STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

CONTENTS

Volume 19, Number 4, December 2018

EDITORIAL OFFICE **ISSN 1234-7655**

Scientific Secretary
Marek Cierpiał-Wolan, e-mail: m.wolan@stat.gov.pl
Secretary
Patryk Barszcz, e-mail: P.Barszcz@stat.gov.pl, phone number 00 48 22 — 608 33 66
Technical Assistant
Rajmund Litkowiec, e-mail: r.litkowiec@stat.gov.pl

Address for correspondence
Statistics Poland, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tel./fax:00 48 22 — 825 03 95

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 569−571

FROM THE EDITOR

With this issue of our quarterly – which is the last for the previous year 2018 – we successfully conclude the systematically intensified activities towards upgrading the Statistics in Transition new series' position in terms of its overall quality assessment, visibility and recognition. Indeed, the SiTns has significantly progressed recently, both as regards the number of new international bases and systems of indexation – amounted to 22 currently – and of the points which count for the impact factors of some of the most prestigious systems/bases, such as Scopus or Index Copernicus, RePec, and others. Such achievements are actually reported on the current basis in the column "Indexing and Abstracting" in the e-SiT-bulletin (on the journal's website: http://stat.gov.pl/en/sit-en).

Since authors of the articles published in the SiTns, as well as reviewers of all the submitted papers, constitute the core of contributors to the journal's achievements, we would like to honour them for their generous input through publishing in this issue their names, respectively, in the "Index of authors" (of all articles published over the past year) and in the "Acknowledgements to reviewers". On behalf of the whole Editorial Office and myself, I would like to express my gratitude and appreciation to all collaborators and supporters, including members of the Editorial Board and the panel of Associate Editors, who provide us with assistance and guidance both in strategic and practical matters on the continuous basis.

*

This issue starts with sampling and estimation section containing three papers. The first one, by **G. N. Singh, Amod Kumar** and **Gajendra K. Vishwakarma** entitled ***Development of chain-type exponential estimators for population variance in two-phase sampling design in presence of random non-response*** presents the results of an investigation aimed at dealing with a unified approach of estimation procedures of population variance in two-phase sampling design under missing at random non-response mechanism circumstances. Using two auxiliary variables, the authors have developed different chain-type exponential estimators of finite population variance for two different set-ups and studied their properties under the different assumption of random non-response. The comparisons of the proposed estimators have been made with some contemporary estimators of population variance under the similar realistic conditions. Numerical illustrations are presented to support the theoretical results. The proposed estimation procedures may be recommended to the survey statisticians for their practical application whenever they intend to deal with the sensitive or stigmatizing attributes such as drinking alcohol, gambling habit, drug addiction, tax evasion, history of induced abortions, etc.

M. R. Irshad's and **R. Maya's** paper ***On a less cumbersome method of estimation of parameters of Lindley distribution by order statistics*** presents U-statistics derived as suitable from a sample of any size exceeding a specified

integer to estimate the location and scale parameters of Lindley distribution. No evaluation of made of means, variances or co-variances of order statistics of an equivalent sample size arising from the corresponding standard form of distribution. The exact variances of the estimators have been also obtained. For practising statisticians the results derived in the paper seem to be helpful, when they look for estimators of parameters of Lindley distribution using ordered random variables.

In the next article, ***Extended exponentiated power Lindley distribution*** by **Vahid Ranjbar, Morad Alizadeh, Gholamhossein Hamedani** a new model, the Extended Exponentiated Power Lindley distribution is introduced, which extends the Lindley distribution and has increasing, bathtub and upside down shapes for the hazard rate function. It also includes the power Lindley distribution as a special case. Several statistical properties of the distribution are explored, such as the density, hazard rate, survival, quantile functions, and moments. Estimation using the maximum likelihood method and inference on a random sample from this distribution are investigated. A simulation study is performed to compare the performance of the different parameter estimates in terms of bias and mean square error. A real data set is applied to illustrate the applicability of the new model as well. Empirical findings show that the proposed model provides better fits than other well-known extensions of Lindley distributions.

The next section, research articles, also contains three articles. **Hikaru Hasegawa's** and **Pink Gao's** paper ***Bayesian spatial analysis of chronic diseases in elderly Chinese people using a STAR model*** addresses the problem of analysing chronic diseases affecting the health of elderly Chinese people, concentrating on the spatial aspect of these diseases and the respective risk factors. A structured additive regression model is applied using the R2BayesX package and data from the Chinese Urban and Rural Elderly Population Surveys for years 2000, 2006, and 2010. The major findings are as follows: (i) the covariates of considerable importance for chronic diseases are gender, smoking, drinking, province, time, age, cultural activities, years of education, and sports activities; (ii) the effect of marital status is negligible; (iii) province is a critical factor, with the highest spatial effect appearing in two types of provinces: economically developed provinces, and economically backward provinces; time also has considerable effects. Authors recommend the need for policies towards further strengthening investment in rural areas and economically backward provinces, and better education of the population on the harmful effects of smoking and drinking alcohol on health.

In the next paper, ***Lindley Pareto distribution*****, Nouara Lazri, Halim Zeghdoudi, Djabrane Yahia** introduce a new Lindley Pareto distribution which offers a more flexible framework for modelling lifetime data. Some of its mathematical properties like density function, cumulative distribution, mode, mean, variance, and Shannon entropy are established. Following a simulation study carried out to examine the bias and mean square error of the maximum likelihood estimators of the unknown parameters, three real data sets are also used. They illustrate the importance and the flexibility of the proposed distribution. According to the authors, the Lindley Pareto distribution can be used quite effectively in analysing real lifetime data and actuarial science.

Marek Walesiak's paper ***The choice of normalization method and rankings of the set of objects based on composite indicator values*** starts with observation that normalization methods lead to different rankings of the set of objects based on composite indicator values. Author considers 18 normalization methods and 5 aggregation measures (composite indicators) showing which of the methods lead to identical rankings of the set of objects, and reducing their number to 10 normalization procedures. A way of separation of groups of normalization methods leading to similar rankings is proposed (using Kendall's tau coefficient and cluster analysis). The simulation results for five composite indicators are complemented by an empirical example.

In the other articles section, the paper by **Dominika Marta Urbańczyk** and **Joanna Małgorzata Landmesser** entitled ***The comparison of income distributions for women and men in Poland using semiparametric reweighting approach*** presents the results of a comparison of the income distributions for women and men in Poland. The gender wage gap can only be partially explained by differences in men's and women's characteristics. The unexplained part of the gap is usually attributed to the wage discrimination. The authors employed the Oaxaca-Blinder decomposition procedure for the pay gap along the whole income distribution and a semiparametric reweighting approach to describe differences between the two income distributions. The reweighting factor was computed for each observation by estimating a logit model for probabilities of belonging to men's or women's group. In effect, the inequalities are decomposed into the explained and unexplained components using data from the EU-SILC for Poland, 2014.

The last section, research communicates and letters, contains the paper by **Urszula Ala-Karvia, Marta Hozer-Koćmiel, Sandra Misiak-Kwit,** and **Barbara Staszko** entitled ***Is Poland becoming Nordic? Changing trends in household structures in Poland and Finland with the emphasis on people living alone***. A comparative analysis of the household structure and its dynamics between post-economic-transformation Poland and Scandinavian-welfare-state Finland is presented with focus on one-person households (OPH). Two interrelated hypotheses concerning similarity-dissimilarity between the household structure in Finland and Poland with suggestion that the differences will be diminishing. At a glance, the analyses based on data for 2005–2015 seemed to confirm that while one- or two-person households are the dominating household structure in Finland, in Poland this structure was more balanced. For instance, the share of OPH among all households in 2015 was noticeably larger in Finland (42%) than in Poland (24%) and the difference between the countries was not diminishing. A simple extrapolation leads to prediction that under the currently observed trend the shares of OPH in the two countries will go further apart (e.g., in 2030, 46 percent of Finnish households and 22 percent of Polish households will be one-person households). In general, the position of people living alone is still different between Poland and Finland, and Poland has not gone Nordic in this respect.

Włodzimierz Okrasa

Editor

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 573

SUBMISSION INFORMATION FOR AUTHORS

Statistics in Transition new series (SiT) is an international journal published jointly by the Polish Statistical Association (PTS) and the Central Statistical Office of 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:
P.Barszcz@stat.gov.pl.,
GUS / Central Statistical Office
Al. Niepodległości 208, R. 287, 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: http://stat.gov.pl/en/sit-en/guidelines-for-authors/

EDITORIAL POLICY

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
- Central and Eastern European Online Library (CEEOL)
- Central European Journal of Social Sciences and Humanities (CEJSH)
- CNKI Scholar (China National Knowledge Infrastructure)
- Current Index to Statistics (CIS)
- Dimensions
- EconPapers
- Elsevier - Scopus
- ERIH Plus
- Google Scholar
- Index Copernicus
- J-Gate
- JournalGuide
- JournalTOCs
- Keepers Registry
- MIAR
- OpenAIRE
- ProQuest - Summon
- Publons
- RePec
- WorldCat
- Zenodo

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 575–596, DOI 10.21307/stattrans-2018-031

DEVELOPMENT OF CHAIN-TYPE EXPONENTIAL ESTIMATORS FOR POPULATION VARIANCE IN TWO-PHASE SAMPLING DESIGN IN PRESENCE OF RANDOM NON-RESPONSE

G. N. Singh[1], **Amod Kumar**[1], **Gajendra K. Vishwakarma**[1]

ABSTRACT

In this paper, an investigation has been carried out to deal with a unified approach of estimation procedures of population variance in two-phase sampling design under missing at random non-response mechanism circumstances. Using two auxiliary variables, we have developed different chain-type exponential estimators of finite population variance for two different set-ups and studied their properties under the different assumption of random non-response considered by Tracy and Osahan (1994). The comparisons of the proposed estimators have been made with some contemporary estimators of population variance under the similar realistic conditions. Numerical illustrations are presented to support the theoretical results. Results are analysed and suitable recommendations are put forward to the survey statisticians.

Key words: two-phase sampling, random non-response, variance estimation, study variable, auxiliary information, bias, mean square error

Mathematics Subject Classification: 62D05

1. Introduction

It is well known that in sample surveys the finite population parameter can be estimated more accurately by making use of information on an auxiliary variable x that is correlated with the study variable y. Sometimes, information on auxiliary variable x is not known in advance for all the units of population, for such a situation two-phase sampling is a well-established technique for generating the valid estimates of unknown population parameters of auxiliary variable x in the first phase sample. Ratio, product and regression methods of estimation are good illustrations in this context. Some pioneer works in this direction have been done by several authors, see Chand (1975), Kiregyera (1980), Mukherjee et al. (1987), Singh and Upadhyaya (1995), Pradhan (2005), Singh and Vishwakarma (2007) and Singh and Majhi (2014) among others, in two-phase sampling set-up.

It may be noted that most of the related work of estimation of population variance in sample surveys is based on the assumption of complete response

[1] Department of applied Mathematics, Indian Institute of Technology (ISM) Dhanbad, Jharkhand-826004, India. E-mail: gnsingh_ism@yahoo.com, amod.ism01@gmail.com, vishwagk@rediffmail.com.

from the sample data such as Das and Tripathi (1978), Srivastava and Jhaji (1980), Isaki (1983), Singh (1983), Upadhyay and Singh (1983), Tripathi *et al.* (1988), Biradar and Singh (1994) and Ahmed *et al.* (2003) among others. However, in some practical situations, it is a common experience in sample surveys that the information cannot always be obtained from all the units selected in the sample. For instance, in the first attempt we are not able to collect information from the selected families while some of them may decline to cooperate with the interviewer even if contacted. This results in incomplete data, and this incompleteness is known as non-response and sometimes a huge amount of non-response can completely deviate from desired estimation. Rubin (1976) recommend three particular causes of non-response: missing at random (MAR), observed at random (OAR), and parameter distribution (PD). The missing at random (MAR) response mechanism is helpful in the estimation of population parameters (means, variance, etc.) in economical way even in the presence of non-response in the survey data. Rubin (1976), Tracy and Osahan (1994), Heitzan and Basu (1996), Singh and Joarder (1998), Singh *et al.* (2000) and Singh and Tracy (2001) have suggested the estimators for estimating the finite population parameters (mean, variance, etc.) under the different type random non-response situation. Singh *et al.* (2003), Singh *et al.* (2012) and Bandyopadhyay and Singh (2015) have developed a class of estimators of population variance in two-phase sampling under the situation of random non-response (MAR). Singh *et al.* (2007) studied the properties of a family of estimators for population mean, ratio and product under the above situation of random non-response. Further improvement in the estimation procedure for population variance in the presence of non-response using multi-auxiliary characters in two-phase sampling was suggested by Ahmad *et al.* (2013) under the strategy given by Hansen and Hurwitz (1946).

In the follow-up to the above work and utilizing two auxiliary variables, we have developed some chain-type exponential estimators for estimation of population variance in the presence of random non-response based on missing at random (MAR) response mechanism under two different set-ups of two-phase sampling and studied their properties. The behaviours of the proposed estimators are studied and results are supported with suitable empirical studies, which are followed by suitable recommendation to the survey practitioners.

2. Two-Phase Sampling Structures

Let $U = (U_1, U_2, ..., U_N)$ be the finite population of N units, the character under study is denoted by variable y and the two auxiliary characters are represented by variables x and z respectively with population means $\bar{Y}, \bar{X}$ and $\bar{Z}$. Let y_i, x_i and z_i be the values of y, x and z for the i-th $(i = 1, 2, ..., N)$ unit in the population. To estimate the population variance $S_y^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (y_i - \bar{Y})^2$ of study variable y in the presence of auxiliary

characters x and z, where the population variance $S_x^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(x_i - \bar{X})^2$ of x is unknown but the information on z is available for all the units of population, we use the following two-phase sampling scheme. To furnish estimate of population variance S_x^2 of auxiliary variable x, a first phase sample S' of size n is drawn by simple random sampling without replacement (SRSWOR) scheme from the entire population U and observed for the auxiliary characters x and z to estimate S_x^2. Further, a second phase sample S of size m (m <n) is drawn from the first phase sample by the method of simple random sampling without replacement and information on the study variable y and x is gathered.

Case-I: Second phase sample S is drawn as a subsample of the first phase sample $(\text{i.e.} S \subset S')$.

Case-II: Second phase sample S is drawn independently of the first phase sample S' $(\text{i.e.} S \subset S')$.

3. Non-Response Probability Model

If random non-response situations occur at the second phase sample S of size m and r $\{r = 0,1,....,(m-2)\}$ denotes the numbers of sampling units on which information could not be obtained due to random non-response, then the observations of the respective variables on which the random non-response occur can be taken from the remaining responding (m - r) units of the second phase sample. Since we are considering the problem of unbiased estimation of finite population variance S_y^2, therefore it is assumed that r is less than $(m-1)$, i.e. $0 \le r \le (m-2)$ and p stands for the probability of non-response among the $(m-2)$ possible values of non-response, hence r is following discrete distribution; see Singh and Joarder (1998).

$$P(r) = \frac{(m-r)}{mq+2p}\,{}^{m-2}C_r p^r q^{m-2-r}\ ,\ r = 0,1,....,(m-2) \qquad (1)$$

where $q = (1-p)$.

Here ${}^{m-2}C_r$ denotes the total number of ways of obtaining r non-response out of total possible $(m-2)$ non-response.

Hence, from now onwards, we use the following notations:

$\bar{Y}$: The population mean of study variable y.

$\bar{X}, \bar{Z}$: The population means of auxiliary variables x and z respectively.

$\bar{y}_m, \bar{x}_m, \bar{x}_n, \bar{z}_m, \bar{z}_n$: The sample means of the respective variables based on the sample sizes shown in the suffices.

$\bar{y}_m^* = \frac{1}{(m-r)}\sum_{i=1}^{m-r} y_i$, $\bar{x}_m^* = \frac{1}{(m-r)}\sum_{i=1}^{m-r} x_i$ and $\bar{z}_m^* = \frac{1}{(m-r)}\sum_{i=1}^{m-r} z_i$: The sample means of the respective variables based on the responding part of the second phase sample S.

$S_z^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(z_i - \bar{Z})^2$: Population variance of the auxiliary variable z.

$s_{y_m}^2 = \frac{1}{(m-1)}\sum_{i=1}^{m}(y_i - \bar{y}_m)^2$: Sample variance of the study variable y based on sample of size m.

$s_{x_m}^2, s_{x_n}^2$, $s_{z_m}^2$ and $s_{z_n}^2$: Sample variance of the auxiliary characters x and z respectively based on the respective sample sizes shown in their subscripts.

$S_{y_m}^{*2} = \frac{1}{(m-r-1)}\sum_{i=1}^{m-r}(y_i - \bar{y}_m^*)$: Sample variance of the study variable y based on the responding part of the second phase sample S.

$S_{x_m}^{*2}$ and $S_{z_m}^{*2}$: Sample variance of the auxiliary characters x and z respectively based on the responding part and sample sizes shown in their subscripts.

4. Proposed Strategies

Following the work of Isaki (1983) and utilizing information on an auxiliary variable x with unknown S_x^2, one may propose the ratio type estimator of population variance S_y^2 in two-phase sampling as

$$t_R = s_{y_m}^2 \frac{s_{x_n}^2}{s_{x_m}^2} \tag{2}$$

Singh and Joarder (1998) have proposed ratio type estimators at the second phase Sample S under random non-response different situation as presented below.

(i) If random non-response occurs only on study variable y at the second phase and the population variance S_x^2 of auxiliary variable x is unknown, then the estimator may be defined as

$$t_1 = s_{y_m}^{*2} \frac{s_{x_n}^2}{s_{x_m}^2} \tag{3}$$

(ii) If random non-response occurs on both variables y and x and the population variance S_x^2 of auxiliary variable x is unknown, then the estimator may be defined as

$$t_2 = s_{y_m}^{*2} \frac{s_{x_n}^2}{s_{x_m}^{*2}} \tag{4}$$

(iii) In this situation, we consider that random non-response occurs on study variable y as well as auxiliary variables x and z at the second phase sample S and the population variance S_z^2 of auxiliary variable z is unknown, then the estimator may be defined as

$$t_3 = s_{y_m}^{*2} \frac{s_{x_n}^2}{s_{x_m}^{*2}} \frac{s_{z_n}^2}{s_{z_m}^{*2}} \tag{5}$$

(iv) In this situation, we assume that random non-response occurs on study variable y as well as auxiliary variable z at the second phase sample S and the population variance S_z^2 of auxiliary variable z is unknown, then the estimator may be defined as

$$t_4 = s_{y_m}^{*2} \frac{s_{x_n}^2}{s_{x_m}^2} \frac{s_{z_n}^2}{s_{z_m}^{*2}} \tag{6}$$

Following the above suggestions, it is assumed that a complete response situation occurs at the first phase sample S' while non-response situation occurs over all variables y, x and z or in different way in the second phase sample S. We have developed different chain-type exponential estimators of population variance S_y^2 in two-phase sampling design when the population variance S_x^2 of auxiliary variable x is unknown, which may be useful for real life situations such as **(i)**. In the household survey, we considered household size as the auxiliary variable for the estimation of family expenditures. Information may be obtained completely on family size, while there may be random non-response on household expenditure **(ii)**. In the agricultural survey, expenditures of fertilizer or pesticides on crop may be used as the auxiliary variable for estimating the production of crop. There may be random non-response on both the variables. We have presented the following strategies I-IV for handling the above real life situations:

Strategies I: In this situation, we assume that the information on variable y could not be obtained for r units while the complete information on variable x is available at the second phase sample S and the population variance S_z^2 of auxiliary variable z is known. Then, the estimators of finite population variance S_y^2 may be obtained as:

$$T_1 = \frac{s_{y_m}^{*2}}{s_{x_m}^2} s_{x_n}^2 \exp\left\{\frac{S_z^2 - s_{z_n}^2}{S_z^2 + s_{z_n}^2}\right\} \tag{7}$$

$$T_2 = s_{y_m}^{*2} \exp\left\{\frac{s_{x_n}^2 - s_{x_m}^2}{s_{x_n}^2 + s_{x_m}^2}\right\}\left(\frac{S_z^2}{s_{z_n}^2}\right) \tag{8}$$

Strategies II: When random non-response occurs on the study variable y as well as auxiliary variable x at the second phase sample S and the population variance S_z^2 of auxiliary variable z is known. Then, the estimators of finite population variance S_y^2 may be defined as:

$$T_3 = \frac{s_{y_m}^{*2}}{s_{x_m}^{*2}} s_{x_n}^2 \exp\left\{\frac{S_z^2 - s_{z_n}^2}{S_z^2 + s_{z_n}^2}\right\} \tag{9}$$

$$T_4 = s_{y_m}^{*2} \exp\left\{\frac{s_{x_n}^2 - s_{x_m}^{*2}}{s_{x_n}^2 + s_{x_m}^{*2}}\right\}\left(\frac{S_z^2}{s_{z_n}^2}\right) \tag{10}$$

Strategies III: In this situation, it is considered that the random non-response occurs on the study variable y as well as on the auxiliary variables x and z in the second phase sample S and the population variance S_z^2 of the auxiliary variable z is unknown. Then, the estimators of finite population variance S_y^2 may be defined as:

$$T_5 = \frac{s_{y_m}^{*2}}{s_{x_m}^{*2}} s_{x_n}^2 \exp\left\{\frac{s_{z_n}^2 - s_{z_m}^{*2}}{s_{z_n}^2 + s_{z_m}^{*2}}\right\} \tag{11}$$

$$T_6 = s_{y_m}^{*2} \exp\left\{\frac{s_{x_n}^2 - s_{x_m}^{*2}}{s_{x_n}^2 + s_{x_m}^{*2}}\right\}\left(\frac{s_{z_n}^2}{s_{z_m}^{*2}}\right) \tag{12}$$

Strategies IV: In this situation, we assume that the random non-response occurs on the study variable y and the auxiliary variable z with unknown population variance S_z^2 while the complete information on the auxiliary variable x is available. Then, the estimators of finite population variance S_y^2 may be obtained as:

$$T_7 = \frac{s_{y_m}^{*2}}{s_{x_m}^2} s_{x_n}^2 \exp\left\{\frac{s_{z_n}^2 - s_{z_m}^{*2}}{s_{z_n}^2 + s_{z_m}^{*2}}\right\} \tag{13}$$

$$T_8 = s_{y_m}^{*2} \exp\left\{\frac{s_{x_n}^2 - s_{x_m}^2}{s_{x_n}^2 + s_{x_m}^2}\right\}\left(\frac{s_{z_n}^2}{s_{z_m}^{*2}}\right) \tag{14}$$

5. Properties of Proposed estimators T_i (i = 1, 2,...., 8)

In this section, we derived the bias and mean square errors of the proposed estimators T_i, (i = 1, 2,...., 8) up to the first order of approximation under large sample assumption by using the following transformations:

$$s_{y_m}^{*2} = S_y^2(1+e_0),\ s_{x_m}^{*2} = S_x^2(1+e_1),\ s_{z_m}^{*2} = S_z^2(1+e_2),\ s_{x_m}^2 = S_x^2(1+e_3),$$

$$s_{x_n}^2 = S_x^2(1+e_4),$$

$$s_{z_n}^2 = S_z^2(1+e_5)$$

Such that $|e_i| < 1 \forall$ (i = 1, 2,...., 5)

We have derived the bias and mean square errors of the proposed estimators T_i, (i = 1, 2,...., 8) separately for the cases I and II of the two-phase sampling structure defined in section 2 and present them below.

5.1. Bias and Mean Square Error of proposed estimators under case I

In this section, we have considered that the second phase sample S of size m is drawn as a subsample from the first phase sample S' of size n and we have the following results.

$$E(e_0^2) = f^* C_0^2,\ E(e_1^2) = f^* C_1^2,\ E(e_2^2) = f^* C_2^2,\ E(e_3^2) = f_1 C_1^2,$$

$$E(e_4^2) = f_2 C_1^2, E(e_5^2) = f_2 C_2^2,\ E(e_0 e_1) = f^* \rho_{01} C_0 C_1,$$

$$E(e_0 e_2) = f^* \rho_{02} C_0 C_2,\ E(e_0 e_3) = f_1 \rho_{01} C_0 C_1,\ E(e_0 e_4) = f_2 \rho_{01} C_0 C_1,$$

$$E(e_0 e_5) = f_2 \rho_{02} C_0 C_2,\ E(e_1 e_2) = f^* \rho_{12} C_1 C_2,\ E(e_1 e_3) = f_1 C_1^2,$$

$$E(e_1 e_4) = f_2 C_1^2,\ E(e_1 e_5) = f_2 \rho_{12} C_1 C_2,\ E(e_2 e_3) = f_1 \rho_{12} C_1 C_2,$$

$$E(e_2 e_4) = f_2 \rho_{12} C_1 C_2,\ E(e_2 e_5) = f_2 C_2^2,\ E(e_3 e_4) = f_2 C_1^2,$$

$$E(e_3 e_5) = f_2 \rho_{12} C_1 C_2,\ E(e_4 e_5) = f_2 \rho_{12} C_1 C_2$$

where $f^{*} = \left(\frac{1}{mq+2p} - \frac{1}{N}\right)$, $f_1 = \left(\frac{1}{m} - \frac{1}{N}\right)$, $f_2 = \left(\frac{1}{n} - \frac{1}{N}\right)$,

$f_3 = \left(\frac{1}{m} - \frac{1}{n}\right)$, $f' = \left(\frac{1}{mq+2p} - \frac{1}{n}\right)$,

$\mu_{rst} = E\left[\left(y_i - \bar{Y}\right)^r \left(x_i - \bar{X}\right)^s \left(z_i - \bar{Z}\right)^t\right]$; $(r,s,t) \geq 0$ are integers,

$\lambda_{rst} = \mu_{rst} \Big/ \left(\mu_{200}^{r/2} \mu_{020}^{s/2} \mu_{002}^{t/2}\right)$, $C_0 = \sqrt{(\lambda_{400} - 1)}$, $C_1 = \sqrt{(\lambda_{040} - 1)}$,

$C_2 = \sqrt{(\lambda_{004} - 1)}$, $\rho_{01} = (\lambda_{220} - 1) \Big/ \sqrt{(\lambda_{400} - 1)(\lambda_{040} - 1)}$,

$\rho_{02} = (\lambda_{202} - 1) \Big/ \sqrt{(\lambda_{400} - 1)(\lambda_{004} - 1)}$,

$\rho_{12} = (\lambda_{022} - 1) \Big/ \sqrt{(\lambda_{040} - 1)(\lambda_{004} - 1)}$.

From the above expectations, it is to be noted that:

(a) If p = 0 (there is absence of non-response), the above expected values coincide with the usual results.

(b) ρ_{01} is the correlation coefficient between $\left(y - \bar{Y}\right)^2$ and $\left(x - \bar{X}\right)^2$. Similarly, ρ_{12} is the correlation coefficient between $\left(x - \bar{X}\right)^2$ and $\left(z - \bar{Z}\right)^2$ and ρ_{02} is the correlation coefficient between $\left(y - \bar{Y}\right)^2$ and $\left(z - \bar{Z}\right)^2$; for instance see Upadhyaya and Singh (2006).

Under the above transformations the estimators T_i, (i = 1, 2,...., 8) take the following forms:

$$T_1 = S_y^2 (1+e_0)(1+e_3)^{-1}(1+e_4)\exp\left\{-\frac{1}{2}e_5\left(1+\frac{e_5}{2}\right)^{-1}\right\} \quad (15)$$

$$T_2 = S_y^2 (1+e_0)(1+e_5)^{-1}\exp\left\{\frac{1}{2}(e_4 - e_3)\left(1+\frac{(e_4+e_3)}{2}\right)^{-1}\right\} \quad (16)$$

$$T_3 = S_y^2 (1+e_0)(1+e_1)^{-1}(1+e_4)\exp\left\{-\frac{1}{2}e_5\left(1+\frac{e_5}{2}\right)^{-1}\right\} \quad (17)$$

$$T_4 = S_y^2\left(1+e_0\right)\left(1+e_5\right)^{-1}\exp\left\{\frac{1}{2}\left(e_4 - e_1\right)\left(1+\frac{\left(e_4+e_1\right)}{2}\right)^{-1}\right\} \tag{18}$$

$$T_5 = S_y^2\left(1+e_0\right)\left(1+e_1\right)^{-1}\left(1+e_4\right)\exp\left\{\frac{1}{2}\left(e_5 - e_2\right)\left(1+\frac{\left(e_5+e_2\right)}{2}\right)^{-1}\right\} \tag{19}$$

$$T_6 = S_y^2\left(1+e_0\right)\left(1+e_2\right)^{-1}\left(1+e_5\right)\exp\left\{\frac{1}{2}\left(e_4 - e_1\right)\left(1+\frac{\left(e_4+e_1\right)}{2}\right)^{-1}\right\} \tag{20}$$

$$T_7 = S_y^2\left(1+e_0\right)\left(1+e_3\right)^{-1}\left(1+e_4\right)\exp\left\{\frac{1}{2}\left(e_5 - e_2\right)\left(1+\frac{\left(e_5+e_2\right)}{2}\right)^{-1}\right\} \tag{21}$$

$$T_8 = S_y^2\left(1+e_0\right)\left(1+e_2\right)^{-1}\left(1+e_5\right)\exp\left\{\frac{1}{2}\left(e_4 - e_3\right)\left(1+\frac{\left(e_4+e_3\right)}{2}\right)^{-1}\right\} \tag{22}$$

Now, we expand the right-hand side of equation (15) binomially, up to the first order of approximations, and we have the following expression of the estimator T_1 as:

$$T_1 = S_y^2\left[1+e_0 - e_3+e_4 - \frac{1}{2}e_5+e_3^2+\frac{3}{8}e_5^2 - e_0e_3+e_0e_4 - e_3e_4 - \frac{1}{2}\left(e_0e_5 - e_3e_5+e_4e_5\right)\right] \tag{23}$$

Similarly, we can express the right-hand side of equations (16)-(22) up to the first order of approximations, and we have the expression of the estimators T_i, (i = 2, 3,...., 8).

Taking expectations on both sides of the equation (23) and similarly processing for equations (16)-(22) and then using the expected values of e_i (i = 0, 1,...., 5) , we obtain the bias B(.) and mean square errors M(.) of estimators T_i (i = 1, 2,...., 8) up to the first order of approximations as:

$$B\left(T_1\right) = S_y^2\left[f_3C_1^2+\frac{3}{8}f_2C_2^2 - f_3\rho_{01}C_0C_1 - \frac{1}{2}f_2\rho_{02}C_0C_2\right] \tag{24}$$

$$B\left(T_2\right) = S_y^2\left[f_2C_2^2+\frac{3}{8}f_3C_1^2 - f_2\rho_{02}C_0C_2 - \frac{1}{2}f_3\rho_{01}C_0C_1\right] \tag{25}$$

$$B\left(T_3\right) = S_y^2\left[f'C_1^2+\frac{3}{8}f_2C_2^2 - f'\rho_{01}C_0C_1 - \frac{1}{2}f_2\rho_{02}C_0C_2\right] \tag{26}$$

$$B(T_4) = S_y^2\left[f_2C_2^2+\frac{3}{8}f'C_2^2 - f_2\rho_{02}C_0C_2 - \frac{1}{2}f'\rho_{01}C_0C_1\right] \quad (27)$$

$$B(T_5) = S_y^2\left[f'C_1^2+\frac{3}{8}f'C_2^2 - f'\rho_{01}C_0C_1 - \frac{1}{2}f'C_2\left(\rho_{02}C_0 - \rho_{12}C_1\right)\right] \quad (28)$$

$$B(T_6) = S_y^2\left[f'C_2^2+\frac{3}{8}f'C_1^2 - f'\rho_{02}C_0C_2 - \frac{1}{2}f'C_1\left(\rho_{01}C_0 - \rho_{12}C_2\right)\right] \quad (29)$$

$$B(T_7) = S_y^2\left[f_3C_1^2+\frac{3}{8}f'C_2^2 - f_3\rho_{01}C_0C_1 - \frac{1}{2}f'\rho_{02}C_0C_2+\frac{1}{2}f_3\rho_{12}C_1C_2\right] \quad (30)$$

$$B(T_8) = S_y^2\left[f'C_2^2+\frac{3}{8}f_3C_1^2 - f'\rho_{02}C_0C_2 - \frac{1}{2}f_3C_1\left(\rho_{01}C_0 - \rho_{12}C_2\right)\right] \quad (31)$$

$$M(T_1) = S_y^4\left[f^*C_0^2+f_3C_1^2+\frac{1}{4}f_2C_2^2 - 2f_3\rho_{01}C_0C_1 - f_2\rho_{02}C_0C_2\right] \quad (32)$$

$$M(T_2) = S_y^4\left[f^*C_0^2+f_2C_2^2+\frac{1}{4}f_3C_1^2 - f_3\rho_{01}C_0C_1 - 2f_2\rho_{02}C_0C_2\right] \quad (33)$$

$$M(T_3) = S_y^4\left[f^*C_0^2+f'C_1^2+\frac{1}{4}f_2C_2^2 - 2f'\rho_{01}C_0C_1 - f_2\rho_{02}C_0C_2\right] \quad (34)$$

$$M(T_4) = S_y^4\left[f^*C_0^2+f_2C_2^2+\frac{1}{4}f'C_1^2 - f'\rho_{01}C_0C_1 - 2f_2\rho_{02}C_0C_2\right] \quad (35)$$

$$M(T_5) = S_y^4\left[f^*C_0^2+f'C_1^2+\frac{1}{4}f'C_2^2 - f'\rho_{02}C_0C_2+f'\rho_{12}C_1C_2 - 2f'\rho_{01}C_0C_1\right] \quad (36)$$

$$M(T_6) = S_y^4\left[f^*C_0^2+f'C_2^2+\frac{1}{4}f'C_1^2 - f'\rho_{01}C_0C_1+f'\rho_{12}C_1C_2 - 2f'\rho_{02}C_0C_2\right] \quad (37)$$

$$M(T_7) = S_y^4\left[f^*C_0^2+f_3C_1^2+\frac{1}{4}f'C_2^2 - f'\rho_{02}C_0C_2+f_3\rho_{12}C_1C_2 - 2f_3\rho_{01}C_0C_1\right] \quad (38)$$

and

$$M(T_8) = S_y^4\left[f^*C_0^2+f'C_2^2+\frac{1}{4}f_3C_1^2 - f_3\rho_{01}C_0C_1+f_3\rho_{12}C_1C_2 - 2f'\rho_{02}C_0C_2\right] \quad (39)$$

5.2. Bias and Mean Square Error of proposed estimators under case II

If the second phase sample S is drawn independently of the first phase sample S', then we have the following results.

$$E\left(e_0^2\right) = f^* C_0^2,\ E\left(e_1^2\right) = f^* C_1^2,\ E\left(e_2^2\right) = f^* C_2^2,\ E\left(e_3^2\right) = f_1 C_1^2,\ E\left(e_4^2\right) = f_2 C_1^2$$

$$,\ E\left(e_5^2\right) = f_2 C_2^2,\ E\left(e_0 e_1\right) = f^* \rho_{01} C_0 C_1,\ E\left(e_0 e_2\right) = f^* \rho_{02} C_0 C_2,$$

$$E\left(e_0 e_3\right) = f_1 \rho_{01} C_0 C_1, E\left(e_1 e_2\right) = f^* \rho_{12} C_1 C_2,\ E\left(e_1 e_3\right) = f_1 C_1^2,$$

$$E\left(e_2 e_3\right) = f_1 \rho_{12} C_1 C_2,\ E\left(e_4 e_5\right) = f_2 \rho_{12} C_1 C_2$$

$$E\left(e_0 e_4\right) = E\left(e_0 e_5\right) = E\left(e_1 e_4\right) = E\left(e_1 e_5\right) = E\left(e_2 e_4\right) = E\left(e_2 e_5\right) = E\left(e_3 e_4\right) = E\left(e_3 e_5\right) = 0$$

Proceeding as section 5.1 and using the expected value as section 5.2, we have derived the expressions for bias B(.) and mean square errors M(.) of the proposed estimators T_i (i = 1, 2,...., 8) to the first order of approximations as:

$$B\left(T_1\right) = S_y^2 \left[f_1 C_1^2 + \frac{3}{8} f_2 C_2^2 - f_1 \rho_{01} C_0 C_1 - \frac{1}{2} f_2 \rho_{02} C_0 C_2 \right] \tag{40}$$

$$B\left(T_2\right) = S_y^2 \left[f_2 C_2^2 + \frac{1}{8} C_1^2 \left(3 f_1 - f_2\right) - \frac{1}{2} \left(f_1 \rho_{01} C_0 C_1 + f_2 \rho_{12} C_1 C_2\right) \right] \tag{41}$$

$$B\left(T_3\right) = S_y^2 \left[f^* C_1^2 + \frac{3}{8} f_2 C_2^2 - f^* \rho_{01} C_0 C_1 - \frac{1}{2} f_2 \rho_{12} C_1 C_2 \right] \tag{42}$$

$$B\left(T_4\right) = S_y^2 \left[f_2 C_2^2 + \frac{1}{8} C_1^2 \left(3 f^* - f_2\right) - \frac{1}{2} C_1 \left(f^* \rho_{01} C_0 + f_2 \rho_{12} C_2\right) \right] \tag{43}$$

$$B\left(T_5\right) = S_y^2 \left[f^* C_1^2 + \frac{1}{8} C_2^2 \left(3 f^* - f_2\right) - f^* \rho_{01} C_0 C_1 + \frac{1}{2} \left\{ \left(f_2 + f^*\right) \rho_{12} C_1 C_2 - f^* \rho_{02} C_0 C_2 \right\} \right] \tag{44}$$

$$B\left(T_6\right) = S_y^2 \left[f^* C_2^2 + \frac{1}{8} C_1^2 \left(3 f^* - f_2\right) - f^* \rho_{02} C_0 C_2 + \frac{1}{2} \left\{ \left(f_2 + f^*\right) \rho_{12} C_1 C_2 - f^* \rho_{01} C_0 C_1 \right\} \right] \tag{45}$$

$$B\left(T_7\right) = S_y^2 \left[f_1 C_1^2 + \frac{1}{8} C_2^2 \left(3 f^* - f_2\right) - f_1 \rho_{01} C_0 C_1 + \frac{1}{2} \left\{ \left(f_1 + f_2\right) \rho_{12} C_1 C_2 - f^* \rho_{02} C_0 C_2 \right\} \right] \tag{46}$$

$$B(T_8) = S_y^2\left[f^*C_2^2 + \frac{1}{8}C_1^2(3f_1 - f_2) - f^*\rho_{02}C_0C_2 + \frac{1}{2}\{(f_1+f_2)\rho_{12}C_1C_2 - f_1\rho_{01}C_0C_1\}\right] \quad (47)$$

$$M(T_1) = S_y^4\left[f^*C_0^2 + (f_1+f_2)C_1^2 + \frac{1}{4}f_2C_2^2 - 2f_1\rho_{01}C_0C_1 - f_2\rho_{12}C_1C_2\right] \quad (48)$$

$$M(T_2) = S_y^4\left[f^*C_0^2 + f_2C_2^2 + \frac{1}{4}(f_1+f_2)C_1^2 - f_1\rho_{01}C_0C_1 - f_2\rho_{12}C_1C_2\right] \quad (49)$$

$$M(T_3) = S_y^4\left[f^*C_0^2 + (f^*+f_2)C_1^2 + \frac{1}{4}f_2C_2^2 - 2f^*\rho_{01}C_0C_1 - f_2\rho_{12}C_1C_2\right] \quad (50)$$

$$M(T_4) = S_y^4\left[f^*C_0^2 + f_2C_2^2 + \frac{1}{4}(f^*+f_2)C_1^2 - f^*\rho_{01}C_0C_1 - f_2\rho_{12}C_1C_2\right] \quad (51)$$

$$M(T_5) = S_y^4\left[f^*C_0^2 + (f^*+f_2)\left(C_1^2 + \frac{1}{4}C_2^2 + \rho_{12}C_1C_2\right) - f^*(\rho_{02}C_0C_2 + 2\rho_{01}C_0C_1)\right] \quad (52)$$

$$M(T_6) = S_y^4\left[f^*C_0^2 + (f^*+f_2)\left(C_2^2 + \frac{1}{4}C_1^2 + \rho_{12}C_1C_2\right) - f^*(\rho_{01}C_0C_1 + 2\rho_{02}C_0C_2)\right] \quad (53)$$

$$M(T_7) = S_y^4\left[f^*C_0^2 + (f_1+f_2)(C_1^2 + \rho_{12}C_1C_2) + \frac{1}{4}(f^*+f_2)C_2^2 - f^*\rho_{02}C_0C_2 - 2f_1\rho_{01}C_0C_1\right] \quad (54)$$

and

$$M(T_8) = S_y^4\left[f^*C_0^2 + (f^*+f_2)C_2^2 + (f_1+f_2)\left(\frac{1}{4}C_1^2 + \rho_{12}C_1C_2\right) - f_1\rho_{01}C_0C_1 - 2f^*\rho_{02}C_0C_2\right] \quad (55)$$

6. Efficiency Comparisons of the Proposed Estimators T_i (i = 1, 2,...., 8)

In this section, we validate the performance of the proposed estimators T_i (i = 1, 2,..., 8) with to respect to the estimators such as population variance $s_{y_m}^{*2}$

(sample variance estimator in presence of random non-response) and t_i (i = 1, 2, ..., 4). The variance/mean square errors of the estimators $s^{*2}_{y_m}$ and t_i (i = 1, 2, ..., 4) up to the first order of approximation under case I and case II are respectively given by:

Case I:

$$V\left(s^{*2}_{y_m}\right) = f^* C_0^2 S_y^4 \tag{56}$$

$$M(t_1) = S_y^4\left[f^* C_0^2 + f_3 C_1^2 - 2f_3\rho_{01}C_0C_1\right] \tag{57}$$

$$M(t_2) = S_y^4\left[f^* C_0^2 + f'\left(C_1^2 - 2\rho_{01}C_0C_1\right)\right] \tag{58}$$

$$M(t_3) = S_y^4\left[f^* C_0^2 + f'\left(C_1^2 + C_2^2\right) - 2f'\left(\rho_{01}C_0C_1 + \rho_{02}C_0C_2 - \rho_{12}C_1C_2\right)\right] \tag{59}$$

and

$$M(t_4) = S_y^4\left[f^* C_0^2 + f_3\left(C_1^2 - 2\rho_{01}C_0C_1 + 2\rho_{12}C_1C_2\right) + f'\left(C_2^2 - 2\rho_{02}C_0C_2\right)\right] \tag{60}$$

Case II:

$$V\left(s^{*2}_{y_m}\right) = f^* C_0^2 S_y^4 \tag{60}$$

$$M(t_1) = S_y^4\left[f^* C_0^2 + \left(f_1 + f_2\right)C_1^2 - 2f_1\rho_{01}C_0C_1\right] \tag{61}$$

$$M(t_2) = S_y^4\left[f^* C_0^2 + \left(f^* + f_2\right)C_1^2 - 2f^*\rho_{01}C_0C_1\right] \tag{62}$$

$$M(t_3) = S_y^4\left[f^* C_0^2 + \left(f^* + f_2\right)\left(C_1^2 + C_2^2 + 2\rho_{12}C_1C_2\right) - 2f^*\left(\rho_{01}C_0C_1 + \rho_{02}C_0C_2\right)\right] \tag{63}$$

and

$$M(t_4) = S_y^4\left[f^* C_0^2 + \left(f_1 + f_2\right)\left(C_1^2 + 2\rho_{12}C_1C_2\right) + \left(f^* + f_2\right)C_2^2 - 2f_1\rho_{01}C_0C_1 - 2f^*\rho_{02}C_0C_2\right] \tag{64}$$

The performances of our proposed estimators T_i (i = 1, 2,...., 8) are compared with the other estimators considered in this paper and their dominance

is examined below through empirical studies carried over three different populations.

7. Numerical Illustration

We have computed the percent relative efficiencies of the proposed estimators T_i (i = 1, 2,...., 8) with respect to $s^{*2}_{y_m}$ and t_i (i = 1, 2, ..., 4) based on three natural populations. The source of the populations, the nature of the variables y, x, z and the values of the various parameters are given as follows.

Population I- Source: Sukhatme and Sukhatme [1970] (page-185)

y: Area under wheat in 1937.

x: Area under wheat in 1936.

z: Total cultivated area in 1931.

N = 34, $C_0 = 1.5959$, $C_1 = 1.5105$, $C_2 = 1.3200$,

$\rho_{01} = 0.6251$, $\rho_{02} = 0.8007$, $\rho_{12} = 0.5342$

Population II- Source: Murthy[1967] (page-399)

y: Area under wheat in 1964.

x: Area under wheat in 1963.

z: Total cultivated area in 1961.

N 33= 34, $C_0 = 1.6510$, $C_1 = 1.3828$, $C_2 = 1.3447$,

$\rho_{01} = 0.9218$, $\rho_{02} = 0.8914$, $\rho_{12} = 0.9346$

Population III- Source: Satici and Kadilar (2011)

This data set is about 923 district of Turkey.

y: Number of successful students.

x: Numbers of teachers.

z: Private teaching institutions.

N = 261, $C_0 = 1.86537$, $C_1 = 1.75941$, $C_2 = 2.02126$,

$\rho_{01} = 0.970$, $\rho_{02} = 0.935$, $\rho_{12} = 0.928$

Table 1. Percent relative efficiency of the estimators T_1 and T_2 with respect to other estimators when non-response situation occur only on study variable y at the second phase sample.

Population I								
Estimators	Case I				Case II			
	T_1		T_2		T_1		T_2	
	$s^{*2}_{y_m}$	t_1	$s^{*2}_{y_m}$	t_1	$s^{*2}_{y_m}$	t_1	$s^{*2}_{y_m}$	t_1
P=0.02	143.93	108.55	164.43	124.02	125.32	103.74	142.77	118.19
P=0.04	142.41	108.26	161.91	123.08	124.55	103.63	141.30	117.57
P=0.06	140.95	107.97	159.49	122.18	123.81	103.52	139.89	116.97
P=0.08	139.54	107.70	157.19	121.32	123.09	103.41	138.53	116.39
P=0.10	138.18	107.43	154.98	120.49	122.38	103.31	137.21	115.83
Population II								
Estimators	Case I				Case II			
	T_1		T_2		T_1		T_2	
	$s^{*2}_{y_m}$	t_1	$s^{*2}_{y_m}$	t_1	$s^{*2}_{y_m}$	t_1	$s^{*2}_{y_m}$	t_1
P=0.02	472.17	132.00	253.54	*	486.82	127.79	226.25	*
P=0.04	433.14	128.65	244.45	*	445.12	124.79	219.54	*
P=0.06	400.53	125.84	236.12	*	410.51	122.30	213.30	*
P=0.08	372.89	123.47	228.45	*	381.30	120.20	207.50	*
P=0.10	349.15	121.42	221.36	*	356.33	118.41	202.09	*
Population III								
Estimators	Case I				Case II			
	T_1		T_2		T_1		T_2	
	$s^{*2}_{y_m}$	t_1	$s^{*2}_{y_m}$	t_1	$s^{*2}_{y_m}$	t_1	$s^{*2}_{y_m}$	t_1
P=0.02	752.54	230.81	352.72	108.18	733.68	216.06	232.38	*
P=0.04	660.37	212.33	334.63	107.59	646.11	200.02	225.98	*
P=0.06	588.63	197.95	318.38	107.07	577.51	187.46	219.96	*
P=0.08	531.19	186.44	303.71	106.59	522.32	177.35	214.27	*
P=0.10	484.18	177.01	290.38	106.16	476.96	169.04	208.89	*

* Indicate, proposed estimator is not preferable over existing estimator.

Table 2. Percent relative efficiency of the estimators T_3 and T_4 with respect to other estimators when non-response situation occur on study variable y and auxiliary variable x at the second phase sample.

Population I								
Estimators	Case I				Case II			
	T_3		T_4		T_3		T_4	
	$s_{y_m}^{*2}$	t_2	$s_{y_m}^{*2}$	t_2	$s_{y_m}^{*2}$	t_2	$s_{y_m}^{*2}$	t_2
P=0.02	145.37	108.64	166.86	124.69	126.41	103.78	144.59	118.70
P=0.04	145.24	108.42	166.63	124.39	126.72	103.69	144.89	118.57
P=0.06	145.12	108.21	166.41	124.09	127.02	103.61	145.19	118.43
P=0.08	145.00	108.00	166.20	123.79	127.31	103.53	145.48	118.30
P=0.10	144.88	107.79	165.99	123.50	127.61	103.45	145.77	118.17
Population II								
Estimators	Case I				Case II			
	T_3		T_4		T_3		T_4	
	$s_{y_m}^{*2}$	t_2	$s_{y_m}^{*2}$	t_2	$s_{y_m}^{*2}$	t_2	$s_{y_m}^{*2}$	t_2
P=0.02	522.02	135.38	263.10	*	539.98	130.82	233.82	*
P=0.04	524.28	134.68	262.71	*	541.95	130.18	234.15	*
P=0.06	526.53	133.97	262.32	*	543.91	129.55	234.46	*
P=0.08	528.77	133.28	261.95	*	545.84	128.93	234.78	*
P=0.10	530.99	132.58	261.58	*	547.77	128.30	235.09	*
Population III								
Estimators	Case I				Case II			
	T_3		T_4		T_3		T_4	
	$s_{y_m}^{*2}$	t_2	$s_{y_m}^{*2}$	t_2	$s_{y_m}^{*2}$	t_2	$s_{y_m}^{*2}$	t_2
P=0.02	884.16	253.69	371.83	106.69	858.24	235.77	240.53	*
P=0.04	893.16	251.93	370.69	104.56	867.27	234.26	241.87	*
P=0.06	902.30	250.15	369.57	102.45	876.44	232.73	243.23	*
P=0.08	911.59	248.34	368.45	100.37	885.76	231.17	244.59	*
P=0.10	921.02	246.50	367.35	98.31	895.24	229.59	245.97	*

Table 3. Percent relative efficiency of the estimators T_5 and T_6 with respect to other estimators when non-response situation occur on study variable y as well as auxiliary variable x and z at the second phase sample

Population I

Estimators	Case I				Case II			
	T_5		T_6		T_5		T_6	
	$s^{*2}_{y_m}$	t_3	$s^{*2}_{y_m}$	t_3	$s^{*2}_{y_m}$	t_3	$s^{*2}_{y_m}$	t_3
P=0.02	146.38	134.61	207.69	190.99	122.06	146.59	175.25	210.47
P=0.04	146.61	134.78	208.44	191.62	122.71	146.51	176.45	210.66
P=0.06	146.83	134.95	209.18	192.25	123.36	146.42	177.65	210.86
P=0.08	147.05	135.11	209.91	192.87	124.01	146.33	178.85	211.05
P=0.10	147.27	135.27	210.64	193.49	124.65	146.25	180.05	211.24

Population II

Estimators	Case I				Case II			
	T_5		T_6		T_5		T_6	
	$s^{*2}_{y_m}$	t_3	$s^{*2}_{y_m}$	t_3	$s^{*2}_{y_m}$	t_3	$s^{*2}_{y_m}$	t_3
P=0.02	305.47	209.83	291.80	200.44	239.88	231.07	232.09	223.57
P=0.04	307.57	210.96	293.68	201.43	242.44	231.67	234.45	224.03
P=0.06	309.68	212.08	295.55	202.41	245.02	232.27	236.81	224.48
P=0.08	311.78	213.20	297.42	203.39	247.62	232.87	239.19	224.94
P=0.10	313.87	214.32	299.28	204.36	250.23	233.47	241.58	225.40

Population III

Estimators	Case I				Case II			
	T_5		T_6		T_5		T_6	
	$s^{*2}_{y_m}$	t_3	$s^{*2}_{y_m}$	t_3	$s^{*2}_{y_m}$	t_3	$s^{*2}_{y_m}$	t_3
P=0.02	217.16	234.38	182.54	197.01	124.44	256.51	103.06	212.44
P=0.04	218.91	236.38	183.57	198.23	126.17	257.50	104.37	212.99
P=0.06	220.67	238.41	184.61	199.45	127.94	258.50	105.70	213.56
P=0.08	222.46	240.46	185.66	200.68	129.75	259.53	107.05	214.13
P=0.10	224.27	242.53	186.71	201.91	131.61	260.58	108.44	214.72

Table 4. Percent relative efficiency of the estimators T_7 and T_8 with respect to other estimators when non-response situation occur on study variable y and auxiliary variable z at the second phase sample

Population I

Estimators	Case I				Case II			
	T_7		T_8		T_7		T_8	
	$s^{*2}_{y_m}$	t_4	$s^{*2}_{y_m}$	t_4	$s^{*2}_{y_m}$	t_4	$s^{*2}_{y_m}$	t_4
P=0.02	147.06	133.29	208.21	188.73	122.53	145.54	175.62	208.60
P=0.04	147.96	132.15	209.49	187.11	123.66	144.40	177.20	206.93
P=0.06	148.86	131.01	210.76	185.50	124.79	143.26	178.79	205.25
P=0.08	149.75	129.88	212.02	183.89	125.92	142.11	180.38	203.56
P=0.10	150.65	128.75	213.29	182.28	127.06	140.96	181.97	201.87

Population II

Estimators	Case I				Case II			
	T_7		T_8		T_7		T_8	
	$s^{*2}_{y_m}$	t_4	$s^{*2}_{y_m}$	t_4	$s^{*2}_{y_m}$	t_4	$s^{*2}_{y_m}$	t_4
P=0.02	300.94	203.60	292.64	197.98	237.08	225.91	232.62	221.67
P=0.04	298.60	198.65	295.37	196.50	236.83	221.43	235.52	220.20
P=0.06	296.33	193.85	298.11	195.01	236.59	217.02	238.45	218.72
P=0.08	294.13	189.20	300.85	193.52	236.35	212.69	241.40	217.23
P=0.10	292.00	184.69	303.61	192.03	236.12	208.43	244.38	215.72

Population III

Estimators	Case I				Case II			
	T_7		T_8		T_7		T_8	
	$s^{*2}_{y_m}$	t_4	$s^{*2}_{y_m}$	t_4	$s^{*2}_{y_m}$	t_4	$s^{*2}_{y_m}$	t_4
P=0.02	217.24	230.09	184.35	195.26	124.47	254.06	103.64	211.54
P=0.04	219.07	227.76	187.27	194.69	126.23	252.54	105.55	211.17
P=0.06	220.93	225.40	190.26	194.11	128.03	250.98	107.52	210.79
P=0.08	222.80	223.02	193.33	193.52	129.87	249.38	109.56	210.39
P=0.10	224.70	220.60	196.49	192.91	131.75	247.75	111.67	209.98

The empirical studies are carried out for different choices of non-response rate p, the performances of the proposed estimators T_i (i = 1, 2,...., 8) have been shown in terms of the percent relative efficiencies with respect to other estimators.

$$\text{PRE} = \frac{M(\delta)}{M(T_i)} \times 100 \text{ , (i = 1, 2 ,..., 8)}$$

8. Interpretations of Empirical Results

The following interpretations can be read out from the present study.

From Tables 1-4, it is visible that almost all the values of percent relative efficiencies are exceeding 100 for all the parametric combinations, which indicate that the proposed estimators are uniformly dominating over the existing estimators as considered in this work.

(a) From Tables 1 and 3, it may be seen that the values of percent relative efficiencies decrease and increase respectively for both the cases as the values of non-response rate p increase.

(b) Further, when the random non-response rate p increases we observe the zig-zag trend in Tables 2 and 4.

9. Conclusions

In this paper, we have studied different chain-type exponential estimators for improving estimation of the population variance under the situation of random non-response. Following the analyses of effective estimation procedures, it has been found that the results are highly desirable, which indicate the proposition of proposed estimators and subsequent estimation procedures. Hence, looking on the nice behaviour, the proposed estimation procedures may be recommended to the survey statisticians for their practical application whenever they intend to deal with the sensitive or stigmatizing attributes such as drinking alcohol, gambling habit, drug addiction, tax evasion, history of induced abortions, etc.

Acknowledgements

The authors are thankful to the reviewers for their valuable suggestions, which helped in improving the quality of this paper. The authors are also grateful to the IIT(ISM), Dhanbad for providing the financial assistance and necessary infrastructure to carry out the present research work.

REFERENCES

AHMED, M. S., ABU-DAYYEH, W., HURAIRAH, A. A. O., (2003). Some estimators for population variance under two phase sampling. Statistics in Transition, 6 (1), pp. 143–150.

AHMAD, Z., ALI, S., HANIF, M., (2013). Variance estimation in two-phase sampling using multi-auxiliary variables in the presence of non-response. Pakistan Journal of Statistics, 29 (4), pp. 487–501.

BANDYOPADHYAY, A., SINGH, G. N., (2015). Estimation of Population Variance in Two-Phase Sampling in Presence of Random Non-Response. Pakistan Journal of Statistics and Operation Research, 11 (4), pp. 525–542.

BIRADAR, R. S., SINGH, H. P., (1994). An alternative to ratio estimator of population Variance. Assam Statistical Review, 8 (2), pp. 18–33.

CHAND, L., (1975). Some ratio-type estimators based on two or more auxiliary variables. Ph. D. dissertation, Iowa State University, Ames, Iowa.

DAS, A. K., TRIPATHI, T. P., (1978). Use of auxiliary information in estimating the finite population variance. Sankhya, 40(C), pp. 139–148.

HANSEN, M. H., HURWITZ, W. N., (1946). The problem of non response in sample surveys. Journal of the American Statistical Association, 41, pp. 517–529.

HEITZAN, D. F., BASU, S., (1996). Distinguish 'Missing at Random' and 'Missing Completely at Random', The American Statistician, 50, pp. 207–217.

ISAKI, C. T., (1983). Variance estimation using auxiliary information. Journal of American Statistical Association, 78, pp. 117–123.

KIREGYERA, B., (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables. Metrika, 27, pp. 217–223.

MUKHERJEE, R., RAO, T. J., VIJAYAN, K., (1987). Regression type estimators using multiple auxiliary information. Australian & New Zealand Journal of Statistics, 29, pp. 244–254.

MURTHY, M. N., (1967). Sampling Theory and Methods. Statistical Publishing Society Calcutta, India.

PRADHAN, B. K., (2005). A chain regression estimator in two-phase sampling using multi-auxiliary information. Bulletin of the Malaysian Mathematical Science Society, 28 (1), pp. 81–86.

RUBIN, D. B., (1976). Inference and missing data. Biometrika, 63 (3), pp. 581–592.

SATICI, E., KADILAR, C., (2011). Ratio estimator for the population mean at the current occasion in the presence of non-response in successive sampling. Hacettepe Journal of Mathematics and Statistics, 40 (1), pp. 115–124.

SINGH, G. N., MAJHI, D., (2014). Some chain-type exponential estimators of population mean in two-phase sampling, Statistics in Transition, 15 (2), pp. 221–230.

SINGH, G. N., UPADHYAYA, L. N., (1995). A class of modified chain type estimators using two auxiliary variables in two-phase sampling, Metron, LIII, pp. 117–125.

SINGH, H. P., TRACY, D. S., (2001). Estimation of population mean in presence of random non-response in sample surveys, Statistica, LXI, pp. 231–248.

SINGH, H. P., CHANDRA, P., SINGH, S., (2003). Variance estimation using multiauxiliary information for random non-response in survey sampling. Statistica, LXIII, pp. 23–40.

SINGH, H. P., CHANDRA, P., JOARDER, A. H., SINGH, S., (2007). Family of estimators of mean, ratio and product of a finite population using random non-response, Test. 16, pp. 565–597.

SINGH, H. P., TAILOR, R., KIM, J. M., SINGH, S., (2012). Families of estimators of finite population variance using a random non-response in survey sampling. The Korean Journal of Applied Statistics, 25 (4), pp. 681–695.

SINGH, H. P., VISHWAKARMA, G. K., (2007). Modified exponential ratio and product estimators for finite population mean in double sampling. Austrian Journal of Statistics, 36 (3), pp. 217–225.

SINGH, R. K., (1983). Estimation of finite population variance using ratio and product method of estimation. Biometrika, 25 (2), pp. 193–200.

SINGH, S., JOARDER, A. H., (1998). Estimation of finite population variance using random non-response in survey sampling. Metrika, 98, pp. 241–249.

SINGH, S., JOARDER, A. H., TRACY, D. S., (2000). Regression type estimators for random non-response in survey sampling. Statistica, LX (1), pp. 39–43.

SRIVASTAVA, S. K., JHAJJ, H. S., (1980). A class of estimators using auxiliary information for estimating finite population variance. Sankhya, C, 42 (1–2), pp. 87–96.

SUKHATME, P. V., SUKHATME, B. V., (1970). Sampling Theory of Survey with Applications. Asia Publishing House. India.

TRACY, D. S., OSAHAN, S. S., (1994). Random non-response on study variable versus on study as well as auxiliary variables. Statistica, 54, pp. 163–168.

TRIPATHI, T. P., SINGH, H. P., UPADHYAYA, L. N., (1988). A generalized method of estimation in double sampling. Journal of Indian Statistical Association, 26, pp. 91–101.

UPADHYAYA, L. N., SINGH, H. P., (1983). Use of auxiliary information in the estimation of population variance. Mathematical Forum, 6 (2), pp. 33–36.

UPADHYAYA, L. N., SINGH, H. P., (2006). Almost unbiased ratio and product-type estimators of finite population variance in sample surveys. Statistics in Transition, 7 (5), pp. 1087–1096.

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 597–620, DOI 10.21307/stattrans-2018-032

ON A LESS CUMBERSOME METHOD OF ESTIMATION OF PARAMETERS OF LINDLEY DISTRIBUTION BY ORDER STATISTICS

M. R. Irshad[1], **R. Maya**[2]

ABSTRACT

In this article, we have derived suitable U-statistics from a sample of any size exceeding a specified integer to estimate the location and scale parameters of Lindley distribution without the evaluation of means, variances and co-variances of order statistics of an equivalent sample size arising from the corresponding standard form of distribution. The exact variances of the estimators have been also obtained.

Key words: Order statistics, Lindley distribution, Best linear unbiased estimator, U-statistics.

1. Introduction

Lindley (1958) suggested a distribution to illustrate the difference between fiducial distribution and posterior distribution, the following probability density function (pdf),

$$f(x) = \frac{\theta^2}{1+\theta}(1+x)e^{-\theta x};\ x > 0,\ \theta > 0. \tag{1}$$

Ghitany et al. (2008) developed different properties of, Lindley distribution and showed that the Lindley distribution fits better than the exponential distribution based on the waiting times before service of the bank customers. Sankaran (1970) used, Lindley distribution as the mixing distribution of a Poisson parameter and the resulting distribution is known as the Poisson-Lindley distribution. Zakerzadeh and Dolati (2009) obtained a generalized Lindley distribution and discussed its various properties and applications. Ghitany et al. (2013) and Nadarajah et al. (2011) recently proposed extensions of the Lindley distribution named the generalized Lindley and power Lindley distributions respectively. A discrete form of Lindley distribution was introduced by Gómez and Ojeda (2011) by discretizing the continuous Lindley distribution. Ali et al. (2013) considered Bayesian analysis of the Lindley model via informative and non-informative priors under different loss functions. Elbatal and Elgarhy (2013) have investigated most of the statistical properties of the transmuted quasi-Lindley distribution. Kadilar and Cakmakyapan (2016) introduced in

[1]Department of Statistics, Cochin University of Science and Technology, Kochi-682 022, India.
E-mail: irshadm24@gmail.com
[2]Department of Statistics, University College, Trivandrum-695 034, India.
E-mail: mayasekharan@gmail.com

the literature, the Lindley family of distributions. Nedjar and Zehdoudi (2016) introduced gamma Lindley distribution and studied some important properties of their proposed generalization. Shibu and Irshad (2016) introduced extended version of generalized Lindley distribution, it includes all the existing Lindley models. Again, Irshad and Maya (2017) developed another form of generalization and elucidated various reliability properties of their proposed model.

Based on reparametrisation of (1.1), Sultan and Thubyani (2016) developed location scale extension of Lindley distribution and derived Best Linear Unbiased Estimators (BLUEs) of location and scale parameters based on order statistics, its pdf is given by

$$f(x) = \frac{\theta^2}{\sigma(1+\theta)}\left(1+\frac{x-\mu}{\sigma}\right)e^{-\theta\left(\frac{x-\mu}{\sigma}\right)}, \quad x > \mu \text{ and } \theta, \mu, \sigma > 0. \tag{2}$$

Even though best linear unbiased estimation of location and scale parameters using order statistics (see, Lloyd (1952)) is a widely accepted method of estimation, one serious difficulty involved in the application of this method is that in order to obtain these estimators one requires the values of means, variances and covariances of the entire order statistics of a random sample of size n arising from the corresponding standard distribution. Thus, the results of Sultan and Thubyani (2016) cannot help one to obtain the BLUEs of μ and σ for larger values of n. However, if one obtains the BLUEs of μ and σ by order statistics based on small or moderate sample of size m and use this as kernel of degree m to construct appropriate U-statistics to estimate μ and σ, then these U-statistics are highly useful as they estimate the parameters explicitly. Moreover, these estimators are highly preferred as they utilize the optimality conditions of BLUE as well as U-statistics. Thomas and Sreekumar (2004) developed the concept of U-statistics by taking BLUE based on the order statistics of a random sample of size two as kernel of degree two to estimate the scale parameter of generalized exponential distribution. Again, Thomas and Sreekumar (2007) generalized the results of Thomas and Sreekumar (2004) to generate estimators based on U-statistics for the location and scale parameters of any distribution, by taking best linear functions of order statistics of a sample of size $m < n$ as kernels.

In the work of Sultan and Thubyani (2016), they did not mentioned the means, variances and the covariances of the order statistics arising from the standard form of (1.2). In the case of location scale family of distributions, a study based on order statistics, it is necessary to evaluate the moments of order statistics arising from the corresponding standard form of the distributions.

Hence, the main objective of this work is to evaluate the moments of order statistics arising from the standard form of (1.2) for some known values of the shape parameter θ. Using these values, we determine the best linear unbiased estimators based on small sample sizes of the location and scale parameters of (1.2) and use them to generate appropriate U-statistics for estimating those parameters for any sample sizes.

2. BLUEs of location and scale parameters of Lindley distribution using order statistics

Let $\mathbf{X} = (X_{1:m}, X_{2:m}, \cdots, X_{m:m})'$ be the vector of order statistics of a random sample of size m drawn from (1.2). Define $Y_{r:m}$=$\frac{X_{r:m}-\mu}{\sigma}, r = 1,2,\cdots,m$. Then, $Y_{r:m}, r = 1,2,\cdots,m$ are distributed as the order statistics of a random sample of size m drawn from the standard form of (1.2) with pdf $f_0(y)$. Let $\alpha = (\alpha_{1:m}, \alpha_{2:m}, \cdots, \alpha_{m:m})'$ and $V = ((v_{r,s:m}))$ be the vector of means and dispersion matrix of the vector of order statistics of a random sample of size m drawn from $f_0(y)$. Then, the BLUEs of μ and σ based on order statistics given by (see, Sultan and Thubyani (2016))

$$\hat{\mu} = -\frac{\alpha' V^{-1}(\mathbf{1}\ \alpha' - \alpha\ \mathbf{1}')V^{-1}}{(\alpha' V^{-1}\alpha)(\mathbf{1}' V^{-1}\mathbf{1}) - (\alpha' V^{-1}\mathbf{1})^2}\mathbf{X}, \tag{3}$$

$$\hat{\sigma} = \frac{\mathbf{1}' V^{-1}(\mathbf{1}\ \alpha' - \alpha\ \mathbf{1}')V^{-1}}{(\alpha' V^{-1}\alpha)(\mathbf{1}' V^{-1}\mathbf{1}) - (\alpha' V^{-1}\mathbf{1})^2}\mathbf{X}, \tag{4}$$

with variances given by

$$Var(\hat{\mu}) = \frac{(\alpha' V^{-1}\alpha)\sigma^2}{(\alpha' V^{-1}\alpha)(\mathbf{1}' V^{-1}\mathbf{1}) - (\alpha' V^{-1}\mathbf{1})^2}, \tag{5}$$

$$Var(\hat{\sigma}) = \frac{(\mathbf{1}' V^{-1}\mathbf{1})\sigma^2}{(\alpha' V^{-1}\alpha)(\mathbf{1}' V^{-1}\mathbf{1}) - (\alpha' V^{-1}\mathbf{1})^2}, \tag{6}$$

where **1** is a column vector of $1's$ of the same dimension as **X**.

2.2 **U-Statistics**

Let $X_1, X_2, \cdots, X_n$ be independent observations coming from a population with distribution function $F(x;\theta)$. Then, the U-statistic for the parameter θ with the symmetric kernel $h(\cdot)$ of degree m is defined as

$$U(X_1, X_2, \cdots, X_n) = \frac{1}{\binom{n}{m}} \sum_{\beta \in B} h(X_{\beta_1}, X_{\beta_2}, \cdots, X_{\beta_m}), \tag{7}$$

where $B = \{\beta / \beta = (\beta_1, \beta_2, \cdots, \beta_m), \beta_1 < \beta_2 < \cdots < \beta_m\}$ is one of the $\binom{n}{m}$ combinations of m integers chosen without replacement from the set $(1,2,\cdots,n)$. Suppose that

$$E[h(X_1, X_2, \cdots, X_m)] = \theta \ \text{ and } \ E[h^2(X_1, X_2, \cdots, X_m)] < \infty. \tag{8}$$

Let $h(X_1, X_2, \cdots, X_\omega, X_{\omega+1}, \cdots, X_m)$ and $h(X_1, X_2, \cdots, X_\omega, X_{m+1}, \cdots, X_{2m-\omega})$ be two random variables having exactly ω samples in common, $\omega = 1,2,\cdots,m$. Let $\xi_\omega^{(m)}$ be

the covariance between these two random variables. Then, the variance of the U-statistic given in (2.5) as (see, Hoeffding (1948))

$$Var[U(X_1,X_2,\cdots,X_n)] = \frac{1}{\binom{n}{m}} \sum_{\omega=1}^{m} \binom{m}{\omega}\binom{n-m}{m-\omega}\xi_{\omega}^{(m)}. \tag{9}$$

3. Estimation of parameters of Lindley distribution using U-Statistics

Let the BLUE of μ as given in (2.1) be represented as

$$h_1(X_1,X_2,\cdots,X_m) = a_1X_{1:m} + a_2X_{2:m} + \cdots + a_mX_{m:m} \tag{10}$$

and that of σ given in (2.2) be represented as

$$h_2(X_1,X_2,\cdots,X_m) = d_1X_{1:m} + d_2X_{2:m} + \cdots + d_mX_{m:m}, \tag{11}$$

where $a_1,a_2,\cdots,a_m$ and $d_1,d_2,\cdots,d_m$ are constants. Now, we can easily write

$$U_{1:n}^{(m)} = \frac{1}{\binom{n}{m}} \sum_{r=1}^{n} \left[\sum_{i=0}^{m-1} \binom{n-r}{m-1-i}\binom{r-1}{i} a_{i+1} \right] X_{r:n} \tag{12}$$

as the U-statistic for estimating μ based on kernel given in (3.1) and

$$U_{2:n}^{(m)} = \frac{1}{\binom{n}{m}} \sum_{r=1}^{n} \left[\sum_{i=0}^{m-1} \binom{n-r}{m-1-i}\binom{r-1}{i} d_{i+1} \right] X_{r:n} \tag{13}$$

as the U-statistic for estimating μ based on kernel given in (3.2), where we define $\binom{r-1}{i} = 0$ for $i \geq r$ and $\binom{n-r}{m-1-i} = 0$ for $n-r < m-1-i$.
If we write
$\xi_{\omega}^{(m)}$=Cov$[h_1(X_1,X_2,\cdots,X_{\omega},X_{\omega+1},\cdots,X_m), h_1(X_1,X_2,\cdots,X_{\omega},X_{m+1},\cdots,X_{2m-\omega})]$, as the covariance between two $h_1(.)$ functions with exactly ω common observations and $\psi_{\omega}^{(m)}$ = Cov$[h_2(X_1,X_2,\cdots,X_{\omega},X_{\omega+1},\cdots,X_m), h_2(X_1,X_2,\cdots,X_{\omega},X_{m+1},\cdots,X_{2m-\omega})]$, as the covariance between two $h_2(.)$ functions with exactly ω common observations for $\omega = 1,2,\cdots,m$, then the variances of $U_{1:n}^{(m)}$ and $U_{2:n}^{(m)}$ are given by

$$Var[U_{1:n}^{(m)}] = \frac{1}{\binom{n}{m}} \sum_{\omega=1}^{m} \binom{m}{\omega}\binom{n-m}{m-\omega}\xi_{\omega}^{(m)} \tag{14}$$

and

$$Var[U_{2:n}^{(m)}] = \frac{1}{\binom{n}{m}} \sum_{\omega=1}^{m} \binom{m}{\omega}\binom{n-m}{m-\omega}\psi_{\omega}^{(m)}. \tag{15}$$

Clearly $\xi_m^{(m)} = V[h_1(X_1, X_2, \cdots, X_m)]$ and $\psi_m^{(m)} = V[h_2(X_1, X_2, \cdots, X_m)]$ and are given in (2.3) and (2.4) respectively. Now, we evaluate the values of $\xi_\omega^{(m)}$ and $\psi_\omega^{(m)}$ for $\omega = 1, 2, \cdots, m-1$, using the methodology developed by Thomas and Sreekumar (2008), as explained in the following steps.
Define the vectors b_{m+k} for $k = 1, 2, \cdots, m-1$ as

$$b'_{m+k} = \left[\frac{\sum_{i=0}^{m-1} \binom{m+k-1}{m-1-i}\binom{o}{i} a_{i+1}}{\binom{m+k}{m}}, \frac{\sum_{i=0}^{m-1} \binom{m+k-2}{m-1-i}\binom{1}{i} a_{i+1}}{\binom{m+k}{m}}, \cdots, \frac{\sum_{i=0}^{m-1} \binom{0}{m-1-i}\binom{m+k-1}{i} a_{i+1}}{\binom{m+k}{m}} \right] \tag{16}$$

and define $w_k = \binom{m+k}{m}(b'_{m+k} V_{m+k} b_{m+k})\sigma^2 - \xi_m^{(m)}, k = 1, 2, \cdots, m-1$ where V_{m+k} is the variance covariance matrix of the vector of order statistics of random sample of size $m+k$ drawn from the distribution with pdf $f_0(y)$ and $\xi_m^{(m)}$ defined as above. Define the matrix

$$H = \begin{bmatrix} 0 & 0 & \cdots & 0 & \binom{m}{m-1}\binom{1}{1} \\ 0 & 0 & \cdots & \binom{m}{m-2}\binom{2}{2} & \binom{m}{m-1}\binom{2}{1} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \binom{m}{1}\binom{m-1}{m-1} & \binom{m}{2}\binom{m-1}{m-2} & \cdots & \binom{m}{m-2}\binom{m-1}{2} & \binom{m}{m-1}\binom{m-1}{1} \end{bmatrix} \times \begin{bmatrix} \xi_1^{(m)} \\ \xi_2^{(m)} \\ \vdots \\ \xi_{m-1}^{(m)} \end{bmatrix} \tag{17}$$

and the vector $w = (w_1, w_2, \cdots, w_{m-1})'$. Then, the components $\xi_\omega^{(m)}, \omega = 1, 2, \cdots, m-1$, involved in (3.5) are solved from the following equations

$$\left(\xi_1^{(m)}, \xi_2^{(m)}, \cdots, \xi_{m-1}^{(m)}\right)' = H^{-1}W. \tag{18}$$

Similarly, the values of $\psi_\omega^{(m)}$, $\omega = 1, 2, \cdots, m-1$ can be obtained as

$$\left(\psi_1^{(m)}, \psi_2^{(m)}, \cdots, \psi_{m-1}^{(m)}\right)' = H^{-1}Z, \tag{19}$$

where $Z' = (z_1, z_2, \cdots, z_{m-1})'$ with $z_k = \binom{m+k}{m}(g'_{m+k} V_{m+k} g_{m+k})\sigma^2 - \psi_m^{(m)}$ and g_{m+k} is obtained from (3.7) just by replacing each a_i by d_i, $i = 1, 2, \cdots, m$.

Once we obtain the values of $\xi_\omega^{(m)}, \psi_\omega^{(m)}, \omega = 1, 2, \cdots, m-1$, from (3.9) and (3.10) respectively, then the exact variances of the U-statistics for estimating μ and σ

based on any sample of size n can be obtained by using (3.5) and (3.6) without any further direct evaluation of moments of order statistics.

Table 1: Means of order statistics arising from the standard form of (1.2) for $n = 2(1)10$ and for $\theta = 0.50(0.50)2$.

n	r	$\theta = 0.50$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$
2	1	1.88889	0.81250	0.49333	0.34722
	2	4.77778	2.18750	1.37333	0.98611
3	1	1.35254	0.56481	0.33798	0.23594
	2	2.96159	1.30787	0.80405	0.56979
	3	5.68587	2.62731	1.65798	1.19427
4	1	1.06481	0.43506	0.25773	0.17896
	2	2.21571	0.95408	0.57870	0.40687
	3	3.70748	1.66166	1.02940	0.73271
	4	6.34534	2.94920	1.86750	1.34813
5	1	0.88297	0.35464	0.20854	0.14425
	2	1.79219	0.75673	0.45452	0.31781
	3	2.85098	1.25011	0.76498	0.54045
	4	4.27847	1.93602	1.20568	0.86088
	5	6.86205	3.20250	2.03296	1.46994
6	1	0.75669	0.29971	0.17522	0.12086
	2	1.51439	0.62928	0.37511	0.26121
	3	2.34780	1.01164	0.61334	0.43101
	4	3.35416	1.48858	0.91661	0.64990
	5	4.74063	2.15975	1.35021	0.96637
	6	7.28634	3.41105	2.16951	1.57065
7	1	0.66343	0.25973	0.15114	0.10402
	2	1.31625	0.53963	0.31971	0.22192
	3	2.00972	0.85341	0.51361	0.35943
	4	2.79857	1.22262	0.74633	0.52644
	5	3.77085	1.68804	1.04432	0.74250
	6	5.12854	2.34843	1.47257	1.05592
	7	7.64597	3.58815	2.28566	1.65644
8	1	0.59150	0.22927	0.13291	0.09131
	2	1.16693	0.47290	0.27876	0.19300
	3	1.76422	0.73981	0.44254	0.30868
	4	2.41889	1.04273	0.63204	0.44402
	5	3.17826	1.40251	0.86061	0.60885
	6	4.12640	1.85937	1.15454	0.82268
	7	5.46259	2.51145	1.57858	1.13367
	8	7.95788	3.74196	2.38667	1.73112
9	1	0.53420	0.20528	0.11862	0.08137
	2	1.04989	0.42118	0.24722	0.17080
	3	1.57658	0.65390	0.38915	0.27070
	4	2.13951	0.91164	0.54933	0.38464

Table 1: Continued

n	r	$\theta = 0.50$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$
9	5	2.76810	1.20660	0.73544	0.51825
	6	3.50639	1.55924	0.96075	0.68134
	7	4.43640	2.00943	1.25144	0.89335
	8	5.75579	2.65488	1.67205	1.20233
	9	8.23314	3.87785	2.47600	1.79722
10	1	0.48740	0.18589	0.10712	0.07339
	2	0.95540	0.37987	0.22215	0.15321
	3	1.42781	0.58645	0.34748	0.24117
	4	1.92371	0.81128	0.48639	0.33963
	5	2.46322	1.06217	0.64373	0.45215
	6	3.07298	1.35102	0.82715	0.58435
	7	3.79534	1.69805	1.04982	0.74600
	8	4.71115	2.14288	1.33784	0.95650
	9	6.01695	2.78288	1.75561	1.26379
	10	8.47938	3.99951	2.55604	1.85650

Table 2: Variances and covariances $v_{r,s:n}$ of order statistics arising from the standard form of (1.2) for $1 \leq r \leq s \leq n$, $n = 2(1)10$ and $\theta = 0.50(0.50)2$.

n	r	s	$\theta = 0.50$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$
2	1	1	2.43210	0.52734	0.20996	0.10860
	1	2	2.08642	0.47266	0.19360	0.10204
	2	2	8.50617	2.02734	0.86062	0.46508
3	1	1	1.26895	0.26123	0.10077	0.05110
	1	2	1.10362	0.23853	0.09463	0.04881
	1	3	0.97997	0.21783	0.08823	0.04618
	2	2	3.03236	0.69148	0.28350	0.14931
	2	3	2.71450	0.63480	0.26528	0.14165
	3	3	8.76918	2.11496	0.90612	0.49298
4	1	1	0.80135	0.15814	0.05957	0.02978
	1	2	0.70230	0.14593	0.05649	0.02869
	1	3	0.63270	0.13553	0.05354	0.02756
	1	4	0.57352	0.12535	0.05032	0.02622
	2	2	1.67832	0.36848	0.14711	0.07609
	2	3	1.51875	0.34312	0.13966	0.07316
	2	4	1.38116	0.31802	0.13147	0.06969
	3	3	3.27371	0.76416	0.31833	0.16944
	3	4	2.99528	0.71127	0.30057	0.16174
	4	4	8.86142	2.15079	0.92644	0.50614
5	1	1	0.56080	0.10681	0.03949	0.01954
	1	2	0.49439	0.09931	0.03770	0.01892
	1	3	0.44832	0.09303	0.03603	0.01831
	1	4	0.41216	0.08737	0.03437	0.01765
	1	5	0.37866	0.08152	0.03250	0.01687

Table 2: Continued

n	r	s	$\theta = 0.50$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$
5	2	2	1.10224	0.23411	0.09149	0.04668
	2	3	1.00239	0.21965	0.08751	0.04519
	2	4	0.92339	0.20655	0.08356	0.04361
	2	5	0.84967	0.19293	0.07907	0.04169
	3	3	1.86983	0.42397	0.17271	0.09046
	3	4	1.72821	0.39952	0.16512	0.08738
	3	5	1.59442	0.37388	0.15646	0.08361
	4	4	3.39454	0.80276	0.33772	0.18102
	4	5	3.14698	0.75391	0.32084	0.17354
	5	5	8.89317	2.16701	0.93675	0.51323
6	1	1	0.41851	0.07732	0.02816	0.01381
	1	2	0.37082	0.07232	0.02701	0.01343
	1	3	0.33770	0.06814	0.02595	0.01306
	1	4	0.31233	0.06447	0.02494	0.01267
	1	5	0.29088	0.06102	0.02391	0.01226
	1	6	0.26976	0.05731	0.02271	0.01175
	2	2	0.79383	0.16373	0.06287	0.03173
	2	3	0.72437	0.15443	0.06044	0.03085
	2	4	0.67089	0.14623	0.05810	0.02996
	2	5	0.62547	0.13850	0.05572	0.02900
	2	6	0.58056	0.13016	0.05295	0.02780
	3	3	1.25600	0.27740	0.11901	0.05736
	3	4	1.16581	0.26301	0.10670	0.05573
	3	5	1.08866	0.24937	0.10241	0.05396
	3	6	1.01185	0.23457	0.09738	0.05176
	4	4	1.97728	0.45681	0.18852	0.09961
	4	5	1.85124	0.43388	0.18114	0.09652
	4	6	1.72441	0.40879	0.17245	0.09267
	5	5	3.46241	0.82558	0.34965	0.18834
	5	6	3.23830	0.78023	0.33365	0.18114
	6	6	8.89922	2.17434	0.94229	0.51735
7	1	1	0.32640	0.05873	0.02112	0.01029
	1	2	0.29052	0.05520	0.02034	0.01004
	1	3	0.26544	0.05224	0.01962	0.00979
	1	4	0.24638	0.04966	0.01894	0.00954
	1	5	0.23077	0.04732	0.01827	0.00928
	1	6	0.21685	0.04504	0.01758	0.00900
	1	7	0.20251	0.04251	0.01676	0.00865
	2	2	0.60587	0.12173	0.04604	0.02304
	2	3	0.55438	0.11528	0.04443	0.02247
	2	4	0.51511	0.10966	0.04291	0.02191
	2	5	0.48283	0.10453	0.04141	0.02132
	2	6	0.45398	0.09954	0.03985	0.02068
	2	7	0.42417	0.09399	0.03800	0.01988

Table 2: Continued

n	r	s	$\theta = 0.50$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$
7	3	3	0.92024	0.19842	0.07807	0.03996
	3	4	0.85639	0.18890	0.07543	0.03897
	3	5	0.80366	0.18020	0.07283	0.03794
	3	6	0.75633	0.17170	0.07012	0.03682
	3	7	0.70722	0.16222	0.06688	0.03540
	4	4	1.34807	0.30481	0.12374	0.06462
	4	5	1.26723	0.29107	0.11956	0.06295
	4	6	1.19420	0.27760	0.11518	0.06111
	4	7	1.11796	0.26249	0.10994	0.05878
	5	5	2.04404	0.47796	0.19905	0.10584
	5	6	1.93035	0.45653	0.19196	0.10282
	5	7	1.81053	0.43230	0.18342	0.09897
	6	6	3.50309	0.84002	0.35749	0.19328
	6	7	3.29717	0.79760	0.34229	0.18635
	7	7	8.89323	2.17717	0.94531	0.51984
8	1	1	0.26288	0.04621	0.01644	0.00797
	1	2	0.23494	0.04361	0.01588	0.00779
	1	3	0.21526	0.04142	0.01537	0.00761
	1	4	0.20032	0.03951	0.01489	0.00744
	1	5	0.18824	0.03781	0.01442	0.00727
	1	6	0.17786	0.03621	0.01396	0.00709
	1	7	0.16822	0.03462	0.01347	0.00689
	1	8	0.15794	0.03280	0.01287	0.00663
	2	2	0.48128	0.09443	0.03526	0.01751
	2	3	0.44144	0.08973	0.03413	0.01713
	2	4	0.41113	0.08565	0.03307	0.01674
	2	5	0.38656	0.08197	0.03204	0.01636
	2	6	0.36540	0.07853	0.03102	0.01595
	2	7	0.34574	0.07510	0.02993	0.01550
	2	8	0.32471	0.07118	0.02862	0.01493
	3	3	0.71207	0.15018	0.05827	0.02957
	3	4	0.66396	0.14343	0.05647	0.02892
	3	5	0.62483	0.13734	0.05474	0.02825
	3	6	0.59102	0.13163	0.05300	0.02756
	3	7	0.55952	0.12593	0.05116	0.02679
	3	8	0.52575	0.11940	0.04893	0.02580
	4	4	0.99933	0.22147	0.08862	0.04584
	4	5	0.94159	0.21223	0.08594	0.04480
	4	6	0.89149	0.20353	0.08325	0.04371
	4	7	0.84464	0.19481	0.08039	0.04250
	4	8	0.79421	0.18481	0.07691	0.04095

Table 2: Continued

n	r	s	$\theta = 0.50$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$
8	5	5	1.40849	0.32343	0.13274	0.06983
	5	6	1.33539	0.31044	0.12866	0.06816
	5	7	1.26666	0.29739	0.12431	0.06630
	5	8	1.19225	0.28234	0.11900	0.06391
	6	6	2.08826	0.49242	0.20644	0.11030
	6	7	1.98435	0.47232	0.19966	0.10736
	6	8	1.87089	0.44900	0.19131	0.10357
	7	7	3.52834	0.84958	0.36289	0.19676
	7	8	3.33691	0.80962	0.34839	0.19009
	8	8	8.88133	2.17755	0.94689	0.52138
9	1	1	0.21698	0.03735	0.01317	0.00635
	1	2	0.19465	0.03538	0.01276	0.00622
	1	3	0.17878	0.03370	0.01237	0.00609
	1	4	0.16672	0.03225	0.01202	0.00597
	1	5	0.15702	0.03095	0.01168	0.00585
	1	6	0.14881	0.02975	0.01134	0.00572
	1	7	0.14150	0.02860	0.01101	0.00559
	1	8	0.13451	0.02744	0.01064	0.00544
	1	9	0.12683	0.02609	0.01020	0.00525
	2	2	0.39367	0.07560	0.02791	0.01378
	2	3	0.36188	0.07204	0.02709	0.01350
	2	4	0.33766	0.06895	0.02631	0.01323
	2	5	0.31816	0.06619	0.02556	0.01295
	2	6	0.30164	0.06364	0.02484	0.01267
	2	7	0.28690	0.06121	0.02410	0.01238
	2	8	0.27278	0.05873	0.02331	0.01205
	2	9	0.25726	0.05584	0.02234	0.01163
	3	3	0.57214	0.11823	0.04531	0.02283
	3	4	0.53434	0.11321	0.04402	0.02237
	3	5	0.50382	0.10871	0.04279	0.02191
	3	6	0.47792	0.10456	0.04158	0.02144
	3	7	0.45475	0.10059	0.04036	0.02095
	3	8	0.43253	0.09654	0.03904	0.02039
	3	9	0.40804	0.09181	0.03742	0.01967
	4	4	0.78068	0.16980	0.06708	0.03441
	4	5	0.73678	0.16314	0.06522	0.03371
	4	6	0.69940	0.15698	0.06340	0.03299
	4	7	0.66588	0.15107	0.06155	0.03224
	4	8	0.63365	0.14504	0.05956	0.03139
	4	9	0.59805	0.13798	0.05710	0.03029

Table 2: Continued

n	r	s	$\theta = 0.50$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$
9	5	5	1.05313	0.23773	0.09631	0.05020
	5	6	1.00070	0.22890	0.09365	0.04915
	5	7	0.95351	0.22040	0.09095	0.04804
	5	8	0.90796	0.21170	0.08804	0.04679
	5	9	0.85748	0.20149	0.08443	0.04516
	6	6	1.45052	0.33672	0.13932	0.07370
	6	7	1.38371	0.32446	0.13537	0.07206
	6	8	1.31893	0.31188	0.13111	0.07021
	6	9	1.24675	0.29706	0.12581	0.06780
	7	7	2.11883	0.50271	0.21184	0.11361
	7	8	2.02279	0.48377	0.20534	0.11076
	7	9	1.91493	0.46133	0.19722	0.10703
	8	8	3.54422	0.85611	0.36673	0.19930
	8	9	3.36461	0.81822	0.35285	0.19286
	9	9	8.86655	2.17654	0.94760	0.52231
10	1	1	0.18260	0.03085	0.01079	0.00518
	1	2	0.16438	0.02931	0.01047	0.00508
	1	3	0.15131	0.02799	0.01018	0.00499
	1	4	0.14135	0.02685	0.00991	0.00490
	1	5	0.13335	0.02583	0.00965	0.00481
	1	6	0.12664	0.02490	0.00940	0.00471
	1	7	0.12078	0.02402	0.00915	0.00462
	1	8	0.11540	0.02317	0.00890	0.00452
	1	9	0.11013	0.02229	0.00862	0.00440
	1	10	0.10420	0.02125	0.00828	0.00425
	2	2	0.32931	0.06200	0.02267	0.01113
	2	3	0.30334	0.05924	0.02204	0.01092
	2	4	0.28349	0.05683	0.02146	0.01072
	2	5	0.26755	0.05468	0.02090	0.01052
	2	6	0.25417	0.05272	0.02035	0.01032
	2	7	0.24245	0.05087	0.01982	0.01011
	2	8	0.23169	0.04908	0.01927	0.00989
	2	9	0.22115	0.04722	0.01868	0.00964
	2	10	0.20928	0.04501	0.01793	0.00931
	3	3	0.47258	0.09582	0.03633	0.01818
	3	4	0.44199	0.09195	0.03536	0.01785
	3	5	0.41736	0.08850	0.03445	0.01751
	3	6	0.39665	0.08534	0.03356	0.01718
	3	7	0.37849	0.08237	0.03268	0.01683
	3	8	0.36180	0.07948	0.03178	0.01647
	3	9	0.34541	0.07648	0.03080	0.01605
	3	10	0.32694	0.07292	0.02958	0.01551

Table 2: Continued

n	r	s	$\theta = 0.50$	$\theta = 1$	$\theta = 1.5$	$\theta = 2$
10	4	4	0.63230	0.13513	0.05278	0.02687
	4	5	0.59751	0.13010	0.05142	0.02638
	4	6	0.56818	0.12550	0.05010	0.02588
	4	7	0.54242	0.12117	0.04880	0.02536
	4	8	0.51870	0.11694	0.04747	0.02481
	4	9	0.49536	0.11256	0.04601	0.02419
	4	10	0.46901	0.10734	0.04419	0.02337
	5	5	0.82859	0.18404	0.07369	0.03811
	5	6	0.78852	0.17761	0.07182	0.03739
	5	7	0.75322	0.17155	0.06997	0.03665
	5	8	0.72064	0.16562	0.06807	0.03586
	5	9	0.68851	0.15946	0.06600	0.03497
	5	10	0.65216	0.15212	0.06341	0.03380
	6	6	1.09177	0.24970	0.10210	0.05356
	6	7	1.04376	0.24130	0.09950	0.05251
	6	8	0.99930	0.23307	0.09683	0.05139
	6	9	0.95532	0.22450	0.09392	0.05012
	6	10	0.90539	0.21426	0.09026	0.04846
	7	7	1.48097	0.34656	0.14430	0.07668
	7	8	1.41930	0.33497	0.14049	0.07507
	7	9	1.35803	0.32286	0.13633	0.07324
	7	10	1.28812	0.30833	0.13108	0.07084
	8	8	2.14058	0.51027	0.21590	0.11614
	8	9	2.05101	0.49234	0.20966	0.11338
	8	10	1.94806	0.47070	0.20176	0.10973
	9	9	3.55411	0.86065	0.36954	0.20121
	9	10	3.38433	0.82453	0.35619	0.19498
	10	10	8.85046	2.17473	0.94776	0.52286

Table 3: Coefficients of $X_{i:n}$ in the BLUE $\hat{\mu}$ and $V_1 = \frac{Var(\hat{\mu})}{\sigma^2}$.

n	θ	Coefficients										
		a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	V_1
2		1.65385	-0.65385									5.77647
3		1.41191	-0.15873	-0.25319								2.19107
4		1.28908	-0.04630	-0.10253	-0.14024							1.20397
5		1.21488	-0.00631	-0.04601	-0.07171	-0.09085						0.77759
6	0.50	1.16548	0.01068	-0.01937	-0.03931	-0.05313	-0.06435					0.55037
7		1.13047	0.01841	-0.00528	-0.02145	-0.03276	-0.04107	-0.04831				0.41330
8		1.10449	0.02197	0.00266	-0.01081	-0.02035	-0.02738	-0.03279	-0.03779			0.32350
9		1.08462	0.02336	0.00736	-0.00412	-0.01237	-0.01842	-0.02310	-0.02685	-0.03047		0.26112
10		1.06900	0.02364	0.01017	0.00021	-0.00696	-0.01231	-0.01646	-0.01971	-0.02243	-0.02515	0.21582
2		1.59091	-0.59091									1.15391
3		1.37351	-0.15579	-0.21772								0.42056
4		1.26736	-0.06025	-0.09082	-0.1163							0.22354
5		1.20501	-0.02638	-0.04537	-0.06009	-0.07317						0.14033
6	1	1.16434	-0.01162	-0.02455	-0.03470	-0.04289	-0.05058					0.09690
7		1.13592	-0.00437	-0.01376	-0.02114	-0.02725	-0.03222	-0.03718				0.07121
8		1.11503	-0.00053	-0.00772	-0.01317	-0.01808	-0.02183	-0.02516	-0.02855			0.05467
9		1.09930	0.00128	-0.00375	-0.00861	-0.01224	-0.01528	-0.01785	-0.02020	-0.02265		0.04336
10		1.08676	0.00250	-0.00145	-0.00548	-0.00839	-0.01101	-0.01300	-0.01492	-0.01659	-0.01842	0.03527

Table 3: Continued

n	θ	Coefficients										
		a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	V_1
2		1.56060	-0.56060									0.44307
3		1.35722	-0.15639	-0.20083								0.15750
4		1.25905	-0.06732	-0.08654	-0.10520							0.08212
5		1.20174	-0.03502	-0.04602	-0.05547	-0.06523						0.05077
6	1.5	1.16417	-0.02017	-0.02718	-0.03343	-0.03881	-0.04457					0.03462
7		1.13813	-0.01263	-0.01739	-0.02167	-0.02526	-0.02873	-0.03246				0.02517
8		1.11867	-0.00787	-0.01177	-0.01495	-0.01730	-0.01992	-0.02215	-0.02471			0.01915
9		1.10428	-0.00623	-0.00731	-0.01074	-0.01269	-0.01410	-0.01616	-0.01755	-0.01949		0.01507
10		1.09199	-0.00323	-0.00577	-0.00783	-0.00918	-0.01075	-0.01188	-0.01326	-0.01432	-0.01576	0.01218
2		1.54347	-0.54347									0.22490
3		1.34885	-0.15753	-0.19132								0.07868
4		1.25518	-0.07150	-0.08463	-0.09904							0.04055
5		1.20010	-0.03918	-0.04681	-0.05315	-0.06096						0.02486
6	2	1.16451	-0.02438	-0.02931	-0.03266	-0.03680	-0.04135					0.01682
7		1.13961	-0.01692	-0.01933	-0.02202	-0.02441	-0.02696	-0.02998				0.01217
8		1.12049	-0.01204	-0.01306	-0.01553	-0.01744	-0.01901	-0.02069	-0.02273			0.00922
9		1.10628	-0.00907	-0.00930	-0.01160	-0.01316	-0.01374	-0.01523	-0.01631	-0.01785		0.00722
10		1.09452	-0.00546	-0.00811	-0.00887	-0.01041	-0.00992	-0.01172	-0.01246	-0.01318	-0.01438	0.00581

Table 4: Coefficients of $X_{i:n}$ in the BLUE $\hat{\sigma}$ and $V_2 = \frac{Var(\hat{\sigma})}{\sigma^2}$.

n	θ	Coefficients										
		a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	V_2
2		-0.34615	0.34615									0.81065
3		-0.31956	0.14123	0.17833								0.39436
4		-0.30376	0.07960	0.10436	0.11980							0.25806
5		-0.29298	0.05104	0.07024	0.08177	0.08993						0.19089
6	0.50	-0.28508	0.03508	0.05081	0.06049	0.06686	0.07184					0.15107
7		-0.27900	0.02514	0.03846	0.04686	0.05243	0.05638	0.05973				0.12479
8		-0.27416	0.01847	0.03002	0.03747	0.04247	0.04601	0.04867	0.05107			0.10619
9		-0.27022	0.01380	0.02394	0.03064	0.03521	0.03844	0.04086	0.04276	0.04457		0.09233
10		-0.26692	0.01037	0.01941	0.02550	0.02969	0.03270	0.03495	0.03668	0.03811	0.03951	0.08163
2		-0.72727	0.72727									0.851234
3		-0.68078	0.30627	0.37451								0.41765
4		-0.65430	0.18069	0.22097	0.25264							0.27504
5		-0.63673	0.12249	0.15124	0.17248	0.19051						0.20448
6	1	-0.62410	0.08983	0.11187	0.12838	0.14121	0.15282					0.16250
7		-0.61453	0.06931	0.08696	0.10041	0.11093	0.11941	0.12752				0.13471
8		-0.60700	0.05542	0.07004	0.08118	0.09026	0.09735	0.10337	0.10938			0.11498
9		-0.60097	0.04566	0.05765	0.06758	0.07522	0.08145	0.08659	0.09108	0.09574		0.10025
10		-0.59584	0.03818	0.04862	0.05718	0.06392	0.06952	0.07400	0.07794	0.08138	0.08511	0.08885

Table 4: Continued

n	θ	Coefficients										V_2
		a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	
2	1.5	-1.13636	1.13636									0.88246
3		-1.07630	0.49268	0.58362								0.43554
4		-1.04318	0.30025	0.34879	0.39414							0.28802
5		-1.02177	0.21044	0.24305	0.27052	0.29777						0.21479
6		-1.00639	0.15937	0.18309	0.20354	0.22106	0.23932					0.17112
7		-0.99510	0.12706	0.14532	0.16110	0.17464	0.18691	0.20007				0.14215
8		-0.98601	0.10438	0.11963	0.13226	0.14304	0.15294	0.16188	0.17187			0.12153
9		-0.97930	0.08942	0.09999	0.11152	0.12061	0.12839	0.13600	0.14270	0.15066		0.10612
10		-0.97258	0.07576	0.08675	0.09576	0.10330	0.11052	0.11646	0.12227	0.12765	0.13410	0.09416
2	2	-1.56521	1.56521									0.90548
3		-1.49633	0.69494	0.80139								0.44863
4		-1.45966	0.43336	0.48533	0.54096							0.29740
5		-1.43583	0.30938	0.34338	0.37411	0.40896						0.22223
6		-1.41966	0.23860	0.26324	0.28398	0.30498	0.32885					0.17729
7		-1.40784	0.19391	0.21121	0.22751	0.24264	0.25744	0.27513				0.14744
8		-1.39773	0.16223	0.17507	0.18858	0.20064	0.21177	0.22295	0.23650			0.12618
9		-1.39056	0.13970	0.14879	0.16078	0.17059	0.17896	0.18782	0.19649	0.20743		0.11024
10		-1.38371	0.11957	0.13164	0.13904	0.14854	0.15352	0.16222	0.16881	0.17566	0.18471	0.09789

Table 5: Values of $\xi_{\omega}^{(m)}$, for $\omega = 1,2,\cdots,m$ and $m = 2(1)5$.

		$\theta = 0.50$	$\theta = 1$	$\theta = 1.50$	$\theta = 2$
m	ω	$\sigma^{-2}\xi_{\omega}^{(m)}$	$\sigma^{-2}\xi_{\omega}^{(m)}$	$\sigma^{-2}\xi_{\omega}^{(m)}$	$\sigma^{-2}\xi_{\omega}^{(m)}$
2	1	1.05101	0.20177	0.07556	0.03792
	2	5.77647	0.15391	0.44307	0.22490
3	1	0.33650	0.06094	0.02213	0.01089
	2	0.99730	0.18551	0.06822	0.03370
	3	2.19107	0.42056	0.15750	0.07868
4	1	0.15227	0.02644	0.00933	0.00460
	2	0.37925	0.06701	0.02399	0.01165
	3	0.71139	0.12854	0.04652	0.02278
	4	1.20397	0.22354	0.08212	0.04055
5	1	0.08272	0.01377	0.00493	0.00217
	2	0.19121	0.03245	0.01138	0.00548
	3	0.33382	0.05766	0.02039	0.00987
	4	0.52247	0.09206	0.03290	0.01598
	5	0.77759	0.14033	0.05077	0.02486

Table 6: Values of $\psi_{\omega}^{(m)}$, for $\omega = 1,2,\cdots,m$ and $m = 2(1)5$.

		$\theta = 0.50$	$\theta = 1$	$\theta = 1.50$	$\theta = 2$
m	ω	$\sigma^{-2}\psi_{\omega}^{(m)}$	$\sigma^{-2}\psi_{\omega}^{(m)}$	$\sigma^{-2}\psi_{\omega}^{(m)}$	$\sigma^{-2}\psi_{\omega}^{(m)}$
2	1	0.23996	0.25864	0.27367	0.28503
	2	0.81065	0.85123	0.88246	0.90548
3	1	0.09638	0.10456	0.11086	0.11557
	2	0.22212	0.23765	0.24960	0.25826
	3	0.39436	0.41765	0.43554	0.44863
4	1	0.05147	0.05610	0.05967	0.06229
	2	0.10969	0.11862	0.12538	0.13021
	3	0.17709	0.19000	0.19981	0.20689
	4	0.25806	0.27504	0.28802	0.29740
5	1	0.03187	0.02768	0.03730	0.03842
	2	0.06611	0.07197	0.07634	0.07946
	3	0.10328	0.11181	0.11821	0.12278
	4	0.14436	0.15541	0.16376	0.16971
	5	0.19089	0.20448	0.21479	0.22223

Table 7: $V_3 = Var(U_{1:n}^{(m)})$ and $V_4 = Var(U_{2:n}^{(m)})$

		$\theta = 0.50$		$\theta = 1$	
m	n	V_3	V_4	V_3	V_4
2	5	1.20825	0.22504	0.23615	0.24031
	10	0.50206	0.10333	0.09721	0.11088
	15	0.31526	0.06714	0.06083	0.07215
	20	0.22954	0.04973	0.04421	0.05349
	30	0.14858	0.03275	0.02856	0.03525
	40	0.10981	0.02442	0.02109	0.02629
	60	0.07214	0.01618	0.01384	0.01743
	80	0.05371	0.01210	0.01030	0.01304
	100	0.04278	0.00967	0.00820	0.01041
3	5	0.91844	0.20162	0.17160	0.21572
	10	0.36945	0.09276	0.06796	0.09996
	15	0.23016	0.06038	0.04212	0.06522
	20	0.16697	0.04478	0.03048	0.04842
	30	0.10771	0.02953	0.01961	0.03196
	40	0.07948	0.02203	0.01445	0.02386
	60	0.05213	0.01461	0.00950	0.01580
	80	0.03878	0.01093	0.00704	0.01184
	100	0.03088	0.00873	0.00560	0.00946
4	5	0.80991	0.19328	0.14754	0.20701
	10	0.30758	0.08809	0.05455	0.09523
	15	0.18913	0.05730	0.03329	0.06213
	20	0.13640	0.04249	0.02393	0.04613
	30	0.08751	0.02802	0.01530	0.03046
	40	0.06441	0.02090	0.01124	0.02274
	60	0.04214	0.01386	0.00734	0.01509
	80	0.03131	0.01037	0.00545	0.01129
	100	0.02491	0.00828	0.00433	0.00902
5	5	0.77759	0.19089	0.14033	0.20448
	10	0.27147	0.08546	0.04681	0.09190
	15	0.16431	0.05550	0.02800	0.05785
	20	0.11771	0.04114	0.01994	0.04164
	30	0.07507	0.02712	0.01264	0.02639
	40	0.05509	0.02023	0.00925	0.01922
	60	0.03595	0.01341	0.00602	0.01241
	80	0.02668	0.01004	0.00446	0.00915
	100	0.02121	0.00802	0.00354	0.00724

Table 7: Continued

m	n	$\theta = 1.50$		$\theta = 2$	
		V_3	V_4	V_3	V_4
2	5	0.08964	0.25245	0.04524	0.26157
	10	0.03671	0.11692	0.01848	0.12147
	15	0.02293	0.07617	0.01153	0.07920
	20	0.01665	0.05650	0.00837	0.05877
	30	0.01075	0.03726	0.00540	0.03878
	40	0.00793	0.02780	0.00398	0.02893
	60	0.00520	0.01843	0.00261	0.01919
	80	0.00387	0.01379	0.00194	0.01436
	100	0.00308	0.01101	0.00155	0.01147
3	5	0.06332	0.22657	0.03135	0.23449
	10	0.02487	0.10551	0.01227	0.10961
	15	0.01537	0.06895	0.00758	0.07171
	20	0.01111	0.05122	0.00547	0.05331
	30	0.00714	0.03384	0.00352	0.03524
	40	0.00526	0.02527	0.00259	0.02632
	60	0.00344	0.01677	0.00169	0.01750
	80	0.00256	0.01255	0.00126	0.01308
	100	0.00204	0.01003	0.00100	0.01045
4	5	0.05364	0.21745	0.02633	0.22499
	10	0.01954	0.10067	0.00954	0.10459
	15	0.01187	0.06581	0.00580	0.06848
	20	0.00851	0.04892	0.00417	0.05094
	30	0.00543	0.03233	0.00266	0.03370
	40	0.00398	0.02415	0.00196	0.02518
	60	0.00260	0.01604	0.00128	0.01673
	80	0.00193	0.01200	0.00095	0.01252
	100	0.00153	0.00959	0.00075	0.01001
5	5	0.05077	0.21479	0.02486	0.22223
	10	0.01656	0.09800	0.00799	0.10178
	15	0.00989	0.06406	0.00470	0.06648
	20	0.00705	0.04764	0.00331	0.04938
	30	0.00448	0.03151	0.00207	0.03261
	40	0.00329	0.02355	0.00150	0.02435
	60	0.00214	0.01564	0.00097	0.01616
	80	0.00159	0.01171	0.00071	0.01209
	100	0.00126	0.00936	0.00057	0.00966

4. Conclusions

The peculiarity of this estimation method is that, if we have the best linear unbiased estimator based on observation of size m as the kernel, then the evaluation of moments of order statistics of sample sizes up to $2m-1$ coming from standard form of (1.2) alone necessary to obtain the variances of the U-statistics defined in (3.5) and (3.6).
For example, in the case of $m=5$, by using the best linear unbiased estimators of μ and σ given in (2.1) and (2.2) respectively , one only needs the moments of order statistics arising from the standard of (1.2) for sample sizes up to 9 to obtain the U-statistic estimators for μ and σ and its variances for any sample of size n and for any given value of θ. Using the values of variances and co-variances of order statistics (given in Table 2) and the coefficients of, BLUEs of μ and σ (given in Table 3 and Table 4), we have obtained the values of $\xi_{\omega}^{(m)}$and $\psi_{\omega}^{(m)}$for $\omega=1,2,\cdots,m-1$, $m=2(1)5$ and $\theta=(0.50)(0.50)2$ (given in Table 5 and Table 6). Also, we have evaluated the variances of the U-statistic estimators for μ and σ which are given in Table 7. For practicing statisticians the results derived in the paper will be helpful, when they look for estimators of parameters of Lindley distribution using ordered random variables.

Acknowledgements

The authors express their gratefulness for the constructive criticism of the learned referees which helped to improve considerably the revised version of the paper.

REFERENCES

ALI, S., ASLAM, M., KAZMI, S. M., (2013). A study of the effect of the loss function on Bayes estimate, posterior risk and hazard function for Lindley distribution. Appl Math Model, 37, pp. 6068–6078.

ELBATAL, I., ELGARHY, M., (2013). Transmuted quasi Lindley distribution: a generalization of the quasi Lindley distribution. Int J Pure Appl Sci Technol, 18, pp. 59–70.

GHITANY, M, E., ATIEH, B., Nadarajah, S., (2008). Lindley distribution and its applications. Mathematics and Computers in Simulation, 78 (4), pp. 493–506.

GHITANY, M, E., Al-MUTAIRI, D, K., BALAKRISHNAN, N., Al-ENEZI, L, J., (2013): Power Lindley distribution and associated inference.Computational Statistics and Data Analysis, 64, pp. 20–33.

GÓMEZ, D, E., OJEDA, E, C., (2011). The discrete Lindley distribution: Properties and applications, Journal of Statistical Computation and Simulation, 81 (11), pp. 1405–1416.

HOEFFDING, W., (1948). A class of statistics with asymptotically normal distributions, The Annals of Mathematical Statistics, 19, pp. 293–325.

IRSHAD, M, R., MAYA, R., (2017). Extended Version of Generalized Lindley Distribution. South African Statistical Journal, 51, pp. 19–44.

KADILAR, G, O., CAKMAKYAPAN, S., (2016). The Lindley family of distributions: Properties and applications, Hacettepe University Bulletin of Natural Sciences and Engineering Series B: Mathematics and Statistics, 46(116), pp. 1–28.

LINDLEY, D, V., (1958). Fiducial distributions and Bayes' theorem, Journal of the Royal Statistical Society, Series B, 20 (1), pp. 102–107.

LOYD, E, H., (1952). Least-squares estimation of location and scale parameters using order statistics. Biometrical Journal, 39, pp. 88–95.

NADARAJAH, S., BAKOUSH, H., TAHMASBI, R., (2011). A generalized lindley distribution, Sankhya B - Applied and Interdisciplinary Statistics, 73, pp. 331–359.

NEDJAR, S., ZEHDOUDI, H., (2016). On gamma Lindley distribution: Properties and simulations, Journal of Computational and Applied Mathematics, 298, pp. 167–174.

SANKARAN, M., (1970). The discrete Poisson-Lindley distribution, Biometrics, 26, pp. 145–149.

SHIBU, D, S., IRSHAD, M, R., (2016). Extended New Generalized Lindley Distribution. Statistica, 76, pp. 42–56.

SULTAN, K, S., Al-THUBYANI, W, S., (2016). Higher order moments of order statistics from the Lindley distribution and associated inference, Journal of Statistical computation and Simulation, 86, pp. 3432–3445.

THOMAS, P, Y., SREEKUMAR, N, V., (2004). Estimation of the scale parameter of generalized exponential distribution using order statistics, Calcutta Statistical Association Bulletin, 55, pp. 199–208.

THOMAS, P, Y., SREEKUMAR, N, V., (2008). Estimation of location and scale parameters of a distribution by U-statistics based on best linear functions of order statistics, Journal of Statistical Planning and Inference, 138, pp. 2190–2200.

ZAKERZADEH, H., DOLATI, A., (2009). Generalized Lindley distribution, Journal of Mathematical Extension, 3 (2), pp. 13–25.

APPENDIX

Here, we use MATHCAD software for all numerical computations. Tables 1 and 2 summarize the values of means, variances and covariances of the order statistics arising from the standard form of (1.2) for $n = 2(1)10$ *and for different values of* θ*. For calculating the U-statistic estimators and its variances, we want the coefficients of* $\boldsymbol{X}$ *in the best linear unbiased estimators of* μ *and* σ *and the corresponding variances. Using the moments of order statistics given in Tables 1 and 2 and using the formulas defined in (2.1) and (2.3), we have evaluated the coefficients of* $\boldsymbol{X}$ *in the BLUE of* μ *and its variances for different values of* θ *and it is given in Table 3. The moments of order statistics given in Tables 1 and 2 and using the formulas defined in (2.2) and (2.4), we have evaluated the coefficients of* $\boldsymbol{X}$ *in the BLUE of* σ *and its variances for different values of* θ *which are given in Table 4.*

For computing the numerical values of variances of the U-statistic estimators defined in (3.5) and (3.6), first we want the numerical values of $\xi_{\omega}^{(m)}$ *and* $\psi_{\omega}^{(m)}$ *for different values of* m *and* ω*. For example if we want to calculate the numerical values of* $\xi_{\omega}^{(m)}$ *we use the formula* $w_k = \binom{m+k}{m}(b'_{m+k}V_{m+k}b_{m+k})\sigma^2 - \xi_m^{(m)}, k = 1,2,\cdots,m-1$*. In particular, if for the case of* $m = 2$*,* $\omega = 2$ *and* $\theta = 0.50$*, we want to calculate only two values* $\xi_1^{(2)}$ *and* $\xi_2^{(2)}$*, where* $\xi_2^{(2)}$ *is nothing but the value of the variance. From Table 2, it is obtained as 5.77647. Using the formula* w_k*, the value of* $\xi_1^{(2)}$ *reduces to* $\xi_1^{(2)} = \frac{3}{2}U - \frac{1}{2}\xi_2^{(2)}$*, where* $U = bVb'$*,* $b = \left(\frac{2a_1}{3}, \frac{a_1}{3} + \frac{a_2}{3}, \frac{2a_2}{3}\right)$*,* a_1 *and* a_2 *are coefficients of* $\boldsymbol{X}$ *in BLUE of* μ *of order statistics of sample of size 2 and* V *is the variance covariance matrix of order 3. From Table 2, the matrix* V *is obtained as*

$$V = \begin{bmatrix} 1.26895 & 1.10362 & 0.97997 \\ 1.10362 & 3.03236 & 2.71450 \\ 0.97997 & 2.71450 & 8.76918 \end{bmatrix}.$$

Also from Table 3, the value of a_1 *is 1.65385 and that of* a_2 *is -0.65385. Using these values, we can easily obtain the value of* $\xi_1^{(2)}$*. In the same way, we can easily obtain the values of* $\xi_3^{(\omega)}$*,* $\xi_4^{(\omega)}$ *and* $\xi_5^{(\omega)}$ *for various values of* ω *and we have evaluated all these values which are given in Table 5. The Table 6 comprises the values of* $\psi_{\omega}^{(m)}$ *for different combinations of* m *and* ω*. The values of* $\psi_{\omega}^{(m)}$ *are obtained when we follow the same steps for obtaining the values of* $\xi_{\omega}^{(m)}$*, the only change is that instead of using the coefficients of* $\boldsymbol{X}$ *in the best linear unbiased estimator of* μ *and its variance, here we use the coefficients of* $\boldsymbol{X}$ *in the best linear unbiased estimator of* σ *and its variance. The numerical values of the variances of U-statistic estimators defined in, (3.5) and (3.6) are given in Table 7 for various values of parameters. If we put* $m = 2$ *the formula (3.5) reduces the following way for various values of* n*. That is*

$Var[U_{1:5}^{(2)}] = \frac{6\xi_1^{(2)}+\xi_2^{(2)}}{10}$, $Var[U_{1:10}^{(2)}] = \frac{16\xi_1^{(2)}+\xi_2^{(2)}}{45}$, $Var[U_{1:15}^{(2)}] = \frac{26\xi_1^{(2)}+\xi_2^{(2)}}{105}$,
$Var[U_{1:20}^{(2)}] = \frac{36\xi_1^{(2)}+\xi_2^{(2)}}{190}$, $Var[U_{1:30}^{(2)}] = \frac{56\xi_1^{(2)}+\xi_2^{(2)}}{435}$, $Var[U_{1:40}^{(2)}] = \frac{76\xi_1^{(2)}+\xi_2^{(2)}}{780}$,

$Var[U^{(2)}_{1:60}] = \frac{116\xi_1^{(2)}+\xi_2^{(2)}}{1770}$, $Var[U^{(2)}_{1:80}] = \frac{156\xi_1^{(2)}+\xi_2^{(2)}}{3160}$ and $Var[U^{(2)}_{1:100}] = \frac{196\xi_1^{(2)}+\xi_2^{(2)}}{4950}$.

Similarly we can find the values of $Var[U^{(3)}_{1:n}]$, $Var[U^{(4)}_{1:n}]$ *and* $Var[U^{(5)}_{1:n}]$ *for various values of* n. *Proceeding in the similar manner we can easily find the values of* $Var[U^{(m)}_{2:n}]$ *for different values of* m *and* n.

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 621–643, DOI 10.21307/stattrans-2018-033

EXTENDED EXPONENTIATED POWER LINDLEY DISTRIBUTION

V. Ranjbar[1], **M. Alizadeh**[2], **G. G. Hamedani**[3]

ABSTRACT

In this study, we introduce a new model called the Extended Exponentiated Power Lindley distribution which extends the Lindley distribution and has increasing, bathtub and upside down shapes for the hazard rate function. It also includes the power Lindley distribution as a special case. Several statistical properties of the distribution are explored, such as the density, hazard rate, survival, quantile functions, and moments. Estimation using the maximum likelihood method and inference on a random sample from this distribution are investigated. A simulation study is performed to compare the performance of the different parameter estimates in terms of bias and mean square error. We apply a real data set to illustrate the applicability of the new model. Empirical findings show that proposed model provides better fits than other well-known extensions of Lindley distributions.

Key words: Power Lindley distribution, Structural properties, Failure-time, Maximum likelihood estimation.

1. Introduction

The statistical analysis and modeling of lifetime data are essential in almost all applied sciences such as, biomedical science, engineering, nance, and insurance, amongst others. A number of one-parameter continuous distributions for modelling lifetime data has been introduced in statistical literature including exponential, Lindley, gamma, lognormal, and Weibull. The exponential, Lindley and Weibull distributions are more popular than the gamma and lognormal distributions because the survival functions of the gamma and the lognormal distributions cannot be expressed in closed forms and both require numerical integration. The Lindley distribution is a very well-known distribution that has been extensively used over the past decades for modeling data in reliability, biology, insurance, and lifetime analysis. It was introduced by Lindley (1985) to analyze failure time data, especially in applications of modeling stress-strength reliability. The motivation for introducing the Lindley distribution arises from its ability to model failure time data with increasing, decreasing, unimodal and bathtub shaped hazard rates. It may also be mentioned that the Lindley distribution belongs to an exponential family and it can be written as a mixture of an exponential and a gamma distributions. This distribution represents a good alternative to the exponential failure time distributions that suffer from not exhibiting unimodal and bathtub shaped failure rates (Bakouch et al. (2012)). The properties and inferential procedure for the Lindley distribution were studied by

[1]Golestan University, Gorgan, Iran. E-mail: vahidranjbar@gmail.com
[2]Persian Gulf University, Bushehr, Iran. E-mail: moradalizadeh78@gmail.com
[3]Marquette University, Milwaukee, USA. E-mail: g.hamedani@mu.edu

Ghitany et al. (2008, 2013). They show via a numerical example that the Lindley distribution gives better modeling than the one based on the exponential distribution when hazard rate is unimodal or bathtub shaped. Furthermore, Mazucheli and Achcar (2011) showed that many of the mathematical properties are more exible than those of the exponential distribution and proposed the Lindley distribution as a possible alternative to exponential or Weibull distributions. The need for extended forms of the Lindley distribution arises in many applied areas. The emergence of such distributions in the statistics literature is quite recent. For some extended forms of the Lindley distribution and their applications, the interested reader is referred to Kumaraswamy Lindley (Cakmakyapan and Ozel, (2014)), beta odd log-logistic Lindley (Cordeiro et al., (2015)), generalized Lindley (Nadarajah et al., (2011)), quasi Lindley distributions (Shanker and Mishra, (2013)).

The probability density function (pdf) and cumulative distribution function (cdf) of the power Lindley distribution are given respectively by

$$f(x) = \frac{\lambda^2}{1+\lambda}\beta x^{\beta-1} e^{-\lambda x^{\beta}},$$
$$F(x) = 1-(1+\frac{\lambda\, x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\, \mathrm{x}^{\beta}}. \tag{1}$$

It can be seen that this distribution is a mixture of Exponential and gamma distributions. Having only one parameter, the Lindley distribution does not provide enough exibility for analyzing different types of lifetime data. To increase the exibility for modeling purposes it will be useful to consider further alternatives to this distribution. Our purpose here is to provide a generalization that may be useful for more complex situations. Once the proposed distribution is quite exible in terms of pdf and hazard rate function (hrf), it may provide an interesting alternative for describing income distributions and can also be applied in actuarial science, nance, bioscience, telecommunications and modeling lifetime data. Therefore, goal is to introduce a new distribution using the Lindley distribution. Alizadeh et.al (2017), introduced a new class of exponentiated distributions which called Extended Exponentiated distribution (EE-G). The cdf and pdf of this family are given by

$$F(x;\alpha,\gamma,\xi) = \int_0^{\frac{G(x;\xi)^{\alpha}}{1-G(x;\xi)^{\gamma}}} \frac{dt}{(1+t)^2}\, dt = \frac{G(x;\xi)^{\alpha}}{G(x;\xi)^{\alpha}+1-G(x;\xi)^{\gamma}} \tag{2}$$

$$f(x;\alpha,\gamma,\xi) = \frac{g(x;\xi)\, G(x;\xi)^{\alpha-1}\, [\alpha+(\gamma-\alpha)\, G(x;\xi)^{\gamma}]}{[G(x;\xi)^{\alpha}+1-G(x;\xi)^{\gamma}]^2}, \tag{3}$$

where $\alpha,\gamma>0$ are two shape parameters and ξ is the vector of parameters for baseline cdf G. For $\alpha=\gamma$, it contains exp-G family of distributions. Taking $G(x;\xi)$ as power Lindley distribution with parameters λ,β, we introduce a new extension of

Exponentiated power Lindley distribution.

The article is outlined as follows: In Section 2, we introduce the EE-PL distribution and provide plots of the density and hazard rate functions. Shapes, quantile function, moments, and moment generating function are also obtained. Moreover, mean deviation, Lorenz and Bonferroni curves, order statistics and finally a simulation study are presented in this section. In section 3, the asymptotic properties and extreme values are obtained. Estimation by the method of maximum likelihood and an explicit expression for the observed information matrix are presented in Section 4. The characterizations of EE-PL distribution are presented in Section 5. The Applications to real data sets are considered in Section 6. Finally, Section 7 offers some concluding remarks.

2. Main properties

2.1. Probability Density and Cumulative Distribution Functions

Inserting (1) in (2), the cdf of the EE-PL with four parameters $(\alpha,\beta,\gamma,\lambda>0)$ is defined by

$$F(x;\alpha,\beta,\gamma,\lambda)=\frac{\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x}^{\beta}}\right]^{\alpha}}{\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x}^{\beta}}\right]^{\alpha}+1-\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x}^{\beta}}\right]^{\gamma}},\quad x\geq 0 \qquad (4)$$

The corresponding pdf for $x>0$ is given by

$$\begin{aligned} f(x;\alpha,\beta,\gamma,\lambda) &= \lambda^{2}\beta\,x^{\beta-1}(1+x^{\beta})\mathrm{e}^{-\lambda\,\mathrm{x}^{\beta}}\left[1-(1+\frac{\lambda\,\mathrm{x}^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x}^{\beta}}\right]^{\alpha-1} \\ &\quad\times\frac{\left\{\alpha+(\gamma-\alpha)\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x}^{\beta}}\right]^{\gamma}\right\}}{(1+\lambda)\left\{\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x}^{\beta}}\right]^{\alpha}+1-\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x}^{\beta}}\right]^{\gamma}\right\}^{2}}, \qquad (5) \end{aligned}$$

where λ is a scale parameter α,β and γ are the shape parameters. Here, α and β govern the skewness of (5). A random variable X with the pdf (5) is denoted by $X\sim EE-PL(\alpha,\beta,\gamma,\lambda)$. It is easy to see that:

- For $\beta=1$, we obtain Extended Generalized Lindley by Ranjbar et al. (2018).
- For $\alpha=\gamma$, we obtain Exponentiated power Lindley.
- For $\alpha=\gamma$ and $\beta=1$, we obtain Generalized Lindley.
- For $\alpha=\gamma=1$, we obtain Power Lindley.
- For $\alpha=\gamma=\beta=1$, we obtain Lindley.

Some of the possible shapes of the density function (5) for the selected parameter values are illustrated in Figure 1. As seen in Figure 1, the density function can take various forms depending on the parameter values. It is evident that the EE-LP distribution is much more flexible than the power Lindley distribution, i.e. the additional shape parameter allows

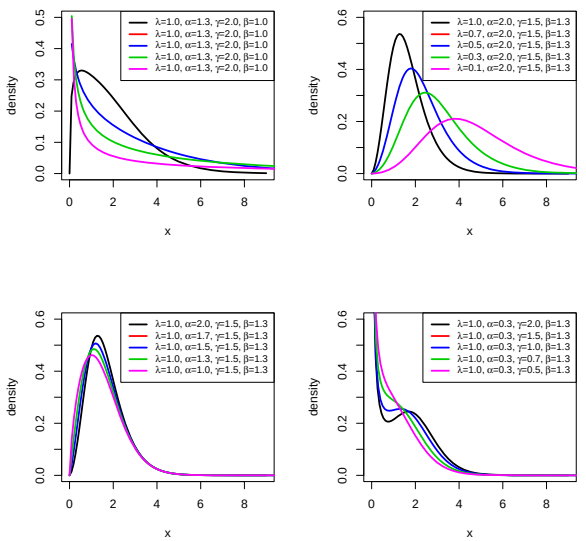


Figure 1: Plots of Pdf of the EE-PL model for selected λ, α, γ and β.

for a high degree of flexibility of the EE-PL distribution. Both unimodal and monotonically decreasing and increasing shapes appear to be possible.

2.2. Survival and Hazard Rate Functions

Central role is played in the reliability theory by the quotient of the pdf and survival function. We obtain the survival function corresponding to (4) as

$$S(x;\lambda,\alpha,\gamma,\beta) = \frac{1-\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x}^{\beta}}\right]^{\gamma}}{\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x}^{\beta}}\right]^{\alpha}+1-\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x}^{\beta}}\right]^{\gamma}}. \tag{6}$$

In reliability studies, the hrf is an important characteristic and fundamental to the design of safe systems in a wide variety of applications. Therefore, we discuss these properties of

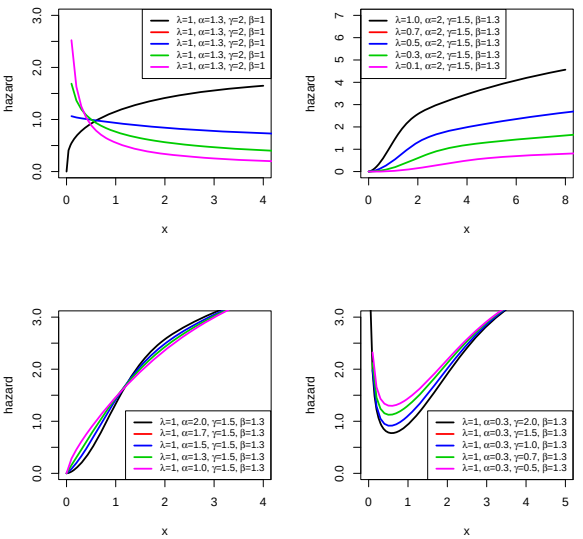


Figure 2: Hazard rate functions of the EE-PL model for selected λ, α, γ and β.

the EE-LP distribution. The hrf of X takes the form

$$
\begin{aligned}
h(x;\lambda,\alpha,\gamma,\beta) \;=\; & \lambda^2\beta x^{\beta-1}(1+x^{\beta})\mathrm{e}^{-\lambda x^{\beta}}\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda x^{\beta}}\right]^{\alpha-1} \\
& \times\left\{\alpha+(\gamma-\alpha)\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda x^{\beta}}\right]^{\gamma}\right\}/ \\
& \left[\left\{\left(1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda x^{\beta}}\right)^{\alpha}+1-\left(1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda x^{\beta}}\right)^{\gamma}\right\}\right. \\
& \left.\times\left\{1-\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda x^{\beta}}\right]^{\gamma}\right\}\right], \quad x>0. \qquad (7)
\end{aligned}
$$

Plots of the hrf of the EE-PL distribution for several parameter values are displayed in Figure 2. Figure 2 shows that the hrf of the EE-PL distribution can have very flexible shapes, such as increasing, decreasing, bathtub followed by upside down bathtub, and bathtub shapes for the selected values of the model parameters. This attractive flexibility makes the hrf of the EE-PL distribution useful and suitable for non-monotone empirical hazard behaviors which are more likely to be encountered or observed in real life situations.

2.3. Mixture representations for the pdf and cdf

In this subsection, we provide alternative mixture representations for the pdf and cdf of X. Some useful expansions for (4) can be derived by using the concept of power series. We

have

$$\begin{aligned}
[1-(1+\frac{\lambda}{1+\lambda}x^{\beta})e^{-\lambda x^{\beta}}]^{\alpha} &= \sum_{i=1}^{\infty}(-1)^{i}\binom{\alpha}{i}[1+\frac{\lambda}{1+\lambda}x^{\beta})e^{-\lambda x^{\beta}}]^{i} \\
&= \sum_{i=1}^{\infty}\sum_{k=0}^{i}(-1)^{i+k}\binom{\alpha}{i}\binom{i}{k}[1-(1+\frac{\lambda}{1+\lambda}x^{\beta})e^{-\lambda x^{\beta}}]^{k} \\
&= \sum_{k=0}^{\infty}\sum_{i=k}^{\infty}(-1)^{i+k}\binom{\alpha}{i}\binom{i}{k}[1-(1+\frac{\lambda}{1+\lambda}x^{\beta})e^{-\lambda x^{\beta}}]^{k} \\
&= \sum_{k=0}^{\infty}a_{k}[1-(1+\frac{\lambda}{1+\lambda}x^{\beta})e^{-\lambda x^{\beta}}]^{k},
\end{aligned}$$

where $a_k = a_k(\alpha) = \sum_{i=k}^{\infty}(-1)^{i+k}\binom{\alpha}{i}\binom{i}{k}$. Also

$$[1-(1+\frac{\lambda}{1+\lambda}x^{\beta})e^{-\lambda x^{\beta}}]^{\alpha}+1-[1-(1+\frac{\lambda}{1+\lambda}x^{\beta})e^{-\lambda x^{\beta}}]^{\gamma}=\sum_{k=0}^{\infty}b_{k}[1-(1+\frac{\lambda}{1+\lambda}x^{\beta})e^{-\lambda x^{\beta}}]^{k},$$

where $b_0 = a_0(\alpha)+1-a_0(\gamma)$ and $b_k = a_k(\alpha)-a_k(\gamma)$ for $k \geq 1$. Then using the ratio of two power series, we can write

$$\begin{aligned}
F(x) &= \frac{\sum_{k=0}^{\infty}a_{k}[1-(1+\frac{\lambda}{1+\lambda}x^{\beta})e^{-\lambda x^{\beta}}]^{k}}{\sum_{k=0}^{\infty}b_{k}[1-(1+\frac{\lambda}{1+\lambda}x^{\beta})e^{-\lambda x^{\beta}}]^{k}} \\
&= \sum_{k=0}^{\infty}c_{k}[1-(1+\frac{\lambda}{1+\lambda}x^{\beta})e^{-\lambda x^{\beta}}]^{k}, \qquad (8)
\end{aligned}$$

where $c_0 = \frac{a_0}{b_0}$ and for $k \geq 1$,

$$c_{k}=\frac{1}{b_{0}}[a_{k}-\frac{1}{b_{0}}\sum_{r=1}^{k}b_{r}c_{k-r}]. \qquad (9)$$

Equation (8) shows that we can write the cdf of EE-PL as a Linear combination of generalized lindly distribution. Then we can write

$$f(x) = \sum_{k=0}^{\infty}c_{k+1}\frac{(k+1)\lambda^{2}\beta x^{\beta-1}(1+x^{\beta})}{1+\lambda}e^{-\lambda x^{\beta}}[1-(1+\frac{\lambda}{1+\lambda}x^{\beta})e^{-\lambda x^{\beta}}]^{k}.$$

2.4. Moments and Moment Generating Function

Some of the most important features and characteristics of a distribution can be studied through moments (e.g. tendency, dispersion, skewness and kurtosis). Now we obtain ordinary moments and the moment generating function (mgf) of the EE-PL distribution. We define and compute

$$A(a_1,a_2,a_3,a_4;\lambda,\beta)=\int_{0}^{\infty}x^{a_1}(1+x^{\beta})^{a_2}\,\mathrm{e}^{-a_3x^{\beta}}\left[1-(1+\frac{\lambda}{1+\lambda}x^{\beta})\mathrm{e}^{-\lambda x^{\beta}}\right]^{a_4}\mathrm{d}x. \qquad (10)$$

Using generalized binomial expansion, one can obtain

$$A(a_1,a_2,a_3,a_4;\lambda,\beta)=\sum_{l,r=0}^{\infty}\sum_{k=0}^{l}(-1)^l\binom{a_4}{l}\binom{l}{k}\binom{a_2}{r}(\frac{\lambda}{1+\lambda})^l\times\frac{\Gamma(\frac{a_1+1}{\beta}+k+r)}{\beta\,(\lambda\,l+a_3)^{\frac{a_1+1}{\beta}+k+r}}. \tag{11}$$

Next, the nth moment of the EE-PL distribution is given by

$$E\left[X^n\right]=\frac{\lambda^2\beta}{1+\lambda}\sum_{k=0}^{\infty}k\,c_k\,A(n+\beta-1,1,\lambda,k;\lambda,\beta). \tag{12}$$

For integer values of k, let $\mu_k'=E(X^k)$ and $\mu=\mu_1'=E(X)$, then one can also find the kth central moment of the EE-PL distribution through the following well-known equation

$$\mu_k=E(X-\mu)^k=\sum_{r=0}^{k}\binom{k}{r}\mu_r'(-\mu)^{k-r}. \tag{13}$$

The moment generating function of a random variable provides the basis of an alternative route to analytical results compared with working directly with its pdf and cdf. Using (12) and (13), we obtain

$$M_X(t)=E\left[\mathrm{e}^{\mathrm{t}X}\right]=\frac{\lambda^2}{1+\lambda}\sum_{k=0}^{\infty}(k+1)\,c_{k+1}A(k+1,\lambda,0,\lambda-t).$$

Using (13), the variance, skewness and kurtosis measures can be obtained. Skewness measures the degree of the long tail and kurtosis is a measure of the degree of tail heaviness. For the EE-PL distribution, The skewness can be computed as

$$S=\frac{\mu_3}{\mu_2^{3/2}}=\frac{\mu_3'-3\mu_2'\mu_1'+2\mu_1'^3}{(\mu_2'-\mu_1'^2)^{3/2}}$$

and the kurtosis is based on octiles as

$$K=\frac{\mu_4}{\mu_2^2}=\frac{\mu_4'-4\mu_1'\mu_3'+6\mu_1'^2\mu_2'-3\mu_1'^4}{\mu_2'-\mu_1'^2}.$$

When the distribution is symmetric $S=0$, and when the distribution is right (or left) skewed $S>0(or\ S<0)$. As K increases, the tail of the distribution becomes heavier. These measures are less sensitive to outliers and they exist even for distributions without moments.
We present first four ordinary moments, skewness and kurtosis of the EE-PL distribution for various values of the parameters in Table 1. Plots for skewness and kurtosis, when $\lambda=2$, are presented in Figure 3.

Next, we define and compute

$$B(a_1,a_2,a_3,a_4;y,\lambda,\beta)=\int_0^y x^{a_1}\,(1+x^\beta)^{a_2}\,\mathrm{e}^{-a_3x^\beta}\left[1-(1+\frac{\lambda}{1+\lambda}\,x^\beta)\mathrm{e}^{-\lambda x^\beta}\right]^{a_4}\mathrm{d}x. \tag{14}$$

Table 1: Moments, skewness, and kurtosis of the EE-PL dist. for the some parameter values.

λ	α	β	γ	μ_1'	μ_2'	μ_3'	μ_4'	*Skewness*	*Kurtosis*
2.0	0.5	0.5	0.5	0.457	1.788	16.97	289.197	7.4082	164.75
2.0	0.5	0.5	1.0	0.413	0.457	0.783	1.7888	2.3382	3.052
2.0	0.5	0.5	3.0	0.608	0.468	0.413	0.4004	0.2704	0.243
2.0	0.5	1.0	0.5	0.752	3.384	33.41	575.819	5.6285	172.40
2.0	0.5	1.0	1.0	0.571	0.752	1.413	3.3844	1.7824	3.070
2.0	0.5	1.0	3.0	0.688	0.592	0.571	0.5947	0.0234	0.261
2.0	0.5	2.0	0.5	1.176	6.223	65.03	1142.245	4.3521	182.28
2.0	0.5	2.0	1.0	0.750	1.176	2.445	6.2232	1.3397	3.107
2.0	0.5	2.0	3.0	0.757	0.716	0.750	0.8367	-0.1566	0.288
2.0	2.0	0.5	0.5	0.655	2.028	17.57	291.944	7.0008	156.70
2.0	2.0	0.5	1.0	0.646	0.655	0.981	2.0282	2.1655	2.572
2.0	2.0	0.5	3.0	0.818	0.708	0.646	0.619	0.4272	0.134
2.0	2.0	1.0	0.5	0.985	3.744	34.46	580.961	5.4788	167.33
2.0	2.0	1.0	1.0	0.806	0.985	1.678	3.7447	1.7687	2.701
2.0	2.0	1.0	3.0	0.882	0.822	0.806	0.8272	0.2475	0.140
2.0	2.0	2.0	0.5	1.437	6.729	66.76	1151.513	4.3391	179.76
2.0	2.0	2.0	1.0	0.986	1.437	2.781	6.7291	1.4076	2.814
2.0	2.0	2.0	3.0	0.943	0.941	0.986	1.0779	0.0508	0.150

From the generalized binomial expansion, we have

$$B(a_1,a_2,a_3,a_4;a,\lambda,\beta)$$
$$= \sum_{l,r=0}^{\infty}\sum_{k=0}^{l}(-1)^l\binom{a_4}{l}\binom{l}{k}\binom{a_2}{r}\left(\frac{\lambda}{1+\lambda}\right)^l \times \frac{\gamma\left(\frac{a_1+1}{\beta}+k+r,\frac{y^{\frac{1}{\beta}}}{\lambda\, l+a_3}\right)}{\beta\,(\lambda\, l+a_3)^{\frac{a_1+1}{\beta}+k+r}}, \tag{15}$$

where $\gamma(\lambda,z)=\int_0^z t^{\lambda-1}\,\mathrm{e}^{-t}\,dt$ denotes the incomplete gamma function. Now, the nth incomplete moment of the EE-PL distribution is found to be

$$m_n(y)=E\left[X^n\,|X<y\right]=\frac{\lambda^2\beta}{1+\lambda}\sum_{k=0}^{\infty}(k+1)\,c_{k+1}\,B(n+\beta-1,1,\lambda,k,y;\lambda,\beta). \tag{16}$$

2.5. Mean Deviations, Lorenz and Bonferroni Curves

Mean deviation about the mean and mean deviation about the median as well as Lorenz and Bonferroni curves for the EE-PL distribution are presented in this section. Bonferroni and Lorenz curves are widely used tool for analyzing and visualizing income inequality. Lorenz curve, L(p) can be regarded as the proportion of total income volume accumulated by those units with income lower than or equal to the volume y, and Bonferroni curve, B(p) is the scaled conditional mean curve, that is, ratio of group mean income of the population.

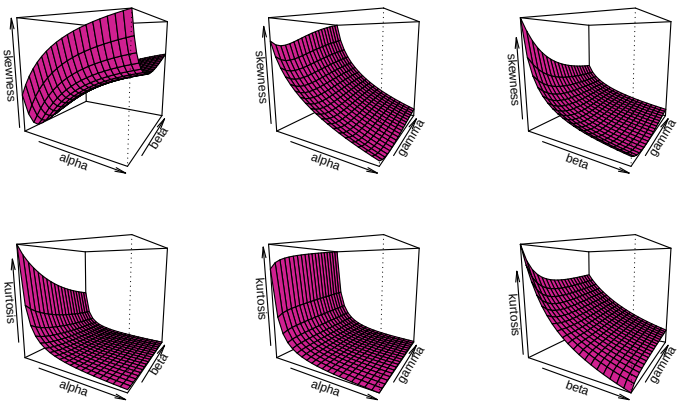


Figure 3: Values of skewness and kurtosis of EE-PL for some values of α, β and γ.

2.5.1 Mean deviations

The amount of scatter in a population may be measured to some extent by deviations from the mean and median. These are known as the mean deviation about the mean and the mean deviation about the median, defined by

$$\delta_1(X) = \int_0^\infty |x-\mu| f(x)\,\mathrm{d}x,$$

and

$$\delta_2(X) = \int_0^\infty |x-M| f(x)\,\mathrm{d}x.$$

respectively, where $\mu = E(X)$ and $M = \text{Median}(X) = Q(0.5)$ denotes the median and $Q(p)$ is the quantile function. The measures $\delta_1(X)$ and $\delta_2(X)$ can be calculated using the relationships

$$\delta_1(X) = 2\mu F(\mu) - 2\int_0^\mu x f(x)\,\mathrm{d}x$$

and

$$\delta_2(X) = \mu - 2\int_0^M x f(x)\,\mathrm{d}x$$

Finally have

$$\delta_1(X) = 2\mu F(\mu) - \frac{\beta\lambda^2}{1+\lambda}\sum_{k=0}^{\infty}(k+1)c_{k+1}A(\beta,1,\lambda,k;\lambda,\beta),$$

and

$$\delta_2(X) = \mu - \frac{2\beta\lambda^2}{1+\lambda}\sum_{k=0}^{\infty}(k+1)c_{k+1}B(\beta,1,\lambda,k;M,\lambda,\beta).$$

2.5.2 Bonferroni and Lorenz curves

The Bonferroni and Lorenz curves have applications in economics as well as other fields like reliability, medicine and insurance. Let $X \sim EE-PL(\lambda,\beta,\alpha,\gamma)$ and $F(x)$ be the cdf of X, then the Bonferroni curve of the EE-PL distribution is given by

$$B(F(x)) = \frac{1}{\mu\, F(x)} \int_0^x t f(t) \mathrm{d}t,$$

where $\mu = E(X)$. Therefore, from (15), we have

$$B(F(x)) = \frac{1}{\mu\, F(x)} \times \frac{\beta\,\lambda^2}{1+\lambda} \sum_{k=0}^{\infty} (k+1) c_{k+1}\, B(\beta,1,\lambda,k;x,\lambda,\beta).$$

The Lorenz curve of the EE-PL distribution can be obtained using the relation

$$L(F(x)) = F(x)B(F(x)) = \frac{1}{\mu} \times \frac{\beta\,\lambda^2}{1+\lambda} \sum_{k=0}^{\infty} (k+1) c_{k+1}\, B(\beta,1,\lambda,k;x,\lambda,\beta).$$

2.6. Order statistics

Order statistics make their appearance in many areas of statistical theory and practice. Suppose $X_1,\ldots,X_n$ is a random sample from any EE-PL distribution. Let $X_{i:n}$ denote the ith order statistic. The pdf of $X_{i:n}$ can be expressed as

$$f_{i:n}(x) = K f(x) F^{i-1}(x)\, \{1-F(x)\}^{n-i} = K \sum_{j=0}^{n-i} (-1)^j \binom{n-i}{j} f(x)\, F(x)^{j+i-1},$$

where $K = 1/B(i,n-i+1)$. We use the result of Gradshteyn and Ryzhik for a power series raised to a positive integer n (for $n \geq 1$)

$$\left(\sum_{i=0}^{\infty} a_i\, u^i \right)^n = \sum_{i=0}^{\infty} d_{n,i}\, u^i, \tag{17}$$

where the coefficients $d_{n,i}$ (for $i = 1,2,\ldots$) are determined from the recurrence equation (with $d_{n,0} = a_0^n$)

$$d_{n,i} = (i\, a_0)^{-1} \sum_{m=1}^{i} [m(n+1)-i]\, a_m\, d_{n,i-m}. \tag{18}$$

We can show that the density function of the ith order statistic of any EGL distribution can be expressed as

$$f_{i:n}(x) = \sum_{r,k=0}^{\infty} m_{r,k}\, f_{EPL}(x;\lambda,\beta,r+k+1), \tag{19}$$

where $f_{EPL}(x;\lambda,\beta,r+k+1)$ denotes the density function of exponentiated power Lindley distribution with parameters λ,β and $r+k+1$,

$$m_{r,k} = \frac{n!\,(r+1)\,(i-1)!\,c_{r+1}}{(r+k+1)} \sum_{j=0}^{n-i} \frac{(-1)^j\, f_{j+i-1,k}}{(n-i-j)!\,j!}.$$

Here, c_r is given by (9) and the quantities $f_{j+i-1,k}$ can be determined given that $f_{j+i-1,0} = c_0^{j+i-1}$ and recursively we have:

$$f_{j+i-1,k} = (k c_0)^{-1} \sum_{m=1}^{k} [m(j+i)-k]\, c_m\, f_{j+i-1,k-m}, k \geq 1.$$

Equation (19) is the main result of this section. It reveals that the pdf of the *i*th order statistic is a triple linear combination of exponentiated power Lindley distributions. Therefore, several mathematical quantities of these order statistics like ordinary and incomplete moments, factorial moments, mgf, mean deviations and others can be derived using this result.

2.7. Simulation study

In this section, we propose Inverse cdf method for generating random data from the EE-PL distribution. If $U \sim U(0,1)$, the solution of non-linear equation

$$u = \frac{\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\, \mathrm{x}^{\beta}}\right]^{\alpha}}{\left[1-(1+\frac{\lambda x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\, \mathrm{x}^{\beta}}\right]^{\alpha} + 1 - \left[1-(1+\frac{\lambda\ x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\, \mathrm{x}^{\beta}}\right]^{\gamma}} \tag{20}$$

has cdf (4).

3. Asymptotic Properties and Extreme Value

One of the main usage of the idea of an asymptotic distribution is in providing approximations to the cumulative distribution functions of the statistical estimators. Moreover, the extreme value theory is a branch of statistics dealing with the extreme deviations from the median of probability distributions. It seeks to assess, from a given ordered sample of a given random variable, the probability of events that are more extreme than any previously observed. Extreme value analysis is widely used in many disciplines,

3.1. Asymptotic properties

The asymptotic of cdf, pdf and hrf of the EE-PL distribution as $x \to 0$ are, respectively, given by

$$\begin{aligned} F(x) &\sim (\lambda x^{\beta})^{\alpha} \quad as \quad x \to 0, \\ f(x) &\sim \alpha\beta\lambda^{\alpha} x^{\alpha\beta-1} \quad as \quad x \to 0, \\ h(x) &\sim \alpha\beta\lambda^{\alpha} x^{\alpha\beta-1} \quad as \quad x \to 0. \end{aligned}$$

The asymptotic of cdf, pdf and hrf of the EE-PL distribution as $x \to \infty$ are, respectively, given by

$$\begin{aligned} 1-F(x) &\sim \frac{\gamma\lambda}{1+\lambda} x^{\beta} e^{-\lambda x^{\beta}} \quad as \quad x \to \infty, \\ f(x) &\sim \frac{\beta\gamma\lambda^2}{1+\lambda} x^{2\beta-1} e^{-\lambda x^{\beta}} \quad as \quad x \to \infty, \\ h(x) &\sim \beta\lambda x^{\beta-1} \quad as \quad x \to \infty. \end{aligned}$$

These equations show the effect of parameters on the tails of the EE-PL distribution.

3.2. Extreme Value

Let $X_1,...,X_n$ be a random sample from (5) and $\bar{X} = (X_1 + ... + X_n)/n$ denote the sample mean, then by the usual central limit theorem, the distribution of $\sqrt{n}(\bar{X} - E(X))/\sqrt{Var(X)}$ approaches the standard normal distribution as $n \to \infty$. Sometimes one would be interested in the asymptotic of the extreme values $M_n = \max(X_1,...,X_n)$ and $m_n = \min(X_1,...,X_n)$. For 4, it can be seen that

$$\lim_{t\to 0} \frac{F(tx)}{F(t)} = x^{\alpha\beta},$$

and

$$\lim_{t\to \infty} \frac{1-F(tx)}{1-F(t)} = \mathrm{e}^{-\alpha\lambda x^{\beta}}.$$

Thus, it follows from Theorem 1.6.2 in Leadbetter et al. (1983) that there must be norming constants $a_n > 0, b_n, c_n > 0$ and d_n such that

$$Pr[a_n(M_n - b_n) \le x] \to \mathrm{e}^{-\mathrm{e}^{-\lambda\alpha x^{\beta}}},$$

and

$$Pr[a_n(m_n - b_n) \le x] \to 1 - \mathrm{e}^{-x^{\alpha\beta}},$$

as $n \to \infty$. Using Corollary 1.6.3 of Leadbetter et al. (1983), we can obtain the form of normalizing constants a_n, b_n, c_n and d_n.

4. Estimation

Several approaches for parameter estimation have been proposed in the literature but the maximum likelihood method is the most commonly employed. Here, we consider estimation of the unknown parameters of the EE-PL distribution by the method of maximum likelihood. Let $x_1, x_2, ..., x_n$ be observed values from the EE-PL distribution with parameters α, β, γ and λ. The log-likelihood function for $(\alpha; \beta; \gamma; \lambda)$ is given by

$$\begin{aligned} \ell_n &= 2n\log(\lambda\beta) + (\beta-1)\sum_{i=1}^{n}\log(x_i) + \beta\sum_{i=1}^{n}\log(1+x_i) - \lambda\sum_{i=1}^{n}x_i \\ &\quad + (\alpha+1)\sum_{i=1}^{n}\log k_i + \sum_{i=1}^{n}\log(\alpha + (\gamma-\alpha)k_i^{\alpha}) - 2\sum_{i=1}^{n}\log(k_i^{\alpha} + 1 - k_i^{\gamma}) \end{aligned}$$

where

$$k_i = 1-(1+\frac{\lambda}{1+\lambda}x_i^{\beta})e^{-\lambda x_i^{\beta}}.$$

The derivatives of the log-likelihood function with respect to the parameters α,β,γ and λ are given respectively, by

$$\begin{aligned}
\frac{\partial \ell_n}{\partial \alpha} &= \sum_{i=1}^{n}\log k_i+\sum_{i=1}^{n}\frac{1-k_i^{\alpha-1}(\alpha+k_i)}{\alpha+(\gamma-\alpha)k_i^{\alpha}}-2\sum_{i=1}^{n}\frac{\alpha k_i^{\alpha-1}}{k_i^{\alpha}+1-k_i^{\alpha}}\\
\frac{\partial \ell_n}{\partial \beta} &= \frac{2n}{\beta}+\sum_{i=1}^{n}\log(x_i(x_i+1))+(\alpha-1)\sum_{i=1}^{n}\frac{k_i^{(\beta)}}{k_i}+\sum_{i=1}^{n}\frac{\alpha(\gamma-\alpha)k_i^{\alpha-1}k_i^{(\beta)}}{\alpha+(\gamma+\alpha)k_i^{\alpha}}\\
&\quad -2\sum_{i=1}^{n}\frac{\alpha k_i^{(\beta)}k_i^{\alpha-1}-\gamma k_i^{(\beta)}k_i^{\gamma-1}}{k_i^{\alpha}+1-k_i^{\gamma}}\\
\frac{\partial \ell_n}{\partial \gamma} &= \sum_{i=1}^{n}\frac{k_i^{\alpha}}{\alpha+(\gamma-\alpha)k_i^{\alpha}}+2\sum_{i=1}^{n}\frac{k_i^{\gamma}\log(k_i)}{k_i^{\alpha}+1-k_i^{\gamma}}\\
\frac{\partial \ell_n}{\partial \lambda} &= \frac{2n}{\lambda}-\sum_{i=1}^{n}x_i+(\alpha-1)\sum_{i=1}^{n}\frac{k_i^{(\lambda)}}{k_i}+\sum_{i=1}^{n}\frac{\alpha(\gamma-\alpha)k_i^{\alpha-1}k_i^{(\lambda)}}{\alpha+(\gamma+\alpha)k_i^{\alpha}}\\
&\quad -2\sum_{i=1}^{n}\frac{\alpha k_i^{(\lambda)}k_i^{\alpha-1}-\gamma k_i^{(\lambda)}k_i^{\gamma-1}}{k_i^{\alpha}+1-k_i^{\gamma}},
\end{aligned}$$

where

$$\begin{aligned}
k_i^{(\beta)} &= \frac{\partial k_i}{\partial \beta}=\left[1+\frac{\lambda}{\lambda+1}(x_i^{\beta}-1)\right]x_i^{\beta}e^{-\lambda x_i^{\beta}}\log x_i\\
k_i^{(\lambda)} &= \frac{\partial k_i}{\partial \lambda}=x_i^{\beta}e^{-\lambda x_i^{\beta}}\left[1+\frac{\lambda}{\lambda+1}x_i^{\beta}+\frac{1}{(1+\lambda)^2}\right].
\end{aligned}$$

The maximum likelihood estimates (MLEs) of $(\alpha;\beta;\gamma;\lambda)$, say $(\widehat{\alpha};\widehat{\beta};\widehat{\gamma};\widehat{\lambda})$, are the simultaneous solution of the equations $\frac{\partial \ell_n}{\partial \alpha}=0;\frac{\partial \ell_n}{\partial \beta}=0;\frac{\partial \ell_n}{\partial \gamma}=0;\frac{\partial \ell_n}{\partial \lambda}=0$.

For estimating the model parameters, numerical iterative techniques should be used to solve these equations.We can investigate the global maxima of the log-likelihood by setting different starting values for the parameters. The information matrix will be required for interval estimation. Let $\theta=(\alpha;\beta,\gamma,\lambda)^T$, then the asymptotic distribution of $\sqrt{n}(\theta-\hat{\theta})$ is $N_4(0,K(\theta)^{-1})$, under standard regularity conditions (see Lehmann and Casella, [**?**] 1998, pp. 461-463), where $K(\theta)$ is the expected information matrix. The asymptotic behavior remains valid if $K(\theta)$) is superseded by the observed information matrix multiplied by $1/n$, say $I(\theta)/n$, approximated by $\hat{\theta}$, i.e. $I(\hat{\theta})/n$. We have

$$I(\theta)=-\begin{bmatrix} I_{\alpha\alpha} & I_{\alpha\beta} & I_{\alpha\gamma} & I_{\alpha\lambda}\\ I_{\beta\alpha} & I_{\beta\beta} & I_{\beta\gamma} & I_{\beta\lambda}\\ I_{\gamma\alpha} & I_{\gamma\beta} & I_{\gamma\gamma} & I_{\gamma\lambda}\\ I_{\lambda\alpha} & I_{\lambda\beta} & I_{\lambda\gamma} & I_{\lambda\lambda}\end{bmatrix}$$

where

$$I_{\alpha\alpha}=\frac{\partial^2\ell_n}{\partial\alpha^2};\quad I_{\alpha\beta}=I_{\alpha\beta}=\frac{\partial^2\ell_n}{\partial\alpha\partial\beta};\quad I_{\alpha\gamma}=I_{\alpha\gamma}=\frac{\partial^2\ell_n}{\partial\alpha\partial\gamma};\quad I_{\alpha\lambda}=I_{\alpha\gamma}=\frac{\partial^2\ell_n}{\partial\alpha\partial\lambda}$$

$$I_{\beta\gamma}=I_{\gamma\beta}=\frac{\partial^2\ell_n}{\partial\beta\partial\gamma};\quad I_{\beta\lambda}=I_{\lambda\beta}=\frac{\partial^2\ell_n}{\partial\beta\partial\lambda};\quad I_{\gamma\lambda}=I_{\lambda\gamma}=\frac{\partial^2\ell_n}{\partial\gamma\partial\lambda}.$$

5. Characterizations

This section deals with various characterizations of EE-PL distribution. These characterizations are presented in four directions: (i) based on the ratio of two truncated moments; (ii) in terms of the hazard function; (iii) in terms of the reverse hazard function and (iv) based on the conditional expectation of certain function of the random variable. It should be noted that characterization (i) can be employed also when the cdf does not have a closed form. We present our characterizations $(i)-(iv)$ in four subsections.

5.1. Characterizations based on truncated moments

Our first characterization employs a theorem due to Glänzel (1986) , see Theorem 1 of Appendix A. The result, however, holds also when the interval H is not closed since the condition of Theorem 1 is on the interior of H. We like to mention that this kind of characterization based on a truncated moment is stable in the sense of weak convergence (see, Glänzel (1990)).

Let $X:\Omega\to(0,\infty)$ be a continuous random variable and let

$$q_1(x)=\frac{\left\{1+\left[1-\left(1+\frac{\lambda x^\beta}{1+\lambda}\right)e^{-\lambda x^\beta}\right]^\alpha-\left[1-\left(1+\frac{\lambda x^\beta}{1+\lambda}\right)e^{-\lambda x^\beta}\right]^\gamma\right\}^2}{\left\{\alpha+(\gamma-\alpha)\left[1-\left(1+\frac{\lambda x^\beta}{1+\lambda}\right)e^{-\lambda x^\beta}\right]^\gamma\right\}}$$

and

$$q_2(x)=q_1(x)\left[1-\left(1+\frac{\lambda x^\beta}{1+\lambda}\right)e^{-\lambda x^\beta}\right]^\alpha$$

for $x>0$. The random variable X belongs to the family (5) if and only if the function η defined in Theorem1 has the form

$$\eta(x)=\frac{1}{2}\left\{1+\left[1-\left(1+\frac{\lambda x^\beta}{1+\lambda}\right)e^{-\lambda x^\beta}\right]^\alpha\right\},\quad x>0.$$

Proof. Let X be a random variable with pdf (2.2), then

$$(1-F(x))E[q_1(X)\mid X\geq x]=\frac{1}{\alpha(1+\lambda)}\left\{1-\left[1-\left(1+\frac{\lambda x^\beta}{1+\lambda}\right)e^{-\lambda x^\beta}\right]^\alpha\right\}$$

and

$$(1-F(x))E[q_2(X)\mid X\geq x]=\frac{1}{2\alpha(1+\lambda)}\left\{1-\left[1-\left(1+\frac{\lambda x^\beta}{1+\lambda}\right)e^{-\lambda x^\beta}\right]^{2\alpha}\right\}.$$

Further,

$$\eta(x)\,q_1(x)-q_2(x)=\frac{q_1(x)}{2}\left\{1-\left[1-\left(1+\frac{\lambda x^{\beta}}{1+\lambda}\right)e^{-\lambda x^{\beta}}\right]^{\alpha}\right\}>0\ for\ \ x>0$$

Conversely, if η is given as above, then

$$s'(x)=\frac{\eta'(x)\,q_1(x)}{\eta(x)\,q_1(x)-q_2(x)}=\frac{\alpha\beta\lambda^2x^{\beta-1}\left(1+x^{\beta}\right)e^{-\lambda x^{\beta}}\left[1-\left(1+\frac{\lambda x^{\beta}}{1+\lambda}\right)e^{-\lambda x^{\beta}}\right]^{\alpha-1}}{1-\left[1-\left(1+\frac{\lambda x^{\beta}}{1+\lambda}\right)e^{-\lambda x^{\beta}}\right]^{\alpha}},\ \ x>0,$$

and hence

$$s(x)=-\lambda\log\left\{1-\left[1-\left(1+\frac{\lambda x^{\beta}}{1+\lambda}\right)e^{-\lambda x^{\beta}}\right]^{\alpha}\right\},\ \ x>0.$$

Now, according to Theorem 1, X has density (2.2).

Let $X:\Omega\to(0,\infty)$ be a continuous random variable and let q_1 be as in Proposition (5.1). Then, X has $pdf\ (2.2)$ if and only if there exist functions q_2 and η defined in Theorem 1 satisfying the differential equation

$$\frac{\eta'(x)\,q_1(x)}{\eta(x)\,q_1(x)-q(x)}=\frac{\alpha\beta\lambda^2x^{\beta-1}\left(1+x^{\beta}\right)e^{-\lambda x^{\beta}}\left[1-\left(1+\frac{\lambda x^{\beta}}{1+\lambda}\right)e^{-\lambda x^{\beta}}\right]^{\alpha-1}}{1-\left[1-\left(1+\frac{\lambda x^{\beta}}{1+\lambda}\right)e^{-\lambda x^{\beta}}\right]^{\alpha}},\ \ x>0.$$

The general solution of the differential equation in Corollary (5.1) is

$$\eta(x)=\left\{1-\left[1-\left(1+\frac{\lambda x^{\beta}}{1+\lambda}\right)e^{-\lambda x^{\beta}}\right]^{\alpha}\right\}^{-1}\times$$
$$\left[-\int\alpha\beta\lambda^2x^{\beta-1}\left(1+x^{\beta}\right)e^{-\lambda x^{\beta}}\left[1-\left(1+\frac{\lambda x^{\beta}}{1+\lambda}\right)e^{-\lambda x^{\beta}}\right]^{\alpha-1}\left(q_1(x)\right)^{-1}q_2(x)\,dx+D\right],$$

where D is a constant. Note that a set of functions satisfying the above differential equation is given in Proposition (5.1) with $D=\frac{1}{2}$.

For $\alpha=\gamma=1$, $q_1(x)\equiv1$ and $q_2(x)=e^{-\lambda x^{\beta}}$, we have $\eta(x)=\frac{1}{2}e^{-\lambda x^{\beta}}$, $x>0$, $s'(x)=\lambda\beta x^{\beta-1}$, $x>0$ and

$$\eta(x)=e^{\lambda x^{\beta}}\left[-\int\lambda\beta x^{\beta-1}\left(1+x^{\beta}\right)e^{-\lambda x^{\beta}}q_2(x)\,dx+D\right].$$

5.2. Characterization in terms of the hazard function

It is known that the hazard function, h_F, of a twice differentiable distribution function, F, satisfies the first order differential equation

$$\frac{f'(x)}{f(x)} = \frac{h'_F(x)}{h_F(x)} - h_F(x).$$

For many univariate continuous distributions, this is the only characterization available in terms of the hazard function. The following characterization establishes a non-trivial characterization of EE-PL,for $\alpha = \gamma = 1$, in terms of the hazard function which is not of the above trivial form.

Let $X : \Omega \to (0,\infty)$ be a continuous random variable. Then, X has pdf (5), for $\alpha = \gamma = 1$, if and only if its hazard function $h_F(x)$ satisfies the differential equation

$$h'_F(x) - (\beta-1)x^{-1}h_F(x) = \lambda^2\beta^2 x^{2(\beta-1)}\left(1+\lambda+\lambda x^{\beta}\right)^{-2},\ x>0,$$

with the initial condition $h_F(0) = 0$ for $\beta > 1$.

Proof. If X has pdf (2.2), for $\alpha = \gamma = 1$, then clearly the above differential equation holds. Now, if the differential equation holds, then

$$\frac{d}{dx}\left\{x^{-(\beta-1)}h_F(x)\right\} = \lambda^2\beta^2\left\{\frac{1+x^{\beta}}{1+\lambda+\lambda x^{\beta}}\right\}$$

or

$$h_F(x) = \frac{\lambda^2\beta x^{\beta-1}\left(1+x^{\beta}\right)}{1+\lambda+\lambda x^{\beta}} \quad x>0,$$

which is the hazard function of (2.2).

5.3. Characterization in terms of the reverse hazard function

The reverse hazard function, r_F, of a twice differentiable distribution function, F , is defined as

$$r_F(x) = \frac{f(x)}{F(x)},\ \ x \in support\ of\ F.$$

In this subsection we present characterization of EE-PL distribution in terms of the reverse hazard function.

Let $X : \Omega \to (0,\infty)$ be a continuous random variable. Then, X has pdf (2.2) if and only if its reverse hazard function $r_F(x)$ satisfies the differential equation

$$r'_F(x)+\lambda\beta x^{\beta-1}r_F(x)=$$
$$\frac{\lambda^2\beta e^{-\lambda x^\beta}}{1+\lambda}\frac{d}{dx}\left\{\frac{x^{\beta-1}\left(1+x^\beta\right)\left\{\alpha+(\gamma-\alpha)\left[1-\left(1+\frac{\lambda x^\beta}{1+\lambda}\right)e^{-\lambda x^\beta}\right]^\gamma\right\}}{\left[1-\left(1+\frac{\lambda x^\beta}{1+\lambda}\right)e^{-\lambda x^\beta}\right]\left\{1+\left[1-\left(1+\frac{\lambda x^\beta}{1+\lambda}\right)e^{-\lambda x^\beta}\right]^\alpha-\left[1-\left(1+\frac{\lambda x^\beta}{1+\lambda}\right)e^{-\lambda x^\beta}\right]^\gamma\right\}}\right\},$$

$x>0$.

5.4. Characterization based on the conditional expectation of certain function of the random variable

In this subsection we employ a single function ψ of X and characterize the distribution of X ,for $\alpha=\gamma=1$, in terms of the conditional expectation of ψ. The following proposition has already appeared in Hamedani's previous work (2013), so we will just state it here which can be used to characterize EE-PL distribution.

Let $X:\Omega\to(a,b)$ be a continuous random variable with cdf F . Let $\psi(x)$ be a differentiable function on (a,b) with $\lim_{x\to a^+}\psi(x)=1$. Then for $\delta\neq1$,

$$E\left[\psi(X)\mid X>x\right]=\delta\psi(x),\qquad x\in(a,b),$$

if and only if

$$\psi(x)=(1-F(x))^{\frac{1}{\delta}-1},\qquad x\in(a,b).$$

For $\alpha=\gamma=1$, $(a,b)=(0,\infty)$, $\psi(x)=\left(1+\frac{\lambda x^\beta}{1+\lambda}\right)e^{-\lambda x^\beta}$ and $\delta=\frac{\lambda}{1+\lambda}$, Proposition 5.4 provides a characterization of the EE-PL distribution.

6. Application

In this section, we illustrate the fitting performance of the EE-PL distribution using a real data set. For the purpose of comparison, we fitted the following models to show the fitting performance of EE-PL distribution by means of real data set:
i) Lindley Distribution, $L(\lambda)$.
ii) Power Lindley distribution, $PL(\beta,\lambda)$.
iii) Generalized Lindley, $GL(\alpha,\lambda)$, (Nadarajah et al. (2011)), with distribution function given by

$$F(x)=\left(1-(1+\frac{\lambda x}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x}}\right)^\alpha.$$

iv) Beta Lindley, $BL(\alpha,\beta,\lambda)$, with distribution function given by

$$F(x)=\int_0^{L(x,\lambda)}t^{\alpha-1}(1-t)^{\beta-1}dt.$$

v) Exponentiated power Lindley distribution, $EPL(\alpha,\beta,\lambda)$, with distribution function given by

$$F(x)=\left(1-(1+\frac{\lambda\, x^{\beta}}{1+\lambda})\mathrm{e}^{-\lambda\, x^{\beta}}\right)^{\alpha}.$$

vi) Odd log-logistic power Lindley distribution $OLL-PL(\alpha,\beta,\lambda)$, (Alizadeh et al. (2017)), with distribution function given by

$$F(x)=\frac{PL(x,\beta,\lambda)^{\alpha}}{PL(x,\beta,\lambda)^{\alpha}+(1-PL(x,\beta,\lambda))^{\alpha}}.$$

vii) Kumaraswamy Power Lindley, $KPL(\alpha,\beta,\gamma,\lambda)$ (Broderick et al. (2012))

$$F(x)=1-[1-PL(x,\beta,\lambda)^{\alpha}]^{\gamma}.$$

viii) Odd Burr-Power Lindley, $OBu-PL(\alpha,\beta,\gamma,\lambda)$ (Altun et al.(2017a))

$$F(x)=1-\left(1-\frac{PL(x,\beta,\lambda)^{\alpha}}{PL(x,\beta,\lambda)^{\alpha}+(1-PL(x,\beta,\lambda))^{\alpha}}\right).$$

ix) Extended Exponential Lindley, $EE-L(\alpha,\gamma,\lambda)$, Ranjbar, et al. (accepted (2018)),

$$F(x)=\frac{L(x,\lambda)^{\alpha}}{L(x,\lambda)^{\alpha}+1-(1-L(x,\lambda))^{\gamma}}.$$

Estimates of the parameters of EE-PL distribution, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Cramer Von Mises and Anderson-Darling statistics (W^* and A^*) are presented for each dataset. We have also considered the Kolmogorov-Smirnov (K-S) statistic and its corresponding p-value and the minimum value of the minus log-likelihood function (-Log(L)) for the sake of comparison. Generally speaking, the smaller values of AIC, BIC, W^* and A^*, the better fit to a data set. All the computations were carried out using the software R.

Note that initial values of model parameters are quite important to obtain the correct MLEs of parameters. To avoid local minima problem, we first obtain the parameter estimate of the Lindley distribution. Then, the estimated parameter of the Lindley distribution is used as the initial value of the parameter of the PL and GL distributions. Then, the estimated parameters of PL distribution, λ and β, is used as the initial values of the EE-PL distribution. This approach is quite useful to obtain correct parameter estimates of extended models.

The data are the exceedances of flood peaks (in m^3/s) of the Wheaton River near Carcross in Yukon Territory, Canada. The data consist of 72 exceedances for the years 1958-1984 rounded to one decimal place. These data were analyzed by Akinsete et al. (2008).

The ML estimates of the parameters and the goodness-of-fit test statistics for the real data set is presented in Table 3 and 4 respectively. As we can see, the smallest values of AIC, BIC, A^*, W^* and $-l$ statistics and the largest p-values belong to the EE-PL distribution. Therefore the EE-PL distribution outperforms the other competitive considered distribution in the sense of this criteria. The OLL-L distribution provides the second best fit and this data set.

Table 2: The data set.

1.7	2.2	14.4	1.1	0.4	20.6	5.3	0.7	1.9	13.0	12.0	9.3	1.4	18.7	8.5	25.5
11.6	14.1	22.1	1.1	2.5	14.4	1.7	37.6	0.6	2.2	39.0	0.3	15.0	11.0	7.3	
22.9	1.7	0.1	1.1	0.6	9.0	1.7	7.0	20.1	0.4	2.8	14.1	9.9	10.4	10.7	30.0
3.6	5.6	30.8	13.3	4.2	25.5	3.4	11.9	21.5	27.6	36.4	2.7	64.0	1.5	2.5	
27.4	1.0	27.1	20.2	16.8	5.3	9.7	27.5	2.5	27.0						

Table 3: Parameter ML estimates and their standard errors (in parentheses) for the data set.

Model	α	β	γ	λ
$Lindley(\lambda)$	–	–	–	0.153 (0.013)
$GL(\alpha,\lambda)$	0.508(0.076)	–	–	0.104 (0.0149)
$PL(\beta,\lambda)$	–	0.700 (0.057)	–	0.338 (0.055)
$BL(\alpha,\beta,\lambda)$	0.555(0.098)	0.274 (0.239)	–	0.333 (0.272)
$EPL(\alpha,\beta,\lambda)$	0.730(0.235)	0.915 (0.595)	–	0.300 (0.279)
$OLLPL(\alpha,\beta,\lambda)$	0.557(0.178)	1.073 (0.244)	–	0.154 (0.091)
$KPL(\alpha,\beta,\gamma,\lambda)$	1.675(2.433)	0.453 (0.432)	7.563 (11.736)	0.279 (0.522)
$OBu(\alpha,\beta,\gamma,\lambda)$	24.91(25.654)	0.024 (0.032)	41.25 (22.520)	0.984 (0.149)
$EEL(\alpha,\gamma,\lambda)$	0.618(0.101)	–	2.770 (1.704)	0.169 (0.028)
$EEPL(\alpha,\beta,\gamma,\lambda)$	4.521(3.067)	0.472 (0.094)	55.07 (58.193)	1.551 (0.643)

Table 4: Goodness-of-fit test statistics for the data set.

Model	*AIC*	*BIC*	$p-value$	W^*	A^*	$-l$
$Lindley(\lambda)$	530.423	532.700	0.001	0.139	0.852	264.211
$GL(\alpha,\lambda)$	509.349	513.902	0.276	0.132	0.822	252.674
$PL(\beta,\lambda)$	508.443	512.996	0.405	0.123	0.766	252.103
$BL(\alpha,\beta,\lambda)$	510.206	517.036	0.297	0.150	0.866	252.221
$EPL(\alpha,\beta,\lambda)$	510.425	517.255	0.395	0.147	0.854	252.212
$OLLPL(\alpha,\beta,\lambda)$	507.937	514.767	0.471	0.093	0.592	250.968
$KPL(\alpha,\beta,\gamma,\lambda)$	512.221	521.328	0.371	0.152	0.866	252.110
$OBu(\alpha,\beta,\gamma,\lambda)$	511.212	520.319	0.401	0.140	0.799	251.606
$EEL(\alpha,\gamma,\lambda)$	508.931	515.761	0.174	0.101	0.662	251.465
$EEPL(\alpha,\beta,\gamma,\lambda)$	500.594	509.701	0.994	0.026	0.180	246.297

In addition, the profile log-likelihood functions of the EE-PL distribution are plotted in Figure 4. These plots reveal that the likelihood equations of the EE-PL distribution have solutions that are maximizers.

Here, we also applied likelihood ratio (LR) tests. The LR tests can be used for comparing the EE-PL distribution with its sub-models. For example, the test of $H_0 : \beta = 1$ against $H_1 : \beta = 1$ is equivalent to comparing the EE-PL and EE-L distributions with each other. For this

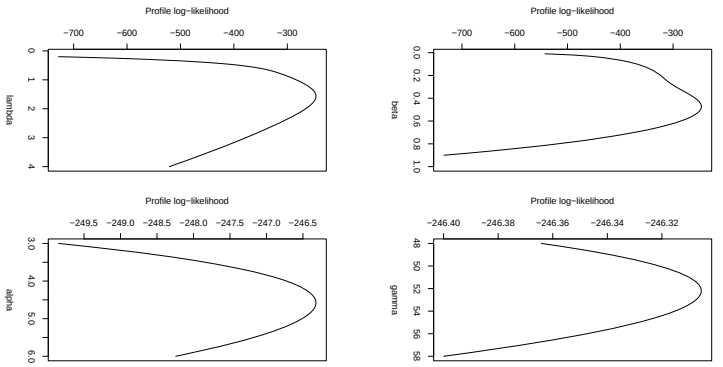


Figure 4: The profile log-likelihood functions of the EE-PL distribution.

test, the LR statistic can be calculated by the following relation

$$LR = \left[l(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}, \widehat{\lambda}) - l(\widehat{\alpha^*}, 1, \widehat{\gamma^*}, \widehat{\lambda^*}) \right],$$

where $\widehat{\alpha^*}, \widehat{\gamma^*}$ and $\widehat{\lambda^*}$ are the ML estimators of α, γ and λ, respectively, obtained under H_0. Under the regularity conditions and if H_0 is assumed to be true, the LR test statistic converges in distribution to a chi square with r degrees of freedom, where r equals the difference between the number of parameters estimated under H_0 and the number of parameters estimated in general, (for $H_0 : \beta = 1$, we have $r = 1$). Table 5 gives the LR statistics and the corresponding p-values.

Table 5: The LR test results.

	Hypotheses	LR	p-value
EE-PL versus Lindley	$H_0 : \alpha = \beta = \gamma = 1$	35.828	< 0.0001
EE-PL versus PL	$H_0 : \alpha = \gamma = 1$	11.612	0.0030
EE-PL versus GL	$H_0 : \alpha = \gamma,\ \beta = 1$	12.754	0.0017
EE-PL versus EPL	$H_0 : \alpha = \gamma$	11.830	0.0005
EE-PL versus EE-L	$H_0 : \beta = 1$	10.336	0.0013

From Table 5, we observe that the computed p-values are too small so we reject all the null hypotheses and conclude that the EE-PL fits the first data better than the considered sub-models according to the LR criterion.

We also plotted the fitted pdfs and cdfs of the considered models for the sake of visual comparison, in Figure 4. Figure 4 suggests that the EE-PL fits the skewed data very well.

7. Conclusion

In this paper, a new distribution called Extended Exponentiated Power Lindley (EE-PL) distribution is introduced. The statistical properties of the EE-PL distribution including the hazard

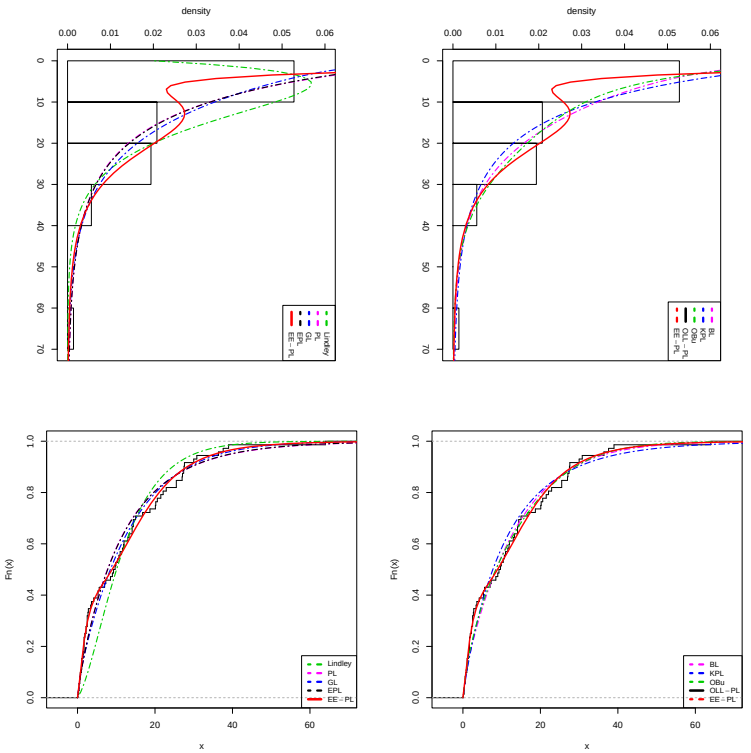


Figure 5: Fitted densities and distribution functions for the data set.

and reverse hazard functions, quantile function, moments, incomplete moments, generating functions, mean deviations, Bonferroni and Lorenz curves, order statistics and maximum likelihood estimation for the model parameters are given. Simulation studies was conducted to examine the performance of the new EE-PL distribution. We also present applications of this new model to a real life data set in order to illustrate the usefulness of the distribution.

REFERENCES

AMROT, W., (2012). Estimation of Finite Population Kurtosis under Two-Phase Sampling for Nonresponse. Statistical Papers, 53, pp. 887–894.

GAMROT, W., (2013). Maximum Likelihood Estimation for Ordered Expectations of Correlated Binary Variables. Statistical Papers, 54, pp. 727–739.

KENNICKELL, A. B., (1997). Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances. In Record Linkage Techniques. W. Alvey and B. Jamerson (eds.) Washington D. C.: National Academy Press, pp. 248–267.

SÄRNDAL, C-E., SWENSSON, B., WRETMAN, J., (1992). Model Assisted Survey Sampling, New York: Springer.

ALIZADEH, M., AFSHARI, M., HOSSEINI, B., RAMIRES, T. G, (2017). Extended Exp-G family of distributions: Properties and Applications. Communication in statistics-simulation and computation, accepted.

ALIZADEH, M., ALTUN, E., OZEL, G., (2017). Odd Burr Power Lindley Distribution with Properties and Applications.Gazi University Journal of Science, Accepted.

AKINSETE, A. FAMOYE, F., LEE, C., (2008) The beta-Pareto distribution, A Journal of Theoretical and Applied Statistics Volume 42, Issue 6, pp. 547–563.

BAKOUCH, H. S., AL-MAHARANI, B. M., AL-SHOMRANI, A. A., MARCHI, V. A. A., LOUZADA, F., (2012): Anextended Lindley distribution, Journal of the Korean Statistical Society, Vol 41 (1), pp. 75– 85.

CAKMAKYAPAN, S., & OZEL, G., (2014). A new customer lifetime duration distribution: the Kumaraswamy Lindley distribution. International Journal of Trade, Economics and Finance, 5, 5, pp. 441–444.

CORDEIRO, G. M.,ALIZADEH, M., TAHIR, M. H., MANSOOR, M., BOURGUIGNON, M., & HAMEDANI, G. G., (2015). The beta odd log-logistic generalized family of distributions, Hacettepe Journal of Mathematics and Statistics, 45, 73, pp. 126-139.

GHITANY, M. E., ATIEH, B. & NADARAJAH, S., (2008). Lindley distribution and its application, Mathematics and Computers in Simulation, 78, pp. 493-506.

GHITANY, M. E., AL-MUTAIRI, D. K., BALAKRISHHNAN, N., & (2013). Al-Enezi, L. J. Power Lindley distribution and associated inference. Computational Statistics and Data Analysis, 64, pp. 20–33.

GLÄNZEL, W., , (1987). A characterization theorem based on truncated moments and its application to some distribution families, Mathematical Statistics and Probability Theory (Bad Tatzmannsdorf, 1986), Vol. B, Reidel, Dordrecht, pp. 75–84.

GLÄNZEL, W., (1990). Some consequences of a characterization theorem based on truncated moments, Statistics: A Journal of Theoretical and Applied Statistics, 21 (4), pp. 613–618.

GRADSHTEYN, I. S., RYZHIK, I. M., (2000). Table of integrals, series, and products. Academic Press, San Diego.

HAMEDANI, G. G., (2013). On certain generalized gamma convolution distributions II, Technical Report No. 484, MSCS, Marquette University.

LEADBETTER, M. R., LINDGREN, G., ROOTZN H., (1983). Extremes and Related Properties of Random Sequences and Processes Springer Statist. Ser., Springer, Berlin.

LEHMANN E. L., CASELLA G., (1998). Theory of Point Estimation, Springer.

LINDLEY, D. V., (1958). Fiducial distributions and Bayesian theorem. Journal of the Royal Statistical Society B, 20, pp. 102–107.

MAZUCHELI, J., ACHCAR J. A., (2011). The Lindley Distribution Applied to Competing Risks Lifetime Data. Computer Methods and Programs in Biomedicine, 104(2), pp. 188-192.

NADARAJAH, S., BAKOUCH, H. S., & TAHMASBI, R., (2011). A generalized Lindley distribution. Sankhya B, 73, pp. 331-359.

OLUYEDE, B. O., YANG, T., & MAKUBATE, B., (2016). A new class of generalized power Lindley distribution with application to lifetime data. Asian Journal of Mathematics and Applications, 6, pp. 1-36.

RANJBAR, V., ALIZADEH, M., Alizade Morad Dr, Extended Generalized Lindley distribution: properties and applications. (Under review)

SHANKER, R., MISHRA, A., (2013). A quasi Lindley distribution, African Journal of Mathematics andComputer Science Research, Vol.6 (4), pp. 64-71.

SHARMA, V, SINGH, S., SINGH., U., & AGIWAL, V., (2015). The inverse Lindley distribution: a stress-strength reliability model with applications to head and neck cancer data. Journal of Industrial and Production Engineering, 32, 3, pp. 162–173.

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 645–670, DOI 10.21307/stattrans-2018-034

BAYESIAN SPATIAL ANALYSIS OF CHRONIC DISEASES IN ELDERLY CHINESE PEOPLE USING A STAR MODEL

Ping Gao[1], **Hikaru Hasegawa**[2]

ABSTRACT

Chronic diseases have become important factors affecting the health of elderly Chinese people. Because the prevalence of chronic diseases varies among the provinces, it is necessary to understand the spatial effects on these diseases, as well as their relationships with potential risk factors. This study applies a structured additive regression model and the `R2BayesX` package to conduct a Bayesian analysis. The data are taken from the 2000, 2006, and 2010 Chinese Urban and Rural Elderly Population Surveys. The findings are as follows: (1) the following covariates have considerable effects on chronic diseases in general, and on specific chronic diseases (hypertension and heart disease) (in descending order): census register (rural or urban), gender, smoking, drinking, province, time, age, cultural activities, years of education, and sports activities; (2) the effect of marital status is negligible; (3) province is a critical factor, with the highest spatial effect appearing in two types of provinces: economically developed provinces, and economically backward provinces; and (4) time also has considerable effects. Based on these findings, the government should further strengthen its investment in rural areas and economically backward provinces as a cost-effective intervention, and should educate the population on the harmful effects of smoking and drinking alcohol on health.

Key words: Bayesian analysis, Markov chain Monte Carlo (MCMC), `R2BayesX`, Spatial effect, Structured additive regression (STAR) models.

1. Introduction

The medical definition of a chronic disease is a disease that persists for a long time. For example, the U.S. National Center for Health Statistics defines a chronic disease as one that lasts for three months or more. In China, more than 70% of elderly people struggle with a chronic disease. In addition, the majority of the elderly suffer from multiple chronic diseases (Thorpe and Howard, 2006; Vogeli et al., 2007; Wolff et al., 2002). The prevalence of chronic diseases is related to many potential risk factors, such as age, lifestyle habits (smoking, drinking), a lack of exercise, and so on. In addition to these influencing factors, the spacial dimension is increasingly being considered as an independent factor, which can be examined using geostatistical methods.

[1]Graduate School of Economics and Business, Hokkaido University. Japan. E-mail: gaoping4069111@gmail.com
[2]Department of Economics, Hokkaido University. Japan. E-mail: hasegawa@econ.hokudai.ac.jp

Kriging is a common geostatistical method that produces a map of a quantity of interest over a geographical region. In addition, as an extension to kriging, universal kriging takes into account the linearity of the covariate (Cressie, 2015, pp.151–172). However, when the effect of the covariate is nonlinear, the universal kriging method is not appropriate. In this case, a geoadditive model may be a better choice. The geoadditive model was introduced by Kammann and Wand (2003), and accounts for non-linear covariate effects under the assumption of additivity. Today, many researchers apply this model in their studies (Basile et al., 2013; Geniaux and Napoléone, 2008; Sauleau et al., 2007; Wand et al., 2011).

Another powerful model used in spatial analyses is the generalized additive model (GAM). The GAM is suitable for modelling nonlinear effects of continuous covariates in regression models with non-Gaussian responses. Furthermore, structured additive regression (STAR) models extend GAM models by including spatial effects, the nonlinear effects of continuous factors, and linear or fixed effects in one model (Kneib, 2006). STAR models include generalized linear models and generalized additive models as special cases, but also allow for a wider class of effects, such as geographical or spatio-temporal effects (Fahrmeir et al., 2004, 2013; Umlauf et al., 2015).

With the rising prevalence of chronic diseases (Freedman and Martin, 2000; Thorpe and Howard, 2006) and the large elderly population (the population aged 60 and over in China was $2,308,600$ in 2016), the number of elderly Chinese people suffering from chronic diseases is very high. Because chronic diseases are essentially permanent, they introduce a heavy economic burden to families and society. In China, the most common cause of death is chronic diseases, rather than infectious diseases (Gu et al., 2009; He et al., 2005). For example, chronic disease-induced deaths accounted for 71.88% of all deaths among residents in Kunming (Yunnan provincial capital) in the period 2007–2010 (Li et al., 2012). The high number of elderly people with severe chronic conditions places a significant burden on medical care. Thus, the Chinese government faces enormous challenges in terms of medical investment. Numerous studies have examined chronic diseases, with many focusing on the factors influencing such diseases.

Fillingim et al. (2009) studied samples from different countries (China, France, Sweden, United States, etc.), and found that the prevalence of the most common forms of pain caused by chronic diseases is higher in women than in men. Furthermore, Zhen (2010) used data on Gansu province (in China), finding that the resistance of females to chronic disease pain is poor and that females' thresholds for discomfort are lower than those of men. As a result, women are more likely to visit a doctor and, thus, are more likely to be diagnosed with a chronic disease. Thus, we hypothesise that gender has a great influence on the reported prevalence of chronic diseases and of specific chronic diseases (e.g., hypertension and heart disease) in elderly Chinese people, and that elderly females are more likely to have chronic diseases and specific chronic diseases than are elderly males.

In this study, we use reported prevalence rather than prevalence, because the prevalence is not the true prevalence. Chronic diseases are usually non-fatal dis-

eases and persist for a long time. Many elderly people may have chronic diseases, but may not be aware of this, in which case, they will report not having a disease, even though they do. Thus, the study can only obtain the reported prevalence.

Woolf et al. (2015) noted that people with a high economic status are more conscious of self-care, and that such individuals are more likely to be diagnosed with chronic diseases. Zhen (2010) found the reported prevalence of chronic diseases is significantly affected by access to health care. Income and medical security are greater among the urban elderly than among the rural elderly. Therefore, we hypothesise that the census register is a critical factor related to chronic diseases and to specific chronic diseases (e.g., hypertension and heart disease), and that the reported prevalence of chronic diseases and specific chronic diseases is higher among the urban elderly than among the rural elderly.

Chen (2005), Ye (2013), and Zhao et al. (2015) found that marital status also has an affect on chronic diseases in China. Furthermore, Ye (2013) analysed data on Jilin province, and found that those who are single have the lowest prevalence of chronic diseases, while the prevalence among divorced/widowed persons is the highest. Thus, we hypothesise that the reported prevalence is higher among divorced and widowed elderly people than among other types of elderly people. The WHO (2005) reported that chronic diseases among the elderly are predominantly attributable to unhealthy habits during youth, such as excessive smoking and drinking. In addition, using data on China, Chen (2005), Jiao et al. (2002), and Zhao et al. (2015) found similar results, namely, that cigarette smoking and alcohol usage are risk factors for chronic diseases. Therefore, we hypothesise that (cigarette) smoking and drinking (alcohol) have a considerable influence on the reported prevalence of chronic diseases and specific chronic diseases in China.

Numerous studies have confirmed that age has a considerable effect on chronic diseases in China (Chen, 2005; Jiao et al., 2002; Lin et al., 2002; Yin, 2011), with the prevalence increasing with age. Furthermore, Jiao et al. (2002), Ye (2013), and Yin (2011) pointed out that education and exercise have nonnegligible effects on chronic diseases, with the prevalence increasing for lower levels of education and less exercise. In this study, we divide exercise into two categories: sports activities and cultural activities. Based on the above studies, we hypothesise that the prevalence increases in older people who get less exercise. However, we hypothesise that people with higher levels of education are more likely to report having chronic diseases because they learn more about the dangers of such diseases and pay more attention to their health.

The above studies are limited to a single province or city. Furthermore, with the exception of some statistical descriptive reports, few studies consider the spatial dimension as an independent factor in chronic disease research in China. There are tremendous differences in economic development levels, medical conditions, and living conditions among the provinces in China, all of which can affect the diagnosis and treatment of a chronic disease. As a result, the prevalence of chronic diseases is quite different among the provinces. Thus, we hypothesise that the province is a critical factor affecting chronic diseases and specific chronic diseases. In addition,

we hypothesise that the reported prevalence is similar to the case of the census register, in that it is higher in economically developed provinces.

In summary, based on past studies and on China's tremendous geographic differences, we hypothesise that the following factors are important factors affecting the prevalence of chronic diseases and specific chronic diseases in elderly Chinese people: gender, census register (urban or rural), marital status, smoking, drinking, age, education years, sports activities and cultural activities, and province. In addition, because most of the samples between surveys in 2006 and 2010 are the same, we further take the effects of time into account.

Prior studies usually only consider the linear effects of the continuous covariates (such as education years) on a chronic disease, even though they may have nonlinear effects. Because STAR models can include spatial effects, the nonlinear effects of continuous factors, and linear or fixed effects in a single model, we apply a STAR model in our empirical study in order to determine which covariates have considerable effects on chronic diseases in China.

In applying this model, we use a fully Bayesian estimation based on Markov chain Monte Carlo (MCMC) simulations, as well as `BayesX`, a standalone software package used to fit general STAR models. Moreover, Umlauf et al. (2015) developed an interactive `R` interface for `BayesX`, called `R2BayesX`, which can be used to specify STAR models using `R`'s formula language. Furthermore, this package adds extensive graphics capabilities for visualizing fitted STAR models.

The rest of this paper is structured as follows. Section 2 introduces the STAR models, and their estimations. Section 3 explains the data. Section 4 presents the STAR model applied in this study, and describes the `R2BayesX` settings. Section 5 discusses the empirical results for chronic diseases, as well as for two specific chronic diseases (hypertension and heart disease) using the `R2BayesX` package. Section 6 presents our conclusions.

2. Estimation Methods

2.1. STAR models

STAR models were first introduced by Fahrmeir et al. (2004), and not only contain generalized linear effects, but also allow for nonlinear effects of continuous covariates and spatial effects.

For generalized linear models, the mean μ of the response variable y is linked to a linear predictor η by

$$\mu = h^{-1}(\eta),\ \eta = x'\gamma, \tag{1}$$

where h is a known link function and γ denotes unknown regression coefficients.

Following Fahrmeir et al. (2004, p.734), in STAR models, the linear predictor is replaced by the following structured additive predictor:

$$\eta = f_1(x_1) + \cdots + f_p(x_p) + w'\gamma, \tag{2}$$

where $x_1,\ldots,x_p$ are nonlinear covariates, and f_j are smooth functions, which can represent potentially non-linear effects of continuous covariates or spatially structured and unstructured effects.

Furthermore, when the response variable is binary, the link function becomes a logit function, and we can consider the following logistic STAR model:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \eta = f_1(x_1) + \cdots + f_p(x_p) + w'\gamma, \tag{3}$$

where p denotes the probability of a specific event occurring (such as the probability of a person having a chronic disease).

2.2. Estimation of STAR models

In this study, the STAR model is estimated using a Bayesian inference. For the Bayesian inference, all components of the STAR models must be supplemented with appropriate prior assumptions.

In the STAR model (2), $w'\gamma$ denotes the fixed effects. In general, a diffuse prior $p(\gamma) \propto const$ is assumed for the fixed effects parameter γ. At the same time, specific priors are given to the functions $f_j(\cdot)$, and depend on the type of the covariate.

For the nonlinear effects of continuous covariates $f_j(\cdot)$, Bayesian P-splines are utilized. P-splines are an improvement over B-splines, and introduce a penalty variable to prevent overfitting.

The basic idea behind P-splines is dividing the data interval into a relatively large number of sub-intervals, and an unknown smooth function f of a covariate x can be approximated by a linear combination of some basis functions. P-splines can be approximated by a polynomial spline of degree l, defined on a set of equally spaced knots $x^{\min} = \zeta_1 < \zeta_2 < \cdots < \zeta_m = x^{\max}$ within the domain of x. Following Fahrmeir et al. (2013, pp.426–431), a spline can be expressed by an adequate linear combination of $d = m + l - 1$ B-spline basis functions:

$$f(x) = \sum_{j=1}^{d} \beta_j B_j(x), \tag{4}$$

where $B_j(x)$ of degree l is defined as follows: for $j = 1,\ldots,d-1$,

$$\begin{cases} B_j^0(x) = I(\zeta_j \le x < \zeta_{j+1}) & l = 0 \\ B_j^1(x) = \dfrac{x-\zeta_{j-1}}{\zeta_j - \zeta_{j-1}} I(\zeta_{j-1} \le x < \zeta_j) + \dfrac{\zeta_{j+1}-x}{\zeta_{j+1}-\zeta_j} I(\zeta_j \le x < \zeta_{j+1}) & l = 1 \\ B_j^l(x) = \dfrac{x-\zeta_{j-l}}{\zeta_j - \zeta_{j-l}} B_{j-1}^{l-1}(x) + \dfrac{\zeta_{j+1}-x}{\zeta_{j+1}-\zeta_{j+1-l}} B_j^{l-1}(x) & l \ge 2, \end{cases} \tag{5}$$

where $I(\cdot)$ is an indicator function.

The crucial choice for P-splines is the number of knots: too few knots may not be flexible enough, while choosing too many knots may overfit the data. To prevent

overfitting, a penalty term is included. The penalty terms are expressed as in Eilers and Marx (1996):

$$P(\lambda) = \frac{1}{2}\lambda \sum_{j=r+1}^{d} (\Delta^r \beta_j)^2, \tag{6}$$

where λ is the smoothing parameter and Δ^r denotes the rth-order differences. In most applications, second-order differences are chosen, which are defined as $\Delta^2\beta_j = \Delta^1\Delta^1\beta_j = \Delta^1\beta_j - \Delta^1\beta_{j-1} = \beta_j - 2\beta_{j-1} + \beta_{j-2}$. Furthermore, when $\lambda \to \infty$, the function estimate for $f(x)$ is close to linear in the case of second-order differences.

In applications, a common choice for P-splines is B-splines of degree $l = 3$, with $m = 20$ equidistant knots. These setting ensure that the estimated function is twice continuously differentiable, which allows sufficient flexibility to capture the typical nonlinear mode.

The general form of the prior of β_j for a P-spline is given by the multivariate normal distribution:

$$p(\beta_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right), \tag{7}$$

where K_j is a penalty matrix, and τ_j^2 is a prior variance, which determines the impact of the prior distribution on the function estimates. For the full Bayesian inference, weakly informative inverse Gamma hyperpriors $\tau_j^2 \sim \mathrm{IG}(a_j, b_j)$ are assigned to τ_j^2, with $a_j = b_j = 0.001$ as a general setting. More detailed information about the Bayesian P-splines can be found in Lang and Brezger (2004).

As we mentioned above, f_j denotes smooth functions that can be used to represent the potentially non-linear effects of continuous covariates or to represent a spatial effect. If f_j represents a spatial effect, it is expressed as $f_{spat}(\cdot)$.

The spatial effect in a STAR model is the effect of a spatial covariate. Usually, this is a proxy for unobserved influential factors, some of which may have a strong spatial correlation (structured), while others may be present only locally (unstructured). Thus, in order to distinguish between these two kinds of spatial effects, $f_{spat}(\cdot)$ is split into a spatially correlated (structured) part $f_{str}(\cdot)$ and spatially uncorrelated (unstructured) part $f_{unstr}(\cdot)$, i.e. $f_{spat}(\cdot) = f_{str}(\cdot) + f_{unstr}(\cdot)$. The structured spatial effects simply indicate that the spatial effects are correlated. There is no specific structure imposed on the spatial effects.

The spatially structured effect $f_{str}(\cdot)$ can be specified using stationary Gaussian random field (GRF) priors. When the place of residence is known exactly, given by geographical x-y coordinates, the spatial analysis can be conducted using a stationary GRF. The estimation of a GRF is based on the centroids of particular regions for geosplines and geokriging terms. More detailed information about stationary GRF priors can be found in Fahrmeir et al. (2013, pp.500–530). The spatially unstructured effect $f_{unstr}(\cdot)$ can be specified using simple Gaussian i.i.d. priors, and denotes the random effect of a covariate.

3. Data

We define a chronic disease using the medical definition: a disease that lasts for a long time (more than three months) and cannot be cured.

The data for this study are taken from the 2000, 2006, and 2010 Chinese Urban and Rural Elderly Population Surveys, conducted by the China Research Center on Aging of the National Committee on Aging. The survey in 2000 only investigated whether people were suffering from chronic diseases, while those in 2006 and 2010 were expanded to include questions on specific chronic diseases. And most of the samples between surveys in 2006 and 2010 are the same. Thus, we also analyse two specific chronic diseases (hypertension and heart disease, both of which are common in China, with a prevalence of more than 10%) in 2006 and 2010.

Moreover, the surveys focused on the following 20 representative provinces, municipalities, and autonomous regions: the eastern region - Beijing, Shanghai, Hebei, Liaoning, Jiangsu, Zhejiang, Fujian, Shandong, and Guangdong; the central region - Heilongjiang, Anhui, Henan, Shanxi, Hubei, and Hunan; and the western region - Sichuan, Yunnan, Shaanxi, Xinjiang, and Guangxi. The selected provinces, municipalities, and autonomous regions are shown in Figure 1. In China, the degree of economic development is closely related to geographical location. In general, provinces in the eastern region are almost economically developed provinces, provinces in the central region are moderately developed provinces, and provinces in the western region tend to be economically backward provinces.

The data sampling method used is the same as that of the Fifth Population Census; based on the distribution of the population aged 60 and older, a quota from each of the regions is determined. Then, stratified sampling is used to confirm that the survey results represent the total elderly population in China (Gao and Li, 2016).

After the survey samples were determined, the interviewers conducted household surveys. Here, interviews were conducted by interviewers, who then completed the questionnaires on behalf of the interviewees, based on their responses. No questionnaires were completed by the interviewees. Then, the interviewers checked and verified the responses after the investigation. As a result, the data accuracy is high. The three surveys generated $20,256$ responses, $19,947$ responses, and $19,986$ responses, respectively.

4. Model

4.1. Model specification

Because the samples include data on whether people suffer from chronic diseases in 2000, 2006, and 2010, the first step is a Bayesian analysis of the geographic distribution of chronic diseases and their relationships with potential risk factors. Moreover, for 2006 and 2010, data are included on common chronic diseases (hypertension and heart disease). Thus, the second step is a Bayesian analysis of the

geographic distribution of these two common chronic diseases and their relationships with potential risk factors. The second step is a refinement of the first step. Therefore, this paper describes how to implement the first step only.

Given a set of observations y_i, $1 \leq i \leq n$, y_i is a binary response for a chronic disease, such that

$$y_i = \begin{cases} 1 & \text{if one has a chronic disease} \\ 0 & \text{otherwise.} \end{cases}$$

Because the responses are binary, we consider a logistic STAR model to estimate the probability of an elderly person having a chronic disease ($y_i = 1$) versus the probability of an elderly person not having a chronic disease ($y_i = 0$):

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \eta_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + w_i'\gamma,\ 1 \leq i \leq n, \tag{8}$$

where $p_i = \Pr(y_i = 1)$, $x_{i1}, \ldots, x_{ip}$ are p continuous covariates, f_j are smooth functions, $w_i = (w_{i1}, \ldots, w_{ir})'$ is a vector of r categorical covariates, and γ is an r-dimensional vector of unknown regression coefficients for the categorical covariates w_i. The response is distributed as a Bernoulli random variable, such that
$f(y_i|\eta_i) = p_i^{y_i}(1-p_i)^{(1-y_i)} = \exp[y_i\eta_i - \log(1+\exp(\eta_i))]$ for $y_i = 0, 1$,
where $\eta_i = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$.

Based on the previous studies mentioned in the Introduction, we analyse the linear effects of the following categorical covariates: gender of the elderly person (female or male), census register (urban or rural), marital status (live with spouse, live differently with spouse, widowed, divorce, and unmarried), smoking (smoked previously, currently smoke, and never smoke), and drinking (drank previously, currently drink, and never drink). In addition, we investigate the potential nonlinear effects of the following continuous covariates: the elderly's age (Age), education years (EY), number of sports activities (SA), and cultural activities (CA). Furthermore, because the observations on chronic diseases are associated with where a person lives, it is important to account for geographical/spatial differences in the analysis. Therefore, by taking spatial and nonlinear effects into account, the predictor shown in (8) is replaced by the following predictor:

$$\eta_i = f_1(Age_i) + f_2(EY_i) + f_3(SA_i) + f_4(CA_i) + f_{spat}(Province_i) + w_i'\gamma, \tag{9}$$

where $f_1(Age_i)$, $f_2(EY_i)$, $f_3(SA_i)$, and $f_4(CA_i)$ are nonlinear smooth effects of the continuous covariates, and $f_{spat}(Province)$ is the effect of the spatial covariate $Province_i$. Here, $Province_i \in \{1, \ldots, S\}$ is an integer, where S is the number of surveyed regions. Every integer indicates the province, municipality, or autonomous region in which the respondent is living. For example, $Province_i = 1$ means the respondent lives in Heilongjiang province, and $Province_i = 8$ denotes a respondent living in Beijing city. In this study, the number of surveyed provinces, municipalities, and autonomous

regions is 20; thus $S = 20$ (i.e., $Province_i \in \{1, \ldots, 20\}$).

In China, the provinces are connected in that they usually have some similarity and correlation, and so they are spatially correlated. Therefore, we consider the structured spatial effect $f_{str}(province)$ rather than the unstructured spatial effect $f_{unstr}(province)$.

Finally, we estimate the following STAR model:

$$\eta_i = \beta_0 + f_1(Age_i) + f_2(EY_i) + f_3(SA_i) + f_4(CA_i) + f_{str}(province_i) + w_i'\gamma, \qquad (10)$$

where $f_{str}(province_i)$ is a structured spatial effect, and $w_i'\gamma$ are the linear effects of the following categorical covariates: gender of the elderly, census register, marital status, smoking, and drinking.

In this study, the estimation of the above STAR model is obtained using a Bayesian inference. In the STAR model (10), for the fixed effects $w_i'\gamma$, a diffuse prior $p(\gamma) \propto const$ is assumed for the parameter γ. For the nonlinear effects of the continuous covariates $f_1(Age)$, $f_2(EY)$, $f_3(SA)$, and $f_4(CA)$, Bayesian P-splines are utilized. In addition, we estimate the structured spatial effect $f_{str}(province)$ using the stationary Gaussian random field (GRF) approach, because the geographical x-y coordinates of every surveyed province in this study are known exactly.

4.2. Model implementation

For each data set (2000, 2006 and 2010), the STAR model (10) is fitted to a chronic disease.

We change all categorical covariates into dummy variates. For example, marital status has five categories: live with spouse; live differently with spouse; widowed; divorce; and unmarried. We use four dummy variates to represent this categorical covariate: marital statusA: live with spouse; marital statusB: live differently with spouse; marital statusC: widowed; marital statusD: divorce.

The model (10) can be implemented in `R2BayesX`, an open `R` package for STAR models. For this model, $25,000$ Markov chain Monte Carlo (MCMC) iterations were carried out after a burn in sample of $2,000$. In general, these random numbers are correlated. Thus, we store every 10th sampled parameter of the Markov chain. The posterior mean, posterior standard deviation, posterior median, and 90% and 95% credible intervals for all parameters, estimated from the posterior distributions, are used to assess the model fit.

When fitting $f_{str}(Province)$ in `R2BayesX`, a "map" argument is needed. It can be an object of class "SpatialPolygonsDataFrame" or an object of class "`bnd`." Spatial polygon data in China can be downloaded as shapefiles. Furthermore, using the function `shp2bnd()` of the `R` package – shapefiles package, the shapefiles can be changed to "`bnd`" objects, – `Chinabnd`. The class "`bnd`" is a `list()` of polygon matrices, with x- and y-coordinates of the boundary points in the first and second columns, respectively, which can be used to calculate the centroids of polygons to estimate the smooth bivariate effects of the resulting coordinates.

5. Empirical Results

We use a logistic STAR model to estimate the probability of an elderly person having a chronic disease ($y_i = 1$) versus the probability of an elderly person not having a chronic disease ($y_i = 0$). It is well known that the coefficients in a logistic regression model do not represent marginal effects, but rather log odds. It is difficult to interpret the coefficients, and this problem becomes even more complex when considering non-linear effects. Therefore, we focus on which group has a greater impact, not the extent of the effect. For example, we examine whether the prevalence of a chronic disease in elderly females is higher than that of elderly males, not the marginal effects of the prevalence of a chronic disease in elderly females and males.

From (8), the structured additive predictor η_i and probability $p_i = \Pr(y_i = 1)$ are positively related. If the coefficient is positive, η_i of the experimental group is larger than η_i of the control group. Then, $p_i = \Pr(y_i = 1)$ of the experimental group is larger, which means the experimental group is more likely to have chronic diseases. Otherwise, if the coefficient is negative, the control group is more likely to have chronic diseases. Furthermore, when all other coefficients remain unchanged, a certain coefficient becomes larger than η_i becomes larger, and the probability $p_i = \Pr(y_i = 1)$ becomes larger, that is, a larger coefficient denotes a greater effect on a chronic disease.

Table 1 displays the variables used in the models and gives their meanings and values. Table 2 compares the hypotheses and the empirical results on the reported prevalence of a chronic disease, hypertension, and heart disease.

5.1. Empirical results for chronic diseases

In Table 2, for the covariates of gender and census register, zero is not included in the 95% credible intervals in 2000, 2006, and 2010. Therefore, we find that these covariates do affect chronic diseases. In addition, the posterior means of gender and census register are positive. This indicates that in comparing elderly females and elderly males (female is 1, male is 0), and urban elderly people and rural elderly people (urban is 1, rural is 0), the reported prevalence of the former groups is higher than that of the latter groups. These results are also consistent with the hypotheses.

In general, marital status may affect the health of elderly people (Kiecolt-Glaser and Newton, 2001). However, being married does not guarantee health benefits. A decline in the quality of marriage has a negative effect on mental and physical health (Wickrama et al., 1997). We find that, marital status has a negligible effect on a chronic disease, because the 95% and 90% credible intervals of marital status include zero in all three years. Thus, we reject the hypothesis of marital status.

For the covariates smoking and drinking, zero was not included in the 95% credible intervals of smokingA (smoked previously), drinkingA (drank previously), and drinkingB (currently drink) in all three years. Therefore, we hold the hypotheses that smoking and drinking do affect chronic diseases. However, smoking and drinking have two different kinds of effects on a chronic disease. The results of

smokingA (smoked previously)/drinkingA (drank previously) are consistent with the following hypotheses: the reported prevalence in elderly people who smoked/drank previously, but no longer do so, is higher than that in elderly people who never smoke/drink. However, the results of smokingB (currently smoke) and drinkingB (currently drink) seem to be counter-intuitive: the reported prevalence in elderly people who currently smoke/drink is lower than that in those who never smoke/drink. One possible reason for this apparent contradiction in the case of drinking may be the following: moderate drinking has little effect on health, and some reports even show that moderate drinking is beneficial to our health (Fillmore et al., 2006).

In addition, as explained above, a larger coefficient denotes a greater effect on chronic diseases. From Table 2, for the fixed effects of the categorical covariates on chronic diseases, the census register has the greatest effect on chronic diseases, followed by gender, smoking, and drinking.

The nonlinear effects of the continuous covariates (age, years of education, sports activities, and cultural activities) are considerable, but they are all very small (mean values are less than 0.012) in the three years. Furthermore, the posterior means are all positive, which indicates a greater reported prevalence for older people with a higher level of education, and who do more sports activities and cultural activities. Age and education are consistent with the hypotheses, but sports activities and cultural activities are contrary to the hypotheses. One possible reason may be as follows: people who do more sports activities and cultural activities pay more attention to their health, and so are more likely to go to the hospital for an examination, and more likely to report having a chronic disease. Figure 2 gives the detailed nonlinear effects of these continuous covariates on a chronic disease, with 95% credible bands in all three years. The tails of all continuous covariates are wide because the numbers of observations in these parts are all very small.

Since a larger coefficient denotes a greater effect on chronic diseases, and the coefficient of province on chronic diseases is relatively large, thus, the spatial effect of province is critical.

In 2000, Xinjiang province's structured spatial effect is the highest, followed by Anhui, Shaanxi, Sichuan, Guangxi, Beijing, and Shanxi provinces. Shandong province's structured spatial effect is the lowest (see Figure 3(a)). In 2006, Xinjiang province's structured spatial effect is still the highest, followed by Anhui, Hubei, and Hunan provinces. Guangdong province's structured spatial effect is the lowest, and Shandong, Guangxi and Fujian provinces' structured spatial effects are very low (see Figure 3(b)). In 2010, Zhejiang province's structured spatial effect is the highest, Hunan province's structured spatial effect is the lowest. In addition, Anhui, Fujian, and Xinjiang provinces' structured spatial effects are relatively high, and Shandong and Liaoning provinces' structured spatial effects are relatively low (see Figure 3(c)).

In conclusion, we find that the province is a critical factor affecting chronic diseases, but that the high reported prevalence of chronic diseases is not only in economically developed provinces (such as Zhejiang, Beijing, Fujian), but also in the economically backward provinces with complex terrain (such as Xinjiang, Guangxi),

as shown in Figure 3. In addition, we should pay special attention to Xinjiang and Anhui provinces, because their structured spatial effects are quite high in all three years.

The high reported prevalence does not necessarily indicate that the health conditions in these provinces are bad. On the contrary, the high reported prevalence may indicate that elderly people pay more attention to their health. However, a high reported prevalence in economically backward provinces with complex terrain may indicate that the health conditions in these provinces are poor, for example, in Xinjiang province.

5.2. Empirical results for hypertension

As shown in Table 2, for the covariates of gender, census register, smokingB (currently smoke), drinkingA (drank previously), drinkingB (currently drink), age, education years, sports activities, cultural activities, and province, zero is not included in the 95% credible intervals in 2006 and 2010. Therefore, we find that they have considerable effects on hypertension, but we reject the hypothesis for marital status.

The empirical results for the fixed effects of the categorical covariates and for the nonlinear effects of the continuous covariates on hypertension are very similar to the results for chronic diseases. Thus, we do not repeat them here. The differences between general chronic diseases and hypertension are mainly reflected in the spatial effects.

Since a larger coefficient denotes a greater effect, and the coefficient of province on hypertension is relatively large, thus, the spatial effect of province on hypertension is critical.

In 2006, Hebei, Beijing, and Zhejiang provinces' structured spatial effects on hypertension are the highest, followed by Jiangsu, Shanghai and Heilongjiang. Guangxi, Yunnan, and Sichuan provinces' structured spatial effects are the lowest, although Guangdong and Liaoning provinces' structured spatial effects are also relatively low (see Figure 4(a)). In 2010, Yunnan province's structured spatial effect is the highest (see Figure 4(b)), showing a marked increase over the low level in 2006.

In conclusion, we find that the province is a critical factor affecting hypertension, but that the highest reported prevalence occurs mainly in economically developed provinces (e.g., Zhejiang, Beijing, and Guangdong), and not in economically backward provinces, as shown in Figure 4. In addition, note that Zhejiang province's structured spatial effects are quite high in both years.

The high reported prevalence in these economically developed provinces simply indicates that the elderly pay more attention to their health. However, the high reported prevalence in Yunnan province in 2010 indicates that the health conditions in this province are poor. Reports in 2013 revealed that for every 10 Kunming (Yunnan provincial capital) residents, two suffer from hypertension, but that about 70% of patients are not aware of this.

5.3. Empirical results for heart disease

As shown in Table 2, for the covariates of gender, census register, smokingA (smoked previously), drinkingB (currently drink), age, education years, sports activities, cultural activities, and province, zero is not included in the 95% and 90% credible intervals in 2006 and 2010. Therefore, we find that they have considerable effects on heart disease, but we reject the hypothesis for marital status.

The empirical results for the fixed effects of the categorical covariates and the nonlinear effects of the continuous covariates on heart disease are very similar to the results for chronic diseases. Thus, we do not repeat them here. The differences between general chronic diseases and heart disease are still mainly reflected in the spatial effects.

As explained above, a larger coefficient denotes a greater effect, and the coefficient of province on heart disease is relatively large, thus, the spatial effect of province on heart disease is critical.

In 2006, Heilongjiang province's structured spatial effect was the highest, followed by Liaoning, Beijing, Hebei, and Xinjiang. Guangdong, Guangxi, Yunnan, and Sichuan provinces' structured spatial effects were the lowest (see Figure 4(c)). In 2010, Shaanxi province's structured spatial effect was the highest, followed by Shanxi and Anhui. Shandong province's structured spatial effect was the lowest, although Xinjiang, Hunan, and Guangxi were relatively low (see Figure 4(d)).

In conclusion, we find that the province is a critical factor affecting heart disease. As shown in Figure 4, in 2006, the highest reported prevalence appeared not only in economically backward provinces (e.g., Xinjiang), but also in the economically developed provinces (e.g., Hebei, Beijing) and the moderately developed provinces (e.g., Heilongjiang, Liaoning). In contrast, in 2010, high spatial effects appeared in moderately developed provinces only (e.g., Shaanxi, Shanxi, Anhui).

5.4. Empirical results of adding time factor

Because most of the samples between surveys in 2006 and 2010 are the same, we further take the effects of time into account. And the estimation results are shown in Table 3.

Table 3 gives the posterior means, posterior standard deviations, and posterior medians for all covariates (including time) of a chronic disease, hypertension, and heart disease. In Table 3, for the covariate of time, zero is not included in the 95% credible intervals. Therefore, time do affect chronic diseases. In addition, the posterior means of time are positive. This indicates that in comparing 2006 and 2010 (2010 is 1, 2006 is 0), the reported prevalence in 2010 is higher than that in 2006.

The higher reported prevalence in 2010 does not necessarily indicate that the health conditions of the elderly people are worse. On the contrary, the high reported prevalence may indicate that elderly people pay more attention to their health with the time going.

6. Conclusions

This study applied a STAR model to determine which covariates have considerable effects on chronic diseases and specific chronic diseases (hypertension and heart disease). STAR models combine spatial effects, nonlinear effects of continuous factors, and the linear or fixed effects into a single model.

The findings are as follows: (1) the following covariates have considerable effects on chronic diseases and specific chronic diseases (hypertension and heart disease) (in descending order): census register (rural or urban), gender, smoking, drinking, province, age, cultural activities, years of education, sports activities; (2) because the 95% and 90% credible intervals of marital status include zero in all three years, thus, we reject the hypothesis for marital status; that is, the effect of marriage is negligible; (3) elderly females, urban elderly people and the elderly who smoke and drink are more likely to report having chronic diseases. In addition, the reported prevalence increases with age, education level, and participation in sports activities and cultural activities; (4) a high reported prevalence may indicate that the health conditions are bad, but may also indicate that the elderly pay more attention to their health; (5) a larger coefficient denotes a greater effect on chronic diseases, thus, province is a critical factor, but the high reported prevalence is not restricted to economically developed provinces. For chronic diseases, a high reported prevalence occurs in economically developed provinces and in economically backward provinces with complex terrain; and (6) because most of the samples between surveys in 2006 and 2010 are the same, we further take the effects of time into account, and find that time also has considerable effects.

Based on the above findings, the government should further strengthen its investment in rural areas and in economically underdeveloped provinces, such as Xinjiang province and Anhui province, as a cost-effective intervention. In addition, the government should educate the population on the harmful effects of smoking and drinking on health. Furthermore, economically developed provinces' highest structured spatial effects on hypertension do not mean that the elderly in these provinces are more likely to have hypertension. However, the government should strengthen its investment in the promotion and diagnosis of hypertension in economically backward areas (e.g., Yunnan province). In the case of heart disease, the government should strengthen its investment in provinces such as Heilongjiang province and Liaoning province.

The nonlinear effects in this study are considerable, but very small. Thus, we should build a better method to reconsider the nonlinear effects. Furthermore, we have data for three years, and conduct separate estimates for the model in each year. In future research, we will include time as a feasible covariate in the regression model to consider the impact of time on chronic diseases.

Acknowledgements

The authors appreciate the comments of anonymous referees, which improve the article greatly. They thank the China Research Center on Aging for their support in data collection and management. The first author would like to thank Hokkaido University President's Fellowship for its financial support. The work of the second author was supported in part by a Grant-in-Aid for Scientific Research (No.16K03589) from the Japan Society for the Promotion of Science (JSPS).

References

AASM, (2017). Sleep disorders affect men and women differently: Women are more likely to feel tired and depressed than men. https://www.sciencedaily.com/releases/2017/05/170523081838.htm. Accessed 23 May 2017.

BASILE, R., BENFRATELLO, L., CASTELLANI, D., (2013). Geoadditive models for regional count data: An application to industrial location. Geographical Analysis 45, pp. 28–48.

CHEN, J. J., (2005). The prevalence study on main chronic diseases and the associated behavioral risk factors among the middle and old aged people in some areas of Hubei province. Master's thesis, Wuhan University (In Chinese).

CRESSIE, N., (2015). Statistics for spatial data. Wiley, New York.

EILERS, P. H. C., MARX, B. D., (1996). Flexible smoothing with B-splines and penalties. Statistical Science 11(2), pp. 89–102.

FAHRMEIR, L., KNEIB, T., LANG, S., MARX, B., (2013). Regression: models, methods and applications. Springer, New York.

FAHRMEIR, L., KNEIB, T., LANG, S., (2004). Penalized additive regression for space-time data: A Bayesian perspective. Statistica Sinica 14, pp. 731–761.

FILLINGIM, R. B., KING, C. D., RIBEIRO-DASILVA, M. C., RAHIM-WILLIAMS, B., RILEY, J. L., (2009) Sex, gender, and pain: A review of recent clinical and experimental findings. Journal of Pain 10, pp. 447–485.

FILLMORE, K. M., KERR, W. C., STOCKWELL, T., CHIKRITZHS, T., BOSTROM, A., (2006). Moderate alcohol use and reduced mortality risk: Systematic error in prospective studies. Addiction Research and Theory 14 (2), pp. 101–132.

FREEDMAN, V. A., MARTIN, L. G., (2000). Contribution of chronic conditions to aggregate changes in old-age functioning. American Journal of Public Health 90, pp. 1755–1760.

GAO, P., LI, H. D., (2016). New characteristics of active life expectancy of the elderly in China. Advances in Aging Research 5, pp. 27–39.

GENIAUX, G., NAPOLÉONE, C., (2008). Semi-parametric tools for spatial hedonic models: An introduction to mixed geographically weighted regression and geoadditive models. In: Baranzini A, Ramirez J, Schaerer C, Thalmann P (eds) Hedonic Methods in Housing Markets. Springer, New York, pp. 101–127.

GU, Z. L., ZHANG, H, HUANG, J. P., MI, Y. P., (2009) Analysis on causes of main chronic diseases and death from 2004 to 2006 in Gangza district of Nantong city. Chinese Journal of Prevention and Control of Chronic Non-Communicable Diseases 17, pp. 89–90 (In Chinese).

HE, J., GU, D., WU, X., REYNOLDS, K., DUAN, X., YAO, C., WANG, J., CHEN, C. S., CHEN, J., WILDMAN, R. P., KLAG, M. J., WHELTON, P. K., (2005). Major causes of death among men and women in China. The New England Journal of Medicine 353, pp. 1124–1134.

JIAO, S. F., YIN, X. J., WANG, Y., XIE, J., GUO, B., (2002). Analysis on behaviors risk factors of chronic disease in residents of Beijing and study its countermeasure. China Public Health 18 (2), pp. 197–198 (In Chinese).

KAMMANN, E., WAND, M. P., (2003). Geoadditive models. Journal of the Royal Statistical Society: Series C (Applied Statistics) 52, pp. 1–18.

KIECOLT-GLASER, J. K., NEWTON, T. L., (2001). Marriage and health: His and hers. Psychological Bulletin 127, pp. 472–503.

KNEIB, T., (2006). Mixed model based inference in structured additive regression. Dissertation, Ludwig-Maximilians-Universität.

LANG, S., BREZGER, A., (2004). Bayesian P-splines. Journal of Computational and Graphical Statistics 13 (1), pp. 183–212.

LI, Z. K., ZHANG, M. R., YANG, Z., SHEN, L., MA, Y., GONG, H. Q., TIAN, R., YAN, F., (2012). The epidemiological characteristics and PYLL analysis of death of chronic diseases among residents in Kunming city in 2007-2010. Chinese Primary Health Care 26, pp. 84–86 (In Chinese).

LIN, H., ZHANG, T. H., YANG, H., WANG, C. B., DUAN, L. R., ZHANG, X. P., GONG, G. C., (2002). Analysis of the chronic disease and its influencing factors among 895 elders in Beijing. Chinese Journal of Prevention and Control of Chronic Non-Communicable Diseases 10 (6), pp. 270–272 (In Chinese).

SAULEAU, E. A., HENNERFEIND, A., BUEMI, A., HELD, L., (2007). Age, period and cohort effects in Bayesian smoothing of spatial cancer survival with geoadditive models. Statistics in Medicine 26, pp. 212–229.

THORPE, K. E., HOWARD, D. H., (2006). The rise in spending among medicare beneficiaries: The role of chronic disease prevalence and changes in treatment intensity. Health Affairs - Web Exclusive 25, pp. w378–w388.

UMLAUF, N., ADLER, D., KNEIB, T., LANG, S., ZEILEIS, A., (2015). Structured additive regression models: An R interface to BayesX. Journal of Statistical Software 63, pp. 1–46.

VOGELI, C., SHIELDS, A. E., LEE, T. A., GIBSON, T. B., MARDER, W. D., WEISS, K. B., BLUMENTHAL, D., (2007). Multiple chronic conditions: Prevalence, health consequences, and implications for quality, care management and costs. Journal of General Internal Medicine 22, pp. 391–395.

WAND, H., WHITAKER, C., RAMJEE, G., (2011). Geoadditive models to assess spatial variation of HIV infections among women in local communities of Durban, South Africa. International Journal of Health Geographics 10, p. 28.

WHO, (2005). Preventing chronic diseases: A vital investment. Technical report, World Health Organization.

WICKRAMA, K., LORENZ, F. O., CONGER, R. D., ELDER JR, G. H., (1997). Marital quality and physical illness: A latent growth curve analysis. Journal of Marriage and the Family 59, pp. 143–155.

WOLFF, J. L., STARFIELD, B., ANDERSON, G., (2002). Prevalence, expenditures and complications of multiple chronic conditions in the elderly. Archives of Internal Medicine 162, pp. 2269–2276.

WOOLF, S. H., ARON, L., DUBAY, L., SIMON, S. M., ZIMMERMAN, E., LUK, K. X., (2015). How are income and wealth linked to health and longevity? Urban Institute.

YE, S. M., (2013). Analysis on non-communicable chronic diseases status and its influencing factors among adults in Gong Zhu-ling of Jilin province in 2012. Master's thesis, Jilin University (In Chinese).

YIN, S. Q., (2011). Analysis of chronic disease prevalence and influencing factors of the elderly in Chongwen district of Beijing. Chinese Primary Health Care 25 (4): 43–46 (In Chinese).

ZHAO, J., WAN, R., LIU, Z., WAN, Q. Q., ZHANG, Q., RUAN, Y., (2015). Prevalence and influencing factors of chronic diseases in residents aged 15 years and above in urban and rural area of Yunnan province, 2010-2012. Chinese Journal of Public Health 31 (7), pp. 878–882 (In Chinese).

ZHEN, S., (2010). Study of the current situation and influencing factors of chronic illnesses in Gansu. Master's thesis, Lanzhou University (In Chinese).

APPENDIX

Table 1: Variables in the models

Variable	Description	Value
Chronic Disease	Whether one has a chronic disease or not.	Having chronic disease is 1, else is 0.
Hypertension	Whether one has hypertension or not.	Having hypertension is 1, else is 0.
Heart Disease	Whether one has heart disease or not.	Having heart disease is 1, else is 0.
Time	Time of investigation.	2010 is 1; 2006 is 0.
Gender	Gender of the elderly person with categories 'male' and 'female'.	Female is 1; Male is 0.
Census Register	Census register with categories 'urban' and 'rural'.	Urban is 1; Rural is 0.
Marital Status	Marital status with categories 'live with spouse', 'live differently with spouse', 'widowed', 'divorce' and 'unmarried'	Using four dummy variables: Marital StatusA: live with spouse; Marital StatusB: live differently with spouse; Marital StatusC: widowed; Marital StatusD: divorce. Yes is 1, no is 0.
Smoking	Smoking condition with categories 'smoked previously', 'currently smoke' and 'never smoke'.	Using two dummy variables: SmokingA: smoked previously; SmokingB: currently smoke; Yes is 1, no is 0.
Drinking	Drinking condition with categories 'drank previously', 'currently drink' and 'never drink'.	Using two dummy variables: DrinkingA: drank previously; DrinkingB: currently drink; Yes is 1, no is 0.
Age	Age of the elderly people in years.	Continuous covariate, minimum value is 60.
Education Years	One's education years.	Continuous covariate, minimum value is 0.
Sports Activities	Number of sports activities one takes part in.	Continuous covariate, value changes from 0 to 5.
Cultural Activities	Number of cultural activities one takes part in.	Continuous covariate, value changes from 0 to 10.
Province	Province where one lives in.	Spatial covariate, integer, value changes from 1 to 20.

Table 2: Comparison of Hypotheses and Empirical Results on the Reported Prevalence

Variable	Hypotheses	Empirical Results (Posterior Mean)						
		Chronic Disease			Hypertension		Heart Disease	
		2000	2006	2010	2006	2010	2006	2010
Gender	considerable influence females > males	0.340**	0.344**	0.247**	0.158**	0.093**	0.489**	0.372**
Census Register	critical influence urban > rural	0.642**	0.631**	0.544**	0.599**	0.462**	0.872**	0.368**
Marital StatusA	considerable influence	-0.037	0.115	-0.100	-0.036	-0.033	0.342	0.111
Marital StatusB	divorced/widowed >	-0.101	0.258	-0.378	-0.045	-0.392	0.268	-0.024
Marital StatusC	other types	-0.034	0.098	-0.213	-0.028	-0.018	0.258	0.056
Marital StatusD		-0.031	0.181	-0.085	-0.024	-0.325	0.193	-0.058
SmokingA	considerable influence	0.430**	0.261**	0.267**	-0.051	-0.009	0.116*	0.117*
SmokingB	smoke > never smoke	0.037	-0.122**	-0.022	-0.237**	-0.178**	-0.003	-0.180**
DrinkingA	considerable influence	0.233**	0.274**	0.181**	0.173**	0.221**	0.018	0.057
DrinkingB	drink > never drink	-0.294**	-0.175**	-0.239**	-0.156**	-0.087**	-0.306**	-0.253**
Age	non-negligible influence older > younger	0.005**	0.005**	0.011**	0.006**	0.003**	0.010**	0.014**
Education Years	non-negligible influence longer > shorter	0.003**	0.005**	0.002**	0.004**	0.002**	0.003**	0.002**
Sports Activities	non-negligible influence more > less	0.003**	0.003**	0.003**	0.003**	0.003**	0.003**	0.004**
Cultural Activities	non-negligible influence more > less	0.004**	0.003**	0.004**	0.005**	0.005**	0.003**	0.003**
Province	critical influence developed > backward	0.416**	0.375**	0.054**	0.144**	0.201**	0.307**	1.975**

[1] "$**$" and "$*$" denote that zero is not included in the 95% and 90% credible intervals, respectively.

Table 3: Posterior Estimates of the Parameters for Chronic Disease (Add Time)

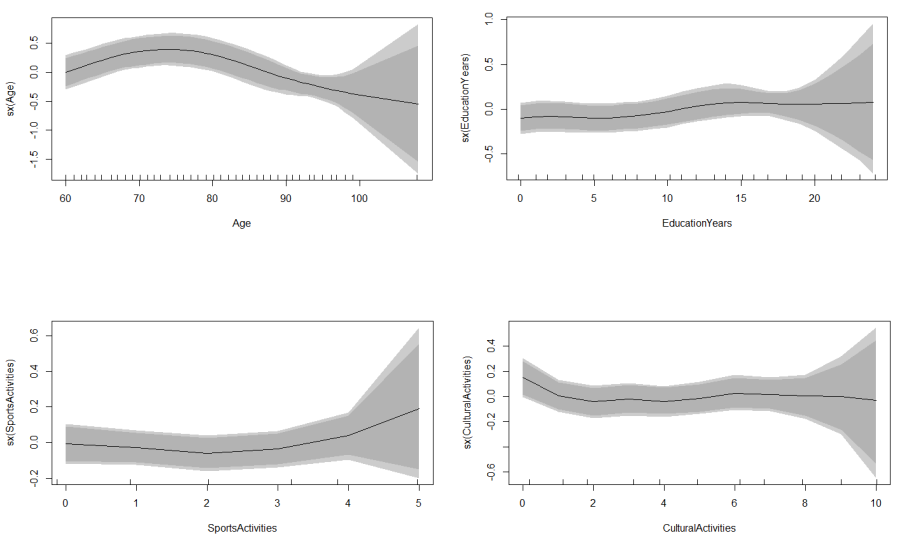
	Chronic Disease			Hypertension			Heart Disease		
Variable	Mean	Sd	Median	Mean	Sd	Median	Mean	Sd	Median
Fixed effects:									
(Intercept)	0.789	0.173	0.784	-0.911	0.191	-0.906	-2.209	0.271	-2.191
Time	0.129	0.027	0.129**	0.329	0.027	0.330**	0.088	0.030	0.088**
Gender	0.292	0.033	0.292**	0.129	0.034	0.129**	0.431	0.036	0.431**
CR	0.590	0.033	0.590**	0.549	0.031	0.549**	0.721	0.035	0.721**
MSA	0.015	0.105	0.017	-0.054	0.123	-0.056	0.193	0.162	0.192
MSB	0.006	0.150	0.003	-0.280	0.166	-0.281*	-0.031	0.211	-0.029
MSC	-0.048	0.107	-0.047	-0.051	0.125	-0.050	0.120	0.163	0.119
MSD	0.067	0.166	0.065	-0.204	0.174	-0.201	0.053	0.219	0.052
SmokingA	0.260	0.045	0.261**	-0.020	0.043	-0.021	0.113	0.047	0.114**
SmokingB	-0.076	0.035	-0.077**	-0.207	0.037	-0.206**	-0.090	0.042	-0.089**
DrinkingA	0.226	0.045	0.225**	0.181	0.042	0.181**	0.022	0.046	0.022
DrinkingB	-0.201	0.034	-0.201**	-0.125	0.036	-0.124**	-0.280	0.041	-0.280**
Nonlinear effects:									
sx(Age)	0.003	0.004	0.002**	0.004	0.005	0.002**	0.011	0.012	0.007**
sx(EY)	0.003	0.006	0.002**	0.002	0.003	0.001**	0.002	0.002	0.001**
sx(CA)	0.003	0.006	0.002**	0.002	0.003	0.002**	0.002	0.003	0.002**
sx(SA)	0.002	0.003	0.001**	0.003	0.007	0.001**	0.002	0.004	0.001**
sx(Province)	0.118	0.057	0.105**	0.106	0.052	0.094**	0.356	0.181	0.312**

[1] sx(·) corresponds to the smooth functions in STAR models.

[2] CR, MSA, MSB, MSC, MSD, CA, EY and SA are the abbreviation of Census Register, Marital StatusA, Marital StatusB, Marital StatusC, Marital StatusD, Cultural Activities, Education Years and Sports Activities, respectively.

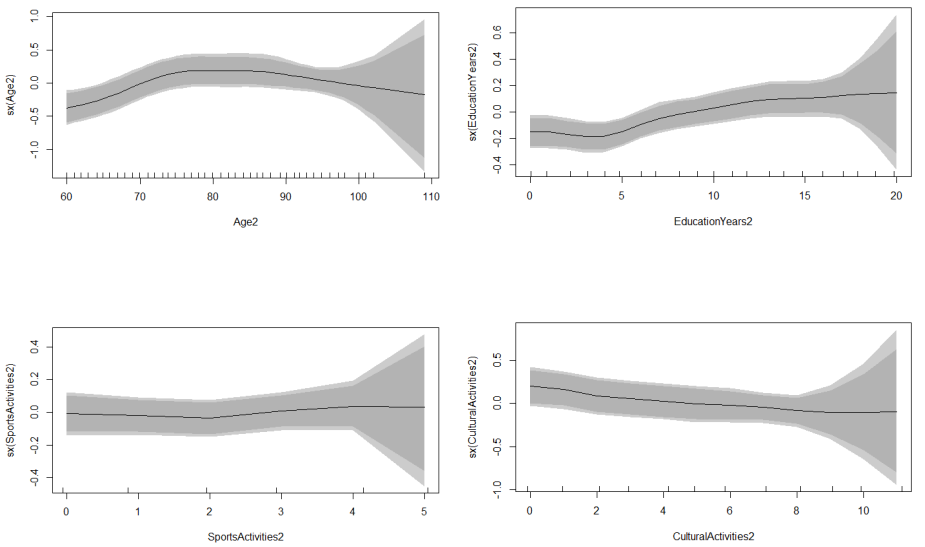
[3] "$**$" and "$*$" denote that zero is not included in the 95% and 90% credible intervals, respectively. Mean, Sd and Median are posterior mean, posterior standard deviation and posterior median, respectively.

Figure 1. 20 selected provinces, municipalities and autonomous regions of China mainland

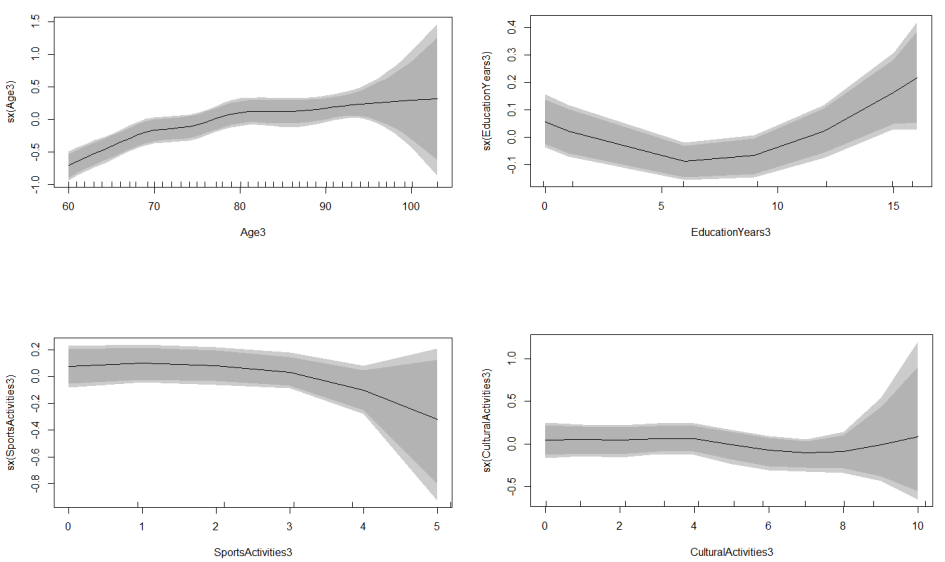


(a) 2000

Figure. 2. Effects of smooth terms on a chronic disease with 90% and 95% credible bands in 2000, 2006 and 2010

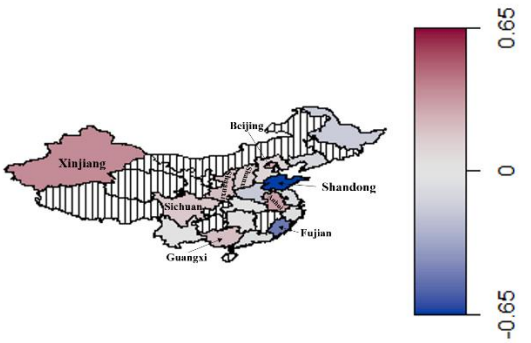


(b) 2006

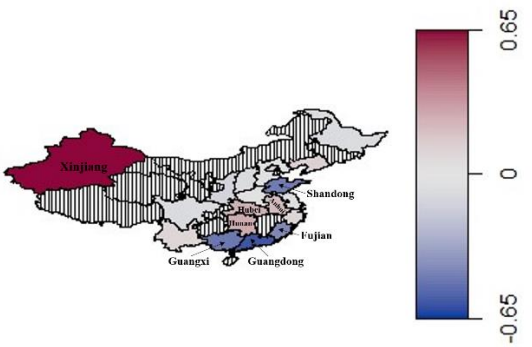


(c) 2010

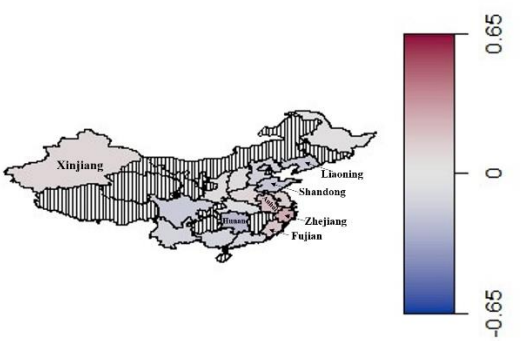
Figure. 2. Effects of smooth terms on a chronic disease with 90% and 95% credible bands in 2000, 2006 and 2010 (cont.)



(a) 2000



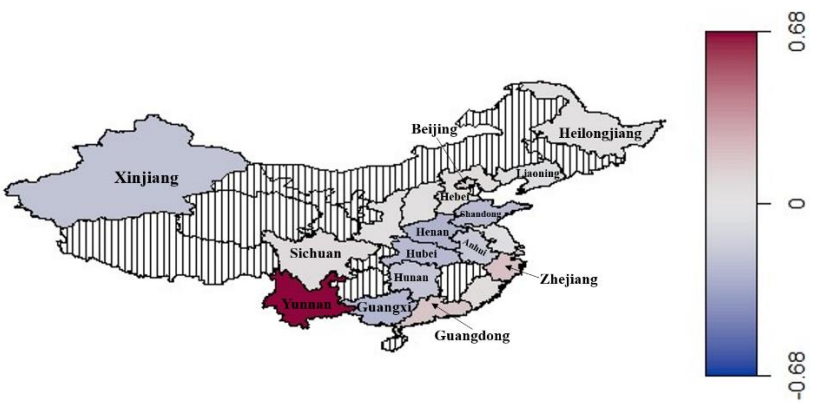
(b) 2006



(c) 2010

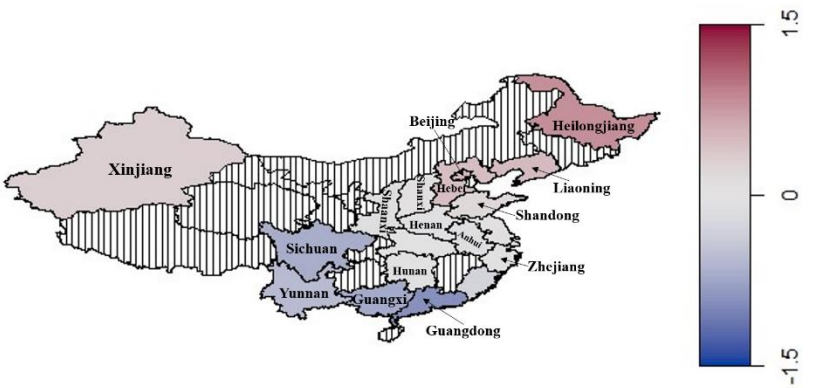
Figure 3. Structured spatial effect on a chronic disease in 2000, 2006 and 2010

(a) Hypertension-2006



(b) Hypertension-2010

Figure 4. Structured spatial effect on hypertension and heart disease in 2006 and 2010



(c) Heart disease-2006

(d) Heart disease-2010

Figure 4. Structured spatial effect on hypertension and heart disease in 2006 and 2010 (cont.)

STATISTICS IN TRANSITION new series, December 2018
Vol. 19 No. 4, pp. 671–692, DOI 10.21307/stattrans-2018-035

LINDLEY PARETO DISTRIBUTION

Halim Zeghdoudi[1], **Lazri Nouara**[2], **Djabrane Yahia**[3]

ABSTRACT

In this paper, we introduce a new Lindley Pareto distribution, which offers a more flexible model for modelling lifetime data. Some of its mathematical properties like density function, cumulative distribution, mode, mean, variance, and Shannon entropy are established. A simulation study is carried out to examine the bias and mean square error of the maximum likelihood estimators of the unknown parameters. Three real data sets are fitted to illustrate the importance and the flexibility of the proposed distribution.

Key words: T-X family, Lindley distribution, Pareto distribution.

1 Introduction

Statistical distributions (Lifetime distributions) are commonly applied to describe real world phenomena and are most frequently used in many applied sciences such as reliability, engineering, actuarial sciences, demography, economics, hydrology, biological studies, insurance, medicine and finance. Recently this issue has received much attention from researchers and practitioners. The quality and effectiveness of the procedures used in a statistical analysis are determined by the assumed probability distribution. Recently, one parameter Lindley distribution has attracted the researchers for its use in stress-strength reliability modelling, and it has been observed in several papers that this distribution has performed excellently. The Lindley distribution was introduced by Lindley (1958) as a new distribution useful to analyze lifetime data. Sankaran (1970) introduced the discrete Poisson-Lindley distribution by combining the Poisson and Lindley distributions. Many generalizations of the Lindley distribution have been proposed in recent years. Asgharzadeh et al. (2013), Ghitany et al. $(2008a, 2008b)$ rediscovered and

[1]LaPS laboratory, Badji-Mokhtar University, Box 12, Annaba, 23000, Algeria. E-mail: halim.zeghdoudi@univ-annaba.dz

[2]LaPS laboratory, Badji-Mokhtar University, Box 12, Annaba, 23000, Algeria. E-mail: lazrinouara.actuariat@yahoo.fr

[3]Laboratory of Applied Mathematics, Mohamed Khider University, Box 145, Biskra, 07000, Algeria. E-mail: yahia.dj@yahoo.fr

studied the new generalizations of Lindley distribution, what they derived is known as Zero-truncated Poisson- Lindley and Pareto Poisson-Lindley distributions. There still remain many important problems where the real data does not follow any of the existing probability distributions. Considerable effort has been expended in the development of large classes of new probability distributions along with relevant statistical methodologies.

Furthermore, Pareto distribution was pioneered by V. Pareto (1896) to explore the unequal distribution of wealth. It is widely used in actuarial science. (e.g. reinsurance) because of its heavy tail properties. To add flexibility to the Pareto distribution, various generalizations of the distribution have been derived, including the generalized Pareto distribution (Pickands, 1975), the beta-Pareto distribution (Akinsete et al., 2008), and the beta generalized Pareto distribution (Mahmoudi, 2011).

The mixing method is one of the most important ideas for obtaining a new distribution. For example, Sharma and Shanker (2013) used a mixture of exponential (θ) and gamma $(2,\theta)$ to create a two-parameter Lindley distribution. Another example includes Zakerzadeh and Dolati (2010), who used gamma (α,θ) and gamma $(\alpha+1,\theta)$ to create a generalized Lindley distribution. Recently, Zeghdoudi and Nedjar $(2016a, 2016b)$ introduced a new distribution, named gamma Lindley distribution, based on mixtures of gamma $(2,\theta)$ and one-parameter Lindley distributions.

Gomes-Silva et al. (2017) introduce a new generator of continuous distributions with one extra positive parameter called the odd Lindley-G family. Some special cases are given (Odd Lindley Weibull, Odd Lindley Kumaraswamy, Odd Lindley half-logistic and Odd Lindley Burr XII), where the hazard rate function of the Odd Lindley Burr XII distribution can be constant, increasing, decreasing, unimodal or bathtub shape. For more details on this last distribution function we refer the reader to Abouelmagd et al. (2018).

In addition, the cumulative distribution function (cdf) of the T-X family of distributions defined by Alzaatreh et al. (2013) is given by

$$G(x)=\int_0^{W(F(x))} r(t)\,dt, \tag{1}$$

where $W(F(x))$ satisfies the following conditions:

- $W(F(x))\in[a,b]$,

-$W(F(x))$ is differentiable and monotonically non-decreasing,

- $W(F(x))\to a$ as $x\to-\infty$ and $W(F(x))\to b$ as $x\to\infty$.

In this paper, we propose a new wider class of continuous distributions called the Lindley Pareto (LP for short) by taking $W(F(x)) = \frac{F(x)}{1-F(x)}$ and $r(t) = \frac{\theta^2}{1+\theta}(1+t)\exp(-\theta t)$, $x > 0$, $\theta > 0$, where $F(x)$ corresponding to Pareto distribution: $F(x) = 1 - \left(\frac{\alpha}{x}\right)^k$, $x > \alpha$. Its cdf is given by

$$G(x) = 1 - \frac{\left(\alpha^k + x^k\theta\right)}{(\theta+1)\alpha^k}\exp\left(-\theta\left(\frac{x^k}{\alpha^k} - 1\right)\right), \tag{2}$$

with corresponding density

$$g(x) = \frac{k\theta^2 e^{\theta} x^{2k-1}}{(\theta+1)\alpha^{2k}}\exp\left(-\theta\left(\frac{x}{\alpha}\right)^k\right), x > \alpha. \tag{3}$$

We can see the plots of the density function and the distribution function of LP distribution for some parameter values in Appendix 1. We refer to the cdf in equation (1) as Lindley Pareto (LP) distribution with parameters θ, α, k, which we denote by $LP(\theta, \alpha, k)$. The objective of this work is to study some mathematical properties of the Lindley Pareto model with the hope that it will attract wider applications in reliability, engineering and other areas of research.

The LP distribution is motivated by the following: the LP distribution use may be restricted to the tail of a distribution, but it is easy to apply. The formulas of the mean, variance, mean deviation, entropy and the quantile function are simple in form and may be used as quick approximations in many cases. Also, the LP distribution can be viewed as a special case of odd Lindley-G family introduced by Gomes-Silva et al.(2017). Also, this new distribution has advantages including a number of parameters (three) which we can modelled physical phenomena inspired in Cooray (2006). Furthermore, LP distribution can be used quite effectively in analyzing many real lifetime data sets: application to waiting times in a queue, Wheaton River Data and application to bladder cancer patients. Moreover, the actuarial literature has discussed hundreds of univariate continuous distributions, of which log-normal, Weibull, multi-parameter Pareto, gamma distributions as well as others.

The remainder of the article is unfolded as follows: in Section 2, various properties of LP distribution are examined, including survival and hazard functions, reliability, mean deviation, entropy and quantile function. The model parameters are estimated via the maximum likelihood estimates (MLEs) and some simulations are proposed in Section 3. In Section 4, the impor-

tance and potentiality of LP distribution are shown using three real lifetime data sets. Finally, some concluding notes are provided in Section 5.

2 Main properties

2.1 Survival and hazard functions

The survival and hazard functions corresponding to the cdf defined in (1) are given by

$$S(x) = 1 - G(x) = \frac{\left(\alpha^k + x^k\theta\right)}{(\theta+1)\,\alpha^k} \exp\left(-\theta\left(\frac{x^k}{\alpha^k} - 1\right)\right)$$

and

$$h(x) = \frac{k\theta^2 x^{2k-1}}{\alpha^k(\theta x^k + \alpha^k)}.$$

2.2 Reliability

The measure of reliability has many applications, especially in the area of engineering. The component fails at the instant that the random stress X_2 applied to it exceeds the random strength X_1, and the component will function satisfactorily whenever $X_1 > X_2$. Hence, $R = P[X_2 < X_1]$ is a measure of component reliability. We derive the reliability R when X_1 and X_2 have independent $LP(\theta_1, \alpha, k)$ and $LP(\theta_2, \alpha, k)$ distributions. The reliability is defined by

$$R = \int_0^\infty g_1(x)\, G_2(x)\, dx = \sum_{i,j,k,l=0} \frac{p_{i,j}(\theta_1)\, q_{k,l}(\theta_2)}{i+j+k+l+2},$$

where

$$p_{i,j}(\theta_1) = \frac{(-1)^j\, \theta_1^{2+j}\Gamma(i+j+3)}{i!\,j!\,(\theta_1+1)\,\Gamma(j+3)}$$

and

$$q_{k,l}(\theta_2) = \frac{(-1)^l\, \theta_2^{2+l}\Gamma(k+l+3)}{k!\,l!\,(\theta_2+1)\,(k+l+1)\,\Gamma(l+3)}.$$

2.3 Mean deviations

The deviation from the mean and the median are used to measure the dispersion and spread in a population from the centre. If the median is denoted by M, then the mean deviation from the mean, $D(\mu)$, and the mean deviation

from the median, $D(M)$, can be written as

$$D(\mu)=\int_{\alpha}^{\infty}|x-\mu|\,g(x)\,dx=2\mu G(\mu)-2\int_{\alpha}^{\mu}xg(x)\,dx,$$

$$D(M)=\int_{\alpha}^{\infty}|x-M|\,g(x)\,dx=\mu-2\int_{\alpha}^{M}xg(x)\,dx.$$

Consider the integral

$$\int_{\alpha}^{b}xg(x)\,dx=\int_{\alpha}^{b}\frac{k\theta^{2}e^{\theta}}{(\theta+1)\,\alpha^{2k}}x^{2k}\exp\left(-\theta\left(\frac{x}{\alpha}\right)^{k}\right)dx=\left(-\frac{e^{\theta}}{(\theta+1)}\frac{\alpha\Gamma(\frac{2k+1}{k},\theta\frac{x^{k}}{\alpha^{k}})}{\theta^{\frac{1}{k}}}\right|_{a}^{b}$$

we obtain,

$$\begin{aligned}D(\mu)&=2\mu G(\mu)-\int_{\alpha}^{\mu}xg(x)\,dx\\&=2\mu G(\mu)-\frac{\alpha e^{\theta}}{(\theta+1)\,\theta^{\frac{1}{k}}}\left(\Gamma(\frac{1}{k}+2,\theta)-\Gamma(\frac{1}{k}+2,\theta\left(\frac{\mu}{\alpha}\right)^{k})\right),\end{aligned}$$

$$\begin{aligned}D(M)&=\mu-2\int_{\alpha}^{M}xg(x)\,dx\\&=\mu-2\frac{\alpha e^{\theta}}{(\theta+1)\,\theta^{\frac{1}{k}}}\left(\Gamma(\frac{1}{k}+2,\theta)-\Gamma(\frac{1}{k}+2,\theta\left(\frac{M}{\alpha}\right)^{k})\right).\end{aligned}$$

2.4 Entropy

The entropy of a random variable X is a measure of variation of uncertainty (see, Rényi, 1961), that of the LP distribution is given by

$$I_{R}(s)=\frac{1}{1-s}\ln\left(\frac{k^{s}\theta^{2s}e^{s\theta}}{\theta^{\frac{1-s}{k}}s^{\frac{2ks-s+1}{k}}(\theta+1)^{s}\alpha^{s-1}}\frac{\Gamma\left(\frac{2ks-s+1}{k},\theta s\right)}{k}\right)\ s>0,s\neq1.$$

Shannon entropy (Shannon, 1948) for a random variable X with density $g(x)$ is defined as $E\{-\ln(g(x))\}$.

$$E\{-\ln(g(X))\}=\ln k+2\ln\theta+\theta-\ln(\theta+1)-2k\ln\alpha+2kE(\ln x)-\frac{\theta}{\alpha^{k}}E\left(x^{k}\right),$$

$$E\left\{-\ln\left(g\left(X\right)\right)\right\}=\theta+\ln k+2\ln\theta-\ln\left(\theta+1\right)+\frac{2\left(1+Ei(\theta)\,e^{-\theta}\right)-\left(\theta^{2}+2\theta+2\right)}{\left(\theta+1\right)}.$$

where, *Ei* is the exponential integral function.

2.5 Quantile function

The quantile function of the LP distribution X is

$$x_{\gamma}=\alpha\left(-\frac{1}{\theta}-\frac{1}{\theta}\mathrm{LAMBERTW(X)}\left(-1,(\gamma-1)\left(\theta+1\right)e^{-\theta-1}\right)\right)^{\frac{1}{k}},\ 0<\gamma<1, \tag{4}$$

where $\theta,\alpha,k>0$ and $\mathrm{LAMBERTW(X)}$ denotes the negative branch of the $\mathrm{LAMBERTW(X)}$ function ($W(z)\exp\left(W(z)\right)=z$, where z is a complex number). For more details we refer the reader to Lazri and Zeghdoudi (2016).

3 Estimation and Simulation

3.1 Maximum Likelihood Estimates (ML)

Let $X_i\sim LP(\theta,\alpha,k)$, $i=1,...,n$ be n random variables. The $\ln$-likelihood function, $\ln l(x_i;\theta,\alpha,k)$ is:

$$L\left(\Theta\right)=\ln l(x;\theta,\alpha,k)=n\ln k+2n\ln\theta+n\theta-2kn\ln\alpha-n\ln\left(\theta+1\right)+(2k-1)\sum_{i=1}^{n}\ln x_i-\theta\sum_{i=1}^{n}\left(\frac{x^{k}}{\alpha^{k}}\right).$$

To simplify, we assume that α is known, the derivatives of $L\left(\Theta\right)$ with respect to θ and k are:

$$\frac{dL\left(\Theta\right)}{d\theta}=\frac{2n}{\theta}+n-\frac{n}{\left(\theta+1\right)}-\frac{1}{\alpha^{k}}\sum_{i=1}^{n}x_i^{k}, \tag{5}$$

$$\frac{dL\left(\Theta\right)}{dk}=\frac{n}{k}-2n\ln\alpha\ +2\sum_{i=1}^{n}\ln x_i-\frac{\theta}{\alpha^{k}}\sum_{i=1}^{n}x_i^{k}\ln x_i+\frac{\theta\ln\alpha}{\alpha^{k}}\sum_{i=1}^{n}x_i^{k}. \tag{6}$$

The two equations (5) and (6) cannot be solved directly, we must used the Fisher scoring method. We have

$$\left[\begin{array}{cc}\frac{\partial^{2}L(\Theta)}{\partial\theta^{2}} & \frac{\partial^{2}L(\Theta)}{\partial\theta\partial k}\\ \frac{\partial^{2}L(\Theta)}{\partial k\partial\theta} & \frac{\partial^{2}L(\Theta)}{\partial k^{2}}\end{array}\right]_{\substack{\hat{\theta}=\theta_0\\ \hat{k}=k_0}}\left[\begin{array}{c}\hat{\theta}-\theta_0\\ \hat{k}-k_0\end{array}\right]=\left[\begin{array}{c}\frac{dL(\Theta)}{d\theta}\\ \frac{dL(\Theta)}{dk}\end{array}\right]_{\substack{\hat{\theta}=\theta_0\\ \hat{k}=k_0}}, \tag{7}$$

where,

$$\frac{\partial^2 L(\Theta)}{\partial \theta^2} = -\frac{2n}{\theta^2} + \frac{n}{(\theta+1)^2},$$

$$\frac{\partial^2 L(\Theta)}{\partial k^2} = \frac{-n}{k^2} - \theta \sum_{i=1}^{n} \left(\frac{x_i^k}{\alpha^k}\right) \ln^2 \frac{x_i}{\alpha},$$

and

$$\frac{\partial^2 L(\Theta)}{\partial \theta \partial k} = \frac{\partial^2 L(\Theta)}{\partial k \partial \theta} = -\sum_{i=1}^{n} \left(\ln \frac{x_i}{\alpha}\right) \left(\frac{x_i^k}{\alpha^k}\right).$$

The equation (7) can be solved iteratively where θ_0, k_0 are the initial values of θ, k.

Existence and uniqueness of the MLE's

Lemma 1. For any given $\eta > 0$, there exists a compact subset $K \equiv K(\eta) \subset (0,\infty) \times (0,\infty)$ such that

$$\{(\theta,k) : L(\Theta) \geq -\eta\} \subset K. \qquad (*)$$

Theorem 2. Suppose that $X_i \sim LP(\theta,\alpha,k)$, $i = 1,...,n$, then the MLEs of parameters θ and k of Pareto Lindley distribution uniquely exist.

Proof. We need only to show that the MLEs of parameters θ and k uniquely exist. According to the results of Mäkeläinen et al. (1981), in order to show the existence and uniqueness of the MLEs of θ and k , it is sufficient to verify the following two conditions:
i) For any given $\eta > 0$, $(*)$ holds.
ii) The Hessian matrix of $L(\Theta)$ is negative definite at every point $(\theta,k) \in (0,\infty) \times (0,\infty)$. Condition i is certainly satisfied by Lemma 1. Therefore, to prove the theorem, we need only to show ii. Then,

$$x^t H x = -2x_1 x_2 \sum_{i=1}^{n} \left(\ln \frac{x_i}{\alpha}\right) \left(\frac{x_i^k}{\alpha^k}\right) + \left(-\frac{2n}{\theta^2} + \frac{n}{(\theta+1)^2}\right) x_1^2 + \left(\frac{-n}{k^2} - \theta \sum_{i=1}^{n} \left(\frac{x_i^k}{\alpha^k}\right) \ln^2 \frac{x_i}{\alpha}\right) x_2^2,$$

where $x^t = (x_1 \;\; x_2)$ and $H = \begin{bmatrix} \frac{\partial^2 L(\Theta)}{\partial \theta^2} & \frac{\partial^2 L(\Theta)}{\partial \theta \partial k} \\ \frac{\partial^2 L(\Theta)}{\partial k \partial \theta} & \frac{\partial^2 L(\Theta)}{\partial k^2} \end{bmatrix}$, we can check that $x^t H x \leq 0$, (H is negative definite).

3.2 Simulation

In this section, we investigate the behaviour of the ML estimators for a finite sample size (n). A simulation study consisting of the following steps is being carried out for each quadruplets (θ,α,k,n), where $\theta = 0.5, 1, 2$, $\alpha = 0.3, 0.5, 1$, $k = 0.75, 1, 2$ and $n = 30, 50, 100$.

- Choose the initial values of θ_0, α_0, k_0 for the corresponding elements of the parameter vector $\Theta = (\theta, \alpha, k)$ to specify LP(θ,α,k) distribution;
- choose sample size n;
- generate N independent samples of size n from LP(θ,α,k);
- compute the ML estimate $\hat{\Theta}_n$ of Θ_0 for each of the N samples;
- compute the mean of the obtained estimators over all N samples,

$$average\ bias\,(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\Theta}_i - \Theta_0\right),$$

and the average square error

$$MSE\,(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\Theta}_i - \Theta_0\right)^2, \text{ see Tables 1and 2.}$$

Table 1. *Average bias* of the simulated estimates

	$\theta=0.75 \quad \alpha=0.3 \quad k=1.5$			$\theta=1.25 \quad \alpha=0.3 \quad k=2$		
	$bias(\theta)$	$bias(\alpha)$	$bias(k)$	$bias(\theta)$	$bias(\alpha)$	$bias(k)$
n=30	0.2034	0.0192	−0.0768	0.3261	0.0071	0.0210
n=50	0.0788	0.0087	0.02108	0.1460	0.0040	0.0589
n=100	0.0653	0.0058	−0.0066	0.0894	0.0022	−0.0117
	$\theta=1 \quad \alpha=1.25 \quad k=1.5$			$\theta=1 \quad \alpha=2 \quad k=5$		
	$bias(\theta)$	$bias(\alpha)$	$bias(k)$	$bias(\theta)$	$bias(\alpha)$	$bias(k)$
n=30	0.5084	0.0494	−0.1130	0.2532	0.0238	−0.0237
n=50	0.1784	0.0269	−0.0220	0.1130	0.0165	0.0132
n=100	0.1048	0.0167	−0.0326	0.0993	0.0066	−0.0567
	$\theta=1.5 \quad \alpha=1 \quad k=1.25$			$\theta=2 \quad \alpha=3 \quad k=1.25$		
	$bias(\theta)$	$bias(\alpha)$	$bias(k)$	$bias(\theta)$	$bias(\alpha)$	$bias(k)$
n=30	0.5046	0.0239	0.1046	0.2366	$2.05867\ 10^{-3}$	$1.6487\ 10^{-2}$
n=50	0.3976	0.0117	0.0826	0.0323	$1.5698\ 10^{-3}$	$6.2404\ 10^{-3}$
n=100	0.2004	0.0073	0.0095	0.0789	$3.7259\ 10^{-5}$	$1.9747\ 10^{-3}$
	$\theta=4 \quad \alpha=3 \quad k=3$			$\theta=1.5 \quad \alpha=5 \quad k=7$		
	$bias(\theta)$	$bias(\alpha)$	$bias(k)$	$bias(\theta)$	$bias(\alpha)$	$bias(k)$
n=30	1.4481	0.0094	0.5102	0.4280	0.0251	−0.3361
n=50	0.7441	0.0071	0.5010	0.2127	0.0136	−0.0616
n=100	0.6058	0.0033	0.1447	0.0499	0.0069	0.2307
	$\theta=0.5 \quad \alpha=0.3 \quad k=0.9$			$\theta=1 \quad \alpha=0.8 \quad k=0.5$		
	$bias(\theta)$	$bias(\alpha)$	$bias(k)$	$bias(\theta)$	$bias(\alpha)$	$bias(k)$
n=30	0.3240	0.0668	−0.0991	0.1971	0.0896	0.0151
n=50	0.1088	0.0394	0.00431	0.1445	0.0680	0.0057
n=100	0.0520	0.0186	−0.0046	0.0521	0.0376	0.0091
	$\theta=0.75 \quad \alpha=0.5 \quad k=1.25$			$\theta=3 \quad \alpha=1.5 \quad k=2$		
	$bias(\theta)$	$bias(\alpha)$	$bias(k)$	$bias(\theta)$	$bias(\alpha)$	$bias(k)$
n=30	0.155	0.0355	−0.0150	1.5423	0.0097	0.1401
n=50	0.1595	0.0258	−0.0374	0.9111	0.0066	0.0729
n=100	0.0827	0.0128	−0.0172	0.5035	0.0032	0.0244

Table 2. *Average MSE* of the simulated estimates

	$\theta=0.75 \quad \alpha=0.3 \quad k=1.5$			$\theta=1.25 \quad \alpha=0.3 \quad k=2$		
	$MSE(\theta)$	$MSE(\alpha)$	$MSE(k)$	$MSE(\theta)$	$MSE(\alpha)$	$MSE(k)$
n=30	0.0414	$3.6786\ 10^{-4}$	$5.9026\ 10^{-3}$	0.1063	$5.0350\ 10^{-5}$	$4.4275\ 10^{-4}$
n=50	$6.2155\ 10^{-3}$	$7.4996\ 10^{-5}$	$4.4426\ 10^{-4}$	$2.1316\ 10^{-2}$	$1.5794\ 10^{-5}$	$3.3969\ 10^{-3}$
n=100	$4.2580\ 10^{-3}$	$3.3405\ 10^{-5}$	$4.3903\ 10^{-5}$	$7.9979\ 10^{-5}$	$5.0401\ 10^{-6}$	$1.3671\ 10^{-4}$
	$\theta=1 \quad \alpha=1.25 \quad k=1.5$			$\theta=1 \quad \alpha=2 \quad k=5$		
	$MSE(\theta)$	$MSE(\alpha)$	$MSE(k)$	$MSE(\theta)$	$MSE(\alpha)$	$MSE(k)$
n=30	0.2584	$2.4364\ 10^{-3}$	$1.2763\ 10^{-2}$	$6.4092\ 10^{-2}$	$5.6487\ 10^{-4}$	$5.6327\ 10^{-4}$
n=50	0.0318	$7.2359\ 10^{-4}$	$4.8501\ 10^{-4}$	$1.2762\ 10^{-2}$	$2.7312\ 10^{-4}$	$1.7385\ 10^{-4}$
n=100	0.0110	$2.7785\ 10^{-4}$	1.065110^{-3}	$9.8506\ 10^{-3}$	$4.3718\ 10^{-5}$	$3.2146\ 10^{-3}$
	$\theta=1.5 \quad \alpha=1 \quad k=1.25$			$\theta=2 \quad \alpha=3 \quad k=1.25$		
	$MSE(\theta)$	$MSE(\alpha)$	$MSE(k)$	$MSE(\theta)$	$MSE(\alpha)$	$MSE(k)$
n=30	0.2546	$5.7258\ 10^{-4}$	1.093510^{-2}	0.2366	$2.05867\ 10^{-3}$	$1.6487\ 10^{-2}$
n=50	0.1581	$1.3610\ 10^{-4}$	$6.8303\ 10^{-3}$	0.0323	$1.5698\ 10^{-3}$	$6.2404\ 10^{-3}$
n=100	0.0401	$5.3779\ 10^{-5}$	$9.0554\ 10^{-5}$	0.0789	$3.7259\ 10^{-5}$	$1.9747\ 10^{-3}$
	$\theta=4 \quad \alpha=3 \quad k=3$			$\theta=1.5 \quad \alpha=5 \quad k=7$		
	$MSE(\theta)$	$MSE(\alpha)$	$MSE(k)$	$MSE(\theta)$	$MSE(\alpha)$	$MSE(k)$
n=30	2.0969	$8.8911\ 10^{-5}$	0.2603	0.1832	$6.3244\ 10^{-4}$	0.1130
n=50	0.5537	$5.0155\ 10^{-5}$	0.2510	$4.5248\ 10^{-2}$	$1.8568\ 10^{-4}$	$3.7919\ 10^{-3}$
n=100	0.3670	$1.1140\ 10^{-5}$	$2.0932\ 10^{-2}$	$2.4916\ 10^{-3}$	$4.7630\ 10^{-5}$	$5.3214\ 10^{-2}$
	$\theta=0.5 \quad \alpha=0.3 \quad k=0.9$			$\theta=1 \quad \alpha=0.8 \quad k=0.5$		
	$MSE(\theta)$	$MSE(\alpha)$	$MSE(k)$	$MSE(\theta)$	$MSE(\alpha)$	$MSE(k)$
n=30	0.1050	$4.4654\ 10^{-3}$	$9.8114\ 10^{-5}$	0.0388	$8.0237\ 10^{-3}$	$2.2877\ 10^{-4}$
n=50	0.0118	$1.5552\ 10^{-3}$	$1.8561\ 10^{-5}$	0.0209	$4.6262\ 10^{-3}$	$3.3000\ 10^{-5}$
n=100	0.0027	$3.4546\ 10^{-4}$	$2.1128\ 10^{-5}$	0.0027	$1.4167\ 10^{-3}$	$8.3529\ 10^{-5}$
	$\theta=0.75 \quad \alpha=0.5 \quad k=1.25$			$\theta=3 \quad \alpha=1.5 \quad k=2$		
	$MSE(\theta)$	$MSE(\alpha)$	$MSE(k)$	$MSE(\theta)$	$MSE(\alpha)$	$MSE(k)$
n=30	0.0240	$1.2623\ 10^{-3}$	$2.2451\ 10^{-4}$	2.3788	$9.3359\ 10^{-5}$	0.01964
n=50	0.0254	$6.6638\ 10^{-4}$	$1.4017\ 10^{-3}$	0.8301	$4.3801\ 10^{-5}$	$5.3191\ 10^{-3}$
n=100	0.0069	$1.6307\ 10^{-4}$	$2.9504\ 10^{-4}$	0.2535	$1.0024\ 10^{-5}$	$5.9578\ 10^{-4}$

Table 1 shows how the four biases vary with respect to n. Table 2 shows how the mean squared errors vary with respect to n. The mean squared errors for each parameter decrease to zero as $n \to \infty$. These numerical results coincide with the established theoretical results.

4 Application to real data sets

In this section, we give the applicability of LP distribution by considering three different data sets used by different researchers: Application to waiting times in a queue, Wheaton River Data, Application to bladder cancer patients, and compare them with different distribution, of which Lindley exponential, Lindley Weibull, Lindley, Power Lindley (see, Cooray, 2006), exponential Pareto, Pareto and gamma Lindley distributions. In each case, the parameters are estimated by maximum likelihood, as described in Section 6, using the R software.

In order to compare the above distributions with Lindley Pareto distribution, we consider criteria like $-2l$, AIC (Akaike information criterion), $AICC$ (corrected Akaike information criterion), BIC (Bayesian information criterion) and $HQIC$ (Hannan-Quinn information criterion) for the data set. The model selection is carried out using the following statistics:

$$AIC = -2LL + 2p, CAIC = -2LL + \frac{2pn}{n-p-1}$$

$$BIC = -2LL + p\log(n) \text{ and } HQIC = -2LL + 2p\log(\log(n))$$

For instance, it is well known that the AIC statistics favours models with large number of parameters in contrast to the Bayesian Information Criterion (BIC), which tends to present a better balance between the (negative) likelihood function and the number of parameters or model complexity.

Remark 3. Kolmogorov Smirnov test cannot be used in this case because the parameters are being estimated.

4.1 Illustration 1: Application to waiting times in a queue

We consider 100 observations on waiting time as a real example that happens before the customer received service in a bank. The data set represents the waiting time (mins) of one hundred (100) bank customers before service is being rendered. This data has previously been used by Ghitany et al. (2008a). Table 3 provides the estimated values of the model parameters. The information criterion values are given in Table 4.

Table 3. Parameter estimates for 100 bank customers

Distribution	Parameters
LP	$\hat{\theta} = 0.1586 \quad \hat{\alpha} = 0.801 \quad \hat{k} = 1.0048$
LE	$\hat{\theta} = 2.6501 \quad \hat{\lambda} = 0.152$
EP	$\hat{k} = 1.5137 \quad \hat{\alpha} = 0.801 \quad \hat{\lambda} = 0.0183$
GaL	$\hat{\theta} = 0.2024 \quad \hat{\beta} = 217.72$
L	$\hat{\theta} = 0.187$
P	$\hat{\alpha} = 0.801 \quad \hat{k} = 0.4367$
LW	$\hat{\theta} = 0.0003 \quad a = 1.0096 \quad b = 0.0014$
PL	$\hat{\theta} = 0.153 \quad \hat{\alpha} = 1.0832$

Table 4. The -LL, AIC, CAIC, BIC, HQIC for 100 bank customers

Distribution	-LL	AIC	CAIC	BIC	HQIC
LP	**308.9731**	**621.9462**	**622.0874**	**627.6346**	**623.9423**
LE	317.005	638.01	638.1337	643.2203	640.1187
EP	312.1154	628.2308	628.372	633.9192	630.2269
GaL	317.3066	638.6132	638.7369	643.8235	640.7219
L	319.00	640.00	640.0408	642.6052	641.0544
P	381.7586	765.5172	765.5637	767.9945	766.5153
LW	317.3267	640.6534	640.9034	648.4689	643.8165
PL	318.3186	640.6372	641.9156	645.8475	642.7459

4.2 Illustration 2: Wheaton River Data

In this subsection we illustrate the flexibility of the new distribution to model both heavy tailed and approximately symmetric data, which correspond to the exceedance of food peaks (in m^3/s) of the Wheaton river near Carcross in Yukon Territory (Canada) of 72 exceedance measures for the years 1958-1984. These data were analyzed by many authors (see for instance, Akinsete et al., 2008). We have chosen the same data in order to compare our results with other models proposed by these authors. Table 5 provides the estimated values. The -LL, AIC, CAIC, BIC and HQIC statistics for each model is provided in Table 6. It can be seen that our proposed distribution leads to a better fit than any of alternative approaches.

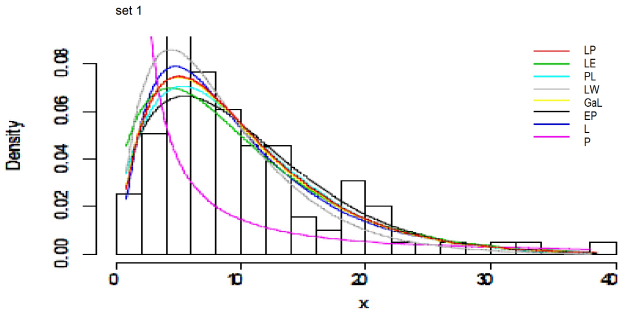


Figure 1: Estimated densities of the models for data set 1

Table 5. Parameter estimates for Wheaton river flood data

Distribution	Parameters		
LP	$\hat{\theta} = 0.1320$	$\hat{\alpha} = 0.1001$	$\hat{k} = 0.5921$
LE	$\hat{\theta} = 1.1210$	$\hat{\beta} = 0.0622$	
EP	$\hat{k} = 0.9320$	$\hat{\lambda} = 0.0115$	$\hat{\alpha} = 0.1001$
GaL	$\hat{\theta} = 0.0821$	$\hat{\beta} = 0.0760$	
L	$\hat{\theta} = 0.1531$		
P	$\hat{\alpha} = 0.1002$	$\hat{k} = 0.2405$	
WL	$\hat{\theta} = 0.0035$	$a = 0.5922$	$b = 0.0002$
PL	$\hat{\theta} = 0.3386$	$\hat{\alpha} = 0.7001$	

Table 6. The statistics -LL, AIC, CAIC, BIC, HQIC for Wheaton river flood data

Distribution	-LL	AIC	CAIC	BIC	HQIC
LP	249.3267	502.6534	502.7502	508.3418	504.9645
LE	251.5364	507.0728	507.1688	512.7769	509.3904
EP	249.3288	502.6576	502.7544	508.346	504.9687
GaL	252.128	508.256	508.352	513.9601	510.5736
L	264.2118	530.4236	530.4553	533.2756	531.5824
P	303.9486	609.8972	609.9292	612.7414	611.0528
LW	252.3039	510.6078	510.8013	519.1639	514.0842
PL	252.2218	508.4436	508.5396	514.1477	510.7612

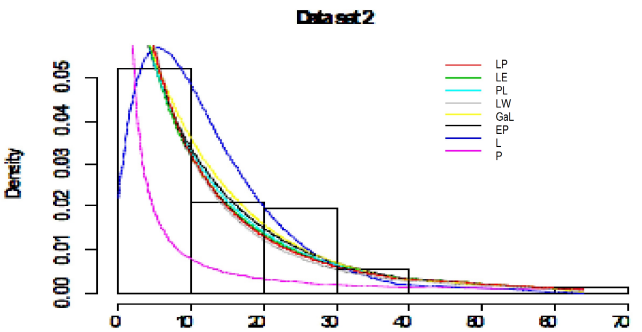


Figure 2: Estimated densities of the models for data set 2

4.3 Illustration 3: Application to bladder cancer patients

We consider a non-controlled data set corresponding to the remission times (in months) of a random sample of (128) bladder cancer patients. This cancer is a disease in which aberrant cells increase without control in the bladder and its application in survival analysis has been identified. The data set was given by Lee and Wang (2003). The results for these data are presented in Tables 7 and 8.

Table 7. Parameter estimates for bladder cancer data

Distribution	Parameters		
LP	$\hat{\theta} = 0.1229$	$\hat{\alpha} = 0.0801$	$\hat{k} = 0.6243$
LE	$\hat{\theta} = 1.2292$	$\hat{\lambda} = 0.0962$	
EP	$\hat{k} = 0.9379$	$\hat{\lambda} = 0.0128$	$\hat{\alpha} = 0.08$
GaL	$\hat{\theta} = 0.1167$	$\hat{\beta} = 0.1045$	
L	$\hat{\theta} = 0.1961$		
P	$\hat{\alpha} = 0.0801$	$\hat{k} = 0.2458$	
WL	$\hat{\theta} = 0.0027$	$a = 0.6316$	$b = 0.0002$
PL	$\hat{\theta} = 0.3855$	$\hat{\alpha} = 0.7443$	

Table 8. The statistics -LL, AIC, CAIC, BIC, HQIC for bladder cancer data

Distribution	-LL	AIC	CAIC	BIC	HQIC
LP	398.0184	800.0368	800.1336	805.7252	802.3479
LE	401.78	807.564	807.656	813.2641	809.8776
EP	400.3128	804.6256	804.7224	810.314	806.9367
GaL	402.9596	809.9192	810.0152	815.6233	812.2368
L	419.52	841.040	841.0717	843.892	842.1988
P	501.1292	1004.258	1004.29	1007.103	1005.414
WL	401.196	808.392	808.5855	816.9481	811.8684
PL	402.2373	808.4746	808.5706	814.1787	810.7922

According to Tables 4, 6, 8 and Figures 1, 2, 3, we can observe that LP distribution provide smallest -LL, AIC, CAIC, BIC and HQIC values as compared to Lindley exponential, Lindley Weibull, Lindley, Power Lindley, exponential Pareto, Pareto and gamma Lindley distributions, and hence best fits the data among all the models considered.

5 Conclusion

This work proposes more properties and simulations of the Lindley Pareto distribution generated by Lindley distribution. We investigate several of its structural properties such as an expansion for the density function and explicit expressions for the quantile function, maximum likelihood estimators of the parameters, mean deviation, and entropy. A simulation study is carried out to examine the bias and mean square error of the maximum likelihood estimators of the parameters. Several applications of the model to a real

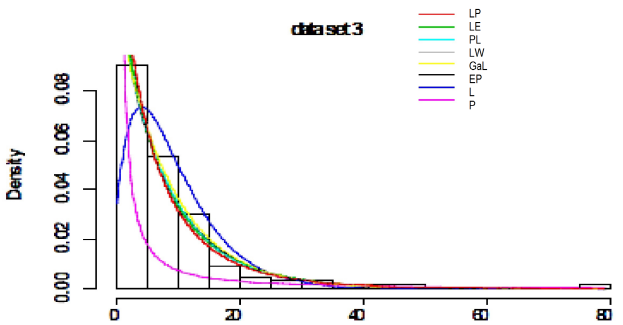


Figure 3: Estimated densities of the models for data set 3

data set are presented finally and compared with the fit attained by some other well-known one, two, three and four parameters. The adequacy of fits was assessed in terms AIC values, BIC values and density plots. We can show that the Lindley Pareto distribution can be used quite effectively in analyzing real lifetime data and actuarial science.

Acknowledgements

The authors acknowledge Editors, Prof. Wlodzimierz Okrasa and Prof. Barszcz Patryk, of this journal for the constant encouragement to finalize the paper. Also, this work was financed by European Mathematical Society.

REFERENCES

ABOUELMAGD, T. H. M, AL-MUALIM, S., AFIFY, A.Z. AHMAD, M., AL-MOFLEH, H., (2018). The odd Lindley Burr XII distribution with applications. Pak. J. Statist. 34(1), pp. 15–32.

AKINSETE, A., FAMOYE, F., LEE, C., (2008). The beta-Pareto distribution, Statistics, 42, pp. 547–563.

ALZAATREH, A., LEE, C., FAMOYE, F., (2013a). A new method for generating families of continuous distributions. Metron. 71(1), pp. 63–79.

ARLOND, B. C., BALAKRISHNAN, N., (1989). Relations, bounds and approximations for order statistics. Lecture Notes in Statistics Vol. 53, Springer-Verlag, New York.

ASGHARZADEH, A., BAKOUCH, H. S., ESMAEILI, L., (2013). Pareto Poisson–Lindley distribution with applications. J. of Applied Statistics, 40(8), pp. 1717–1734.

COORAY, K., (2006). Generalization of the Weibull Distribution: The Odd Weibull Family. Statistical Modelling, 6, pp. 265–277.

GHITANY, M. E., AL-MUTAIRI, D. K., NADARAJAH S., (2008a). Zero-truncated Poisson-Lindley distribution and its application, Math. Comput. Simulation, 79, pp. 279–287.

GHITANY, M. E., ATIEH, B. NADARAJAH, S., (2008b). Lindley distribution and its applications. Math. Comput. Simulation, 78, pp. 493–506.

GOMES-SILVA, F.S., PERCONTINI, A., DE BRITO, E., RAMOS, M. W., VENÂNCIO, R., CORDEIRO, G. M., (2017). The odd Lindley-G family of distributions. Austrian Journal of Statistics, 46(1), pp. 65–87.

LEE, E.T., WANG, J.W., (2003). Statistical Methods for Survival Data Analysis, 3rd edn. Wiley, Hoboken.

LEHMANN, E.L., SCHEFFÉ, H., (1950). Completeness, similar regions, and unbiased estimation. Sankhyā, 10, pp. 305–340.

LINDLEY, D. V., (1958). Fiducial distributions and Bayes' theorem. Journal of the Royal Society, series B,20, pp. 102–107.

MÄKELÄINEN, T., SCHMIDT, K., STYAN, G.P.H., (1981). On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples, The Annals of Statistics, 9(4), pp. 758–767.

MAHMOUDI, E., (2011). The beta generalized Pareto distribution with application to lifetime data. Mathematics and Computers in Simulation, 81, pp. 2414–2430.

PARETO, V., (1896). Essai sur la courbe de la répartition de la richesses. Faculté de droit à l'occasion de l'exposition nationale suisse, Genève, Université de Lausanne.

PICKANDS, J., (1975) Statistical inference using extreme order statistics. Annals of Statistics, 3, pp. 119–131.

RÉNYI, A., (1961). On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, I, University of California Press, Berkeley, pp. 547–561.

SHANNON, C. E., (1948). A mathematical theory of communication. Bell System Technical Journal, 27, pp. 379–432.

SANKARAN, M., (1970). The discrete Poisson-Lindley distribution. Biometrics, 26, pp. 145–149.

SHARMA, M., SHANKER, R., (2013). A two-parameter Lindley distribution for modeling waiting and survival times data, Applied Mathematics, 4, 363-368.

ZAKERZADAH, H. , DOLATI, A., (2010). Generalized Lindley distribution. J. Math. Ext, 3(2), pp. 13–25.

ZEA, L.M., SILVA, R.B., BOURGUIGNON, M., SANTOS, A.M., CORDEIRO, G.M.,(2012). The beta exponentiated Pareto distribution with application to bladder cancer susceptibility. International Journal of Statistics and Probability, 1, pp. 8–19.

ZEGHDOUDI, H., NEDJAR, S., (2016). Gamma Lindley distribution and its application. Journal of Applied Probability and Statistics. 11(1), pp. 129–138.

ZEGHDOUDI, H., NEDJAR, S., (2016). On gamma Lindley distribution: Properties and Simulations. Journal of Computational and Applied Mathematics, 298, pp. 67–174.

ZEGHDOUDI, H., LAZRI, N., (2016). On Lindley-Pareto Distribution: Properties and Application. Journal of Mathematics, Statistics and Operations Research (JMSOR), Vol. 3, No. 2.

APPENDIX

(1) Power Lindley distribution

$$f_1(x)=\frac{\alpha\theta^2}{(\theta+1)}(1+x^\alpha)x^{\alpha-1}\exp(-\theta x^\alpha)$$

(2)Lindley Weibull Distribution

$$f_5(x)=\frac{\alpha\theta^2}{b(\theta+1)}\left(\frac{x}{b}\right)^{\alpha-1}\left(1+\left(\frac{x}{b}\right)^\alpha\right)\exp\left(-\theta\left(\frac{x}{b}\right)^\alpha\right)$$

(3) Lindley Distribution

$$f_3(x)=\frac{\theta^2}{1+\theta}(1+x)\exp(-\theta x)$$

(4) Lindley Exponential distribution

$$f_4(x)=\frac{\lambda\theta^2\exp(-\lambda x)}{(\theta+1)}(1-\exp(-\lambda x))^{\theta-1}(1-\ln(1-\exp(-\lambda x)))$$

(5) Pareto Distribution

$$f_6(x)=k\frac{\alpha^k}{x^{k+1}}$$

(6) Exponential Pareto Distribution

$$f_6(x)=\frac{\lambda\alpha}{k}\left(\frac{x}{k}\right)^{\alpha-1}e^{-\lambda\left(\frac{x}{k}\right)^\alpha}$$

(7) Gamma Lindley distribution

$$f_2(x)=\frac{\theta^2((\beta+\beta\theta-\theta)x+1)e^{-\theta x}}{\beta(1+\theta)}$$

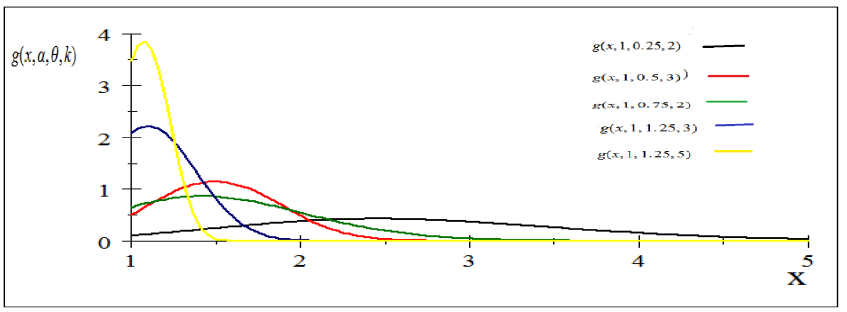


Figure 4: PDF plot for various values of parameters

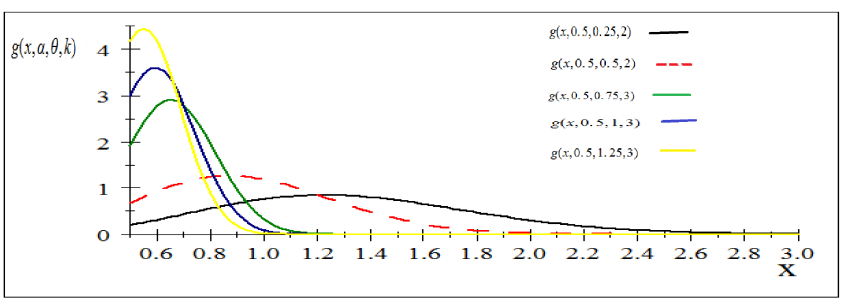


Figure 5: PDF plot for various values of parameters

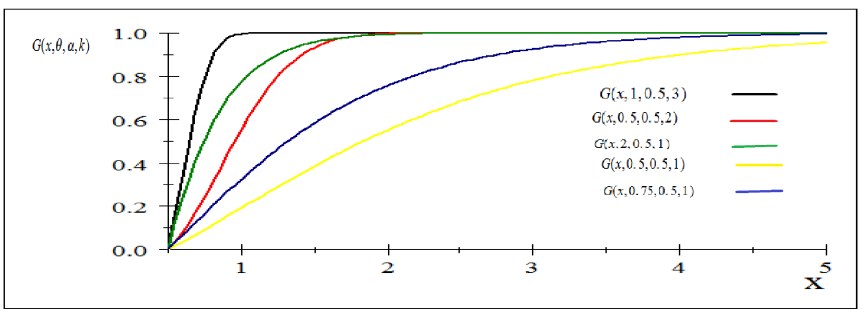


Figure 6: CDF plot for various values of parameters

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 693–710, DOI 10.21307/stattrans-2018-036

THE CHOICE OF NORMALIZATION METHOD AND RANKINGS OF THE SET OF OBJECTS BASED ON COMPOSITE INDICATOR VALUES

Marek Walesiak[1]

ABSTRACT

The choice of the normalization method is one of the steps for constructing a composite indicator for metric data (see, e.g. Nardo et al., 2008, pp. 19-21). Normalization methods lead to different rankings of the set of objects based on composite indicator values. In the article 18 normalization methods and 5 aggregation measures (composite indicators) were taken into account. In the first step the groups of normalization methods, leading to identical rankings of the set of objects, were identified. The considerations included in Table 3 reduce this number to 10 normalization methods. Next, the article discusses the procedure which allows separating groups of normalization methods leading to similar rankings of the set of objects separately for each composite indicator formula. The proposal, based on Kendall's tau correlation coefficient (Kendall, 1955) and cluster analysis, can reduce the problem of choosing the normalization method. Based on the suggested research procedure the simulation results for five composite indicators and ten normalization methods were presented. Moreover, the proposed approach was illustrated by an empirical example. Based on the analysis of the dendrograms three groups of normalization methods were separated. The biggest differences in the results of linear ordering refer to methods n2, n9a against the other normalization methods.

Key words: variables normalization, rankings, composite indicators, Kendall's tau correlation coefficient, cluster analysis.

1. Introduction

Simulation studies, allowing the alignment of linear ordering (ranking) of the set of objects, via composite indicators values, procedures (the procedure takes into account weights of variables, selected normalization methods and selected constructions of aggregation measures), from the perspective of determining the correctness (quality) of aggregated variables, were conducted by (Grabiński, 1984) and (Bąk, 1999). T. Grabiński (see Grabiński, 1984, pp. 58–62; Grabiński,

[1] Wroclaw University of Economics, Department of Econometrics and Computer Science, Jelenia Góra, Poland. E-mail: marek.walesiak@ue.wroc.pl.

Wydymus and Zeliaś, 1989, pp. 122–123) suggested five groups of correctness (quality) measures for determining the value of aggregated variables:

1. Measures of compatibility of distance matrices calculated for objects in *m*-dimensional space of the variables and 1-dimensional space of the aggregated variable (3 measures).
2. Measures based on Pearson product-moment correlation coefficient between *m* variables and the aggregated variable (2 measures).
3. Measures based on Spearman's rank correlation coefficient between *m* variables and the aggregated variable (3 measures).
4. Measures determining an average taxonomic distance of the aggregated variable from the *m* variables (2 measures).
5. Measures characterizing the variability level and concentration for the aggregated variable (2 measures).

The better a given linear ordering procedure (taking into account weights of variables, selected normalization methods and selected constructions of aggregation measures), the lower are the values of these measures (Grabiński, Wydymus and Zeliaś, 1989, p. 125). The author does not justify substantively the introduced measures. The doubts related to their application are presented based on two groups of measures.

The first group of measures covers the selected compatibility functions applied in multidimensional scaling (e.g. STRESS-1 function – see Borg and Groenen, 2005, p. 42). Based on the distance matrix between objects in *m*-dimensional space, such mapping of the set of objects into a set of points in *r*-dimensional space is sought ($r < m$, in linear ordering $r = 1$ – of the aggregated variable values), which allows achieving the best possible compatibility. The objects distant from each other in *m*-dimensional space shall also remain distant in *r*-dimensional space (1-dimensional). The situation is, however, different in the case of linear ordering. A distant object, from the perspective of the initial set of *m* variables, can be found at the same distance from the pattern object. Therefore, the distance between them, in terms of the aggregated variable, may equal zero.

In the second group, e.g. the measure of linear correlation of the aggregated variable with diagnostic variables was suggested, which takes the following form (the so-called uncertainty coefficient):

$$M_4 = 1 - \frac{1}{m}\sum_{j=1}^{m} r_{\cdot j}, \tag{1}$$

where: $r_{\cdot j}$ – linear correlation coefficient for *j*-th variable with the aggregated variable,

$j = 1, \ldots, m$ – variable number.

The most preferred value of this measure is 0, when all correlation coefficients of diagnostic variables with the aggregated variable equal 1. Such approach is missing substantive justification in the case of linear ordering.

Due to an ambiguous interpretation of correctness (quality) measures for determining values of aggregated variables a different approach was used in the article.

There is a growing demand for various rankings of the set of objects (e.g. countries, regions) due to, for example, their competitiveness, tourist attractiveness, social cohesion, socio-economic development, and environmental

pollution. Analyses using aggregate measures require normalization of variable values. Normalization methods lead to different rankings of the set of objects based on aggregation measures (composite indicators) values. The proposed approach allows to objectivize the results of analyses in this area.

In the article 18 normalization methods and 5 aggregation measures (composite indicators) were taken into account. Two elements of the article should be considered innovative:

- identification of groups of normalization methods resulting in identical values and identical orderings for the aggregation measures (see Table 3),
- the proposal of the procedure allowing the separation of the groups of normalization methods leading to similar rankings of the set of objects (see section 3).

The proposal, based on Kendall's tau correlation coefficient and cluster analysis, can reduce the problem of choosing the normalization method. Based on the suggested research procedure the simulation results were presented. Moreover, the proposed approach was illustrated by an empirical example.

2. Steps for Constructing a Composite Indicator

The general procedure in linear ordering (ranking) of the set of objects via composite indicators values, carried out based on metric data (measured on an interval scale and ratio scale)[2], takes the following form (see Grabiński, Wydymus and Zeliaś, 1989, p. 92; Pawełek, 2008, pp. 110–111; Nardo et al., 2008, pp. 19–21):

a) for methods based on pattern object (there are two types of pattern objects: upper pattern – ideal object, upper pole, lower pattern – anti-ideal object, lower pole):

$$P \rightarrow A \rightarrow X \rightarrow \left[x_{ij}\right] \rightarrow SDN \rightarrow T_w \rightarrow N \rightarrow SM_w \rightarrow R, \tag{2}$$

where:

P – choice of a complex phenomenon (the overriding multidimensional phenomenon for ordering *A* set elements, which is not subject to direct measurement),

A – choice of objects,

X – selection of variables,

$\left[x_{ij}\right]$ – collecting data and the construction of data matrix (x_{ij} – value for *j*-th variable for *i*-th object),

SDN – identifying preferential variables (stimulant, destimulant, nominant). M_j variable is a stimulant (see Hellwig, 1981, p. 48), when for every two of its observations x_{ij}^S, x_{kj}^S referring to objects A_i, A_k, it takes $x_{ij}^S > x_{kj}^S \Longrightarrow A_i \succ A_k$ ($\succ$ means A_i object domination over A_k object). M_j variable is a destimulant (see Hellwig, 1981, p. 48), when for every two of its observations x_{ij}^D, x_{kj}^D referring to objects A_i, A_k take $x_{ij}^D > x_{kj}^D \Longrightarrow A_i \prec A_k$ ($\prec$ means A_k object domination over

[2] The characteristics of measurement scales is presented, e.g. in the studies by (Stevens, 1946; Walesiak, 1995; Walesiak, 2011, pp. 13–16).

A_i object). Therefore, M_j variable represents a unimodal nominant (see Borys, 1984, p. 118), when for every two of its observations x_{ij}^N, x_{kj}^N referring to objects A_i, A_k (nom_j means the nominal level of *j*-th variable) if $x_{ij}^N, x_{kj}^N \leq nom_j$, then $x_{ij}^N > x_{kj}^N \Longrightarrow A_i \succ A_k$; if $x_{ij}^N, x_{kj}^N > nom_j$, then $x_{ij}^N > x_{kj}^N \Longrightarrow A_i \prec A_k$,

T_w – transformation of nominants into stimulants (required for an anti-ideal object only). Transformation formulas can be found, e.g. in the study by (Walesiak, 2011, p. 18),

N – normalization of variable values,

SM_w – composite indicator calculation by aggregating normalized variables – the application of distance measures from pattern object using weights. The coordinates of upper pattern object covers the most preferred variable values (maximum for a stimulant, minimum for a destimulant). The coordinates of lower pattern object cover the least preferred variable values (minimum for a stimulant, maximum for a destimulant),

R – ordering of objects (ranking) in accordance with the composite indicator values.

b) for methods not based on pattern object:

$$P \rightarrow A \rightarrow X \rightarrow \left[x_{ij}\right] \rightarrow SDN \rightarrow T_b \rightarrow N \rightarrow SM_b \rightarrow R, \tag{3}$$

where:

T_b – transformation of destimulants and nominants into stimulants. Transformation formulas are presented, e.g. in the study by (Walesiak, 2011, p. 18),

SM_b – composite indicator calculation by aggregating normalized variables – averaging normalized variable values using weights.

In linear ordering, carried out based on metrical data, the choice of the normalization method for variable values remains one of the stages. The purpose of normalization is to adjust the size (magnitude) and the relative weighting of the input variables (see, e.g. Milligan and Cooper, 1988, p. 182). An overview of normalization methods for variable values is presented in the study by (Walesiak, 2014b). Table 1 presents normalization methods of linear transformation (see e.g. Jajuga and Walesiak, 2000, pp. 106–107; Zeliaś, 2002, p. 792):

$$z_{ij} = b_j x_{ij} + a_j = \frac{x_{ij} - A_j}{B_j} = \frac{1}{B_j} x_{ij} - \frac{A_j}{B_j} \, (b_j > 0), \tag{4}$$

Table 1. Normalization methods

Type	Method	Parameter		Measurement scale of variables	
		B_j	A_j	before normalization	after normalization
n1	Standardization	s_j	$\bar{x}_j$	ratio or interval	Interval
n2	Positional standardization	mad_j	med_j	ratio or interval	Interval
n3	Unitization	r_j	$\bar{x}_j$	ratio or interval	Interval
n3a	Positional unitization	r_j	med_j	ratio or interval	Interval

Table 1. Normalization methods (cont.)

n4	Unitization with zero minimum	r_j	$\min_i\{x_{ij}\}$	ratio or interval	Interval
n5	Normalization in range [–1; 1]	$\max_i\lvert x_{ij} - \bar{x}_j\rvert$	$\bar{x}_j$	ratio or interval	Interval
n5a	Positional normalization in range [–1; 1]	$\max_i\lvert x_{ij} - med_j\rvert$	med_j	ratio or interval	Interval
n6	Quotient transformations	s_j	0	ratio	Ratio
n6a		mad_j	0	ratio	Ratio
n7		r_j	0	ratio	Ratio
n8		$\max_i\{x_{ij}\}$	0	ratio	Ratio
n9		$\bar{x}_j$	0	ratio	Ratio
n9a		med_j	0	ratio	Ratio
n10		$\sum_{i=1}^{n} x_{ij}$	0	ratio	Ratio
n11		$\sqrt{\sum_{i=1}^{n} x_{ij}^2}$	0	ratio	Ratio
n12	Normalization	$\sqrt{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}$	$\bar{x}_j$	ratio or interval	Interval
n12a	Positional normalization	$\sqrt{\sum_{i=1}^{n} (x_{ij} - med_j)^2}$	med_j	ratio or interval	Interval
n13	Normalization with zero being the central point	$\frac{r_j}{2}$	m_j	ratio or interval	Interval

$\bar{x}_j$ – mean for *j*-th variable, s_j – standard deviation for *j*-th variable, r_j – range for *j*-th variable, $m_j = \frac{\max_i\{x_{ij}\}+\min_i\{x_{ij}\}}{2}$ – mid-range for *j*-th variable, $med_j = \operatorname{med}_i(x_{ij})$ – median for *j*-th variable, $mad_j = \operatorname{mad}_i(x_{ij})$ – median absolute deviation for *j*-th variable.

Source: (Walesiak, 2014b, pp. 364–365).

where:

x_{ij} – value for *j*-th variable for *i*-th object,

z_{ij} – normalized value for *j*-th variable for *i*-th object,

A_j – shift parameter to arbitrary zero for *j*-th variable,

B_j – scale parameter for *j*-th variable,

$a_j = -A_j/B_j$, $b_j = 1/B_j$ – parameters for *j*-th variable presented in Table 1.

Column 1 in Table 1 presents the type of normalization formula adopted as the function data.Normalization of clusterSim package (see Walesiak and Dudek, 2018) of R program (R Core Team, 2018).

An aggregation measure SM_i represents the tool for linear ordering methods as a sub-function aggregating partial information contained in particular variables and determined for each object from the set of objects. Generally, the constructions of aggregation measures (composite indicators) can be divided as follows (cf. e.g. Grabiński 1984, p. 38):

– based on pattern object (e.g. Hellwig's measure of development; GDM1 distance; TOPSIS measure),
– not based on pattern object (arithmetic mean, harmonic mean, geometric mean; median).

Table 2 presents five constructions of aggregation measures (composite indicators) (four based on pattern object ones to be followed by one not based on pattern object) applied for metric data to be used later in the article.

Table 2. Constructions of aggregation measures (composite indicators) used for linear ordering (ranking) of objects

No.	Name	SM_i
1	GDM1 distance (Walesiak, 2002; Jajuga, Walesiak and Bąk, 2003)	$1-GDM1_i^+ = \frac{1}{2} + \frac{\sum_{j=1}^{m}\alpha_j(z_{ij}-z_{wj})(z_{wj}-z_{ij}) + \sum_{j=1}^{m}\sum_{\substack{l=1 \\ l\neq i,w}}^{n}\alpha_j(z_{ij}-z_{lj})(z_{wj}-z_{lj})}{2\left[\sum_{j=1}^{m}\sum_{l=1}^{n}\alpha_j(z_{ij}-z_{lj})^2 \cdot \sum_{j=1}^{m}\sum_{l=1}^{n}\alpha_j(z_{wj}-z_{lj})^2\right]^{0,5}}$
2	Measure of development (Hellwig, 1968; 1972)	$1-\frac{d_{iw}^+}{\bar{d}_{.w}^+ + 2s_d}$
3	TOPSIS measure (Hwang and Yoon, 1981)	$\frac{d_{iw}^-}{d_{iw}^- + d_{iw}^+}$
4	GDM1_TOPSIS – TOPSIS measure with GDM1 distance (Walesiak, 2014a)	$\frac{GDM1_i^-}{GDM1_i^- + GDM1_i^+}$
5	Arithmetic mean	$\sum_{j=1}^{m}\alpha_j z_{ij}$

SM_i – aggregation measure (composite indicator) value for *i*-th object (the resulting aggregate variable has stimulant interpretation), $i,l=1,\ldots,n$ – object number, w – pattern object number, $j=1,\ldots,m$ – variable number, z_{wj} – *j*-th coordinate of a pattern object, α_j – weight for *j*-th variable ($\alpha_j \in [0;1]$ and $\sum_{j=1}^{m}\alpha_j = 1$), $d_{iw} = \sqrt{\sum_{j=1}^{m}\alpha_j^2(z_{ij}-z_{wj})^2}$ – weighted Euclidean distance for *i*-th object from a pattern object, $GDM1_i^-$ and $GDM1_i^+$ – GDM1 distance for *i*-th object from the lower pole (anti-ideal object) and the upper pole (ideal object), d_{iw}^- and d_{iw}^+ – weighted Euclidean distance for *i*-th object from the lower and upper pole, $\bar{d}_{.w} = \frac{1}{n}\sum_{i=1}^{n}d_{iw}^+$, $s_d = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_{iw}^+ - \bar{d}_{.w})^2}$.

Source: Author's compilation.

3. Research Procedure Allowing Separation of the Groups of Normalization Methods Resulting in a Similar Linear Ordering of a Set of Objects

The research procedure allowing separation of the groups of normalization methods for variable values resulting in a similar linear ordering (ranking) of a set of objects covers the following steps:

1. Linear ordering of the set of objects is performed in accordance with the general procedure used in linear ordering methods illustrated in section 2 (scheme (2) or (3)). Any acceptable methods presented in Table 1 are used in the normalization of variable values (for ratio variables 18 normalization methods are possible and for interval variables – 10 normalization methods).
2. Object ordering obtained for the acceptable normalization methods is compared with the application of Kendall's tau correlation coefficient Γ_{rs} (see Kendall and Buckland, 1986, p. 266; Kendall, 1955, p. 19; Walesiak, 2011, pp. 36-38). Kendall's tau correlation coefficient takes values in interval $[-1; 1]$. The value of 1 means complete compatibility of orderings, whereas the value -1 implies their complete opposition. For the purposes of cluster analysis, Kendall's tau correlation coefficients are transformed into distances using the following formula:
$$d_{rs} = \frac{1}{2}(1 - \Gamma_{rs}), \tag{5}$$
where:
$d_{rs} \in [0; 1]$, $d_{rs} = 0$, when $\Gamma_{rs} = 1$ and $d_{rs} = 1$, when $\Gamma_{rs} = -1$,
r, s – numbers of normalization methods.
3. Cluster analysis is carried out based on the distance matrix $[d_{rs}]$, which allows separating groups of normalization methods for variable values resulting in similar linear ordering of a set of objects. In this case it is possible to use one of many classification methods (see, e.g. Everitt et al., 2011). The agglomerative hierarchical method of the farthest neighbour clustering was applied in the article.

Certain observations can be put forward regarding normalization methods presented in Table 3 for aggregation measures (SM_i) obtained using the following distance measures: GDM1, Hellwig's measure of development, TOPSIS, GDM1_TOPSIS and SM_i taking the form of an arithmetic mean.

Table 3. Groups of normalization methods resulting in identical values and identical orderings for the aggregation measures (SM_i) from Table 2

Groups of methods	Identical SM_i values		Identical orderings (rankings)
	Distances: GDM1 and GDM1_TOPSIS	Hellwig's measure of development, TOPSIS	all SM_i constructions from Table 2
A	n3, n3a, n4, n7, n13	n3, n3a, n4, n7	n3, n3a, n4, n7, n13
B	n1, n6, n12	n1, n6	n1, n6, n12
C	n2, n6a	n2, n6a	n2, n6a
D	n9, n10	–	n9, n10

Source: Author's compilation.

Identical SM_i values (and thus identical orderings) for A, B, C and D groups of formulas in the case of GDM1 and GDM1_TOPSIS distance measures result from the fact that GDM1 distance does not depend on the shift parameter used in normalization methods (4). Furthermore, multiplying normalized values by a constant does not change GDM1 distance:

– for n13 formula the constant equals 2:

$$z_{ij} = \frac{x_{ij}}{r_j/2} - \frac{m_j}{r_j/2} = 2\left(\frac{x_{ij}}{r_j} - \frac{m_j}{r_j}\right), \tag{6}$$

– for n12 formula the constant equals $\sqrt{\frac{1}{n-1}}$:

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}} - \frac{\bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}} = \sqrt{\frac{1}{n-1}}\left(\frac{x_{ij}}{s_j} - \frac{\bar{x}_j}{s_j}\right), \tag{7}$$

– for n10 formula the constant equals $1/n$:

$$z_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} = \frac{1}{n}\frac{x_{ij}}{\bar{x}_j}. \tag{8}$$

Identical SM_i values (and thus identical orderings), in the case of Hellwig's measure of development and TOPSIS, result from the fact that Euclidean distance applied in these measures does not depend on the shift parameter used in normalization methods, but only on the scale parameter which is identical for A, B and C groups of methods (see Pawełek, 2008, p. 94).

Additionally, n13 is present in A group of normalization methods, while in group B – n12 formula. Two normalization methods n9 and n10 result in identical object ordering. For n13, n12 and n10 formulas normalized values are multiplied by a constant. This causes a change in Euclidean distance, however, does not change the ordering of objects.

In the case of a aggregation measure, taking the form of an arithmetic mean, identical orderings result from the fact that the shift parameter, used in normalization methods, does not change the order of objects (in fact a constant is subtracted from SM_i value of each object). Multiplying SM_i value by a constant does not alter the order of objects either. For example, for n1, n6 and n12 formulas from group B the following is obtained:

$$\text{for n1: } SM_i = \sum_{j=1}^{m}\left(\frac{x_{ij}}{s_j} - \frac{\bar{x}_j}{s_j}\right) = \sum_{j=1}^{m}\frac{x_{ij}}{s_j} - \sum_{j=1}^{m}\frac{\bar{x}_j}{s_j}, \tag{9}$$

$$\text{for n6: } SM_i = \sum_{j=1}^{m}\frac{x_{ij}}{s_j}, \tag{10}$$

$$\text{for n12: } SM_i = \sum_{j=1}^{m}\left(\frac{x_{ij}}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}} - \frac{\bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}\right) = \sqrt{\frac{1}{n-1}}\left(\sum_{j=1}^{m}\frac{x_{ij}}{s_j} - \sum_{j=1}^{m}\frac{\bar{x}_j}{s_j}\right). \tag{11}$$

The order of objects, determined in line with n6 normalization method, does not change n1 formula (subtracting a constant from each SM_i value obtained for n6 formula) and for n12 formula (here subtracting a constant takes place and next multiplying by a different constant).

4. The Results of Simulation Analyses

The research procedure discussed in section 3 was used in simulation analyses, which allows separating groups of normalization methods for variable values resulting in similar linear orderings of a set of objects using a specific aggregation measure (SM_i):

1. Multivariate normal distribution was used to generate data (function rmnorm from the mnormt package – see Genz and Azzalini, 2016) based on models presented in Table 4. A simplifying assumption was adopted that the set of analysed variables includes stimulants only. Correlation with aggregate variable (vector of SM_i values) is positive for stimulants (see Grabiński, 1992, p. 138). Due to the transitivity of variables correlation[3] (Hellwig, 1976) it was adopted that the correlation between stimulants will also be positive. Therefore, models in Table 4 take values of correlation coefficients from 0.2 to 0.95 between variables in data matrix. The generated data differ in terms of variables' order of magnitude (see mean values for variables) and the variability measured by the coefficient of variation (0.20, 0.16, 0.24, 0.10).

Table 4. The characteristics of models in simulation analysis

No.	Mean values for variables	Covariance matrix Σ	Correlation matrix $[r_{jl}]$
1	$(10, 125, 250, 1000)$	$\begin{bmatrix} 4 & 14 & 42 & 70 \\ 14 & 400 & 420 & 700 \\ 42 & 420 & 3600 & 2100 \\ 70 & 700 & 2100 & 10000 \end{bmatrix}$	$r_{jj} = 1, r_{jl} = 0.35$ $1 \leq j, l \leq 4$
2	$(10, 125, 250, 1000)$	$\begin{bmatrix} 4 & 26 & 78 & 130 \\ 26 & 400 & 780 & 1300 \\ 78 & 780 & 3600 & 3900 \\ 130 & 1300 & 3900 & 10000 \end{bmatrix}$	$r_{jj} = 1, r_{jl} = 0.65$ $1 \leq j, l \leq 4$
3	$(10, 125, 250, 1000)$	$\begin{bmatrix} 4 & 38 & 114 & 190 \\ 38 & 400 & 1140 & 1900 \\ 114 & 1140 & 3600 & 5700 \\ 190 & 1900 & 5700 & 10000 \end{bmatrix}$	$r_{jj} = 1, r_{jl} = 0.95$ $1 \leq j, l \leq 4$
4	$(10, 125, 250, 1000)$	$\begin{bmatrix} 4 & 36 & 90 & 120 \\ 36 & 400 & 1080 & 1000 \\ 90 & 1080 & 3600 & 3600 \\ 120 & 1000 & 3600 & 10000 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.9 & 0.75 & 0.6 \\ 0.9 & 1 & 0.9 & 0.5 \\ 0.75 & 0.9 & 1 & 0.6 \\ 0.6 & 0.5 & 0.6 & 1 \end{bmatrix}$
5	$(10, 125, 250, 1000)$	$\begin{bmatrix} 4 & 8 & 60 & 140 \\ 8 & 400 & 480 & 1200 \\ 60 & 480 & 3600 & 1800 \\ 140 & 1200 & 1800 & 10000 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.2 & 0.5 & 0.7 \\ 0.2 & 1 & 0.4 & 0.6 \\ 0.5 & 0.4 & 1 & 0.3 \\ 0.7 & 0.6 & 0.3 & 1 \end{bmatrix}$

Source: Author's compilation.

2. Normalization of variables was carried out using the methods from Table 1. Due to the fact that the groups of A, B, C and D normalization methods result

[3] Let Y represent the aggregated variable, whereas X_1 and X_2 two variables from the data matrix. For $r_{X_1Y} = 0.9$ and $r_{X_2Y} = 0.95$ correlation coefficient $r_{X_1X_2}$ can only take values in the interval $0.719 \leq r_{X_1X_2} \leq 0.991$. On the other hand, for $r_{X_1Y} = 0.6$ and $r_{X_2Y} = 0.8$, correlation coefficient $r_{X_1X_2}$ can only take values in the interval $0 \leq r_{X_1X_2} \leq 0.96$.

in identical ordering, further analysis covered first methods from the indicated groups (n1, n2, n3, n9) and the other methods (n5, n5a, n8, n9a, n11, n12a).

3. Linear ordering was conducted using five aggregation measures (SM_i) listed in Table 2 (equal weights were used for variables).
4. For each individual aggregation measure (SM_i) the ordering of objects was compared by applying 10 normalization methods. Kendall's tau correlation coefficient Γ_{rs} was used to compare the ordering of objects, which gave 10 x 10 matrix.
5. Cluster analysis of normalization methods for variable values was carried out for 10 x 10 matrix. For the purposes of the analysis Kendall's tau correlation coefficient was transformed into distances using formula (5). The agglomerative hierarchical method of the farthest neighbour clustering was applied to separate groups of normalization methods for variable values resulting in similar linear orderings of the set of objects using a specific SM_i.

20 sets of data were generated for each model from Table 4, the procedure was conducted in accordance with points 2-5 divided into 2, 3 and 4 classes and next the obtained classification results of five aggregation measures (SM_i) from Table 2 were compared using the adjusted Rand index (see Hubert and Arabie, 1985). Table 5 presents the outcome of compatibility comparison for cluster analysis results of normalization methods for five aggregation measures (SM_i) taking the mean value of the adjusted Rand index.

Table 5. Compatibility comparison of cluster analysis results of normalization methods for five aggregation measures (SM_i) taking the mean value of the adjusted Rand index

Model 1

	1	2	3	4	5
1	1.000	0.914	0.886	0.891	0.870
2	**0.914**	1.000	**0.916**	**0.922**	0.865
3	0.886	0.916	1.000	0.890	**0.908**
4	0.891	**0.922**	0.890	1.000	0.859
5	0.870	0.865	0.908	0.859	1.000

Model 2

	1	2	3	4	5
1	1.000	**0.922**	0.841	0.826	0.813
2	**0.922**	1.000	0.833	0.839	0.834
3	0.841	0.833	1.000	**0.899**	0.824
4	0.826	0.839	**0.899**	1.000	**0.867**
5	0.813	0.834	0.824	0.867	1.000

Model 3

	1	2	3	4	5
1	1.000	**0.865**	0.800	0.824	**0.823**
2	**0.865**	1.000	0.801	0.774	0.808
3	0.800	0.801	1.000	**0.853**	0.806
4	0.824	0.774	**0.853**	1.000	0.806
5	0.823	0.808	0.806	0.806	1.000

Model 4

	1	2	3	4	5
1	1.000	**0.884**	0.885	0.893	0.862
2	0.884	1.000	0.861	0.817	0.844
3	0.885	0.861	1.000	0.892	0.873
4	**0.893**	0.817	**0.892**	1.000	**0.896**
5	0.862	0.844	0.873	**0.896**	1.000

Model 5

	1	2	3	4	5
1	1.000	**0.955**	0.879	0.898	0.869
2	**0.955**	1.000	0.870	**0.930**	0.877
3	0.879	0.870	1.000	0.908	0.857
4	0.898	0.930	**0.908**	1.000	**0.878**
5	0.869	0.877	0.857	0.878	1.000

Mean (models 1-5)

	1	2	3	4	5
1	1.000	**0.908**	0.858	0.866	0.847
2	**0.908**	1.000	0.856	0.856	0.845
3	0.858	0.856	1.000	**0.888**	0.853
4	0.866	0.856	**0.888**	1.000	**0.861**
5	0.847	0.845	0.853	0.861	1.000

1 – GDM1 distance, 2 – Hellwig's measure of development, 3 – TOPSIS measure, 4 – TOPSIS measure with GDM1 distance, 5 – arithmetic mean. Minimum values are underlined, maximum values are in **bold** (excluding the main diagonal).

Source: Author's compilation using R program.

Having analysed the obtained results of compatibility comparison for cluster analysis of normalization methods for five aggregation measures (SM_i), taking the mean value of the adjusted Rand index, the following conclusions can be drawn:

1. Values of the adjusted Rand index for models 1-5 vary in the interval $[0.774, 0.955]$. Mean values of the adjusted Rand index taken from five models are in the interval $[0.845, 0.908]$. Therefore, the results of cluster analysis of normalization methods for the analysed aggregation measures (SM_i) are similar to each other.
2. Dendrite of cluster analysis results' similarity of normalization methods for five aggregation measures is presented in Figure 1 (developed based on the matrix for models 1-5 from Table 5).

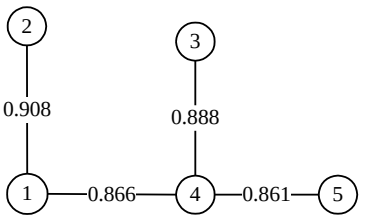


Figure 1. Dendrite of cluster analysis results' similarity of normalization methods for five aggregation measures

1 – GDM1 distance, 2 – Hellwig's measure of development, 3 – TOPSIS measure, 4 – TOPSIS measure with GDM1 distance, 5 – arithmetic mean.

Source: Author's compilation.

5. Empirical Research Results

The evaluation of tourism competitiveness level of the Sudety communes in Poland, covering 52 out of 169 communes in Lower Silesia region, was carried out in the article (Gryszel and Walesiak, 2018). The Sudety communes are located in the geographical area of the Sudety in the southern part of Lower Silesia region. They are characterized by the most valuable tourism advantages, where the tourism function either dominates or is of great importance among other economic functions in a commune. The following variables were used in the study:

x1 – beds in hotels per 1 km^2 of a commune area,

x2 – beds in other accommodation facilities per 1 km^2 of a commune area,

x3 – number of nights of resident tourists (Poles) falling per day per 1000 inhabitants of a commune,

x4 – number of nights of foreign tourists falling per day per 1000 inhabitants of a commune,

x5 – communal expenditure for tourism per 1000 inhabitants in PLN,

x6 – funds obtained from the European Union and from the state budget to finance the EU programs and projects per 1 inhabitant in PLN,

x7 – number of tourist economy entities per 1000 inhabitants of a commune (natural persons conducting economic activity),

x8 – number of tourist economy entities per 1000 inhabitants of a commune (legal persons and organizational entities without legal personality).

All variables represent stimulants. Statistical data were collected in 2012 and retrieved from the Local Data Bank (LDB). The research procedure discussed in section 3 was used in the article, which allows separating the groups of methods for the normalization of variable values, resulting in similar linear ordering of the set of communes in terms of their tourism competitiveness level. The analysed variables are measured using ratio scale, therefore all normalization methods listed in Table 1 are acceptable.

The results of linear ordering compatibility for 52 Sudety communes, in terms of their tourism competitiveness level, using 18 normalization methods and 5 SM_i from Table 2 are presented in Figure 1. Due to the fact that the groups of A, B, C and D normalization methods result in identical orderings, further analysis covered first formulas from the indicated groups (n1, n2, n3, n9) and other formulas (n5, n5a, n8, n9a, n11, n12a).

Regardless of the adopted SM_i construction, the results of linear ordering compatibility for 52 Sudety communes, in terms of their tourism competitiveness level, using 10 normalization methods are analogical in this case. Table 6 contains compatibility comparison of cluster analysis results of normalization methods (after splitting the dendrograms from Figure 2 into 2, 3 and 4 clusters) for five aggregation measures (SM_i) taking the mean value of the adjusted Rand index.

Table 6. Compatibility comparison of cluster analysis results of normalization methods (after splitting the dendrograms into 2, 3 and 4 clusters) for five aggregation measures (SM_i) taking the mean value of the adjusted Rand index.

Aggregation measure	1	2	3	4	5
1	1.0000	1.0000	0.7674	0.7674	0.7674
2	1.0000	1.0000	0.7674	0.7674	0.7674
3	0.7674	0.7674	1.0000	1.0000	1.0000
4	0.7674	0.7674	1.0000	1.0000	1.0000
5	0.7674	0.7674	1.0000	1.0000	1.0000

1 – GDM1 distance, 2 – Hellwig's measure of development, 3 – TOPSIS measure, 4 – TOPSIS measure with GDM1 distance, 5 – arithmetic mean.

Source: Author's compilation.

Taking into account results from Table 6 the following conclusions can be drawn:

1. The results of cluster analysis of normalization methods (after splitting the dendrograms into 2, 3 and 4 clusters) for TOPSIS measure, TOPSIS measure with GDM1 distance, and arithmetic mean are the same.
2. The results of cluster analysis of normalization methods (after splitting the dendrograms into 2, 3 and 4 clusters) for GDM1 distance and Hellwig's measure of development are the same.
3. The differences between groups of aggregation methods listed in points 1 and 2 relate to the division of dendrograms into two clusters.

Based on the analysis of the dendrograms in Figure 2 three groups of normalization methods were separated:

group 1 (6 methods): n1, n3, n5, n5a, n8, n12a,

group 2 (2 methods): n2, n9a,

group 3 (2 methods): n9, n11.

The results presented in Figure 2 regarding the adopted SM_i construction differ, for the separated groups of normalization methods, in the level of class links in a dendrogram.

The biggest differences in the results of linear ordering refer to methods n2, n9a against the other normalization methods.

The presented proposal allows reducing the problem of a normalization method selection. Significant differences between the results of linear ordering appear in the analysed case for the normalization methods from different groups.

In the current practice, not considering the proposed procedure, selecting 18 methods for normalizing variable values for metric data, we had 18 proposals to choose from (see Table 1). The considerations included in Table 3 reduce this number to 10 normalization methods. The choice still becomes arbitrary and difficult to justify. The proposed approach does not completely solve the problem, but it allows distinguishing groups of normalization methods leading to similar results of linear ordering (rankings) of objects. In the analysed example, we already have three types of normalization methods to choose from (normalization methods in the same groups give similar results of linear ordering of objects). Therefore, the presented proposal allows limiting the problem of selecting the normalization method of variable values.

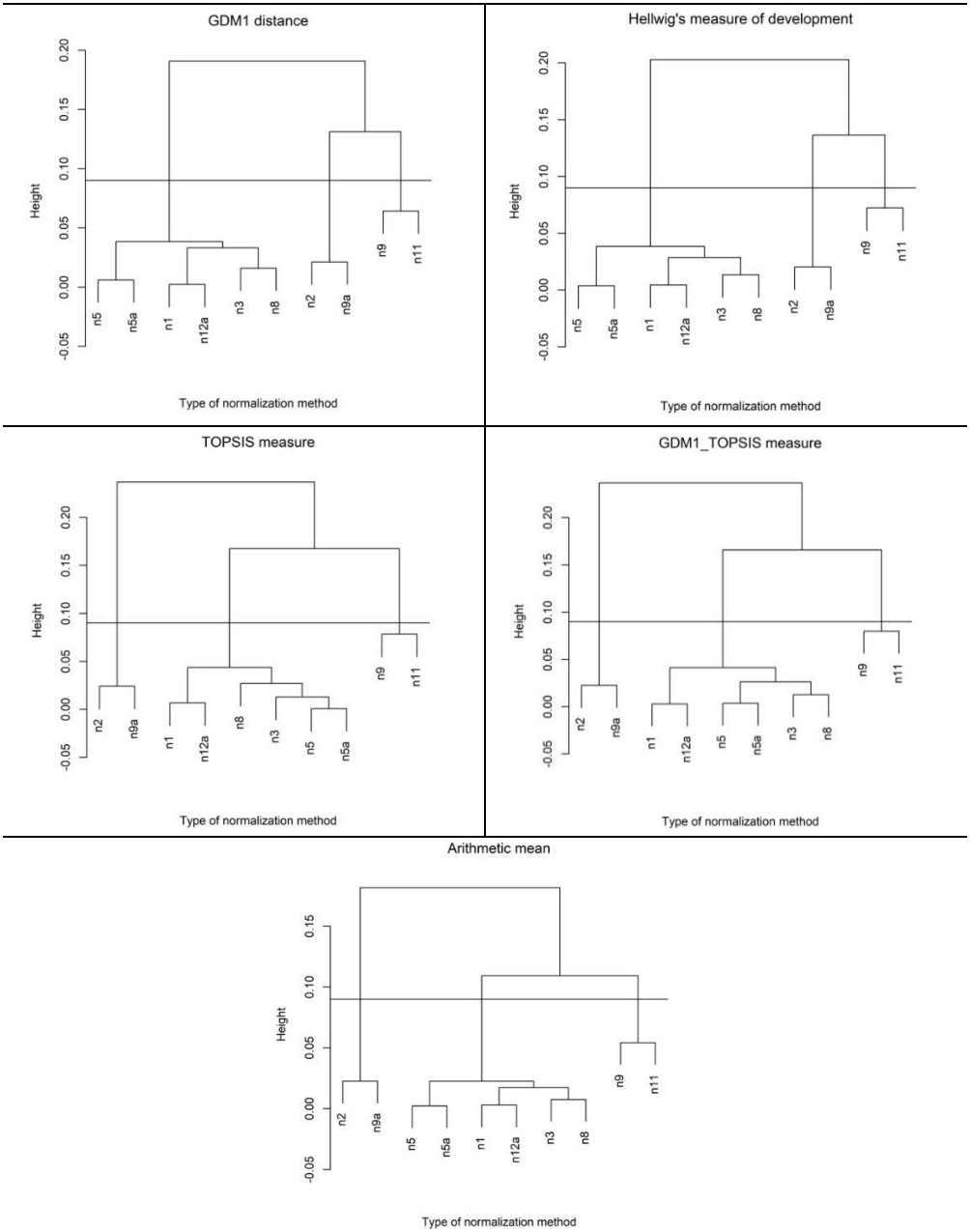


Figure 2. The results of linear ordering compatibility for 52 Sudety communes in terms of their tourism competitiveness level using 10 normalization methods and 5 aggregation measures (dendrograms of normalization methods similarity)

Source: Author's compilation using R program.

6. Conclusions

Normalization methods lead to different rankings of the set of objects based on aggregation measure (composite indicator) values. The study includes 18 normalization methods and 5 aggregation measures (composite indicators).

The groups of normalization methods were indicated, which results in identical SM_i values and identical orderings for SM_i obtained using the following distance measures: GDM1, Hellwig's measure of development, TOPSIS, GDM1_TOPSIS and aggregation measure (SM_i) taking the form of arithmetic mean. Due to the fact that the groups of A, B, C and D normalization methods result in identical ordering (see Table 3), further analysis covered 10 methods of normalization: n1, n2, n3, n5, n5a, n8, n9, n9a, n11, n12a.

The article discusses the proposal of research procedure (section 3), based on Kendall's tau correlation coefficient and cluster analysis, which allows reducing the problem of normalization method selection for variable values.

The effects of simulation studies for 5 aggregation measures and 10 normalization methods were presented (section 4). Mean values of the adjusted Rand index taken from five models are in the interval $[0.845, 0.908]$. Therefore, the results of cluster analysis of normalization methods for the analysed aggregation measures are similar to each other (dendrite of cluster analysis results' similarity of normalization methods for five aggregation measures is presented in Figure 1).

The results of conducted research were illustrated by an empirical example presenting the application of five aggregation measures and ten normalization methods in linear ordering of Lower Silesian districts in terms of their tourism attractiveness level. Based on the analysis of the dendrograms three groups of normalization methods were separated. The biggest differences in the results of linear ordering refer to methods n2, n9a against the other normalization methods.

The author's own scripts, prepared in R environment, were applied in the calculations.

REFERENCES

BĄK, A., (1999). Modelowanie symulacyjne wybranych algorytmów wielowymiarowej analizy porównawczej w języku C++ [Simulation modeling of selected algorithms of multivariate comparative analysis with C++ language], Wrocław: Wydawnictwo Akademii Ekonomicznej we Wrocławiu, ISBN: 8370114016.

BORG, I., GROENEN, P. J. F., (2005). Modern multidimensional scaling, New York: Springer. ISBN: 978-0387-25150-9, http://dx.doi.org/10.1007/0-387-28981-X.

BORYS, T., (1984). Kategoria jakości w statystycznej analizie porównawczej [Category of quality in statistical comparative analysis], Prace Naukowe Akademii Ekonomicznej we Wrocławiu No. 284, Series: Monografie i opracowania No. 23. ISBN: 83-7011-000-0.

EVERITT, B. S., LANDAU, S., LEESE, M., STAHL, D., (2011). Cluster analysis, Chichester: Wiley, ISBN: 978-0-470-74991-3.

GENZ, A., AZZALINI, A., (2016). mnormt: The Multivariate Normal and t Distributions. *R package*, version 1.5-5, https://CRAN.R-project.org/package=mnormt.

GRABIŃSKI, T., (1984). Wielowymiarowa analiza porównawcza w badaniach dynamiki zjawisk ekonomicznych [Multivariate comparative analysis in research over the dynamics of economic phenomena], Zeszyty Naukowe Akademii Ekonomicznej w Krakowie, Special series: Monografie No. 61, ISSN: 0209-1674.

GRABIŃSKI, T., (1992). Metody taksonometrii [Taxonometric methods], Kraków: Wydawnictwo Akademii Ekonomicznej w Krakowie.

GRABIŃSKI, T., WYDYMUS, S., ZELIAŚ, A., (1989). Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych [Numerical taxonomy methods in modeling socioeconomic phenomena], Warszawa: PWN, ISBN 83-208-0042-0.

GRYSZEL, P., WALESIAK, M., (2018). The application of selected multivariate statistical methods for the evaluation of tourism competitiveness of the Sudety communes, Argumenta Oeconomica, No. 1 (40), pp. 147–166, https://doi.org/10.15611/aoe.2018.1.06.

HELLWIG, Z., (1968). Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju i strukturę wykwalifikowanych kadr [Procedure of evaluating high level manpower data and typology of countries by means of the taxonomic method], Przegląd Statystyczny, Tom 15, z. 4, pp. 307–327.

HELLWIG, Z., (1972). Procedure of Evaluating High-Level Manpower Data and Typology of Countries by Means of the Taxonomic Method, [in:] Gostkowski Z. (ed.), Towards a system of Human Resources Indicators for Less Developed Countries, Papers Prepared for UNESCO Research Project, Ossolineum, The Polish Academy of Sciences Press, Wrocław, pp. 115–134.

HELLWIG, Z., (1976). Przechodniość relacji skorelowania zmiennych losowych i płynące stąd wnioski ekonometryczne [Transitivity of correlation and some econometric implications], Przegląd Statystyczny, Tom 23, z. 1, pp. 3–20.

HELLWIG, Z., (1981). Wielowymiarowa analiza porównawcza i jej zastosowanie w badaniach wielocechowych obiektów gospodarczych [Multivariate comparative analysis and applications in research of multifeature economic objects], In: W. Welfe (ed.), Metody i modele ekonomiczno-matematyczne w doskonaleniu zarządzania gospodarką socjalistyczną [Economic and mathematical methods and models in the improvement of socialist economy management], Warszawa: PWE, 46-68. ISBN 83-208-0042-0.

HUBERT, L., ARABIE, P., (1985). Comparing partitions, Journal of Classification, No. 1, pp. 193–218.

HWANG, C. L., YOON, K., (1981). Multiple attribute decision making – methods and applications. A state-of-the-art. Survey, New York: Springer-Verlag. ISBN: 978-3-540-10558-9, http://dx.doi.org/10.1007/978-3-642-48318-9.

JAJUGA, K., WALESIAK, M., (2000). Standardisation of Data Set under Different Measurement Scales, In: Decker, R., Gaul, W., (Eds.), Classification and Information Processing at the Turn of the Millennium, pp. 105–112, Springer-Verlag, Berlin, Heidelberg, http://dx.doi.org/10.1007/978-3-642-57280-7_11.

JAJUGA, K., WALESIAK, M., BĄK, A., (2003). On the General Distance Measure, in Schwaiger, M., Opitz, O., (Eds.), *Exploratory Data Analysis in Empirical Research*. Berlin, Heidelberg: Springer-Verlag, pp. 104–109, https://dx.doi.org/10.1007/978-3-642-55721-7_12.

KENDALL, M. G., (1955). Rank correlation methods, London: Griffin.

KENDALL, M. G., BUCKLAND, W. R., (1986). Słownik terminów statystycznych [A dictionary of statistical terms], Warszawa: PWE, ISBN: 83-208-0504-X.

MILLIGAN, G. W., COOPER, M. C., (1988). A study of standardization of variables in cluster analysis, Journal of Classification, Vol. 5, No. 2, pp. 181–204.

NARDO, M., SAISANA, M., SALTELLI, A., TARANTOLA, S., HOFFMANN, A., GIOVANNINI, E., (2008). Handbook on Constructing Composite Indicators. Methodology and User Guide, Paris: OECD Publishing, ISBN: 978-92-64-04345-9.

PAWEŁEK, B., (2008). Metody normalizacji zmiennych w badaniach porównawczych złożonych zjawisk ekonomicznych [Normalization of variables methods in comparative research on complex economic phenomena], Kraków: Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie, ISBN: 978-83-7252-398-3.

R CORE TEAM, (2018). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria,

https://cran.r-project.org.

STEVENS, S. S., (1946). On the theory of scales of measurement, Science, Vol. 103, No. 2684, pp. 677–680.

WALESIAK, M., (1995). The analysis of factors influencing the choice of the methods in the statistical analysis of marketing data, Statistics in Transition, June, Vol. 2, No. 2, pp. 185–194.

WALESIAK, M., (2002). Uogólniona miara odległości w statystycznej analizie wielowymiarowej [The Generalized distance measure in multivariate statistical analysis], Wrocław: Wydawnictwo Akademii Ekonomicznej we Wrocławiu, ISBN: 83-7011-583-7.

WALESIAK, M., (2011). Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R [The Generalized distance measure GDM in multivariate statistical analysis with R], Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, ISBN: 978-83-7695-132-4.

WALESIAK M., (2014a). Wzmacnianie skali pomiaru w statystycznej analizie wielowymiarowej [Reinforcing measurement scale for ordinal data in multivariate statistical analysis], Taksonomia 22, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, No. 327, pp. 60–68.

WALESIAK M., (2014b). Przegląd formuł normalizacji wartości zmiennych oraz ich własności w statystycznej analizie wielowymiarowej [Data normalization in multivariate data analysis. An overview and properties], Przegląd Statystyczny, Tom 61, z. 4, pp. 363–372.

WALESIAK M., DUDEK A., (2018). clusterSim: Searching for Optimal Clustering Procedure for a Data Set. *R package*, version 0.47-2, http://CRAN.R-project.org/package=clusterSim.

ZELIAŚ, A., (2002). Some notes on the selection of normalization of diagnostic variables, Statistics in Transition, Vol. 5, No. 5, pp. 787–802.

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 711–723, DOI 10.21307/stattrans-2018-037

THE COMPARISON OF INCOME DISTRIBUTIONS FOR WOMEN AND MEN IN POLAND USING SEMIPARAMETRIC REWEIGHTING APPROACH

Dominika Marta Urbańczyk[1], **Joanna Małgorzata Landmesser**[2]

ABSTRACT

In this paper, we compare the income distributions for women and men in Poland. The gender wage gap can only be partially explained by different men's and women's characteristics. The unexplained part of the gap is usually attributed to the wage discrimination. The objective of the study is to extend the Oaxaca-Blinder decomposition procedure for the pay gap along the whole income distribution. To describe differences between two distributions of incomes we use a semiparametric reweighting approach (DiNardo, Fortin, Lemieux, 1996). The reweighting factor is computed for each observation by estimating a logit model for probabilities of belonging to men's or women's group. Then, we estimate probability density functions, including the counterfactual density function, using kernel density methods. This allows us to decompose the inequalities into the explained and unexplained components. The analysis is based on the EU-SILC data for Poland in 2014.

Key words: gender wage gap, differences in distributions, decomposition methods.

1. Introduction

There is now a growing number of papers analysing the differences in income distributions for women and men. The past studies in Poland were mostly focused on a simple comparison of average values for incomes by using the Oaxaca-Blinder method. The findings of these studies show that males earn substantially higher wages than females (e.g. Słoczyński, 2012; Śliwicki, Ryczkowski, 2014). Differences in income distributions have been studied by Newell, Socha (2005), Rokicka, Ruzik (2010), Landmesser, Karpio, Łukasiewicz (2015), Landmesser (2016). They utilized such a decomposition method as a quantile regression method (Machado, Mata, 2005). The obtained results showed that differences between wages of men and women are the biggest in the right part of the distribution. Also, the other methodological approaches have been suggested in the economic decomposition literature: the residual imputation approach (Juhn, Murphy, Pierce, 1993), hazard model approach (Donald, Green, Paarsch, 2000),

[1] Warsaw University of Life Science, Poland. E-mail: dominika_urbanczyk@sggw.pl.
[2] Warsaw University of Life Science, Poland. E-mail: joanna_landmesser@sggw.pl.

RIF-regression (recentered influence function) method (Firpo, Fortin, Lemieux, 2009).

The objective of this study is to extend the income gap analysis to the whole distribution and to decompose the income inequalities between women and men in Poland into the explained and unexplained components. In our paper, we suggest to describe differences between two distributions using a semiparametric reweighting approach proposed by DiNardo, Fortin, Lemieux (1996). In this method the counterfactual density function is estimated employing the reweighting factor. The analysis will be based on the EU-SILC data for Poland in 2014.

2. Analysis method

This section outlines the applied methodology. First, the Oaxaca-Blinder decomposition of mean wages differences is presented. Then, we explain the idea of the reweighting approach to the decomposition that allows analysing the differences along the whole distribution.

2.1. Oaxaca-Blinder Decomposition of Mean Wages Differences

The Oaxaca-Blinder decomposition method may be applied whenever there is a need to explain the differences between the expected values of dependent variable in two comparison groups (Oaxaca, 1973; Blinder, 1973).

Let two groups *A* and *B*, an outcome variable *y* and a set of predictors *X* be given. In this case the variable *y* may present log wages and predictors *X* may concern such individual characteristics of people as age, education level or work experience. The expected value of y conditionally on X is a linear function of X:

$$y_g = X_g \beta_g + v_g, \qquad g = A, B, \tag{1}$$

where X_g are the characteristics of people in group *g* and β_g are the coefficients related to these characteristics. The estimated expected value of income $\hat{y}$ in each group is:

$$\hat{y}_g = X_g \hat{\beta}_g, \qquad g = A, B \tag{2}$$

The idea of the Oaxaca-Blinder decomposition of the difference between expected values of incomes in each of groups $\hat{y}_A$ and $\hat{y}_B$ is as follows:

$$\hat{\Delta}^{\mu} = \overline{X}_A \hat{\beta}_A - \overline{X}_B \hat{\beta}_B = \underbrace{(\overline{X}_A - \overline{X}_B)\hat{\beta}_A}_{\hat{\Delta}^{\mu}_{\text{explained}}} + \underbrace{\overline{X}_B(\hat{\beta}_A - \hat{\beta}_B)}_{\hat{\Delta}^{\mu}_{\text{unexplained}}} \tag{3}$$

The above equation is based on one group's characteristics and the estimated coefficients of another group's equation. The first term on the right-hand side of the equation gives the effect of characteristics and expresses the

difference of the potentials of both groups (the so-called explained, endowments or composition effect). The second term represents the effect of differences in the estimated parameters (unexplained by characteristics of groups). This is typically interpreted as discrimination.

One important disadvantage of the Oaxaca-Blinder decomposition method is that it focuses only on average effects, and this may lead to a misleading assessment if the effects of covariates vary along the entire distribution (Salardi, 2012).

2.2. Decomposition Along the Entire Distribution

The idea to avoid the drawback of the Oaxaca-Blinder decomposition method may be to extend the mean decomposition to the case of differences between distributions or density functions of income. This approach is the basis of most decomposition methods. It requires the counterfactual distribution to be considered. In general, the counterfactual distribution is interpreted as a distribution for people from group *B* if they were described by characteristics of people from group *A* (in our case this is the distribution of income for women with characteristics of men).

In terms of density functions the difference can be expressed as follows:

$$\hat{\Delta}^{f} = \hat{f}_M(y) - \hat{f}_W(y) = \underbrace{[\hat{f}_M(y) - \hat{f}_C(y)]}_{\hat{\Delta}^{\mu}_{\text{explained (structureeffect)}}} + \underbrace{[\hat{f}_C(y) - \hat{f}_W(y)]}_{\hat{\Delta}^{\mu}_{\text{unexplained (composition effect)}}} \tag{4}$$

where $f_M(y)$ is the density function of income for men, $f_W(y)$ and $f_C(y)$ are the density functions for women and counterfactual distribution respectively.

In turn, the application of the cumulative distribution function of incomes allows writing the difference between the men and women density function of income $\hat{F}_M(y) - \hat{F}_W(y)$ with the counterfactual distribution $\hat{F}_C(y)$ in the following form:

$$\hat{\Delta}^{F} = \hat{F}_M(y) - \hat{F}_W(y) = \underbrace{[\hat{F}_M(y) - \hat{F}_C(y)]}_{\hat{\Delta}^{\mu}_{\text{explained (structureeffect)}}} + \underbrace{[\hat{F}_C(y) - \hat{F}_W(y)]}_{\hat{\Delta}^{\mu}_{\text{unexplained (composition effect)}}} \tag{5}$$

2.3. Semiparametric Reweighting Approach

The semiparametric reweighting approach to the decomposition of distribution differences was introduced by DiNardo, Fortin and Lemieux in 1996 (DiNardo, Fortin, Lemieux, 1996). The method allows performing the decomposition of differences along the entire distributions in terms of density function (according to expression (4)).

The method requires the estimation of probability density functions for groups and for the counterfactual distribution. For this purpose, the kernel density

estimation methods are applied. The kernel estimator of the density function for each group (in the case *g* = *W* for women and *g* = *M* for men) is as follows:

$$\hat{f}_g(y) = \frac{1}{h \cdot N_g} \sum_{i \in g} K\left(\frac{Y_i - y}{h}\right) \tag{6}$$

where *K* is the kernel function, *N* is the number of people in the group and *h* is a smoothing parameter called bandwidth. The value of *h* is chosen to minimize the mean squared error. In this method the counterfactual density function is also estimated employing the kernel density estimation but, additionally, the reweighting factor $\hat{\Psi}(X)$ is required. Then, the kernel density estimator for the counterfactual density is:

$$\hat{f}_C(y) = \frac{1}{h \cdot N_W} \sum_{i \in W} \hat{\Psi}(X_i) K\left(\frac{Y_i - y}{h}\right) \tag{7}$$

The counterfactual distribution interpretation in the reweighting approach is different than in most decomposition methods. In this case, the counterfactual distribution is the distribution for women that consists of the influence of the whole sample characteristics.

The impact of the characteristics of the whole sample is ensured by the construction of the reweighting factor $\hat{\Psi}(X)$, which is defined as (Fortin, Lemieux, Firpo, 2010):

$$\hat{\Psi}(X) = \frac{d\hat{F}_{X_M}(X)}{d\hat{F}_{X_W}(X)} = \frac{\hat{P}(X|D_M = 1)}{\hat{P}(X|D_M = 0)} \tag{8}$$

where D_M = 1means that the person is a man and D_M = 0 is a woman.

By applying Bayes' rule the reweighting factor can be written as:

$$\hat{\Psi}(X) = \frac{\hat{P}(D_M = 1|X) / \hat{P}(D_M = 1)}{\hat{P}(D_M = 0|X) / \hat{P}(D_M = 0)} \tag{9}$$

The reweighting factor value $\hat{\Psi}(X)$ can be computed for each observation by estimating a logit or probit model for conditional probabilities of belonging to groups *M* and *W* ($\hat{P}(D_M = 1|X)$ and $\hat{P}(D_M = 0|X) = 1 - \hat{P}(D_M = 1|X)$) and from the classical definition of probability using the sample proportions in both groups ($\hat{P}(D_M = 1)$ and $\hat{P}(D_M = 0)$).

The advantages of the reweighting approach are the opportunity to compare the differences along the whole distribution as well as simplicity and efficiency. On the other hand, a limitation of this method is that it is impossible to extend this approach to the case of the detailed decomposition due to the estimation of the logit (or probit) model.

3. Results of Empirical Analysis

This section is devoted to introduce the results of the empirical analysis. First, the data used for analysis are presented. Then, we provide estimated density functions for women and men as well as the construction of the counterfactual distribution. Finally, the results of the decomposition of the difference in incomes in both groups are discussed.

3.1. Database

We employ data from the European Union Statistics on Income and Living Conditions (EU-SILC) for Poland in 2014[3]. It is the source of microdata on income, poverty, social exclusion and living conditions. The EU-SILC belongs to the European Statistical System (ESS).

Our data consist of a sample of 4727 women and 5177 men containing information on annual income, natural logarithm annual income as well as on persons' attributes such as age, gender, marital status, education level, information if it is full-time or part-time job and other describing the type of contract. The applied variables with description and possible values are presented in the table below.

Table 1. Description of the variables

Variable	Description and possible values
age	age in years
men	sex, 1 – man, 0 – woman
married	marital status, 1 – married, 0 – unmarried
educlevel	educational level, 1 – primary, …, 5 - tertiary
parttime	1 – person working part-time, 0 – person working full-time
big	number of persons working at the local unit, 1 – more than 10 persons, 0 – less than 11 persons
permanent	type of contract, 1 – permanent job/work contract of unlimited duration, 0 – temporary contract of limited duration
manager	managerial position, 1 – supervisory, 0 – non-supervisory
yearswork	number of years spent in paid work
income	gross annual income in € (including benefits)
ln_income	natural logarithm gross annual income in €

3.2. Density functions of income for men and women

We apply the kernel estimation method to obtain the density function of income for women and men. In our analysis the logarithm of the annual income is the outcome variable. Two kernel functions – Epanechnikov and Gaussian – are

[3] This database was obtained under Eurostat project number 234/2016-EU-SILC.

applied. We prefer Epanechnikov kernel for the reason it is optimal in a mean square error sense (Epanechnikov, 1969). The kernel function is as follows:

$$K = \begin{cases} \frac{3}{4}\left(1 - x^2\right) & x \in [-1, 1] \\ 0 & x \notin [-1, 1] \end{cases} \tag{10}$$

The estimated density functions of income for men and women are compared in Figure 1a. The income distribution for men is shifted to the higher values of the logarithm of income related to the distribution for women. This fact may be interpreted as meaning that men earn more than women.

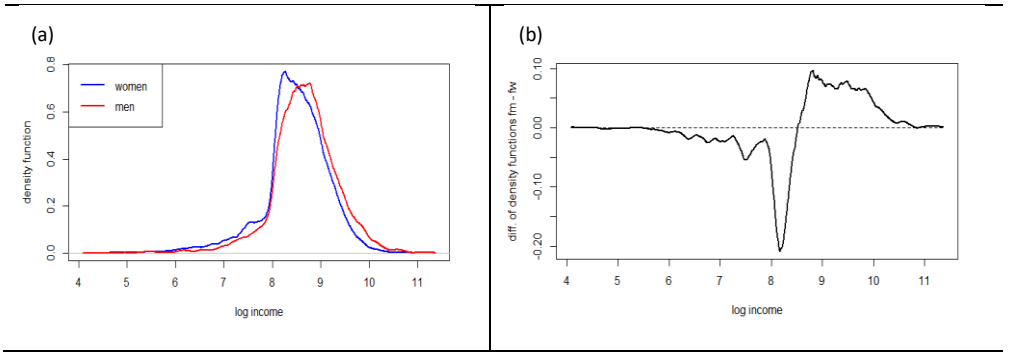


Figure 1. The estimated density functions of the logarithm of income for women and men (a) and the difference between the density function of men and women (b)

The difference between the density function for men and women $\hat{\Delta}^f = \hat{f}_M(y) - \hat{f}_W(y)$ is presented in Figure 1b. We can see a greater participation of women in the case of lower wages. On the other hand, there are more men for higher values of income. This is also the evidence that men earn more.

3.3. Reweighting Factor Computation

For the aim of the estimation of the counterfactual density function, the reweighting factor $\hat{\Psi}(X)$ is required. It may be written as in formula (9):

$$\hat{\Psi}(X) = \frac{\hat{P}(D_M = 1 | X) / \hat{P}(D_M = 1)}{\hat{P}(D_M = 0 | X) / \hat{P}(D_M = 0)}$$

The probabilities $\hat{P}(D_M = 1)$ and $\hat{P}(D_M = 0) = 1 - \hat{P}(D_M = 1)$ are computed from the classical definition of probability using the groups and sample size as follows: $\hat{P}(D_M = 1) = \frac{5177}{9904} \approx 0{,}5227$ and $\hat{P}(D_M = 0) = 1 - \frac{5177}{9904} \approx 0{,}4773$.

To determine the conditional probability $\hat{P}(D_M = 1|X)$, the logit model is estimated. The logarithm of the maximum likelihood function is −6359.066, AIC = 12736. In the Likelihood ratio test the hypothesis that model coefficients are equal to 0 was rejected (*p-value* < $2.2 \cdot 10^{-16}$). The estimated parameters for each of variables are presented in Table 2.

Table 2. Results of logit model estimation

Variable	Parameter	p-value
age	-0.089004	< 2e-16***
educlevel	-0.422461	< 2e-16***
married	0.095614	0.05466 .
yearswork	0.085015	< 2e-16***
permanent	-0.143589	0.00505 **
parttime	-0.861151	< 2e-16***
manager	0.488480	< 2e-16***
big	0.162550	0.00378 **
constance	3.593760	< 2e-16***

where significance levels codes are as follows: *** 0,001; ** 0,01; * 0,05; . 0,1.

Based on the above results, it can be easily seen that all the variables in the model are statistically significant. The positive values of parameters indicate that an increase in the value of the corresponding variable increases the probability that the person is a man with the fixed values of the other variables. The interpretation of negative parameter values is analogical.

In this way the conditional probability $\hat{P}(D_M = 1|X)$ is estimated by the logit model. Using probability values, obtained as described above, the reweighting factor is computed separately for each person from the sample.

3.4. Counterfactual Distribution and Decomposition for Density Functions

In the next step, using the reweighting factor obtained earlier for each person in the sample, we estimate the counterfactual distribution. It is worth emphasizing that the interpretation of the counterfactual distribution is different in comparison with typical decomposition methods. In most approaches the counterfactual distribution mixes the distribution of outcome variable *Y* for women and explanatory variables *X* for men. In this case the counterfactual distribution may be understood as the distribution for women reweighted by the effect of characteristics of both groups, which is contained in the reweighting factor.

We also apply the kernel estimation method with Epanechnikov kernel to obtain the density function for the counterfactual distribution. The estimated density functions of logarithm of income for women, men and counterfactual distribution are presented in Figure 2.

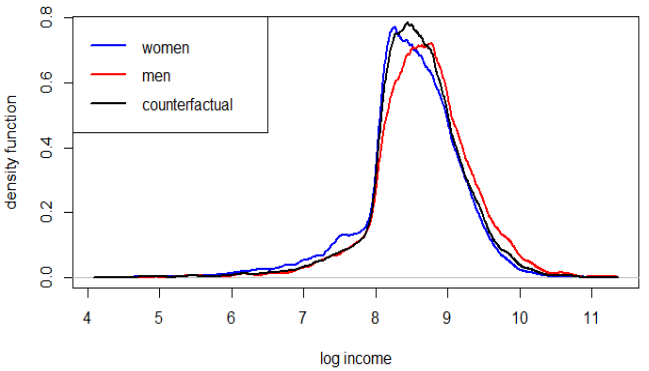


Figure 2. The estimated density function of the counterfactual distribution in comparison with the density functions of the logarithm of income for women and men

Subsequently, we decompose the inequalities of income in men's and women's group into the explained and unexplained components. This procedure is performed in terms of probability density functions, which allows for the analysis along the entire distribution. The results illustrating the formula (4) are presented in Figure 3.

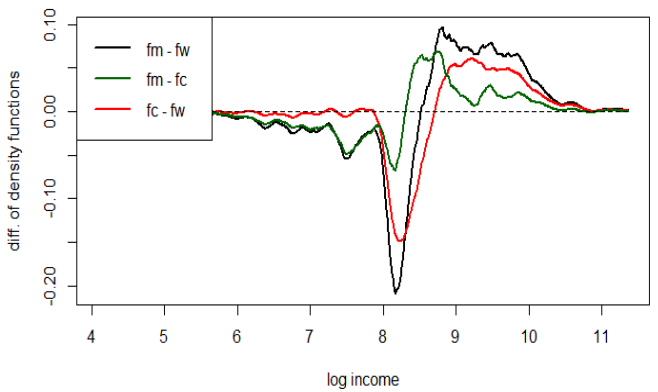


Figure 3. The results of the decomposition of income inequalities for men and women. The explained and unexplained components are indicated respectively by green and red line

Analyzing the results of the decomposition, it is easy to notice that the dominance of women in the group with lower incomes is explained. This may be related to the fact that women are much more likely to work part-time than men. On the other hand, the dominance of men in the group of the higher income is mainly due to the discrimination. It is worth taking into account that the significant dominance of men is explained only for the values of the logarithm of wages from 8 to 9, which corresponds to the income of 3000 to 8000 €. Moreover, the occurrence of the unexplained part leads to the shift of the distribution for men into higher incomes. However, it should be noticed that the fact of including benefits in the gross annual income increases the gender pay gap in the upper quantiles of the distributions. In general, the better-paid men receive higher bonuses.

3.5. Distribution Function and Decomposition for Quantiles

It is worth considering that the comparison of distributions in terms of probability density functions gives only a partial insight into the analysis of the wage gap. The decomposition of differences in distributions using quantiles allows considering the income inequalities completely.

Using the estimated density functions, the cumulative distribution functions (CDFs) may be determined by the trapezoidal numerical integration method. Figure 4 presents the cumulative distribution functions. The cumulative distribution function curve for women's income is above this for the men's one. From this fact, and on the basis of CDF definition, we can conclude that women earn less.

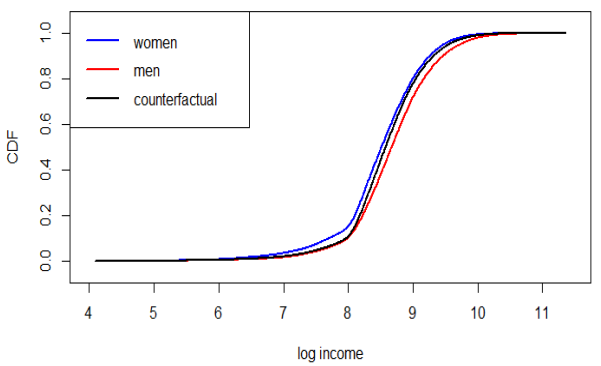


Figure 4. The distribution functions for women's and men's income as well as the counterfactual distribution

In the next step, the quantiles for distributions of men's, women's income and counterfactual distribution are determined. The precise values of quantiles $\hat{Q}_{g,\tau} = \hat{F}_{Y_g}^{-1}(\tau)$ are computed by linear interpolation. This allows decomposing the wage gap for quantiles. The results are presented in Table 3.

Table 3. Decomposition of difference in income distributions in terms of quantiles

τ	$\hat{Q}_W$	$\hat{Q}_M$	$\hat{Q}_C$	$\hat{Q}_M - \hat{Q}_W$	$\hat{Q}_M - \hat{Q}_C$	$\hat{Q}_C - \hat{Q}_W$
0.1	7.7009	7.9848	7.9629	0.2839	0.0219	0.2620
0.2	8.1009	8.2205	8.1775	0.1196	0.0430	0.0766
0.3	8.2415	8.3843	8.3154	0.1428	0.0689	0.0739
0.4	8.3745	8.5297	8.4458	0.1552	0.0840	0.0713
0.5	8.5127	8.6704	8.5756	0.1577	0.0948	0.0629
0.6	8.6574	8.8102	8.7119	0.1528	0.0984	0.0544
0.7	8.8140	8.9642	8.8588	0.1502	0.1054	0.0448
0.8	8.9952	9.1635	9.0403	0.1683	0.1232	0.0451
0.9	9.2563	9.4603	9.3180	0.2040	0.1423	0.0616
1	10.7675	11.1478	10.7270	0.3803	0.4208	-0.0405

This approach also allows determining the explained and unexplained components of the difference in terms of quantiles (see Table 4). For an easier analysis, the logarithmic values are converted to income in euro.

We can see that for a quantile of the order of 0.1 the difference is high. This is also accompanied by a large share of the explained part in the wage gap. This may be due to the much greater share of women working part-time than men. Starting with the quantile of the order of 0.2 the wage gap grows with the order of the quantile as well as with the amount of income.

Table 4. Wage gap for women's and men's group and share of explained and unexplained part of difference

τ	$\hat{Q}_W$ [€]	$\hat{Q}_M$ [€]	$\hat{Q}_M - \hat{Q}_W$ [€]	unexplained part [%]	explained part [%]
0.1	2210.32	2935.94	725.62	7.72%	92.28%
0.2	3297.41	3716.50	419.09	35.94%	64.06%
0.3	3795.28	4377.97	582.69	48.24%	51.76%
0.4	4335.02	5063.03	728.02	54.08%	45.92%
0.5	4977.52	5827.94	850.42	60.13%	39.87%
0.6	5752.78	6702.53	949.75	64.37%	35.63%
0.7	6727.68	7817.99	1090.31	70.15%	29.85%
0.8	8064.33	9542.33	1478.00	73.21%	26.79%
0.9	10470.53	12839.69	2369.16	69.78%	30.22%
1	47453.68	69411.97	21958.30	110.65%	-10.65%

It is also worth noticing that the unexplained component of the wage gap increases with the amount of income. This demonstrates that the discrimination is

more evident for higher values of wages. The interesting result is the negative value of the explained component of the income difference in the group of the best earning people. It may be associated with the fact that women in this group should earn more than men. However, it is worth mentioning that there is far fewer people in this group in comparison with the others (for the reason there are more people having incomes about mean level than in the tail of income distribution), which causes that the result may be misleading.

4. Conclusions

The aim of this study was to perform a decomposition of income inequalities between women and men. It was achieved by using the semiparametric reweighting DFL method (DiNardo, Fortin, Lemieux, 1996). It allows extending the income gap analysis to the whole distribution rather than just the average level of wages as in the case of the Oaxaca-Blinder decomposition method. Furthermore, the chosen approach leads to more accurate results than the Oaxaca-Blinder decomposition method for the average value because the DFL decomposition method is not based on the linear regression.

The major drawback of the applied method is that it is not suitable for the detailed (taking into account the individual explanatory variables) decomposition of inequalities between distributions. This is because all of these variables are included during the estimation of the logit model.

In this work the decomposition of the wage gap between women and men was performed in terms of the density function. Moreover, the explained and unexplained (associated with discrimination) components of the difference were determined. Furthermore, for the aim of the more accurate analysis of the inequalities in women's and men's incomes, the cumulative distribution functions and quantiles were calculated. This allowed decomposing the wage gap in terms of quantiles and the "horizontal analysis" of differences between distributions.

In the light of the results obtained, we found that the share of the unexplained part of inequalities is higher than the explained one and it tends to increase with the rising values of income. This is the evidence that the discrimination in wages is significant. However, it should be noticed that this study was based only on factors from EU-SILC database. The inclusion in the model of the additional explanatory variables, describing in more detail the job position or the employment environment, could influence the results and lead to reduction of the unexplained component. In addition, we should be aware of the effect of the increase in the wage gap in the upper part of earnings distribution by including higher bonuses in annual income.

The obtained results are consistent with those for Poland reported in the literature. Other researchers also notice the higher level of incomes for men (Kompa, Witkowska, 2013; Matuszewska-Janica, 2014; Witkowska, 2014). A significant unexplained part of the wage gap and the larger inequality at the top of distribution are observed (Śliwicki, Ryczkowski, 2014; Rokicka, Ruzik, 2010).

It is worth considering performing an analogous analysis of the difference in income distributions for women and men according to the individual levels of education. The expected result is to obtain information about the relation between the level of employees' education level and the occurrence of discrimination.

REFERENCES

BLINDER, A., (1973). Wage Discrimination: Reduced Form and Structural Estimates. Journal of Human Resources, 8, pp. 436–455.

DINARDO, J., FORTIN, N. M., LEMIEUX, T., (1996). Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. Econometrica, 64, pp. 1001–1044.

DONALD, S. G., GREEN, D. A., PAARSCH, H. J., (2000). Differences in Wage Distributions between Canada and the United States: An Application of a Flexible Estimator of Distribution Functions in the Presence of Covariates. Review of Economic Studies, 67 (4), pp. 609–633.

EPANECHNIKOV, V. A., (1969). Non-parametric estimation of a multivariate probability density. Theory of Probability and its Applications, 14, pp. 153–158.

FIRPO, S., FORTIN, N. M., LEMIEUX, T., (2009). Unconditional Quantile Regressions. Econometrica, 77 (3), pp. 953–973.

FORTIN, N., LEMIEUX, T., FIRPO, S., (2010). Decomposition methods in economics. Cambridge: NBER Working Paper, No. 16045.

JUHN, CH., MURPHY, K. M., PIERCE, B., (1993). Wage Inequality and the Rise in Returns to Skill. Journal of Political Economy, 101, pp. 410–442.

KOMPA, K., WITKOWSKA, D., (2013). Application of Classification Trees to Analyze Income Distribution in Poland. Quantitative Methods in Economics, XIV (1), pp. 265–275.

LANDMESSER, J. M., KARPIO, K., ŁUKASIEWICZ, P., (2015). Decomposition of Differences Between Personal Incomes Distributions in Poland. Quantitative Methods in Economics, XVI (2), pp. 43–52.

LANDMESSER, J., (2016). Decomposition of Differences in Income Distributions Using Quantile Regression, Statistics in Transition - new series, Vol. 17, No. 2, pp. 331–348.

MACHADO, J. F., MATA, J., (2005). Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression. Journal of Applied Econometrics, 20, pp. 445–465.

MATUSZEWSKA-JANICA, A., (2014). Wages Inequalities Between Men and Women: Eurostat Ses Metadata Analysis Applying Econometric Models. Quantitative Methods in Economics, XV (1), pp. 113–124.

NEWELL, A., SOCHA, M., (2005). The Distribution of Wages in Poland. IZA Discussion Paper, No. 1485, Bonn.

OAXACA, R., (1973). Male-Female Wage Differentials in Urban Labor Markets. International Economic Review, 14, pp. 693–709.

ROKICKA, M., RUZIK, A., (2010). The Gender Pay Gap in Informal Employment in Poland. CASE Network Studies and Analyses, No. 406, Warsaw.

SALARDI, P., (2012). Wage Disparities and Occupational Intensity by Gender and Race in Brazil: An Empirical Analysis using Quantile Decomposition Techniques. Brighton: University of Sussex, Job Market Paper.

SŁOCZYŃSKI, T., (2012). Wokół międzynarodowego zróżnicowania międzypłciowej luki płacowej. International Journal of Management and Economics, 34, pp. 169–185.

ŚLIWICKI, D., RYCZKOWSKI, M., (2014). Gender Pay Gap in the micro level – case of Poland. Quantitative Methods in Economics, XV (1), pp. 159–173.

WITKOWSKA, D., (2014). Determinants of Wages in Poland. Quantitative Methods in Economics, XV (1), pp. 192–208.

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 725–742, DOI 10.21307/stattrans-2018-038

IS POLAND BECOMING NORDIC? CHANGING TRENDS IN HOUSEHOLD STRUCTURES IN POLAND AND FINLAND WITH THE EMPHASIS ON PEOPLE LIVING ALONE

Urszula Ala-Karvia[1], **Marta Hozer-Koćmiel**[2], **Sandra Misiak-Kwit**[3], **Barbara Staszko**[4]

ABSTRACT

This paper presents a comparative analysis of the household structure and its dynamics between post-economic-transformation Poland and Scandinavian-welfare-state Finland, with a focus on one-person households (OPH). Based on the literature, two research hypotheses were formulated: (H1) strong differences in the household structure in Finland and Poland still occur, and (H2) the share of one-person households is at very different levels in the two countries. However, due to the globally growing popularity of solo living, the difference is diminishing. Finally, an estimate was made for the time when the shares of one-person households will be equal in both countries if the changing trends from 2005–2015 stay the same.

The first research hypothesis was proven to be correct. Small, one- or two-person households dominate the household structure in Finland, while in Poland the household structure by size was considerably more balanced. The second hypothesis was confirmed only partially. The share of OPH among all the households in 2015 was significantly larger in Finland (42%) than in Poland (24%). However, the difference between the countries was not diminishing. The share in Finland is increasing, while it is decreasing in Poland. This allowed the assumption that if the changing trends from the studied period are maintained, the shares of OPH in the two countries will not equalize, but will instead grow further apart. An estimate was made that in 2030 46% of Finnish households and 22% of Polish households will be one-person households.

Key words: household structure, people living alone, one-person households, comparative analyses and forecast.

[1] University of Helsinki, Ruralia Institute, Finland. E-mail: urszula.ala-karvia@helsinki.fi.
[2] University of Szczecin, Faculty of Economics and Management, Department of Statistics, Poland. E-mail: mhk@wneiz.pl.
[3] University of Szczecin, Faculty of Economics and Management, Department of Human Capital Management, Poland . E-mail: s.misiak@wneiz.pl.
[4] University of Szczecin, Faculty of Economics and Management, Poland. E-mail: bstaszko@wp.pl.

1. Introduction

Many social, economic and demographic processes are becoming similar among different European countries. The countries that underwent economic transformation from a centrally planned to a market-oriented economy in the 1980s and 1990s have experienced particularly strong structural changes. In many respects, these countries have become similar to countries in Western and Northern Europe (Batóg & Batóg, 2006; Hozer-Koćmiel & Lis, 2016; Zimoch, 2013).

This paper reviews household structures in two relatively different countries, Poland and Finland. Poland represents post-economic transformation countries, while Finland is a prosperous Nordic welfare state country. The paper compares household structures and their dynamics in Poland and Finland, with a particular focus on one-person households (OPH), whose relative percentages among all households have been growing across Europe.

Additionally, an attempt has been made to estimate when the shares of OPH will be the same in Poland and Finland if the changing trends during the analysed period 2005–2015 are maintained.

One-person households and persons living alone correspond to the same target group. The terms being used may depend on the source of the data, e.g. households or population statistics, market or demography studies. Additionally, using basic descriptive analysis both terms can be used alternatively. However, while calculating shares, it is important to distinguish the share of OPH, which is the number of OPH divided by the total number of households and the share of population living alone, standing for the number of people living alone out of a total population (aged 15+). In this paper the terms OPH and persons living alone are used to provide the reader with a maximum understanding of the changing trends in household structures.

2. Literature review

Households are a basic economic market element and they play a key role in consumption. The main purpose of a household is to fulfil all the needs of its members. ‘A household is understood as a single or multi-person economic entity, usually based on family ties, operating in the sphere of consumption, whose purpose is to meet the needs of all members, thanks to the common disposition of income earned by all or only some of them’ (Zalega, 2007, pp. 8: translated by S.M.-K.).

Similarly, according to the United Nations Economic Commission for Europe (UN, 2011), a private household is either a one-person household (i.e. a person living alone in a separate housing unit, or occupying, as a lodger, one or more separate rooms in a housing unit, but not joining with other occupants to form a multi-person household), or a multi-person household (i.e. a group of two or more persons occupying a housing unit, or a part of it, jointly providing themselves with food and other essentials). Both of the above household categories represent housekeeping concepts in which joint providing for common goods plays an essential part. UN also distinguished that countries with register-based data often

use a household-dwelling concept, instead, then the number of households and dwellings is equal.

Samuelson and Nordhaus (2005) mention that the terms family and household are often used alternatively. However, according to these authors, there are big differences between these concepts due to their different functions: the role of a family is to maintain biological and cultural continuity, while a household has economic functions, depending on the scope of its members' needs. In economics, it is assumed that these functions result from the main goal of the household, i.e. utility maximization or maximizing the fulfilment of needs (Kopycińska, 2011).

Thus, households consist of members who not only live together, but also decide and act together based on their own preferences and existing restrictions. As a statistical unit, however, a household has socio-economic rather than biological features (Latuch, 1980). Referring to Statistics Poland's definition of a household, one of the criteria distinguishing a household is its common economic management, with the condition of joint residence or family ties, thus it follows the housekeeping concept. The same source determines a one-person household as a person who is self-dependent and lives alone. In Finland, up until the 1980 census Statistics Finland (OSF) used the housekeeping concept of the household, which was then substituted by the concept of a household-dwelling unit. The household-dwelling unit consists of the permanent occupants of a dwelling. Persons classified in the Population Information System of the Population Register Centre as institutionalized, homeless or living abroad are excluded. Additionally, living in a residential home that does not meet the criteria of a dwelling (intended for year-round habitation, at least 7 m^2, furnished with at least a cooking area and its own entrance) is also not categorized as a household-dwelling unit. Statistics Finland recognizes two categories of a household dwelling: 1) family household-dwelling units that comprise one or more families, with or without other persons, or one family and other persons, and 2) other household-dwelling units, including people living alone, and two or more people of the same or different sex.

Changing trends in partnership and childbearing patterns from the last decades have influenced the household structure across Europe (Oláh, 2015). According to Eurostat (2015), a rising share of people living alone, declining fertility rates, higher divorce rates, and a shift in household structures away from multigenerational living have visibly shrunk the average size of households in the European Union in recent decades. One-person households have become a dominant household type in many regions of Europe. The dominance of OPH has been somewhat overlooked by public policy and social research, which in the last decades was focused on bigger households and on families.

Living alone is a growing trend worldwide, noticed decades ago in North America and Europe (Hall et al., 1997; Jacobsen et al., 2012). Bennet and Dixon (2006) called the rising number of people living alone one of the most important demographic shifts in recent decades. According to the following brief literature review, the increase in separate living can be seen both as a cause and as a result of changing household and family composition. There are relatively many literature sources from the 80s analysing the changes in family and household structures, which in some countries started already after the Second World War.

Next to reporting the facts that the number of people living alone has been increasing, or that the average size of a household has decreased, researchers have tried to analyse the factors that influence the choice of living alone along with other structural changes in the family. Pampel (1983) linked separate living with increased income and changes in norms and tastes as well as changes in the relationship between parents and children. While some of these linkages were more obvious than others, throughout his modelling analysis, time seemed to have the strongest additive result on propensity to live alone. The rising importance of one-person households was grouped by Keilman (1988) according to demographic factors: delayed marriage at a young age, divorce without a new relationship in middle age, lower male mortality at elder ages, and cultural factors following the above-mentioned shift in propensity to live alone. He also stated that higher living standards have made it easier for an individual person to set up an independent, solitary household. Similarly, Keyfitz and Caswell (2005) stated three supporting aspects for separate living: the desire for privacy and independence highly correlated with income; an absence of kin correlated with low fertility rates; and finally, personal changed preferences. In their probabilistic, dynamic household forecast, Christiansen and Keilman (2013) observed several features concerning the status of living alone: young people living alone are likely to enter into cohabitation; at all ages, the status from cohabiting is more likely to change toward living alone than to living with a spouse; a high increase of living alone of previously single parents in their fifties due to the adulthood of their children; living alone starting at an advanced age (e.g. after the death of the spouse) is a common state.

In its series of Statistics Explained, Eurostat (2016) published a summary of European household composition focusing on the size and types of households across 28 countries in the EU. With a timeline between 2005 and 2015, single households, i.e. people living alone, recorded the greatest increase between those years and was the most common household type (EU-28 average of 33.4% in 2015). The same publication clearly showed how living alone varies across different countries, including our countries of interest, namely Poland and Finland.

Iacovou and Skew (2011) present several indicators of the household structure in the enlarged EU. They marked Finland in the Nordic cluster of countries of the EU15 with attributes such as, on average, a small household, early residential independence of young people and extended residential independence of the elderly. Poland, on the other hand, is classified among the new member states along the Hajnal line, a line that runs from St. Petersburg (Russia) to Trieste (Italy), which historically was characterized by an early marriage and multigenerational households. They also conclude that Poland belongs to the four Eastern European countries that stand out from the rest of Europe by having the largest households, an absence of separate living among young people, extended multigenerational co-residence and relative scarcity of lone-parent families.

Also, a current paper by Habartova (2018) presents a cross-country analysis of recent household trends. Based on the 2011 census, Habartova presents the average size of households (Poland having the second highest and Finland the lowest value in Europe) and analyses particular types of households in more detail. According to cluster analysis, the household structure in Poland is similar

to the traditional structure (i.e. fewer lone parents, large family size, etc.) observed in Southern Europe (e.g. Portugal, Spain, Italy), while Finland among other Nordic countries (plus France and the Netherlands) presents, inter alia, a high intensity of new forms of living arrangements and a great number of people living alone.

Changing trends in living arrangements of men and women from the late 80s in Europe were analysed by Fokkema & Liefbroer (2008). They refer to the Second Demographic Transition developed by Lesthaeghe and Van de Kaa in the 80s, as an explanation for the weakening of the institution of the family through the strengthening economic independence of people and the rise of self-development ideologies. They summarize their findings on people living alone as very age-specific trends concentrated on the elderly, taking diverse forms in different parts of Europe. They also point out that age patterns are different between men and women, with women being, in general, less likely to live alone at a younger age and more likely at an older age.

Nowak-Sapota (2007) analysed regional differences in household structures and shares of living alone in Poland up to 2002. As a reason for separate living, next to economic factors and marital status, she pointed out that living alone does not specifically stand for being unmarried or widowed, however it is highly correlated. It is important to note that the majority of single households (over 75%) were located in urban areas. Nowak-Sopota (2008) also forecasted that in 2030, corresponding to the year 2002, the number of people living alone in Poland will increase by 55% (meaning every third household in Poland being OPH) and the majority (61.6%) of people living alone will be aged 60 and over.

Forecasting the number of households and their composition according to Alho and Keilman (2010) is an essential action from the policy perspective, for example when planning social support expenditures or evaluating the demand for new dwellings or electricity consumption. They forecasted that among all household types the share of people living alone will steadily grow, while Keilman (2016) estimated that the growth of people living alone would even be as high as 40% for the period 2011–2041 for selected European countries.

Based on the review of the literature, two hypotheses were formulated:

H1: Strong differences in household structure can still be observed in Finland and Poland,

H2: The share of OPH among all households is far larger in Finland, yet the difference was diminishing.

3. Research methods and data

Descriptive analysis was carried out based on (1) shares of OPH among all household types and structural differences among all OPH, and (2) shares of people living alone in the total population.

To measure the similarities among household structures in Poland and Finland, the Renkonen similarity index was used (Renkonen, 1938; Bąk et al., 2015) in its basic form:

$$\omega_p = \sum_{i=1}^{k} min(\omega_{1i}, \omega_{2i}), \qquad 0 \leq \omega_p \leq 1,$$

where $\omega_{1i} = {}^{n_i}/_{\sum n_i}$ is a relative (proportional) representation of characteristic n_i in the total population $\sum n_i$.

Age pyramids were used to capture the structure of the total population (see Fig.4). This simple tool presents graphically the population structure by age and sex (Holzer, 2003; Okólski, 2005).

Changing trends within the OPH structures by socio-economic characteristics were presented graphically as the difference of shares between the two study periods. The single-base increments showing changes in the shares were determined:

$$\Delta y_{t/0} = y_t - y_0,$$

where y_t is the variable value in the later observation period, and y_0 the variable value in the initial period.

Linear trend models describing changes in the shares over time were also built:

$$\hat{y}_t = a_1 t + a_0, \qquad t = 1,2, \dots, n.$$

In the above equation, $\hat{y}_t$ is the dependent variable, *t* is the time variable, and a_1 and a_0 are coefficients. Based on the trend models for both Poland and Finland, the shares of OPH until 2030 were estimated (Bąk et al., 2015; Hozer, 1997; Weinbach & Grinnell, 2007).

Data on the composition of household structures in 2005 and 2015 in Poland and Finland were obtained from Eurostat data on private households. The vital and population statistics were obtained from Eurostat, Statistics Poland and Statistics Finland. Despite having two different concepts of households, Eurostat database is a reputable source of comparable data. However, this has influenced the choice of study period for the openly available and comparable data across different themes. Additionally, in 2005, both countries were already part of the European Union, thus the descriptive comparison occurs in a similar political setting.

4. Presentation of the obtained results

The first analytical step was to answer the questions: What is the current household structure in Finland and Poland? How did it change in the last few years? And are Polish and Finnish structures similar or different? Based on official statistics, households were divided into six groups, depending on their size, from one-person households to six and more persons in a household. While comparing the structures, it is also worth considering the difference in the number of households in Finland and Poland: in 2015 there were 2.6 million households in Finland and 13.5 million households in Poland (also see Fig. 3).

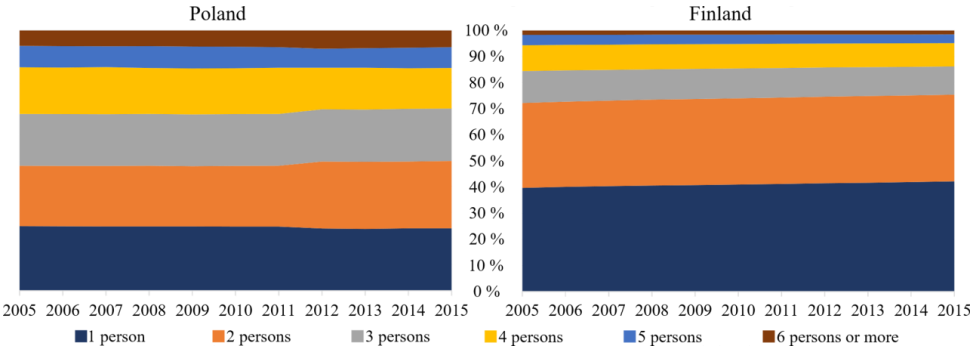


Figure 1. Cumulative household structures in Poland and Finland from 2005 to 2015 by number of persons in each household

Source: Calculations based on Eurostat data: Distribution of private households by household size [ilc_lvph03].

Figure 1 shows clear differences in the household structure by the household size between the studied countries, and also presents the changing trends over time. In Poland, the household structure by size has been more even, with no dominant groups. In Finland, one- and two-person households together accounted for nearly 75% of all households in 2015, while Polish households of the same size accounted for less than 50%. One-person households (almost 25%) were the most numerous household type in Poland until 2011, and then two-person households became most numerous. In Poland, through the research period, three-person households accounted for about 20% of all households, which can be called a constant due to an increase of as little as 1%. A visible difference between Finland and Poland was also seen in households with four members, for in Poland this type of living was significantly more popular. However, their shares dropped within the ten-year period from 18% to 15.6%. In both countries, the least numerous household types were those of five and six or more persons.

Comparing two countries and their household structures required an assessment of the similarity between these two populations. The Renkonen similarity index presented in Table 1 indicates not only that Polish and Finnish household structures are not similar (with an index lower than 1), but also indicates that the differences increased with time (lower index value for 2015 than 2005).

Table 1. The Renkonen similarity index of the household structure in Poland and Finland according to size

Proportional shares by household size	1 pers.	2 pers.	3 pers.	4 pers.	5 pers.	6+ pers.	Renkonen similarity index
ω_{PL2005}	0.25	0.23	0.20	0.18	0.08	0.06	0.76
ω_{FI2005}	0.40	0.33	0.12	0.10	0.04	0.02	
min	0.25	0.23	0.12	0.10	0.04	0.02	
ω_{PL2015}	0.24	0.26	0.20	0.16	0.08	0.06	0.74
ω_{FI2015}	0.42	0.33	0.11	0.09	0.03	0.01	
min	0.24	0.26	0.11	0.09	0.03	0.01	

Source: Calculation based on Eurostat data: Distribution of private households by household size [ilc_lvph03].

4.1. Dynamics of OPH shares among all households

After analysing the household structures in both countries, the focus was placed on OPH, in order to answer the question of the shares of OPH and its changing trends (Table 2) in the studied period in Poland and Finland. Between 2005 and 2015, the share of OPH in Poland dropped steadily, while in Finland the share kept growing. The changes, although not strong (1% for Poland, and 2.5% for Finland), went in the opposite direction.

Table 2. Shares of one-person households out of all household types, percent

TIME	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*	*2015*
OPH_{PL}	24.80	24.70	24.70	24.70	24.70	24.60	24.60	23.90	23.70	23.90	23.90
OPH_{FI}	39.71	40.09	40.37	40.61	40.74	41.01	41.20	41.47	41.66	41.94	42.22

Source: Calculation based on Eurostat data: Distribution of private households by household size [ilc_lvph03].

4.2. Structure of one-person households by selected socio-economic variables

Further analysis divides one-person households by age, sex, employment and education (Figure 2). In both countries, women were the majority among people living alone, with greater gender differences in Poland. In both Poland and Finland, the share of men among OPH slightly increased from 2005 to 2015. Overall, in 2015, women in Poland constituted 66% and in Finland 56% of all OPH.

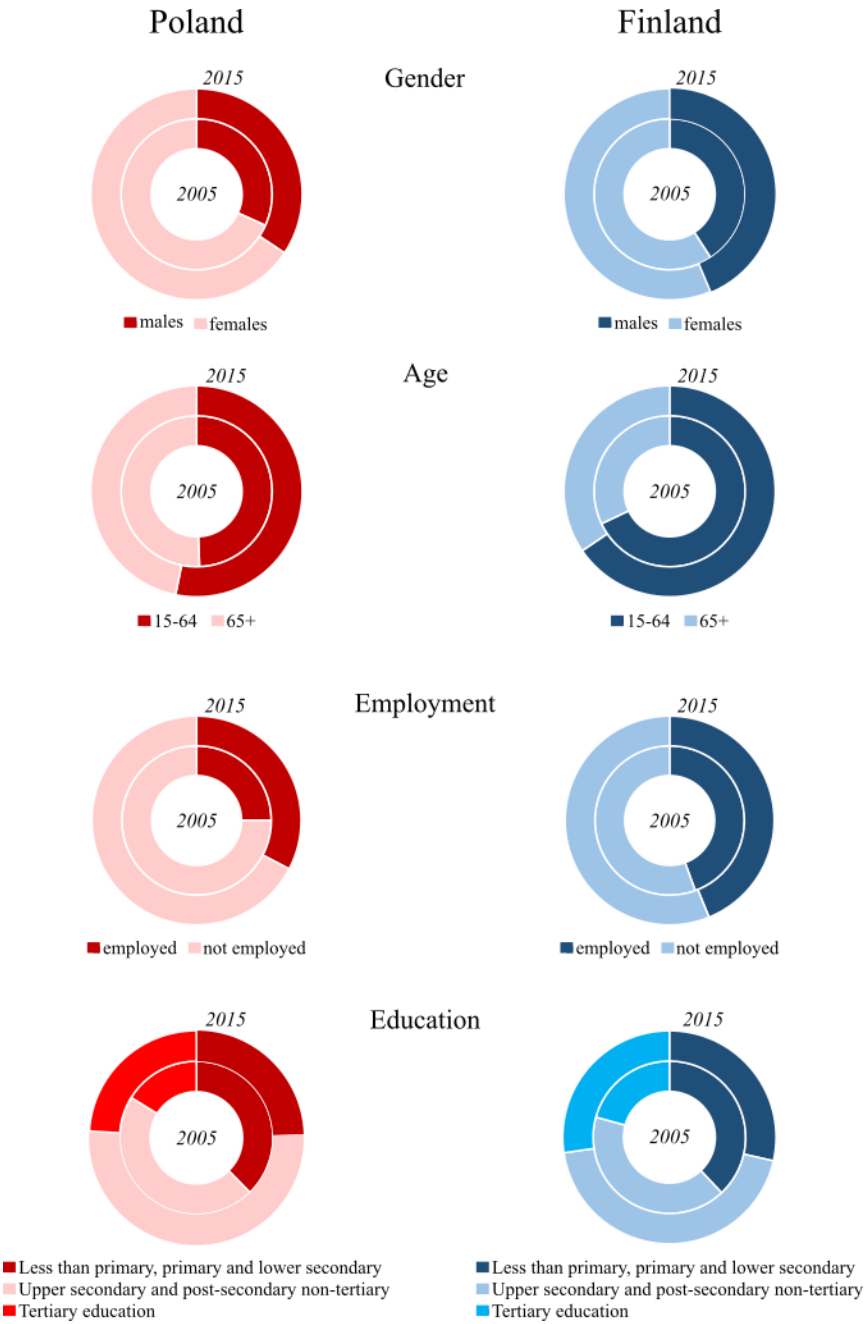


Figure 2. Comparison of structures of one-person households by selected socio-economic attributes

Source: Analysis based on Eurostat data on private household characteristics by type of household [hbs_car_t313].

Dividing OPH into two age groups, corresponding roughly to the working age (15–64) and the retirement age (65+), shown that the majority of people living alone in both countries belong to the younger age group. In Finland, the share of people aged 15–64 among those living alone was visibly higher than in Poland; however, it slightly dropped from 2005 to 2015, while in Poland that share grew.

In both Poland and Finland the majority of OPH were not employed. However, Polish employed OPH grew between 2005 and 2015, while the structure in Finland stayed the same. This finding was confirmed by the fact that many people living alone are of retirement age; thus, they are outside the labour force (see Figure 3).

The analysis of the distribution of OPH by the level of education presented in both countries shows that most people living alone had an upper secondary and post-secondary (non-tertiary) education. Also, for both studied countries, between 2005 and 2015 a decrease in shares of OPH with lower education levels was noticeable, as was an increase in tertiary education. The general education level of OPH hence increased.

4.3. Population pyramids

Figure 3 presents the shares of people living alone among the total population, i.e. the second research approach. Having the total population of each country categorized by sex and age groups, the number of people living alone with the same attributes was collected. Therefore, it first shows the age distribution of the population; second, it emphasizes the difference in size of the Polish and Finnish populations. Finally, it presents the number of each age group that lives alone.

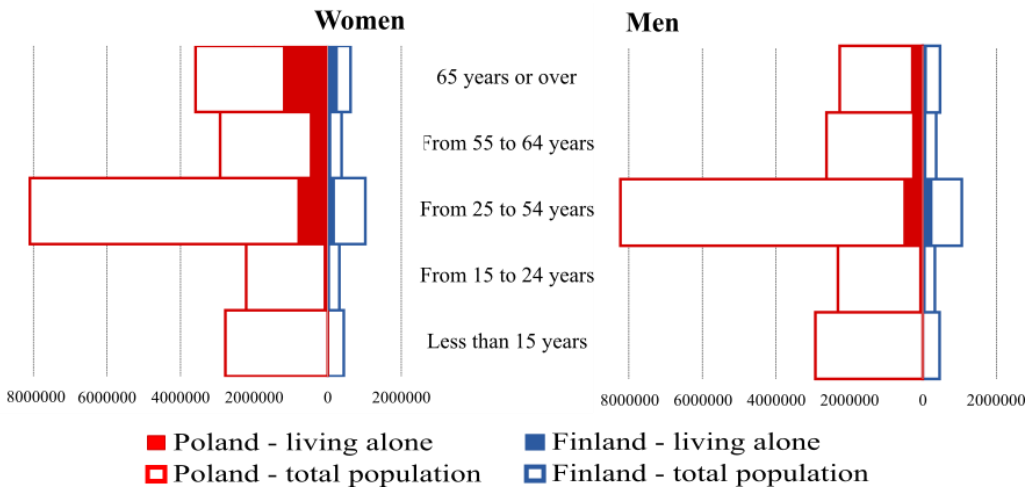


Figure 3. Population pyramids of total and living alone population in 2015

Source: Analysis based on population statistics from Statistics Poland and Statistics Finland.

Due to the vast difference in size of the populations, the share of people living alone per sex and age is additionally presented in Figure 4. Shares of living alone in Finland across every age group above 15 are higher in Finland than in Poland. Similarly, in both countries, living alone is more common among women and the elderly population, while an interesting difference is the almost non-existent OPH of ages 15–24 in Poland compared to every fifth Finnish woman of that age, and almost as many young Finnish men lived alone in 2015. Another difference is the age group 25–54; in Poland, women have higher shares of separate living, while in Finland men have higher shares.

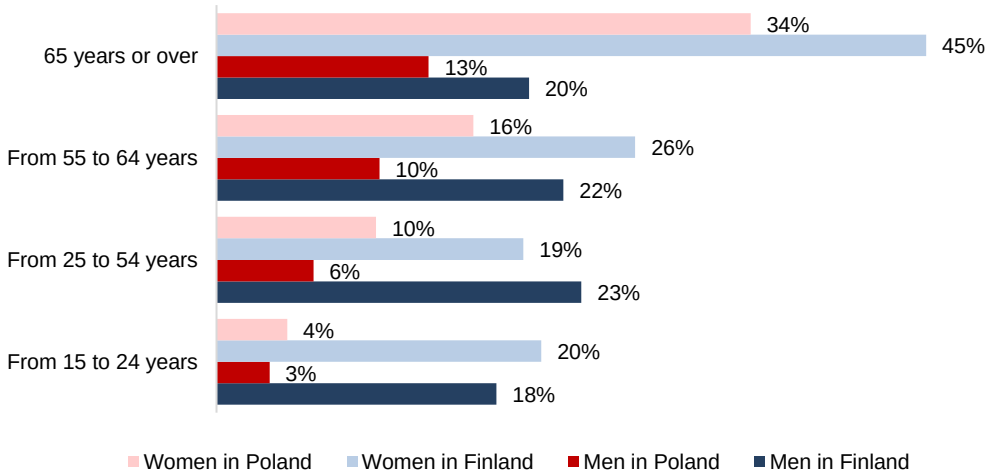


Figure 4. Share of the population living alone among total population by age and sex

Source: Analysis based on population statistics from Statistics Poland and Statistics Finland.

4.4. Selected demographic indicators

The literature mostly links living arrangements, especially OPH, with the economic variables. However, living alone as a process is affected by several marital and vital life-events. Different shares of people living alone at different age indicate that several demographic indicators could also have an explanatory role. For example, the difference in mortality, the longer life expectancy of women over men could, at least partially, describe the high differences in shares of OPH in the oldest age group.

Table 3. Selected descriptive demographic indicators for Poland and Finland

	Male life expectancy at age 0		Female life expectancy at age 0		Difference in life expectancy of male and females		Crude birth rate		Crude death rate	
year	2005	2015	2005	2015	2005	2015	2005	2015	2005	2015
PL	70.8	73.5	79.3	81.6	8.5	8.1	9.5	9.7	9.6	10.4
FI	75.6	78.7	82.5	84.4	6.9	5.7	11.0	10.1	9.1	9.6

	Total fertility rate		Mean age of women at birth of first child		Mean age at first marriage woman/man		Crude marriage rate		Crude divorce rate	
PL	1.24	1.32	25.7	27.0	25.3/27.7	26.9/29.3	5.4	5	1.8	1.8
FI	1.80	1.65	27.9	28.8	29.4/31.5	31.0/33.4	5.6	4.5	2.6	2.5

Source: Eurostat data on demographic indicators: [demo_gind], [demo_find], [demo_nind], [demo_mlexpec].

Furthermore, the lower mean age at the events of first child birth and first marriage in Poland also shows that young Poles start family life sooner, thus they are less likely to live alone. At the same time, in the studied period there was an increase in both the crude birth and crude death rates for Poland (i.e. occurring event per 1,000 of population), while the crude birth rate for Finland decreased.

The causality between demographic indicators and shares of people living alone is not targeted by the paper, one of the reasons being a short time series and lack of individual base data. The subject, however, is considered being of future interest to the authors.

4.5. When will the shares of OPH be equal in both countries?

Inspired by the literature, the last point of the analysis was to estimate when the shares of OPH will have the same values in Poland and Finland if the changing trends from 2005 – 2015 stayed the same. For that purpose, the changing trends of the shares of OPH in both countries were presented and described with a linear trend model (Fig. 5). In Finland, from 2005 to 2015, the annual share of OPH among all household types increased by 0.002 percent, while in Poland the share decreased by 0.001 percent.

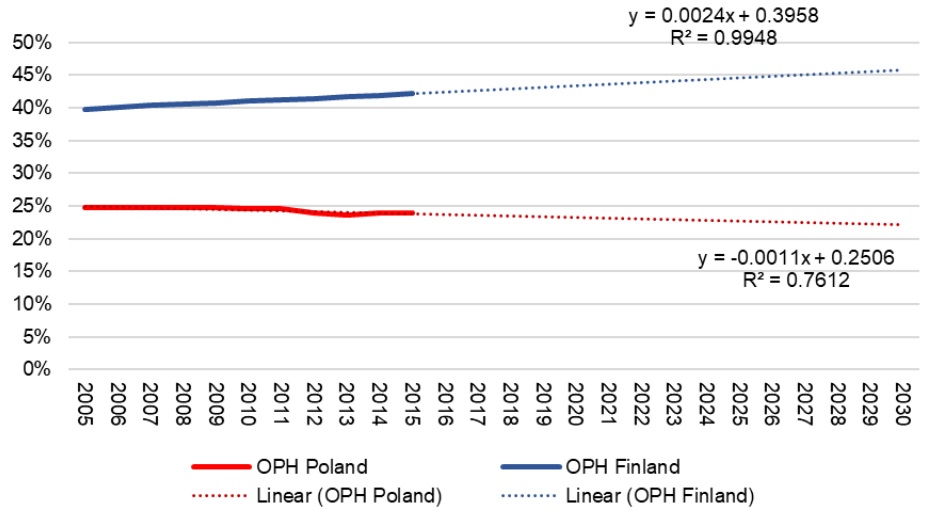


Figure 5. Linear trend estimation for shares of OPH in Finland and Poland.

Source: Analysis based on Eurostat data: Distribution of households by household size [ilc_lvph03].

Table 4. Quality assessment of the model fit

	Regression Coefficient a1 (Error)	t-value for a1	Regression Coefficient a0 (Error)	t-value for a0	R-square	Standard Error
PL	-0.001 (0.000)	-5.356	0.251 (0.001)	175.542	0.761	0.002
FI	0.002 (0.000)	41.633	0.396 (0.000)	1028.010	0.995	0.001

Table 4 presents an assessment of fitting the linear trend model to the changing shares of OPH in household structures. Both regression coefficients proved to be statistically significant. R-square and standard error both indicate a good model fit. As the values of the slope (a1) are close to 0, yearly changes are minor. However, having different signs means that the values for Finland are increasing and the values for Poland are decreasing.

The estimate of the shares of OPH up to 2030 was conducted with the linear trend model. Assuming the trend changes remained the same over the 2005 – 2015 period, the share of OPH in Poland and Finland will never be equal. This means that, regardless of the ongoing development and despite aspiring to the level of its Western neighbouring countries, *Poland has not become Nordic* with respect to living solo. According to the forecast, in Finland in 2030, one-person households will account for 46% of all households, and in Poland they will account for 22%.

5. Discussion

The household structure is an important indicator of well-being in a society. As presented in Eurostat (2013), national household structures are strongly linked to the level of income, with a clear connection between joint living arrangements and lower incomes (Kuijsten, 1995, 1999). Labour market indicators, such as the rate of economically active people, or the employment rate, put Finland in a better position than Poland (Ramb, 2008; Misiak et al., 2014). In addition, the proportion of people at risk of poverty or social exclusion indicates that Finland is a country with higher living standards compared to Poland (Misiak-Kwit et al., 2016). However, as the employment rate of Finnish women in 2015 was 71.8 and its full-time equivalent was 66.8, for Finnish men the values were 73.9 and 71.5, and the employment of Polish women accounted for 60.9 with full-time employed 59.2, and Polish men were working at the rate of 74.7 and 75.0 (EC, 2016). These differences support the theory of the growing difference of only female OPH shares.

According to Czapiński and Panek (2009), 72% of households in Poland were finding it a little hard, hard or very hard to make ends meet. A visible share of Polish households (28%) claimed that their income was not sufficient to cover their basic needs; however, the share was decreasing over the last few years. To be able to cover the costs of necessities, 55% of households in Poland lived economically or very economically. When incomes were low, people's ability to meet their own needs was compromised: 86% of households had to choose what needs to satisfy and what not, while another 39% asked other family members for support and 36% took loans.

According to Eurostat, GDP per capita in 2015 in Finland was 9% above the EU-28 average (securing 9th position in the EU), while the Polish equivalent was 32% below the average. This, however, does not explain why Finland has an exceptionally high share of OPH among the EU countries. Therefore, next to income level, other factors influence household structures.

Another factor that could explain the high share of OPH is low fertility rates (TFR). However, both Poland and Finland currently struggle with low TFR, while the differences in OPH shares increase. In theory, changes in civil status, such as getting married, lowers the OPH shares, while getting a divorce increases separate living. The presented data on selected demographic indicators indeed supports the national differences. However, with the divorce rate being almost constant in Finland over the recent years, and growing in Poland, the shares of people living alone in Poland should be increasing and not decreasing, as it is shown here. Linking child birth with the change of OPH status can also partially explain the different trend direction for Poland and Finland, as the crude birth rate was increasing in Poland and decreasing in Finland.

The Finnish welfare state system provides a housing subsidy for students, among other things. Together, the financial independence and high social acceptance of separate living is considered by the authors to be the main reason for the great difference in shares of living alone among the population aged 15–24 in Poland and Finland.

Similarly, the elderly care system in Finland could be seen as more complex than in Poland. Not only can Finnish pensioners afford separate living, e.g. after

the death of a spouse, but in the case of higher care needs (with no need for hospitalization), a person can move into independent dwellings (and therefore still live separately) in a direct neighbourhood or into a hospital where daily care can be easily provided.

Finland is a bilingual country with both Finnish and Swedish as the national languages. There is, however, a lack of available data on language-based household structure. The population of Swedish-speaking Finns in 2015 constituted 5.3%, thus rather a minor share, however, together with a foreign language population of 6% in 2015, the language-based household composition is a potentially interesting subject, once the data is available.

The decreasing share of people living alone in Poland is in contradiction to the statement in the literature that living solo is a growing phenomenon across Europe. For example, Nowak-Sapota (2008) forecasted that in 2030 OPH will constitute one out of three households in the country. However, according to this paper, the shares of OPH are undergoing a decreasing trend and in 2030 they will account for only 22%. On the basis of this result it can be stated that the forecast made by Nowak-Sapota cannot be confirmed, but the important research question instead became: What are the reasons for this situation? The authors of this article recommend further in-depth research on the subject. Other research question also arise: Do people in Poland live in bigger households by choice or out of necessity? Are Poles and Finns satisfied with their accommodation? It should be noted that overall life satisfaction and average satisfaction with living accommodation tended to be highest in the Nordic countries. Moreover, those living in rural areas were clearly more satisfied with their accommodation than those living in cities (Misiak-Kwit et al., 2016).

While writing this paper, both Statistics Poland and Statistics Finland have published data on private household composition for 2017. Keeping in mind the conceptual difference, the data showed the share of OPH in Poland in 2017 was 23.5% (i.e. still decreasing) and in Finland 43.4% (i.e. still increasing). These values support the findings of this paper.

6. Summary and Conclusions

The paper compared the household structure and its dynamics in Finland and Poland, with the focus on one-person households, in the form of a descriptive analysis. The first research hypothesis was confirmed, namely that there are strong differences in household structures in Finland and Poland. Small, one- or two-person households dominate the household structure in Finland to as high as 75%. In Poland, the household structure by size has been more even and households up to three persons together account for about 70%. The differences are considered to have both a cultural and an economic background. Living solo has reached a much higher social acceptance in Finland, while high economic development and the Nordic welfare state model is also supportive.

The second hypothesis was confirmed only partially. The share of OPH among all households has been significantly larger in Finland (42%) than in Poland (24%). However, the distance between these countries has not been diminishing. The OPH share in Finland is increasing, while in Poland it is decreasing. This has allowed for the calculation that, if the changing trends from

the studied period are maintained, the shares of OPH in these two countries will not equalize, but will instead grow further apart. An estimate was made that in 2030, 46% of Finnish households and 22% of Polish households will be one-person households.

Summing up, regardless of the progressive convergence that is diminishing difference gaps between different European regions and countries, the position of people living alone is still different between Poland and Finland. *Poland has not gone Nordic* in this aspect.

In the next paper, the authors plan to expand the comparative analysis to all European countries, empirically and spatially analysing changing trends in the shares of one-person households across Europe. Statistical analysis of casualty between demographic indicators as well as economic indicators is also planned in order to better understand why transnational differences occur.

REFERENCES

ALHO, J. M., KEILMAN, N., (2010). On future household structure. Journal of the Royal Statistical Society A Series, 173, pp. 117–143.

BENNETT, J., DIXON, M., (2006). Single person households and social policy. Looking forwards. York: Joseph Rowntree Foundation.

BĄK, I., MARKOWICZ, I., MOJSIEWICZ, M., WAWRZYNIAK, K., (2015). Wzory i tablice. Metody statystyczne i ekonometryczne, CeDeWu, Warszawa.

BATÓG, B., BATÓG, J., (2006). Analysis of Income Convergence in the Baltic Sea Region, "Baltic Business Development: Regional Development SME Management and Entrepreneurship", University of Szczecin, Szczecin.

CHRISTIANSEN, S. G., KEILMAN, N., (2013). Probabilistic households forecast based on register data – the case of Denmark and Finland. Demographic Research, 28, pp. 1263–1302.

CZAPIŃSKI, J., PANEK, T. ed., (2009). Diagnoza społeczna 2009. Warunki i jakość życia Polaków. Raport, Warszawa: Rada Monitoringu Społecznego.

EUROPEAN COMISSION, (2016). Labour force participation of women. European Semester Thematic Factsheet.

EUROSTAT, (2013). Household composition, poverty and hardship across Europe. Eurostat – Statistical working papers.

EUROSTAT, (2015). People in the EU: Who are we and how do we live? Eurostat Statistical books.

EUROSTAT, (2016). http://ec.europa.eu/eurostat/statisticsexplained/index.php/Household_composition_statistics (Accessed 8 August 2016).

FOKKEMA, T., LIEFBROER, A. C., (2008). Trends in living arrangements in Europe: Convergence or divergence? Demographic Research, 19 (36), pp. 1351–1418.

HABARTOVA, P., (2018). Recent Household Trends in Europe: A Cross-Country Analysis. Demografie, Vol. 2/2018, Český statistický úřad.

HALL, R., OGDEN, P. E., HILL. C., (1997). The pattern and structure of one-person households in England and Wales and France. International Journal of Population Geography, 3, pp. 161−181.

HOLZER, J. Z., (2003). Demografia. Warszawa: PWE.

HOZER, J., ed., (1997). Ekonometria. Uniwersytet Szczeciński, Szczecin.

HOZER-KOĆMIEL, M., LIS CH., (2016). Examining similarities in time allocation amongst European countries, in Statistics in Transition, Vol. 17, No. 2/2016, pp. 317–330.

IACOVOU, M., SKEW, A. J., (2011). Household composition across the new Europe: Where do the new member states fit in? Demographic Research, 25 (14), pp. 465−490, DOI:10.4054/DemRes.2011.25.14.

JACOBSEN, L. A., MATHER, M. M., DUPUIS, G., (2012). Household Change in the United States. Population Bulletin, Population Reference Bureau, Vol. 67, No. 1, pp 1–16.

KEILMAN, N., (1988). Recent trends in family and household composition in Europe. European Journal of Population, 3 (3/4), pp. 297−325, DOI: 10.1007/BF01796903.

KEILMAN, N., (2016). Household forecasting: Preservation of age patterns. International Journal of Forecasting, 32, pp. 726–735.

KEYFITZ, N., CASWELL, H., (2005). Applied mathematical demography. Statistics for biology and health. Third edition. Springer-Verlag New York.

KOPYCIŃSKA, D., ed., (2011). Mikroekonomia, Szczecin: Kreos, 2011.

KUIJSTEN, A., (1995). Recent trends in household and family structures in Europe: An over-view, [in:] E. van Imhoff, A. Kuijsten, P. Hooimeijer, L. J. G. van Wissen ed., Household demography and household modeling, Plenum Press, New York–London: pp. 53–84.

KUIJSTEN, A., (1999). Households, families, and kin networks. In: L. J. G. van Wissen, P. A. Dykstra ed., Population issues. An interdisciplinary focus, Kluwer Academ-ic/Plenum Publisher, New York: pp. 87–122.

LATUCH, M., (1980). Demografia społeczno-ekonomiczna. Warszawa: PWE.

MISIAK, S., HOZER–KOĆMIEL, M., TOMASZEWSKA, K., (2014). The life of women and men in South Baltic countries, Economic approach, Malmo: Winnet.

MISIAK-KWIT, S., HOZER-KOĆMIEL, M., TOMASZEWSKA, K., BĄKOWSKA, S., (2016). Rural women and men in Baltic Sea Region, Overview, statistics, recommendations. Sweden: Winnet.

NOWAK-SAPOTA, W., (2007). Gospodarstwa jednoosobowe w Polsce – analiza przestrzenna. In: A. Rączaszek ed., Uwarunkowania demograficzne rozwoju społeczno-gospodarczego na przykładzie województwa śląskiego i opolskiego, AE, Katowice, pp. 217–233.

NOWAK-SAPOTA, W., (2008). Osoby starsze w strukturze nierodzinnych gospodarstw domowych w Polsce. In: J. T. Kowaleski, and P. Szukalski ed., Starzenie się ludności Polski – między demografią a gerontologią społeczną, UŁ, Łódź 2008, pp. 27–47.

OFFICIAL STATISTICS OF FINLAND (OSF). Household-dwelling units and housing conditions [e-publication]. Helsinki: Statistics Finland [referred: 12.10.2018], Access method: http://www.stat.fi/til/asuolo/kas_en.html

OKÓLSKI, M., (2005). Demografia. Podstawowej pojęcia, procesy i teorie w encyklopedycznym zarysie. Warszawa: Wydawnictwo Naukowe Scholar.

OLÁH, L. S., (2015). Changing families in the European Union: Trends and policy implications. In: United Nations Expert Group Meeting "Family policy development: Achievements and challenges" New York.

PAMPEL, F. C., (1983). Changes in the propensity to live alone: Evidence from consecutive cross-sectional surveys, 1960–1976, Demography, 20, pp. 433–501.

RAMB, F., (2008). Population and social conditions, Statistics in focus, 99/2008, Eurostat.

RENKONEN, O., (1938). Statisch-ökologische Untersuchungen über die terrestrische Käferwelt der finnischen Bruchmoore. Ann. Zool. Soc. Bot. Fenn. Vanamo 6, pp. 1–231

SAMUELSON, P. A., NORDHAUS, W. D., (2005). Ekonomia 1, Warszawa: PWN.

SZUKALSKI, P., (2004). Rodziny i gospodarstwa domowe w Polsce i w krajach UE. In: W. Warzywoda-Kruszyńska, and P. Szukalski, ed. Rodzina w zmieniającym się społeczeństwie polskim, Łódź: Wydawnictwo Uniwersytetu Łódzkiego, pp. 23–47.

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE, (2011). Measurement of different emerging forms of households and families. New York and Geneva.

WEINBACH, R. W. GRINNELL, R. M., (2007). Statistics for Social Workers. 7th edition, Pearson, Boston.

ZALEGA, T., (2007). Gospodarstwo domowe jako podmiot konsumpcji, Studia i Materiały – Wydział Zarządzania Uniwersytetu Warszawskiego, 1, pp. 7–24.

ZIMOCH, U., (2013). A Comparative Analysis of the Information Culture as the Information Society Indicator in Poland and Finland, Transformacje, 3-4, pp. 70–93.

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 743–745, DOI 10.21307/stattrans-2018-039

REPORT

The XXXVII International Conference on Multivariate Statistical Analysis 5–7 November 2018, Łódź, Poland

The 37th edition of the International Conference on **Multivariate Statistical Analysis** (MSA) was held in **Łódź, Poland on November 5–7, 2018**. The MSA conference was organized by the **Department of Statistical Methods** of the University of Łódź, the **Institute of Statistics and Demography** of the University of Łódź, the **Polish Statistical Association** and the **Committee on Statistics and Econometrics of Polish Academy of Sciences**. Its organization was financially supported by the **National Bank of Poland**, the **Polish Academy of Sciences** and **Satsoft Polska Sp. z o.o.** The Organizing Committee was headed by **Professor Czesław Domański** and the scientific secretary was **Katarzyna Bolonek-Lasoń**, Assistant Professor from the Department of Statistical Methods of the University of Łódź.

The Multivariate Statistical Analysis conference constituted a forum for discussion and exchanging opinions about development of statistics. Participants presented the latest theoretical achievements in the field of the multivariate statistical analysis, its practical aspects and applications. The scientific programme covered a wide range topics of statistical mathematics and multivariate statistical methods including multivariate distributions, statistical tests, nonparametric inference, factor analysis, cluster analysis, discrimination analysis, Bayesian methods, stochastic analysis and application of statistical methods in finance, economy, capital market and risk management.

The conference was attended by 80 participants from many academic centres in Poland (Białystok, Gdańsk, Katowice, Kraków, Lublin, Łódź, Poznań, Toruń, Szczecin, Warszawa, Wrocław) and from abroad (Italy, Lithuania). Representatives of Statistics Poland and the Statistical Office in Lodz and Poznań were also participants of the 2018 MSA conference. In 18 sessions (plenary and parallel) 57 papers were presented including 3 invited lectures.

The conference was opened by the Head of the Organizing Committee, **Professor Czesław Domański**. The subsequent speakers at the conference opening included **Professor Antoni Różalski**, Rector of the University of Łódź, and **Professor Michał Mackiewicz**, the Deputy Dean of the Faculty of Economics and Sociology of the University of Łódź.

The first plenary session was devoted to eminent representatives of statistical thought. **Professor Krzysztof Jajuga** (Wrocław University of Economics) was the chairman of the historical session. **Professor Mirosław Krzyśko** (Adam Mickiewicz University) presented paper titled “Mieczysław Warmus – a memory on the centenary of birth”. **Professor Józef Pociecha** (Cracov University of Economics) talked about professor Juliusz Leo. **Professor Czesław Domański**

(University of Łódź) presented two papers. The first paper concerned the role of Scientific Societies in the organization of independent Polish structures, the second one was dedicated to Tadeusz Korzon.

One of the session tilted „New century of official statistics" was organized together with Statistical Office in Łódź. **Professor Józef Pociecha** was a chairman of this session. The first speaker, **Dominik Rozkrut**, the President of Statistics Poland, presented a paper „*Academic education in the field of data science in Poland*". The next speakers included **Professor Czesław Domański** and **Professor Alina Jędrzejczak** (University of Łódź), who addressed the problem of „*Ethical dilemmas of statisticians in the face of new data sources*". Consecutive speakers (representatives of Statistical Office in Łódź) included **Tomasz Piasecki**, who talked about the application of mathematical methods in the official statistics, **Katarzyna Szkopiecka**, who presented a paper "*Educational activities of the Statistical Office in Łódź*", and **Anna Jaeschke**, who drew a picture of "*Population census yesterday and today*".

Invited lectures were presented by **Professor Krzysztof Jajuga** (Wrocław University of Economics), **Professor Włodzimierz Okrasa** (Cardinal Stefan Wyszyński University in Warsaw) and **Professor Mirosław Szreder** (University of Gdańsk).

Other sessions were chaired respectively by:

Session III	Professor Bronisław Ceranka (Poznań University of Life Sciences)
Session IVa	Professor Grażyna Dehnel (Poznań University of Economics and Business)
Session IVb	Professor Alina Jędrzejczak (University of Łódź)
Session Va	Professor Grzegorz Kończak (University of Economics in Katowice)
Session Vb	Professor Andrzej Dudek (Wrocław University of Economics)
Session Vc	Professor Wojciech Gamrot (University of Economics in Katowice)
Session VI	Professor Czesław Domański (University of Łódź)
Session VII	Professor Mirosław Krzyśko (Adam Mickiewicz University in Poznań)
Session VIIIa	Professor Grażyna Trzpiot (University of Economics in Katowice)
Session VIIIb	Professor Agata Szczukocka (University of Łódź)
Session IXa	Professor Iwona Markowicz (University of Szczecin)
Session IXb	Professor Jerzy Korzeniewski (University of Łódź)
Session X	Professor Marek Walesiak (Wrocław University of Economics)
Session XIa	Professor Tomasz Żądło (University of Economics in Katowice)
Session XIb	Professor Iwona Bąk (West Pomeranian University of Technology)
Session XII	Professor Józef Dziechciarz (Wrocław University of Economics)

During the conference, a meeting of the members of the **Main Board of the Polish Statistical Association** was also held. **Professor Czesław Domański** (President of Polish Statistical Association) chaired this meeting.

The 2018 MSA conference was closed by the Chairman of the Organizing Committee, **Professor Czesław Domański**, who summarized the conference

and thanked the guests for arriving and taking active participation in the conference.

The next edition of **Multivariate Statistical Analysis Conference MSA 2019** is planned on **November 4–6, 2019** and will be held in **Łódź**, **Poland**.

Prepared by

Katarzyna Bolonek-Lasoń
Department of Statistical Methods, University of Łódź

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 747–749

ABOUT THE AUTHORS

Ala-Karvia Urszula is a doctoral student at the Doctoral Programme in Social Sciences and a member of the research staff of Ruralia Institute, University of Helsinki, Finland. People living alone in urban-rural areas of Finland and their statistical analyses is the focus of her ongoing PhD research. With her dual nationality, Polish and Finnish, she is highly interested in statistical comparisons between the two countries. She is a member of the Finnish Demographic Society.

Alizadeh Morad is PhD in statistics at Persian Gulf university of Bushehr, Iran. He is working as academic member. His current areas of research are distribution theory and lifetime distributions. He has published (accepted) over 80 research articles.

Gao Ping is a PhD student at the Graduate School of Economics and Business in Hokkaido University. Her research interests are Bayesian analysis, spatial statistic, multivariate statistical analysis, statistical inference, and health of elderly people. She is a member of the Japanese Economic Association. She has published three research papers and one conference paper.

Hamedani G. G. is Wehr Professor of Mathematics and Statistics at the Department of Mathematics, Statistics and Computer Science, Marquette University. His current areas of research are in distribution theory and characterizations of distributions. He is the Editor of the Journal of Statistical Theory and Applications and a member of the Editorial Board of seven journals. He has published over 250 research articles and Research Books.

Hasegawa Hikaru is a Professor at the Department of Economics in Hokkaido University, Japan. His main areas of research interest include Bayesian statistics, discrete choice models, economic inequality and income distribution. Currently, he is a member of the Japanese Economic Association and the Japan Statistical Society.

Hozer-Koćmiel Marta, PhD, is an Assistant Professor at the Department of Statistics, Faculty of Economics and Management, University of Szczecin, Poland. Her main research interest is the usage of quantitative methods to study the differences and similarities of economic behaviour of women and men. She is also interested in time use surveys, methods of valuation of household work, entrepreneurship and sustainable development from gender perspective.

Irshad M. R. is an Assistant Professor in the Department of Statistics, Cochin University of Science and Technology, Kochi, Kerala, India. His research interests are order statistics, distribution theory, entropy, record values, ranked set sampling and concomitants ordered random variables. He has published over 20 research articles in various international/national journals. He has served as a reviewer of various international statistical journals.

Kumar Amod is a senior research fellow in the Department of Applied Mathematics, Indian Institute of Technology (ISM) Dhanbad, India. He is pursing Ph.D. in Applied Statistics. His main areas of research are sample surveys and statistical inference. He has published his research work in international Journals repute.

Landmesser Joanna is an Associate Professor at the Department of Econometrics and Statistics, Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences. She received her PhD degree in Economics in 2002 at the Bundeswehr University Munich in Germany and habilitated in 2014 in Economics ("The use of duration analysis methods for the survey of economic activity of people in Poland") at the Nicolaus Copernicus University in Toruń, Poland. Her research interests focus on evaluation of different policies on the labour market, counterfactual scenarios analysis, comparing the income distributions, decomposition of income inequalities.

Maya R. is working as an Assistant Professor on contract basis in the PG and Research Department of Statistics, University College, Thiruvananthapuram, Kerala, India. Her research interests are order statistics, distribution theory, reliability theory, information theory, concomitants of record values and non-parametric estimation. She has published over 15 research articles in various international/national journals. She has served as a reviewer of various statistical journals of international repute.

Misiak-Kwit Sandra, PhD, is an Assistant Professor at the Department of Human Capital Management, Faculty of Economic and Management, University of Szczecin, Poland. Her main research areas are entrepreneurship, women, self-employment, human resource management, business ethics, negotiations, stress management, business relations and communication.

Nouara Lazri is doctoral student of mathematics class at Badji-Mokhtar University Annaba - Algeria. She received her master degree in mathematics from Badji Mokhtar University. Her research areas are in: applied statistics and actuarial science.

Ranjbar Vahid is an Assistant Professor of statistics at the Department of Statistics in Golestan University, Goran, Iran. His research interests are asymptotic properties, distribution theory and lifetime distributions. He has published (accepted) over 20 research articles.

Staszko Barbara studies at the University of Szczecin, Poland, at the Faculty of Economics and Management. Her main interest is real estate market and its operating rules. Currently, she acquires knowledge, develops skills and gathers experience.

Urbańczyk Dominika M. is a PhD student at the Faculty of Economic Sciences, Warsaw University of Life Sciences. Simultaneously she also cooperates with the Department of Econometrics and Statistics at the Faculty of Applied Informatics and Mathematics, WULS. Her research interests include the application of mathematical statistics and data analysis methods to economic problems, such as enterprises survival on the market, inequalities in income distributions. The research subject of her doctoral dissertation concentrates on the survival analysis of Polish enterprises.

Vishwakarma Gajendra Kumar is an Assistant Professor at the Department of Applied Mathematics, Indian Institute of Technology (ISM) Dhanbad, India. He has worked as Visiting Scientist in Indian Statistical Institute (ISI), India. He is an elected Fellow of the Society of Earth Scientists (FSES) India. He is an Elected Member of International Statistical Institute, Netherlands. His research experience covers both applied as well as theoretical provinces that include estimation techniques, statistical models for clinical trials and data analysis. He has published a number of research papers in journals of international repute. He is a reviewer and editorial board member of peer reviewed journals. He has delivered various invited talks at industrial as well as academic forum. He received VIFRA-2015 Young Scientist Award from Center for Advanced Research and Design, Chennai, India.

Walesiak Marek is a Professor at Wroclaw University of Economics in Department of Econometrics and Computer Science. He is a member of the Methodological Commission in Statistics Poland (GUS) and an active member of many scientific professional bodies. His main areas of interest include: classification and data analysis, multivariate statistical analysis, marketing research, computational techniques in R. Currently, he is a member of two editorial boards: Przegląd Statystyczny (Statistical Review), Econometrics. Advances in Applied Data Analysis.

Yahia Djabrane is faculty member at the Department of Mathematics at The University of Mohamed Khider, Biskra-Algeria. He has received his PhD degree in Mathematics and the highest academic degree (HDR) specializing in Probability and Statistics from Mohamed Khider, Biskra-Algeria. His research areas are in applied probability and applied statistics. He has published over 20 research papers in international journals and conferences.

Zeghdoudi Halim is faculty member at the Department of Mathematics at The University of Badji-Mokhtar, Annaba-Algeria. He has received his PhD degree in Mathematics and the highest academic degree (HDR) specializing in Probability and Statistics from Badji-Mokhtar University, Annaba-Algeria. He also did his Post Doc at Waterford Institute of Technology- Cork Rd, Waterford, Ireland. His research areas are in actuarial science, particles systems, dynamics systems and applied statistics. He has published over 60 research papers in international journals and conferences. Currently, he is a member of two editorial boards: Asian Journal of Probability and Statistics and Journal of Advanced Statistics and Probability.

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 751–754,

ACKNOWLEDGEMENTS TO REVIEWERS

The Editor and Editorial Board of the Statistics in Transition new series wish to thank the following persons who served from 31 December 2017 to 31 December 2018 as peer-reviewers of manuscripts for the ***Statistics in Transition new series – Volume 18, Numbers 1–4***; the authors` work has benefited from their feedback.

Adebanji Atinuke, Department of Mathematics, Kwame Nkrumah University of Science and Technology, Ghana

Adichwal Nitesh Kumar, Department of Statistics, Banaras Hindu University, India

Afify Ahmed, Department of Statistics, Mathematics & Insurance, Benha University, Egypt

Agunloye Oluokun Kasali, Obafemi Awolowo University, Department of Mathematics, Faculty of Science, Nigeria

Ayhan H. Öztaş, Department of Statistics, Middle East Technical University, Turkey

Arayal T. R., Department of Statistics, T.B. University, Nepal

Baszczyńska Aleksandra, Department of Statistical Methods, University of Lodz, Poland

Bouza Carlos, Universidad de La Habana, Cuba

Bhatia Mrigesh, Department of Health Policy, London School of Economics and Political Science (LSE), United Kingdom

Caliński Tadeusz, Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, Poland

Chaturvedi Anoop, Department of Statistics, University of Allahabad, India

Dihidar Kajal, Indian Statistical Institute, Kolkata, India

Djerf Kari, Statistics Finland, Finland

Domański Czesław, Department of Statistical Methods, University of Lodz, Poland

Elkasabi Mahmoud A., ICF International, Maryland, USA

Ezzebsa Abdali, Guelma University, Algeria

Fink Paul, Department of Statistics- Ludwig Maximilian University of Munich, Munich

García Luengo, Amelia Victoria, Department of Mathematics, University of Almería, Spain

Gil-Alana Luis, Department of Economics, School of Economics and Business Administration, Universidad de Navarra

Górecki Tomasz, Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznań, Poland

Grzenda Wioletta, Institute of Statistics and Demography, Warsaw School of Economics, Poland

Gunnarsson Ewa, Luleå University of Technology, Department of Engineering Sciences and Mathematics, Sweden

Hampel David, Department of Statistics and Operation Analysis, Mendel University in Brno, Czech Republic

Horňák Marcel, Department of Human Geography and Demography, Faculty of Natural Sciences, Comenius University in Bratislava, Slovakia

Kadilar Cem, Department of Statistics, Hacettepe University, Turkey

Karadeniz Pinar Gunel, Faculty of Medicine, Department of Biostatistics, Sanko University, Turkey

Keilman Nico, Department of Economics, University of Oslo, Norway

Komornicki Tomasz, Institute of Geography and Spatial Organization, Polish Academy of Sciences, Poland

Kosiorowski Daniel, Department of Statistics, Cracow University of Economics, Poland

Kośny Marek, Department of Econometrics and Operations Research , Wroclaw University of Economics, Poland

Kowalczyk Barbara, Institute of Econometrics, Warsaw School of Economics, Poland

Kozek Andrzej, Department of Statistics, Macquarie University, Australia

Krzyśko Mirosław, Faculty of Probability and Mathematical Statistics , Adam Mickiewicz University, Poland

Han-Dong Li, Beijing Normal University, China

Laaksonen Seppo, Department of Social Research, University of Helsinki, Finland

Lehtonen Risto, Department of Social Research, University of Finland, Finland

Lengyel Tamas, Mathematics Department, Occidental College, USA,

Leonida Tekie Asehun, Department of Applied Mathematics, University of Twente, Netherlands

Thakur Narendra Singh, Department of Mathematics and Statistics, Banasthali University, India

Tiensuwan Montip, Department of Mathematics, Faculty of Science, Mahidol University, Thailand

Touati Ali Bey, Department of Mathematics, Badji Mokhtar - Annaba University, Algeria

Vencálek Ondřej, Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science of the Palacky University in Olomouc, Czech Republic

Verma Vijay, Department of Economics and Statistics, University of Siena, Italy

Yaldiz Hatice, Department of Mathematics, Faculty of Science and Arts, Duzce University, Turkey

Wenhao Gui, Department of Mathematics and Statistics, University of Minnesota Duluth, USA,

Więckowska Barbara, Department of Computer Science and Statistics, Poznan University of Medical Sciences, Poland

Witkowska Dorota, Department of Finance and Strategic Management, University of Lodz, Poland

Wojnar Jolanta, Department of Economics, University of Rzeszów, Poland

Wolny-Dominiak Alicja, Department of Statistical and Mathematical Methods in Economics, University of Economics in Katowice, Poland

Wołyński Waldemar, Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland

Wunsch Guillaume, Royal Academy of Belgium &Demography, University of Louvain, Belgium

Žežula Ivan, Institute of Mathematics, Pavol Jozef Šafárik University in Košice, Slovakia

Zhao Yang, Department of Mathematics and Statistics, University of Regina, Canada

Zwiech Patrycja, Wydział Nauk Ekonomicznych i Zarządzania, University of Szczecin, Poland

STATISTICS IN TRANSITION new series, December 2018
Vol. 19, No. 4, pp. 755–758

INDEX OF AUTHORS, VOLUME 19, 2018

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:* http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***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 CAPITALS**. 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 italics) 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.**, **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).