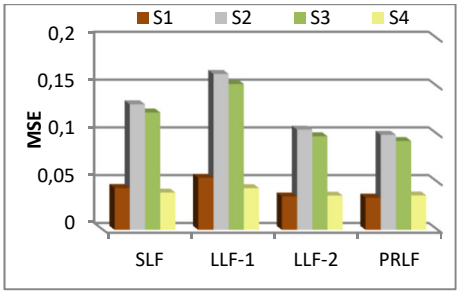


Figure 9. MSE of DWCRTE for different loss functions when $t = 1.5$

Source: created by researchers utilizing Microsoft Excel.

16- As can be seen in Figures 10 and 11, as n and m increase, the MSEs of $R^*(c)$ under different loss functions are decreasing.

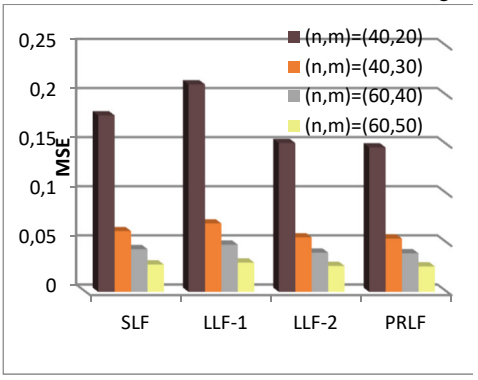


Figure 10. MSE of DWCRRE for different estimates under S_3 at $\delta = 1.5, \lambda = 2.5$ and t =1.5

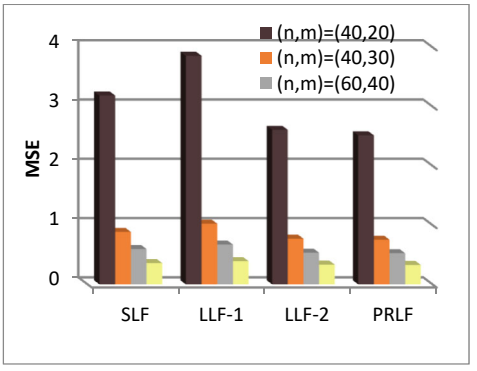


Figure 9. MSE of DWCRRE for different estimates under S_3 at$\delta = 1.5, \lambda = 2.5$ and t =0.5

Source: created by researchers utilizing Microsoft Excel.

8. Concluding Remarks

This article introduces the DWCRHC as an additional measure of uncertainty surrounding the residual lifetime function, particularly relevant in fields like survival analysis and reliability. The DWCRHC measure is formally defined within this work. We investigate the estimation of the DWCRHC, along with its related measures, DWCRRE and DWCRTE, for the BXIID under PT-IIC. Maximum and Bayesian estimation methods are employed. For the Bayesian estimation, we utilize the MCMC approach with the M-H algorithm, assuming a gamma prior distribution and considering three different loss functions. The article features the application, simulation studies, and an evaluation of the accuracy of the DWCRTE, DWCRRE, and DWCRHCE estimates for the BXIID.

Simulation results demonstrate that the BE of DWCRTE converges to the true value as the sample size increases. Generally, BEs under the PRLF exhibit the lowest MSE values, followed by the LLF-2, making them preferable over competing estimates. Furthermore, Sch. 1, compared to other schemes, often yields the lowest MSE values, followed in most cases by Sch. 2. The conclusions drawn from the simulated data are corroborated by the examination of actual data, particularly the water capacity data from the Shasta reservoir. These results are helpful in making well-informed decisions about the management of water resources. Future research could explore the application of the E-Bayesian technique to estimate other uncertainty metrics, such as dynamic weighted cumulative residual Shannon entropy.

References

Abd-Elfattah, A. M., Hassan, A. S. and Nassr, S. G., (2008). Estimation in step-stress partially accelerated life tests for the Burr Type XII distribution using type I censoring. *Statistical Methodology*, 5(6), pp. 502–514.

Abo-Eleneen, Z. A., (2011). The entropy of progressively censored samples. *Entropy*, 13 (2), pp. 437–49.

Ahmadini, A. H., Hassan, A. S., Zaky, A. N. and Alshqaq, S. S., (2020). Bayesian inference of dynamic cumulative residual entropy from Pareto II distribution with Application to COVID-19. *AIMS Mathematics*, 6(3), pp. 2196–2216.

Al-Babtain, A. A., Hassan, A.S., Zaky, A. N., Elbatal, I. and Elgarhy, M., (2021). Dynamic cumulative residual Rényi entropy for Lomax distribution: Bayesian and non-Bayesian methods. *AIMS Mathematics*, 6(3), pp. 3889–3914.

AlmarashiI, M., Algarni, A., Hassan, A. S., Zaky, A. S. and Elgarhy, M., (2021). Bayesian analysis of dynamic cumulative residual entropy for Lindley distribution. *Entropy*, 23, 1256. https://doi.org/10.3390/ e23101256.

Alyami, S. A., Hassan, A. S., Elbatal, I., Elgarhy, M., A. R. and El-Saeed, A. R., (2023). Bayesian and non- Bayesian estimates of the DCRTE for moment exponential distribution under progressive censoring type II. *Open Physics*. 21, 20220264.

Asadi, M., Zohrevand, Y., (2007). On the dynamic cumulative residual entropy. *Journal of Statistical Planning and Inference*, 137, pp. 1931–1941.

Balakrishnan, N., Aggrawala, R., (2000). *Progressive Censoring, Theory, Methods and Applications*. Birkhauser, Boston.

Baratpour, S., Ahmadi, J. and Arghami, N. R., (2007). Entropy properties of record statistics. *Statistical Papers*, 48, pp. 197–213.

Belis, M., Guiasu, S., (1968). A quantitative-qualitative measure of information in cybernetic systems. *IEEE Transactions on Information Theory*, IT-4, pp. 593–594.

Burr, W. I., (1942). Cumulative frequency functions. *Annals of Mathematical Statistics*, 13(2), pp. 215–232.

Belzunce, F., Navarro, J., Runiz, J. M. and Aguila, Y., (2004). Some results on residual entropy function. *Metrika*, 59, pp. 147–161.

Cho, Y., Sun, H. and Lee, K., (2015). Estimating the entropy of a Weibull distribution under generalized progressive hybrid censoring. *Entropy*, 17, pp. 102–122.

Cohen, A. C., (1963). Progressively censored samples in life testing. Technometrics, 5(3), pp. 327–339.

Di Crescenzo, A. D, Longobardi, M., (2006). On weighted residual and past entropies, Scientiae Mathematicae Japonicae, 64(3), pp. 255–266.

Evans, R. A., Simons, G., (1975). Research on statistical procedures in reliability engineering. ARL TR 75-0154, AD A013687.

Guiasu, S., (1986). Grouping data by using the weighted entropy. *Journal of Statistical Planning and Inference*, 15, pp. 63–69.

Gupta, P. L., Gupta, R. C. and Lvin, S. J., (1996). Analysis of failure time data by Burr distribution. *Communications in Statistics-Theory and Methods*, 25, pp. 2013–2024.

Hassan, A. S., Zaky, A. N., (2021). Entropy Bayesian estimation for Lomax distribution based on record. *Thailand Statistician*,19(1), pp. 96–115.

Hassan, A. S., Zaky, A. N., (2019). Estimation of entropy for inverse Weibull distribution under multiple censored data. *Journal of Taibah University for Science*, 13, pp. 331–337.

Hassan, A. S., Assar, S. M. and Ali, K.A., (2024a). Efficient estimation of the Burr XII distribution in presence of progressive censored samples with Binomial random removal. *Thailand Statistician*, 22(1), pp. 121–141.

Hassan, A. S., Alsadat, N., Balogu, O. S. and Helmy, B. A., (2024b). Bayesian and non-Bayesian estimation of some entropy measures for a Weibull distribution. *AIMS Mathematics*, 9(11), 32646–32673. DOI: 10.3934/math.20241563

Hassan, A. S., Elsherpieny, E. A. and Mohamed, R. E., (2022). Estimation of information measures for power-function distribution in presence of outliers and their applications. *Journal of Information and Communication Technology*, 21 (1), pp. 1–25.

Havrda, J., Charvat, F., (1967). Quantification method of classification process: Concept of structural α-entropy, *Kybernetika,* 3, pp. 30–35.

Helmy, B.A., Hassan, A. S. and El-Kholy, A. K., (2021). Analysis of uncertainty measure using unified hybrid censored data with applications. *Journal of Taibah University for Science*, 15(1), pp. 1130–1143.

Helmy, B. A., Hassan, A. S. and El-Kholy, A. K., (2023). Analysis of Uncertainty Weighted Measures for Pareto II distribution. *Reliability Theory & Applications*, 18(3), pp. 81–196.

Kayal, S., Balakrishnan, N., (2023). Weighted fractional generalized cumulative past entropy and its properties. *Methodology Computing Applied Probability*, 25, p 61. https://doi.org/10.1007/s11009-023-100350

Khammar, A.H., Jahanshahi, A. M. A., (2018). On weighted cumulative residual Tsallis entropy and its dynamic version. *Physica A: Statistical Mechanics*, 491, pp. 678–692.

Lee, K., (2017). Estimation of entropy of the inverse Weibull distribution under generalized progressive hybrid censored data. *Journal of the Korean Data & Information Science Society*, 28(3), pp. 659–668.

Li, X., Shi, Y., Wei, J. and Chai, J., (2007). Empirical Bayes estimators of reliability performances using LINEX loss under progressively type-II censored samples. *Mathematics and Computers in Simulation*, 73(5), pp. 320–326.

Misagh, F., Yari, G. H., (2011). On weighted interval entropy. *Statistics & Probability Letters*, 81, pp. 188–194.

Mohamed, M. S., (2022). On cumulative residual Tsallis entropy and its dynamic version of concomitants of generalized order statistics. *Communications in Statistics-Theory and Methods*, 51(8), pp. 2534–2551.

Mousa, M. A., Jaheen, Z. F., (2002). Statistical inference for the Burr model base on progressively censored data. *Computers & Mathematics with Applications*, 43, pp. 1441–1449.

Nair, R. S., Sathar and E. I. A., (2024). Bivariate dynamic weighted cumulative residual entropy. *Japanese Journal of Statistics and Data Science*, 7, pp. 83–100 (2024). https://doi.org/10.1007/s42081-023-00232-z

Nanda, A. K., Paul, P., (2006). Some results on generalized residual entropy, Information Sciences, 176, pp. 27–47.

Nadar, M., Papadopoulos and A. Kızılaslan, F., (2013). Statistical analysis for Kumaraswamy‘s distribution based on record data. *Statistical Papers*, 54, pp. 355–369.

Navarro, J., Del Aguila, Y. and Ruiz, J. M., (2001). Characterizations through reliability measures from weighted distributions. *Statistical Papers*, 42, pp. 395–402.

Panahi, H., Sayyareh, A., (2014). Parameter estimation and prediction of order statistics for the Burr type XII distribution with type II censoring, *Journal of Applied Statistics*, 41, pp. 215–232.

Qin, X., Gui, W., (2020). Statistical inference of Burr-XII distribution under progressive type-II censored competing risks data with binomial removals. *Journal of Computational and Applied Mathematics*, 378, 112922.

Rao, M., Chen, Y., Vemuri, B. C. and Wang, F., (2004). Cumulative residual entropy: a new measure of information. *IEEE Transactions Information Theory*, 50, pp. 1220–1228.

Renjini, K. R., Abdul Sathar, E. I. and Rajesh, G., (2016a). Bayes estimation of dynamic cumulative residual entropy for Pareto distribution under type-II right censored data. *Applied Mathematical Modelling*, 40, pp. 8424–8434.

Renjini, K. R., Abdul Sathar, E. I. and Rajesh, G., (2016b). A study of the effect of loss functions on the Bayes estimates of dynamic cumulative residual entropy for Pareto distribution under upper record values. *Journal of Statistical Computation and Simulation*, 86, pp. 324–339.

Renjini, K. R., Abdul Sathar, E. I. and Rajesh, G., (2018). Bayesian estimation of dynamic cumulative residual entropy for classical Pareto distribution. *American Journal of Mathematical and Management Sciences*, 37, pp. 1–13.

Rény, A., (1961). On measures of entropy and information in proceeding of the fourth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press: Berkeley, CA, USA, 1, pp. 547–561.

Sati, M. M., Gupta, N., (2015). Some characterization results on dynamic cumulative residual Tsallis entropy, *Journal of Probability and Statistics*. 23 http://dx.doi.org/10.1155/2015/694203.ArticleID 694203, p. 8.

Shannon, C. E., (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, pp. 379–423.

Smitha, S., Sudheesh, K. K. and Sreedevi, E. P., (2024). Dynamic cumulative residual entropy generating function and its properties. *Communications in Statistics - Theory and Methods*, 53(16), pp. 5890–5909.

Soliman, A. A., (2005). Estimation of parameters of life from progressively censored data using Burr-XII model. *IEEE Transactions on Reliability*, 54(1), pp. 34–42.

Sunoj, S.M., Linu, M. N., (2012). Dynamic cumulative residual Rényi's entropy. *Statistics*, 46(1), pp. 41–56.

Sunoj, S.M., Maya, S. S., (2006). Some properties of weighted distribution in the context of repairable systems. *Communications in Statistics—Theory and Methods*, 35(2), pp. 223–228.

Tsallis, C., (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52, pp. 479– 487.

Wang, M-C., (1996). Hazards regression analysis for length-biased data. *Biometrika,* 83, pp. 343–354.

Wang, X., Gui, W., (2021). Bayesian estimation of entropy for Burr Type XII distribution under progressive Type-II censored data. *Mathematics,* 9, p. 313. https://doi. org/10.3390/math9040313.

Wingo, D. R., (1993). Maximum likelihood methods for fitting the Burr type XII distribution to multiply (progressively) censored life test data. *Metrika*, 40, pp. 203–210.

Wu, J. W., Yu, H. Y., (2005). Statistical inference about the shape parameter of the Burr type XII distribution under the failure-censored sampling plan. *Applied Mathematics and Computation*, 163(1), pp. 443–482.

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 85–111, https://doi.org/10.59139/stattrans-2025-017
Received – 18.06.2024; accepted – 14.02.2025

Synchronization and similarity between regional and sectoral output gaps in the Polish manufacturing industry

Mirosław Błażej[1], **Mariusz Górajski**[2], **Magdalena Ulrichs**[3]

Abstract

The notion of the output gap is crucial in understanding underlying economic trends. The macroeconomic stabilization policy can only be effective for all sectors or regions of the economy if regional or sectoral output gaps are sufficiently coherent. This study proposes a new microeconometric model to calculate regional and sectoral output gaps in Poland's manufacturing industry over the period 2009–2020. The model is estimated using official data sources, which include capacity utilization from business tendency surveys and annual enterprise activity reports. To address potential endogeneity issues, the estimation employs control function methods. We assess the synchronization and similarity of the sectoral and regional output gaps in Poland's manufacturing industry. We also analyze the impact of the COVID-19 pandemic on the coherence of regional and sectoral output gaps.

Key words: output gap coherence, capacity utilization, control function method, sectoral and regional output gaps, COVID-19.

1. Introduction

Understanding underlying economic trends requires recognizing the potential outputs and output gaps for the whole economy and regional and sectoral levels. These measures of the business cycle have become indispensable instruments for assessing the cyclical position and productive capacity of a sector, region, or the whole economy. Potential output measures the ability of the economy, sector, or region to generate long-term, non-inflationary growth. At the same time, the output gap indicates the degree of

[1] Macroeconomic Studies and Finance Statistics Department, Statistics Poland, Poland. E-mail: m.blazej@stat.gov.pl. ORCID: https://orcid.org/0000-0003-4482-8996.

[2] Department of Econometrics, Faculty of Economics and Sociology, University of Lodz, Lodz, Poland; Macroeconomic Studies and Finance Statistics Department, Statistics Poland, Poland. E-mail: mariusz.gorajski@uni.lodz.pl. ORCID: https://orcid.org/0000-0002-2591-8657.

[3] Department of Econometrics, Faculty of Economics and Sociology, University of Lodz; Lodz, Poland Macroeconomic Studies and Finance Statistics Department, Statistics Poland, Poland. E-mail: magdalena.ulrichs@uni.lodz.pl. ORCID: https://orcid.org/0000-0002-9630-5460.

overheating or slack compared to the potential output (Havik et al. 2014; Blondeau et al. 2021). A negative output gap may indicate a recession or the early stages of recovery, whereas a positive output gap suggests a period of economic overheating. The macroeconomic literature emphasizes the output gap as a primary indicator of inflationary pressures in the Phillips curves and a valid fiscal and monetary policy rule response variable (Woodford 2003; Walsh 2010; Gali 2009). Despite its importance in monetary and fiscal policy, there has yet to be an agreement on the methodology for estimating and assessing unobservable potential production at sectoral or regional levels (Orphanides and van Norden 2002; Ódor and Jurasekova Kucserova 2014; Blagrave et al. 2015; Edge and Rudd, 2016; Álvarez and Gómez-Loscos 2018; Quast and Wolters 2020; Pu et al. 2023).

In this study, using microdata from the manufacturing industry in Poland, we estimate microdata-based regional and sectoral output gaps and calculate their coherence with several overall output gap measures in Poland. The standard macroeconomic policy is only optimal for all economic sectors or regions if the regional or sectoral output gaps are sufficiently coherent. When we experience low values of similarity and synchronization between industries and regions in terms of a measure of business activity, we should adjust the sector- or region-specific structural policy. However, measurements of output gaps at regional or sectoral levels and assessments of their coherence still need to be made available. Peykov (2021), using the Hodrick Prescot filter, identified the sectors that determine the cyclical state of the Bulgarian economy over the period 2000-2019. The author found the contribution and correlations of individual sector output gaps to the overall economic output gap. The new output gap coherence measures have been proposed by Mink et al. (2012) and de Haan et al. (2023) and applied to examine whether the standard monetary policy is equally optimal for all countries in the Euro area. The authors proposed the synchronization and similarity measures between output gaps that take differences in the signs and amplitudes of the output gaps more adequately into account than the correlation coefficient. These measures are also used in Jokubaitis and Celov (2023) to analyze the EU's sectoral and regional business cycle synchronization.

This study employs a novel micro-econometric model to estimate sectoral and regional output gaps based on business tendency surveys and annual enterprise activity reports. We apply this approach to Poland's manufacturing industry using survey microdata that involves the firm's subjective judgment on their productive capacity (Statistics Poland 2023) and firm-level data from reports on enterprises' activities (Statistics Poland 2019) to estimate microdata-based output gap measures. However, we are distinguishing two separate stages in the estimation procedure. We use the firm-level data from annual enterprise activity reports in the first stage. We apply the Olley and Pakes (1996) model with Ackerberg et al. (2015) correction to estimate the sectoral

production functions and determine individual enterprises' total factor productivity (TFP). There are only several studies that estimate production function for the Polish economy using firm-level data (see Gosińska et al. 2024, Górajski and Błażej 2020; Hagemejer and Kolasa 2011 and references therein). The second stage of our approach describes the relationship between firm-level TFP and capacity utilization that decomposes firm-level unadjusted TFP indices into business cycle components and technology levels. We estimate this equation using a dynamic panel data model based on data from the business tendency surveys (Kripfganz 2020). Next, we calculate the potential values of all inputs by extracting statistical trends from firm-level data and use them in production functions to construct aggregate indices of potential outputs and output gaps. We measure the enterprise's potential levels of capacity utilization, labor, and capital by their trend component obtained by applying the Hodrick-Prescott filter. Finally, we employ the synchronization and similarity output gap measures as in (Mink et al. 2012) to calculate the level of sectoral and regional output gap coherence for the Polish regions and manufacturing industry sectors. We also confirm the impact of the COVID-19 pandemic on output gaps in all regions and sectors of the manufacturing industry in Poland. These new measures of economic activity are essential for policymakers to adjust macroeconomic policy tools to the characteristics of a particular industry or region. Comparing regional and sectoral business cycles, we consider both the sign and the amplitude of the output gap and assess the level of coherence between the Polish regions and sectors of the manufacturing industry in Poland. We employ a microdata-based approach to estimate the output gap using information about an individual firm's level of capacity utilization, a leading indicator of potential growth. Capacity utilization is a core item in the European Union's harmonized business tendency survey conducted in several countries. As the results are typically available before the end of the period, they are a valuable tool for monitoring the current economic situation, and they contain information unavailable from official sources. TFP relates to an available technology's labor and capital efficiency levels and depends on labor and capital capacity utilization rates. We observe that capacity utilization positively impacts the TFP cycle. Thus, we decompose firm-level unadjusted TFP indices into business cycle components, represented by capacity utilization rates and actual technology levels in the sector. The literature has investigated the relationship between capacity utilization and production (Gradzewicz et al. 2017; Havik et al. 2014; Planas et al. 2013, Fernald, 2012; Graff and Strum, 2010; Greenwood et al. 1988; Wen, 1998). Moreover, Graff and Strum (2010) considered a regression of capacity utilization on the output gap for 22 OECD countries to confirm that using survey data can produce estimates significantly closer to later releases of output gap estimates.

In this study, we use Statistics Poland's Business Tendency Surveys (BTS, see Statistics Poland 2023) and Annual Non-Financial Enterprises Survey (ANFES, see Statistics Poland 2019) data regarding Polish enterprises in the manufacturing industry that employ at least ten employees, thus excluding micro-enterprises. We measure the total economic output in the manufacturing industry using the sum of all firm-level gross value-added.

The remainder of this article is structured as follows: Section 2 describes our methodology for the estimation of regional and sectoral output gaps and defines the coherence measures. Section 3 presents the empirical application of this study. Section 4 concludes with a discussion of the further study's opportunities for other applications.

2. Research methods

2.1. Production function, TFP, and capacity utilization

We assume that the gross value-added Y_{it} for enterprise i in period t is determined by the Cobb-Douglas function:

$$Y_{it} = TFP_{it} K_{it}^{\beta_{k,d}} L_{it}^{\beta_{l,d}} e^{\varepsilon_{it}}, i \in S_d, \tag{1}$$

where TFP_{it} is the unadjusted total factor productivity; L_{it}, K_{it} are the quantities of labor and capital, respectively; and L_{it} is the number of employees at the end of period t, $e^{\varepsilon_{it}}$ is output shock. The variables Y_{it} and K_{it} express production and capital values, respectively, and are not fully observable. Nevertheless, they can be measured by setting actual real gross value-added and physical capital levels in the enterprise. The parameters $\beta_{k,d}$ and $\beta_{l,d}$ denote the gross value-added elasticities of capital and labor, respectively, for homogenous groups of firms S_d standing for the sector of the economy. Moreover, TFP_{it} is an autoregressive (Markov) process:

$$TFP_{it} = A_{it} g(TFP_{it-1}), i \in S_d. \tag{2}$$

The level of individual technology A_{it} used in the production process of sector d is an unobservable variable that is decomposed into the product of the average productivity of companies in the given economic region and for a given year $e^{\lambda_{r,d}+\lambda_{t,d}}$ (where $\lambda_{r,d}$ is a regional dummy variable; and $\lambda_{t,d}$ is a dummy variable for the year) and the independent white-noise idiosyncratic component $e^{\xi_{it}}$. Consequently, we obtain

$$A_{it} = e^{\lambda_{r,d}+\lambda_{t,d}+\xi_{it}}. \tag{3}$$

Equations (1) through (3) provide the foundation of our two-stage micro-econometric model for output gap estimation. Our model differs from the standard neo-classical specification of the production function by including the exogenous rate of capacity utilization (Greenwood et al., 1988), among others.

Hereafter, let $y_{it}, l_{it}, k_{it},$ and ω_{it} denote the logarithms of variables Y_{it}, L_{it}, K_{it} , and TFP_{it}, respectively. Then, production equation (1) can be presented in a log-linear form as

$$y_{it} = \omega_{it} + \beta_{k,d} k_{it} + \beta_{l,d} l_{it}. \tag{4}$$

The productivity coefficient ω_{it} is often interpreted as a state variable in the company-decision problem, which involves the selection of production factors. We determine the enterprise's individual TFP by finding the output elasticities in Equation (4). The firm-level production function in Equation (4) can be estimated using control function methods, such as Olley and Pakes, (1996) model (OP model) and Levinsohn and Petrin, (2003) model (LP model), both of which can be enhanced by the correction made by Ackerberg et al., (2015). Control function methods use different proxy variables to approximate productivity shocks and estimate a company's probability of survival in the market. Moreover, the productivity coefficients acknowledge the Markovian structure. As a result, the OP and LP models produce robust estimates of output elasticities against the endogeneity of explanatory variables and attrition (Van Beveren, 2012). As we use the OP model due to data availability restrictions to estimate the enterprise production function, we assume that company investment expenditures control for unobserved TFP indices.

Estimating the production equation in the OP model is a two-stage procedure. In the first stage, we approximate the unobservable productivity shocks using a polynomial function of capital and proxy variables represented by investment outlays. Finally, we substitute the results from the first step in calculating the production function to obtain a non-linear regression equation for the gross value-added for the enterprises that survived in the market. We employ the correction in the control function approach developed by Ackerberg et al. (2015). These authors prove that labor input may not vary independent of the productivity approximation function estimated using a low-order polynomial of capital and proxy variables. To avoid this collinearity problem, they estimate labor and capital coefficients in the second stage using the generalized method of moments (GMM) approach.

We built on works by Fernald (2012) and Gradzewicz et al. (2017) to define firm-level capacity utilization as the aggregated capital and labor utilization rates. While labor utilization can be measured as the number of hours worked per capita during the period, capital utilization combines at least two essential dimensions: the number of hours the plant operates per period and its intensity expended during the period. Many studies suggest that the capacity utilization rate may be necessary to understand business cycles (e.g. Greenwood et al., 1988; Wen, 1998; Fernald, 2012; Gradzewicz et al., 2017). Using the control function method, Blazej et al. (2025) estimate sector-specific production functions for the divisions of the manufacturing sector, which include three inputs: labor, capital, and capacity utilization. Here, we focus on two-input production functions and assume that capacity utilization is a cyclical component of TFP.

We assume similar to Fernald (2012) that unadjusted total factor productivity TFP_{it} in (2) is decomposed into a product of lagged firm-level unadjusted productivity TFP_{it-1}^{ρ}, and cyclical component represented by capacity utilization term ${U_{it}}^{u}$. Thus, we approximate the firm-level TFP in (2) by

$$\widehat{\omega}_{it} = \lambda_t + \rho_d \widehat{\omega}_{it-1} + \beta_{u,d} u_{it} + \xi_{it}, \tag{5}$$

where we assume that the firm-level unadjusted total factor productivity coefficient $\widehat{\omega}_{it} = \log\widehat{TFP}_{it}$, is derived from the OP model according to Equation (1). The set of explanatory variables for the unadjusted TFP coefficient includes the capacity utilization U_{it}, lagged productivity coefficient $\widehat{\omega}_{it-1}$, and a complete set of time dummies λ_t for the estimation period. Time dummies reflect trends in TFP, which can be associated with technological changes.

2.2. Potential variables and output gaps

Similar to the OECD's methodology, we define the company-level capacity utilization gap as the log difference between a firm's individual capacity utilization and potential capacity utilization (Chalaux and Guillemette, 2019). The trend component measures the potential level of capacity utilization U_{it}^{POT} obtained by the Hodrick-Prescott filter (HP) (Phillips and Shi, 2021). Therefore, the firm-level potential TFP coefficients are derived from Equation (5) and take the following form:

$$\widehat{\omega}_{it}^{POT} = \hat{\rho}_d \widehat{\omega}_{it-1} + \hat{\lambda}_t + \hat{\beta}_{u,d} u_{it}^{POT}, \tag{6}$$

where $u_{it}^{POT} = \log U_{it}^{POT}$; U_{it}^{POT} is the HP trend of U_{it}; $\widehat{\omega}_{it-1}$ is the lagged productivity coefficient of company $i \in S_d$, $\hat{\beta}_{u,d}$; and $\hat{\lambda}_t$, $\hat{\rho}_d$ are parameter estimates of the dynamic panel data model in Equation (5).

We use the HP filter for every company i to calculate the potential levels of labor and capital stock L_{it}^{POT} and K_{it}^{POT}, respectively. In substituting the potential levels of TFP coefficients (6) and firm-level potential labor and capital stock into the production function in Equation (4), we obtain the following specifications of firm-level potential outputs for $i \in S_d$:

$$y_{it}^{POT\,UKL} = \widehat{\omega}_{it}^{POT} + \hat{\beta}_{k,d} k_{it}^{POT} + \hat{\beta}_{l,d} l_{it}^{POT}, \tag{7}$$

$$y_{it}^{POT\,K} = \widehat{\widehat{\omega}}_{it} + \hat{\beta}_{k,d} k_{it}^{POT} + \hat{\beta}_{l,d} l_{it}, \tag{8}$$

$$y_{it}^{POT\,L} = \widehat{\widehat{\omega}}_{it} + \hat{\beta}_{k,d} k_{it} + \hat{\beta}_{l,d} l_{it}^{POT}, \tag{9}$$

$$y_{it}^{POT\,U} = \widehat{\omega}_{it}^{POT} + \hat{\beta}_{k,d} k_{it} + \hat{\beta}_{l,d} l_{it}, \tag{10}$$

where $\hat{\beta}_{k,d}$ and $\hat{\beta}_{l,d}$ are the estimates of output elasticities in sector d; k_{it}^{pot} and l_{it}^{pot} are the logs of K_{it}^{POT} and L_{it}^{POT} respectively; and $\widehat{\widehat{\omega}}_{it} = \hat{\rho}_d \widehat{\omega}_{it-1} + \hat{\lambda}_t + \hat{\beta}_{u,d} u_{it}$ is the theoretical value of firm-level productivity. We extract the unpredictable component ξ_{it} from the output, and consider the theoretical values of the outputs defined by

$$\hat{y}_{it} = \widehat{\widehat{\omega}}_{it} + \hat{\beta}_{k,d} k_{it} + \hat{\beta}_{l,d} l_{it}. \tag{11}$$

Based on the potential and theoretical output values, we propose the following definition of the microdata-based output gaps at the sectorial (d) and regional (r) and the whole industry (S) levels:

$$\hat{x}_{t,s}^{UKL} = \log \hat{Y}_{st} - \log Y_{st}^{POT\ UKL}, s \in \{S, d, r\} \tag{12}$$

$$\hat{x}_{t,s}^{K} = \log \hat{Y}_{st} - \log Y_{st}^{POT\ K}, s \in \{S, d, r\} \tag{13}$$

$$\hat{x}_{t,s}^{L} = \log \hat{Y}_{st} - \log Y_{st}^{POT\ L}, s \in \{S, d, r\} \tag{14}$$

$$\hat{x}_{t,s}^{U} = \log \hat{Y}_{st} - \log Y_{st}^{POT\ U}, s \in \{S, d, r\} \tag{15}$$

where $\hat{Y}_{dt} = \sum_{i \in S_d} \hat{Y}_{it}$, $\hat{Y}_{rt} = \sum_{i \in S_r} \hat{Y}_{it}$, and $\hat{Y}_{St} = \sum_d \hat{Y}_{dt} = \sum_r \hat{Y}_{rt}$ are aggregated firm-level theoretical outputs $\hat{Y}_{it} = e^{\hat{y}_{it}}$; and $Y_{dt}^{POT\ j} = \sum_{i \in S_d} Y_{it}^{POT\ j}$, $Y_{rt}^{POT\ j} = \sum_{i \in S_r} Y_{it}^{POT\ j}$, and $Y_{St}^{POT\ j} = \sum_d Y_{dt}^{POT\ j} = \sum_r Y_{rt}^{POT\ j}$ are aggregated firm-level potential outputs $Y_{it}^{POT\ j} = e^{y_{it}^{POT\ j}}$ for $j = \{UKL, U, K, L\}$. Finally, we decompose the microdata-based output gap $\hat{x}_t^{UKL}$ into three main, sectorial, and regional components:

$$\hat{x}_{t,S}^{UKL} = \hat{x}_{t,S}^{U} + \hat{x}_{t,S}^{K} + \hat{x}_{t,S}^{L} + res_t \tag{16}$$

$$\hat{x}_{t,S}^{UKL} = \sum\nolimits_d w_{t,d} \hat{x}_{t,d}^{UKL} + res_{t,sectoral} \tag{17}$$

$$\hat{x}_{t,S}^{UKL} = \sum\nolimits_r w_{t,r} \hat{x}_{t,r}^{UKL} + res_{t,reginal}, \tag{18}$$

where $\hat{x}_{t,S}^{U}$, $\hat{x}_{t,S}^{K}$, and $\hat{x}_{t,S}^{L}$ capture capacity utilization, and the effects of capital and labor gaps on the total output gap, respectively; $res_t, res_{t,sectoral}$, and $res_{t,regional}$ reflect the mixed or interactive effects; and $w_{t,d}, w_{t,r}$ are gross value-added-based weights for the sectoral and regional decomposition of the economy-wide output. Further, we call $\hat{x}_{t,S}^{UKL}$, $\hat{x}_t^{K}$, $\hat{x}_{t,S}^{L}$, and $\hat{x}_{t,S}^{U}$ the microdata-based output gap, capital-based output gap, labor-based output gap, and capacity utilization-based output gap, respectively.

2.3. Output gap coherence measures

We assess the level of coherence between the Polish regions and sectors of the manufacturing industry in Poland by synchronization and similarity measures proposed in Mink et al. (2012). Denoting the output gap of region or sector i at time t by $\hat{x}_{t,i}^{UKL}$ and the reference output gap by x_t, we calculate synchronicity between individual region $i = r$ or sector $i = d$ with the whole economy in period t as:

$$\varphi_{it} = \frac{\hat{x}_{t,\mathrm{i}}^{UKL} x_t}{\left| \hat{x}_{t,\mathrm{i}}^{UKL} x_t \right|}. \tag{19}$$

Measure defined by Equation (25) is specified on the [-1,1] scale, where the value of 1 indicates that output gap i has the same sign as the reference. The overall synchronicity of all regions or sectors with the reference output gap is defined as $\varphi_t = \frac{1}{n}\sum_i \varphi_{it}$,
where n is the total number of regions or sectors analyzed. This measure is defined on a $[-1+2/n\,,1]$ scale, it equals 1 when all output gaps have the same sign as the reference.

The similarity measure of coherence considers the differences in the amplitudes of output gaps.

$$\gamma_{it} = 1 - \frac{\left|\hat{x}_{t,\mathrm{i}}^{UKL} - x_t\right|}{\sum_i \left|\hat{x}_{t,\mathrm{i}}^{UKL}\right|/n}. \tag{20}$$

The overall similarity of all regions or sectors with the reference output gap is defined as $\gamma_t = \frac{1}{n}\sum_i \gamma_{it}$,:

This measure is defined on a $[2-n, 1]$ scale, which equals 1 when all output gaps are identical.

3. Results

In this section, we calculate novel microdata-based output gaps for the sectors and regions of the Polish manufacturing industry. The firm-level output gap estimation uses capacity utilization, labor, physical capital, and production output (gross value added) data. Data on the gross value-added, capital, and labor originate from the ANFES (2008–2020), which examines the business activity of Polish enterprises in the manufacturing industry employing at least ten employees. Enterprise capacity utilization data is collected from quarterly BTS in Poland. The enterprise's gross value-added Y is the difference between its global output and intermediate consumption. Physical capital is defined as the enterprise's average annual fixed assets. The final measurement for variables Y and K is determined by calculating the enterprise's real gross value-added and real physical capital at constant average prices from 2015. For this purpose, we use capital and gross value-added deflators in the two-digit NACE sectors. As an approximation of capacity utilization, we use a survey-based measure of answers to the question regarding capacity utilization in each company. The BTS in Poland's manufacturing sector is conducted based on a monthly questionnaire and includes additional questions regarding capacity utilization in each company: "What percentage of your company's total production capacity is currently used?" Essential characteristic of these survey data is that they are generally not subject to revisions but become final when the survey is complete. Since the data are final upon survey completion, output gap estimates based on these data remain stable over time. This

property of microdata-based output gap contrasts with other macroeconomic indicators (e.g. GDP), often revised, leading to potential re-evaluations of past output gaps.

3.1. Production function estimation

We estimate the production functions defined by Equations (1) and (2), where sectors S_d are defined by NACE Rev. 2 divisions of the manufacturing industry in Poland ($d = 10,11, \ldots ,33$) and regional dummy variables $\lambda_{r,d}$ indicate the region ($r = 1,2, \ldots 16$; alphabetical order of voivodeships). The estimations were conducted separately for each division on a total sample of 199,745 observations for 40,025 firms.

Figure 1 summarizes the production function estimation results for all manufacturing industry divisions. We employ the OP model with time and regional dummies for each homogenous division. The estimators' standard errors are determined using a bootstrap procedure.

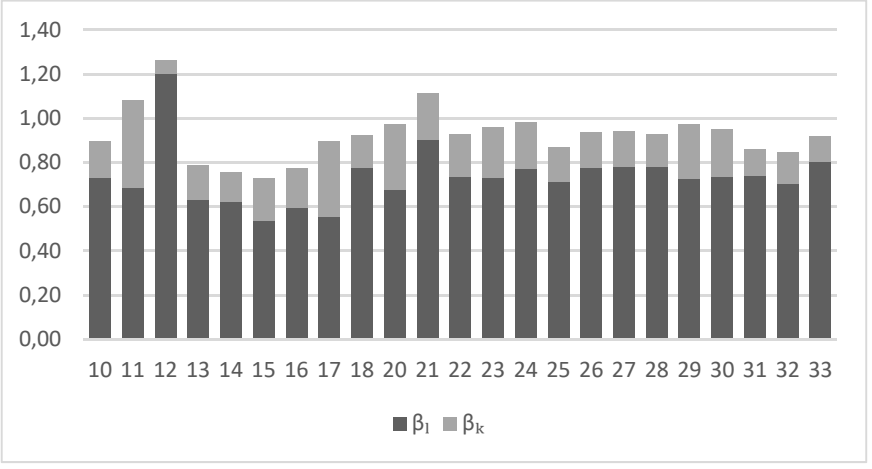


Figure 1: Production function estimation (OP model): input elasticities by division

Source: own calculations based on ANFES data (2008–2020). Note: β_k and β_l denote the gross value-added elasticities of capital and labor, respectively, for the sectors of the economy defined by NACE Rev. 2 divisions of the Polish manufacturing industry. Numbers 10, 11, …, 33 indicate the NACE divisions.

In all analyzed regressions, Student's t-tests indicate that labor and capital statistically significantly impact the companies' gross value-added (all p-values < 0.01). We observe an increasing return to scale only in the production of beverages ($d = 11$), tobacco ($d = 12$), and pharmaceuticals ($d = 21$). Figure 18 in Appendix D provides estimates of the time-effects relative to 2020; in nearly all divisions the growth time effects on the TFP indices are confirmed. Figure 19 in Appendix D displays the heterogeneity of the regional impacts on the firm-level TFP.

The estimated parameters of the production function serve to determine the company's unadjusted TFP. TFP captures how efficiently firms transform inputs (e.g. capital and labor) into outputs. TFP is interpreted as part of the firm-level value-added at the firm level that inputs cannot explain. The TFP estimation based on micro-level data allows for the inclusion of underlying productivity heterogeneity across different industries and individual determinants of productivity.

3.2. Capacity utilization and adjusted TFP

This section analyzes the relationship between TFP and capacity utilization of enterprises in the Polish manufacturing industry. We estimate the influence of capacity utilization level on TFP. We assume that the firm-level TFP $\widehat{\omega}_{it} = \log\widehat{TFP}_{it}$ is derived from the OP model (1). The set of explanatory variables includes capacity utilization U_{it} and a complete set of time dummies, or $\lambda_t = 1$ for $t = 2009, \ldots, 2020$. The model equation for the enterprises' TFP is derived from (2) and takes the form of (5). We solve the problem of regressor endogeneity in (5) by applying the system-generalized method of moments (sGMM) (Blundell and Bond, 2000), in which instrumental variables were constructed using the first-difference transformation of the response variable with collapsing (see Kripfganz, 2020).

Table 1 lists the estimation results of Equation (5) which were obtained by using the sGMM estimator for all firms in the manufacturing industry.

Table 1: The TFP equation estimation results

Explanatory variable	$\widehat{\omega}_{it-1}$	u_{it}
Coefficient	0.60***	1.24***

t	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
$\widehat{\lambda_t}$	-3.87	-3.73	-3.74	-3.81	-3.74	-3.78	-3.81	-3.78	-3.79	-3.74	-3.76	-3.64
	***	***	***	***	***	***	***	***	***	***	***	***

*Note: *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.*

The Arellano-Bond tests confirm the lack of autocorrelation of the error term. We verify the set of instrumental variables by performing the Sargan-Hansen test of over-identifying moment restrictions to confirm that the selection of instrumental variables is correct (p-value > 0.01). The influence of capacity utilization is positive and exceeds one; all explanatory variables and time effects were statistically significant (all p-value < 0.01).

3.3. Microdata-based output gap decomposition

We define the company's i capacity utilization gap as the difference between observed individual capacity utilization and the potential level of capacity utilization as measured by the trend component U_{it}^{pot}. Trend components are estimated by the HP filter. Using the HP filter we also calculate the potential levels of labor and capital stock L_{it}^{POT} and K_{it}^{POT}, respectively, for every company i. We then substitute the potential levels of TFP coefficients in Equation (6) and firm-level potentials of labor and capital stock into the production function Equation (4) to obtain the firm-level potential outputs; see Equations (7) to (10). After all, we aggregate the firm-level outcomes in the manufacturing industry S and calculate the microdata-based output gaps using log differences, as in Equations (12) to (15).

Figure 2 presents the levels of microdata-based output gaps: $\hat{x}_{t,S}^{UKL}$(black solid line), $\hat{x}_{t,S}^{L}$ (L gap), $\hat{x}_{t,S}^{K}$ (K gap), and $\hat{x}_{t,S}^{U}$ (U gap) calculated according to Equations (12) to (15). A positive output gap $\hat{x}_{t,S}^{UKL}$ occurred for 2011 and from 2015 to 2019, but the output gap $\hat{x}_{t,S}^{UKL}$ was negative between 2009-2010, 2012-2014 and in 2020. The decline in 2009-2010 was related to the Global Financial Crisis. From 2014 to 2019, Poland experienced a solid economic recovery and expansion, with a dynamic GDP growth of 4-5% per year. During this period, the inflow of EU funds under the new EU budget perspective for 2014-2020 was an important factor supporting the economy. The COVID-19 pandemic has caused the most significant economic crisis. In 2020, Poland's GDP fell by only 2%, significantly deviating from previous growth trends.

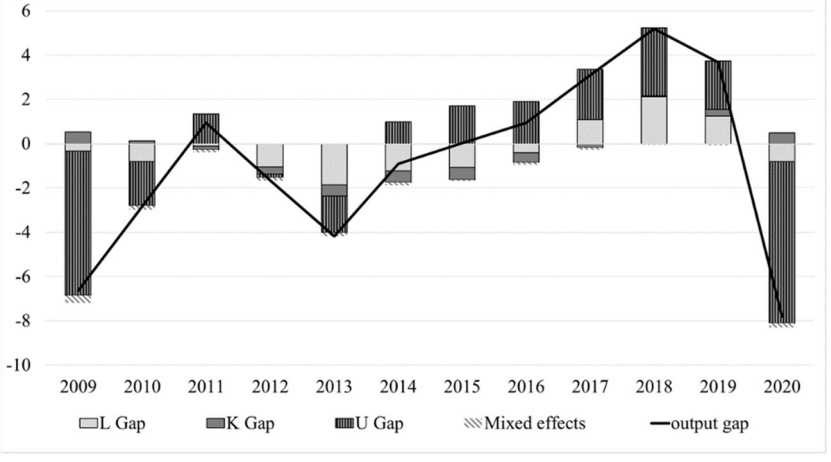


Figure 2: Decomposition of microdata-based output gap estimates in the Polish manufacturing industry (2009–2020)

Note: Calculations based on Equations (12) to (15): $\hat{x}_{t,S}^{UKL}$(microdata-based output gap, black solid line), $\hat{x}_{t,S}^{L}$ (labor-based output gap, light grey bars), $\hat{x}_{t,S}^{K}$ (capital based output gap, dark grey bars), $\hat{x}_{t,S}^{U}$ (capacity-utilization based output gap, vertical striped bars), res_t (mixed effects: slanting striped bars).

Moreover, we consider the potential and theoretical output values to decompose the output gaps from Equation (16) into four components: the labor- , capital-, and capacity utilization-based output gaps; and the interaction term (mixed-effects). Figure 2 also illustrates this decomposition (bars on the chart). In 2012-2014, the labor-based output gap had the largest share in generating the gap in the manufacturing industry's gross value-added, while for the periods of 2009-2011 and 2015-2020 the capacity utilization-based output gap was the most significant component of the output gap.

Understanding the output gap sources can help to adjust policy instruments. For instance, policy measures could prioritize optimizing capacity and investing in productive fixed capital. An example of targeted policy tool is Special Economic Zones, which have proven effective in increasing employment, encouraging capital investment, and improving the utilization of labor and capital.

One primary advantage of using a microdata-based output gap is the possibility of calculating it at different levels of aggregation, such as sectoral, regional, or even more in-depth levels. Figure 3 presents the output gap levels for the two-digit NACE sectors.

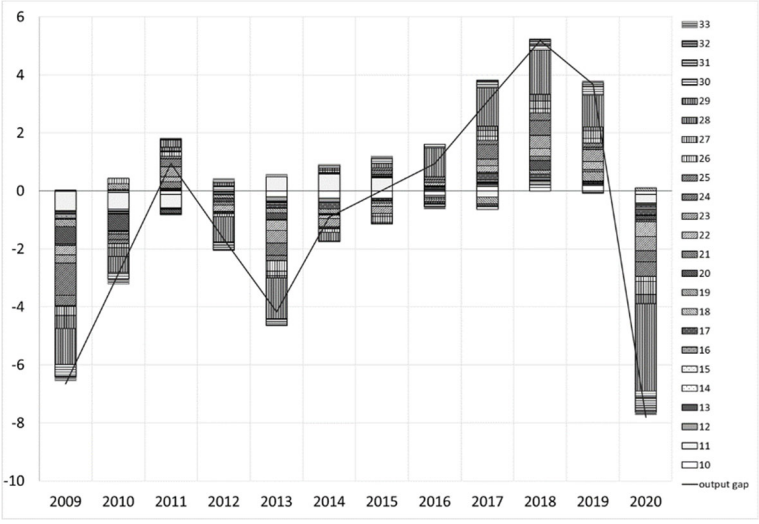


Figure 3: Sectoral output gap decomposition in the Polish manufacturing industry (2009–2020)

Note: The sectoral output gap decomposition is given by Equation (17), where d = 10, 11, …, 33 indicates the NACE divisions.

From Figure 3, we can conclude that the motor vehicle, trailer, and semi-trailer production sector ($d = 29$) — which provides one of the largest contributions to the gross value-added in Poland's manufacturing sector — also has the greatest impact on the global output gap. Further, the sign of its output gap is always the same as the output gap for the entire industry. In contrast, beverage production ($d = 11$) has the largest negative influence on the global output gap. In the period 2009-2019, the sign of the output gap was opposite to global values.

Poland's manufacturing sector analysis reveals the need for policies enhancing resilience and supporting growth in key areas. Export-oriented and high-tech sectors, like motor vehicles and electronics, require investments in R&D and market diversification to reduce vulnerability to global shocks. Consumer-discretionary industries like textiles and furniture would benefit from measures to boost domestic demand and diversify products. Essential sectors like food and chemicals, which showed stability, should be strengthened through supply chain security and process modernization. Investments in green and digital transitions are crucial for sectors like metals and machinery to align with sustainability goals. Additionally, improving supply chain resilience, supporting SMEs, and providing labor market programs will ensure adaptability to economic shifts. Preparing for future crises through fiscal buffers and industrial diversification is critical to safeguarding long-term financial stability.

Figure 4 illustrates the regional decomposition of the output gap in the manufacturing industry for Poland's 16 regions.

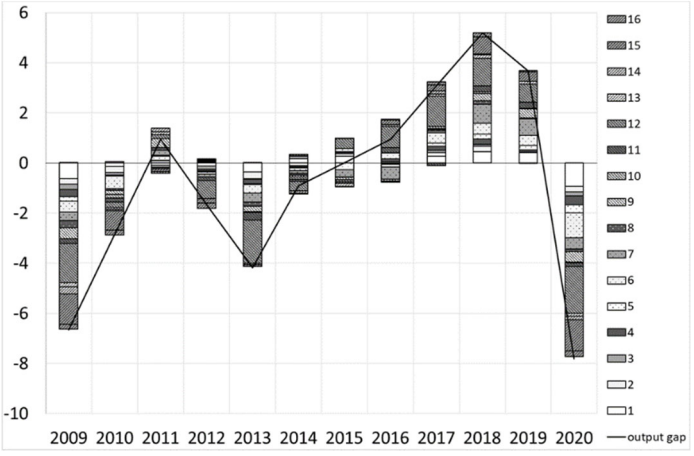


Figure 4: Regional microdata-based output gap decomposition in the Polish manufacturing industry (2009–2020)

Note: The regional output gap decomposition is given by Equation (18); r = 1, 2, …, 16 indicate the regions (alphabetical order of voivodeships).

Several essential implications follow. From 2009 to 2020, Poland had two regions driving business cycle fluctuations: Mazowieckie ($r = 7$) and Śląskie ($r = 12$). Lubuskie ($r = 4$), Małopolskie ($r = 6$) and Śląskie ($r = 12$) are examples of regions in which output gap was always similar to global values.

During the first year of the COVID-19 pandemic (2020), all two-digits sectors in the Polish manufacturing industry experienced negative output gaps, with a maximal value of 21.4% in the leather manufacturing ($d = 15$) industry and -19.1% in the

industry ($d = 29$) that produces motor vehicles, trailers and semi-trailers excluding motorcycles (see Figure 5).

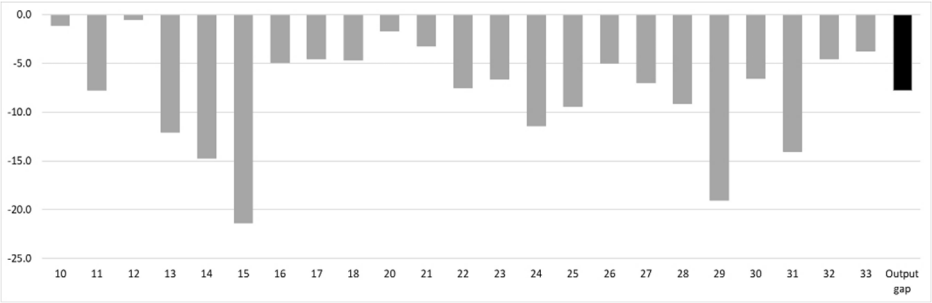


Figure 5: Sectoral microdata-based output gaps during the COVID-19 shock in 2020

Note: The sectoral output gaps in 2020, where d = 10, 11, …, 33 indicates the NACE divisions.

Figure 6 shows that all regional output gaps during the pandemic year (2020) were negative. The slightest deviation of output from its potential level was observed in the Podlaskie region ($d = 10,\ -1.6\%$) and the most significant -14.3% in the Małopolskie (r=6) and Lubuskie regions ($r = 4$).

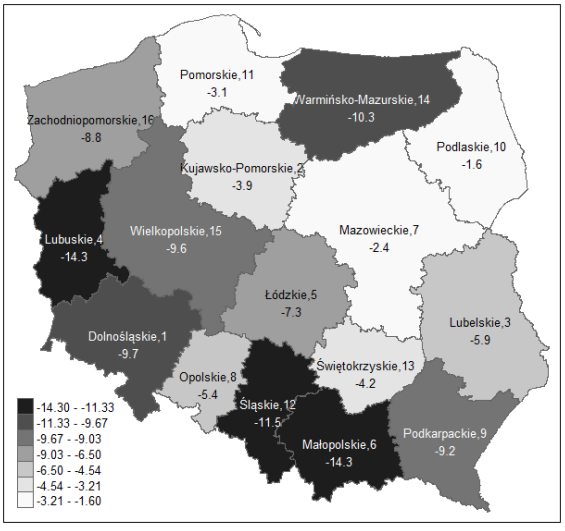


Figure 6: Regional microdata-based output gaps during the COVID-19 shock in 2020

Note: The regional output gaps in 2020; r = 1, 2, …, 16 indicate the regions (alphabetical order of voivodeships).

Figure 3-8 indicate that sectoral and regional output gaps produce different time patterns depending on their intrinsic characteristics. In a given year, we can identify

several sectors and regions below their potential outputs and those in which investments in production capacity are needed to prolong fast sustainable growth. Thus, our analysis helps adjust macroeconomic policy tools to the characteristics of a particular industry or region.

Analyzing regional and sectoral output gaps is vital for identifying economic disparities and resource utilization. These insights help policymakers design targeted interventions for regions with negative gaps through investments in infrastructure, business incentives, and labor programs. Promoting balanced growth by boosting private investment, innovation, and skill development in lagging regions is key to reducing disparities with economically stronger areas. A strategic allocation of resources can strengthen regional and sectoral competitiveness while enhancing the efficient use of production capacities.

3.4. Synchronization and similarity of regional and sectoral output gaps

We apply the synchronicity and similarity measures defined by (26) and (28) to evaluate regional and sectoral output gap coherence in the Polish manufacturing industry. As a reference series, we use the overall microdata-based output gap in the manufacturing industry. During the sample analysis, we found significant differences in the signs and amplitudes of the output gaps across sectors and regions. The COVID-19 pandemic made the signs of the sectoral and regional output gaps in Poland more similar. Still, the highest level of similarity of output gaps in sectors and regions was reported in 2018.

Figure 7 shows strong positive synchronicity and similarity in sectoral output gaps at the end of the sample, but in 2014-2016, the sectoral output gap coherence dropped to zero. The graph in the left panel shows the number of sectors that have a synchronicity measure of one for each year. The highest level of sectoral output gap synchronicity is achieved in the pandemic crisis year 2020, but the highest level (0.58) of sectoral similarity was observed in 2018; in 2020, similarity dropped to 0.46. In 2020, all sectors produced below their potential level.

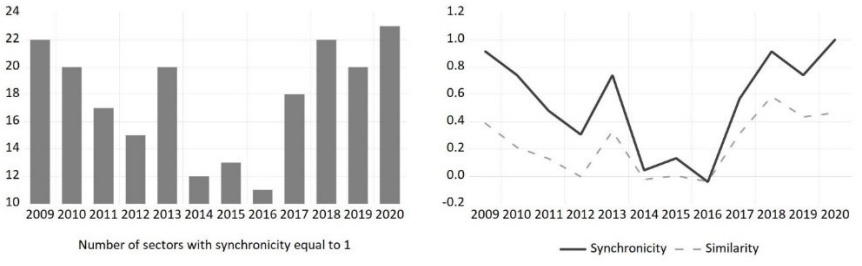


Figure 7: Sectoral output gaps synchronization and similarity in the Polish manufacturing industry

The regional output gaps admit positive synchronicity and similarity with the overall output gap (see Figure 8), but their values change over time. The lowest values were over 2011–2012 and 2015–2016. The graph in the left panel of Figure 8 illustrates the number of regions in each year that have a synchronicity measure of one.

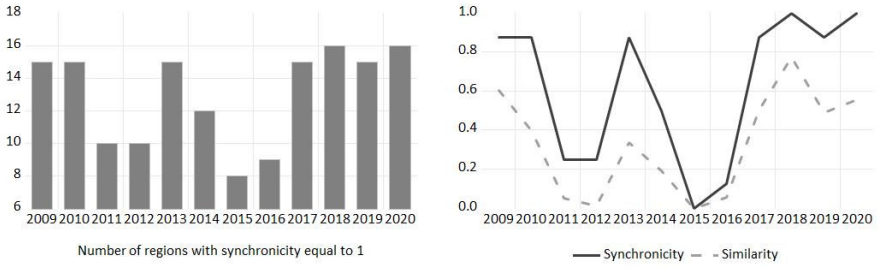


Figure 8: Regional output gaps synchronization and similarity in the Polish manufacturing industry

The number of regions with a synchronicity equal to 1 decreased significantly between 2011–2012 and 2015–2016. The highest regional output gap synchronicity levels are archived in 2018 and 2020. In 2018 all regions are above potential level, whereas in 2020 all regions exhibit a negative output gap. During 2009–2010, 2013 and 2017–2020 most regions are fully synchronized with the reference series. The similarity measure for regions dropped to zero in 2015 and then increased, reaching a maximum value in 2018. Finally, in 2020, the overall similarity dropped to 0.55.

Further, we present the similarity and synchronicity of sectoral microdata-based output gaps separately for each manufacturing industry sector. Figure 9 illustrates that only three out of 23 manufacturing industry sectors are fully synchronized with the microdata-based output gap. This group includes the manufacture of metals ($d = 24$), manufacture of fabricated metal products except machinery and equipment ($d = 25$), and manufacture of motor vehicles, trailers and semi-trailers excluding motorcycles ($d = 29$). Additionally, beverages and machinery repair and installation sectors ($d = 11, d = 33$) experienced lack of output gap synchronization. Figure 10 reveals that only three sectors, wood and wood products ($d = 16$), printing and reproduction of recorded media ($d = 18$), and fabricated metal products ($d = 25$), experienced positive similarity of output gaps. Six sectors, including beverages ($d = 11$), computer, electronic, optical products ($d = 26$), tobacco ($d = 12$), leather ($d = 15$), transport equipment ($d = 30$), and machinery repair and installation ($d = 33$), experienced negative similarity measures over half of sample period.

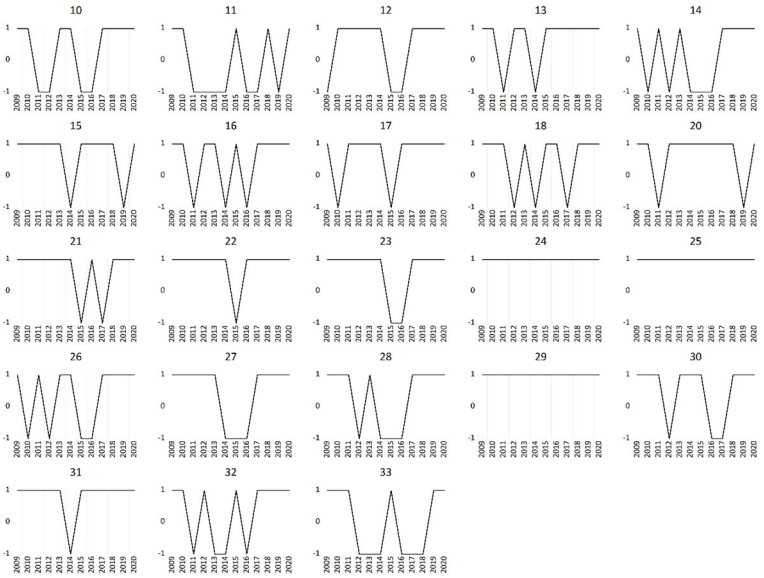


Figure 9: Output gap synchronicity for each sector in the manufacturing industry in Poland

Note: 10, 11, …, 33 are the NACE divisions.

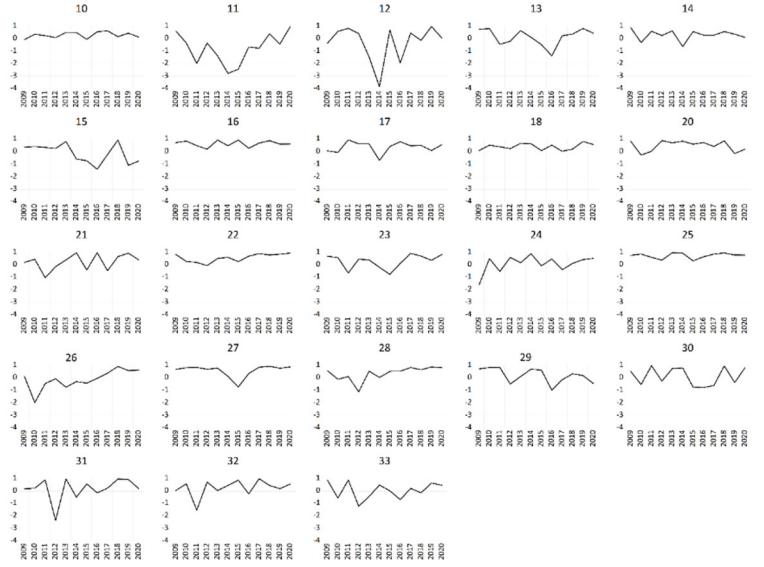


Figure 10: Output gap similarity for each sector in the manufacturing industry in Poland

Note: 10, 11, …, 33 are the NACE divisions.

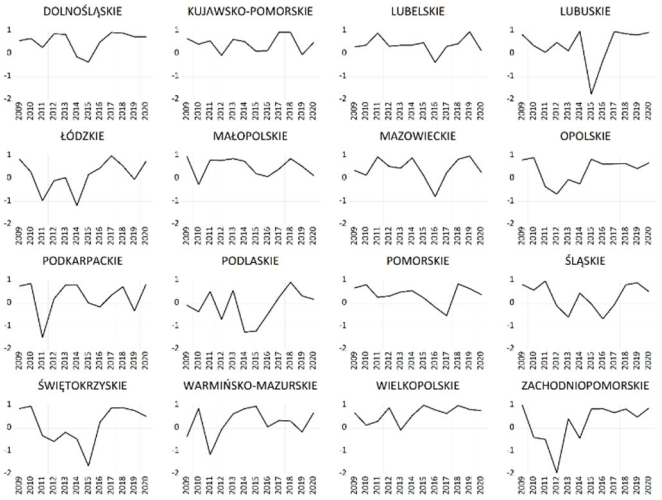


Figure 11 and

Figure **12** illustrate the similarity and synchronicity of regional microdata-based output gaps for each region in Poland. In three regions, Lubuskie ($r = 4$), Małopolskie ($r = 6$), and Śląskie ($r = 12$), regional output gaps were fully synchronized with the microdata-based output gap during the sample period. Moreover, Lubuskie, Małopolskie, Mazowieckie, and Wielkopolskie only during one year in the sample achieved negative similarity. Considering the average similarity measure, the regional output gap in Wielkopolskie is the most similar to the microdata-based output gap in the manufacturing industry. In contrast, the output gap in Podlaskie was most often incoherent with the microdata-based output gap in the manufacturing industry. The similarity and synchronicity measures in Podlaskie take negative values during 5 and 6 years, respectively.

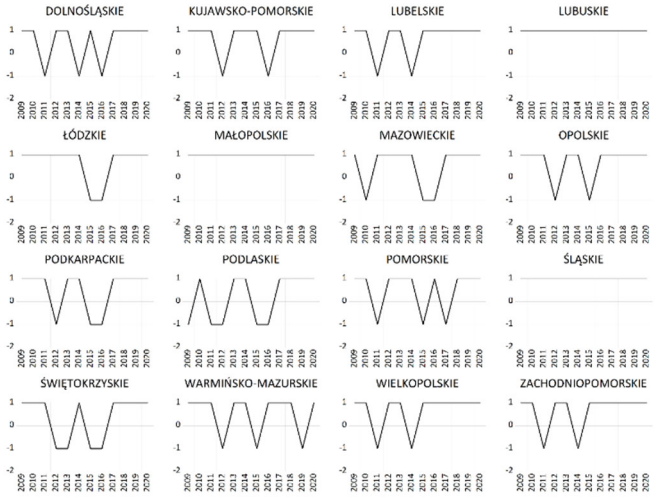


Figure 11: Output gap synchronicity for each region in Poland

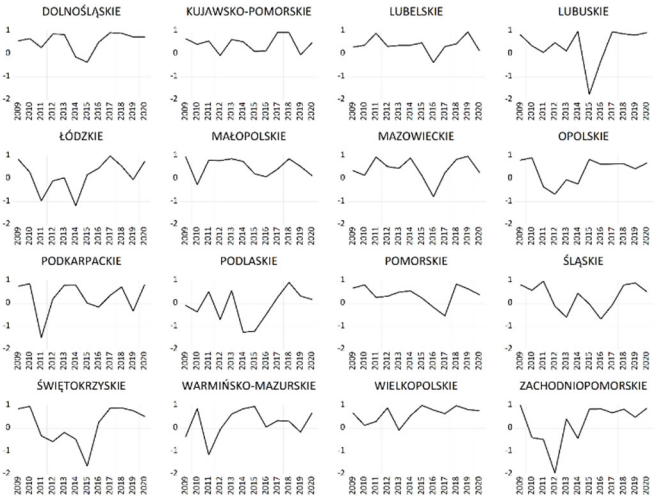


Figure 12: Output gap similarity for each region in Poland

3.5. Robustness analysis

This section summarizes the robustness of similarity and synchronicity of regional and sectoral output gaps in the Polish manufacturing industry. We investigate these coherence measures for different reference time series. In Figure 13, we compare different measures of output gap. In the first part of this exercise, we employ the medians per year of regional and sectoral output gaps as the reference output gap. Both medians are very close to the microdata-based output gap in the manufacturing industry (see Figure 14, Figure 15). Hence, the coherence of regional and sectoral output gaps concerning medians is similar to our baseline scenario. In the second part, we use European Commission experts' output gap estimation (EC output gap, (Havik et al. 2014, Blondeausi et al. 2021) as a reference time series (see Figure 16, Figure 17). In 2009–2010 the EC output gap estimates in Poland are inconsistent with microdata-based estimates. As a result, the regional and sectoral synchronization and similarity measures take negative values. However, in 2011–2020 the synchronization and similarity of regional and sectoral output gaps with respect to EC output gap are very similar to these measures calculated with respect to the microdata-based output gap.

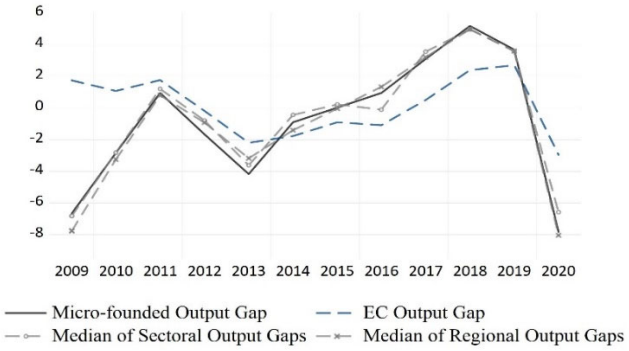


Figure 13: Output gap measures

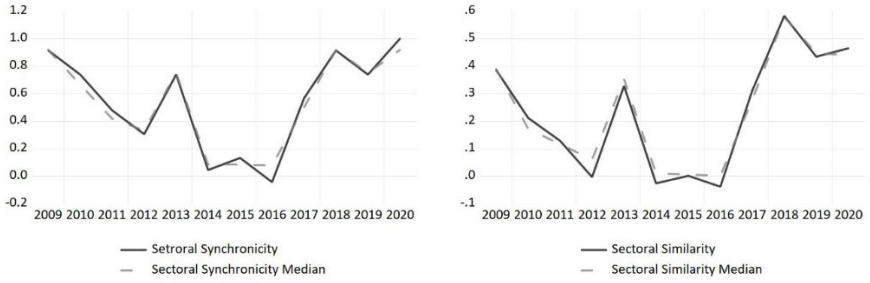


Figure 14: Sectoral similarity and synchronicity measures with respect to the microdata-based output gap and median of sectoral output gaps

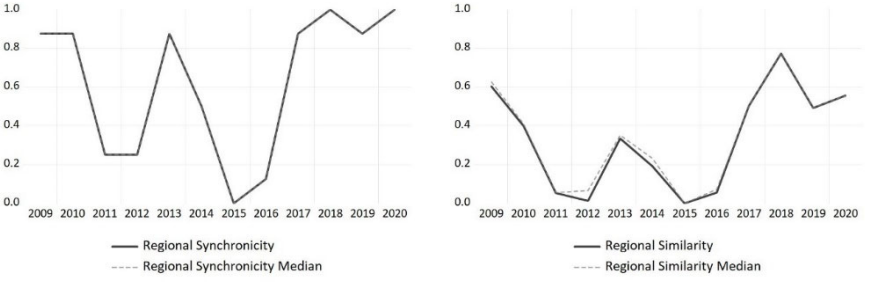


Figure 15. Regional similarity and synchronicity measures with respect to the microdata-based output gap and median of regional output gaps

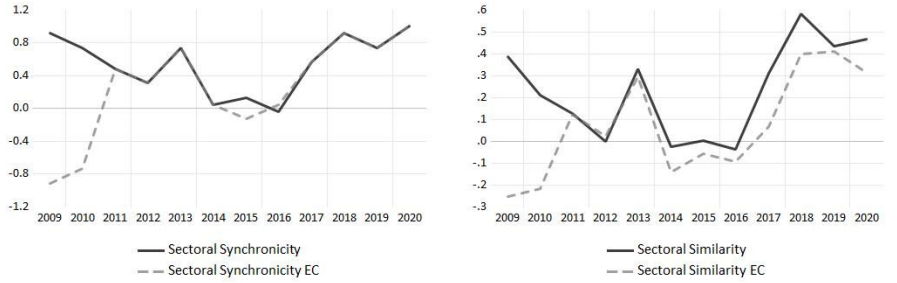


Figure 16: Sectoral similarity and synchronicity measures with respect to the microdata-based output gap and European Commission output gap

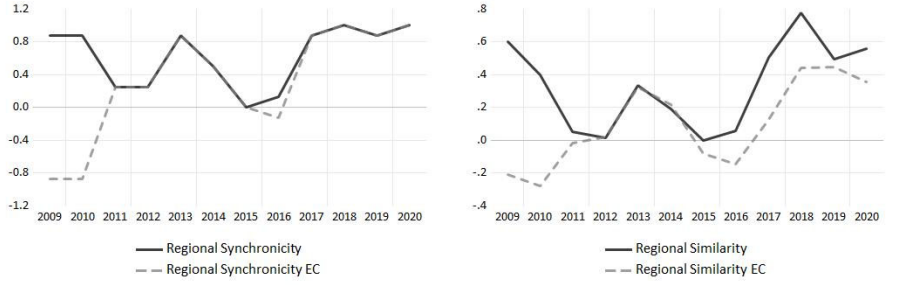


Figure 17. Regional similarity and synchronicity measures with respect to the microdata-based output gap and European Commission output gap

4. Discussion and Conclusions

This study proposes a new methodology for estimating regional and sectoral output gaps based on a firm-level approach. Our study responds to the need for a uniform method to estimate and assess levels of unobservable potential production. We used a two-stage micro-econometric procedure to estimate the output gap based on business tendency surveys and annual reports on business activity. In the first stage, we estimated the production function to determine the enterprise's individual TFP by applying the Olley-Pakes model. In the second stage, we evaluated the influence of capacity utilization on firm-level TFP. Our results demonstrate that capacity utilization positively impacts the TFP cycle. Based on these estimates, we proposed novel formulas for firm-level potential outputs. Finally, we aggregated these potential values to derive several output gap-related measures for the sectors and regions in the manufacturing sector.

This study offers a new method relevant from both policy and firm perspectives and can be applied to different national contexts to examine output gaps in other economies worldwide. Moreover, our firm-level approach decomposes the output gap

into three main components: capital-, labor-, and capacity utilization-based output gaps. We recognized that the latter two components dominate the total output gap in the Polish manufacturing sector. Regional and sectoral output gap decompositions also illuminate the economy's most substantial regions and sectors in terms of business cycle fluctuations.

During the sample analysis, we found significant differences in the signs and amplitudes of the output gaps across sectors and regions. The COVID-19 pandemic made the signs of the sectoral and regional output gaps in Poland more similar. However, despite the same signs of output gaps during the pandemic, their substantial heterogeneity was observed. Still, the highest level of similarity of output gaps in sectors and regions was reported in 2018. Subsequently, we identified the sectors and regions below their potential outputs in a given year and those in which production factors generated above-average value added. Thus, our analysis may help policymakers adjust macroeconomic policy tools to the characteristics of a particular industry or region.

References

Ackerberg, D. A., Caves K. and Frazer G., (2015). Identification Properties of Recent Production Function Estimators. *Econometrica,* 83(6), pp. 2411–51. https://doi.org/10.3982/ECTA13408.

Álvarez, L. J., Gómez-Loscos, A., (2018). A Menu on Output Gap Estimation Methods. *Journal of Policy Modeling*, 40(4), pp. 827–50. https://doi.org/10.1016/j.jpolmod.2017.03.008.

Blagrave, P., Garcia-Saltos, R., Laxton D. and Zhang F., (2015). A Simple Multivariate Filter for Estimating Potential Output. *IMF Working Papers*, 15(79), pp. 1. https://doi.org/10.5089/9781475565133.001.

Blażej, M., Górajski, M. and Ulrichs, M., (2025). Microdata-based Output Gap Estimation Using Business Tendency Surveys. *Journal of Economic Dynamics and Control*, 174, pp. 1–18. https://doi.org/10.1016/j.jedc.2025.105068.

Blondeau, F., Planas, C. and Rossi A., (2021). *Output Gap Estimation Using the European Union's Commonly Agreed Methodology: Vade Mecum & Manual for EUCAM Sotfware*, vol. 148. Luxembourg: European Commission's Directorate-General for Economic and Financial Affairs. https://doi.org/10.2765/217592.

Blundell, R., Bond, S., (2000). GMM Estimation with Persistent Panel Data: An Application to Production Functions. *Econometric Reviews*, 19(3), pp. 321–40. https://doi.org/10.1080/07474930008800475.

Chalaux, T. Guillemette, Y., (2019). The OECD Potential Output Estimation. *OECD Working Papers*, ECO/WKP/32, pp. 1–28.

De Haan, J., Jacobs, Jan P.A.M. J and Zijm, R. (2024). Coherence of Output Gaps in the Euro Area: The Impact of the COVID-19 Shock. *European Journal of Political Economy*, no. 84. https://doi.org/10.1016/j.ejpoleco.2023.102369.

Edge, R., M, Rudd, J., B., (2016). Real-Time Properties of the Federal Reserve's Output Gap. *Review of Economics and Statistics*, 98(4), pp. 785–91. https://doi.org/10.1162/REST_a_00555.

Fernald, J. G., (2012). A Quarterly, Utilisation-Adjusted Series on Total Factor Productivity. *Federal Reserve Bank of San Francisco, Working Paper Series* 2012–19, pp. 01–28. https://doi.org/10.24148/wp2012-19.

Gali, J., (2009). *Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework*. Princeton University Press.

Gosińska, E., Górajski, M. and Ulrichs, M., (2024). Micro-firms' productivity growth in Poland before and during COVID-19: Do industry and region matter? Entrepreneurial Business and Economics Review, 12(2), pp. 177–200.

Górajski M., Błażej M., (2020). A control function approach to measuring the total factor productivity of enterprises in Poland Bank i Kredyt, 51, pp. 293–316

Gradzewicz, M., J. Growiec, M. Kolasa, Ł. Postek and P. Strzelecki, (2017a). Poland's Uninterrupted Growth Performance: New Growth Accounting Evidence. *Post-Communist Economies*, 30(2), pp. 238–72. https://doi.org/10.1080/14631377.2017.1398519.

Graff, M., Strum, J.-E., (2010). KOF Working Papers. *KOF Working Papers*, no. 269, pp. 2–32. https://doi.org/10.3929/ethz-a-006070989.

Greenwood, J., Zvi, H. and Huffman G. W., (1988). American Economic Association Investment, Capacity Utilisation, and the Real Business Cycle. *American Economic Review,* 78(3), pp. 402–17.

Havik, K., Mc Morrow, K., Orlandi, F., Planas, C., Raciborski, R., Röger, W., Rossi, A., Thum-Thysen, A. and Vandermeulen, V., (2014). *The Production Function Methodology for Calculating Potential Growth Rates & Output Gaps*, Vol. 3187. https://doi.org/10.2765/71437.

Hagemejer, J., Kolasa, M. (2011). Internationalisation and economic performance of enterprises: Evidence from polish firm-level data. *World Economy*, 34(1), pp. 74–100. https://doi.org/10.1111/j.1467-9701.2010.01294.x

Jokubaitis, S., Celov, D., (2023). *Business Cycle Synchronization in the EU: A Regional-Sectoral Look through Soft-Clustering and Wavelet Decomposition*. *Journal of Business Cycle Research*. Springer International Publishing. https://doi.org/10.1007/s41549-023-00090-4.

Kripfganz, S., (2020). Generalized Method of Moments Estimation of Linear Dynamic Panel-Data Models.

Levinsohn, J., Petrin, A., (2003). Estimating Production Functions Using Inputs to Control for Unobservables. *Review of Economic Studies,* 70(2), pp. 317–41. https://doi.org/10.1111/1467-937X.00246.

Mink, M., Jacobs, Jan P.A.M. and De Haan, J., (2012). Measuring Coherence of Output Gaps with an Application to the Euro Area. *Oxford Economic Papers*, 64(2), pp. 217–36. https://doi.org/10.1093/oep/gpr049.

Ódor, L., Kucserová, J. J., (2014). Finding Yeti: More Robust Estimates of Output Gap in Slovakia. *Council for Budget Responsibility Working Paper* 2.

Olley, G S., Pakes, A., (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica*, 64(6), p. 263. https://doi.org/10.2307/2171831.

Orphanides, A., Van Norden, S., (2002). The Unreliability of Output-Gap Estimates in Real Time. *The Review of Economics and Statistics*, 84(4), pp. 569–83.

Planas, C., Roeger, W. and Rossi, A., (2013). The information content of capacity utilisation for detrending total factor productivity. *Journal of Economic Dynamics and Control*, 37(3), pp. 577–590.

Peykov, N., (2021). Sectoral Output Gaps – Estimates for Bulgaria. *Economic Alternatives*, no. 1, pp. 5–26. https://doi.org/10.37075/EA.2021.1.01.

Phillips, P. C. B., Shi, Z., (2021). Boosting: Why You Can Use the HP Filter. *International Economic Review*, 62(2), pp. 521–70. https://doi.org/10.1111/ iere.12495.

Pu, Z., Fan, X., Xu, Z. and Skare, M., (2023). A Systematic Literature Review on Business Cycle Approaches: Measurement, Nature, Duration. *Oeconomia Copernicana*, 14(3), pp. 935–76. https://doi.org/10.24136/oc.2023.028.

Quast, J., Wolters, M. H., (2020). Reliable Real-Time Output Gap Estimates Based on a Modified Hamilton Filter. *Journal of Business and Economic Statistics*, 0(0), pp. 1–34. https://doi.org/10.1080/07350015.2020.1784747.

Statistics Poland, (2019). Methodological Report. Non-financial Enterprises Survey, Warsaw.

Statistics Poland, (2023). Methodological Handbook. Business Tendency Survey, Warsaw.

Van Beveren, I., (2012). Total Factor Productivity Estimation: A Practical Review. *Journal of Economic Surveys*, 26(1), pp. 98–128. https://doi.org/10.1111/j.1467-6419.2010.00631.x.

Walsh, Carl E., (2010). *Monetary Theory and Policy. The MIT Press Cambridge Massachusetts. Third Dition*. Cambridge: The MIT Press. https://doi.org/10.1007/s11293-007-9065-y.

Wen, Y., (1998). Capacity Utilisation under Increasing Returns to Scale. *Journal of Economic Theory*, 81(1), pp. 7–36. https://doi.org/10.1006/jeth.1998.2412.

Woodford, M., (2003). *Interest and Prices. Interest and Prices*. Princeton: Princeton University Press.

Appendix

Supplementary materials to:

Synchronization and Similarity of Regional and Sectoral Output Gaps in the Polish Manufacturing Industry

Annex D. Time and regional effects on TFP.

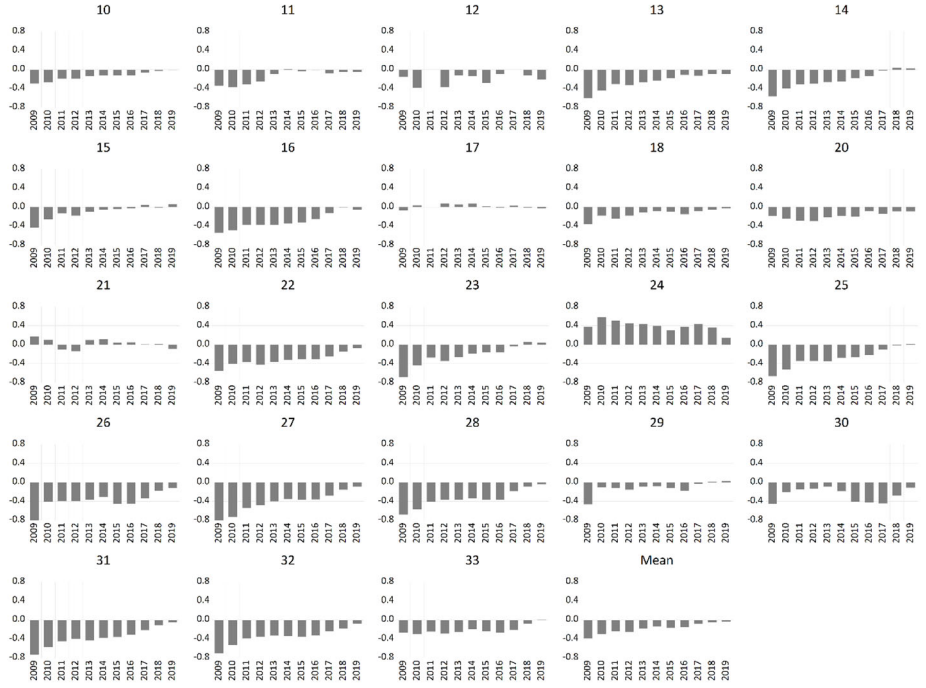


Figure 18: Production function estimation: time effects on TFP by division

Note: 10, 11, …, 33 are the NACE divisions; the baseline year is 2020.

Figure 18 illustrates the time effects on TFP over 2009–2019 for various NACE divisions in the manufacturing sector relative to 2020. Most divisions (except 24 – pharmaceutical products and 21 – metals) exhibit negative TFP deviations in the earlier years, indicating lower productivity compared to 2020, followed by a general upward trend toward convergence with the 2020 level. Some divisions demonstrate steady progress, while others display irregular patterns, highlighting sector-specific dynamics.

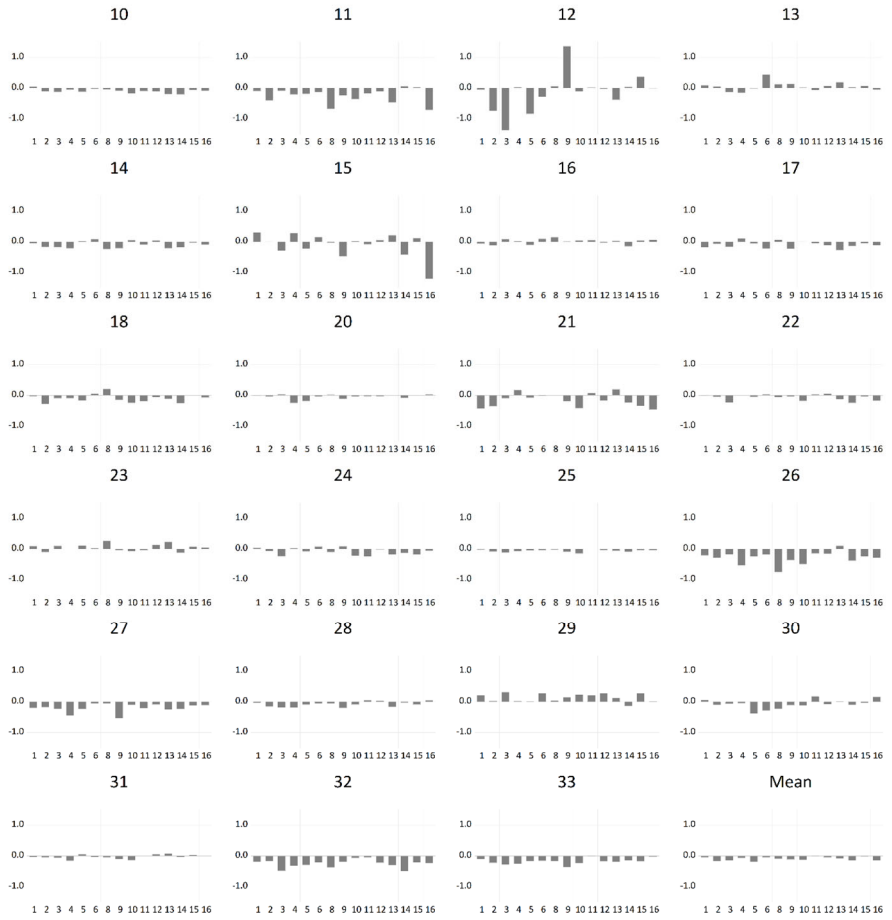


Figure 19: Production function estimation: regional effects on TFP by division

Note: 10, 11, …, 33 are the NACE divisions; 1, …, 16 denote the regions; the baseline voivodeship is 7 (Mazowieckie region).

Figure 19 shows the deviations in total factor productivity (TFP) for different manufacturing divisions in Poland's regions compared to Mazowieckie. Significant variability is evident across regions and divisions, with some divisions experiencing consistent negative deviations while others show mixed patterns. This highlights regional disparities in productivity within Poland's manufacturing sector. It can be observed that when analyzing deviations in all divisions, most of the regions have a lower TFP than Mazowieckie. The manufacture of motor vehicles (29) is an example of a sector that has higher TFP in almost all regions (except Warmińsko-Mazurskie (14)) than Mazowieckie.

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 113–128, https://doi.org/10.59139/stattrans-2025-018
Received – 24.11.2023; accepted – 23.12.2024

Clustering based on poverty indicator data using K-Means cluster with Density-Based Spatial Clustering of Application with Noise

Sapriadi Rasyid[1], **Siswanto Siswanto**[2], **Sitti Sahriman**[3]

Abstract

The Indonesian government has implemented poverty alleviation programs, including assistance programs for the poor. Despite these efforts, the number of impoverished individuals in South Sulawesi continues to rise. To address this issue, a statistical method is necessary to cluster the poor based on error indicators for each region, serving as a reference for providing assistance. The appropriate statistical method is cluster analysis by minimizing object differences within one cluster and maximizing object differences between clusters. This study employs two methods, namely K-Means and Density-Based Spatial Clustering of Application with Noise (DBSCAN), to compare their effectiveness based on the Silhouette Coefficient. The data used for the analysis included eight poverty indicators for the South Sulawesi province in 2022. The K-Means method yielded two optimal clusters, with cluster 1 comprised of 23 regencies and cities, and cluster 2 only of Makassar City. The results of further analysis on cluster 1 consisted of eight new clusters and produced a Silhouette Coefficient of 0.507. In contrast, the DBSCAN method yielded one cluster, that encompassed 23 regencies and cities, with Makassar City identified as noise. The results of the further analysis on the clusters consisted of one cluster with three noises and produced a Silhouette Coefficient of 0.318. The study concludes that K-Means provides a higher Silhouette Coefficient and a more accurate representation of poverty clusters in South Sulawesi, which renders it a more effective tool for targeted poverty alleviation efforts.

Key words: Cluster, DBSCAN, poverty, K-Means, Silhouette Coefficient.

1. Introduction

Cluster analysis is a statistical method that can be used to cluster several objects into a class. Cluster analysis aims to maximize the similarity among objects within

[1] Department of Statistics, Hasanuddin University, Indonesia. E-mail: sapriadirasyid@gmail.com. ORCID: https://orcid.org/0000-0006-2972-7125.

[2] Corresponding author. Department of Statistics, Hasanuddin University, Indonesia. E-mail: siswanto@unhas.ac.id. ORCID: https://orcid.org/0000-0003-1934-5343.

[3] Department of Statistics, Hasanuddin University, Indonesia. E-mail: sittisahriman@unhas.ac.id. ORCID: https://orcid.org/0000-0002-9614-7132.

a cluster while minimizing the similarity between objects in different clusters (Pramana et al., 2018). However, cluster analysis assumes no multicollinearity (Hair et al., 2010). To address multicollinearity, principal component analysis (PCA) is necessary to reduce data dimensions into mutually independent principal components (Jhonson & Wichern, 2018). Thus, a combination of PCA is required for more optimal clustering results (Granato et al., 2018). Cluster analysis consists of various algorithms, two of which are K-Means and density-based spatial clustering of application with noise (DBSCAN).

K-Means is a method which needs an appropriate number of clusters denoted as k. K-Means is susceptible to noise (Huang et al., 2023). In contrast, DBSCAN is a method which clusters data based on distance density, thus identifying noise (Jing et al., 2010). Density, in this context, refers to the quantity of points found within a designated radius (Pu et al., 2021). But, DBSCAN cannot determine the number of clusters. By using the same data, the procedure and results of K-Means with DBSCAN will be different. According to Dewi & Pramita (2019), the quality of cluster analysis can be measured using the Silhouette Coefficient. Therefore, the Silhouette Coefficient test can be a reference for comparing of K-Means and DBSCAN result. This method can help the government in identifying poverty indicators in each region. This is important because the number of impoverished individuals in South Sulawesi increased by 16.86 thousand people in September 2022 compared to the previous year (BPS, 2023). By using the appropriate clustering method, the government can provide more effective and targeted assistance based on relevant poverty indicators.

Research on K-Means with the DBSCAN method has been carried out by many researchers, such as research conducted by Rais et al. (2021), which optimized K-Means clustering with PCA, resulting in two principal components and two clusters with a small Davies-Bouldin Index indicating good clustering results. Meanwhile, research on DBSCAN was performed using K-Nearest Neighbor to determine the epsilon parameter. This study used four variables and the result obtained 5 noises and produced one cluster consisting of 19 object (Akbar et al., 2021). What distinguishes this research from previous studies is that this research combines PCA to obtain comparison results of K-Means with DBSCAN using the Silhouette Coefficient on poverty indicator data of South Sulawesi Province. Besides that, determining epsilon in DBSCAN is done based on hierarchy principles. Based on this case, this study is focused on obtaining clustering results based on poverty indicators data for each region as a reference providing assistance. Results of this

research can, as hoped, assist the government in addressing poverty cases in each region in South Sulawesi.

2. Methodology

2.1. Data and Research Variables

This research is quantitative in nature and covers 24 regencies and cities in the South Sulawesi Province, using secondary data sourced from the Central Statistics Agency (Badan Pusat Statistik) of South Sulawesi in 2022, which is available on the official website *sulsel.bps.go.id*. The research variables consist of eight poverty indicators following Rais et al. (2021), including the human development index (X_1), population size (X_2), labor force (X_3), percentage of poor population (X_4), labor force participation (X_5), unemployment rate (X_6), population density (X_7), and per capita expenditure (X_8). These variables were selected because they are commonly used in poverty research and are considered to be significant factors in determining the socio-economic conditions of a region.

2.2. Cluster Analysis

Cluster analysis is an effort to identify groups of similarity of data objects within one cluster while minimizing similarity to other clusters. In general, clustering algorithms can be divided into different categories, such as partitional, hierarchical, and density-based (Cardeiro de Amorim & Makarenkov, 2023). Furthermore, cluster analysis methods require a measure of dissimilarity or distance. Typically, the distance often used is the Euclidean distance as defined in Equation (1) below.

$$\mathbf{d_{p,q}} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{m}} \left(\mathbf{x_{ip}} - \mathbf{x_{iq}}\right)^2} \tag{1}$$

with: $d_{p,q}$ is the distance between object p and q, x_{ip} is the i variable of object p, x_{iq} is the i variable of q , and m is the number of variables.

2.3. Assumption of Multicollinearity

Cluster analysis generally assumes that the data to be analyzed do not exhibit strong correlations between two or more variables with other variables. This is because strong correlations can lead to multicollinearity (Hair et al., 2010). In regression analysis, the Variance Inflation Factor (VIF) is used to identify the presence of multicollinearity among independent variables. Generally, multicollinearity is considered significant when VIF > 10, indicating a strong linear correlation between variables in the model

(Salmerón et al., 2020). According to Hair et al., (2010), one way to identify multicollinearity is by calculating the Variance Inflation Factor (VIF) value, which is formulated based on Equation (2) as follows:

$$\mathbf{VIF_i} = \frac{\mathbf{1}}{\mathbf{Tolerance}} = \frac{\mathbf{1}}{\mathbf{1 - R_i^2}} \tag{2}$$

with: R_i^2 is coefficient of determination from regressing variable i with others.

2.4. Principal Component Analysis

Principal Component Analysis (PCA) is an effective statistical technique for addressing multicollinearity by reducing data dimensions and the elimination of highly correlated variable, because PCA is useful for extracting data to be new variables (Kurita, 2019). These variables are called Principal Components (PC) (Festa et al., 2023). PCA can be employed for data preprocessing before applying complex statistical methods (Kherif & Latypova, 2019). In PCA, PC are new variables obtained through a linear combination of the original variables. Before transforming the data into *Principal Components*, standardization is necessary to ensure that all variables have a uniform scale. Suppose there is a set of original variables denoted as $X = (X_1, X_2, \dots, X_m)$, the standardization process is then applied to transform these variables into standardized variables $Z = (Z_1, Z_2, \dots, Z_m)$. This standardization aims to eliminate differences in scale among variables, ensuring that the analysis remains accurate and is not affected by variations in measurement units. Meanwhile, the values of Z are obtained according to Equation (3) below (González et al., 2022).

$$\mathbf{Z_i} = \frac{\mathbf{X_i - \mu_i}}{\sqrt{\boldsymbol{\sigma_i^2}}} \tag{3}$$

with: X_i is the value of variable i on the object, μ_i is the mean value of variable i, σ_i^2 is the variance of variable i, and Z_i is the standardized value of i (Bari & Kindzierski, 2018). Therefore, Equation (4) can be obtained as follows:

$$\mathbf{PC_f} = \mathbf{b_{11}Z_1} + \cdots + \mathbf{b_{mg}Z_m} \tag{4}$$

with: PC_f is the f^{th} principal component, b_{mg} is the eigenvalue of m on the g^{th} principal component. Meanwhile, the eigenvalue can be obtained using the following Equation (5):

$$\mathbf{|A - \lambda I|} = \mathbf{0} \tag{5}$$

Equation (5) yields characteristic roots λ_l such that $\lambda_1 > \lambda_2 > \cdots > \lambda_v$, so each characteristic λ_l depends on the value of $\mathbf{b}$ (Astutik et al., 2018). To determine the

principal components, one can follow the criteria by weighting the cumulative proportions as described in Equation (6) below (Abdi & Williams, 2010):

$$\frac{\sum_{l=1}^{u} \boldsymbol{\lambda_l}}{\sum_{l=1}^{v} \boldsymbol{\lambda_l}} > \mathbf{0.80}\ ;\ \mathbf{u} \leq \mathbf{v} \tag{6}$$

with: $\sum_{l=1}^{u} \lambda_l$ is total variance of the first u principal component and $\sum_{l=1}^{v} \lambda_l$ is total variance.

2.5. K-Means

K-Means is a type of the non-hierarchical cluster analysis techniques as it is susceptible to the selection of the initial clustering center (Liu et al., 2023). In simple terms, the K-Means algorithm can be performed as follows (Huang et al., 2023):

1. Determine *k* as the number of clusters to be formed.
2. Randomly initialize *k* parameters, which are the initial centroids of the clusters.
3. Calculate the distance of each data point to each selected centroid using Equation (1). Each data point chooses the nearest centroid.
4. Calculate the mean value of the data points that chose the same centroid. This value becomes the new centroid.
5. Repeat step 3 and 4 if the position of the new centroid and the old centroid is not the same. However, if the positions of the new and old centroids are the same, the clustering process is considered complete.

2.6. Density-Based Spatial Clustering of Application with Noise

Density-Based Spatial Clustering of Application with Noise (DBSCAN) is often considered the most popular density-based clustering algorithm (Chowdhury et al., 2023). DBSCAN is a clustering method based on the concept of density, which can be determined by a single density condition. The following are terms in DBSCAN: Epsilon is the DBSCAN radius used to determine density connectivity. As core points, the number of points within their neighborhood must be greater than or equal to the minimum points (Jing et al., 2019). Point p is density-reachable from a point q if there is a chain of points $p_1, p_2, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{1+1} is direct density-reachable from p_1 (Zhang et al., 2022).

In simple terms, the DBSCAN algorithm requires two parameters: epsilon and minimum points (Starczewski & Cader, 2019). When determining the minimum points, according to Hahsler et al. (2019), it is typically at least the number of variables in the analyzed dataset plus one. This approach aims to adjust the parameter according to the complexity of the data dimensions to ensure that each formed cluster has a sufficiently significant density. However, DBSCAN has evolved by creating

a representation without requiring an epsilon radius by incorporating hierarchical clustering processes within the density concept (Stewart & Al-Khassaweneh, 2022). The DBSCAN algorithm is as follows (Nurhaliza & Mustakim, 2021):

1. Initialize the minimum points parameter.
2. Choose a random starting point, p.
3. Calculate all point distances using Equation (1) for density reachability with respect to p. If a point satisfies the core point condition, the number of points within its neighborhood is equal to or greater than the minimum points parameter, it forms a cluster. If p is a border point and no points are density-reachable from p, move to the next point.
4. Repeat 2-4 step for each observed point until all objects are identified.

2.7. Silhouette Coefficient

The *Silhouette Coefficient* (SC) is a widely used metric for evaluating clustering performance by assessing the compactness and separation of clusters. It measures how well each data point fits within its assigned cluster compared to other clusters (Řezanková, 2019). Moreover, SC is used to evaluate the clustering results as per the following Equation (7):

$$\mathbf{SC} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{p=1}}^{\mathbf{n}} \frac{\mathbf{b(p)} - \mathbf{a(p)}}{\mathbf{max(a(p); b(p))}} \tag{7}$$

The *Silhouette Coefficient* (SC) is a widely used metric for evaluating clustering performance by assessing the compactness and separation of clusters. It measures how well each data point fits within its assigned cluster compared to other clusters

with: $a(p)$ is the average distance of object p's characteristics to all objects within the same cluster, and $b(p)$ is the average distance of object p's characteristics to all objects within a different cluster (Batool & Hening, 2021).

SC has an interval range of $-1 \leq SC \leq 1$, and the evaluation criteria for the SC method can be shown in Table 1 as follows (Batool & Hening, 2021):

Table 1: Evaluation Criteria for SC Method

Interval SC Score	Interpretation	Interval SC Score	Interpretation
$0.70 < SC \leq 1.00$	Strong Structure	$0.25 < SC \leq 0.50$	Weak Structure
$0.50 < SC \leq 0.70$	Medium Structure	$SC \leq 0.25$	No Structure

Source: Batool & Hening, 2021.

3. Result and Discussion

3.1. Multicollinearity Test

Cluster analysis must satisfy the assumption of non-multicollinearity to ensure that the weights of each variable are balanced. Therefore, VIF calculations are performed based on Equation (2), which can be presented in Table 2 as follows:

Table 2: VIF Score Each Variable

Variable	VIF	Description	Variable	VIF	Description
X_1	4.152	Not Significant	X_5	3.031	Not Significant
X_2	128.993	Significant	X_6	34.406	Significant
X_3	130.826	Significant	X_7	19.184	Significant
X_4	2.202	Not Significant	X_8	4.097	Not Significant

Source: data processed.

Based on Table 2 show the VIF values indicate that for variables $X_2 = 128.993 > 10$, $X_3 = 130.826 > 10$, $X_6 = 34.406 > 10$, and $X_7 = 19.184 > 10$, and the general criterion that multicollinearity is considered significant when the VIF > 10. Therefore, it is concluded that there is multicollinearity in the data. The presence of multicollinearity leads to imbalanced weights in the analysis results, making the presented information inaccurate, such as in the calculation of distances between objects. The solution to this problem is to use PCA.

3.2. Principal Component Analysis

The data on poverty indicators in South Sulawesi Province have different units of measurement. Therefore, to determine the Principal Components (PC), a correlation matrix is used. Additionally, differences in units of measurement can result in inconsistent cluster analysis results. The solution to this problem is to transform the data into the same units of measurement using the PCA according to Equation (3). The number of PC formed is based on the cumulative diversity proportion of the PC variables, which should be at least around 80%. The calculation of the cumulative diversity proportion is calculated according to Equation (6), and the results of the cumulative diversity proportion calculation are shown in Table 3 as follows:

Table 3: Cumulative Diversity Proportion Each PC

Principal Component	λ	Cumulative Diversity Proportion	Principal Component	λ	Cumulative Diversity Proportion
PC_1	4.778	59.723%	PC_5	0.191	97.906%
PC_2	**1.686**	**80.797%**	PC_6	0.144	99.710%
PC_3	0.767	90.383%	PC_7	0.019	99.952%
PC_4	0.411	95.519%	PC_8	0.004	100.00%

Source: data processed.

The cumulative diversity proportion in Table 3 show value greater than 80% indicates that these two principal components capture most of the information contained in the original variables, allowing the analysis to proceed with just these two components without significant loss of information. Hence, it can be concluded that PC_1 and PC_2 meet the criteria for forming two principal components that explain a total variance of 80.797% in the original variables. The principal components are as follows according to Equation (4):

$$\mathbf{PC_1 = 0.349Z_1 - 0.374Z_2 - 0.368Z_3 - 0.255Z_4 - 0.203Z_5 + 0.431Z_6 + 0.427Z_7 + 0.359Z_8} \quad (8)$$

$$\mathbf{PC_2 = 0.284Z_1 - 0.377Z_2 - 0.407Z_3 - 0.461Z_4 - 0.502Z_5 + 0.207Z_6 + 0.110Z_7 - 0.301Z_8} \quad (9)$$

Equation (8) is formed from an eigenvalue of 4.778, while Equation (9) is formed from an eigenvalue of 1.686. So, Equations (8) and (9) become the new variables that will be further analyzed using both K-Means cluster and DBSCAN. The results of this research can provide information that K-Means clustering and DBSCAN can be combined with PCA to create new mutually independent variables called PC.

3.3. K-Means

The initial step in K-Means is to determine the optimal k value or the number of clusters, as shown in Table 4 below:

Table 4: SC Score Each k

k	SC Score	k	SC Score	k	SC Score	k	SC Score
1	0.000	4	0.400	7	0.279	10	0.459
2	0.772	5	0.382	8	0.411	11	0.431
3	0.436	6	0.426	9	0.483	12	0.313

Source: data processed.

Based on Table 4, the best SC is obtained when $k = 2$. The next step is to determine the initial centroids randomly. Given that the selection of centroids is random, this study adopts an approach by selecting the minimum and maximum values of each Principal Component (PC). Specifically, the first centroid is determined based on the minimum value of each PC, while the second centroid is determined based on the maximum value of each PC, as shown in Table 5 below:

Table 5: Initial Centroid

k	o_1	o_2
1	-1.958	-2.486
2	9.439	1.990

Source: data processed.

The next step is to calculate the distance between objects based on Equation (1) for each data point based on the selected initial centroids. Each data point selects the nearest centroid. After that, the new centroids are determined by calculating the average of the data points that selected the same centroid, as shown in Table 6 below:

Table 6: Centroid of the second iteration

k	o_1	o_2
1	-0.410	-0.060
2	9.439	1.371

Source: data processed.

Based on Table 6, it is found that the closest distance values have not changed, indicating that the cluster iteration process is stopped, meaning that the clustering results have been obtained. The result is that cluster 1 consists only of Makassar City, while the other 23 cities and regencies are in cluster 2. This indicates that Makassar City has data characteristics in the poverty indicators that are significantly different from the others. Therefore, analysis continues with clustering without including Makassar City. The analysis is carried out with the same procedure, resulting in an optimal number of clusters of eight. The results of K-Means cluster can be presented in Figure 1 below:

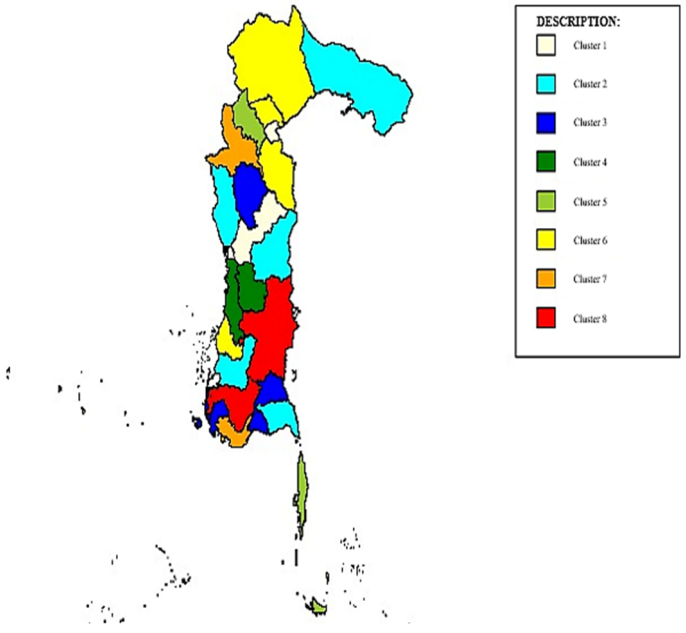


Figure 1: Further clustering results using K-Means

Source: data processed.

Based on Figure 1, it can be concluded that further clustering using K-Means cluster produce eight clusters, with cluster 1 consisting of Sidrap, Palopo city, and Pare-Pare city. Cluster 2 consists of Bulukumba, Maros, Wajo, Pinrang, and Luwu Timur. Cluster 3 consists of Takalar, Sinjai, Bantaeng, and Enrekang. Cluster 4 consists of Barru and Soppeng. Cluster 5 consists of Kepulauan Selayar and Toraja Utara. Cluster 6 consists of Luwu, Luwu Utara, and Pangkep. Cluster 7 consists of Tana Toraja and Jeneponto. Cluster 8 consists of Gowa and Bone.

3.4. Density-Based Spatial Clustering of Application with Noise

The initial step in DBSCAN is to determine the minimum points based on the number of analyzed variables, which in this case is two PC plus one, resulting in a minimum of three points. The determination of the hierarchy can be done by gradually selecting object p^*. The initial object p^* is the one with the smallest epsilon when the minimum points are three. Based on the data, the initial p^* object is Luwu Timur. Then, the next p^* has the second smallest epsilon, which is 0.437, in the case of Wajo and Maros. This hierarchy results in Wajo as the first new core point, having border points: Maros, Luwu Timur, and Pinrang. Meanwhile, Maros, as the second new core point, has border points: Wajo, Luwu Timur, and Bulukumba. The next step is to determine the clusters formed based on the principles of density achieved (reachable) and connected density (connectivity). Therefore, the direct arrived density for each core point in the second hierarchy is:

$p^*_{20} = \{p_8, p_{13}, p_{15}\}$;
$p^*_{13} = \{p_8, p_{20}, p_{15}\}$;
$p^*_8 = \{p_{20}, p_{13}, p_2\}$.

Object $p^*_{20}, p^*_{13}, p^*_8$ are core points, and if $\exists\ p^*_{20}$ is not a member of p^*_8 and vice versa $\in p^*_{20} \cup p^*_8$, then p^*_{20} and p^*_8 are considered as density reached. Based on the connectivity principle, if p^*_{20} and p^*_8 are considered density reached, then all members of p^*_{20} and p^*_8 are considered as connected density. The same process in determining clusters is repeated until all objects have been determined. Thus, based on the concept of the second hierarchy, it still forms one cluster with Luwu Timur Regency, Maros Regency, and Wajo Regency as core points. Meanwhile, the border points consist of Bulukumba and Pinrang. Additionally, 19 other regencies and cities are considered as noise.

The same procedure is repeated to determine core points by choosing the smallest third object and so on, carried out step by step based on the reachable density and

connectivity principles until all objects have been declared as core points. The cluster extraction process is further illustrated in Figure 2 below:

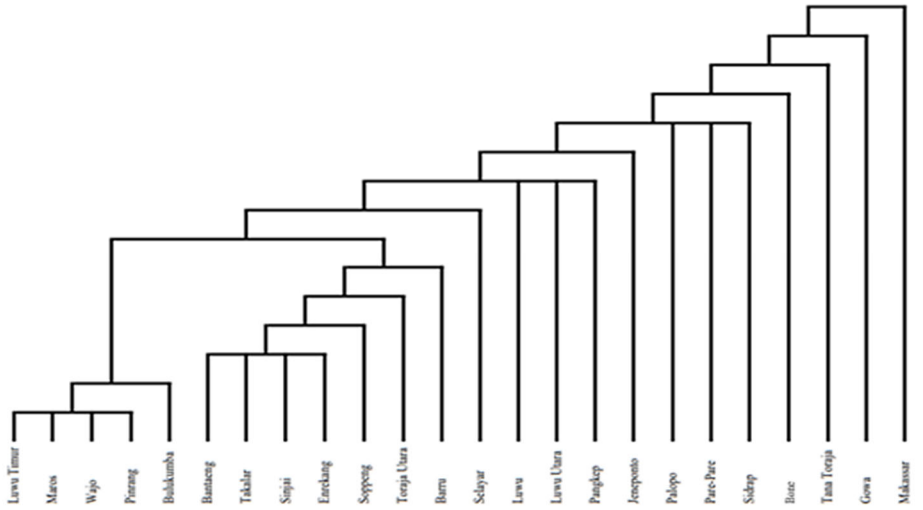


Figure 2: Extraction of DBSCAN Based on Hierarchy

Source: data processed.

Based on Figure 2, the SC values for each hierarchy level (denoted as h) can be shown in Table 7 below:

Table 7: SC Score Each Hierarchy Level

h	SC Score	h	SC Score	h	SC Score	h	SC Score
1	0.000	4	0.521	7	0.310	10	0.027
2	0.772	5	0.472	8	0.202	11	0.009
3	0.627	6	0.341	9	0.179	12	-0.015

Source: data processed.

Based on Table 7, the best hierarchy level is level 2, which results in one noise point, namely Makassar city, while the other 23 districts and cities form a single cluster. This indicates that Kota Makassar has different poverty indicator characteristics compared to the others. Therefore, the analysis continues with clustering without including Kota Makassar. The analysis is performed using the same procedure, the best hierarchy at the fourth level with one cluster and three noise points consisting of Bone, Gowa, and Tana Toraja. So, the results of this research can provide information that determining

epsilon in DBSCAN can be represented through the hierarchical principle, whereas the DBSCAN results in this research can be presented in Figure 3 below:

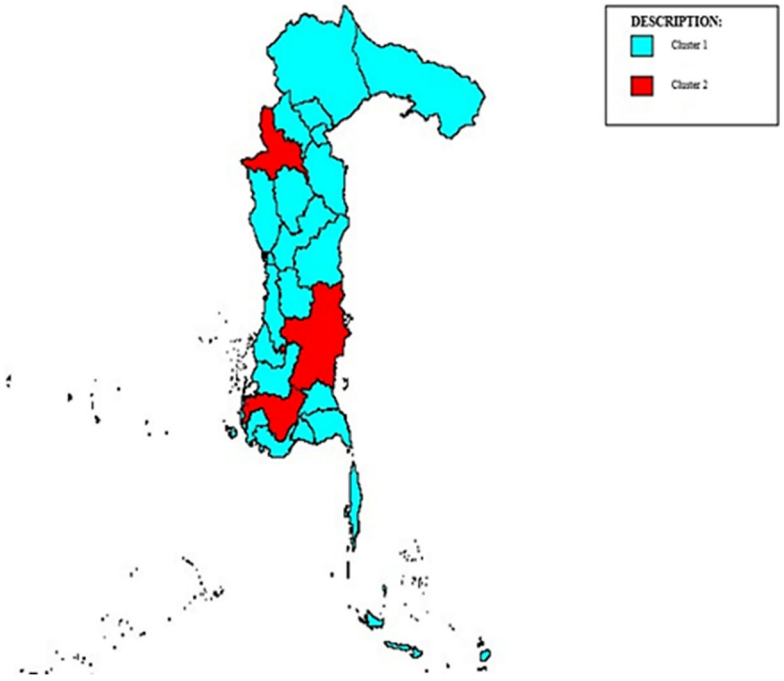


Figure 3. Further clustering results using DBSCAN

Source: data processed.

3.5. Silhouette Coefficient

The next step after the clustering process of each method is to evaluate the clustering results obtained with the Silhouette Coefficient according to Equation (7). The purpose is to assess how well the clusters that have been constructed perform. The comparison of the K-Means and DBSCAN hierarchical-based methods using the Silhouette Coefficient without Makassar City can be presented in Table 8 as follows:

Table 8. Comparison of SC Values for K-Means and DBSCAN

Method	SC Value
K-Means	0.507
DBSCAN	0.318

Source: data processed.

Based on Table 8, referring to Table 1, it can be concluded that K-Means achieves a good clustering result because it falls within the range of $0.50 < SC \leq 0.70$. On the other hand, DBSCAN obtains a weak clustering result as it falls within the range of $0.25 < SC \leq 0.50$.

3.6. Profiling Cluster

Based on the clustering results, the best method is K-Means. The profiles and explanations of the characteristics of each cluster are as follows:

a. Cluster 1 consists of Sidrap, Kota Palopo, and Kota Pare-Pare. This cluster has the highest population density. However, it also has the highest Human Development Index (IPM), per capita expenditure, and relatively low poverty indicators, making it the most prosperous cluster among the others.
b. Cluster 2 consists of Bulukumba, Maros, Wajo, Pinrang, and Luwu Timur. This cluster has a relatively high IPM and per capita expenditure, as well as relatively low poverty indicators, categorizing it as a prosperous cluster.
c. Cluster 3 consists of Takalar, Sinjai, Bantaeng, and Enrekang. This cluster has relatively low IPM and per capita expenditure. It also has relatively low poverty indicators, making it moderately prosperous.
d. Cluster 4 consists of Barru and Soppeng. This cluster has relatively low IPM and per capita expenditure compared to other clusters. It also has the lowest average percentage of poor population and relatively low poverty indicators, categorizing it as prosperous based on these indicators.
e. Cluster 5 consists of Kepulauan Selayar and Toraja Utara. Despite having the lowest population, workforce, and unemployment figures, this cluster has a relatively low IPM and per capita expenditure compared to other clusters, categorizing it as relatively poor.
f. Cluster 6 consists of Luwu, Luwu Utara, and Pangkep. This cluster has the highest percentage of poor population and the second-highest workforce participation rate, categorizing it as a poor region.
g. Cluster 7 consists of Tana Toraja and Jeneponto. This cluster has the lowest IPM and per capita expenditure and the second-highest percentage of poor population and workforce participation rate, categorizing it as poor.
h. Cluster 8 consists of Gowa and Bone. This cluster has the highest population, workforce, and unemployment figures and the second-lowest IPM, categorizing it as a poor region.

Based on these characteristics, clusters 5, 6, 7, and 8 require special attention from the government based on the analysis. Meanwhile, cluster 1 is the most prosperous cluster. The results of this research are useful for the government to identify poverty indicators for each region in South Sulawesi.

4. Conclusions

This study obtained that K-Means cluster obtained the optimal number of clusters, which is two. Cluster 1 consists only of Kota Makassar, while cluster 2 consists of 23 districts and cities. Further analysis revealed eight optimal clusters. On the other hand, the Density-Based Spatial Clustering of Application with Noise produced one large cluster with 23 districts and cities, classifying Kota Makassar as noise. Further analysis at the best hierarchy level yielded three noise points, including Gowa, Bone, and Tana Toraja, while the other districts and cities belong to a single cluster. K-Means clustering provided more effective groupings of the poverty indicator data for South Sulawesi Province in 2022 compared to Density-Based Spatial Clustering of Application with Noise. This is evident in the higher Silhouette Coefficients obtained, with values of 0.507 and 0.318, respectively. K-Means clustering achieved better groupings, while Density-Based Spatial Clustering of Application with Noise resulted in weaker clustering. So, the study concludes that the K-Means method is more effective than DBSCAN in helping the government to group the poverty characteristics of each region so that it can overcome poverty cases in South Sulawesi Province.

Acknowledgement

Thanks to the Department of Statistics, Universitas Hasanuddin, along with my colleagues and academic staff, for their invaluable support and contributions in the preparation of this research article.

References

Abdi, H., Williams, L. J., (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), pp. 433–459.

Akbar, T., Tinungki, G. M. and Siswanto, (2023). Performance of K-Medoids and Density Based Spatial Clustering of Application with Noise Using Silhouette Coefficient Test. *Barekeng: J. Math. & App*, 17(3), pp. 1605–1616.

Astutik, S., Solimun and Darmanto, (2018). Analisis Multivariat: Teori dan Aplikasinya dengan SAS. UB Press.

Bari, M. A., Kindzierski, W. B., (2018). Ambient volatile organic compounds (VOCs) in Calgary, Alberta: Sources and screening health risk assessment. *Science of the Total Environment*, 631, pp. 627–640.

Batool, F., Hennig, C., (2021). Clustering with the Average Silhouette Width. *Computational Statistics and Data Analysis*, 158(107190), pp. 1–18.

BPS, (2023, March). *Profil Kemiskinan di Sulawesi Selatan*. *https://sulsel.bps.go.id*

Chowdhury, S., Helian, N. and Cordeiro de Amorim, R., (2023). Feature weighting in DBSCAN using reverse nearest neighbours. *Pattern Recognition*, 137(109314), pp. 1–15.

Cordeiro de Amorim, R., Makarenkov, V., (2023). On k-means iterations and Gaussian clusters. *Neurocomputing*, 553(126547), pp. 1–10.

Dewi, D. A. I. C., Pramita, D. A. K., (2019). Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. *Jurnal Matrix*, 9(3), pp. 102–109.

Festa, D., Novellino, A., Hussain, E., Bateson, L., Casagli, N., Confuorto, P., Soldato, M. D. and Raspini, F., (2023). Unsupervised detection of InSAR time series patterns based on PCA and K-means clustering. *International Journal of Applied Earth Observation and Geoinformation*, 118, pp. 1–13

González, C. A. D, Calderón, Y. M. M, Cruz, N. A. M and Sandoval, L. E. P., (2022). Typologies of Colombian off-grid localities using PCA and clustering analysis for a better understanding of their situation to meet SDG-7. *Cleaner Energy Systems*, *3*(100023), pp. 1–16.

Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L. and Maggio, R. M., (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science and Technology,* 72, pp. 83–90.

Hahsler, M., Piekenbrock, M. and Doran, D., (2019). dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software,* 91, pp. 1–30.

Hair, J. F. J. R., Black, W. C., Babin, B. J. and Anderson, R. E., (2010). Multivariate Data Analysis (7th ed.). Pearson Education Inc.

Huang, Q., Chen, S. and Li, Y., (2023). Selection of seismic noise recording by K-means. *Case Studies in Construction Materials*, 19 (e02363), pp 1–16.

Jing, W., Zhao, C. and Jiang, C., (2019). An Improvement Method of DBSCAN Algorithm on Cloud Computing. *Procedia Computer Science*, 147, pp. 596–604.

Johnson, R. A., Wichern, D. W., (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall.

Kherif, F., Latypova, A., (2019). Principal Component Analysis. *In Machine Learning: Methods and Applications to Brain Disorders*, pp. 209–225.

Kurita, T., (2019). Principal Component Analysis (PCA). *In Computer Vision: a Reference Guide*, pp. 1–4.

Liu, G., Ji, F., Sun, W. and Sun, L., (2023). Optimization design of short-circuit test platform for the distribution network of integrated power system based on improved K-means clustering. *Energy Reports*, 9, pp. 716–726.

Nurhaliza, N., Mustakim, (2021). Pengelompokan Data Kasus Covid-19 di Dunia Menggunakan Algoritma DBSCAN. *IJIRSE*, *1*(1), 1–8.

Pramana, S., Yuniarto, B., Mariyah, S., Santoso, I. and Nooraeni, R., (2018). Data Mining Dengan R (Konsep Serta Implementasi). *In Media*.

Pu, G., Wang, L., Shen, J. and Dong, F., (2021). A hybrid unsupervised clustering-based anomaly detection method. *Tsinghua Sci Technol*, 26(2), pp. 146–153.

Rais, M., Goejantoro, R. and Prangga, S., (2021). Optimalisasi K-Means Cluster dengan Principal Component Analysis pada Pengelompokan Kabupaten/Kota di Pulau Kalimantan Berdasarkan Indikator Tingkat Pengangguran Terbuka. *Jurnal Eksponensial*, 12(2), pp. 129–135.

Řezanková, H. A. N. A., (2018). Different approaches to the silhouette coefficient calculation in cluster evaluation. *In 21st International Scientific Conference AMSE Applications of Mathematics and Statistics in Economics*, pp. 1–10.

Salmerón, R., García, C. B. and García, J., (2020). Variance Inflation Factor and Its Influence on Regression Models. *Journal of Statistical Computation and Simulation, 90*(12), pp. 1–15.

Starczewski, A., Cader, A., (2019). Determining the eps parameter of the DBSCAN algorithm. *In Artificial Intelligence and Soft Computing: 18th International Conference*, pp. 420–430.

Stewart, G., Al-Khassaweneh, M., (2022). An Implementation of the HDBSCAN Clustering Algorithm. *Applied Sciences*, 12(2405), pp. 1–21.

Zhang, R., Qiu, J., Guo, M., Cui, H. and Chen, X., (2022). An Adjusting Strategy after DBSCAN. *IFAC-PapersOnLine*, *55*(3), pp. 219–222.

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 129–149, https://doi.org/10.59139/stattrans-2025-019
Received – 23.10.2023; accepted – 12.12.2024

Testing for multinormality with goodness-of-fit tests based on phi divergence measures

Mbanefo S. Madukaife[1], **Uchenna C. Nduka**[2], **Everestus O. Ossai**[3]

Abstract

In this paper, a beta transform of multivariate normal datasets is obtained. The phi divergence measure, $D_\Phi(F,G)$ between two distributions F and G is used to obtain a goodness-of-fit test to multivariate normality (MVN) based on the theoretical density function of the beta transformed random variable and a window-size-spacing-based sample density function. Three versions of the statistic are derived from three known phi divergence measures that are based on a sum of squares. The empirical critical values of the statistics are obtained and the empirical type-one-error rates as well as powers of the statistics in comparison with those of other well-known competing statistics are computed through extensive simulation study. The study shows that the new statistics have good control over type-one-error and are highly competitive with the existing well-known ones in terms of power performance. The applicability of the new statistics is also carried out in comparison with three other efficient techniques using four different datasets, and all the competing statistics agreed perfectly in their decisions of rejection or otherwise of the multivariate normality of the datasets. As a result, they can be regarded as appropriate statistics for assessing multinormality of datasets especially, in large samples.

Key words: beta transform of multivariate normal observation, empirical critical value, entropy estimator, phi divergence measure, power of a test.

1. Introduction

The search for a more tractable, highly powerful and generally acceptable goodness-of-fit techniques for assessing the normality of a set of data has continued to receive the attraction of cross-generational researchers in the field of statistical methodology. Since the pioneer work of Pearson (1900), more than ten scores of such techniques at both univariate and multivariate spheres have been introduced in the literature from diverse unique characterizations of the normal distribution. These characterizations range from the distribution functions, generating functions (moment generating function, characteristic function and Laplace transform), skewness and kurtosis and entropy, to mention but a few, to other characterizations of various transformations of the normal distribution.

[1]University of Nigeria, Nsukka, Nigeria. E-mail: mbanefo.madukaife@unn.edu.ng. ORCID: https://orcid.org/0000-0003-2823-4223.
[2]University of Nigeria, Nsukka, Nigeria. E-mail: uchenna.nduka@unn.edu.ng. ORCID: https://orcid.org/0000-0001-5931-2840.
[3]University of Nigeria, Nsukka, Nigeria. E-mail: everestus.ossai@unn.edu.ng. ORCID: https://orcid.org/0000-0001-9742-2389.

Of particular attention are tests for multivariate normality (MVN). This is probably due to the fact that most classical multivariate statistical techniques, with diverse applications in many areas of study such as machine learning, econometrics and genomics, require multivariate normality. Suppose $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n; \mathrm{x}_j \in R^d$, $j = 1,2,\ldots,n$ is a sequence of n independent and identically distributed (iid) d-dimensional random vectors from an unknown distribution $F(\mathrm{x})$; where $d \geq 2$ is an integer. The problem of testing for MVN is that of testing the null hypothesis

$$H_0 : F(x) \in F_N \tag{1}$$

against an alternative that $F(x) \notin F_N$; where F_N is a class of nondegenerate d-dimensional multivariate normal distributions with mean vector μ and nondegenerate covariance matrix Σ. Examples of tests devoted to (1) in the literature include Healy (1968), Mardia (1970, 1974), Malkovich and Afifi (1973), Small (1978), Royston (1983), Srivastava (1984), Baringhaus and Henze (1988), Henze and Zirkler (1990), Singh (1993), Romeu and Ozturk (1993), Henze and Wagner (1997), Hwu et al. (2002), Szekeley and Rizzo (2005), Pudelko (2005), Doornik and Hansen (2008), Liang et al. (2009), Cardoso de Oliveira and Ferreira (2010), Hanusz and Tarasinska (2012), Zhou and Shao (2014), Thulin (2014), Korkmaz et al. (2014), Tenreiro (2017), Madukaife and Okafor (2018), Madukaife and Okafor (2019), Henze and Jimenez-Gamero (2019), Henze et al. (2019), Henze and Visagie (2020), Dorr et al. (2020a, 2020b). For extensive reviews on different tests for MVN in their various classes, see Henze (2002), Mecklin and Mundfrom (2004), Ebner and Henze (2020) as well as Chen and Genton (2023).

Some of the developed techniques are direct extension of tests for univariate normality to their multivariate counterparts. For instance, Epps and Pulley (1983) developed a test for univariate normality as an integral of the squared difference between the theoretical and empirical characteristic functions of the univariate normal distribution. They showed that the test was very consistent against all fixed alternatives and affine invariant (invariant with respect to changes in location and scale) with competitive high power performance. Because of its interesting properties, Baringhaus and Henze (1988) developed its multivariate counterpart. Since then, several versions of it have been developed and they are coined Baringhaus-Henze-Epps-Pulley (BHEP) class of tests for multivariate normality by Csorgo (1989). In a like manner, Shapiro and Wilks (1965) obtained an omnibus test for assessing univariate normality of a dataset, $x_1, x_2, \ldots, x_n; x_j \in R$, $j = 1,2,\ldots,n$, which they defined as a ratio of two variance estimators obtained from the dataset and stated that if the dataset is drawn from a normal distribution, then the two estimators would amount to the same value, thereby approaching 1. With the intension of obtaining a multivariate test that inherits the good power performance of the Shapiro and Wilks (1965) test, Villasenor and Gonzalez-Estrada (2009) extended it to the multivariate sphere and the resultant statistic shows an appreciable good power performance. Recently, Tavakoli et al. (2020) applied the sample entropy measure of Vasicek (1976) to estimate phi divergence measures $D_\Phi(F,G)$ between a normal distribution, F and an unknown distribution, G, from where a random sample $x_1, x_2, \ldots, x_n; x_j \in R$, $j = 1,2,\ldots,n$ is drawn. They argued that if G is also a normal distribution, then, $D_\Phi(F,G)$ will be a minimum. With this, they introduced consistent and affine invariant tests for univariate normality which are very tractable and have high power

performances with good control over type-I-error. Since the search for more tractable tests for MVN with relatively high competitive performances is an open research in the literature, it suffices that extension of the Tavakoli et al. (2020) procedures to their multivariate counterparts with some one-to-one transformations would, no doubt, retain the properties and hence serve as highly competitive tests for MVN. This is the purpose of the present paper. The rest of the paper is presented as follows: the statistics are developed in Section 2, with their properties. Section 3 gives the empirical critical values as well as the empirical size and power comparisons. Section 4 gives some real-life applications of the new statistics in comparison with some other statistics while the paper is concluded in Section 5.

2. The test statistic

Suppose $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n; \mathrm{x}_j \in R^d$, $j = 1,2,\ldots,n$, and $d \geq 2$ is a d-dimensional random sample from a continuous distribution F. Healy (1968) obtained the sample Mahalanobis squared distances of the observations, which he defined as squared radii, as

$$y_j = (\mathbf{x}_j - \overline{\mathbf{x}}_n)^T S_n^{-1} (\mathbf{x}_j - \overline{\mathbf{x}}_n); \; j = 1, 2, \ldots, n \tag{2}$$

where $\overline{\mathbf{x}}_n = n^{-1} \sum_{j=1}^n \mathbf{x}_j$ is the sample mean vector and $S_n = (n-1)^{-1} \sum_{j=1}^n (\mathbf{x}_j - \overline{\mathbf{x}}_n)(\mathbf{x}_j - \overline{\mathbf{x}}_n)^T$ is the sample covariance matrix. Under the null distribution of multivariate normality of F, Healy (1968) stated that the squared radii are asymptotically distributed as chi-squared observations with d degrees of freedom. Gnanadesikan and Kettenring (1972) obtained a transform of the squared radii as

$$z_j = \frac{n}{(n-1)^2} (\mathbf{x}_j - \overline{\mathbf{x}}_n)^T S_n^{-1} (\mathbf{x}_j - \overline{\mathbf{x}}_n); \; j = 1, 2, \ldots, n \tag{3}$$

and stated that z_j's are exact independent univariate observations from the beta distribution of the first kind, $B(a,b)$ with parameters $a = d/2$ and $b = (n-d-1)/2$ under the MVN of F, where n and d have their usual meanings. The exactness of this assumption was proved by Bilodeau and Brenner (1999) and it has since been used in goodness-of-fit statistics such as Small (1978), Hanusz and Tarasinska (2012) and Madukaife (2017). It is interesting to note here that the transformations in (2) and (3) are functions of d-dimensional observations, $d > 1$. However, even when the observations emanate from a univariate distribution, $d = 1$, it is natural to still obtain z_j's as beta distributed independent observations, $B(a,b)$, with $a = 1/2$ and $b = (n-2)/2$. As a result, the statistics obtained in this study can also apply to univariate normality testing.

Now, the phi divergence measure between any two distributions F_X and G_X, with density functions $f(x)$ and $g(x)$ respectively, is defined by

$$D_\Phi(F_X, G_X) = \int_{-\infty}^{\infty} \Phi\left(\frac{g(x)}{f(x)}\right) f(x)\, dx \tag{4}$$

where $\Phi(x)$ is a convex function such that $\Phi(1) = 0$ and $\Phi''(1) > 1$. At different times, a number of works have independently obtained different convex functions satisfying the

conditions of the $\Phi(x)$ in (4) and these works have led to different phi divergence functions. Some of them include the following, as listed in Tavakoli et al. (2019) and Tavakoli et al. (2020):

Kullback-Leibler divergence measure, $\Phi(x) = xlog(x)$.

Pearson divergence measure, $\Phi(x) = (x-1)^2$;

Hellinger divergence measure, $\Phi(x) = 1/2(\sqrt{x}-1)^2$;

Triangular divergence measure, $\Phi(x) = \frac{(1-x)^2}{1+x}$;

Lin-Wong divergence measure, $\Phi(x) = xlog\left(\frac{2}{1+x}\right)$;

Jeffreys divergence measure, $\Phi(x) = (x-1)log(x)$;

Total variation divergence measure, $\Phi(x) = |x-1|$; and

Balakrishnan-Sanghvi divergence measure, $\Phi(x) = \left(\frac{x-1}{x+1}\right)^2$. For more divergence measures and more details on them, readers are referred to Lin (1991).

Now, using the method of estimating the entropy of a random variable by Vasicek (1976), Tavakoli et al. (2020) obtained $D_\Phi(F,G)$ in (4) when G is normal with mean μ and variance σ^2 and F is unknown as:

$$D_\Phi(F_X, G_X) = \int_0^1 \Phi\left(\frac{(2\pi\sigma^2)^{-1/2} exp\left(-\left(F^{-1}(p)-\mu\right)^2/2\sigma^2\right)}{(dF^{-1}(p)/dp)-1}\right)^2 dp \tag{5}$$

where $F(x) = p \implies F^{-1}(p) = inf\{x : F(x) = p\}; p \in (0,1)$. Replacing F in (5) with F_n (the empirical distribution function) and using the difference operator in place of differential operator, they obtained an estimator, V_Φ of $D_\Phi(F,G)$ as a generic statistic for testing the normality of a set of n observations. The statistic is given as:

$$V_\Phi = \frac{1}{n}\sum_{j=1}^{n} \Phi\left(\frac{n}{\sqrt{2\pi\hat{\sigma}^2}} exp\left\{-\frac{(X_{(j)}-\hat{\mu})^2}{2\hat{\sigma}^2}\right\}\frac{\left(X_{(j+m)}-X_{(j-m)}\right)}{2m}\right) \tag{6}$$

where $X_{(j)}$ is the jth order statistic, $j = 1, 2, , \ldots, n$, of the random sample, $X_1, X_2, \ldots, X_n$ such that $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$; $\hat{\mu} = \overline{X}$; $\hat{\sigma}^2 = n^{-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$ and m, known as the window size or spacing, is an integer such that $m \leq \frac{n}{2}$. They proved that the statistic is consistent against fixed alternatives and that it is affine invariant. The test rejects the null hypothesis of normality for large values of the statistic and it is said to be generic because it is ammenable to any specific phi function in the class of phi divergence measures.

It is very clear from Vasicek (1976) as well as Tavakoli et al. (2020) that the development of the theory behind (5) and (6) does not depend on the normality of F_X and G_X. As a result, a plug-in method is possible for goodness-of-fit statistics to statistical distributions. Therefore, let $z_1, z_2, \ldots, z_n$ be the beta transforms of the random sample according to (3) and let G_Z be beta distributed with parameters a and b such that

$$g(z) = \frac{1}{B(a,b)} z^{a-1}(1-z)^{b-1}; 0 < z < 1 \tag{7}$$

Then, replacing the normal density function in (5) with that of the beta in (7), $D_\Phi(F,G)$ can be presented as:

$$D_\Phi(F_Z,G_Z) = \int_0^1 \Phi\left(\frac{B(a,b)^{-1}\left[F^{-1}(p)\right]^{a-1}\left[1-F^{-1}(p)\right]^{b-1}}{\left(dF^{-1}(p)/dp\right)^{-1}}\right)^2 dp \tag{8}$$

This gives rise to a new generic goodness-of-fit statistic obtained similar to (6) by replacing F with F_n and using difference operator in place of differential operator. It is given as:

$$M_{n,\Phi} = \frac{1}{n}\sum_{j=1}^{n}\Phi\left(\frac{n\Gamma(a+b)(Z_{(j)})^{a-1}(1-Z_{(j)})^{b-1}\left(Z_{(j+m)}-Z_{(j-m)}\right)}{2m\Gamma(a)\Gamma(b)}\right) \tag{9}$$

where $a = \frac{d}{2}$; $b = \frac{(n-d-1)}{2}$ and m has its usual meaning. The test rejects the null hypothesis of MVN for large values of the statistic. Also, it is invariant with respect to changes in the scale and location of the observation vectors. This is because the transformations in (2) and (3) are standardized transformations that result in beta distributed observations with constant parameters for each n and d such that no matter the affine transformation in $\mathbf{x}_j$'s, $j = 1,2,\ldots,n$, z_j's have a specified beta distribution with specified parameters.

Theorem 2.1: Suppose $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n; \mathrm{x}_j \in R^d$, $j = 1,2,\ldots,n$ *is a random sample from an unknown continuous distribution* $F(\mathbf{x})$ *with a probability density function* $f(\mathbf{x})$. *The statistic* $M_{n,\Phi}$ *obtained from the observation vectors is invariant with respect to changes in scale and location of the observation vectors.*

Proof:

Let C be defined as a dxd nonsingular matrix of constants and $\mathbf{u}$ a d-component vector of constants. The affine invariance of $M_{n,\Phi}$ stems from the affine invariance of the Mahalanobis squared distance. That is, for $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$, the sample mean vector is $\overline{X}_n$ and the sample covariance matrix is S_n. Also, for affine transformed observation vectors $C\mathrm{x}_1 \pm \mathbf{u}, C\mathrm{x}_2 \pm \mathbf{u}, \ldots, C\mathrm{x}_n \pm \mathbf{u}$, the sample mean vector is $C\overline{X}_n \pm u$ and the sample covariance matrix is CS_nC. Then the sample Mahalanobis squared distance of the affine transformed observation vectors is given by:

$$\left[C\mathbf{x}_j \pm u - C\overline{X}_n \pm u\right]^T (CS_nC)^{-1}\left[C\mathbf{x}_j \pm u - C\overline{X}_n \pm \mathbf{u}\right]$$

$$\left[C\left(\mathbf{x}_j - \overline{X}_n\right)\right]^T (CS_nC)^{-1}\left[C\left(\mathbf{x}_j - \overline{X}_n\right)\right]$$

$$\left(\mathbf{x}_j - \overline{X}_n\right)^T C^T\left(C^T\right)^{-1} S_n^{-1}C^{-1}C\left(\mathbf{x}_j - \overline{X}_n\right) = \left(\mathbf{x}_j - \overline{X}_n\right)^T S_n^{-1}\left(\mathbf{x}_j - \overline{X}_n\right)$$

Hence,

$$\frac{n}{(n-1)^2}\left(\mathbf{x}_j - \overline{X}_n\right)^T S_n^{-1}\left(\mathbf{x}_j - \overline{X}_n\right) = z_j$$

$$= \frac{n}{(n-1)^2}\left(C\mathbf{x}_j \pm \mathbf{u} - C\overline{X}_n \pm \mathbf{u}\right)^T (CS_nC)^{-1}\left(C\mathbf{x}_j \pm \mathbf{u} - C\overline{X}_n \pm \mathbf{u}\right)\ j = 1,2,\ldots,n.$$

Since $Z \sim B(a,b)$, where $a = \frac{d}{2}$ and $b = \frac{(n-d-1)}{2}$ which do not depend on any sample observation vector $\mathbf{x}_j$, the invariance property is proved.

The invariance property of the $M_{n,\Phi}$ statistic, as proved in Theorem 2.1 is because the transformations in (2) and (3) are standardized transformations that result in beta distributed observations with constant parameters for each n and d such that no matter the affine transformation in $\mathbf{x}_j$'s, $j = 1,2,\ldots,n$, z_j's have a specified beta distribution (with specified parameters). As a result, under any null distribution of MVN, the value of the statistic is unaffected at any fixed sample size n and variable dimension d.

Theorem 2.2: Suppose $F(\mathbf{x})$ is an unknown continuous distribution in a d-dimensional real space, R^d, with a probability density function $f(\mathbf{x})$, having unknown mean vector and unknown covariance matrix. Then, the test based on $M_{n,\Phi}$ is consistent.

Proof:

Under the null distribution of multivariate normality, $Z \sim B(a,b)$, where Z is the random variable from where z_j in (3) is assumed to have come from. Hence, as $n, m \to \infty$ and $m/n \to 0$,

$$F_n\left(z_{(j+m)}\right) - F_n\left(z_{(j-m)}\right) \simeq F\left(z_{(j+m)}\right) - F\left(z_{(j-m)}\right)$$
$$\simeq \frac{f\left(z_{(j+m)}\right) + f\left(z_{(j-m)}\right)}{2}\left(z_{(j+m)} - z_{(j-m)}\right).$$

Now, it is obvious that the a and b in the distribution of Z are consistent since a is fixed and b is based on sample size.

Hence, $E\left(M_{n,\Phi}\right) = E\left(\frac{1}{n}\sum_{j=1}^{n}\Phi\left(\frac{n\Gamma(a+b)(Z_{(j)})^{a-1}(1-Z_{(j)})^{b-1}\left(Z_{(j+m)}-Z_{(j-m)}\right)}{2m\Gamma(a)\Gamma(b)}\right)\right)$

$$= E\left\{\Phi\left(\frac{n\Gamma(a+b)(Z_{(j)})^{a-1}(1-Z_{(j)})^{b-1}\left(Z_{(j+m)}-Z_{(j-m)}\right)}{2m\Gamma(a)\Gamma(b)}\right)\right\}.$$

Again, $Z_{(j-m)}$ and $Z_{(j+m)}$ belong to an interval where $f(z)$ is both positive and continuous. Then according to Vasicek (1976) and Tavakoli et al. (2020), there exists $z_j^* \in \left(Z_{(j-m)}, Z_{(j+m)}\right)$ such that

$$\frac{F\left(Z_{(j+m)}\right) - F\left(Z_{(j-m)}\right)}{Z_{(j+m)} - Z_{(j-m)}} = f(z_j^*).$$

Therefore, $M_{n,\Phi} \to D_\Phi\left(F_Z, G_Z\right)$ and hence, $M_{n,\Phi}$ is consistent.

3. Simulation study

In this section, extensive simulations are carried out to obtain the critical values of the proposed test as well as to determine their relative performance. For these purposes, it is important to first determine an appropriate window size, m for each sample size, n and number of variables, d in a multivariate dataset. Wieczorkowski and Grzegorzewski (1999) have proposed an optimal value of m for estimating the entropy of a distribution to be a function of the sample size as $m = [\sqrt{n} + 0.5]$, where $[x]$ is the integer part of x. However, it

has been shown that an appropriate m also depends on the underlying distribution in addition to the sample size as against the suggestion of Wieczorkowski and Grzegorzewski (1999). Again, one serious problem with the application of Tavakoli et al. (2020) statistic is lack of operational function for determining m.

Now, the statistic is based on the Vasicek (1976) estimator of the Shannon (1941) entropy of a random variable. Therefore, the empirical mean squared error (EMSE) of the Vasicek (1976) estimator was computed for all the possible values of m, $m \leq \frac{n}{2}$ under the beta distribution with parameters $a = \frac{d}{2}$; $b = \frac{(n-d-1)}{2}$. This is carried out for $n = 5\,(5)\,100\,(10)\,150$ and $d = 2, 5,$ and, 10. In each combination of n and d, an appropriate m is obtained as the one with the smallest EMSE and we used the selected m values to obtain a linear trend equation of m for each combination of n and d as $m = 3.2349 + 0.0808n - 0.2929d$, with an R^2 value of 89 percent, see Figure 1. The Vasicek (1976) estimator is given by $H_{mn} = \frac{1}{n}\sum_{j=1}^{n} log\left\{\frac{n}{2m}\left(X_{(j+1)} - X_{(j-m)}\right)\right\}$ and the EMSE is based on 10,000 replications of each sample size, n drawn from the beta distribution with parameters $a = \frac{d}{2}; b = \frac{(n-d-1)}{2}$.

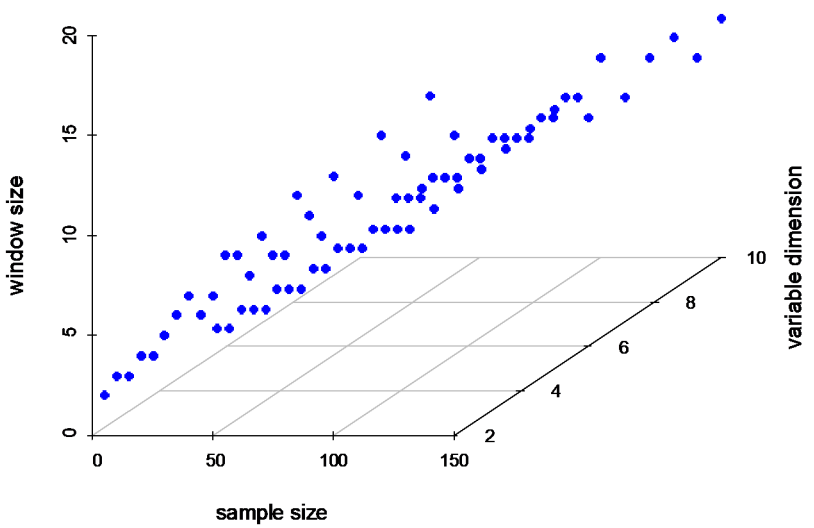


Figure 1: 3D scatter plot of n, d and associated m

3.1. Empirical critical values of the test

The statistic proposed in this work is generic in nature. Therefore, its critical value, application and performance depend on the specific phi divergence measure being used. We

have stated that there are several phi divergence measures but our study in this section is limited to only three which are based on sum of squares. They are the Pearson; Hellinger; and Balakrishnan-Sanghvi divergence measures and the proposed statistic for them are $M(P)$, $M(H)$, and $M(BS)$ respectively.

For each combination of sample sizes n = 10 (5) 100 (10) 150 and random vector dimensions d = 2, 5, and 10, 5 percent level critical values were evaluated. To achieve this, a total of N = 100,000 samples for each combination of n and d from the standard multivariate normal distribution were generated and each generated sample was transformed into a beta sample according to (3). Then, each of the three sum of square versions of the statistic is computed from each of the beta transformed samples to arrive at N = 100,000 values of each statistic. The 5 percent level critical value is calculated as the 95 percentile of the N values in each version of the statistic. The critical values are presented in Table 1. The test then proposes to reject the MVN of a dataset with sample size n and number of variables d if the computed value of the statistic is greater than the corresponding critical value at 5 percent level of significance.

Table 1: Empirical critical values at $\alpha = 0.05$

	$M(P)$			$M(H)$			$M(BS)$		
n	$d=2$	$d=5$	$d=10$	$d=2$	$d=5$	$d=10$	$d=2$	$d=5$	$d=10$
10	0.5773	0.4318	-	0.1687	0.1695	-	0.1988	0.2195	-
15	0.6623	0.4764	0.8508	0.1604	0.1528	0.2220	0.1697	0.1853	0.2491
20	0.6173	0.4042	0.8188	0.1423	0.1250	0.2072	0.1422	0.1490	0.2281
25	0.5633	0.4044	0.4720	0.1262	0.1154	0.1311	0.1240	0.1324	0.1502
30	0.5281	0.3591	0.4713	0.1156	0.1019	0.1253	0.1112	0.1159	0.1402
35	0.4789	0.3473	0.3612	0.1047	0.0953	0.0997	0.1014	0.1067	0.1133
40	0.4464	0.3248	0.3557	0.0961	0.0884	0.0952	0.0932	0.0979	0.1064
45	0.4291	0.3141	0.3500	0.0914	0.0828	0.0918	0.0868	0.0916	0.1013
50	0.3956	0.2983	0.2953	0.0851	0.0790	0.0800	0.0812	0.0862	0.0896
55	0.3891	0.2848	0.2921	0.0826	0.0743	0.0775	0.0773	0.0813	0.0853
60	0.3626	0.2757	0.2622	0.0767	0.0714	0.0709	0.0728	0.0769	0.0784
65	0.3480	0.2656	0.2557	0.0728	0.0689	0.0683	0.0690	0.0741	0.0750
70	0.3434	0.2589	0.2518	0.0714	0.0653	0.0663	0.0667	0.0705	0.0725
75	0.3252	0.2490	0.2331	0.0682	0.0634	0.0622	0.0634	0.0677	0.0684
80	0.3243	0.2443	0.2284	0.0669	0.0617	0.0606	0.0619	0.0659	0.0660
85	0.3100	0.2368	0.2171	0.0642	0.0598	0.0581	0.0594	0.0632	0.0636
90	0.2980	0.2353	0.2144	0.0615	0.0587	0.0564	0.0569	0.0621	0.0615
95	0.2979	0.2306	0.2100	0.0608	0.0570	0.0549	0.0559	0.0600	0.0593
100	0.2876	0.2235	0.2041	0.0589	0.0550	0.0532	0.0537	0.0579	0.0580
110	0.2835	0.2163	0.1946	0.0565	0.0527	0.0509	0.0514	0.0553	0.0552
120	0.2741	0.2127	0.1880	0.0549	0.0510	0.0490	0.0495	0.0530	0.0528
130	0.2708	0.2093	0.1818	0.0536	0.0497	0.0468	0.0475	0.0512	0.0503
140	0.2542	0.2049	0.1786	0.0505	0.0483	0.0454	0.0450	0.0496	0.0485
150	0.2527	0.1975	0.1752	0.0495	0.0461	0.0442	0.0436	0.0474	0.0473

3.2. Description of the competing tests

Primarily, the three versions of our proposed test are according to the phi divergence measures due to Pearson, Hellinger, as well as Barakrishnan and Sanghvi. They are

$$M(P) = \frac{1}{n}\sum_{j=1}^{n}(U_j - 1)^2$$

$$M(H) = \frac{1}{2n}\sum_{j=1}^{n}\left(\sqrt{U_j} - 1\right)^2$$

$$M(BS) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{U_j - 1}{U_j + 1}\right)^2$$

where $U_j = \frac{n\Gamma(a+b)Z_{(j)}^{a-1}(1-Z_{(j)})^{b-1}\left(Z_{(j+m)}-Z_{(j-m)}\right)}{2m\Gamma(a)\Gamma(b)}$; $a = d/2$; $b = (n-d-1)/2$; $Z_{(j)}$ is the jth order statistic of the Z-transformed dataset such that $Z_{(j+m)} = Z_{(n)}$ for all $j+m \geq n$ and $Z_{(j-m)} = Z_{(1)}$ for all $j-m \leq 1$.

The existing statistics considered in this work for a proper comparison with these new ones include the Henze and Zirkler (HZ) test for MVN of Henze and Zirkler (1990); the Madukaife (M) test for MVN of Madukaife (2017); and the Henze and Jimenez-Gamero (HJG) test for MVN of Henze and Jimenez-Gamero (2019). The choice of the three competing tests is not completely arbitrary. First, they are all affine invariant and consistent L^2-type tests for MVN with good power performances. Secondly, in most comparative studies on powers of tests for MVN, the $HZ-$statistic has remained a reference point while the $HJG-$statistic is similar to it. In fact, any test for MVN that competes favourably with the $HZ-$statistic is generally regarded as a good statistic for assessing MVN of datasets. Again, the $M-$statistic is also based on beta transform of multivariate datasets. Also, since the choice of d presented in this work is $d = 2, 5, 10$, which represents multivariate datasets, comparison with good univariate tests for normality such as Jargue and Bera (1987) as well as Bayoud (2021) is not discussed here. It may be the interest of a future work. In what follows, therefore, the three competing statistics are described.

3.2.1 Henze and Zirkler HZ test

Henze and Zirkler (1990) introduced a smoothing parameter, β in the weight function of the consistent and affine invariant statistic due to Baringhaus and Henze (1988) to obtain a highly regarded test for MVN of multivariate datasets. The statistic is given as:

$$HZ = n\left(4I\left\{S_n \text{ is singular}\right\} + D_{n,\beta}I\left\{S_n \text{ is nonsingular}\right\}\right)$$

where $D_{n,\beta} = \left(1+2\beta^2\right) + n^{-2}\sum_{j,k=1}^{n} exp\left\{-\frac{\beta^2\left\|y_j - y_k\right\|^2}{2}\right\}$
$-2\left(1+\beta^2\right)^{-d/2}n^{-1}\sum_{j=1}^{n} exp\left\{-\frac{\beta^2\left\|y_j\right\|^2}{2\left(1+\beta^2\right)}\right\}$; $\beta > 0$ and $I\{.\}$ is an indicator function. The test is universally consistent, affine invariant and rejects MVN of datasets for large values of the statistic, with appropriate $\beta = \frac{((2d+1)n/4)/(1/(d+4))}{\sqrt{2}}$.

3.2.2 Madukaife M test

Madukaife (2017) obtained a statistic to formalize the graphical test of Small (1978), using the sum of squared differences between expected and sample order statistics according to Madukaife and Okafor (2018), who also formalized the geometric procedure of Hanusz and Tarasinska (2012) to a classical test procedure. The statistic, which is the sum of squared differences between observed and expected order statistics of beta transformed observations, is given as

$$M = \sum_{j=1}^{n} \left(z_{(j)} - c_j\right)^2$$

where $z_{(j)}$ is the jth order statistic of the beta transformed observations and c_j is the corresponding jth expected order statistic from the beta distribution with parameters $a = d/2$ and $b = (n-d-1)/2$. The consistent and affine invariant test rejects the null hypothesis of MVN for large values of the statistic.

3.2.3 Henze and Jimenez-Gamero HJG test

Henze and Jimenez-Gamero (2019) obtained a statistic for assessing MVN based on the empirical moment generating function. It is a weighted squared integral of the difference between the theoretical and empirical moment generating functions respectively of the standard multivariate normal distribution and a multivariate dataset. The statistic is given as

$$HJG = \pi^{d/2}\left(\frac{1}{n}\sum_{j,k=1}^{n}\frac{1}{\beta^{d/2}}exp\left\{\frac{\left\|Y_{n,j}+Y_{n,k}\right\|^2}{4\beta}\right\}+\frac{n}{(\beta-1)^{d/2}}\right) - 2\pi^{d/2}\left(\sum_{j=1}^{n}\frac{1}{(\beta-\frac{1}{2})^{d/2}}exp\left\{\frac{\left\|Y_{n,j}\right\|^2}{4\beta-2}\right\}\right),$$

where $\beta > 1, Y_{n,j}$ is the jth d-dimensional standardized multivariate data point contained in the standardized sample of size n and $\|.\|$ is a vector norm. The HJG test rejects the null distribution of MVN for large values of the statistic.

3.3. Size and power comparison of the competing tests

The power of a test, which is the ability of the test to reject a wrong null hypothesis, and the size of a test, which is the maximum probability of rejecting a true null hypothesis, are among the most important properties of a test. Although they can be obtained theoretically when the true null distribution of the test statistic is known, the sizes and powers of the three specific versions of the $M(\Phi)$ statistic however are obtained empirically and compared with those of other well-known statistics in the literature. To achieve the objective of power comparison in this work, four different classes of distributions other than the multinormal distribution are identified. They are short-tailed symmetric distributions as group I; heavy-tailed symmetric distributions as group II; short-tailed asymmetric distributions as group III;

and heavy-tailed asymmetric distributions as group IV. Three distributions were considered from each of the four groups in this study and they include the following:

Group I

Standard multivariate Laplace distribution (MVL)

Products of the univariate Laplace distribution $\left(L^d(0,1)\right)$

Products of the univariate Laplace and the symmetric beta distribution $\left(L^p(0,1)\otimes B^{d-p}(1.5,1.5)\right)$

Group II

Multivariate Cauchy distribution (MVC)

Multivariate t distribution with 2 degrees of freedom $(MVt(2))$

Products of the univariate t with 5 degrees of freedom and the Cauchy distributions $\left(t^p(5)\otimes C^{d-p}(0,1)\right)$

Group III

Products of the standard exponential distribution $\left(Exp^d(1)\right)$

Products of the gamma distribution $\left(G^d(1,3)\right)$

Products of the gamma and Gumbel distributions $\left(Ga^p(1,3)\otimes Gu^{d-p}(0,1)\right)$

Group IV

Products of the Pareto distribution $\left(P^d(1,2)\right)$

Products of the standard lognormal distribution $\left(LN^d(0,1)\right)$

Products of the Weibull distribution $\left(W^d(1,2)\right)$

where p is an integar less than d.

A total of 10,000 datasets from each of the 12 distributions grouped into I-IV and the standard multivariate normal distribution were simulated in each combination of sample sizes $n = 10, 25, 50,$ and 100 and variable dimensions $d = 2$ and 5. For each of the combinations of sample size and variable dimension, the values of each of the competing statistics were calculated and the estimated power performance of each statistic was obtained as the percentage of the 10,000 simulated samples that is rejected by the statistic at 5% level of significance. The null distribution is the multivariate normal distribution. Therefore the power performances of the statistics obtained from it are the empirical probabilities of committing the error of type one, also known as the size of a test, which in this work are expected to be equal to 5%. The type-one-error rates of the competing statistics are presented in Table 2. Also, their power performances are presented in Tables 3 and 4 for sample sizes $n = 10, 25, 50,$ and 100 respectively.

From the results in Table 2, all the six tests considered showed very good control over type-one-error. This is because, none of them recorded a type-one-error of more than the 5% level of significance in any of the combinations of sample size, n and variable dimension, d. Again, while all the other tests, including the new techniques, maintained a type-one-error of 5% ($4.5\% \leq \alpha < 5.5\%$) in all the combinations of n and d considered, the HZ test maintained a conserved state (less than 5%) in all the variable dimensions considered at sample sizes up to 50. This, however, is not a disadvantage to the technique, it rather assures that the power performance of the statistic is completely devoid of the error of type-one. The

Table 2: Empirical type-I-error rate of the competing statistics

n	d	HZ	M	HJG	$M(P)$	$M(H)$	$M(BS)$
	2	2.5	4.9	5.1	4.6	5.1	5.0
10	5	0.9	5.0	5.2	4.7	5.0	4.9
	10	-	-	-	-	-	-
	2	4.2	5.2	5.2	4.8	5.0	5.2
25	5	3.3	4.9	4.5	4.9	5.0	5.2
	10	3.3	5.1	5.2	4.8	5.0	5.3
	2	4.4	5.0	5.2	5.1	4.8	4.8
50	5	4.1	4.9	5.2	5.2	4.8	4.6
	10	4.2	5.2	4.8	5.4	5.0	4.8
	2	4.7	5.0	5.2	5.0	4.9	5.0
100	5	4.7	4.9	4.7	4.6	4.9	4.8
	10	4.8	5.0	4.8	4.9	5.0	5.2
	2	4.8	5.0	5.0	5.0	4.8	4.9
150	5	4.8	5.1	4.9	4.9	5.0	5.0
	10	4.9	5.0	5.2	5.2	4.9	4.8

error rates of all the statistics considered at sample size, $n = 10$ and variable dimension, $d = 10$ are not obtained due to the fact that such a dataset is known to be singular. Based on the results in Table 2, the new phi-divergence statistics can be said to have a good control over type-one-error and hence can be recommended, at that instance, as a good technique for testing MVN of datasets.

From Tables 3 and 4, it is observed that the new statistics are generally slightly more powerful than the other competing techniques considered in this work under the alternative symmetric distributions in Table 3, especially at large sample sizes of $n > 25$. The only exception, however, is the products of the univariate Laplace and symmetric beta distributions where the HZ statistic is observed to be slightly more powerful than the rest of the techniques considered, including the new statistics. Among the three new statistics obtained from the sample phi-divergence measure statistic in (9), the $M(BS)$ generally recorded least power performance at small sample size but most powerful, together with the $M(H)$, at large sample sizes of $n \geq 25$ under these alternative symmetric distributions.

Conversely, under the asymmetric alternative distributions in groups III and IV as presented in Table 4, it is observed that the new statistics are generally slightly less powerful than the other three L^2-type statistics, especially at large sample sizes. The only exception is the products of the univariate gamma and Gumbel distributions where the new statistics are as good as the other statistics. It is, however, expected that at large sample sizes of $n > 100$, the power performances of all the competing statistics would be equal. Again, it can be seen that under these alternative distributions in groups III and IV, the $M(H)$ performed better in their powers than the other two versions of the new phi-divergence technique.

Table 3: Empirical power comparisons of competing tests for MVN under alternative symmetric distributions in groups I and II, $\alpha = 0.05$

Group	Distributions	n	d	HZ	M	HJG	$M(P)$	$M(H)$	$M(BS)$
I	MVL	10	2	16.6	27.0	24.1	26.0	24.8	21.7
	MVL		5	14.6	39.4	38.9	35.8	33.6	29.7
	MVL	25	2	50.2	57.4	49.0	52.0	60.7	62.3
	MVL		5	85.6	93.0	80.1	83.3	92.7	93.0
	MVL	50	2	82.8	82.2	71.3	74.6	86.1	91.0
	MVL		5	99.7	99.7	95.4	98.8	99.9	100.0
	MVL	100	2	98.8	92.3	89.2	91.5	98.5	92.8
	MVL		5	100.0	100.0	99.9	100.0	100.0	100.0
	$L^d(0,1)$	10	2	11.3	21.5	17.6	19.5	19.3	16.2
	$L^d(0,1)$		5	3.4	16.3	15.3	14.3	12.6	11.4
	$L^d(0,1)$	25	2	34.4	45.1	39.8	40.4	45.9	46.4
	$L^d(0,1)$		5	28.8	52.7	43.3	38.6	48.8	47.2
	$L^d(0,1)$	50	2	63.3	68.6	58.2	59.4	71.2	76.1
	$L^d(0,1)$		5	64.3	82.9	67.5	64.1	80.2	83.7
	$L^d(0,1)$	100	2	91.1	90.4	78.6	77.7	91.8	96.3
	$L^d(0,1)$		5	94.6	98.1	87.2	84.5	97.6	98.8
	$L(0,1)\otimes B(1.5,1.5)$	10	2	5.4	8.3	10.1	7.0	7.3	5.5
	$L^3(0,1)\otimes B^2(1.5,1.5)$		5	1.8	7.5	8.6	6.9	6.6	6.8
	$L(0,1)\otimes B(1.5,1.5)$	25	2	19.5	17.7	20.0	16.1	15.2	9.6
	$L^3(0,1)\otimes B^2(1.5,1.5)$		5	15.1	24.5	25.4	18.1	19.8	14.1
	$L(0,1)\otimes B(1.5,1.5)$	50	2	43.8	29.7	33.9	32.3	30.1	19.4
	$L^3(0,1)\otimes B^2(1.5,1.5)$		5	40.8	44.6	43.2	36.0	38.1	32.6
	$L(0,1)\otimes B(1.5,1.5)$	100	2	81.2	45.5	49.8	52.5	53.0	42.0
	$L^3(0,1)\otimes B^2(1.5,1.5)$		5	80.7	69.0	66.0	51.6	61.6	58.8
II	MVC	10	2	70.6	75.3	72.2	74.4	73.5	67.9
	MVC		5	50.9	73.1	82.0	67.6	70.9	67.8
	MVC	25	2	98.7	98.8	97.6	98.5	99.0	99.0
	MVC		5	99.9	100.0	99.9	99.9	100.0	100.0
	MVC	50	2	100.0	100.0	100.0	100.0	100.0	100.0
	MVC		5	100.0	100.0	100.0	100.0	100.0	100.0
	MVC	100	2	100.0	100.0	100.0	100.0	100.0	100.0
	MVC		5	100.0	100.0	100.0	100.0	100.0	100.0
	$MVt(2)$	10	2	29.4	38.7	43.8	36.7	36.3	36.1
	$MVt(2)$		5	13.7	34.5	50.8	30.5	30.0	33.3
	$MVt(2)$	25	2	71.1	81.2	79.8	78.1	79.6	83.9
	$MVt(2)$		5	89.0	96.5	96.2	92.1	95.2	96.6
	$MVt(2)$	50	2	93.4	96.7	95.8	95.3	96.7	98.7
	$MVt(2)$		5	99.6	100.0	99.9	99.8	99.9	100.0
	$MVt(2)$	100	2	99.8	99.9	99.9	99.7	99.9	100.0
	$MVt(2)$		5	100.0	100.0	100.0	100.0	100.0	100.0
	$t(5)\otimes C(0,1)$	10	2	41.9	51.0	50.6	47.3	47.4	38.8
	$t^3(5)\otimes C^2(0,1)$		5	12.3	30.3	44.1	26.6	30.5	26.1
	$t(5)\otimes C(0,1)$	25	2	88.5	88.9	86.3	86.5	88.3	87.1
	$t^3(5)\otimes C^2(0,1)$		5	90.3	95.5	94.2	90.7	93.5	90.7
	$t(5)\otimes C(0,1)$	50	2	99.2	99.2	98.4	98.6	99.1	99.3
	$t^3(5)\otimes C^2(0,1)$		5	99.9	100.0	99.8	99.6	99.9	99.9
	$t(5)\otimes C(0,1)$	100	2	100.0	100.0	99.0	100.0	100.0	100.0
	$t^3(5)\otimes C^2(0,1)$		5	100.0	100.0	100.0	100.0	100.0	100.0

Table 4: Empirical power comparisons of competing tests for MVN under alternative skewed distributions in groups III and IV, $\alpha = 0.05$

Group	Distributions	n	d	HZ	M	HJG	M(P)	$M(H)$	$M(BS)$
III	$Exp^d(1)$	10	2	35.0	32.3	37.0	32.0	30.4	25.6
	$Exp^d(1)$		5	11.4	19.3	29.6	18.8	17.7	16.6
	$Exp^d(1)$	25	2	92.4	66.0	72.4	69.9	67.8	58.5
	$Exp^d(1)$		5	92.2	77.2	76.8	69.4	71.0	66.4
	$Exp^d(1)$	50	2	96.4	88.6	94.6	90.3	91.3	88.5
	$Exp^d(1)$		5	99.9	96.6	95.1	92.9	94.6	93.9
	$Exp^d(1)$	100	2	100.0	98.9	98.0	98.8	99.4	99.4
	$Exp^d(1)$		5	100.0	100.0	99.8	99.6	99.9	99.9
	$Ga^d(1,3)$	10	2	34.6	33.1	35.5	31.5	30.3	25.0
	$Ga^d(1,3)$		5	12.2	19.1	29.0	19.3	17.3	16.1
	$Ga^d(1,3)$	25	2	92.6	65.1	71.8	69.4	68.2	59.2
	$Ga^d(1,3)$		5	92.2	76.9	75.3	68.3	71.3	65.0
	$Ga^d(1,3)$	50	2	99.9	89.0	94.2	90.9	91.1	88.1
	$Ga^d(1,3)$		5	100.0	96.5	94.8	92.5	94.4	93.6
	$Ga^d(1,3)$	100	2	100.0	99.0	99.0	99.0	99.4	99.3
	$Ga^d(1,3)$		5	100.0	99.9	99.9	99.6	99.9	99.9
	$Ga(1,3) \otimes Gu(0,1)$	10	2	10.0	23.7	26.4	20.9	22.1	16.7
	$Ga^3(1,3) \otimes Gu^2(0,1)$		5	2.3	14.1	21.7	13.7	12.7	11.3
	$Ga(1,3) \otimes Gu(0,1)$	25	2	36.4	51.0	57.5	51.5	49.0	39.5
	$Ga^3(1,3) \otimes Gu^2(0,1)$		5	28.8	63.2	63.5	53.9	56.5	49.4
	$Ga(1,3) \otimes Gu(0,1)$	50	2	67.0	74.7	84.6	75.0	75.5	67.8
	$Ga^3(1,3) \otimes Gu^2(0,1)$		5	67.7	90.6	88.5	80.4	86.1	82.3
	$Ga(1,3) \otimes Gu(0,1)$	100	2	93.2	93.5	98.7	91.3	93.4	93.0
	$Ga^3(1,3) \otimes Gu^2(0,1)$		5	95.8	99.4	99.0	96.3	98.5	98.3
IV	$Pa^d(1,2)$	10	2	73.4	67.5	68.7	65.9	65.7	57.7
	$Pa^d(1,2)$		5	50.8	57.3	74.3	55.8	56.1	52.3
	$Pa^d(1,2)$	25	2	99.9	95.7	97.8	96.9	96.6	95.2
	$Pa^d(1,2)$		5	100.0	99.5	99.5	99.1	99.4	99.2
	$Pa^d(1,2)$	50	2	100.0	99.7	100.0	99.9	99.9	99.9
	$Pa^d(1,2)$		5	100.0	100.0	100.0	100.0	100.0	100.0
	$Pa^d(1,2)$	100	2	100.0	100.0	100.0	100.0	100.0	100.0
	$Pa^d(1,2)$		5	100.0	100.0	100.0	100.0	100.0	100.0
	$LN^d(0,1)$	10	2	57.7	54.3	56.8	52.3	52.8	44.5
	$LN^d(0,1)$		5	31.0	40.1	56.2	38.9	37.4	35.9
	$LN^d(0,1)$	25	2	98.9	90.0	92.4	90.8	90.5	86.5
	$LN^d(0,1)$		5	99.7	97.6	96.7	95.5	96.5	95.5
	$LN^d(0,1)$	50	2	100.0	99.3	99.9	99.4	99.4	99.1
	$LN^d(0,1)$		5	100.0	100.0	100.0	100.0	100.0	100.0
	$LN^d(0,1)$	100	2	100.0	100.0	100.0	100.0	100.0	100.0
	$LN^d(0,1)$		5	100.0	100.0	100.0	100.0	100.0	100.0
	$W^d(1,2)$	10	2	33.7	33.0	35.8	31.3	30.2	24.9
	$W^d(1,2)$		5	11.7	19.9	28.1	19.0	17.4	15.9
	$W^d(1,2)$	25	2	72.9	65.2	73.7	69.9	68.1	58.7
	$W^d(1,2)$		5	92.4	76.5	75.6	67.4	70.8	65.7
	$W^d(1,2)$	50	2	100.0	89.1	94.8	90.6	91.4	87.9
	$W^d(1,2)$		5	100.0	96.5	95.0	92.3	94.8	93.4
	$W^d(1,2)$	100	2	100.0	98.7	99.9	98.9	99.5	99.6
	$W^d(1,2)$		5	100.0	100.0	99.9	99.6	99.9	99.9

Table 5: Mean empirical powers of competing tests for MVN, under alternative distributions, $n \geq 25$ and $\alpha = 0.05$

Distribution group	HZ	M	HJG	$M(P)$	$M(H)$	$M(BS)$
I	65.2611	66.3000	60.9889	59.5278	66.1778	64.6722
II	96.0778	97.5944	97.0444	96.6000	97.2833	97.5111
III	86.3611	84.7833	86.8611	82.7222	83.7389	80.1111
IV	97.9889	94.8778	95.8444	94.4611	94.8222	93.3667

In order to give a clearer picture of the competitive nature of the new tests, the mean empirical power performances of the six competing statistics are presented in Table 5 and it is evident from the table that the new statistics can be recommended as good tests for assessing MVN of datasets, especially at large samples as well as when the dataset is known to be symmetric.

4. Data application

In this section, the applicability of the new statistics is presented in comparison with those of the other three competing techniques. This is carried out on a set of four multivariate datasets, which are retrieved from https://openmv.net/tag/multivariate. The datasets are as follows:

Brittleness index dataset: This is a 3-component dataset, comprising of 18 observation vectors. It is obtained as measures of brittleness of plastic products produced in three parallel reactors, TK104, TK105 and TK107 as the components.

Film thickness dataset: This is a 4-component dataset obtained as thickness measurements taken at four different positions of 160 plastic films after being cut. The measurement positions which make up the data components included top right, top left, bottom right and bottom left.

Room temperature dataset: This is another 4-component dataset which is obtained as temperature measurements, in Kelvin, taken at four corners of a room. The measurement corners which form the data components included front left, front right, back left and back right and the measurements were taken 144 times, giving rise to a 144 rows by 4 columns dataset.

Solvents dataset: The solvents dataset is a 9-component dataset which consists of physical properties of a sample of 103 chemical solvents. The properties which form the data components included melting point, boiling point, dielectric, dipole moment, refractive index, ET30, density, logP and solubility.

The four datasets are tested independently for MVN using the six competing techniques considered in this study, each at 5% $\alpha-$level. The results comprise of their test statistics, critical values and decisions of either rejection or otherwise of their MVN reached by comparing the test statistics values with their corresponding critical values. They are presented in Table 6.

From the results in Table 6, all the six tests show perfect agreement in their decisions. Specifically, none of the six competing techniques could reject MVN of the brittleness in-

Table 6: Results of tests for MVN of some real-life datasets, $\alpha = 0.05$

Dataset	Test Components	HZ	M	HJG	$M(P)$	$M(H)$	$M(BS)$
Brittleness Index	Computed value of statistic	0.428886	0.018040	0.113696	0.255089	0.079972	0.095587
	Critical value	0.804912	0.055778	0.465075	0.495144	0.137537	0.154695
	Decision	Do not reject	Do not reject	Do not reject	Do not reject	Do not reject	Do not reject
Film Thickness	Computed value of statistic	0.992884	0.000945	2.477551	0.168497	0.036121	0.032316
	Critical value	1.009868	0.002093	3.076987	0.202478	0.045625	0.045660
	Decision	Do not reject	Do not reject	Do not reject	Do not reject	Do not reject	Do not reject
Room Temperature	Computed value of statistic	2.029179	0.025055	2134.348	0.507064	0.079452	0.053684
	Critical value	1.008710	0.002518	3.051355	0.212408	0.047989	0.048165
	Decision	Reject	Reject	Reject	Reject	Reject	Reject
Solvents	Computed value of statistic	1.750361	0.108639	8604.996	0.351212	0.104571	0.107069
	Critical value	0.997587	0.006947	9.441462	0.203694	0.053017	0.056799
	Decision	Reject	Reject	Reject	Reject	Reject	Reject

dex and film thickness data while, on the other hand, they all rejected the MVN of room temperature and solvents datasets. The result in this section further shows that the three new statistics can be regarded as good statistics for testing MVN of datasets.

5. Conclusion

The plug-in techniques developed in this study for assessing MVN of multivariate datasets have shown, through their size and power performances, that they can be regarded as good statistics. Their affine invariance and consistency properties have been proved. Also, the statistics can be adapted for goodness-of-fit test to other continuous distributions. Besides good power performances, the new statistics are computationally less tedious since they are based on univariate transform of multivariate datasets. Finally, it is not difficult to implement the new statistics developed in this paper to statistical software such as R so that users can access them for applicability to real-life situations. As a result, they are recommended as good techniques for testing normality of d-dimensional datasets, $d \geq 1$.

Acknowledgements

The authors wish to thank the editorial team and the anonymous reviewers for their useful suggestions, which have improved the quality of the paper.

References

Baringhaus, L., Henze, N., (1988). A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35, pp. 339–348.

Bayoud, H. A., (2021). Tests of normality: new test and comparative study. *Communications in Statistics - Simulation and Computation*, 50, pp. 4442–4463.

Bilodeau, M., Brenner, D., (1999). *Theory of multivariate statistics*. New York: Springer-Verlag.

Cardoso de Oliveira, I. R. C. and Ferreira, D. F., (2010). Multivariate extension of chi-squared univariate normality test. *Journal of Statistical Computation and Simulation*, 80, pp. 513–525.

Chen, W., Genton, M. G., (2023). Are you all normal? It depends! *International Statistical Review*, 91, pp. 114–139.

Csorgo, S. (1989). Consistency of some tests for multivariate normality. *Metrika*, 36, pp. 107–116.

Doornik, J. A., Hansen, H., (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, 70, pp. 927–939.

Dorr, P., Ebner, B. and Henze, N., (2020). Testing multivariate normality by zeros of the harmonic oscillator in characteristic function spaces. *Scandinavian Journal of Statistics*, doi.org/10.1111/sjos.12477.

Dorr, P., Ebner, B. and Henze, N., (2020). A new test of multivariate normality by a double estimation in a characterizing PDE. *Metrika*, doi.org/10.1007/s00184-020-00795-x.

Ebner, B., Henze, N., (2020). Tests for multivariate normality – a critical review with emphasis on weighted L^2-statistics. *Test*, 29, pp. 847–892.

Epps, T. W., Pulley, L. B., (1983). A test for normality based on the empirical characteristic function. *Biometrika*, 70, pp. 723–726.

Gnanadesikan, R., Kettenring, J. R., (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28, pp. 81–124. doi:10.2307/2528963.

Hanusz, Z., Tarasinska, J., (2012). New tests for multivariate normality based on Small's and Srivastava's graphical methods. *Journal of Statistical Computation and Simulation*, 82, pp. 1743–1752.

Healy, M. J. R., (1968). Multivariate normal plotting. *Applied Statistics*, 17 (2), pp. 157–161. doi:10.2307/2985678.

Henze, N., (2002). Invariant tests for multivariate normality: a critical review. *Statistical Papers*, 43, pp. 467–506.

Henze, N., Jimenez-Gamero, M. D., (2019). A new class of tests for multinormality with i.i.d. and GARCH data based on the empirical moment generating function. *Test*, 28, pp. 499–521.

Henze, N., Jimenez-Gamero, M. D. and Meintanis, S. G., (2019). Characterizations of multinormality and and corresponding tests of fit, including for GARCH models. *Economic Theory*, 35, pp. 510–546.

Henze, N., Visagie, J., (2020). Testing for normality in any dimension based on a partial differential equation involving the moment generating function. *Annals of the Institute of Statistical Mathematics*, 72, pp. 1109–1136.

Henze, N., Wagner, T., (1997). A new approach to the BHEP tests for multivariate normality. *Journal of Multivariate Analysis*, 62, pp. 1–23.

Henze, N., Zirkler, B., (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics - Theory and Methods*, 19, pp. 3595–3618.

Hwu, T., Han, C. and Rogers, K. J., (2002). The combination test for multivariate normality. *Journal of Statistical Computation and Simulation*, 72, pp. 379–390.

Jarque, C. M., Bera, A. K., (1987). A test for normality of observations and regression residuals. *International Statistical Review*, 55, pp. 163–172.

Korkmaz, S., Göksülük, D. and Zararsiz, G., (2014). MVN: An R package for assessing multivariate normality. *R Journal*, 6(2), pp. 151–162.

Liang, J., Fang, M. L. and Chang, P. S., (2009). A generalized Shapiro-Wilk W statistic for testing high – dimensional normality. *Computational Statistics & Data Analysis*, 53, pp. 3883–3891.

Lin, J., (1991). Divergence measures based on the Shannon entropy. *Information Theory–IEEE Transactions on Reliability*, 37 (1), pp. 145–151.

Madukaife, M. S., (2017). A new affine invariant test for multivariate normality based on beta probability plots. *Journal of the Nigerian Statistical Association*, 29, pp. 58–70.

Madukaife, M. S., Okafor, F. C., (2018). A powerful affine invariant test for multivariate normality based on interpoint distances of principal components. *Communications in Statistics - Simulation and Computation*, 47, pp. 1264–1275.

Madukaife, M. S., Okafor, F. C., (2019). A new large sample goodness of fit test for multivariate normality based on chi squared probability plots. *Communications in Statistics-Simulation and Computation*, 48(6), pp. 1651–1664.

Malkovich, J. F., Afifi, A. A., (1973). On tests for multivariate normality. *Journal of the American Statistical Association*, 68(341), pp. 176–179.

Mardia, K. V., (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 573, pp. 519–530.

Mardia, K. V., (1974). Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies. *Sankhya*, 36, pp. 115–128.

Mecklin, C. J., Mundfrom, D. J., (2004). An appraisal and bibliography of tests for multivariate normality. *International Statistical Review*, 72, pp. 123–138.

Pearson, K., (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,". *Philosophical Magazine*, 5th Series 50, pp. 157–175.

Pudelko, J., (2005). On a new affine invariant and consistent test for multivariate normality. *Probability and Mathematical Statistics*, 25, pp. 43–54.

Romeu, J. L., Ozturk, A., (1993). A comparative study of goodness-of-fit tests for multivariate normality. *Journal of Multivariate Analysis*, 46(2), pp. 309–334.

Royston, J. P., (1983). Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 32(2), pp. 121–133.

Shannon, C. E., (1948). A mathematical theory of communications. *Bell System Technical Journal*, 27, pp. 379–423, 623–656. doi:10.1002/bltj.1948.27.issue-3.

Shapiro, S. S., Wilk, M. B., (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52 (3 and 4), pp. 591–611. doi:10.1093/biomet/52.3-4.591.

Singh, A., (1993). Omnibus robust procedures for assessment of multivariate normality and detection of multivariate outlier, in Multivariate Environmental Statistics, G.P. Patil and C.R. Rao eds., Amsterdam: North Holland.

Small, N. J. H., (1978). Plotting squared radii. *Biometrika*, 65 (3), pp. 657–658.

Srivastava, M. S., (1984). A measure of skewness and kurtosis and a graphical method for assessing multivariate normality. *Statistics and Probability Letters*, 2, pp. 263–267.

Szekely, G. J., Rizzo, M. L., (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, 93, pp. 58–80.

Tavakoli, M., Alizadeh Noughabi, H. and Borzadaran, G. R. M., (2020). An estimation of phi divergence and its application in testing normality. *Hacettepe Journal of Mathematics and Statistics*, 49 (6), pp. 2104-2118.

Tavakoli, M., Arghami, N. and Abbasnejad, M., (2019). A goodness of fit test for normality based on Balakrishnan-Sanghvi information. *Journal of the Iranian Statistical Society*, 18 (1), pp. 177–190.

Tenreiro, C., (2017). A new test for multivariate normality by combining extreme and nonextreme BHEP tests. *Communications in Statistics - Theory and Methods*, 46, pp. 1746–1759.

Thulin, M., (2014). Tests for multivariate normality based on canonical correlations. *Statistical Methods & Applications*, doi:10.1007/s10260-013-0252-5.

Vasicek, O., (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society B*, 38, pp. 54–59.

Villasenor, J. A., Gonzalez-Estrada, E., (2009). A generalization of Shapiro-Wilks test for multivariate normality. *Communications in Statistics-Theory and Methods*, 38 (11), pp. 1870–1883.

Wieczorkowski, R., Grzegorzewsky, P., (1999). Entropy estimators improvements and comparisons. *Communication in Statistics Simulation and Computation*, 28(2), pp. 541–567. doi:10.1080/03610919908813564.

Zhou, M., Shao, Y., (2014). A powerful test for multivariate normality. *Journal of Applied Statistics*, 41, pp. 1–13.

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 151–175, https://doi.org/10.59139/stattrans-2025-020
Received – 18.07.2024; accepted – 27.02.2025

Households' invisible input to the economy: a review of its measurement methods and results

Marta Marszałek[1]

Abstract

Unpaid domestic work is the main part of non-market household production which is not covered by national statistics (GDP). The monetary value of unpaid work is identified within the gross value added (GVA), which is 60–80% of (the invisible) non-market household production. GVA of unpaid work provides significant information about the household sector and its impact on the national economy even if some part of that production is unobserved. After long discussions, a consensus was achieved and the input method was approved and used in the estimates of unpaid work in the Household Satellite Accounts (HHSA). However, the consensus is still in the process of household estimations. This paper shows that different wages used in input methods do not change the final proportion of the GVA of unpaid work to total household production. The analysis also confirms that, in accordance with UNECE and Eurostat, a regular implementation of the HHSA alongside the core system – the European System of Accounts – is a valuable and comprehensive tool for assessing the total output of household production (both market and non-market).

Key words: household production, input method, unpaid work, satellite accounts, time use survey.
JEL: D13, J10, J13, J16.

1. Introduction

The household is a dynamic economic unit where the division of responsibilities plays a key role in everyday functioning. Everyone is involved in the household structure, even if it is an individual or collective household. In the System of National Accounts (SNA) and European System of Accounts (ESA) households are treated as an institutional sector, which is one of 5 sectors in the economy. Therefore, the household is a special sector which covers the whole territory of the country. Any other sector does not include each unit of the category. In the household, regardless of whether it is family

[1] Department of Applied Statistics, SGH Warsaw School of Economics, Warsaw, Poland.
E-mail: mmars1@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-6810-7977.

or non-family, one- or multi-person, individual or collective, people share their duties to provide all basic functions to realize individual or group (social) needs.

For ages, people have allocated their time to activities that can be classified as paid work, unpaid work, and no work (personal care, leisure time). "No work time" is more various than paid and unpaid work because it can be divided into different groups of activities, e.g. personal care, leisure time, work as a volunteer (without money), help for others (neighbors, family members in other households).

Historical research of time distribution was carried out in England and France at the end of 19th century, based on analysis of 'duty time'. The duty time covered only the time spent on paid work (Poissonnier and Roy 2013; Soinne 2021; Szép 2003; Varjonen and Aalto 2010, 2006; Varjonen and Hamunen 2014; 1999; Błaszczak-Przybycińska and Marszałek 2021, 2020, 2015). The analyses were provided to register the time spent in factory by employees because the workers demanded shortening working hours and increasing the salaries (Błaszczak-Przybycińska 2020, 2008).

Housework was separated in a different group of activities in the late 1950s. The first time use surveys were conducted to the analyses of time distribution in households. In Polish, research of time budgets showed the analogy between increasing of efficiency in paid work as well as in housework. Since the 1960s, a prior assumption of the substantial amount of research has been considered to undertake the differences in the socio-economic status between men and women. At this time in United Nations, many governments debated to eliminate all forms of discrimination against sex, nationality.

For decades, scientific findings have pointed out that gender disparities in paid and unpaid work, is a contributing factor to promoting not only gender inequality, but also economic growth and development. Time use surveys and current scientific research provide valuable guidance regarding the balance and satisfaction associated with this allocation.

In this article, we will focus on the analysis of methods and different approaches to estimating the value of unpaid work in order to present the non-market household production, the most invisible part of productive results which are made for own use at home. Household production (both market and non-market) is an important component of total economic output to growth and development in national and international perspective. This study is drawing on research conducted priory in European countries, with the international context of the household economy. Our results are consistent with previous works about regular compilation of satellite accounts which are the most appropriate statistical tool to better understand the overview of social and economic condition of households.

The article is organized as follows. We begin with the theoretical framework of Becker's model of time allocation as a foundation to further statistical models and estimations of unpaid work and non-market production which is invisible in official

statistics. Conventional economic statistics, such as national accounts, are supplemented with time use survey data, which provides an economic perspective of unpaid domestic work estimation. The monetary value of unpaid work is identified with the gross value added (GVA), which is the significant part of the household production, and covers 60-80% of that production. In the paper, methodological differences between various approaches of calculation were also described based on current analyses provided in European countries. The final results of estimates confirmed that the input method delivers the most accurate data to the international comparisons and the core national accounts.

2. Time allocation model and time distribution research

2.1. Time use survey

The theoretical background of time use research had its beginning in 18th century (Luszniewicz 1982). Before that time any household duties were not observed or noted as productive activities realized at home for all household's members.

The breakthrough for research of time distribution was Becker's theories of the time allocation model and dual role of household: production and consumption. The theoretical and practical foundations for Becker's classic theory of time allocation took place in other analytical studies. Mitchell (1912) claimed that if households are compared to a company which produces goods or services for the market selling, the households are insufficient in producing domestic services. Reid (1934) recognized in the early 1930s that both paid and unpaid work should be treated as total household production which generates comprehensive overview of household productive activities. She also underlined that national economy does not cover the important components of production, unpaid work and service work as a main part of household production. Kuznets (1934) also confirmed that system of national account (SNA), which recommend the structure and calculations of national income, GDP and macroeconomic indicators, omitted significant part of household production.

While statistical studies on Becker's model (Becker 1965) were performed and developed (Kuznets 1934; Gorman 1959; Gronau 1977, 1986, and 1997; Graham and Green 1984; Koreman and Kapteyn 1987; Heckman 1988 and 2015; Fitzgerald 1996), economists discussed on GDP limitations in describing the socio-economic development (Stiglitz et al. 2009; Folbre 2006; Gershuny 2005 and 2000). The national official statistics do not contain important effects of economic production which was made outside the market, in households. Economists agreed that household products (goods and services) made by themselves have the important economic value and enormous quantity even if they are not registered in macroeconomic indicators such as GDP, national income, value added.

Also, United Nations Economic Commission for Europe (UNECE) identified unpaid work as one of the most informative area, with other sources not providing sufficient data (UNECE 2017). A lack of information on domestic household service work might lead to inadequate policy conclusions. An increase of childcare or long-term care services available in the private or public (governmental) sector, generates also the increase in gross value added of goods and services in this area. It reflects a growth of production, but it also illustrates the shifts from household sector to the market. UNECE provided the guideline for valuing own-use household work of services, and methods to compile the Household Satellite Accounts and HETUS recommendations for European countries.

European Union countries carry out the time use survey regularly in harmonized waves. HETUS 2000 (Harmonised Time Use Survey) was the first round, which was conducted between 1998 and 2006 in 15–EU countries. HETUS 2010 gathered data between 2008 and 2015 in 18–EU countries. HETUS 2020 is ongoing round in which 20 countries plan to conduct the survey. The final collection of microdata is planned before 2027. Harmonization of methodological guidelines standardize the survey design, structure, content, statistical classifications, timing, frequency, but some local differences remain.

The monetary value of unpaid work, as a main component of household production, is measurable and countable if we use it in unconventional economic statistics, e.g. sample surveys, such as time use survey (TUS). Time use survey provides the information about time distribution in households. It also collects various data about socio-economic status of households, their demographical structure, and overview of daily schedule of activities. Respondents register each single activity in 10-minute periods during 24 hours by 2 days (a day: from Monday-Friday, and one weekday: Saturday or Sunday). Those activities were gathered into 10 different groups, such as: personal care (sleeping, eating, washing, dressing), paid work, household and family care (unpaid work), study, voluntary work, social life and entertainment, sports, hobbies, mass media, travel.

2.2. Value of domestic labor

When we begin to compare unpaid work estimates, we start from the point whether unpaid work is economic work or non-economic work. Pigou noted that "if a man marries his housekeeper or his cook, the national dividend is diminished" (Pigou 1920). If someone hires the housekeeper, employer will pay the salary for her or him, so this transaction will be able visible and reflected in GDP. In other case, if the same person realizes the same productive activities without remuneration, it will not be reflected in economic measures and indicators.

Currently, towards to SNA 2025, the update program of SNA 2008 for national statistics, efforts are underway to develop new guidelines tailored to the market and economic context, with the aim of permanently integrating the valuation of unpaid work and household production into national accounts[2].

According to European System of Accounts 2010 (legal act for EU members), which is based on System of National Accounts (SNA 2008), it provides the conceptual framework that sets the international statistical standard for the measurement and classification of economic activities, and economic aggregates such as Gross Domestic Product or Gross National Income. ESA 2010 also indicates the household work activities, which are deemed "market" or "economic work", e.g. paid work. Market work is included in "SNA production boundary". Other unpaid work activities are classified as "non-market" or "non-economic" so they are "invisible" and missed in official measures and indicators in the economy (Table 1.).

Table 1: The overlap of paid and unpaid work in SNA/non-SNA work

SNA work (production boundary) Visible, recorded in GDP	1. Paid work (for the market	2. Unpaid work (for the market): (a) owner occupiers' imputed rents (housing services of equivalent rented accommodation); (b) own-account house constructions; (c) paid domestic staff; (d) agricultural production for own use (hunting, fishing, picking berries and mushrooms); (e) collection of raw materials for income generating activities like handicrafts, and other manufacturing	3. Unpaid work for the household (non-market)
Non-SNA work (outside the production boundary) Invisible in GDP			D. Unpaid work (non-market, household maintenance, care work, and volunteer work)

Source: own work based on Eurostat 2013.

[2] Towards the 2025 SNA, https://unstats.un.org/unsd/nationalaccount/towards2025.asp (accessed: 03/02/2025).

ESA 2010, as well as previous version ESA 1995, indicates that some unpaid domestic activities are treated as economic because they are measured and included in annual estimates of GDP. These comprised of: (a) owner occupiers' imputed rents (housing services of equivalent rented accommodation); (b) own-account house constructions; (c) paid domestic staff, (d) agricultural production for own use (hunting, fishing, picking berries and mushrooms), and (e) collection of raw materials for income generating activities like handicrafts, and other manufacturing. Accordingly, unpaid economic work consists of activities producing for own use, as well as for the market. In practice, data collection, identification and classification of each type of household unpaid productive activity in Gross National Income (GNI) and GDP is very difficult (Table 1.).

Other types of unpaid domestic work are deemed by the SNA 2008 (and previous SNA 1993) to be "non-economic", treated also as "invisible" in GDP, and are relegated "outside the SNA production boundary". Non-SNA household work consists of 5 groups of productive activities: (1) housing services (cleaning, repairing, and other maintenance), (2) food preparation (cooking, dish-washing, cleaning, shopping), (3) providing and repairing clothes, (4) care work for infants, children (active and passive care), care for dependent people (ill or temporarily sick, elder and disabled), and (5) all volunteer work for family members living in separate household and other community services.

The Eurostat (ESA and SNA) and the UNECE proposals recommend to compile the parallel (satellite) accounts as a supporting and comprehensive to the national accounts. Household Production Satellite Account (Household Satellite Account, HPSA, HHSA) is an additional account which provide the information of SNA/non-SNA household production. Table 1. presents a composition of relationship between paid and unpaid work and SNA/non-SNA production boundaries. To briefly sum up, work is unpaid in 2. and 3. cell. The 2. unpaid work is registered in GDP, because it is produced to the market, but 3. unpaid work is produced and consumed by households themselves, so it is produced for their own use, not to the market.

The estimation of unpaid household work can be done using two approaches: *the input method* and *the output method*. Both methods, *input* and *output* provide the sufficient information about housework. *The input method* is more often applied to estimating the value of unpaid work (Varjonen et al. 2014). Two different approaches are used: (1.) *the replacement cost* and (2.) *the opportunity (alternative) cost* (Figure 1.).

The replacement cost approach provides three options: (1a) is to use the wages of specialized workers in market enterprises. It can be reasoned that specialized workers in certain occupations perform similar activities to those done in households, e.g. a cook in a restaurant, a teacher at school, a task manager at an enterprise, etc. The difficulties start when we consider the productivity and working conditions in market

enterprises, which are different from those prevailing in the household, e.g. capital investment is higher, production is organized according to specialization of skills, task (mass production). It is also difficult to choose the adequate level of qualification of the jobs in the market (variety is large, e.g. from the chef de cuisine to the kitchen maid or trainee). In housework several tasks are performed simultaneously: the main activity (e.g. childcare) and second or third activities (cooking, cleaning), whereas in enterprises work may be more like line production (Goldschmidt-Clermont 1994).

The second option (1b) is to apply the wages of specialized workers at home. One can purchase the services of a specialized worker who comes to work in a household as a cleaner, window cleaner, plumber, gardener, private teacher, nurse, dog walker, etc. Workers who come to the home may use tools and materials of their own or those available in the household. The working conditions come closer to those in housework, except that these specialized workers focus on one task at a time. The payments by households to these specialized workers, however, are higher than the wages for workers in enterprises because the former include also other costs than just wages. This aspect must be taken into account in measuring household work and production (Eurostat 2003).

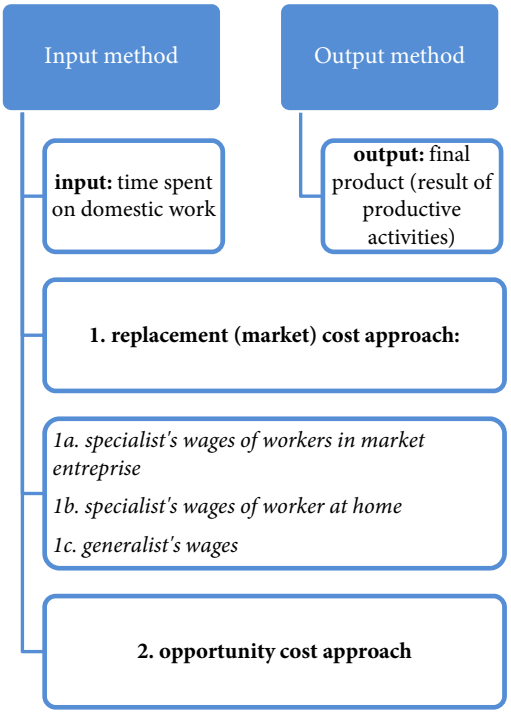


Figure 1: Domestic work estimation methods

Source: own compilation.

The third possibility (1c) is to apply the wages of generalist workers. Household requires may be provided by professional paid staff. These are workers who may or may not have received a special qualification for their job and who are most often responsible for supporting and visiting elderly people or helping when a parent caring for home and children are ill. However, the paid domestic staff usually provide some household tasks, except household management, volunteer or community work (Goldschmidt-Clermont 2000 and 1994).

The opportunity cost method (2.) is complicated to apply because the problem concerns the different values of similar products depending on who did the housework. Many researchers argue that this approach should not be used to value household production (e.g. Goldschmidt-Clermont 1994, 2000; Blades 1997). This method may be relevant for the recognition of utility maximation at the micro-level concepts of the core (national) accounts.

The precedent assumption in the replacement (market) cost method is that housework done by household members could be realized by a hired person, referring to the "third party criterion" or "productivity criterion", which provides the distinguishing between productive and unproductive household activities (Eurostat, 1999). Therefore, each type of domestic work can be valuated using the market cost of parallel services. The replacement cost method can be based on generalist's or specialist's wages (Varjonen et al. 2014). The generalist's wages is an easier approach to estimate the value of unpaid work, because it includes all kinds of domestic work that is done in households. The average wage per hour, which is used to estimate the value added of unpaid domestic work, provide the simplified final results. The specialised worker's wages need consideration in defining the standards for the work done at home. It requires the decisions of, e.g. whether to use the wage of cook or kitchen helper to calculate the value of preparing food, planning menu, deciding ingredients or making up the dishes in households (Soinne 2021, Varjonen et al 2014).

The specialist's wages approach entails multiplying the average time spent on housework by the average hourly wage for professions and specialists related to household activities. The total value of domestic work is reflected in aggregation of the values estimated for different types of activities.

In the empirical results and in the literature dozens of detailed activities are gathered into 4 main groups of unpaid work, such as household upkeep, food preparation, making and caring for textiles (clothes and shoes), childcare and adult care. According to Eurostat recommendations also several activities related to volunteer work were taken into account in the estimation: unpaid help to other households, neighborly help, informal work in organizations, etc. (Eurostat 1999).

The average hourly wages of professions correspond to the services purchased by households on the market. If household members buy meals (lunch), they pay for the

total service (wages of canteen staff, cooks, waiters, etc.). The parallel grouping of activities in households and proper market wages were adjusted to each type of domestic work (Eurostat, 1999).

The daily, weekly and monthly formulas to estimate housework value by sex, activity on the labor market, family status by number of children, the level of education were calculated separately for all selected groups of respondents (Błaszczak-Przybycińska 2007, Varjonen et al 2014).

Daily value of domestic labor[3]

$$ {}_F\bar{t}^z{}_{laj} = \frac{\sum_{i=1}^{n_1} {}_F t^z_{ilaj}}{n_F} \tag{4} $$

$$ {}_M\bar{t}^z{}_{laj} = \frac{\sum_{i=1}^{n_2} {}_M t^z_{ilaj}}{n_M} \tag{5} $$

where:

- ${}_F\bar{t}^z{}_{laj}$ – duration of the *a-th* activity in the *j-th* group for *i-th* women from the *l-th* class in the *z-th* day of the week,
- n_F – the number of women in a subsample,
- ${}_M\bar{t}^z{}_{laj}$ – duration of the *a-th* activity in the *j-th* group for *i-th* men from the *l-th* class in the *z-th* day of the week,
- n_M – the number of men in a subsample,
- z – the day of the week; $z = 1,2,3$, where: 1 – Monday-Friday, 2 – Saturday, 3 – Sunday,
- j – group of domestic activities, $j = 1,2,3,4,5$.

*Weekly value of domestic labor***

$$ {}_F\bar{t}_{laj} = [\frac{5}{7}\,{}_F\bar{t}^1{}_{laj} + \frac{1}{7}({}_F\bar{t}^2{}_{laj} + {}_F\bar{t}^3{}_{laj})] * 7 \tag{6} $$

$$ {}_M\bar{t}_{laj} = [\frac{5}{7}\,{}_M\bar{t}^1{}_{laj} + \frac{1}{7}({}_M\bar{t}^2{}_{laj} + {}_M\bar{t}^3{}_{laj})] * 7 \tag{7} $$

where:

- ${}_F\bar{t}_{laj}$, ${}_M\bar{t}_{laj}$ – the average week time duration of the *a-th* activity in the *j-th* group for the *i-th* women and men from the *l-th* class,
- ${}_F\bar{t}^1{}_{laj}$, ${}_M\bar{t}^1{}_{laj}$ – the average duration of the *a-th* activity in the *j-th* group for women and men from the *l-th* class (weekdays from Monday to Friday),
- ${}_F\bar{t}^2{}_{laj}$, ${}_M\bar{t}^2{}_{laj}$ – the average duration of the *a-th* activity in the *j-th* group for women and men from the *l-th* class (on Saturdays),
- ${}_F\bar{t}^3{}_{laj}$, ${}_M\bar{t}^3{}_{laj}$ – the average duration of the *a-th* activity in the *j-th* group for women and men from the *l-th* class (on Sundays).

[3] *, **, *** – formulas are based on Błaszczak-Przybycińska 2007, Błaszczak-Przybycińska and Marszałek 2020, 2019.

*Monthly value of domestic labor****

$$ {}_{F}H_{l} = \frac{52}{12}\sum_{j=1}^{4}\sum_{a=1}^{n_a} {}_{F}\bar{t}_{laj}s_{aj} \quad (8) $$

$$ {}_{M}H_{l} = \frac{52}{12}\sum_{j=1}^{4}\sum_{a=1}^{n_a} {}_{M}\bar{t}_{laj}s_{aj} \quad (9) $$

where:

- ${}_{F}H_{l}$ – monthly housework value for a woman from the *l-th* class,
- ${}_{M}H_{l}$ – monthly housework value for a man from the *l-th* class,
- ${}_{F}\bar{t}_{laj}$ – average duration per week of the *a-th* activity in the *j-th* group for women from the *l-th* class,
- ${}_{M}\bar{t}_{laj}$ – average duration per week of the *a-th* activity in the *j-th* group for men from the *l-th* class,
- s_{aj} – hourly wage calculated for the *a-th* activity in the *j-th* group.

The International Standard Classification of Occupations (ISCO-08), which is applied in most countries, can be useful in defining the wages of professions or specialists. Whether we consider the generalist's wages according to ISCO-08 codes is proper to use wages of housekeepers (9111 – Domestic Cleaners and Helpers[4]). In Finland home-helpers or housekeepers wages are available (based on ISCO-88). The problems of housekeepers, who are employed by private households, are also related to "black market activities", which means that data and statistics on wages are not available.

Monetary value of unpaid domestic work is the most important part of the total non-SNA/non-market household production. The unpaid domestic work identified with *the gross value added in households* is calculated at 60–80% of non-market household production depending on the level of national development and social factors, e.g. tradition, gender. In more traditional societies the value of unpaid work will be higher because of providing care of elderly person or infants and children aged 6 and less by themselves at home.

Results from many different studies in Finland (Soinne 2021 and Varjonen et al. 2014, 2010, 2006, 1999), France (Poissonnier and Roy 2013), Germany (Schäfer 2004), Hungary (Szép 2003), Poland (Błaszczak-Przybycińska and Marszałek 2021, 2020, 2019), and Spain (Duran 2007) present that *the generalist's* and *specialist's wages* deliver equivalent estimates.

2.3. Consumption: intermediate consumption

According to ESA 2010, the intermediate consumption means that it was recorded as "completely used in the production process" at the end of the period (ESA 2010).

[4] Internet: UE ISCO-08 codes: https://esco.ec.europa.eu/pl/classification/occupation_main#overlayspin (accessed: 07/04/2025) and International Labor Organization ISCO-8 classification: https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/ (accessed: 07/04/2025).

In the Household Satellite Account the intermediate consumption was calculated separately for each group of household unpaid work. The monthly household's expenditures for different goods and services were selected and grouped into the same categories as unpaid work. The data of expenditures was obtained in household budget survey (HBS).

The intermediate consumption in the HHSA is calculated according to formula (10):

$$IC = i_{pop} \sum \bar{e}_a \tag{10}$$

where:

IC – total intermediate consumption in households for people aged 15 years and more;

i_{pop} – index of population for people aged 15 years and more (without people with disability),

$\sum \bar{e}_a$ – yearly sum of average expenditures on consumption for *a-group* of activities (household upkeep, food preparation, making and caring for textiles (clothes and shoes), childcare and adult care, volunteer work).

The intermediate consumption is an important part of consumption, because the products which are used in the production process are changed into new products (good or service). Households decide which products are treated as final products, and which take part indirect in the production process.

The second type of consumption is final consumption, which means the actual use of a product: wearing clothes or eating food. These products are used directly with their intended use.

2.4. Capital

Capital services made in household consist of two items: *consumption of fixed capital* (depreciation of cars, machinery and other equipment), and interest corresponding to the acquisition of capital. The Household Satellite Account contain only the consumption of fixed capital (Marszałek 2015, Varjonen and Aalto 2006, Eurostat 2003).

Varjonen and Alto indicated that total output of domestic services is increased by the consumption of fixed capital which reflects depreciation of home equipment, furnishing.

Polish HHSA 2011 presents the estimation of consumption of fixed capital in accordance with below formulas (Marszałek 2015):

$$C = \sum c_i * hh * hh_{xi} * d_{xi} * p_{xi} \tag{11}$$

where:

c_i – yearly capital consumption for people aged 15 years and more in *i-class,*

hh – number of households,

hh_{xi} – percentage of households having *x-good* in *i-class,*
d_i – percentage of depreciation *x-good* in *i-class,*
p_i – average price of *x-good* in *i-class*.

More appropriate for comparisons with national statistics (GDP, GVA) is *the output method*. The output method assumes that only results of productive activities which can be observed as the market final products (goods or services) are registered in the estimation of non-market household production. The crucial problem is how to define the results of each kind of activity. For example, the final product of cooking is meal, cleaning – clean apartment, ironing – ironed clothes. But what is the product of childcare or elder care, management of the household, preparing menu and ingredients for cooking, etc.? The findings are ambiguous and more expensive than the input method because they need more detailed surveys of final products in different fields of household unpaid work. The output method was applied in UK's Household Production Satellite Account that was dismissed after the experimental estimation of accounts in 2002 (Holloway, Short and Tamplin 2002).

3. Analysis and results

The analysis presents the most significant aggregates of Finnish and Polish Household Satellite Accounts. In Finland, the average *generalist's wage* of average monthly salary was used to calculate the value of unpaid work. In Poland, the gross value added of domestic work was estimated based on the *specialist's wages*[5] according to different types of activities and corresponding market wages per hour. The average wages per hour, used in Finnish estimates, amounted 13.44 EUR (in 2009), 14.53 EUR (2012), and 15.36 EUR (2016). These wages constituted 72.8%, 73% and 72.9% of average hourly wages by employer sector[6]. Also in Polish analyses the adequate proportion of professional wages (69.8%, 66.8% and 72.0%) was applied to estimate the unpaid household work by various approaches to consider which method is more accurate to the final estimation (Błaszczak-Przybycińska and Putkowska 2024).

The most important aggregates present *the gross value added*, *total household production* (SNA + non-SNA household production), and the share of household production in the national economy (and in the expanded GDP). According to data obtained from 2 waves of Harmonized European Time Use Survey[7] (2001 and 2009 for

[5] The different approaches were used in that analysis because various data and different calculation methods are available per country. Both approaches may be applied – *the generalist's wages* and *the specialist's wages* – as both are defined as comparable *input methods* (see Figure 1).

[6] Estimates are based on StatFin/Index of wage and salary earnings / 122k – Average monthly earnings by sector and gender, 2000-2023.

[7] So far, Eurostat conducted 2 waves of Harmonised Time Use Survey: HETUS 2000 (round 1, 1998–2006) conducted in 15 EU countries; HETUS 2010 (round 2, 2008–2015) conducted in 15 EU countries and 3 non-EU countries: Norway, Serbia and Turkey). The current wave of the European TUS is ongoing and has started in 2020.

Finland; 2003/2004 and 2013 for Poland), the gross value added of household production (market and non-market household production) in Finland 2009 amounted to more than 78 billion EUR, in Poland 2011 achieved 192 billion EUR (Table 2. and 3.). Of this, national accounts recognized almost 12.6 billion EUR in Finland, and 31 billion EUR in Poland. The remaining 65.8 billion EUR was excluded from national accounts in Finland, and 180.9 billion EUR in Poland 2011. This sum of household production would increase GDP by 39.9% (Finland), and 45.2% (Poland). The total amount of domestic labor in both countries depends on population. The population size in Poland is near 7 times larger than in Finland, so not monetary value but share of macroeconomic indicators might be compared in both countries. The wages and structure of non-market household production could be also compiled.

The gross value of domestic unpaid work is a major part of the non-market household production, and it achieved from 60.2% to 61.4% in Finland, and 58.1% to 58.7% in Poland. The similar proportion of the results confirm that the estimation using input method guarantee the international comparisons even of different approaches (the generalist's or specialist's wages) which were used in calculations.

In Finland, from 2009 to 2012, the gross value added (GVA) of household production at market prices increased by 9.7%, and from 2012 to 2016 increased by 19%. For Poland: 2011 vs. 2013 recorded the 4.9% decline and from 2013 to 2016 the 6.7% increase. GVA is the most informative measure of household impact on the economy. It reflects the monetary value of caregiving activities realized at home for children and other dependent people. Even if some part of care work was noted as secondary work in Time Use Survey and not fully valued in HPSA, then the total overview of proportion in market price of that work regular increases for Finland (in billion EUR): 65.8 in 2009, 73.1 in 2012, 78.8 in 2016, and for Poland (billion EUR): 31.1 in 2011, 31.6 in 2016 (Table 2. and 3.).

Furthermore, the value of domestic work related to childcare increased not only due to annual wage and inflation growth but primarily because of social recognition of this type of work as important for societal development. In previous decades (30–40 years ago), these tasks were not particularly appreciated, also because Poland had a relatively high birth rate and fertility rate (Szałtys and Cierniak-Piotrowska 2022).

Growing awareness of the importance of childcare in the context of social and economic development is evident. Currently, childcare focuses more on shaping attitudes, nurturing emotional development, enhancing children's potential, promoting educational and cognitive growth, and providing health services – particularly preventive care – rather than solely meeting basic and caregiving needs such as food provision, housing, and essential necessities.

The most interesting result of that analysis is the percentage of non-market household production in relation to GDP. In Finland, we observed almost stabile

relation: 39.9% in 2009, 40.1% in 2012 and 39.8% in 2016. For Poland, it was estimated at 45.2% in 2011, 44.4% in 2013 and 45.1% in 2016. This confirms that the domestic work made for own use at home without any market payment is a huge invisible productive resource. If we consider the concept of expanded GDP, which covers the traditional GDP measure plus non-market household production, then the share of total household production (both market and non-market) in extended GDP is calculated at 35% for Finland and 36–37% for Poland.

To conclude: the same level of household production share in the extended GDP obtained from independently conducted estimations confirms that both the valuation method (input methods) used and the choice of a specific approach (the generalist's or specialist's wages) allow for reliable assessment of the non-market contribution of domestic production to the national economy. Both approaches provide the basis for comparison between macroeconomic measures which are estimated in extended tables in the core national accounts as well as in the satellite accounts. The most significant difference between the generalist's and specialist's wage is that the first of them is simpler to calculate the value of domestic labor because of the use of average wages of specialist's to estimate it.

The generalist's wage is a more time-consuming and labor-intensive approach because it requires more intensive work and estimates to pair each specific domestic work (time use survey code of activity) with corresponding codes of occupations.

Other summarizing items in the sequence of extended tables in the Household Production Satellite Account are calculated according to the same formulas in both methods.

Table 2: SNA and non-SNA household production in Finland, 2009-2016 (EUR million)

Specification	2009	2012	2016
GDP (ESA2010)	181 029	199 793	216 111
Household production, total, EUR million	108 007	121 239	128 384
Gross value added of SNA household production	**12 563**	**13 784**	**16 398**
Gross value added of non-SNA household production	65 822	73 139	78 816
Gross value added of voluntary work (non-SNA)	6 368	6 973	7 286
Sum of non-SNA household production and voluntary work (non-SNA)	72 190	80 112	86 102
Total gross value added of household production (SNA + non-SNA)	**78 386**	**86 923**	**95 214**
Total gross value added of household production (SNA + non-SNA) and voluntary work (non-SNA)	84 753	93 896	102 500
Share of non-SNA household production of GDP (%)	36.4%	36.6%	36.5%

Table 2: SNA and non-SNA household production in Finland, 2009-2016 (EUR million) (cont.)

Specification	2009	2012	2016
Share of sum of non-SNA household production and voluntary work (non-SNA) of GDP (%)	**39.9%**	**40.1%**	**39.8%**
Expanded GDP (= sum of GDP and gross value added of non-SNA household production)	246 851	272 932	294 927
Share of household production of expanded GDP	31.8%	31.8%	32.3%
Expanded GDP including voluntary work (= sum of GDP and gross value added of non-SNA household production and gross value added of voluntary work).	253 219	279 905	302 213
Share of household production (including voluntary work) of expanded GDP	**34.3%**	**34.4%**	**34.8%**
The wage used in calculations (eur / hour)	13.44	14.53	15.36
Value of domestic work (unpaid work)	70 014	77 205	83 325
Output	114 375	128 212	135 670
Share of gross value added (unpaid work) of household production (%)	**61.2%**	**60.2%**	**61.4%**

Source: K. Soinne calculations, Statistics Finland.

Table 3: SNA and non-SNA household production in Poland, 2011-2016 (EUR million)

Specification	2011[a]	2013[b]	2016[b]
GDP (ESA2010)	377 189	388 356	424 803
Household production, total	277 456	278 438	308 005
Gross value added of SNA household production	**31 146**	**29 635**	**31 610**
Gross value added of non-SNA household production	160 880	162 214	180 631
Gross value added of voluntary work (non-SNA)	9 789	10 182	10 883
Sum of non-SNA household production and voluntary work (non-SNA)	170 668	172 395	191 514
Total gross value added of household production (SNA + non-SNA)	**192 026**	**191 849**	**212 241**
Total gross value added of household production (SNA + non-SNA) and voluntary work (non-SNA)	201 815	202 031	223 124
Share of non-SNA household production of GDP (%)	42.7%	41.8%	42.5%

Table 3: SNA and non-SNA household production in Poland, 2011-2016 (EUR million) (cont.)

Specification	2011[a]	2013[b]	2016[b]
Share of sum of non-SNA household production and voluntary work (non-SNA) of GDP (%)	**45.2%**	**44.4%**	**45.1%**
Expanded GDP (= sum of GDP and gross value added of non-SNA household production)	538 069	550 570	605 434
Share of household production of expanded GDP	35.7%	34.8%	35.1%
Expanded GDP including voluntary work (= sum of GDP and gross value added of non-SNA household production and gross value added of voluntary work)	547 857	560 752	616 317
Share of household production (including voluntary work) of expanded GDP	**37.5%**	**36.7%**	**36.9%**
The wage used in calculations (eur / hour)	-	-	-
Value of domestic work (unpaid work)	161 098	162 728	180 774
Output	277 456	278 438	308 005
Share of gross value added (unpaid work) of household production (%)	**58.1%**	**58.4%**	**58.7%**

a – estimates based on Time use survey Poland 2003/2024,
b – estimates based on Time use survey Poland 2013.

Source: own calculations.

The following figures are based on Finnish satellite accounts for 2006 (Varjonen and Alto 2010) and Polish Household Production Satellite Account for 2011 (Marszałek 2015). In 2016 the household sector in Finland delivered to the economy the value of unpaid work which was calculated at 83.3 billion EUR. Of course, it is invisible for the market value produced for own use or for other family members. If we compare that value of domestic labor to the household production, we will observe that it is 74.5% of the total output (Table 4). Finally, if we sum both market and non-market household production, we will achieve 61.4% of total output. It confirms that unpaid work is a significant part of the total result of productive activities of households. Estimates of the value of domestic work, which is a major component of gross value added (GVA), is also the most informative component of new production in the national accounts, and is calculated at 75-76% (Table 4.).

Table 4: Main aggregates of extended household accounts compared to SNA-based household accounts in Finland, 2016 (EUR million)

Components of household production	Household production		
	SNA (market production)	**Non-SNA** (non-market production)	**Total (SNA+non-SNA)**
Value of labor (number of hours x hourly wages)	◦	83 325	83 325
Paid domestic staff	◦	◦	◦
Housing services produced by owner occupiers (rents of equivalent rented accommodation)	8 444	◦	8 444
Own-account house construction	739	◦	739
Agricultural production for own use (hunting, fishing, picking berries and mushrooms)	160		160
Vehicle tax (part)		287	
Taxes on production	813	1	814
Subsidies on production	◦	-1 565	-1 565
Net value added	**10 155**	**82 048**	**92 204**
Consumption of fixed capital (depreciation)	6 243	4 054	10 297
Gross value added (GVA)	**16 398**	**86 102**	**102 500**
Intermediate consumption	6 818	26 352	33 170
Total output (household production)	**23 217**	**112 454**	**135 670**
Share of total output of total household production (SNA + non-SNA) (%)	17.1%	82.9%	100%
Share of housework of total output (household production) (%)	◦	74.1%	61.4%
Share of GVA of total output (household production) (%)	◦	76.6%	75.6%

Source: own calculations based on K. Soinne estimations, Statistics Finland.

Satellite accounts combine the interactions between market and households and explain the flows between household sector and the rest of the economy. The monetary value and amount of domestic goods and services can be compared to similar market products offered by private or public services. The significant disproportion between market and non-market household production is observed mainly in domestic work.

We made calculations for value of unpaid work (based on input method) and paid domestic staff (from national accounts). For Poland, we achieved more than 180 billion EUR of value of unpaid work and 526 million EUR of paid domestic services in 2016 (Table 5.). Only small part of total domestic work is registered in national accounts. Some of this production is non-registered (called "grey zone") because households do not employ officially workers, they pay for one day in a week as a support of house cleaning or for other domestic services.

Table 5: Main aggregates of extended household accounts compared to SNA-based household accounts in Poland, 2016 (EUR million)

Components of household production	**Household production**		
	SNA (market production)	**Non-SNA** (non-market production)	**Total (SNA+non-SNA)**
Value of labour (number of hours x hourly wages)	◦	180 774	180 774
Paid domestic staff	526	◦	526
Housing services produced by owner occupiers (rents of equivalent rented accommodation)	13 912	◦	13 912
Own-account house construction	11 807	◦	11 807
Agricultural production for own use (hunting, fishing, picking berries and mushrooms)	1 758	926	2 684
Taxes on production	1 598	251	1 849
Subsidies on production	-5 274	-5 784	-11 058
Net value added	**24 328**	**176 166**	**200 495**
Consumption of fixed capital (depreciation)	7 282	15 348	22 629
Gross value added	**31 610**	**191 514**	**223 124**
Intermediate consumption	40 844	44 036	84 881
Total output (household production)	**72 455**	**235 550**	**308 005**
Share of total output of total household production (SNA + non-SNA) (%)	23.5%	76.5%	100%
Share of housework of total output (household production) (%)	◦	76.7%	58.7%
Share of GVA of total output (household production) (%)	◦	81.3%	72.4%

Source: own calculations.

When considering the ratio of domestic work to the value of home production in Poland, analogous proportions to Finland were obtained. 76.7% of total output of non-market production was allocated in unpaid work, and 58.7% for output as a sum of SNA and non-SNA household production (for Finland: 74.1% and 61.4% respectively). In Poland the highest values were calculated in the share of GVA of total output (household production): 81.3% and 72.4% (Table 5.). This indicates that the value and amount of unpaid work in Poland has a higher impact on total household production than other categories, such as agricultural products, housing services produced by ourself, housing construction for own use, subsidies to domestic products or taxes on production. We also observe from the analysis that in Poland households spend more time on unpaid work, and the proportion of that monetary value to total household production is also higher than in Finland because GDP per capita in Poland is also lower than in Finland. If GDP per capita increases then people spend more time on paid work or leisure time activities (sports, hobby, travelling) than on domestic work, so they spend more money for market products than they produce goods or services by themselves.

Total output is a sum of gross value added and consumption of fixed capital (depreciation). Considering only the market (official) production by household sector, official statistics miss the additional 76.5% of total output (household production, both market and non-market) in Poland (Table 5.) and 82.9% in Finland (Table 4.). These calculations confirm that Household Production Satellite Accounts provide overall view and wider perspective of real household impact to the economy. The increase in the value in care and family care, as well as other domestic works in comparison to prices of market services, also affects the growth of social awareness that unpaid work is important as paid work. Therefore, regular estimates of labor and home production provide a better understanding of the economic interactions between market (public or private institutions) and households.

4. Conclusion and discussion

In particular, the issue of other kinds of estimating methods of non-observed productive activities (household production) in the economy remains one of the future research issues. Previous studies and analyses use different methods: input and output method, and in various approaches: replacement (market) cost or alternative cost method adopted the best practices and solutions to produce comparable figures. However, the core system of national account is still progressing – it is possible to choose the most appropriate harmonized method for group of countries, especially in the EU zone. Finnish, German, French and Polish recommendations are consistent with the Eurostat guidelines, which means that the input method is the most proper to

that kind of calculations and it should be implemented and developed. The analysis of this paper confirms that generalist's or specialist wages do not have crucial impact on final proportions of value of domestic work in the share of total output (household production). Also, share of GVA of SNA (market production) or non-SNA (non-market production) to total household production obtained in this estimation reflects similar proportions even if we used different wages in the input approach. The most valuable of that analysis is the confirmation that the input method is absolutely the most appropriate approach to provide further calculations as a sequence of accounts. If the input method is officially confirmed by National Statistics, then the further consideration of satellite accounts might be adjusted to the system of national accounts (SNA/ESA) and implemented.

The sector satellite account of household production – Household Production Satellite Account (HHSA) may fulfill the gap in core statistics because it is a comprehensive statistical tool to estimate the real economic value of domestic work and household production, both visible and invisible in the market and national accounts. The HHSA covers the market products and non-market goods and services which are made and provided in households. Moreover, satellite accounts generate a complete overview of final interactions and financial flows between market and households, and provide the European harmonized comparable figures to core national accounts (ESA).

Our recommendation is that the input method should take the first place in the satellite accounts and other detailed solutions should be developed and discussed, such as: definitions of household output, whole sequence of accounts (to economic analyses and forecasts), and the value of labor. The estimates of value of labor should be adjusted to available data, resources and regional (national) specifics. The approach based on the replacement (market) cost is recommended according to one of different types of calculation: generalist's or specialist's wages, depending on the available statistics in a given country.

References

Balestra, C., Boarini, R. and Ruiz, N., (2018). Going beyond GDP: empirical findings [in:] Handbook of research on economic and social well-being (edt. Conchita D'Ambrosio). *Edward Elgar Publishing*, pp. 52–103.

Becker, G. S., (1960). Demographic and Economic Change in Developed Countries. New York. NY: *Columbia University Press. An Economic Analysis of Fertility*, pp. 209–40.

Becker, G. S., (1962). Irrational behavior and economic theory. *Journal of Political Economy*, 70(1), pp. 1–13.

Becker, G. S., (1964). Human Capital: A Theoretical and Empirical Analysis. With Special Reference to Education. *National Bureau of Economic Research*: New York.

Becker, G. S., (1965). A Theory of the Allocation of Time. *Economic Journal*, 65, pp. 493–517.

Becker G. S. (1973). A theory of marriage: Part I. *Journal of Political Economy*, 81(4), pp. 813–46.

Becker, G. S., (1974). A theory of social interactions. *Journal of Political Economy*, 82(6): pp. 1063–93.

Becker, G. S., (1981). A Treatise on the Family. Cambridge. MA: *Harvard University Press*.

Becker, G. S., (2007). Economic Theory. New Brunswick. NJ: *Transaction Publishers*.

Blades, D., (1997). A proposal for the measurement of non-market household production. Session paper. *IATUR*. Stockholm 8–10 October 1997.

Błaszczak-Przybycińska, I., (2008). Produkcja gospodarstw domowych jako czynnik dochodotwórczy. Warsaw: *Oficyna Wydawnicza SGH*.

Błaszczak-Przybycińska, I., (2007). Estimation of unpaid work in Polish households. *Statistics in transition-new series*, 8(6), pp. 547–561.

Błaszczak-Przybycińska, I., Marszałek, M., (2021). Transfery czasu pracy i przepływy międzypokoleniowe w gospodarstwach domowych w Polsce. In: J. Szczepaniak-Sienniak. Ekonomia polityki rodzinnej. Wybrane zagadnienia. Wrocław: *Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu*, pp. 61–78.

Błaszczak-Przybycińska, I., Marszałek, M., (2020). Gospodarowanie czasem. In: Statystyka społeczna. T. Panek (ed.). Warszawa: *PWE*, pp. 405–436.

Błaszczak-Przybycińska I., Marszałek M., (2019). Satellite Account of Household Production. Methodological remarks and results for Poland. *Econometrics*, 23(1), pp. 61–76. https://doi.org/10.15611/eada.2019.1.05; OAI dbc.wroc.pl:61993.

Błaszczak-Przybycińska I., Marszałek M., (2015). Wycena wartości pracy własnej gospodarstw domowych na podstawie badania budżetu czasu (Own work of households value estimation on the basis of time use survey). In: Budżet czasu ludności 2013. Część I (Time use survey 2013. Part I). *GUS*, pp. 131–183.

Błaszczak-Przybycińska I., Putkowska A., (2024). The impact of selected solutions in the field of wage rates on the results of the valuation of domestic work using the input method (in print).

Duran. M.-A., (2007). The Satellite account for unpaid work in the community of Madrid. *La Suma de Todos. Community de Madrid 36*.

Eurostat, (2003). Household Production and Consumption Proposal for a Methodology of Household Satellite Accounts. Luxembourg: *Office for Publications of the European Communities*.

Eurostat, (2013). European system of accounts. ESA 2010. Luxembourg: *Publications Office of the European Union*. https://doi.org/10.2785/16644

Fitzgerald, J. M., Swenson M. S. and Wicks J. H., (1996). Valuation of Household Production at Market Prices and Estimation of Production Functions. *Review of Income and Wealth No. 2*.

Folbre, N., (2006). The Invisible Heart: Economics and Family Values. *New Press*.

Gershuny, J., (2000). The European Union's Time Use Survey: Implications for Household Production and the Role of Time in Economic Analysis. *Review of Income and Wealth*, 46(1), pp. 5–30.

Giovannini, E., Rondinella, T., (2018). Going beyond GDP: theoretical approaches. In: Handbook of research on economic and social well-being (edt. Conchita D'Ambrosio). *Edward Elgar Publishing*, pp. 1–51.

Goldschmidt-Clermont, L., (1994). Monetary Valuation of Unpaid Work. In Proceedings of the International Conference on the Measurement and Valuation of Unpaid Work. Ottawa, April 28–30, 1993. *Statistics Canada and the Status of Women in Canada*, pp. 67–77.

Goldschmidt-Clermont, L., (2000). Household production and income: Some preliminary issues. *Bulletin of labour statistics 2000-2*, ILO, Geneva.

Gorman, W. M., (1959). Separable utility and aggregation. *Econometrica*, 27(3), pp. 469–81.

Gorman, W. M., (1980). A possible procedure for analysing quality differentials in the egg market. *Review of Economic Studies*, 47(5), pp. 843–56.

Graham J. W., Green, C. A., (1984). Estimating the parameters of a household production function with joint products. *The Review of Economics and Statistics*, 66(2), pp. 277-282.

Green, H. A. J., (1964). Aggregation in Economic Analysis; an Introductory Survey. Princeton, New York: *Princeton University Press*.

Gronau, R., (1977). Leisure. home production and work-the theory of the allocation of time revisited. *Journal of Political Economy*, 85(6), pp. 1099–23.

Gronau, R., (1986). Home production–a survey In: Ashenfelter O. Layard R. editors. Handbook of Labor Economics, Amsterdam: *Elsevier Science Publishers*, pp. 273–304.

Gronau, R., (1997). The theory of home production: the past ten years. *Journal of Labor Economics*, 15(2), pp. 197–205.

Gronau, R., (2008). Household production and public goods. In: Durlauf Steven N. Blume Lawrence E. editors. The New Palgrave Dictionary of Economics. Basingstoke. UK: *Palgrave Macmillan*, 2nd edn. https://doi.org/10.1057/9780230226203.0750.

Heckman, J. J., (1988). Time constraints and household demand functions' In: Schultz TP. editor. Research in Population Economics: A Research Annual. Greenwich. CN: *JAI Press*, pp. 3–14.

Heckman, J. J., (2015). Introduction to A Theory of the Allocation of Time by Gary Becker. *Economic Journal* (London), 125(583), pp. 403–409.

Holloway, S., Short, S. and Tamplin, S., (2002). Household Satellite Account (experimental) Methodology. https://www.ons.gov.uk/peoplepopulationandcommunity/housing/methodologies/householdsatelliteaccountexperimental.

Koreman, P., Kapteyn, A., (1987). A Disaggregated Analysis of the Allocation of Time Within Households. *Journal of Political Economy*, 95, pp. 223–249.

Kuznets, S., (1934). National income. *NBER bulletin 49*. New York. NY, 1929–1932.

Kuznets, S., (1955). Economic growth and income inequality. *American Economic Review*, 45, pp. 1–28.

Kuznets, S., (1962). How to judge quality. *New Republic*, 147, pp. 29–31.

Luszniewicz, A., (1982). Statystyka społeczna. Warszawa: *PWE*.

Marszałek, M., (2020). The Unobserved Economy – Invisible Production in Households. The Household Production Satellite Account and the National Time Transfer Account. *Statistics in Transition new series*, 21(3), pp. 149–169. https://doi.org/10.21307/stattrans-2020-049.

Marszałek, M., (2015). Rachunek produkcji domowej w Polsce w koncepcji systemu statystyki społecznej. Warszawa: *Oficyna Wydawnicza SGH*.

Mitchell, W. C., (1912). The Backward Art of Spending Money. *The American Economic Review*, 2(2), pp. 269–281.

OECD, (2010). Incorporating estimates of household production of non-market services into international comparisons of material well-being. *Working Party of National Accounts*, STD/CSTAT/WPNA (2010) 9.

Poissonnier, A., Roy, D., (2013). Household Satellite Account for France in 2010. Methodological issues on the assessment of domestic production. France: *Institut National de la Statistique et des Études Économiques*.

Reid, M., (1934). *Economics of Household Production*. New York: Wiley.

Rüger, Y., Varjonen, J., (2008). Value of household production in Finland and Germany: analysis and recalculation of the Household Satellite Account System in both countries. Helsinki: *National Consumer Research Centre*.

Stiglitz, J. E., Sen. A. and Fitoussi. J. P., (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*. https://www.researchgate.net/publication/258260767

Schäfer, D., (2004). Unbezahlte Arbeit und Bruttoinlandsproduct 1992 und 2001. Neuberechnung des Haushalts-Satellitensystems. Wiesbaden: *Statistiches Bundesamt.*

Soinne, K., (2021). Value of household production 2020. Annual national accounts 2021. *Statistics Finland*: Helsinki.

Szałtys, D., Cierniak-Piotrowska, M., (2022). Wybrane wskaźniki z zakresu dzietności w latach 1980–2021. *Departament Badań Demograficznych GUS*. Konferencja naukowa: Uwarunkowania dzietności. Warszawa: 19–20.10.2022 r. https://www.gov.pl/ attachment/68660185-68a0-4c4d-ac9e-56ae2f91ce5a.

Szép, K., (2003). Összefoglalás helyett – A Nemzeti Számlákban nem Kimutatott Háztartási Termelés Termelési Számlája és a Jövöbeli Feladatok. [in:] A Háztartási Termelés Ertéke a Mai Magyarországon (Household Statistical Office. Satellite Accounts. Statistical Sampling and Methodology Section). Budapest: *Hungarian Central.*

UNECE, (2017). *Guide on Valuing Unpaid Household Service Work*. New York and Geneva.

Varjonen, J., Aalto, K., (2006). Household production and consumption in Finland 2001. Household Satellite Account. Helsinki: *Statistics Finland*. *National Consumer Research Centre*.

Varjonen, J., Aalto, K., (2010). Kotitalouksien palkaton tuotanto ja ostopalvelujen käyttö. (Households' unpaid production and use of market services). Helsinki: *National Consumer Research Centre*.

Varjonen, J., Hamunen, E. and Soinne, K., (2014) Satellite Accounts on Household Production: Eurostat Methodology and Experiences to Apply It. *Working Papers* 1/2014. *Statistics Finland* (ISSN 2323–1998).

Varjonen, J., Hamunen, E., (1999). Proposal for a Satellite Account of Household Production. Agenda Item 1. Statistics Finland. *OECD Meeting of National Accounts Experts*, Paris.

Varjonen, J., Niemi, I., Hamunen, E., Sandström, T. and Pääkkönen, H., (1999). Proposal for a Satellite Account of Household Production. *Eurostat Working Papers*. 9/1999/A4/11: 92.

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 177–198, https://doi.org/10.59139/stattrans-2025-021
Received – 28.08.2022; accepted – 13.01.2025

Ratio regression type estimators of the population mean for missing data in sample surveys

Prachi Garg[1], **Namita Srivastava**[2], **Manoj Kumar Srivastava**3

Abstract

In this article, new ratio regression type estimators with imputation have been proposed as means to overcome the problem of missing data relating to a studied variable in a sample survey. It has been shown that the suggested estimators are more efficient than the mean method of imputation, the ratio method of imputation, the regression method of imputation, and the estimators given by Singh and Horn (2000), Singh and Deo (2003), Singh (2009), Diana and Perri (2010) and Gira (2015). The biases and their mean square errors of the suggested estimators are derived. A comparative study is conducted using real and simulated data. The results are found to be encouraging showing improvement of all the methods discussed in this article.

Key words: imputation methods, Bias, Mean square error (MSE), Efficiency, Ratio-Regression type estimators.

1. Introduction

Missing data or missing values occur when no data value is stored for a variable in an observation. Even in a well-designed and controlled study, missing data occurs in almost all research. Missing data is commonly described as major issue in most scientific research domains that may originate from such a mishandling sample, measurement error, non-response or deleted aberrant value. To get precise estimates of population parameters we seek information on every selected unit of the sample. Imputation means replacing a missing value with other value based on a reasonable estimate. Information on the related auxiliary variable is generally used to recreate the

[1] Department of Statistics, St. John's College Agra, Dr. B.R. Ambedkar University, Agra (U.P.), India. E-mail: prachigarg2093@gmail.com. ORCID: https://orcid.org/0009-0001-8809-4464.

[2] Department of Statistics, St. John's College Agra, Dr. B.R. Ambedkar University, Agra (U.P.), India. E-mail: drnamita.sjc@gmail.com. ORCID: https://orcid.org/0000-0001-8695-9148.

[3] Shaheed Mahendra Karma Vishwavidyalaya, Bastar, Chhattisgarh, India. E-mail: mksriv@gmail.com. ORCID: https://orcid.org/0000-0002-8256-1439.

missing values for completing datasets. Incomplete data is usually categorized into three different response mechanisms: Missing Completely At Random (MCAR); Missing At Random (MAR); and Missing Not At Random (MNAR or NMAR) (Little & Rubin,2002). In Missing Completely at Random (MCAR) missing data is randomly distributed across the variable and unrelated to other variables. In Missing at Random (MAR) the missing observations are not randomly distributed but they are accounted for by other observed variables. In Missing Not at Random (MNAR) category, the missing data systematically differ from the observed values. In the present article we are assuming MCAR response mechanism of missing data.

Auxiliary information is important for a survey practitioner as it is utilized to improve the performance of the methods in finite sample survey. At the estimation stage the auxiliary information is utilized for suggesting imputation methods which results in ratio, product and regression estimators. Many imputation methods have been proposed utilizing the auxiliary information. Several researchers (Lee, Rancourt, and Sarndal 1994,1995; Singh and Horn 2000; Singh et. al; Diana and Perri 2010; Pandey, Thakur, and Yadav 2015; Singh et. al. 2016; Bhushan and Pandey 2018; Prasad 2017, 2018, 2019; Singh and Khalid 2019; K Chodjuntug and N Lawson (2022)[4]; K Chodjuntug and N Lawson (2022)[5]; N Lawson (2023) [12]; N Lawson (2023) [13]; N Thongsak and N Lawson (2023) [14] etc.) among others assumed MCAR mechanism to develop several imputation methods and resultant estimators to estimate population mean in the case of missing data problems. The imputation methods proposed by Singh and Horn (2000), Singh and Deo (2003), Singh (2009), Diana and Perri (2010), and Gira (2015) result in different estimators, but they all lead to the same Mean Squared Error (MSE) formula, which are same as regression method of imputation. Therefore, in this paper we compared and simulated our estimator with the mean, ratio and regression estimators after proposing the new imputation strategy and the resulting estimator. The proposed estimators come out to be more efficient than the usual mean, ratio and regression (Diana & Perri's regression) method for handling missing observations to estimate the population mean.

This article proposes three ratio-regression type imputation methods to inadequate the annoyance outcome of nonresponse in survey sampling. The resulting classes of point estimators that can be used to estimate the population mean have been discussed in detail. The bias and Mean Square Error (MSE) properties of the proposed estimators have been derived. An empirical study was conducted to assess their performance in comparison with existing estimators, and the findings have been presented. These are designed as follows.. In Section 2,the sample structure and notations are considered and in Section 3, we have reviewed several imputation techniques of finite population mean under non-response that are available in the literature suggested by various authors. In Section 4, construction of the suggested alternative method of imputation

is carried out and the bias mean square error equations for this estimator is obtained. In Section 5, we have proposed a new method of imputation and obtained their bias mean square error equations for this estimator. In Section 6, we have conducted efficiency comparison of alternative method of imputation. In Section 7, we do computational study by using real and artificial populations, respectively. Section 8 summaries the main findings and conclusions.

2. Sample Structure and Notations:

Consider a finite population U = $\{U_1, U_2, \ldots\ldots, U_N\}$ of size N for which a random sample s = $\{u_1, u_2, \ldots\ldots, u_N\}$ of size n under simple random sampling without replacement scheme is drawn to estimate the population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ of the study variable y. Let y_i and x_i be the values of the study variable y and auxiliary variable x , respectively for the i^{th} unit of a finite population of size N. The information on x can be available on the entire population through knowledge of x_i, ∀ iϵU, or its population mean $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$. Let s be a simple random sample without replacement (SRSWOR) of size n $(n < N)$ drawn from U to estimate Y. Let r be the number of responding units out of sampled n units. Let the set of responding units be denoted by R and that of non-responding (n-r) units be denoted by R^C. For every unit, $i \in R$ the value y_i is observed. However, for the units, $i\epsilon R^C$, the y_i values are missing and imputed values are derived. Imputation is performed by employing the auxiliary variable x where values are believed to be known for each sampled unit iϵs.

The structure of the general method of imputation in the case of complete dataset under nonresponse is defined as:

$$y_{.i}=\begin{cases} y_i & if \quad i\epsilon R \\ \hat{y}_i & if \quad i\epsilon R^c \end{cases}$$

where $\hat{y}_i$ is the imputed value for the i^{th} non-responding unit. Using the above data, we get the following form of the general point estimator of the population mean ($\bar{Y}$)

$$\bar{y}_s = \frac{1}{n}\sum_{i\epsilon s_n} y_{.i} = \frac{1}{n}[\sum_{i\epsilon R} y_{.i} + \sum_{i\epsilon R^c} y_{.i}] = \frac{1}{n}[\sum_{i\epsilon R} y_i + \sum_{i\epsilon R^c} \hat{y}_i].$$

Here, $\hat{y}_i$ takes a different value for a different imputation method.

The following notations have been adopted for further use:

$\bar{X}$,$\bar{Y}$: The population mean of the auxiliary variable x and study variable y respectively,

$\bar{y}_r$: Sample mean of responding units,

$\bar{x}_n$: Sample mean of all units,

$\bar{x}_r$: Sample mean of responding units,

$\rho_{yx} = \frac{S_{yx}}{S_y S_x}$: The correlation coefficient between the variables y and x,

$S_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X})$: The covariance between y and x,
$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$: The population mean square of y,
$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$: The population mean square of x,
$C_{y=}\frac{S_y}{\bar{Y}}$ & $C_{x=}\frac{S_x}{\bar{X}}$: The coefficients of variation of y and x, respectively,

we define,

$$\bar{y}_r = \bar{Y}\,(1+\varepsilon_o),\ \bar{x}_r = \bar{X}\,(1+\delta_o),\ \ \bar{x}_n = \bar{X}\,(1+\eta_o)$$

using the above notation, we have

$$\mathrm{E}\,(\varepsilon_o) = E(\delta_o) = E(\eta_o) = 0$$

and,

$$\mathrm{E}(\varepsilon_o^2) = \left(\frac{1}{r}-\frac{1}{N}\right)C_y^2,\ \mathrm{E}(\delta_o^2) = \left(\frac{1}{r}-\frac{1}{N}\right)C_x^2,\ \mathrm{E}(\varepsilon_o\delta_o) = \left(\frac{1}{r}-\frac{1}{N}\right)\rho_{xy}C_xC_y,$$
$$\mathrm{E}(\eta_o^2) = \left(\frac{1}{n}-\frac{1}{N}\right)C_x^2,\mathrm{E}(\delta_o\eta_o) = \left(\frac{1}{n}-\frac{1}{N}\right)C_x^2,\mathrm{E}(\varepsilon_o\eta_o) = \left(\frac{1}{n}-\frac{1}{N}\right)\rho_{xy}C_xC_y.$$

3. Review of Some existing estimators:

In this section, we discuss some of the classical and existing imputation methods for estimating the population mean in sample surveys.

3.1. Mean method of imputation:

In the mean method of imputation the form of data by Lee, Rancourt and Sarndal (1994) is treated as

$$y_{i,m} = \begin{cases} y_i & if \quad i\epsilon R \\ \bar{y}_r & if \quad i\epsilon R^c \end{cases} \tag{3.1}$$

The mean estimator under the new data (3.1) is given by

$$\bar{y}_m = \frac{1}{r}\sum_{i\epsilon R} y_i = \bar{y}_r \tag{3.2}$$

The variance of the response sample mean $\bar{y}_m$ is given by

$$V\,(\bar{y}_m) = V\,(\bar{y}_r) = \left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2C_y^2 \tag{3.3}$$

3.2. Ratio method of imputation:

The ratio method of imputation, based on information from the auxiliary variable x, was proposed by Lee, Rancourt, and Särndal (1994). Under this method, the imputed data are adjusted using the known relationship between the study variable and the auxiliary variable.

$$y_{i,R} = \begin{cases} y_i & if \quad i\epsilon R \\ \hat{b}x_i & if \quad i\epsilon R^c \end{cases} \tag{3.4}$$

where $\hat{b} = \frac{\sum_{i\,\epsilon\,R} y_i}{\sum_{i\epsilon R} x_i}$.

The ratio estimator in the case of imputation method (3.4), is defined as

$$t_R = \bar{y}_r \left(\frac{\bar{x}_n}{\bar{x}_r}\right) = \bar{y}_{RAT} \tag{3.5}$$

The bias and mean square error of the estimator t_R are obtained under MCAR response mechanism up to the first order approximation, and are given by

$$B(t_R) = \left(\frac{1}{r} - \frac{1}{n}\right) \bar{Y}\, \mathrm{C}_x(C_x - \rho C_y) \tag{3.6}$$

and $$\mathrm{MSE}\,(\bar{y}_{RAT}) = V(\bar{y}_r) + \left(\frac{1}{r} - \frac{1}{n}\right) \bar{Y}^2 \mathrm{C}_x(C_x - 2\,\rho C_y) \tag{3.7}$$

The ratio method of imputation is better to choose over the mean method of imputation whenever $(\rho C_y/C_x)$>1 / 2.

3.3. Regression method of imputation:

In this method, the data after imputation becomes

$$y_{i,REG} = \begin{cases} y_i & if \quad i\epsilon R \\ a + b_{yx}x_i & if \quad i\epsilon R^c \end{cases} \tag{3.8}$$

where, $a = \bar{y}_r - b_{yx}\,\bar{x}_r$ and $b_{yx} = \frac{S_{yx}}{S_x^2}$

The point estimator of population mean $\overline{\mathrm{Y}}$

$$\bar{y}_{REG} = \bar{y}_r + b_{yx}(\bar{x}_n - \bar{x}_r) \tag{3.9}$$

The bias and mean square error of the estimator $\bar{y}_{REG}$ are obtained under MCAR response mechanism up to the first order approximation, and are given by

$$B(\bar{y}_{REG}) = \frac{\rho_{yx}C_y}{C_x\bar{X}} \left(\frac{1}{r} - \frac{1}{n}\right) \overline{\mathrm{Y}} \left(\frac{\mu_{300}}{\mu_{200}} - \frac{\mu_{210}}{\mu_{110}}\right) \tag{3.10}$$

where, $\mu_{abc} = \sum_{i=1}^{N}(x_i - \bar{X})^a\,(y_i - \bar{Y})^b(z_i - \bar{Z})^c$

$$\mathrm{M}(\bar{y}_{REG}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right) S_y^2\rho_{yx}^2 \tag{3.11}$$

3.4. Singh and Horn (2000) Estimator

Singh and Horn (2000) introduced this method, the data after imputation becomes

$$y_{.i} = \begin{cases} (\alpha n/r)y_i + (1-\alpha)\hat{b}x_i & if \quad i\epsilon R \\ (1-\alpha)\hat{b}x_i & if \quad i\epsilon R^c \end{cases} \tag{3.12}$$

The point estimator of population mean is given as

$$\bar{y}_{comp} = \left[\alpha\bar{\mathrm{y}}_{\mathrm{r}} + (1-\alpha)\bar{\mathrm{y}}_{\mathrm{r}}\frac{\bar{\mathrm{x}}_{\mathrm{n}}}{\bar{\mathrm{x}}_{\mathrm{r}}}\right] \tag{3.13}$$

where α is an appropriate constant with optimum value $\alpha^* = 1 - \rho_{yx}\left(\frac{C_y}{C_x}\right)$: The bias of the estimator $\bar{y}_{comp}$ is given by

$$B(\bar{y}_{COMP}) = (1-\alpha)\left(\frac{1}{r} - \frac{1}{n}\right) \overline{\mathrm{Y}}\, \mathrm{C}_x\left(C_x - \rho_{yx}\,C_y\right) \tag{3.14}$$

using α^* we get the minimum MSE of $\bar{y}_{comp}$ as

$$\mathrm{M_{min}}(\bar{\mathrm{y}}_{\mathrm{COMP}}) = \mathrm{MSE}(\bar{y}_{RAT}) - \left(\frac{1}{r} - \frac{1}{n}\right)\overline{\mathrm{Y}}^2(C_x - \rho_{yx}\, C_y)^{\,2} \tag{3.15}$$

3.5. Singh and Deo (2003) Estimator:

Singh and Deo (2003), using power transformation, this method gives the following form of the data after imputation

$$y_{.i} = \begin{cases} y_i & if \quad i\epsilon R \\ \bar{\mathrm{y}}_{\mathrm{r}}\left[\mathrm{n}\left(\frac{\bar{x}_n}{\bar{x}_r}\right)^{\alpha} - \mathrm{r}\right]\frac{x_i}{\sum_{\mathrm{i}\epsilon \mathrm{R}^{\mathrm{c}}} x_i} & if \quad i\epsilon R^c \end{cases} \tag{3.16}$$

The resultant estimator of the population mean is given as

$$\bar{y}_{SD} = \bar{\mathrm{y}}_{\mathrm{r}}\left(\frac{\bar{\mathrm{x}}_{\mathrm{n}}}{\bar{\mathrm{x}}_{\mathrm{r}}}\right)^{\alpha} \tag{3.17}$$

where α is a suitably chosen constant and the optimum value α is $\alpha^*=\rho_{yx}\left(\frac{C_y}{C_x}\right)$

The bias of the estimator $(\bar{y}_{SD})$ obtained by Singh and Deo is given by

$$B(\bar{y}_{SD}) = \left(\frac{1}{r} - \frac{1}{n}\right)\overline{\mathrm{Y}}\left(\frac{\beta(\beta-1)}{2}\mathrm{C_x^2} - \rho_{yx}\, C_y C_x\right) \tag{3.18}$$

using optimum value α^*, the minimum MSE of $\bar{y}_{SD}$ is given

$$\mathrm{MSE_{min}}(\bar{\mathrm{y}}_{\mathrm{SD}}) = \mathrm{MSE}(\bar{y}_{RAT}) - \left(\frac{1}{r} - \frac{1}{n}\right)\mathrm{S_x^2}\left(\frac{S_{yx}}{\mathrm{S_x^2}} - \frac{\overline{\mathrm{Y}}}{\overline{\mathrm{X}}}\right)^{\mathbf{2}} \tag{3.19}$$

3.6. Singh (2009) Estimator:

This method of imputation is an alternative technique to estimate population mean $\overline{\mathrm{Y}}$ in the presence of non-response. The study variate after imputation takes the following form

$$y_{.i} = \begin{cases} y_i & if \quad i\epsilon R \\ \bar{y}_r\left[\frac{(n-r)\bar{x}_n + \alpha r(\bar{x}_n - \bar{x}_r)}{\alpha\bar{x}_r + (1-\alpha)\bar{x}_n}\right]\frac{x_i}{\sum_{i\epsilon R^c} x_i} & if \quad i\epsilon R^c \end{cases} \tag{3.20}$$

The point estimator of population mean as following

$$\bar{y}_{singh} = \frac{\bar{y}_r\bar{x}_n}{\alpha\bar{x}_r + (1-\alpha)\bar{x}_n} \tag{3.21}$$

where α is an appropriate constant with optimum value $\alpha^* = \rho_{yx}$

The bias of the estimator $(\bar{y}_{singh})$ is given as

$$B(\bar{y}_{singh}) = \overline{\mathrm{Y}}\left[\left(\frac{1}{n} - \frac{1}{N}\right)\rho_{yx}\, C_y C_x + \alpha^2\left(\frac{1}{r} - \frac{1}{n}\right)c_x^2 + (1-\alpha)^2\left(\frac{1}{n} - \frac{1}{N}\right)c_x^2 - \alpha\left\{\left(\frac{1}{r} - \frac{1}{n}\right)\rho_{yx}\, C_y C_x + \left(\frac{1}{n} - \frac{1}{N}\right)c_x^2\right\} + 2\alpha(\alpha-1)\left(\frac{1}{n} - \frac{1}{N}\right)c_x^2 - (1-\alpha)\left(\frac{1}{n} - \frac{1}{N}\right)\left(\rho_{yx}\, C_y C_x + c_x^2\right)\right] \tag{3.22}$$

using optimum value of α is α^*, the minimum MSE of $\bar{y}_{Singh}$ is given

$$\mathrm{MSE_{min}}(\bar{\mathrm{y}}_{\mathrm{Singh}}) = \mathrm{MSE}(\bar{y}_{RAT}) - \left(\frac{1}{r} - \frac{1}{n}\right)\mathrm{S_x^2}\left(\frac{S_{yx}}{\mathrm{S_x^2}} - \frac{\overline{\mathrm{Y}}}{\overline{\mathrm{X}}}\right)^{\mathbf{2}} \tag{3.23}$$

3.7. Diana and Perri (2010) Estimators

Diana and Perri (2010) propounded three regression-type imputation methods for missing data as

$$y_{i,DP1} = \begin{cases} \frac{n\,y_i}{r} + b(\overline{\mathrm{X}} - x_i) & if \quad i\epsilon R \\ b(\overline{\mathrm{X}} - x_i) & if \quad i\epsilon R^c \end{cases} \tag{3.24}$$

$$y_{i,DP2} = \begin{cases} \frac{n\,y_i}{r} - \mathrm{b}\,\frac{n\,x_i}{r} & if \quad i\epsilon R \\ b\,\frac{n\overline{\mathrm{X}}}{n-r} & if \quad i\epsilon R^c \end{cases} \tag{3.25}$$

$$y_{i,DP3} = \begin{cases} \frac{n\,y_i}{r} - \mathrm{b}\,\frac{n\,x_i}{r} & if \quad i\epsilon R \\ b\,\frac{n\bar{x}_n}{n-r} & if \quad i\epsilon R^c \end{cases} \tag{3.26}$$

The subsequent estimators under the imputation methods are respectively, given as

$$\bar{y}_{DP1} = \bar{y}_r + b(\overline{\mathrm{X}} - \bar{x}_n) \tag{3.27}$$

$$\bar{y}_{DP2} = \bar{y}_r + b(\overline{\mathrm{X}} - \bar{x}_r) \tag{3.28}$$

$$\bar{y}_{DP3} = \bar{y}_r + b(\bar{x}_n - \bar{x}_r) \tag{3.29}$$

and

$$\mathrm{MSE}(\bar{y}_{DP1}) = S_y^2\left[\left(\frac{1}{n} - \frac{1}{N}\right)(1 - \rho_{yx})^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\right] \tag{3.30}$$

$$\mathrm{MSE}(\bar{y}_{DP2}) = S_y^2\left(\frac{1}{r} - \frac{1}{N}\right)(1 - \rho_{yx})^2 \tag{3.31}$$

$$\mathrm{MSE}(\bar{y}_{DP3}) = S_y^2\left[\left(\frac{1}{n} - \frac{1}{N}\right) + \left(\frac{1}{r} - \frac{1}{n}\right)(1 - \rho_{yx})^2\right] \tag{3.32}$$

3.8. Gira (2015) Estimator

Gira (2015) proposed a ratio type imputation procedure where the study variate after imputation becomes

$$y_{.i} = \begin{cases} y_i & if \quad i\epsilon R \\ \bar{y}_r\left[n\left(\frac{\alpha - \bar{x}_r}{\alpha - \bar{x}_n}\right)\right]\frac{x_i}{\sum_{i\epsilon R^c} x_i} & if \quad i\epsilon R^c \end{cases} \tag{3.33}$$

where α is a suitably chosen constant, such that the MSE of the resultant estimator is minimum. Note that if $\alpha = 0$ then $\bar{\mathrm{y}}_{\mathrm{Gira}} = \bar{\mathrm{y}}_{\mathrm{Ratio}}$. The resultant estimator is obtained as

$$\bar{y}_{Gira} = \bar{y}_r\,\frac{\alpha - \bar{x}_r}{\alpha - \bar{x}_n} \tag{3.34}$$

The bias of the above estimator is

$$B(\bar{y}_{Gira}) = -\frac{\bar{X}\bar{Y}}{\alpha - \bar{X}}\left(\frac{1}{r} - \frac{1}{n}\right)\rho_{yx}\,C_y C_x \tag{3.35}$$

Using the optimum value of $\alpha = \bar{X}\{C_x(\rho_{yx}\,C_y)^{-1} - 1\}$ and the optimum MSE of $\bar{y}_{Gira}$ as follows.

$$\mathrm{M}(\bar{\mathrm{y}}_{\mathrm{Gira}}) = \mathrm{V}(\bar{\mathrm{y}}_{\mathrm{m}}) - \left(\frac{1}{r} - \frac{1}{n}\right)\overline{\mathrm{Y}}^2\mathrm{C}_y^2\rho_{yx}^2 \tag{3.36}$$

4. An Alternative Method of Imputation

The estimators rely on three different ratio-regression type methods of imputation as follows.

Case I: Auxiliary information on X is completely available, i.e., $\bar{X}$ is known and corresponding estimates $\bar{x}_n$ are used in the imputation technique

$$y_{.i1} = \begin{cases} y_i & if \quad i\epsilon R \\ \dfrac{n\bar{y}_r}{n-r}\left[\left\{2-\left(\dfrac{\bar{x}_n}{\bar{X}}\right)^k\right\}+\beta_1(\bar{X}-\bar{x}_n)-r\right] & if \quad i\epsilon R^c \end{cases}$$

The resultant estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{KB1} = \bar{y}_r\left\{2-\left(\frac{\bar{x}_n}{\bar{X}}\right)^k\right\}+\beta_1(\bar{X}-\bar{x}_n) \tag{4.1}$$

Case II: Auxiliary information on X is completely available i.e., $\bar{X}$ is known and corresponding estimates $\bar{x}_r$ are used in the imputation technique.

$$y_{.i2} = \begin{cases} y_i & if \quad i\epsilon R \\ \dfrac{n\bar{y}_r}{n-r} & \left[\left\{2-\left(\dfrac{\bar{x}_r}{\bar{X}}\right)^k\right\}+\beta_1(\bar{X}-\bar{x}_r)-\mathrm{r}\right] \quad if \quad i\epsilon R^c \end{cases}$$

The resultant estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{KB2} = \bar{y}_r\left\{2-\left(\frac{\bar{x}_r}{\bar{X}}\right)^k\right\}+\beta_1(\bar{X}-\bar{x}_r) \tag{4.2}$$

Case III: Auxiliary information on X is not available at population level, i.e., $\bar{X}$ is not known and we use corresponding estimates $\bar{x}_n$, $\bar{x}_r$ is used in the imputation technique.

$$y_{.i3} = \begin{cases} y_i & if \quad i\epsilon R \\ \dfrac{n\bar{y}_r}{n-r}\left[\left\{2-\left(\dfrac{\bar{x}_r}{\bar{x}_n}\right)^k\right\}+\beta_1(\bar{x}_n-\bar{x}_r)-r\right] & if \quad i\epsilon R^c \end{cases}$$

The resultant estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{KB3} = \bar{y}_r\left\{2-\left(\frac{\bar{x}_r}{\bar{x}_n}\right)^k\right\}+\beta_1(\bar{x}_n-\bar{x}_r) \tag{4.3}$$

Therefore, the expression of Bias and Mean Squared Error (MSE) of proposed estimator $(\bar{y}_{KBi,(i=1,2\&3)})$ discussed as follows.

Theorem (4.1): The bias of the proposed ratio regression type estimators $\bar{y}_{Mi}$, $i = 1,2\ and\ 3$ is given by:

$$\text{Bias}\,(\bar{y}_{KBi}) = \bar{Y}f_i C_x^2\frac{k}{2}\left[1-k-2\rho_{xy}\frac{C_y}{C_x}\right] \tag{4.4}$$

where $k = \frac{\rho_{xy}S_y-\beta_1 S_x}{RS_x}$

and $f_1 = f_n, f_2 = f_r,\ f_3 = f_{rn}$.

***Proof*:** Proof is given in Appendix-1.

Theorem (4.2): The minimum mean square error of the proposed estimators T_{KB_i}, $i = 1,2,3$ is given by

$$MSE(T_{KBi}) = \left[f_r S_y^2 + f_i\left(S_x^2(kR+\beta_1)^2 - 2\rho_{xy}S_xS_y(kR+\beta_1)\right)\right]$$
$$\text{i} = 1,2,3 \ldots\ldots\ldots \tag{4.5}$$

For the optimum value k given by

$$k = \frac{\rho_{xy}S_y - \beta_1 S_x}{RS_x}$$

where,

$$R = \frac{\bar{Y}}{\bar{X}}\ , \beta_1 = \frac{S_x^2}{S_xS_y} \text{ and}$$

$$f_n = \left(\frac{1}{n} - \frac{1}{N}\right), f_r = \left(\frac{1}{r} - \frac{1}{N}\right) \ \& \ f_{rn} = \left(\frac{1}{r} - \frac{1}{n}\right).$$

The minimum MSE of the proposed estimator is given by

$$\text{Min MSE}(\bar{y}_{KBi}) = S_y^2\left[f_r - f_i * {\rho_{xy}}^2\right].$$

***Proof*:** Proof is given in Appendix-1.

5. A New Method of Imputation

The estimators rely on three different ratio-regression type methods of imputation as follows.

Case I: Auxiliary information on X is completely available, i.e., $\bar{X}$ is known and corresponding estimates $\bar{x}_n$ are used in the imputation technique.

$$y_{.i1} = \begin{cases} y_i & if \quad i\epsilon R \\ \dfrac{1}{n-r}\left[\left[n\gamma_1\bar{y}_r\left\{2 - \left(\dfrac{\bar{x}_n}{\bar{X}}\right)^k\right\}\right] + \beta_1(\bar{X} - \bar{x}_n) - \bar{y}_r\, r\right] & if \quad i\epsilon R^c \end{cases}$$

The resultant estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{KN1} = \left[\gamma_1\bar{y}_r\left\{2 - \left(\frac{\bar{x}_n}{\bar{X}}\right)^k\right\}\right] + \beta_1(\bar{X} - \bar{x}_n) \tag{5.1}$$

Case II: Auxiliary information on X is completely available i.e., $\bar{X}$ is known and corresponding estimates $\bar{x}_r$ are used in the imputation technique.

$$y_{.i2} = \begin{cases} y_i & if \quad i\epsilon R \\ \dfrac{1}{n-r}\left[n\gamma_1\bar{y}_r\left\{2 - \left(\dfrac{\bar{x}_r}{\bar{X}}\right)^k\right\} + \beta_1(\bar{X} - \bar{x}_r) - \bar{y}_r\, \text{r}\right] & if \quad i\epsilon R^c \end{cases}$$

The resultant estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{KN2} = \left[\gamma_1 \bar{y}_r \left\{2 - \left(\frac{\bar{x}_r}{\bar{X}}\right)^k\right\}\right] + \beta_1(\bar{X} - \bar{x}_r) \tag{5.2}$$

Case III: Auxiliary information on X is not available at the population level, i.e., $\bar{X}$ is not known and we use corresponding estimates $\bar{x}_n$, $\bar{x}_r$ in the imputation technique.

$$y_{.i3} = \begin{cases} y_i & if \quad i\epsilon R \\ \dfrac{1}{n-r}\left[n\gamma_1 \bar{y}_r \left\{2 - \left(\dfrac{\bar{x}_r}{\bar{x}_n}\right)^k\right\} + \beta_1(\bar{x}_n - \bar{x}_r) - \bar{y}_r r\right] & if \quad i\epsilon R^c \end{cases}$$

The resultant estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{KN3} = \left[\gamma_1 \bar{y}_r \left\{2 - \left(\frac{\bar{x}_r}{\bar{x}_n}\right)^k\right\}\right] + \beta_1(\bar{x}_n - \bar{x}_r) \tag{5.3}$$

Therefore, under the above situations, the properties of imputation methods discussed are as follow.

Theorem (5.1): The bias of the proposed ratio regression type estimators $\bar{y}_{KNi}$, $i = 1,2\ and\ 3$ is given by:

$$\text{Bias}\ (\bar{y}_{KNi}) = \left[(\gamma_1 - 1) - k f_i \left\{C_y^2 - \frac{(k-1)}{2} C_x^2\right\}\right] \tag{5.4}$$

where $k = \frac{\rho_{xy} S_y - \beta_1 S_x}{R S_x}$

and $f_1 = f_n, f_2 = f_r,\ f_3 = f_{rn}$.

Proof: Proof is given in Appendix-2.

Theorem (5.2): The minimum mean square error of the proposed ratio regression type estimators $\bar{y}_{KNi}$ $i = 1,2$ and 3 is given by

$$\text{MSE}\ (\bar{y}_{KNi}) = \bar{Y}^2(\gamma_1^2 A_i - 2\gamma_1 B_i + C_i) \qquad \text{i} = 1,2,3 \ldots\ldots \tag{5.5}$$

For the optimum value k given by

$$k = \frac{\rho_{xy} S_y - \beta_1 S_x}{R S_x}$$

where, $A_i = 1 + f_r C_y^2 + f_i(k C_x^2 - 4k\rho_{xy} C_x C_y)$

$$B_i = 1 - f_i\left(k\rho_{xy} C_x C_y - k\beta_1 \frac{\bar{X}}{\bar{Y}} C_x^2 + \beta_1 \frac{\bar{X}}{\bar{Y}} \rho_{xy} C_x C_y - \frac{k(k-1)}{2} C_x^2\right)$$

$$C_i = 1 + \beta_1^2 \frac{\bar{X}^2}{\bar{Y}^2} f_i C_x^2$$

and $\gamma_{1opt} = \frac{B_i}{A_i}$

The minimum MSE of the proposed estimator is given by

$$Min\ MSE(\bar{y}_{KNi}) = \bar{Y}^2\left(C_i - \frac{B_i^2}{A_i}\right).$$

Proof: Proof is given in Appendix-2.

6. Efficiency Comparison

The following conditions are derived for the theoretical comparison of the Mean Squared Error (MSE) of the proposed estimator with other existing estimators.

Strartegy I:

$$V(\bar{y}_r) - MSE(\bar{y}_{KB1}) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{REG}) - MSE(\bar{y}_{KB1}) = \left(\frac{2}{n} - \frac{1}{r} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{DP1}) - MSE(\bar{y}_{KB1}) = 0$$
$$MSE(\bar{y}_{DP2}) - MSE(\bar{y}_{KB1}) = \left(\frac{1}{n} - \frac{1}{r}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{DP3}) - MSE(\bar{y}_{KB1}) = \left(\frac{2}{n} - \frac{1}{r} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{GIRA}) - MSE(\bar{y}_{KB1}) = \left(\frac{2}{n} - \frac{1}{r} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$

Strategy II:

$$V(\bar{y}_r) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{r} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{REG}) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{DP1}) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{r} - \frac{1}{n}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{DP2}) - MSE(\bar{y}_{KB2}) = 0$$
$$MSE(\bar{y}_{DP3}) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{GIRA}) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$

Strategy III:

$$V(\bar{y}_r) - MSE(\bar{y}_{KB3}) = \left(\frac{1}{r} - \frac{1}{n}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{REG}) - MSE(\bar{y}_{KB3}) = 0$$
$$MSE(\bar{y}_{DP1}) - MSE(\bar{y}_{KB3}) = \left(\frac{1}{r} + \frac{1}{N} - \frac{2}{n}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{DP2}) - MSE(\bar{y}_{KB3}) = \left(\frac{1}{N} - \frac{1}{n}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{DP3}) - MSE(\bar{y}_{KB3}) = 0$$
$$MSE(\bar{y}_{GIRA}) - MSE(\bar{y}_{KB3}) = 0$$

Comparing the proposed estimators, even if they involve different source of information, after simple algebra we note that:

$$MSE(\bar{y}_{KB1}) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{r} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{KB3}) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{KB1}) - MSE(\bar{y}_{KB3}) = \left(\frac{1}{r} + \frac{1}{N} - \frac{2}{n}\right)\rho_{xy}^2 S_y^2 \geq 0$$

This means that $\bar{y}_{KB2}$ is always more efficient than both $\bar{y}_{KB1}$ and $\bar{y}_{KB3}$, whereas $\bar{y}_{KB3}$ performs better than $\bar{y}_{KB1}$ if the condition $\boldsymbol{r < \frac{nN}{2N-n}}$ is satisfied. The results are

valuable because they highlight the role of the auxiliary information in improving the estimates and afford sampling practitioners a useful indication on a profitable collection of auxiliary information in the case of missing data. The choice among competing estimators can be certainly facilitated by awareness of the information at hand.

7. Computational Study

We have divided the computations into two categories, namely, with real data and artifically generated data.

7.1. Empirical study using real data

In this section, an empirical study is performed in the presence of auxiliary variable where the performance of the proposed methods of imputation is compared with competing methods based on MSE and PRE. This study is carried out on six real data sets. We have computed and reported MSEs and percentage relative efficiency (PREs) of the proposed imputation methods with respect to the conventional methods to compare the proposed imputation methods with that of the existing imputation method that utilizes auxiliary information.

$$PRE(\bar{y}_{KBi}, \bar{y}_r) = \frac{MSE(\bar{y}_r)}{MSE(\bar{y}_{KBi})} \times 100 \quad and \quad PRE(\bar{y}_{KNi}, \bar{y}_r) = \frac{MSE(\bar{y}_r)}{MSE(\bar{y}_{KNi})} \times 100$$

Six different real data sets have been considered in the present empirical study. Data set 1 is taken from, Kadilar & Cingi (2008) with details on y as the level of apple production, and x as the number of apple trees.Data set 2 is taken from Diana & Perri (2010) with information on the Survey of Households Income and Wealth conducted by the Bank of Italy for the year 2002, y as the household's net disposable income, x as the number of household income earners. Data set 3 is taken from Source: [7] Page 228. The source of data set 4 is Singh (2009). The data set 5 is taken from Srivastava et. al. (1989) pp. 3922): y of weight of children, x as the skull circumference of children. Data set 6 is taken from ICMR, Department of Pediatrics, BHU, during 1983-84 of school children with study variable y as height (in kg) of the children, x, variable related to weight. The required values of the parameters for all six data sets are given in table 1.

Table 1: Population Parameters of Six Different Real Population.

Parameter	Population 1	Population 2	Population 3	Population 4	Population 5	Population 6
N	19	8011	80	3055	82	95
n	10	400	20	611	43	35
r	8	250	16	520	25	10
$\bar{Y}$	575	28229.43	51.8264	308582.4	11.90	115.9526
$\bar{X}$	13573.68	1.69	2.8513	56.5	39.80	19.4968
S_y	858.36	22216.56	18.3569	425312.8	0.5792685	5.966921
S_x	12945.38	0.78	2.7041	72.3	0.8581212	3.27346
ρ_{xy}	0.88	0.46	0.9150	0.677	0.009	0.713

These population have varying amount of correlation between study variate(y) and auxiliary variate(x) as shown in the table 1.

Table 2: Mean Square Errors of the Existing and Suggested Estimators

Estimator	Population 1	Population 2	Population 3	Population 4	Population 5	Population 6
Case I						
$\bar{y}_r$	53319.74	191268.93	16.85	288655815.71	0.00932998	3.185634
$\bar{y}_{RAT1}$	47051.53	1683371.53	76.31	290511133.40	0.01066795	3.471515
$\bar{y}_{REG1}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{SH1}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{SD1}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{SINGH1}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{DP1}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{GIRA1}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{KB1(proposed)}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{KN1(proposed)}$	28448.79	1209049.49	8.84	190771840.36	0.00890512	2.949472
Case II						
$\bar{y}_r$	53319.74	191268.93	16.85	288655815.71	0.009329982	3.185634
$\bar{y}_{RAT2}$	43743.31	1538549.33	96.14	290916984.15	0.01269343	4.603127
$\bar{y}_{REG2}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{SH2}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{SD2}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{SINGH2}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{DP2}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{GIRA2}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{KB2(proposed)}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{KN2(proposed)}$	18811.33	767278.91	6.20	169535367.50	0.008266157	2.017615
Case III						
$\bar{y}_r$	53319.74	191268.93	16.85	288655815.71	0.009329982	3.185634
$\bar{y}_{RAT3}$	50011.52	1767867.72	36.67	289061666.45	0.01135546	4.317246
$\bar{y}_{REG3}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{SH3}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{SD3}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{SINGH3}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{DP3}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{GIRA3}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{KB3(proposed)}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{KN3(proposed)}$	36826.3	1466695.66	14.11	266571886.52	0.008688242	2.253037

Table 3: Percentage Relative Efficiency of the Considered Estimators under Six Different populations

Estimator	Population 1	Population 2	Population 3	Population 4	Population 5	Population 6
Case I						
$\bar{y}_r$	100	100	100	100	100	100
$\bar{y}_{RAT1}$	152.7992	113.6226	22.07836	99.36136	87.45805	91.76495
$\bar{y}_{REG1}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{SH1}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{SD1}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{SINGH1}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{DP1}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{GIRA1}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{KB1(proposed)}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{KN1(proposed)}$	216.947	158.1978	190.683	151.3094	104.7709	108.0069
Case II						
$\bar{y}_r$	100	100	100	100	100	100
$\bar{y}_{REG2}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{SH2}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{SD2}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{SINGH2}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{DP2}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{GIRA2}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{KB2(proposed)}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{KN2(proposed)}$	435.1715	249.2822	271.9538	170.2629	112.8697	157.8911
Case III						
$\bar{y}_r$	100	100	100	100	100	100
$\bar{y}_{RAT3}$	122.305	108.1919	45.94658	99.8596	82.16294	73.78856
$\bar{y}_{REG3}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{SH3}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{SD3}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{SINGH3}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{DP3}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{GIRA3}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{KB3(proposed)}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{KN3(proposed)}$	152.4202	130.4081	119.4088	108.2844	107.3863	141.3929

7.2. Artificial population

The artificial population has been generated as described below

Population1. A population of size $N = 500$ with one study variable y and one auxiliary variable x is generated from the bivariate normal distribution where study variable y is correlated with auxiliary variables with various amount of $\rho_{yx} = 0.6, 0.7, 0.8$ and 0.9. The variables (y, x) are generated using MVNORM package in R software. A sample of size $n = 50$ is drawn from the population, with the number of responding units assumed to be $r = 30$.

Population 2. An artificial population is generated of size $N = 200$ which involves one study variable y and auxiliary variable x. Study variable y is correlated with auxiliary variables with various amount of $\rho_{yx} = 0.6, 0.7, 0.8$ and 0.9 The variables (y, x) is generated using MVNORM package in R software. From this population we draw sample of size $n = 26$, responding units are $r = 21$.

The percentage relative efficiencies (PRE) of the proposed estimators are computed through 50,000 repeated samples of size n as per imputation technique. In which we (i) draw a random sample of size n from population size N, (ii) from each selected sample $(n - r)$ units are dropped randomly, and (iii) the estimators and their MSE's are calculated for each sample and then averaged over all 50,000 samples .

The mean square error and percent relative efficiencies are given by

$$MSE(T_j) = \frac{1}{50000} \sum_{i=1}^{50000} \left(T_j(s_i) - \bar{Y}\right)^2 \quad j = 0, 1, 2, 3$$

$$PRE(T_j) = \frac{MSE(T_0)}{MSE(T_j)} \times 100 \quad j = 1, 2, 3$$

based on 50,000 repeated samples.

Table 4: Mean square error and percentage relative efficiency based on

Population 1 (Artificially generated normal population)								
(N=500, n=50, r=30)								
Correlation	0.9		0.8		0.7		0.6	
Estimator	MSE	PRE	MSE	PRE	MSE	PRE	MSE	PRE
Mean per unit($\bar{y}_r$)	3.1962	100	3.0434	100	3.23365	100	2.62266	100
Ratio method ($\bar{y}_{RAT}$)	2.3856	133.978	2.4749	122.973	2.83633	114.008	2.34778	111.7081
Regression Method ($\bar{y}_{REG}$)	2.2519	141.9321	2.4259	125.455	2.85815	113.138	2.35706	111.2686
Proposed imputation ($\bar{y}_{KNi}$)	2.1520	148.5223	2.2456	135.529	2.44280	132.375	2.13541	122.8176

Table 4: Mean square error and percentage relative efficiency based on (cont.)

Population 2 (Artificially generated normal population)								
(N=200, n=26, r=21)								
Correlation	0.9		0.8		0.7		0.6	
Estimator	MSE	PRE	MSE	PRE	MSE	PRE	MSE	PRE
Mean per unit($\bar{y}_r$)	4.7080	100	4.5813	100	4.5712	100	4.0428	100
Ratio method ($\bar{y}_{RAT}$)	4.1007	114.810	4.1094	111.484	4.1587	109.9177	3.8485	105.049
Regression Method ($\bar{y}_{REG}$)	4.0259	116.943	4.0721	112.504	4.1552	110.0103	3.8514	104.972
Proposed imputation ($\bar{y}_{KNi}$)	3.7094	126.921	3.5807	127.944	3.7578	121.6439	3.4201	118.208

8. Interpretations of the Computational Results

In this article, it is clear that MSE of the proposed alternative estimator is more efficient than the mean estimator. In addition, the proposed estimator is always more efficient than the usual ratio estimator. We note that the proposed method is free from the assumptions of a model for the ratio method of imputation. In addition, MSE is similar to the other mentioned estimators MSE ($\bar{y}_{SH}$) = MSE ($\bar{y}_{SD}$) = MSE ($\bar{y}_{Singh}$) = MSE($\bar{y}_{DPi}$) = MSE($\bar{y}_{GIRA}$) = MSE($\bar{y}_{KBi}$). A new method of imputation is introduced that remains more efficient than conventional and existing imputation methods in the presence of auxiliary variables. The following interpretations are made based on empirical results summarized in Table 3.

1. We introduce an alternative method of imputation and the resultant estimator in the presence of non-response. The performance of the proposed estimator is justified theoretically and numerically. Table 2 & 3 expressed that the relative efficiency of the proposed estimator ($\bar{\boldsymbol{y}}_{\boldsymbol{KB1}}, \bar{\boldsymbol{y}}_{\boldsymbol{KB2}}$ *and* $\bar{\boldsymbol{y}}_{\boldsymbol{KB3}}$) performs better than the mean and the ratio estimators and is equivalent to other mentioned estimators.
2. For all the populations 1 to 6, Table 2 & 3 exhibit the superiority of the proposed imputation method ($\bar{\boldsymbol{y}}_{\boldsymbol{KN1}}, \bar{\boldsymbol{y}}_{\boldsymbol{KN2}}$ *and* $\bar{\boldsymbol{y}}_{\boldsymbol{KN3}}$) over the mean and ratio type imputation method. Also, the proposed method of imputation ($\bar{\boldsymbol{y}}_{\boldsymbol{KN1}}, \bar{\boldsymbol{y}}_{\boldsymbol{KN2}}$ *and* $\bar{\boldsymbol{Y}}_{\boldsymbol{KN3}}$) is superior to the Diana and Perri regression type imputation method and the proposed alternative method ($\bar{\boldsymbol{y}}_{\boldsymbol{M1}}, \bar{\boldsymbol{y}}_{\boldsymbol{M2}}$ *and* $\bar{\boldsymbol{y}}_{\boldsymbol{M3}}$) of imputation.
3. The proposed new imputation method ($\bar{\boldsymbol{y}}_{\boldsymbol{KN1}}, \bar{\boldsymbol{y}}_{\boldsymbol{KN2}}$ *and* $\bar{\boldsymbol{y}}_{\boldsymbol{KN3}}$), in all the populations 1 to 6, as shown in Table 2 & 3, has achieved considerable gain

in performance over the conventional imputation method for all the three cases, namely $\bar{y}_{KN1}$ is considerably better than $\bar{y}_{SH}$(I), $\bar{y}_{SD}$(I), and $\bar{y}_{DP}$(I), in case I. Similarly, $\bar{y}_{KN2}$ is considerably superior to $\bar{y}_{SH}$(II), $\bar{y}_{SD}$(II), and $\bar{y}_{DP}$(II), imputation methods in case II and $\bar{y}_{KN3}$ is considerably superior to $\bar{y}_{SH}$(III), $\bar{y}_{SD}$(III), and $\bar{y}_{DP}$(III), in case III.

4. It is important to note that Table 2 & 3 exhibit that the proposed imputation method $(\bar{\boldsymbol{y}}_{\boldsymbol{KN1}}, \bar{\boldsymbol{y}}_{\boldsymbol{KN2}}\ \boldsymbol{and}\ \bar{\boldsymbol{y}}_{\boldsymbol{KN3}})$ when applied to six real data sets, and compared with conventional and recent imputation methods, attains considerable gain in efficiency over competing for imputation methods for the case II, and gives better results over other cases, namely, case I and case III.
5. This empirical study confirms the superiority of the proposed imputation method over Diana & Perri's (2010) imputation method and other landmark imputation methods that use auxiliary information.

It can be noted that the proposed method of imputation $(\bar{\boldsymbol{y}}_{\boldsymbol{KN1}}, \bar{\boldsymbol{y}}_{\boldsymbol{KN2}}\ \boldsymbol{and}\ \bar{\boldsymbol{y}}_{\boldsymbol{KN3}})$ is easy to use and rewarding in terms of efficiency and deals with the problem of non-response. Survey practitioners can use the proposed imputation method to deal with the problem of nonresponse and get a high gain in efficiency when one has access to auxiliary information.

References

Kadilar, C., Cingi, H., (2008). Estimators for the Population Mean in the Case of Missing Data. *Communications in Statistics—Theory and Methods*, 37, pp. 2226–2236. http://dx.doi.org/10.1080/03610920701855020.

Diana, G., Perri, P. F., (2010). Improved Estimators of the Population Mean for Missing Data, *Communications in Statistics—Theory and Methods*, 39, pp. 3245–3251. http://dx.doi.org/10.1080/03610920903009400.

Gira, Abdeltawab, A., (2015). Estimation of population mean with a new imputation methods. *Applied Mathematical Sciences*, 9(34), pp. 1663–1672.

Chodjuntug K., Lawson, N., (2022). Imputation for estimating the population mean in the presence of nonresponse, with application to fine particle density in Bangkok. *Mathematical Population Studies*, 29(4), pp. 204–225.

Chodjuntug K., Lawson, N., (2022). A chain regression exponential type imputation method for mean estimation in the presence of missing data. *Songklanakarin Journal of Science and Technology*, 44(4), pp. 1109–1118.

Heitzan, D. F., Basu, S., (1996). Distinguishing 'Missing at Random' and 'Missing Completely At Random'. *The American Statistician,* 50, pp. 207–213.

Kalton, G., Kasprzyk, D., (1982). Imputing for missing survey responses, In: Proceedings of the section on survey research method. *American Statistical Association*, pp. 22–31.

Kalton, G., Kasprzyk, D. and Santos, R., (1981). Issues of nonresponse and imputation in the survey of income and program participation, In: Krewski D, Platek R, Rao JNK (eds) Current topics in survey sampling. *Academic Press, New York*, pp. 455–480.

Lee, H., Rancourt, E. and Särndal, C. E., (1994). Experiments with Variance Estimation from Survey Data with Imputed Values. *Journal of Official Statistics,* 10, pp. 231–243.

Lee, H., Rancourt, E., Sarndal, C. E., (1995). Variance estimation in the presence of imputed data for the generalized estimation system. *In: Proceedings of the section on survey research methods*, American Statistical Association.

Murthy, M. N., (1967). Sampling theory and methods. *Statistical publishing Society*, Calcutta, India.

Lawson, N., (2023). New imputation method for estimating population mean in the presence of missing data. *Lobachevskii Journal of Mathematics*, 44(9), pp. 3740–3748.

Lawson, N., (2023). A class of population mean estimators in the presence of missing data with applications to air pollution in Chiang Mai, Thailand. *Lobachevskii Journal of Mathematics*, 44(9), pp. 3749–3757.

Thongsak, N., Lawson, N., (2023). A new imputation method for population mean in the presence of missing data based on a transformed variable with applications to air pollution data in Chiang Mai, Thailand. *Journal of Air Pollution and Health*, 8(3), pp. 285–298.

Rao, J. N. K, Sitter, R. R., (1995). Variance Estimation under Two-Phase Sampling with Application to Imputation for Missing Data. *Biometrika*, 82, pp. 453–460. http://dx.doi.org/10.1093/biomet/82.2.453

Rubin, R. B., (1976). Inference and missing data. *Biometrika*, 63, pp. 581–592. http://dx.doi.org/10.1093/biomet/63.3.581

Singh, S., (2009). A new method of imputation in survey sampling. Statistics: *A Journal of Theoretical and Applied Statistics*, 43, pp. 499–511. http://dx.doi.org/ 10.1080/ 02331880802605114.

Singh, S., Deo, B., (2003). Imputation by Power Transformation. *Statistical Papers*, 44, pp. 555–579. http://dx.doi.org/10.1007/bf02926010.

Singh, S., Horn, S., (2000). Compromised Imputation in Survey Sampling. *Metrika*, 51, pp. 267–276. http://dx.doi.org/10.1007/s001840000054.

Srivenkataramana, T., Tracy, D. S., (1980). An Alternative to Ratio Method in Sample Surveys. *Annals of the Institute of Statistical Mathematics*, 32, pp. 111–120. http://dx.doi.org/10.1007/ bf02480317.

Appendix-1

Proof: proof of theorem 4.1(Bias of the proposed estimator). The line of proof here is worked out for the estimator defined under case 1, The estimator $\bar{y}_{KN1}$ can be written be as follows.

$$\begin{aligned}\bar{y}_{KN1} &= \bar{Y}(1+\varepsilon_o)[2-(1+\eta_o)^K] - \beta_1\bar{X}\eta_o \\ &= \bar{Y}(1+\varepsilon_o)\left[2-(1+K\eta_o+\frac{K(K-1)}{2}\eta_o^2+\cdots)\right] - \beta_1\bar{X}\eta_o \\ &= \bar{Y}(1+\varepsilon_o)\left(1-K\eta_o-\frac{K(K-1)}{2}\eta_o^2+\cdots\right) - \beta_1\bar{X}\eta_o \qquad (A) \\ &= \bar{Y}\left(1+\varepsilon_o-K\eta_o-K\varepsilon_o\eta_o-\frac{K(K-1)}{2}\eta_o^2-\beta_1\frac{\bar{X}}{\bar{Y}}\eta_o+O(\varepsilon_o^2)\right)\end{aligned}$$

Neglecting the higher order of approximation, the bias

$$\begin{aligned}B(\bar{y}_{KN1}) &= E(\bar{y}_{KN1}-\bar{Y}) \\ &= \bar{Y}E\left(\varepsilon_o-K\eta_o-K\varepsilon_o\eta_o-\frac{K(K-1)}{2}\eta_o^2-\beta_1\frac{\bar{X}}{\bar{Y}}\eta_o\right) \qquad (B)\end{aligned}$$

Taking the expectations of (B), we get (4.4) for i $= 1$, which prove theorem (4.1).

The derivation of other estimators $\bar{y}_{KNi}$ $(i = 2\ \&\ 3)$ can be carried out in a similar way.

Proof: proof of theorem 4.2

The MSE of $\bar{y}_{M1}$ can be found up to the first order of approximation by rewriting as follow:

$$\begin{aligned}MSE\ (\bar{y}_{KN1}) &= E(\bar{y}_{KN1}-\bar{Y})^2 \\ &= \bar{Y}^2E\left[\varepsilon_o-K\eta_o-K\varepsilon_o\eta_o-\frac{K(K-1)}{2}\eta_o^2-\beta_1\frac{\bar{X}}{\bar{Y}}\eta_o\right]^2 \\ &= \bar{Y}^2E\left[\varepsilon_o^2+K^2\eta_o^2+\beta_1^2\frac{\bar{X}^2}{\bar{Y}^2}\eta_o^2-2K\varepsilon_o\eta_o+2K\beta_1\frac{\bar{X}}{\bar{Y}}\eta_o^2-2\beta_1\frac{\bar{X}}{\bar{Y}}\varepsilon_o\eta_o\right] \\ &= \left[\left(\frac{1}{r}-\frac{1}{N}\right)S_y^2+\left(\frac{1}{n}-\frac{1}{N}\right)\left(S_x^2(KR+\beta_1)^2-2\rho_{xy}S_xS_y(KR+\beta_1)\right)\right] \quad \text{…..(C)}\end{aligned}$$

where, R $= \frac{\bar{Y}}{\bar{X}}$

Differentiating equation (C) with respect to K and equating to zero, we get

$$K = \frac{\rho_{xy}S_y-\beta_1S_x}{RS_x} = K(optimum)$$

then substitute the value of optimum K in equation (4.8), thus the resulting minimum mean square error of $\bar{y}_{M1}$ is given by

$$\text{Min MSE}(\bar{y}_{KN1}) = S_y^2\left[f_r-f_n*{\rho_{xy}}^2\right]$$

Appendix-2

Proof: Proof of theorem 5.1

The estimator $\bar{y}_{KB1}$ can be written be as follows.

$$\begin{aligned}\bar{y}_{KB1} &= \gamma_1\bar{Y}(1+\varepsilon_o)[2-(1+\eta_o)^K]-\beta_1\bar{X}\eta_o \\ &= \gamma_1\bar{Y}(1+\varepsilon_o)\left[2-(1+K\eta_o+\tfrac{K(K-1)}{2}\eta_o^2+\cdots)\right]-\beta_1\bar{X}\eta_o \\ &= \gamma_1\bar{Y}(1+\varepsilon_o)\left(1-K\eta_o-\tfrac{K(K-1)}{2}\eta_o^2+\cdots\right)-\beta_1\bar{X}\eta_o \qquad \text{(D)} \\ &= \bar{Y}\left(\gamma_1+\gamma_1\varepsilon_o-\gamma_1K\eta_o-\gamma_1K\varepsilon_o\eta_o-\gamma_1\tfrac{K(K-1)}{2}\eta_o^2-\beta_1\tfrac{\bar{X}}{\bar{Y}}\eta_o+O(\varepsilon_o^2)\right)\end{aligned}$$

Neglecting the higher order of approximation, the bias

$$\begin{aligned}B(\bar{y}_{KB1}) &= (\bar{y}_{KB1}-\bar{Y}) \\ &= \bar{Y}E\left(\gamma_1+\gamma_1\varepsilon_o-\gamma_1K\eta_o-\gamma_1K\varepsilon_o\eta_o-\gamma_1\tfrac{K(K-1)}{2}\eta_o^2-\beta_1\tfrac{\bar{X}}{\bar{Y}}\eta_o-1\right) \qquad \text{(E)}\end{aligned}$$

Taking the expectations of (E), we get the (5.4) for $i=1$, which proves theorem (5.1)

The derivation of other estimators $\bar{y}_{KBi}$ $(i=2\ \&\ 3)$ can be carried out in a similar way

Theorem (5.2): The minimum mean square error of the proposed ratio regression type estimators $\bar{y}_{KB1}$ up to the first order approximation is given by

$$\text{MinMSE}\ (\bar{y}_{KB1}) = \bar{Y}^2\left[C_1-\left(\frac{B_1^2}{A_1}\right)\right] \qquad (5.7)$$

For the optimum value K given by

$$K=\frac{\rho_{xy}S_y-\beta_1S_x}{RS_x}$$

Proof: The MSE of $\bar{y}_{M11}$ can be found up to the first order of approximation by rewriting as follows:

$$MSE\ (\bar{y}_{KB1}) = E(\bar{y}_{KB1}-\bar{Y})^2$$

$$= \bar{Y}^2E\left[\gamma_1-\gamma_1K\eta_o-\gamma_1\tfrac{K(K-1)}{2}\eta_o^2+\gamma_1\varepsilon_o-\gamma_1K\varepsilon_o\eta_o-\beta_1\tfrac{\bar{X}}{\bar{Y}}\eta_o-1\right]^2$$

$$= \bar{Y}^2E\left[1+{\gamma_1}^2\left(1+\varepsilon_o^2+K\eta_o^2-4K\varepsilon_o\eta_o\right)-2\gamma_1\left(1-K\varepsilon_o\eta_o-K\beta_1\tfrac{\bar{X}}{\bar{Y}}\eta_o^2-\tfrac{K(K-1)}{2}\eta_o^2+\beta_1\tfrac{\bar{X}}{\bar{Y}}\varepsilon_o\eta_o\right)+\beta_1^2\tfrac{\bar{X}^2}{\bar{Y}^2}\eta_o^2\right]$$

$$= \bar{Y}^2\left[\begin{matrix}1+{\gamma_1}^2\{1+f_rC_y^2+f_n(KC_x^2-4K\rho_{xy}C_xC_y)\}-2\gamma_1 \\ \left\{1-f_n\left(K\rho_{xy}C_xC_y-K\beta_1\tfrac{\bar{X}}{\bar{Y}}C_x^2+\beta_1\tfrac{\bar{X}}{\bar{Y}}\rho_{xy}C_xC_y-\tfrac{K(K-1)}{2}C_x^2\right)\right\}+\beta_1^2\tfrac{\bar{X}^2}{\bar{Y}^2}f_nC_x^2\end{matrix}\right]..(F)$$

Differentiate equation (F) with respect to K when $\gamma_1$equating to 1, we get

$$K = \frac{\rho_{xy}S_y - \beta_1 S_x}{RS_x}$$

For optimum value of γ_1 differentiating the equation(G) with respect to and equating to zero, we get $\gamma_{1opt} = \frac{B_1}{A_1}$

Substituting the optimum value of γ_{1opt} in equation (G), we get minimum MSE

$$\text{Min. MSE}(\bar{y}_{K1}) = \overline{Y}^2 \left(C_1 - \frac{B_1^2}{A_1}\right)$$

we get the (5.7) for $i = 1$ that prove theorem (5.2)

The derivation of other estimators $T_{KBi\ (i=2\ \&\ 3)}$ can be drive on similar lines.

In general, we have

$$MSE\ (\bar{y}_{KBi}) = \overline{Y}^2(\gamma_1^2 A_i - 2\gamma_1 B_i + C_i)$$

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 199–213, https://doi.org/10.59139/stattrans-2025-022
Received – 03.03.2025; accepted – 11.04.2025

Harnessing AI for business transformation: strategies for effective implementation and market advantage

Jakub Kubiczek[1], Elżbieta Roszko-Wójtowicz[2], Julianna Koczy[3], Izabela Waszkiewicz[4], Klaudia Woś[5]

Abstract

The rapid digitalization of consumer behavior presents businesses with unprecedented opportunities and challenges. Artificial Intelligence (AI) has emerged as a key driver of efficiency, enabling companies to analyze vast amounts of consumer data, personalize experiences and enhance decision-making processes. The paper aims to examine how AI-driven tools contribute to business transformation, focusing on their impact on operational efficiency, customer engagement and market competitiveness. The research employs a multi-method approach, including literature reviews, secondary data analysis and case studies of AI implementation in enterprises operating on the Polish market. Findings highlight the dual nature of AI adoption: while it enhances productivity, accuracy and sustainability, businesses must also navigate the risks related to data security, compliance and financial feasibility. The study underscores the importance of dynamic capabilities in leveraging AI for strategic growth while mitigating the associated challenges. The results contribute to the discourse on AI's role in shaping modern e-commerce, offering practical insights for companies seeking to integrate AI-driven solutions effectively.

Key words: artificial intelligence (AI), digital consumer behavior, dynamic capabilities, AI implementation in business, e-commerce and AI-driven personalization.
JEL: M15; L81; O33.

[1] University of Economics in Katowice, Katowice, Poland. E-mail: jakub.kubiczek@uekat.pl. ORCID: https://orcid.org/0000-0003-4599-4814.

[2] University of Lodz, Lodz, Poland. E-mail: elzbieta.roszko@uni.lodz.pl. ORCID: https://orcid.org/0000-0001-9337-7218.

[3] University of Economics in Katowice, Katowice, Poland. E-mail: Julianna.koczy@edu.uekat.pl. ORCID: https://orcid.org/ 0009-0000-8266-4019.

[4] Independent researcher, Poland. E-mail: izakwapisz@wp.pl. ORCID: https://orcid.org/0009-0002-8700-7370.

[5] University of Economics in Katowice, Katowice, Poland. E-mail: klaudia.wos@edu.uekat.pl. ORCID: https://orcid.org/0009-0008-0545-4446.

1. Introduction

Changes in the business environment have always shaped the conditions under which companies operate. These changes impact various aspects of organizational functioning, requiring firms to remain flexible and adaptive. Business transformation can be understood through two key dimensions: turbulence and complexity (Mason and Staude, 2009; Reed, 2022). Turbulence refers to short-term, dynamic shifts with immediate consequences, while complexity encompasses long-term developments such as technological and societal advancements. These dimensions are interconnected, as seen in the COVID-19 pandemic, which accelerated the digitalization of consumer behavior (Roszko-Wójtowicz et al., 2024b; Borsiak-Dańska et al., 2024; Roszko-Wójtowicz et al., 2024a; Amankwah-Amoah et al., 2021).

In response to these evolving conditions, organizations must develop dynamic capabilities, i.e. the ability to integrate, build and reconfigure internal and external resources to adapt to the changing environments. In the era of rapid technological advancements, artificial intelligence (AI) has become a key enabler of these dynamic capabilities, allowing companies to respond effectively to shifting consumer behaviors and market demands (Dubey et al., 2020). As digital consumer behavior continues to evolve, businesses face both opportunities and challenges in leveraging AI for enhanced operational efficiency and strategic decision-making.

A major advantage of AI is its ability to process vast amounts of consumer data, providing companies with insights into customer preferences, behaviors and emerging market trends (Mikalef et al., 2021). Some researchers argue that AI is reshaping competitive landscapes by fostering innovation and improving business performance (Dwivedi et al., 2021). With the use of AI, entrepreneurs can gain deeper insights into consumer expectations, fulfill those expectations more effectively and identify new market segments (Haleem et al., 2022). By automating processes, enhancing decision-making and enabling hyper-personalization, AI-driven solutions offer businesses a pathway to increased productivity and customer engagement. However, these technologies also pose significant risks, including data security concerns, legal compliance issues and potential overdependence on automation.

While the academic discourse largely focuses on the risks associated with AI, particularly in the areas of cybersecurity and legal regulations aimed at protecting consumers, relatively little attention is given to examining successful AI implementations and their tangible benefits for businesses. In particular, there is a research gap in studies that explore how well-designed and effectively deployed AI solutions can create value for organizations. An essential aspect of these considerations is the digital context, as AI currently influences organizational value primarily in the digital realm. This paper explores the characteristics of digitized consumer purchasing behaviors and illustrates

how companies can enhance operational efficiency in the digital age through the utilization of AI. It incorporates a retrospective perspective, acknowledging the dynamic nature of technological change, where what was considered cutting-edge a year ago may already seem outdated. However, the true value of the paper does not rest solely on the timeliness of the information but rather on the real-world examples that showcase AI applications in modern business environments, addressing the existing research gap.

The paper aims to examine how AI-driven tools contribute to business transformation, focusing on their impact on operational efficiency, customer engagement, and market competitiveness, particularly in organizations operating in the rapidly evolving Polish digital market, where AI adoption is accelerating. To achieve this objective, the study explores the following key questions:

- How do digitized consumer purchasing behaviors create opportunities for modern technology adoption?
- How can AI be utilized to enhance business efficiency?
- What are the primary benefits and risks associated with AI implementation?
- What is the overall impact of AI on business operations?

The paper explores digitized consumer purchasing behaviors and their influence on AI-driven business strategies. It examines the role of dynamic capabilities in AI adoption, focusing on how businesses leverage AI for adaptive decision-making. The discussion then shifts to the benefits and risks of AI, highlighting both the opportunities and challenges of its use. The study employs a multi-method approach, including literature reviews, secondary data analysis and case studies of AI implementation in enterprises present on the Polish market. Such methods are feasible due to the availability of information on successful AI deployments. The paper concludes by analyzing the real-world impact of these implementations and discussing their broader implications and future research directions.

2. Digital consumer behavior: e-commerce, AI, and social media influence

In the current era of digitalization, e-commerce platforms have become an integral part of daily life, leading consumers to expect not only convenient and efficient transactions but also personalized and engaging experiences (He and Liu, 2024). The widespread use of devices such as computers and smartphones, which enable consumers to access online sales offers, has made online shopping the preferred choice for many. Young consumers, particularly Generation Z, have grown up in a world where the Internet is an inseparable part of reality. Technology has always been present in their lives, allowing them to adapt seamlessly to new digital solutions. In their e-commerce choices, they trust reviews and recommendations from other users and

often seek shopping inspiration on social media (Paczka, 2020). According to SW Research (2024), 86.2% of Generation Z consumers shop online at least once a month or more frequently. Additionally, research by Gemius (2023) shows that 79% of all Polish Internet users shop online. Therefore, a company's online presence is not merely a complement to physical sales but a crucial sales channel in its own right.

Beyond this generational shift, another significant factor contributing to the growth of online shopping was the COVID-19 pandemic (Kowal and Świątek, 2023; Dańska-Borsiak et al., 2024; Roszko-Wójtowicz et al., 2024a). According to research by Namogoo (2020), the pandemic prompted a surge in online shopping, with 56% of respondents increasing their online spending and 14% making online purchases for the first time. The crisis forced businesses to rapidly adapt to new consumer expectations and market conditions. As reported by Adobe (2020), online sales in the USA reached $813 billion in 2020, reflecting a 42% year-over-year increase. Meanwhile, PostNord (2021) found that in 2020 Europe saw the highest e-commerce market growth in countries such as Spain (44%), Belgium (41%) and Italy (37%), with Poland experiencing a 33% increase.

E-commerce is dominated by global platforms such as Amazon and Shopify, which provide businesses with extensive market reach (Ballerini et al., 2024). In Poland, Allegro is the largest e-commerce platform, leading in both the number of active users and the variety of products offered (Kubiczek et al., 2024). By leveraging AI-driven recommendation systems, these platforms enhance online sales and assist consumers in finding products aligned with their preferences (Zhu et al., 2022). Similar recommendation systems are widely employed by entertainment service platforms such as Netflix and Spotify, which use AI to personalize their content for users.

When analyzing digitalized purchasing behaviors, it is essential to consider the growing role of social media, which has evolved from a mere communication tool to an integral sales channel. Research by the Reuters Institute for the Study of Journalism (2023) indicates that 41% of respondents aged 18–24 cite social media as their primary source of information, marking a 23% increase since 2015. At the same time, researchers emphasize that these rapid transformations primarily affect younger demographics. The relationship between social media and e-commerce is increasingly symbiotic, as businesses integrate AI-driven strategies to target consumers through personalized advertising, influencer marketing and direct sales on social platforms.

3. Dynamic capabilities and the adoption of modern technologies

A proper understanding of environmental changes is fundamental to determining the appropriate method of adaptation. In a dynamic environment, agile and flexible

approaches become increasingly important, whereas in a stable and predictable environment, well-established processes may suffice (Mason and Staude, 2009). The COVID-19 pandemic has underscored the volatility of the e-commerce sector, demonstrating how unexpected disruptions can reshape business operations. Furthermore, globalization, internationalization and digitalization have interconnected markets to such an extent that even minor turbulence in one area can have immediate ripple effects on other. Therefore, the ability to respond swiftly and effectively to the changing conditions is essential for organizations operating in today's unstable and highly competitive business environment (Dyduch et al., 2021).

In an era of widespread digitalization, online business operations may seem like a natural progression. However, research suggests that digitalization is not merely an inevitable development but rather a strategic factor that enhances value creation and capture (Dyduch et al., 2023; Saura et al., 2022). To gain a competitive edge, companies often prioritize external partnerships, particularly strategic alliances, while ensuring they can agilely reallocate resources to seize any emerging opportunities. From a technological standpoint, dynamic capabilities extend beyond digital transformation and encompass the integration of advanced technologies such as AI. AI can be seen as the next evolutionary stage of industrial digitalization and digital servitization (Sjödin et al., 2023).

The increasing adoption of AI by competitors necessitates organizational adaptation to prevent the risk of falling behind. The democratization of AI has made the technology widely accessible, eliminating the need for advanced programming expertise to utilize complex algorithms. As a result, companies that successfully anticipate and embrace change are not only able to survive but also thrive in evolving market conditions. Industrial AI capabilities leverage digital technologies to unlock new value creation opportunities and drive revenue growth (Parida et al., 2019). Organizations that effectively integrate AI into their processes can gain a substantial competitive advantage, respond more effectively to the shifting consumer needs and better capitalize on the emerging market trends. Thus, AI enables businesses to transform challenges into strategic opportunities, strengthening their position on the market.

It is essential to recognize that modern technologies should not be viewed as standalone solutions capable of automatically enhancing company performance (Schweikl and Obermaier, 2023). As with business digitalization, organizations must adopt a systemic approach that prioritizes value creation and resource reconfiguration (Amit and Han, 2017). This requires a management strategy that carefully balances the benefits and potential risks associated with AI implementation, ensuring its effective and sustainable integration into business operations.

4. AI in organizational operations: key benefits and challenges

The implementation of AI-based tools can provide organizations with numerous benefits, ranging from increased efficiency and cost reduction to improved customer service and enhanced competitiveness through innovation, which is becoming increasingly important from the employers' perspective (Florczak-Strama, 2024, Nabila et al., 2021). These solutions not only enable the creation of new, technologically advanced products and services but also significantly enhance organizational competitiveness by directly improving operational efficiency and enabling better adaptation to market demands. Table 1 outlines the main benefits of AI implementation in enterprises.

Table 1: Main benefits of AI implementation

Dimension	Description
Productivity	AI can assist in automating both repetitive tasks and the decision-making process flow, thereby optimizing the received information and enhancing efficiency and productivity.
Accuracy	AI can not only monitor human work to reduce the number of human errors but also perform individual tasks with greater precision.
Cost efficiency	By increasing the efficiency and accuracy of processes, AI can help minimize costs.
Customer satisfaction	AI can more precisely tailor offers to individual customer expectations and needs through the analysis of historical data.
Innovation	AI can help develop new, technologically advanced products and services, directly contributing to creating an attractive work environment.
Sustainability	Within supported processes, AI can monitor energy use and assist in sustainable resource management. AI can support the design of products with their lifecycle in mind, reducing the amount of raw materials used and thus minimizing waste.

Source: authors' based on Dey, 2024; Makarius et al., 2020; McKinsey & Company & Forbes Polska, 2017; Ransbotham et al., 2019.

The implementation of AI-based solutions offers numerous opportunities to enhance organizational efficiency. These systems can automate routine tasks, minimize human errors, improve customer relations, support the design of technologically advanced products and contribute to the sustainable development of enterprises. Organizations can maximize these benefits by developing comprehensive AI implementation strategies, including the creation of their own innovation roadmaps. This structured approach helps companies effectively identify areas with the greatest potential for AI deployment.

To ensure responsible, safe and effective AI implementation, several preliminary actions must be undertaken. It is essential to ensure data protection, provide employee training and manage organizational change processes. Additionally, potential risks must be identified, assessed and managed proactively. Table 2 presents the main risks associated with AI implementation.

Table 2: Main risks of AI implementation

Dimension	Description
Privacy and data security	AI-based solutions utilize vast amounts of data for training and optimizing their algorithms. Processing the data involves the risk of violating privacy and data security.
Overdependence on technology	High automation within an enterprise can lead to a loss of human capital. This, in turn, poses significant risks during crisis situations that require human intervention.
Higher than expected costs	Developing, purchasing and implementing AI systems may require substantial capital investments, including debt and interest costs, which may not be compensated in the short term. Additionally, inefficient implementations can result in additional costs.
Insufficient process efficiency	There is a risk of improperly integrating AI into business processes or discovering that AI does not achieve the expected efficiency.
Compliance with legal regulations	AI regulations are continually evolving, necessitating regular monitoring and adaptation to new laws.

Source: authors' work based on Deloitte, 2021; Westerman et al., 2024.

Despite the many benefits of AI-based solutions, their implementation also presents several significant challenges for enterprises. Some of these challenges could potentially offset the benefits if not properly managed. One of the most pressing issues is proper data management, which requires data standardization and secure processing. Since high-quality data are fundamental to effective AI solutions, any deficiencies in data quality can undermine the performance of AI.

Another challenge is over-reliance on AI systems, which must always be properly supervised by humans to prevent costly errors and operational disruptions in the event of system failures. While AI can help optimize costs, it also entails significant financial challenges, such as high initial costs, integration expenses and the rising cost of AI-related services, which may negatively impact profitability. Additionally, inefficiencies in AI automation can arise from poorly selected processes or improper system configurations.

Ensuring compliance with evolving AI regulations is another critical concern. Since AI-related laws and regulations are continuously developing, organizations must actively monitor and adapt to regulatory changes to maintain legal compliance.

Predicting how enterprises will integrate AI into their processes in the future remains challenging. However, it is undeniable that the success of businesses will increasingly depend on AI-driven processes. The use of advanced technologies is already essential for companies aiming to maintain competitiveness in the market (Nowakowska, 2024; Saura et al., 2022). Therefore, enterprises must adopt a comprehensive approach to AI implementation, carefully balancing short- and long-term benefits and risks. To achieve this, it is crucial to analyze current AI implementations to gain a deeper understanding of the integration process and its impact on business operations.

5. AI solutions in practice: implementation examples and business applications

Currently, some of the most popular AI-driven approaches used in consumer data analysis and purchasing trend prediction include neural networks and machine learning (ML). These algorithms enable real-time processing of vast amounts of data used for classifying customers and products (Bielińska-Dusza, 2022). Deep learning (DL), a subset of ML that utilizes multi-layered neural networks, is widely employed in recommendation systems to personalize offers (Steck et al., 2021). The effectiveness of these advanced algorithms is largely driven by the abundance of consumer preference and behavior data generated through digital interactions. As a result, ML- and DL-based recommendation systems can deliver highly personalized content and product suggestions.

Another category of AI tools includes chatbots, which enhance communication between businesses and customers. The use of chatbots not only reduces operational costs for businesses but also facilitates customer acquisition (Schneider and Janowska, 2020). Chatbots are particularly effective in intensifying customer engagement and enhancing user experience, thereby increasing customer satisfaction and loyalty (Kaczorowska-Spychalska, 2019).

Many organizations also utilize sentiment analysis to improve decision-making efficiency in response to rapidly changing market trends and consumer preferences. This subfield of natural language processing (NLP) collects and analyzes subjective information, opinions, thoughts and impressions from consumers regarding specific products or services. By analyzing sentiment, these algorithms help businesses understand and interpret consumer emotions expressed in text (Turek, 2017). Sentiment analysis is particularly valuable in social media monitoring, enabling companies to track public opinions and sentiments regarding new product launches, marketing campaigns or brand-related events.

The implementation of these AI solutions is most visible among leading companies, which leverage AI to gain a competitive advantage. Table 3 presents selected examples of AI applications in enterprises.

Table 3: Selected uses of AI

Company	Description
Allegro	Utilizes algorithms, including machine learning algorithms to personalize shopping offers through user interactions with ML models. From a customer-centric perspective, AI predicts order delivery times, estimates Allegro Pay limits and supports Visual Search.
Amazon	Uses ML models to process vast data resources, large language models (LLMs) with NLP to detect data errors indicating inauthentic reviews and deep neural networks to verify complex relationships and behavior patterns. Due to the growing importance of ESG (Environmental, Social and Governance) in business and increasing customer awareness in this area, Amazon leverages AI for sustainable development, including reducing returns through highly personalized reviews, measuring the carbon footprint of products and automating cloud infrastructure. New models reduce energy consumption by 29% and achieve 50% greater savings.
Netflix	Netflix's recommendation systems are based on algorithms that analyze user interactions with the platform, including the genres of watched films, viewing times and durations, preferred language and the device used. Predictive models match new movies and series that may interest the subscriber.
Spotify	Uses ML for personalizing and suggesting content to users. The most popular recommendation models, creating the best-matched playlists for listeners and creators, include Discover Weekly, Release Radar and Made for You Mixes. Additionally, the AI DJ feature, currently in testing, uses voice synthesis technology as a personalized AI guide. The potential of AI is evidenced by the fact that over half of the listeners who used this feature did so again the next day. User opinions on playlists, songs and podcasts are collected for sentiment analysis.
Shopify	Offers its clients a range of AI-based tools to gain a competitive edge in e-commerce. The Shopify Magic tool uses LLM models to automatically generate texts. Shopify Inbox, part of Shopify Magic, provides automated customer interactions via a chat-bot. The Sidekick assistant helps sellers start and efficiently manage their businesses, supports content creation, performs repetitive tasks, answers questions, creates reports and proposes marketing campaigns.

Source: authors' work based on Amazon, 2023; Augustyniak and Jadczak, 2023; Home Depot, 2024; Hurst, 2024; Lasek, 2024; Piech, 2022.

Based on the AI implementations presented in Table 3, it is evident that AI plays a crucial role in enhancing business efficiency. By analyzing consumer data and predicting purchasing trends, companies can optimize operations and meet the emerging consumer needs in an increasingly digitalized marketplace. Each of the analyzed companies

utilizes machine learning, natural language processing, chatbots and recommendation systems and considers them the key components of their AI-driven strategies.

6. Conclusions

The rapidly evolving business environment requires companies to implement adaptive strategies to maintain competitiveness. Dynamic capabilities play a crucial role in enabling organizations to effectively and efficiently seize the emerging opportunities. As consumer purchasing behaviors continue to digitize, the online sales channel, where competition is exceedingly high, has become a core business component. To remain competitive, businesses must continuously monitor industry trends and adopt cutting-edge technological solutions, with AI being one of the most transformative innovations. This is particularly relevant in markets undergoing rapid digital transformation, such as Poland, where AI adoption is accelerating and influencing how companies operate and engage with digital consumers.

A prominent trend among organizations is the adoption of AI-based solutions which offer numerous advantages but also present significant challenges. While AI enhances efficiency, automation and strategic decision-making, its implementation entails certain risks. Algorithms can exhibit bias, lead to technological dependency and require constant updates, often incurring substantial financial and operational costs. Nevertheless, with effective risk management and a strategic approach, businesses can maximize AI's benefits while mitigating its potential downsides.

AI contributes substantially to business performance by fostering innovation, enhancing precision and improving overall efficiency and productivity. Its implementation can optimize internal processes, automate repetitive tasks and reduce costs while supporting predictive analytics for tasks such as risk assessment and market forecasting. Moreover, AI-driven customization enhances personalization and customer relationship management, allowing businesses to better segment their markets, improve service quality and tailor products to individual consumer needs. These capabilities collectively increase business efficiency and help build a sustainable competitive advantage.

However, the successful adoption of AI requires a comprehensive, strategic approach that balances benefits with the potential risks. Organizations must ensure data security, comply with evolving regulations and invest in AI literacy among employees. Ethical considerations, particularly regarding bias in AI algorithms and consumer privacy, must also be proactively addressed to maintain trust and regulatory compliance.

While this article highlights those AI implementations that prove successful, understanding the reasons behind unsuccessful deployments still remains crucial. However, businesses rarely disclose failed implementations, making such information difficult to obtain. To bridge this gap, future studies should rely on primary research methods to uncover the barriers, challenges and missteps that hinder AI-driven transformations.

References

Adobe, (2020). *Adobe Digital Economy Index*. Retrieved from: https://www.adobe.com/content/dam/dx/us/en/experience-cloud/digital-insights/pdfs/adobe_analytics-digital-economy-index-2020.pdf.

Amankwah-Amoah, J., Khan, Z., Wood, G. and Knight, G., (2021). COVID-19 and digitalization: The great acceleration. *Journal of Business Research*, 136, pp. 602–611. doi: 10.1016/j.jbusres.2021.08.011.

Amazon, (2023). *Jak Amazon stosuje sztuczną inteligencję, aby zapewnić wiarygodne recenzje produktów*. Retrieved from: https://www.aboutamazon.pl/wiadomosci/wiadomosci-i-poglady-dotyczace-polityki/jak-amazon-stosuje-sztuczna-inteligencje-aby-zapewnic-wiarygodne-recenzje-produktow.

Amit, R., Han, X., (2017). Value Creation through Novel Resource Configurations in a Digitally Enabled World: Novel Resource Configurations in a Digitally Enabled World. *Strategic Entrepreneurship Journal*, 11(3), pp. 228–242. doi: 10.1002/sej.1256.

Augustyniak, S., Jadczak, A., (2023). *W jaki sposób algorytmy AI wykorzystują liderzy cyfryzacji w Polsce*. Retrieved from: https://itwiz.pl/w-jaki-sposob-algorytmy-ai-wykorzystuja-liderzy-cyfryzacji-w-polsce.

Ballerini, J., Ključnikov, A., Juárez-Varón, D. and Bresciani, S., (2024). The e-commerce platform conundrum: How manufacturers' leanings affect their internationalization. *Technological Forecasting and Social Change*, 202, 123199. doi: 10.1016/j.techfore.2023.123199.

Bielińska-Dusza, E., (2022). Transformacja technologiczna przedsiębiorstw jako skutek zastosowania sztucznej inteligencji. *Organizacja i Kierowanie*, 2(191).

Borsiak-Dańska, B., Grzelak, M. and Roszko-Wójtowicz, E., (2024). The development of innovation and infrastructure in the European countries fostering the growth of the e-commerce sector. In: *Financial Stability, Economic Growth and Sustainable Development*, pp. 251–278. Routledge.

Deloitte, (2021). *Potrzeba matką wynalazku, czyli skąd się wzięło Data Governance?* Retrieved from: https://www2.deloitte.com/pl/pl/pages/technology/articles/skad-sie-wzielo-data-governance.html.

Dey, S., (2024). *How AI Can Promote ESG*. Retrieved from: https://www.forbes.com/sites/forbestechcouncil/2023/04/28/how-ai-can-promote-esg/.

Dubey, R., Gunasekaran, A., Childe, S. J., Bryde, D. J., Giannakis, M., Foropon, C., Roubaud, D. and Hazen, B. T. (2020). Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations. *International Journal of Production Economics*, 226, 107599. doi: 10.1016/j.ijpe.2019.107599.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., … Williams, M. D., (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994. doi: 10.1016/j.ijinfomgt.2019.08.002.

Dyduch, W., Chudziński, P., Cyfert, S. and Zastempowski, M., (2021). Dynamic capabilities, value creation and value capture: Evidence from SMEs under Covid-19 lockdown in Poland. *PLOS ONE*, 16(6), e0252423. doi: 10.1371/journal.pone.0252423.

Dyduch, W., Dominiczewska, M. and Kubiczek, J., (2023). Value creation value capture revisited: Resource, entrepreneurial and relational orientations. *Forum Scientiae Oeconomia*, 11. doi: 10.23762/FSO_VOL11_NO4_3.

Florczak-Strama, M., (2024). Robotyzacja – rewolucja i jej wpływ na nasze życie. *Zeszyty Naukowe Wyższej Szkoły Bankowej w Poznaniu*, 103(4), pp. 95–109. doi: 10.58683/dnswsb.1955.

Gemius, (2023). *E-commerce w Polsce 2023*. Retrieved from: https://www.gemius.pl/wszystkie-artykuly-aktualnosci/id-79-internautow-kupuje-online-raport-e-commerce-w-polsce-2023-juz-dostepny.html.

Haleem, A., Javaid, M., Asim Qadri, M., Pratap Singh, R. and Suman, R., (2022). Artificial intelligence (AI) applications for marketing: A literature-based study. *International Journal of Intelligent Networks*, 3, pp. 119–132. doi: 10.1016/j.ijin.2022.08.005.

He, X., Liu, Y., (2024). Knowledge evolutionary process of Artificial intelligence in E-commerce: Main path analysis and science mapping analysis. *Expert Systems with Applications,* 238, 121801. doi: 10.1016/j.eswa.2023.121801.

Home Depot, (2024). *5 Technologies Changing How We Shop*. Retrieved from: https://corporate.homedepot.com/news/company/5-technologies-changing-how-we-shop.

Hurst, K., (2024). *7 sposobów Amazon na wykorzystanie AI do budowania zrównoważonej przyszłości*. Retrieved from: https://www.aboutamazon.pl/wiadomosci/technologie/7-sposobow-amazon-na-wykorzystanie-ai-do-budowania-zrownowazonej-przyszlosci.

Kaczorowska-Spychalska, D., (2019). How chatbots influence marketing. *Management*, 23(1), pp. 251–270. doi: 10.2478/manment-2019-0015.

Kowal, B., Świątek, I., (2023). Marketing cyfrowy w branży surowcowej – Case Study. *Inżynieria Mineralna*, 1(1). doi: 10.29227/IM-2023-01-39.

Kubiczek, J., Hadasik, B., Krawczyńska, D., Przedworska, K., Madarász, E. Z. and Ryczko, A., (2024). Perspective of Created Value in Consumer Choice: Comparison of Economic and Ecological Dimensions. *Sage Open*, 14(1), 21582440241238516. doi: 10.1177/21582440241238516.

Lasek, S., (2024). *Jak sztuczna inteligencja wpływa na rozwój e-handlu?* Retrieved from: https://www.linkedin.com/pulse/jak-sztuczna-inteligencja-wp%C5%82ywa-na-rozw%C3%B3j-e-handlu-s%C5%82awomir-lasek-zilue/?originalSubdomain=pl.

Makarius, E. E., Mukherjee, D., Fox, J. D. and Fox, A. K., (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, 120, pp. 262–273. doi: 10.1016/j.jbusres.2020.07.045.

Mason, R. B., Staude, G., (2009). An exploration of marketing tactics for turbulent environments. *Industrial Management & Data Systems*, 109(2), pp. 173–190. doi: 10.1108/02635570910930082.

McKinsey & Company & Forbes Polska, (2017). *Rewolucja AI. Jak sztuczna inteligencja zmieni biznes w Polsce.*

Mikalef, P., Conboy, K. and Krogstie, J., (2021). Artificial intelligence as an enabler of B2B marketing: A dynamic capabilities micro-foundations approach. *Industrial Marketing Management*, 98, pp. 80–92. doi: 10.1016/j.indmarman.2021.08.003.

Nabila, E. A., Santoso, S., Muhtadi, Y. and Tjahjono, B., (2021). Artificial Intelligence Robots and Revolutionizing Society in Terms of Technology, Innovation, Work And Power. *IAIC Transactions on Sustainable Digital Innovation (ITSDI),* 3(1), pp 46–52. Retrieved from http://aptikom-journal.id/index.php/itsdi/article/view/526.

Namogoo, (2020). Consumer Survey: How COVID-19 Online Shopping Habits are Shaping the Customer Journey. *Namogoo*. Retrieved from: https://www.namogoo.com/resources/ebook/how-covid-19-online-shopping-habits-are-shaping-the-customer-journey/.

Nowakowska, P., (2024). Nowe technologie w rozwoju i zarządzaniu przedsiębiorstwem. *Zarządzanie Innowacyjne w Gospodarce i Biznesie*, 1/36, pp. 63–74. doi: 10.25312/2391-5129.36/2023_05pno.

Paczka, E., (2020). Zmiana zachowań rynkowych pokolenia Z. *Ekonomia*, 26(1), pp. 21–34. doi: 10.19195/2658-1310.26.1.2.

Parida, V., Sjödin, D. and Reim, W., (2019). Reviewing Literature on Digitalization, Business Model Innovation, and Sustainable Industry: Past Achievements and Future Promises. *Sustainability*, 11(2), 391. doi: 10.3390/su11020391.

Piech, M., (2022). *Allegro korzysta ze sztucznej inteligencji. Jak A.I. pomaga nam robić zakupy?* Retrieved from: https://businessinsider.com.pl/biznes/allegro-korzysta-ze-sztucznej-inteligencji-jak-ai-pomaga-nam-robic-zakupy/358prnv.

PostNord, (2021). *E-commerce in Europe 2020*. Retrieved from: https://www.postnord.se/siteassets/pdf/rapporter/e-commerce-in-europe-2021.pdf.

Ransbotham, S., Khodabandeh, S., Fehling, R., LaFountain, B. and Kiron, D., (2019). *Winning With AI*. MIT Sloan Management Review and Boston Consulting Group.

Reed, J. H., (2022). Operational and strategic change during temporary turbulence: evidence from the COVID-19 pandemic. *Oper Manag Res*, 15, pp. 589–608. doi: 10.1007/s12063-021-00239-3.

Reuters Institute for the Study of Journalism, (2023). *Digital News Report 2023*. Retrieved from: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023.

Roszko-Wójtowicz, E., Deep Sharma, G., Dańska-Borsiak, B., Grzelak, M. M., (2024a). Innovation-driven e-commerce growth in the EU: An empirical study of the propensity for online purchases and sustainable consumption. *Sustainability*, 16(4), 1563. doi: 10.3390/su16041563.

Roszko-Wójtowicz, E., Pleśniarska, A. and Grzelak, M. M., (2024b). Determinants of Digital Economy Development in the EU Member States: The Role of Technological Infrastructure, Human Capital, and Innovation. *Ekonomia i Prawo. Economics & Law*, 23(4), pp. 611–635.

Saura, J., R., Skare, M. and Riberio-Navarrete, S. (2022). How Does Technology Enable Competitive Advantage? Reviewing State of the Art and Outlining Future Directions. *Journal of Competitiveness*, 14(4), 172–188. doi: 10.7441/joc.2022.04.10.

Schneider, A., Janowska, A., (2020). *Chatboty w Polsce 2020*. UX UPGRADE, SYMETRIA. Retrieved from: https://symetria.pl/chatboty-w-polsce/Chatboty-w-Polsce-2020.pdf.

Schweikl, S., Obermaier, R., (2023). Lost in translation: IT business value research and resource complementarity—an integrative framework, shortcomings and future research directions. *Management Review Quarterly*, 73(4), pp. 1713–1749. doi: 10.1007/s11301-022-00284-7.

Sjödin, D., Parida, V. and Kohtamäki, M., (2023). Artificial intelligence enabling circular business model innovation in digital servitization: Conceptualizing dynamic capabilities, AI capacities, business models and effects. *Technological Forecasting and Social Change*, 197, 122903. doi: 10.1016/j.techfore.2023.122903.

Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., and Basilico, J., (2021). Deep learning for recommender systems: A Netflix case study. *AI Magazine*, 42(3), pp. 7–18. doi: 10.1609/aimag.v42i3.18140.

Turek, T., (2017). Possibilities of using sentiment analysis in prosumption processes. *Ekonomiczne Problemy Usług*, 126, pp. 285–294. doi: 10.18276/epu.2017.126/2-29.

Westerman, G., Ransbotham, S. and Farronato, C., (2024). *Find the AI Approach that Fits the Problem You're Trying to Solve*. Retrieved from: https://hbr.org/2024/02/find-the-ai-approach-that-fits-the-problem-youre-trying-to-solve.

Zhu, Z., Wang, S., Wang, F. and Tu, Z., (2022). Recommendation networks of homogeneous products on an E-commerce platform: Measurement and competition effects. *Expert Systems with Applications,* 201(C). doi: 10.1016/j.eswa.2022.117128.

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 215–234, https://doi.org/10.59139/stattrans-2025-023
Received – 26.08.2024; accepted – 29.01.2025

The integrity of the innovation process on the example of EU countries: a PLS-SEM approach

Mateusz Borkowski[1], **Ewa Gruszewska**[2]

Abstract

The purpose of the article is to assess the integrity of the elements of innovation processes and to measure their efficiency based on the Schumpeter trilogy concept. The research was conducted with regard to European Union (EU) countries. The paper applies the partial least squares structural equation modelling (PLS-SEM) method that allows the analysis of latent variables (LVs). On the basis of the PLS-SEM models for 2010 and 2020, it was concluded that innovation processes were proceeding in an integrated manner in EU countries. Not only did the modelling results indicate a positive and moderate effect of the invention inputs LV on the innovation efficiency LV, but also a positive and strong influence of innovation efficiency LV on the innovation diffusion LV in the analyzed countries in both researched years. The technological process integrity of the EU economies was lower in 2020, than in 2010. In order to improve the functioning of innovation activities it is necessary to increase technology inputs and the efficiency of their use in R&D activities. Intensification of the collaboration between scientific and research institutes and entrepreneurs is recommended. The PLS-SEM model made it possible to measure its elements and assess the integrity of technological change.

Key words: the Schumpeter trilogy, innovation, technology diffusion, PLS-SEM.

1. Introduction

Innovativeness has been considered as one of the main driving forces of the contemporary economic development. Not only do innovations allow an increase in the productivity of production factors, but also lead to qualitative changes in the economy. Every new technological solution results from an innovation process, which consists of three phases: invention, innovation and imitation – it is the, i.e. the Schumpeter trilogy (Curlee & Goel, 1989, p. 3). The integrity of the indicated elements

[1] Department of Political Economics, Faculty of Economics and Finance, University of Bialystok, Poland. E-mail: m.borkowski@uwb.edu.pl, ORCID: https://orcid.org/0000-0003-0644-4764.

[2] Department of Political Economics, Faculty of Economics and Finance, University of Bialystok, Poland. E-mail: gruszew@uwb.edu.pl, ORCID: https://orcid.org/0000-0002-4943-0250.

of technological change is essential for an innovative activity to generate the greatest benefits in the economy.

The theory of innovation stems from the work of Schumpeter, while research of such scientists as Rogers, Freeman, Rosenberg, Porter, Rothwell, Lundvall or Nelson contributed to this theory development (Fagerberg et al., 2012, p. 1144). The intensive development of research on innovations, the process of their creation and the technology transfer has led innovation theory to become a self-contained stream separated from production theory. It includes both elements of micro- (enterprise and production theory) and macroeconomics (growth and development theory), in addition to those from the field of management (see: Fernández, 2023).

The analysis of the economic data indicates that the rate of creation and implementation of innovations varies among contemporary economies. One can clearly distinguish those economies that are at the top of innovation rankings (innovative leaders) and on the other hand, countries with a low rate of internal innovativeness, that only import new solutions from others or imitate extraneous innovations.

The purpose of the article is to assess the integrity of the elements of innovation processes and to measure their efficiency with regard to the concept of the Schumpeter trilogy (invention, innovation and imitation). The research was conducted on the basis of data from 26 EU economies for the years 2010 and 2020. The paper applies the partial least squares structural equation modelling (PLS-SEM) method that allows the analysis of relationships between latent variables (LVs).

The article consists of three parts. The subsequent section contains a review of relevant and topical literature on innovation, innovativeness and entrepreneurship. The second part includes a presentation of the research methodology applied in the study. This section presents the research method, i.e. the partial least squares structural equation modelling (PLS-SEM), and the econometric specification of PLS-SEM model used in our research. On the basis of literature review and the specified model, two hypothesis have been formulated. The fourth section provides modeling results and discussion. The paper closes with conclusions, in addition to which the limitations of the research and future research directions are specified.

2. Theoretical framework

The theory of innovation has its roots in the works of Schumpeter, who defined innovation not only as the revolutionary introduction of a new product or production method but also as the opening of a new market or even the acquisition of a new source of supply (Schumpeter, 1949, p. 66). The contemporary concept of innovation does not differ significantly from that proposed by Schumpeter. Innovation is new (radical) or

improved (incremental) technology solutions (product or business, but also a combination of these) that are significantly different from the previous ones and "*has been introduced on the market or brought into use*" (OECD & European Union, 2018, p. 68).

The process of technological change consists of three subsequent stages. First, there is an idea, which is the result of the application of knowledge and/or technical information to solve a problem. Creativity enables human capital to be transformed into new technical and product or organizational solutions; into inventions that can become innovations. The second stage of technological change is the emergence of innovation. This occurs when there are resources to support new solutions. Therefore, innovation is the first commercial application of a certain set of knowledge. The entity that has the right to use the idea can profit from the practical application of the invention. Successful innovation enables businesses to achieve increasingly high profits, which can contribute to relatively rapid expansion into new markets. *"(I)nnovation is a central determinant of longer-run success and failure for manufacturing firms. Moreover, most industry shattering innovations do not spring from the established competitors in an industry but from new firms or from the established firms entering a new arena"* (Utterback, 1994, p. xxvii).

Proposition 1: *Technology inputs, both financial and human, are essential in increasing the efficiency of the innovation activities of enterprises in the economy.*

The greater the benefits, the faster the next stage of technological, i.e. the diffusion of innovation, will come. It is a process of the continuous spread of a new technological solution across companies, regions or even countries. As Rogers (1983, pp. 34–35) asserts, it is a "*special type of communication concerned with the spread of messages that are new ideas*". Entrepreneurs may attempt to implement external innovations in their own production process of goods and services. This results in the emergence of imitations, which diminish the power of the company that launched a particular innovation on the market. As Kurz (2008, p. 276) wrote: "*In the course of the diffusion process the new methods of production are generalized throughout the system as a whole, thereby establishing a new set of relative prices and gradually eroding the (extra) profits reaped by the innovators and the first generation of followers, while late adopters run the risk of being driven out of the market*".

The diffusion of innovations is inevitable; its cause is the innovative motives of companies and the competitive strength. Imitation processes lead to the better satisfaction of people's needs. The increase in quantities and the reduction in prices of both old and new goods and services enables people to access and use new solutions quicker (Diamond, 2019, p. 65). The diffusion leads to the expansion of benefits achieved from the innovation process. It also contributes to the creation of subsequent generations of technology and products (Vargo et al., 2020, pp. 527–528).

The degree of the diffusion of innovation within highly developed economies is stronger than in other countries (Keller, 2010, p. 806). Nevertheless, the importance of spreading technical solutions to other countries, especially those with medium and low levels of development, cannot be overestimated. In the presence of capital constraints, foreign direct investment may be the only available channel for of new technical solutions and a way to increase efficiency. Positive effects can be achieved by building higher-productivity human capital in foreign companies and then adopting similar solutions in domestic companies (the FDI spillovers through worker rotation).

The diffusion of a new technology is a time- and capital-consuming process. The rate of innovation spread varies and is changing (shortening) dynamically over time due to the progress of communication and information transmission technology advances. The proof of the success of the diffusion process is the occurrence of the horizontal and vertical technology spillovers (Keller, 2010, p. 824). Not every innovation is diffused effectively, which means that only some percentage of new technology solutions completes the innovation process successfully (Dosi & Nelson, 2010, pp. 91–92). Therefore, it is essential that the analysis of the process of technological change include all the three above-mentioned elements. However, nowadays there is a tendency to focus only on the "middle" part of this process (Potts, 2019, pp. 53–54).

Proposition 2: *The ability and propensity to create and implement innovations is a necessary condition for the diffusion of new technological solutions locally and globally.*

The technological process is similar for most new technological solutions, but the reasons why they are created vary. The very first models of the development of innovations indicated that the cause of their emergence could be a supply or demand factor (Rothwell, 1994, pp. 7-9). Innovations could be "pushed" by the science or "pulled" by the market. Over time, linear models were superseded by more complex, non-linear models (the presence of interactions, as well as feedback loops between factors and elements of the technological process), which were a synthesis and development of the previous ones (Ahmed & Shepherd, 2010, pp. 169-172). Subsequent generations of innovation models were based on the belief that new technological solutions are induced by both scientific developments and changes in market needs (Kline & Rosenberg, 1986, p. 290; Rothwell & Zegveld, 1985, p. 50). The high dynamics of innovation processes and the active role of entrepreneurs in the search for optimal solutions in the 21st century led to the formation of a model emphasizing the importance of openness in the innovation process. The essence of open innovation is in the "*purposeful inflows and outflows of knowledge to accelerate innovation internally while also expanding the markets for the external use of innovation*" (Chesbrough, 2006).

Innovativeness, i.e. the ability of enterprises to use existing knowledge to create, implement, and then spread (diffuse) new technological solutions (Salavou, 2004, p. 35), is at the core of entrepreneurship. "*Innovation is the specific instrument of entrepreneurship. It is the act that endows resources with a new capacity to create wealth*" (Drucker, 1993, p. 30).

Entrepreneurship, which is the basis of the entire capitalist world, externalizes itself in a continuous series of disruptive events (Schumpeter, 2003, pp. 82–83). These result from creativity, a willingness to bear risks, openness to change, and curiosity in the search for new technological solutions (Kirzner, 2009, p. 148). Entrepreneurs are the architects of the new order; through the creation of innovation, they continuously revolutionize the economic structure from within, constantly destroying what is old, relentlessly creating new value and quality (Aydin, 2010, p. 21). The emergence of new breakthrough applications of knowledge is the trigger for "*creative destruction*", which clears the market of unnecessary products and inefficient production methods, making room for new, better solutions. Such disruption of the equilibrium causes a tendency for it to reappear, but at a higher level (Dahms, 1995, p. 6). "*If we are open to innovative dynamism and allow entrepreneurs to innovate, we will have bounty. If we are closed to innovative dynamism and bind entrepreneurs, we will have stagnation*" (Diamond, 2019, p. 3).

3. Research methodology

3.1. Research method – PLS-SEM

The article applies structural equation modelling (SEM) as a research method. It is an econometric technique for modelling relationships between latent variables (LVs), that "*makes full use of theoretical and empirical knowledge*" (Skrodzka, 2016, p. 283). In general, two SEM model estimation methods can be distinguished:

- Jöreskog's (1970) covariance based (CB) and
- Wold's (1980) partial least squares based (PLS).

While both methods lead to similar results (Sarstedt et al., 2016, p. 4005), the choice should be supported by substantive and statistical reasons. We decided to choose PLS-SEM instead of CB-SEM on the basis of three main premises (Hair et al., 2011, p. 144; Hair, Matthews, et al., 2017, p. 118): the research uses a small data sample ($N < 100$), indicators do not follow a normal distribution, and the study uses values of latent variables as synthetic measures.

The PLS-SEM proceeds in three subsequent stages (Hair et al., 2016; Hair, Sarstetd, et al., 2017):

- model specification:
  - structural (theoretical, inner) model specification that involves identifying relationships between latent variables in the model;

- measurement (outer) model specification that consists in determining how latent variables are defined (the selection of observable indicators reflect or form them) in the model;

– estimation – PLS-SEM algorithm involves (Lohmöller, 1989, p. 29): the iterative estimation of the values of weights, the estimation of structural parameters and factor loadings using OLS and the determination of location parameters for both inner and outer relations;
– model verification:
  - substantive validation – coincidence and consistency with theory (initial assumptions) assessment;
  - statistical evaluation, that consists of using verification measures.

In the PLS-SEM, values of latent variables (weighted sums of manifest variables) can be used in subsequent analysis. As they are not original in every estimation, they can be treated as synthetic measures (Ćudić & Skrodzka, 2021, p. 76).

2.2. PLS-SEM model specification and hypothesis evaluation

The structural (inner) model consists of two stochastic equations (1, 2), and includes three latent variables – the level of invention inputs (INP), the innovation efficiency (IE) and the scale of innovation diffusion (ID). The Schumpeter trilogy was the basis for determining both internal and external relations in the applied PLS-SEM model. The level of invention inputs latent variable (INP) was lagged by one year due to the substantive assumption that technology outlays need time to be transformed into innovation effects.

$$\mathrm{IE_t} = \alpha_1 \cdot \mathrm{INP_{t-1}} + \alpha_0 + \varepsilon_t, \tag{1}$$

$$\mathrm{ID_t} = \beta_1 \cdot \mathrm{IE_t} + \beta_0 + \zeta_t. \tag{2}$$

where:

IE_t	is the innovation efficiency in the year t;
INP_{t-1}	is the level of invention inputs in the year t–1;
ID_t	is the scale of innovation diffusion in the year t;
α_1, β_1	is the structural parameters of the model;
α_0, β_0	is the location parameters for structural relations;
ε_t; ζ_t	is the random errors (with expected value equal to 0).

All of the latent constructs were defined deductively, which implies that they are reflective in nature. Table 1 contains the final specification of the measurement model. Indicators were chosen on the basis of substantive (theoretical premises) and statistical criteria (the discriminatory abilities of diagnostic variables and the quality of the esti-

mated PLS-SEM model). The statistical data was retrieved from international organizations' databases (Eurostat, ILOSTAT, UNCTAD) for 2010 and 2020. The modelling was performed on the basis of 26 EU economies – Greece was excluded from the research due to substantial data gaps.

Every innovation process starts with invention that requires a certain number of outlays (Ciborowski, 2017, pp. 276–277). This phase is represented by the level of invention inputs latent variable (INP), which is defined by four indicators. Business enterprise R&D expenditures as a % of GDP (INP_1) and government budget allocations for R&D as a % of GDP (INP_2) are related to financial technology inputs, while the percentage of R&D personnel in labour force (INP_3) and the percentage of scientists and engineers in the population aged from 25 to 64 (INP_4) reflect the human contribution to technology development.

Then, invention inputs are transformed into new technology solutions, which are sold and used in business activities. This stage of innovation process is reflected by the innovation efficiency latent variable (IE), which is specified by five diagnostic variables. Not only does the successful innovation activities result in a larger scale of creating technology solutions (IE_1, i.e. the number of patent applications to the EPO per million inhabitants) and higher turnover (IE_2, i.e. the total turnover of innovative enterprises in EUR per one innovative enterprise), but also in higher productivity (Eaton & Kortum, 1999, p. 542) of companies (labour, i.e. IE_3, energy, i.e. IE_4, and resource, i.e. IE_5).

Table 1: The specification of measurement (outer) model

	Description of a diagnostic variable	Data source
The level of invention inputs (INP) latent variable		
INP_1	Business enterprise R&D expenditures (% of GDP)	Eurostat
INP_2	Government budget allocations for R&D (% of GDP)	Eurostat
INP_3	R&D personnel (% of labour force)	Eurostat
INP_4	Scientists and engineers (% of population at the age from 25 to 64 years)	Eurostat
The innovation efficiency (IE) latent variable		
IE_1	Patent applications to the EPO per million inhabitants	Eurostat
IE_2	Total turnover of innovative enterprises (EUR per innovative enterprise)	Eurostat (CIS[a])
IE_3	Labour productivity (output per hour worked)	ILOSTAT
IE_4	Energy productivity (euro per kilogram of oil equivalent)	Eurostat
IE_5	Resource productivity (euro per kilogram used materials)	Eurostat
The scale of innovation diffusion (ID) latent variable		
ID_1	High-tech export (thous. EUR per inhabitant)	Eurostat
ID_2	Foreign Direct Investments, stock outward (thous. USD per capita)	UNCTAD

Note: [a] Community Innovation Survey.
Source: authors' work.

The final phase is imitation, a spread of new and existing technology solutions throughout enterprises located in other regions and countries. This stage is reflected by the scale of innovation diffusion latent variable (ID), that is defined by two indicators. There are two main channels of innovation diffusion (Roszkowska, 2013, p. 58): trade (ID_1 – a value of high technology products exports per capita) and investment (ID_2 – a stock value of outward FDI per capita).

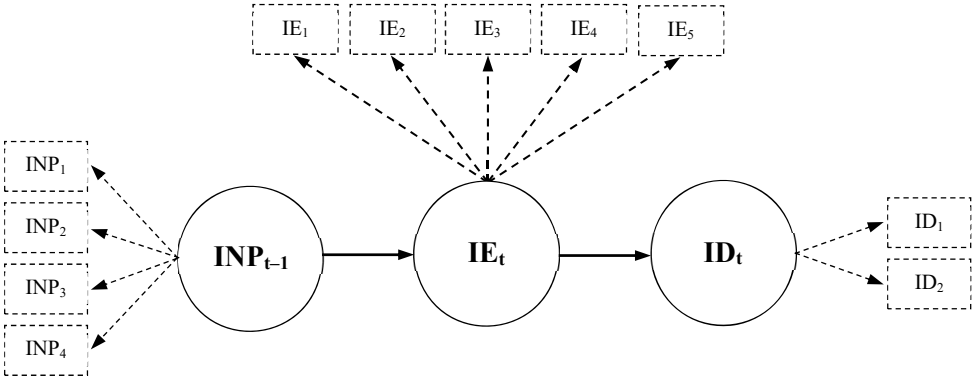


Figure 1: Diagram of the PLS-SEM model applied in this study

Source: authors' work.

The final specification of the applied PLS-SEM model is presented in Figure 1. The hypotheses correspond with propositions, and are formulated as follows:

H1. The level of invention inputs LV correlates with the innovation efficiency LV in a positive, strong (≥ 0.700) and statistically significant ($p < 5\%$) manner.

H2. The innovation efficiency LV correlates with the scare of innovation diffusion LV in a positive, strong (≥ 0.700) and statistically significant ($p < 5\%$) manner.

4. Results and discussion

The measurement model estimates are presented in Table 2. Since all latent variables in the model were defined deductively (reflective indicators), the convergent validity, the internal consistency reliability and the discriminant validity of the measurement model are evaluated (Hair et al., 2019, p. 15). All the factor loading values are above 0.400 and are statistically significant at the level of $p < 5\%$. In addition, the average variance extracted (AVE) values are higher than 50%, indicating their convergent validity. The values of the composite reliability measure are in the interval

from 0.600 to 0.950, which confirm the internal consistency reliability of each latent variable. On the basis of cross-loadings analysis, discriminant validity was established. Moreover, the model is coincident and consequently consistent with initial assumptions, which are based on an economic theory. Therefore, the measurement model estimated for 2010 as well as 2020 data can be considered positively verified.

Table 2: The measurement model results

Latent variable	Indicator	Convergent validity				Reliability of internal consistency		Discriminant validity	
		Loadings[a]		Average variance extracted		Composite reliability		Cross loadings	
		≥ 0.400		≥ 0.400		0.600 – 0.950			
		2010	2020	2010	2020	2010	2020	2010	2020
INP_{t-1}	INP_1	0.955***	0.914***	0.901	0.759	0.939	0.926	✓	✓
	INP_2	0.841***	0.825***					✓	✓
	INP_3	0.939***	0.943***					✓	✓
	INP_4	0.825***	0.794***					✓	✓
IE_t	IE_1	0.931***	0.847***	0.717	0.658	0.926	0.905	✓	✓
	IE_2	0.916***	0.873***					✓	✓
	IE_3	0.783***	0.627***					✓	✓
	IE_4	0.709***	0.815***					✓	✓
	IE_5	0.874***	0.869***					✓	✓
ID_t	ID_1	0.941***	0.680**	0.819	0.512	0.901	0.676	✓	✓
	ID_2	0.869***	0.749***					✓	✓

Note: [a] 5,000 samples in bootstrapping procedure; t-Student test; *** $p \leq 0.01$; ** $p \leq 0.05$
Source: authors' work.

In the model for 2010, the level of invention inputs latent variable (INP) is most strongly reflected by the business enterprise R&D expenditures variable (INP_1; 0.955). The R&D personnel indicator (INP_3; 0.939) is also very strongly correlated with this latent variable. The other two indicators of the INP latent variable, namely INP_2 (0.841; government budget allocations for R&D) and INP_4 (0.825; scientists and engineers in population), reflect its changes in a strong way. The model estimated for 2020 yields similar results. The notable difference is in the manifest variable that is most strongly correlated with this latent construct (INP_3; 0.943).

In the model for 2010, the innovation efficiency latent variable (IE) is reflected by indicator of patent applications to the EPO (IE_1; 0.931) in the strongest way. Changes in the IE latent construct is also very strongly reflected by the changes in the turnover of innovative enterprises (IE_2; 0.916). The correlation of the IE variable with sequentially: IE_5 (0.874; resource productivity), IE_3 (0.783; labour productivity) and IE_4

(0.709; energy productivity) is strong. In the model for 2020, results are slightly different. None of the variables correlate very strongly with the innovation efficiency latent variable (IE). This latent variable is most strongly correlated with the measure of turnover of innovative companies (IE_2; 0.873) and this relationship is strong. IE_5 (0.874; resource productivity), IE_1 (0.847; number of patent applications to EPO per capita) and IE_4 (0.815; energy productivity) also reflect this LV changes in a strong manner. The labour productivity (IE_3; 0.623) has the lowest value of factor loading, indicating that correlation of this measure with the IE latent construct is moderate. Considering the two years analysed, it can be concluded that the increase in innovation efficiency is most strongly reflected in the increase in the turnover of innovative businesses and the rise in the number of patent applications. "*Enterprises increasingly tend to use patent protection as an effect of the incurred costs of R&D and as a necessity to secure the results of their intramural research*" (Ciborowski & Skrodzka, 2020, p. 1360).

The measurement model for 2010 indicates that the scale of innovation diffusion latent variable (ID) is most strongly correlated with the high-tech export indicator (ID_1; 0.941). The FDI variable (ID_2; 0.869) reflects changes in these LV values strongly. However, the results for 2020 are slightly different. Both manifest variables correlate slightly weaker with ID than in 2010. Moreover, the FDI measure (ID_2; 0.749) reflects this LV to a larger degree than hi-tech export variable (ID_1; 0.680).

The increase in the relevance of the FDI measure in reflecting innovation diffusion LV over the years indicates changes in the preferred diffusion channels. Entrepreneurs increasingly often choose to rely on more stable, sustained technology diffusion streams, abandoning the one-off, "contract" ones. The FDI enables companies both to create and to accumulate stable assets abroad. Moreover, the resources build through FDI remain in the recipient country even when the investor withdraws from a particular market. The effects of changes in products, manufacturing processes, labour organization, or customer access channels are not confined to a single company. They create positive externalities, including increasing the competitiveness of domestic companies and thereby raising the pressured innovative changes in other companies. Considering the recent socio-economic events in the world (the COVID-19 pandemic, the war in Ukraine), one can expect an even greater increase in the importance of the investment diffusion channel.

$$\begin{aligned} IE_{2010} &= 0.683^{***} \cdot INP_{2009} - 0.174 \\ R^2 &= 0.467 \end{aligned} \qquad \begin{aligned} IE_{2020} &= 0.642^{***} \cdot INP_{2019} - 0.526 \\ R^2 &= 0.412 \end{aligned} \tag{3}$$

$$\begin{aligned} ID_{2010} &= 0.798^{***} \cdot IE_{2010} - 0.431 \\ R^2 &= 0.637 \\ Q^2 &= 0.282 \end{aligned} \qquad \begin{aligned} ID_{2020} &= 0.776^{***} \cdot IE_{2020} - 0.406 \\ R^2 &= 0.603 \\ Q^2 &= 0.118 \end{aligned} \tag{4}$$

As the measurement model has been considered to be positively verified, one can proceed to the structural model validation. That consists in the evaluation of collinearity, the significance of path coefficients, and the exploratory and predictive power of internal relations (Hair et al., 2019, pp. 15–16). Formulas (3) and (4) represent the estimation of the structural model. In 2010, the level of invention inputs variable (INP) had a moderate, positive, and statistically significant ($p < 1\%$) impact (0.683) on the innovation efficiency construct (IE). Moreover, the correlation (0.798) between innovation efficiency LV and the scale of innovation diffusion LV was positive, strong, and statistically significant ($p < 1\%$). The exploratory power of the structural model for 2010 can be considered as satisfactory (R^2 values). The general Q^2 (Stone-Geisser's test) value (0.282; 10 blindfolds) indicates that the model has good predictive power.

The structural model estimates for 2020 are relatively similar to those obtained for 2010. The level of invention inputs LV moderately, positively and significantly ($p < 1\%$) influenced (0.642) the innovation efficiency LV. What is more, the scale of innovation diffusion LV depends (0.776) on the innovation efficiency LV – this relationship is strong, positive, and statistically significant ($p < 1\%$). The exploratory power of structural equations is satisfactory. The model has fairly good predictive power ($Q^2 = 0.118$; 10 blindfolds). Therefore, the structural model for both 2010 and 2020 can be regarded as positively verified. On the basis of structural relations it can be concluded that innovation processes in EU countries followed the pattern of Schumpeterian trilogy in the studied period.

Formulated statistical hypotheses can be verified by means of structural model equations. Even though there are conditions for the negative verification of the first hypothesis, due to the minor difference from the assumption, conditionally it can be considered positively verified. The level of invention inputs LV positively correlates with the innovation efficiency LV, approximating a strong (≥ 0.600) and statistically significant ($p < 1\%$) manner both in 2010 and 2020. The second statistical hypothesis is also considered to be positively verified. The innovation efficiency LV corelated with the scale of diffusion LV in a positive, strong (0.700) and statistically significant ($p < 5\%$) manner in 2010 and 2020.

However, it should be also noted, that this integrity of the technological change is getting weaker over time (from 2010 to 2020). It seems this has been caused by the gradual merging of the second (innovation) and third (diffusion) phases of the innovation process. For instance, the joint innovation policy of the European Union makes innovative entities apply for legal protection not in their domestic offices, but directly at the European Patent Office. Hence, not only is obtaining such patent a confirmation of innovation, but also a step towards diffusion into a wider space than a single economy. In addition, due to the high capital intensity of innovation activity, new ideas in the process of transformation into innovations are often financed from

foreign sources. This is particularly the case with innovations of great importance to the economy, or with high "profit-creating" potential. Such new solutions arouse the interest of large corporations as early as during their creation. This means that the moment an innovation is implemented, it is already internationalized (diffused). Therefore, it can be concluded that globalization processes have led to the shortening of the Schumpeter trilogy into the Schumpeter dilogy. The shortening of the technological process has been demonstrated by studies of several authors, e.g. (Bento & Wilson, 2016; Ellwood et al., 2017; Fischer et al., 2015).

Table 3: Rankings of the level of latent variables in the model in EU countries

Country	INP_{t-1}			IE_t			ID_t		
	2009	2019	change	2009	2019	change	2009	2019	change
Austria	9.	6.	+3	9.	10.	−1	8.	7.	+1
Belgium	6.	5.	+1	7.	9.	−2	6.	4.	+2
Bulgaria	24.	22.	+2	26.	26.	=	26.	26.	=
Croatia	19.	19.	=	22.	24.	−2	25.	23.	+2
Cyprus	21.	23.	−2	14.	14.	=	4.	5.	−1
Czechia	14.	11.	+3	16.	16.	=	12.	6.	+6
Denmark	2.	2.	=	2.	4.	−2	10.	10.	=
Estonia	13.	13.	=	24.	22.	−2	15.	13.	+2
Finland	1.	3.	−2	11.	11.	=	14.	18.	−4
France	7.	10.	−3	6.	5.	+1	13.	15.	−2
Germany	5.	4.	+1	4.	6.	−2	9.	9.	=
Hungary	18.	18.	=	20.	21.	−1	11.	12.	−1
Ireland	8.	12.	−4	8.	2.	+6	3.	2.	+1
Italy	15.	15.	=	10.	7.	+3	18.	20.	−2
Latvia	22.	25.	−3	18.	20.	−2	24.	19.	+5
Lithuania	17.	20.	−3	23.	23.	=	20.	17.	+3
Luxembourg	4.	8.	−4	1.	1.	=	1.	1.	=
Malta	26.	21.	+5	13.	15.	−2	5.	8.	−3
Netherlands	11.	7.	+4	3.	3.	=	2.	3.	−1
Poland	20.	16.	+4	15.	18.	−3	22.	21.	+1
Portugal	16.	14.	+2	21.	19.	+2	21.	24.	−3
Romania	25.	26.	−1	25.	25.	=	23.	25.	−2
Slovakia	23.	24.	−1	19.	17.	+2	17.	16.	+1
Slovenia	10.	9.	+1	17.	13.	+5	16.	14.	+2
Spain	12.	17.	−5	12.	12.	=	19.	22.	−3

Source: authors' work.

As both the measurement and structural model have been positively verified, one can proceed to the analysis of the values of latent variables. Table 3 presents rankings of the EU countries in terms of latent variable scores in 2010 and 2020.

In 2009, the highest level of invention inputs LV was recorded in Finland, and the lowest in Malta. Meanwhile, in 2019, Sweden was the leader of INP ranking, while Romania was at the end of the list. Two countries changed their ranks notably – Malta progressed from the 26th place in 2009 to 21st in 2019, while Spain moved downwards from the 12th place in 2009 to the 17th in 2019. The improvement of Malta's performance in terms of INP ranking results from the increase in the number of scientists and engineers as a % of population aged from 25 to 64 (INP_4). The reason for Spain's decline in the INP ranking was a major decrease in the government budget allocations for R&D as a % of GDP (INP_2).

In 2010 as well as in 2020, the highest innovation efficiency was reported in Luxemburg, and the lowest in Bulgaria. The biggest IE rank changes were observed in Slovenia (ranked 17th in 2010 and 13th in 2020), and in Ireland (8th in 2010 and 2nd in 2020). Both countries' performance in terms of all of the innovation efficiency LV manifest variables was better in 2020 than in 2010.

Technology diffusion occurred on the largest scale in Luxembourg and on the smallest in Bulgaria, both in 2010 and 2020. Two EU economies achieved a significant progress in the ID ranking, i.e. Latvia (24th rank in 2010 and 19th in 2020) and Czechia (12th in 2010 and 6th in 2020). Both countries achieved higher hi-tech export (ID_1) and outward FDI (ID_2) value per capita in 2020, compared to 2020.

EU countries can be classified into four typological groups on the basis of the results for innovation efficiency (IE) and the innovation diffusion (ID) latent variables:

- **dynamic innovators (technology pioneers)** – countries in which companies achieve high innovation efficiency and large-scale of innovation diffusion ($IE_{it} \geq 0$ and $ID_{it} \geq 0$),
- **internal (local) innovators** – countries in which business entities achieve high innovation efficiency and small-scale of innovation diffusion ($IE_{it} \geq 0$ and $ID_{it} < 0$),
- **innovation intermediaries** – countries in which business entities achieve low innovation efficiency and large-scale of innovation diffusion ($IE_{it} < 0$ and $ID_{it} \geq 0$),
- **imitators** – countries in which businesses achieve low level of innovation efficiency and small-scale of innovation diffusion ($IE_{it} < 0$ and $IE_{it} < 0$)

where: i is the number of EU country (i = 1, 2, 3, …, 26) and t is a year (t = 2010, 2020)

Figure 2 presents the division of EU economies into four typological groups according to their innovation status in 2010 and 2020. In 2010, five countries were classified as dynamic innovators: Belgium, Ireland, Luxembourg, the Netherlands and

Sweden. The internal innovators group was comprised of seven economies: Austria, Denmark, Finland, France, Germany and Italy. Only two EU countries were characterized as innovation intermediaries in 2010, namely Cyprus and Malta. The remaining countries (Bulgaria, Croatia, Czechia, Estonia, Hungary, Latvia, Lithuania, Poland, Portugal, Romania, Slovakia, Slovenia, Spain) were classified as technology imitators.

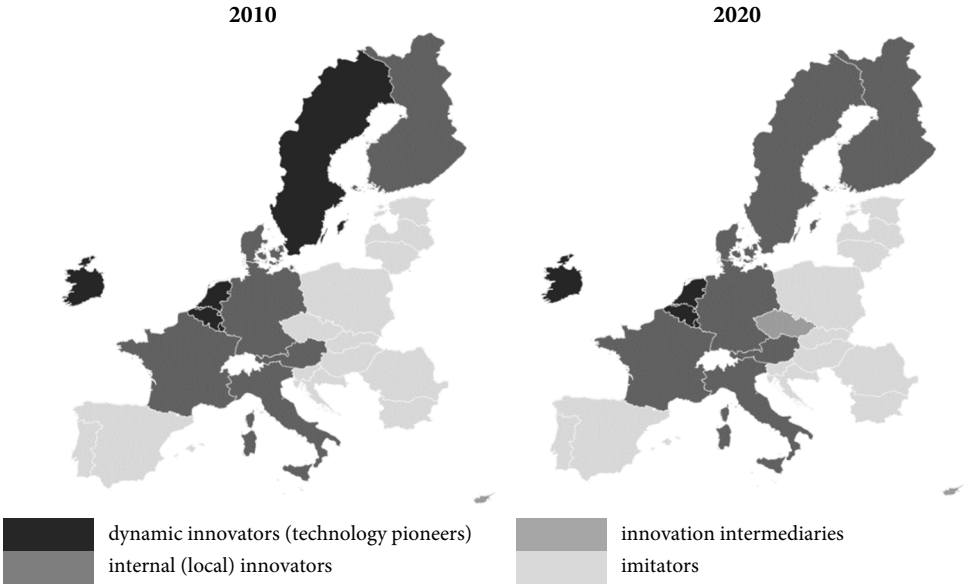


Figure 2: Typological groups of innovation status in 2010 and 2020 among EU countries
Source: authors' work.

The classification of objects in 2020 is similar to the one in 2010. Only three EU economies changed their group throughout the analyzed research period. In 2010, Sweden was among technological pioneers, while in 2020, it belonged to the group of internal innovators. Malta also went down in the ranking: in 2010, it was in the group of innovation intermediates, while in 2020 it descended to the group of technology imitators. The only EU country to improve its innovation status was Czechia, which advanced from imitators to the group of innovation intermediates.

Based on the results, one can observe that countries of Central and Eastern Europe (CEE-11) are much less innovative than other EU economies relatively not long ago (in the 1990s). There are many reasons for this situation. The CEE-11 economies begun to build market economies. The weakness and incompatibility of formal and informal institutions provided a fragile base for entrepreneurship and especially innovative activity. A sound institutional framework is essential in stimulating innovation processes.

Another factor here is comparatively small scale of R&D investment and insufficient involvement of the corporate sector in innovation processes. Based on the Mann-Whitney test[3], significant ($p < 5\%$) differences were identified between the CEE-11 and EU-15 economies for the values of most indicators reflecting invention of inputs in both 2010 and 2020 (Table 4).

Table 4: The average value of invention inputs variables in EU-15 and CEE-11

Countries	INP_1		INP_2		INP_3		INP_4	
	2009	2019	2009	2019	2009	2019	2009	2019
EU-15 (N=15[a])	1.27	1.31	0.70	0.59	1.24	1.49	4.63	7.71
CEE-11 (N=11[b])	0.41	0.72	0.46	0.41	0.68	0.97	3.24	5.49
Mann-Whitney p-value	**0.001**	**0.032**	**0.015**	0.054	**0.003**	**0.005**	**0.027**	**0.003**

Note: [a] Austria, Belgium, Cyprus, Denmark, Finland, France, Germany, Ireland, Italy, Luxembourg, Malta, the Netherlands, Portugal, Spain, Sweden; [b] Bulgaria, Croatia, Czechia, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, Slovenia
Source: authors' work.

Adequate availability of input streams would make it possible to increase the scale of research in the public and private sectors. Financing of innovation activities could be supported by foreign capital. The demand channel in the creation of innovation is weaker (competition stimulates innovation) due to the limited accumulation capacity of companies. Under such conditions, it may be cheaper to acquire solutions from abroad than to perform research locally. All this causes the CEE-11 countries to predominantly be imitators.

5. Conclusions

The study described in this article concentrated on the assessment of the integrity of the phases of innovation processes and the measurement of their efficiency on the basis of the concept of the Schumpeter trilogy. On the basis of the literature review, two hypotheses were formulated, which were then subjected to verified using an econometric research technique.

The first hypothesis, according to which the level of invention inputs LV correlates with the innovation efficiency in a positive, strong and statistically significant manner, was conditionally positively verified. The level of invention inputs variable (INP) was close to having a strong, positive, and statistically significant ($p < 1\%$) impact on the innovation efficiency construct (IE) both in 2010 (0.683) and 2020 (0.642). The second

[3] Mann-Whitney test is the non-parametric equivalent of the parametric t-Student test. It is used to compare medians, not means, between the nondependent samples. Since the data does not meet the assumptions of a normal distribution, nonparametric tests should be used (Hart, 2001, p. 392).

hypothesis states that the innovation efficiency LV correlates with the scale of innovation diffusion LV in a positive, strong and statistically significant manner. On the basis of the second equation of the structural model, this hypothesis was verified positively. Both in 2010 (0.798) and 2020 (0.776), the PLS-SEM model indicated that the innovation efficiency LV correlates with the scale of diffusion LV in a positive, strong and statistically significant manner. On this basis it can be concluded that in innovation processes in EU economies follow the pattern delineated in the Schumpeter trilogy theory.

Results yielded by the PLS-SEM model were also used to make a typological division of EU countries into four groups in terms of the innovation status. The first group was labelled "dynamic innovators". These countries that achieve a high degree of efficiency in the innovation process and at the same time export innovative solutions and benefit from it. In this group of countries, the innovation process run in an integrated manner and at an advanced level. The second group (internal innovators) is comprised of those economies that have achieved relatively high innovation efficiency, but a low scale of technology diffusion. Those economies achieve only internal benefits from innovation. Despite the adequate development of innovation processes in the two initial stages of the Schumpeter trilogy, barriers emerge in the last stage that hinder expansion to other countries. The third group (innovation intermediaries) consists of economies that, despite their limited potential for innovation, are exporters of new solutions. These are predominantly small economies that import technology and then resell it, reaping the benefits. Such an activity can, through learning-by-doing, contribute to building a country's intellectual capital and create a basis for its own innovation activity. Such processes mostly concern the services sector (van der Boor et al., 2014, pp. 1595–1596). The countries from the fourth group are characterized by both low innovation potential and low propensity to export innovations to other countries. They are only recipients and imitators of new technologies. Actions that need to be taken to increase the level of innovation should foster the enlargement of the internal potential, as well as the creation of clear channels for the technology transfer into the economy and then out of the country.

Results of this study allowed the identification of the problem of innovative activity of contemporary economies, namely the insufficient scale of innovation inputs. Therefore, it is crucial not only to increase the financial inputs but also the quality of human contribution to innovation processes. It is also necessary to close the gap in education, experience and qualifications. Moreover, it is important to support private channels for funding innovation, as they, compared to public ones, tend to more effective (Ciborowski & Skrodzka, 2019, p. 403). Pro-entrepreneurial attitudes, e.g. openness to change, risk-taking, are of a great importance in innovative activities, and need to be continuously supported by the state. Another problem innovative activity

faces in modern economies is the insufficient level of collaboration between R&D centres (such as: universities, technology parks) and enterprises.

The most significant limitation to the research is the low availability of statistical data related to innovative activity. This made it impossible to construct a model based on data on the EU economies for earlier years, i.e. before 2010. A comparison of the modeling results for the later periods, would have allowed more precise conclusions about technological changes in EU countries. Research limitations also resulted from the econometric technique used. Its biggest drawback, apart from the difficulty of calculations without specialised software, is the lack of inter-period comparisons in terms of changes in the values of latent variables. Only increases and decreases in rankings are subject to interpretation, and this results in simplified inferences about changes.

Innovativeness (with its components, changes and determinants), is an important and contemporary topic in the economic theory. The authors' future research will be the continuation of the analyses conducted in the study presented in this article. We plan to conduct a similar study, but one taking into account more countries, also from other parts of the world, to check if one can speak of the integrity of innovation processes in, e.g., Asian countries.

References

Ahmed, P. K., Shepherd, C. D., (2010). Innovation management: Context, strategies, systems and processes. *Pearson*, London.

Aydin, D., (2010). "Destructive" and "Creative" Results of Dynamic Analytical Frameworks of Marx and Schumpeter. *Business and Economics Research Journal*, Vol. 1, No. 2, pp. 17–26.

Bento, N., Wilson, C., (2016). Measuring the duration of formative phases for energy technologies. *Environmental Innovation and Societal Transitions*, Vol. 21, pp. 95–112.

Chesbrough, H., (2006). Open Innovation: A New Paradigm for Understanding Industrial. In H. Chesbrough, W. Vanhaverbeke, & J. West (eds.), Open innovation: Researching a new paradigm. *Oxford University Press*, Oxford, pp. 15–33.

Ciborowski, R., (2017). Territorial transfer of knowledge in terms of creative destruction. *Studies in Logic, Grammar and Rhetoric*, Vol. 50, No. 1, pp. 269–287.

Ciborowski, R., Skrodzka, I., (2019). International Technology Transfer, Innovation and Economic Development of European Union Countries. *European Research Studies Journal*, Vol. XXII, pp. 384–404.

Ciborowski, R., Skrodzka, I. (2020). International technology transfer and innovative changes adjustment in EU. *Empirical Economics*, Vol. 59, pp. 1351–1371.

Ćudić, B., Skrodzka, I., (2021). 'Soft' support infrastructure and the performance of small and medium-sized enterprises in European countries. *Economic Annals*, Vol. 66, No. 230, pp. 67–99.

Curlee, T. R., Goel, R. K., (1989). The Transfer and Diffusion of New Technologies: A Review of the Economics Literature. *Oak Ridge National Laboratory*, Oak Ridge.

Dahms, H. F., (1995). From Creative Action to the Social Rationalization of the Economy: Joseph A. Schumpeter's Social Theory. *Sociological Theory*, Vol. 13, No. 1, pp. 1–13.

Diamond, A. M. J., (2019). Openness to Creative Destruction: Sustaining Innovative Dynamism. *Oxford University Press*, Oxford.

Dosi, G., Nelson, R. R., (2010). Technical Change and Industrial Dynamics as Evolutionary Processes. In B. H. Hall & N. Rosenberg (eds.), Handbook of the Economics of Innovation. *North-Holland*, Amsterdam, pp. 51–127.

Drucker, P. F., (1993). Innovation and Entrepreneurship. Practice and Principles. *Harper Collins Publishers*, New York.

Eaton, J., Kortum, S., (1999). International Technology Diffusion: Theory and Measurement. *International Economic Review*, Vol. 40, No. 3, pp. 537–570.

Ellwood, P., Grimshaw, P. and Pandza, K., (2017). Accelerating the Innovation Process: A Systematic Review and Realist Synthesis of the Research Literature. *International Journal of Management Reviews*, Vol. 19, No. 4, pp. 510–530.

Fagerberg, J., Fosaas, M. and Sapprasert, K., (2012). Innovation: Exploring the knowledge base. *Research Policy*, Vol. 41, No. 7, pp. 1132–1153.

Fernández, I. A., (2023). Innovation and international business: A systematic literature review. *Heliyon*, Vol., 9 No. 1, e12956.

Fischer, C., Lušić, M., Bönig, J., Hornfeck, R. and Franke, J., (2015). Shortening Innovation Cycles by Employee Training Based on the Integration of Virtual Validation into Worker Information Systems. *Procedia CIRP*, Vol. 37, pp. 65–70.

Hair, J. F., Hult, G. T. M., Ringle, C. M. and Sarstedt, M., (2016). A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). *SAGE Publications*, New York.

Hair, J. F., Matthews, L. M., Matthews, R. L. and Sarstedt, M., (2017). PLS-SEM or CB-SEM: Updated guidelines on which method to use. *International Journal of Multivariate Data Analysis*, Vol. 1, No. 2, pp. 107–123.

Hair, J. F., Ringle, C. M. and Sarstedt, M., (2011). PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory and Practice*, Vol. 19, No. 2, pp. 139–152.

Hair, J. F., Risher, J. J., Sarstedt, M. and Ringle, C. M., (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, Vol. 31, No. 1, pp. 2–24.

Hair, J. F., Sarstetd, M., Ringle, C. M. and Gudergan, S. P., (2017). Advanced Issues in Partial Least Squares Structural Equation Modeling. *SAGE Publications*, New York.

Hart, A., (2001). Mann-Whitney test is not just a test of medians: Differences in spread can be important. British Medical Journal, 323(7309), 391–393.

Jöreskog, K. G., (1970). A General Method for Estimating a Linear Structural Equation System*. *ETS Research Bulletin Series*, Vol. 1970, No. 2, pp. 1–41.

Keller, W., (2010). International Trade, Foreign Direct Investment, and Technology Spillovers. In B. H. Hall & N. Rosenberg (Eds.), Handbook of the Economics of Innovation. *North-Holland*, Amsterdam, pp. 793–829.

Kirzner, I. M., (2009). The alert and creative entrepreneur: A clarification. *Small Business Economics*, Vol. 32, No. 2, pp. 145–152.

Kline, S. J., Rosenberg, N., (1986). An Overview of Innovation. In R. Landau & N. Rosenberg (Eds.), The Positive Sum Strategy: Harnessing Technology for Economic Growth. *National Academy Press*, Washington. pp. 275–307.

Kurz, H. D., (2008). Innovations and profits: Schumpeter and the classical heritage. *Journal of Economic Behavior & Organization*, Vol. 67, No. 1, pp. 263–278.

Lohmöller, J.-B., (1989). Latent Variable Path Modeling with Partial Least Squares. *Springer*, Berlin.

OECD, European Union, (2018). Oslo Manual 2018: Guidelines for collecting, reporting and using data on innovation.

Potts, J., (2019). Innovation Commons: The Origin of Economic Growth. *Oxford University Press*, Oxford.

Rogers, E. M., (1983). Diffusion of Innovations. *The Free Press*.

Roszkowska, D., (2013). Approaches to International Technology Transfer Measurement – An Overview. *Optimum. Studia Ekonomiczne*, Vol. 5, No. 65, pp. 51–63.

Rothwell, R., (1994). Towards the Fifth-generation Innovation Process. *International Marketing Review*, Vol. 11, No. 1, pp. 7–31.

Rothwell, R., Zegveld, W., (1985). Reindustrialization and Technology. *Longman*, London.

Salavou, H., (2004). The concept of innovativeness: Should we need to focus? *European Journal of Innovation Management*, Vol. 7, No. 1, pp. 33–44.

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O. and Gudergan, S. P., (2016). Estimation issues with PLS and CB-SEM: Where the bias lies! *Journal of Business Research*, Vol. 69, No. 10, pp. 3998–4010.

Schumpeter, J. A., (1949). The theory of Economic Development. *Harvard University Press*, Cambridge.

Schumpeter, J. A., (2003). Capitalism, Socialism and Democracy. *Routledge*, London.

Skrodzka, I., (2016). Knowledge-Based Economy in the European Union: Cross-Country Analysis. *Statistics in Transition*, Vol. 17, No. 2, pp. 281–294.

Utterback, J. M., (1994). Mastering the Dynamics of Innovation: How Companies Can Seize Opportunities in the Face of Technological Change. *Harvard Business School Press*, Cambridge.

van der Boor, P., Oliveira, P. and Veloso, F., (2014). Users as innovators in developing countries: The global sources of innovation and diffusion in mobile banking services. *Research Policy*, Vol. 43, No. 9, pp. 1594–1607.

Vargo, S. L., Akaka, M. A. and Wieland, H., (2020). Rethinking the process of diffusion in innovation: A service-ecosystems and institutional perspective. *Journal of Business Research*, Vol. 116, pp. 526–534.

Wold, H., (1980). Soft modelling: Intermediate between traditional model building and data analysis. *Banach Center Publications*, Vol. 6, No. 1, pp. 333–346.

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 235–236

In Memoriam

Professor Janusz Witkowski

It is with great sadness that we inform you – also on behalf of the editorial bodies and collaborators of SiTns – about the sudden passing of Professor Janusz Witkowski, which has caused an inconsolable grief among the statistical community.

Professor Janusz Witkowski was an outstanding specialist in demography, social policy and official statistics. He served as Vice President of Statistics Poland from 1996 to 2011 and its President from 2011 to 2016. He was also Co-Chairman and Chairman of the Editorial Board of SiTns. His term of office coincided with a period of profound methodological and operational transformations taking place in official statistics, involving the development of modern IT systems, an increased application of administrative data in statistical research and the modernization and digitalization of data collection and sharing processes. Professor Witkowski greatly contributed to these achievements, while playing a key role in overcoming the challenges encountered during these processes.

He was a renowned and respected scientist and author of 350 scientific papers. He worked at the Institute of Statistics and Demography of the SGH Warsaw School of Economics and at the Faculty of Economics of the Professor Edward Lipinski School of Economics, Law and Medical Sciences in Kielce.

Professor Witkowski actively participated in the work of many international organizations, also in the role of an expert of the European Commission, OECD, Council of Europe and the International Labour Organization. He was very active within the international scientific community, attending conferences and congresses organized by the International Statistical Institute and other organizations.

He also held important positions in multiple scientific and advisory bodies, including the Committee of Statistics and Econometrics of the Polish Academy of Sciences and served as the Vice-Chairman of the Government Population Council.

For many years, Professor Witkowski was the Vice President of the Polish Demographic Society, actively participating in its work. He was awarded the Bronze and Gold Cross of Merit, the Knight's Cross of the Order of Polonia Restituta and the Waclaw Szubert Medal for his outstanding performance.

His contribution to official statistics and construction of a modern national statistical system will remain an invaluable achievement and last for years to come.

We would like to convey my deepest condolences to the family and relatives of Professor Janusz Witkowski.

Marek Cierpiał-Wolan
President
Statistics Poland

Włodzimierz Okrasa
Editor-in-Chief
Statistics in Transition new series

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 237–238

In Memoriam

Professor Achille Lemmi

With deep sorrow, we bid farewell to Professor Achille Lemmi, an outstanding scientist, researcher and mentor, who played a key role in the development of modern statistics, especially in the study of income distribution, poverty and social inequalities. His contribution to the advancement of statistical methods and engagement in international scientific cooperation have made left a lasting impact on the academic community.

Professor Lemmi was an Honorary Fellow of the ASESD Tuscan Universities Research Centre "Camilo Dagum" and retired Professor Emeritus of economic statistics at the University of Siena. Throughout his career, he held played numerous academic and organizational roles, including that of Director of the Centro Interdipartimentale di Ricerca sulla Distribuzione del Reddito (CRIDIRE) and member of prestigious research institutions, such as the Centro di Ricerca sull'Integrazione Europea (CRIE) at the University of Siena. He was also member of the Selection Panel Network ECASS at the University of Essex in the United Kingdom and actively contributed to scientific institutions such as the Tes-Institute (Luxembourg) and Laboratorio per l'Intelligenza Artificiale e la Statistica Applicata (LIASA). His scientific achievements included approximately 85 monographs and research articles published in international journals and by renowned publishers.

Professor Lemmi made a significant contribution to the scientific collaboration between Poland and Italy. In the 1980s, he was one of the initiators of a close collaboration between the Italian Statistical Society, the Polish Statistical Association, Statistics Poland and the SGH Warsaw School of Economics, which resulted in numerous international research projects.

Professor Lemmi collaborated with the Statistics in Transition journal as an Associate Editor from its very beginning, actively supporting its development. His activities also involved active participation in international statistical conferences, organized by Statistics Poland and the SGH Warsaw School of Economics, where he served as member of scientific committees, a panelist and a speaker.

His role as a mentor was invaluable – Professor Lemmi had a significant impact on the academic careers of many Polish statisticians. He organized research internships for them at Italian universities, supported their scientific growth and invited them to prestigious conferences organized by the Italian Statistical Society and leading Italian universities.

Professor Achille Lemmi will be remembered as an outstanding scholar, dedicated researcher and mentor who inspired generations of statisticians. His passing is a great loss for the statistical community.

We extend our sincere condolences to the family and loved ones of Professor Achille Lemmi.

Tomasz Panek
Associate Editor
Warsaw School of Economics

Włodzimierz Okrasa
Editor-in-Chief
Statistics in Transition new series

STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. 239–244

About the Authors

Aghel Wesal Emhemed Ramadhan is a Professor of Statistics at the Department of Statistics, Zawia University Faculty Sciences, Libya. She received her PhD degree in Statistics from the Faculty of Graduate Studies for Statistical Research (Institute of Statistical Studies & Research (previously)), Cairo University, Egypt. Her main research interests are: probability distributions, record values, ranked set sampling, goodness of fit tests and information measures and accelerated life tests. She acts as a reviewer for several high-impact journals in the field of statistics.

Bandyopadhyay Arnab is currently working as an Associate Professor in the Department of Mathematics, Basic Science and Humanities at Dr. B. C. Roy Engineering College, Durgapur, West Bengal, India. He has more than sixteen years teaching experience as Assistant Professor in Mathematics. His research field is statistical sample surveys, sata snalysis and statistical inference. Dr. Bandyopadhyay is also a member of editorial board/reviewers for many international journals. He has published more than 55 international journals of repute and 4 international books on sampling theory.

Beresovsky Vladislav is a research mathematical statistician working for the Office of Survey Methods Research at the U.S. Bureau of Labor Statistics. Prior to that, he worked as a mathematical statistician at the Centers for Disease Control, National Center for Health Statistics. His research focuses on estimation from probability survey data and nonprobability data sources, survey nonresponse adjustment and spatio-temporal small area estimation. His work has been presented at multiple conferences and published in journals and conference proceedings.

Mirosław Błażej completed studies in the field of physics and engineering at the Technical University of Wroclaw and PhD studies at the Warsaw School of Economics. He is also a graduate of the National School of Public Administration. He is currently a director of the Macroeconomic Studies and Finance Statistics Department, Statistics Poland, and was a deputy director of Financial Policy and Analysis, Ministry of Finance. His research interest concentrates mainly on macroeconomic analysis, especially public finances, productivity, and business cycle analysis, including methods and business tendency surveys.

Borkowski Mateusz (PhD in social sciences in the discipline of economics and finance) is an Assistant at the Department of Political Economics, Faculty of Economics and

Finance, University of Bialystok. His main areas of interest include: economic theory (in particular: macroeconomics, theory of innovation and institutional economics), methodology of economics, econometrics (in particular: structural equation modelling - SEM). He is a member of the Polish Economic Society.

Elsherpieny E.A. is a Full Professor at Cairo University, within the Faculty of Graduate Studies for Statistical Research and is a renowned expert in the field of mathematical statistics. His research interests encompass bivariate distributions, life testing, competing risks, and accelerated life testing. He actively contributes to the scholarly community as a reviewer for numerous esteemed publications in his field. Furthermore, he serves as the Chief Editor for the Computational Journal of Mathematical and Statistical Sciences.

Garg Prachi is a research scholar in the Department of Statistics at the Department of Statistics, St. John's College, Agra, affiliated to Dr. Bhim Rao Ambedkar University Agra. Her main fields of interest include survey sampling, missing data imputation, and statistical modeling. She is currently pursuing research focusing on super population models and imputation strategies in the context of sample surveys. She has presented her work at national conferences. She is proficient in statistical programming using R. Her ongoing doctoral research aims to develop efficient estimators for handling non-response in complex survey designs.

Gershunskaya Julie is a mathematical statistician with the Statistical Methods Staff of the Office of Employment and Unemployment Statistics at the U.S. Bureau of Labor Statistics. Her main areas of interest include statistical data integration, small area estimation, and treatment of influential observations, with application to the U.S. Current Employment Statistics Program.

Górajski Mariusz gained a PhD in mathematics and a habilitation degree in economics and finance at the University of Lodz. He is an Associate Professor in the Department of Econometrics at the University of Lodz and a consultant at the Macroeconomic Studies and Finance Statistics Department, Statistics Poland. His recent principal scientific interests have focused on optimal monetary and macroprudential policy rules, the total factor productivity of enterprises, and business cycle analysis. He is a co-author of more than 40 papers that have appeared in international journals and the proceedings of national and international conferences. He is a member of the Polish Mathematical Society.

Gruszewska Ewa (PhD with habilitation in economics) is an Associate Professor at the Department of Political Economics, Faculty of Economics and Finance, University of Bialystok. Her main areas of interest include: macroeconomics and institutional eco-

nomics. She examines factors of economic growth and development, including the impact of institutions on the economies of Poland and Central and Eastern European countries. She is a member of the Polish Economic Society and the Forum of Institutional Thought.

Gupta Arindam is currently working as Professor of Statistics in the Department of Statistics, The University of Burdwan, Burdwan, West Bengal, India. He has more than eighteen years teaching as well as research experience. His research field is Demography, Occupational Mobility, Operations Research, Growth Curve Modelling, Multivariate Analysis, Bio-Statistics etc. He has published more than 37 research papers in international/national journals and conferences. Dr. Gupta is also a life member of The Calcutta Statistical Association and Indian Association for the Study of Population. He got the prestigious award "K. Srinivasan Award" form Indian Association for Study of Population in 2023.

Hassan Amal S. is a Full Professor at Cairo University, within the Faculty of Graduate Studies for Statistical Research, and is a renowned expert in the field of mathematical statistics. Her scholarly research is centered on the development of novel statistical distributions, reliability analysis. Her significant contributions to the field are evidenced by notable publications such as "Optimum Step-Stress Accelerated Life Test Plan for Lomax Distribution" and "Type II Half Logistic Family of Distributions." With over 3,664 citations, her work has demonstrably impacted statistical methodologies. She actively engages in collaborative research with colleagues worldwide, establishing her as a prominent figure in the advancement of statistical science. Furthermore, she serves as an editor for two statistics journals and contributes her expertise as a reviewer for numerous prestigious publications within her discipline.

Koczy Julianna – is a student of Economics at the Faculty of Economics at the University of Economics in Katowice, Chairperson of the Management Student Research Group "Menedżer", Deputy Chairperson of the Market and Consumption Student Research Group "SprzedaJEMY!", and Deputy Chairperson of the Student Research Group "HR-owców". She is a member of the Programme Council for the Economics programme for the period 2024-2028. For her research activities, she was awarded the National Bank of Poland Scholarship "Złote Indeksy NBP" for the 2024/2025 academic year.

Kubiczek Kubiczek PhD, is an Assistant Professor at the Department of Economic and Financial Analysis at the University of Economics in Katowice. His research focuses on financial decision-making, green consumer behavior, and sustainability, particularly in the context of how organizations respond to evolving consumer expectations. He has been involved in numerous research projects, including those funded by the National Science Centre (NCN) in Poland, exploring the interplay between consumer behavior,

ecological awareness, and economic motivations. In his teaching, Dr. Kubiczek specializes in quantitative methods — such as statistics and econometrics — with a strong emphasis on their practical application to real-world economic and business challenges.

Madukaife Mbanefo S. is an Associate Professor in the Department of Statistics, Faculty of Physical Sciences at the University of Nigeria, Nsukka. His main research interests include goodness-of-fit techniques, characterizations of multivariate statistical distributions, multivariate statistical inference, multivariate computational statistics, classification and clustering techniques, as well as sample survey methodology and nonparametric estimation. Dr. Madukaife is an active member of many scientific professional bodies.

Majumder Sanjoy is currently doing PhD (Statistics) from Aliah University, Kolkata, West Bengal, India. He has more than 4 years teaching experience as Statistician Cum Tutor in the Department of Community Medicine in different medical colleges. His research interests are statistical sample surveys, survival snalysis, ciostatistics. He has published three research papers in international/national journals and conferences.

Marszałek Marta is an Assistant Professor at the SGH Warsaw School of Economics, Institute of Statistics and Demography. Analyst and expert collaborating with Statistics Poland in the field of social statistics and satellite accounts, including the Household Production Satellite Account and the Social Economy Satellite Account. Her research interests and professional experience focus on: domestic unpaid work, non-market household production, the care economy, generational economics, time transfers, satellite accounts, the measurement of gender inequalities, care work, and the future of work.

Nduka Uchenna Chinedu is a Senior Lecturer in the Department of Statistics, University of Nigeria, Nsukka. He holds Bachelor's, Master's, and Doctorate degrees in Statistics. His research interests include time series analysis, computational statistics, statistical modelling, and data science. A seasoned data analyst and scientist, Dr. Nduka is skilled in predictive analytics and proficient in R and Python. He has published in peer-reviewed journals and presented at international conferences, reflecting a strong commitment to advancing statistical knowledge and practice.

Ossai Everestus O. is an Associate Professor at the Department of Statistics, Faculty of Physical Sciences, University of Nigeria, Nsukka. His research interests are stochastic processes, Markovian modelling of dynamic systems and statistical manpower planning and control. He has published many research papers in international/national journals and conferences.

Rasyid Sapriadi is a private sector professional working as a Procurement Officer. He is responsible for updating the categorization of over 70,000 merchandise items and

analysing more than 300 types of case categories available across all company branches. His analytical work serves as a reference for providing procurement recommendations. In addition, he manages Purchase Orders using an ERP system based on item requests.

Roszko-Wójtowicz Elżbieta is an Associate Professor at the University of Lodz (Department of Economic and Social Statistics), specialises in the measurement of socio-economic phenomena, including innovativeness, entrepreneurship, education, digitalisation, and tourism. She leads the Academy of Sustainable Tourism project, funded by the Ministry of Education and Science under the "Science for Society II" programme. She is Editor-in-Chief of *Folia Oeconomica* and serves as a regular reviewer for national and international academic journals.

Savitsky Terrance is a Senior Research Statistician in the Office of Survey Methods Research in the U.S. Bureau of Labor Statistics. His main areas of interest include Bayesian hierarchical models for: 1. Respondent-level survey data; 2. Small domain estimation; 3. Combining randomized and non-randomized data; 4. Production of synthetic data equipped with a formal privacy guarantee

Siswanto Siswanto is an Associate Professor at the Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Hasanuddin. His research field is spatial statistics, especially spatial in the fields of health sciences, economics and geostatistics. To date, he has written 60 articles. He is also active in participating in statistics forum activities, namely FORSTAT, which is an adequate forum for institutions providing higher education in the field of statistics, of course to improve the quality of statistics providers in Indonesia.

Sahriman Sitti is a faculty member at the Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University. Her expertise lies in applied statistics, with a special emphasis on time series modelling and downscaling techniques. She has been actively involved in numerous research projects and publications related to these topics. In addition, she regularly engages in statistical community activities, including participation in FORSTAT—a key forum that connects higher education institutions in statistics and supports efforts to improve the quality of statistical education and practice in Indonesia.

Srivastava Manoj. K. served as Professor of Statistics in Institute of Social Sciences, Agra affiliated to Dr. Bhim Rao Ambedkar University, Agra. Presently, he is working as Vice Chancellor of Shaheed Mahendra Karma University Bastar, Chattisgarh, India. His research interests are sampling theory, statistical inference, Bayesian analysis, multivariate analysis and data analysis. Professor Srivastava has published research papers in international/national journals and conferences. He has also published two books.

Srivastava, Namita is a Professor and Head at the Department of Statistics, St. John's College, Agra, affiliated to Dr. Bhim Rao Ambedkar University Agra. Her main areas of interest include: sampling theory, robust estimation, superpopulation model and inferential sampling.

Ulrichs Magdalena holds a PhD in Economics from the University of Lodz. She is an Assistant Professor in the Department of Econometrics at the University of Lodz and a consultant in the Department of Macroeconomic Studies and Finance Statistics at Statistics Poland. Her research interests primarily focus on applied econometric methods in business cycle analysis, total factor productivity of enterprises, and modelling structural changes in the macroeconomy.

Izabela Waszkiewicz is an early-career professional specializing in the development and regulation of artificial intelligence. Since 2022, she has been part of a dedicated team within the public sector, actively contributing to initiatives related to AI policy and implementation. She holds a postgraduate degree in Business AI: Artificial Intelligence Project Management from Kozminski University (2023).

Woś Klaudia is a graduate of the University of Economics in Katowice. She received a master's degree in Finance and Accounting and a bachelor's degree in Economics. She is interested in digital technologies in finance, statistics and monetary policy.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* https://sit.stat.gov.pl/ForAuthors.

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **Bold**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, Key words (in bold) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper.
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, (**1.1.**, **1.2.** …), **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).