Probability Distribution Modeling of Scanner Prices and Relative Prices

Piotr Sulewski[1] | *Pomeranian University in Słupsk, Słupsk, Poland*
Jacek Białek[2] | *University of Lodz, Lodz, Poland*

Received 25.3.2022, **Accepted** (reviewed) 14.4.2022, **Published** 16.9.2022

Abstract

The article deals with the problem of the proper selection of the theoretical distribution to describe the empirical distribution of scanner prices. In the empirical study we use scanner data from one retail chain in Poland, i.e. monthly data on natural yoghurt, drinking yoghurt, long grain rice and coffee powder sold in 212 outlets in January and February 2022. Prices and price relatives were modeled using selected ten probability distributions with non-negative support, including two, three and four-parameter family of distributions In addition to the visual assessment in the form of empirical PDF and CDF figures, numerical criteria were used. These include information criteria values such as AIC, BIC, HQIC and p values calculated for the K-S, AD and CVM goodness-of-fit tests. Our research showed that at least two models could be distinguished as very accurate, which provides a good background for simulation research on price indices or for the construction of so-called population price indices.[3]

Keywords	***DOI***	***JEL code***
Data modeling, scanner data, price distributions	*https://doi.org/10.54694/stat.2022.14*	*C43, E31, C13*

INTRODUCTION

Scanner data are a relatively new and at the same time cheap alternative data source in inflation measurement. The volume of scanner data is enormous compared to the datasets obtained as part of the traditional data collection and they provide detailed information about the products sold at the barcode level. As these data are usually obtained with high frequency (monthly, weekly, and in some countries even daily), it enables effective modeling of scanner prices. In turn, having well-matched theoretical probability distributions to empirical price distributions, we have a good background for simulation research on price indices or for the construction of so-called population price indices. This article addresses the problem of proper adjustment of the theoretical probability distribution to the distribution of real scanner prices.

[1] Institute of Exact and Technical Sciences, Pomeranian University in Słupsk, Słupsk, Poland. E-mail: piotr.sulewski@apsl.edu.pl.
[2] Department of Statistical Methods, University of Lodz, Lodz, Poland. E-mail: jacek.bialek@uni.lodz.pl. Also Central Statistical Office in Poland, Department of Trade and Services, Al. Niepodległości 208, 00-925 Warsaw, Poland. E-mail: J.Bialek@stat.gov.pl.
[3] This publication if financed by the National Science Centre in Poland (grant No. 2017/25/B/HS4/00387).

The article attempts to model the prices of food products, inter alia, due to the multitude of their representatives. In the empirical study we use scanner data from one retail chain in Poland, i.e. monthly data on natural yoghurt, drinking yoghurt, long grain rice and coffee powder sold in 212 outlets in January and February 2022. Prices and price relatives were modelled using selected ten probability distributions with non-negative support, including two, three and four-parameter family of distributions In addition to the visual assessment in the form of empirical PDF and CDF figures, numerical criteria were used. These include information criteria values such as AIC, BIC, HQIC and p values calculated for the K-S, AD and CVM goodness-of-fit tests. Our research showed that at least two models could be distinguished as very accurate.

Our paper consists of following sections. Section 1 describes the importance of scanner data, with the main advantages but also methodological challenges related to the implementation of these data in inflation measurement. This section also explains why scanner price modeling can be of great practical and theoretical importance. Section 2 presents probability distributions with non-negative support selected to price data modeling and two numerical criteria for comparisons of the quality of data modeling, i.e. the information criteria such as AIC, BIC and HQIC and p-values calculated while goodness-of-fit testing. Section 3 describes main stages of the implemented scanner data processing and it presents and describes results obtained for the set of selected probability distributions and applied goodness-of-fit tests, i.e. the Kolmogorov-Smirnow (K-S), Anderson-Darling (AD) and Cramer von Mises (CVM) tests. Final section lists general conclusions which can be drawn from our empirical study.

1 SCANNER DATA IN INFLATION MEASUREMENT

Scanner data mean transaction data that specify turnover and numbers of items sold by barcodes, e.g. GTIN (Global Trade Article Number), formerly known as the EAN (European Article Number) code (International Labour Office, 2004). These data are a quite new data source for statistical agencies and the availability of electronic sales data for the calculation of the Consumer Price Index (CPI) has increased over the past 20 years. They can be obtained from a wide variety of retailers (supermarkets, home electronics, Internet shops, etc.). However, the use of scanner data in the inflation measurement is associated with a number of methodological challenges discussed in the work.

1.1 The genesis, advantages and disadvantages of scanner data

We distinguish several basic sources of scanner data. The most valuable source of this type of data seems to be direct suppliers, i.e. points of sale with particular emphasis on supermarket chains. Supermarkets are powerful potential providers of scanned data - a typical supermarket has a database of 10 000–25 000 barcodes for products sold, most of which are food and drink. Theoretically similar providers of scanner data can also be smaller supermarkets, small retailers, pharmacies, travel agencies or even online stores, as long as they archive sales data taking into account product coding. The second, alternative source of scanner data may be companies specialized in market research. For case, some countries use the scanner data provided by Nielsen or GfK companies and include it in their national CPI estimates, nevertheless this is an expensive solution.

Listing main advantages of using scanner data we should note that: a) using scanner data is relatively cheap, automatic and based on huge data volumes; b) these data sets are complete at the lowest level of aggregation, i.e. they provide information both on product prices and their sales value at the elementary level; c) this data can be obtained at a high frequency at the barcode level, which in turn enables precise modeling of product price distributions even at the lowest level of data aggregation. The advantage listed in point "c" is precisely the main topic of this article.

Nevertheless, the decision to use scanner data is associated with a number of technological, IT and methodological problems (Białek, 2020). It is necessary to correctly and highly automatically classify products into COICOP groups and there is need also to precise match products over time. Some countries

implement data filtering before price index calculation (e.g. removal of products with extreme price changes). Ultimately, the Statistical Office is faced with the choice of the appropriate formula of the price index and the method of aggregation of indicators obtained on the basis of various data sources (Chessa, 2016).

1.2 Scanner data processing: classification and matching products

After downloading, formatting to the required form and pre-clearing the scanner data (deletion of records with missing data, deletion of deduplicates), all products should be classified into the appropriate elementary groups (COICOP 5 level) or their local subgroups (national COICOP 6 or lower). The classification of products can be carried out basically by two methods, and their selection can be determined by the content of the scanner data. Complete scanner data are transaction data, where at the level of GTIN codes we have information about the price of the product, the volume of its sales, sales unit, product label (detailed description), its weight, sometimes the material of execution, VAT or the size of the discount. Such a structure of information means that effective classification can be carried out based on machine learning methods, which, however, requires manual preparation of learning and test trials (Białek and Beręsewicz, 2021). The second effective solution in the process of automatic product classification is to prepare dictionaries of keywords and phrases that uniquely identify the COICOP group to which the tested product belongs.

After the products have been correctly classified into the appropriate homogeneous segments, the products sold in the compared months should be matched. For proper matching of products, the product code (internal code, broadcast over the retail chain and external code, such as EAN or GTIN) and their labels are most often used, if they are sufficiently precise. Comparison of product labels, both at the stage of product classification and in the process of matching products over time, requires the use of text mining methods and appropriate measures of distance between text strings.

1.3 Why do we need fitted scanner price distributions?

This article addresses the problem of proper adjustment of the theoretical probability distribution to the empirical distribution of scanner prices. Of course, there is a natural question about the desirability of this type of consideration.

In order to justify undertaking the research problem, let us note at the beginning that knowledge of the distribution of prices and the distribution of relative prices allows for the construction of the so-called *population price indices*. It is possible then to generalize the so-called *sample elementary indic*es (the Dutot index, 1738; the Carli index, 1764; or the Jevons index, 1865) to the entire population of products from a given segment by determining the so-called *population elementary price indices* (Silver and Heravi, 2008; Białek, 2022). With certain technical assumptions about consumption levels (quantity distributions), it is also possible to infer the population Laspeyres price index (Białek, 2015).

Another argument may be the fact that by having accurate probabilistic price models, we are able to effectively construct simulation experiments to study the nature of price indices. For example, knowing the expected values of such distributions and using theorems about the distribution of sums and quotients of random variables, we can formulate expectations for price indices understood as random variables, and then check whether the indices determined on the basis of empirical data are close to these expectations. The above approach was used, for instance, in the papers by Białek and Bobel (2019) or Białek and Beręsewicz (2021) to optimize the choice of a multilateral price index.

2 THEORETICAL PROBABILITY DISTRIBUTION CONSIDERED

2.1 The list of considered probability distributions

Continuous distributions related to the support can be divided into distributions supported on a bounded interval, supported on the whole real axis, supported to semi-infinite intervals (usually $[0,\infty)$)

and distributions with variable support. Due to the topic presented in the article, we limited ourselves only to distributions with a non-negative support.

We considered ten distributions divided into three groups according to the number of parameters. Distributions with two parameters are: the beta prime (BPr) (Johnson et al., 1995), Gompertz (Gom) (Johnson et al., 1995), inverse normal (InvN) (Chhikara and Folks, 1989), lognormal (Gaddum, 1945), log-Laplace (LLap) (Lindsey, 2004), Nakagami (Nak) (Nakagam, 1960) and Shifted Gompertz (SGom) (Bemmaor, 1994). Distributions with three parameters are: the inverse Weibull (InvW) (Drapella, 1993) and generalized gamma (GG) (Stacy, 1962). Distribution with four parameters is the generalized beta of the second kind (GB2) (McDonald, 1984).

The GG and GD2 distributions are actually distribution families. These families consist of other, more or less known distributions, which are referred to as their special cases (see Tables 1 and 2). In fact, there are much more models which could be incorporated in price data modeling. The selected PDFs are presented in Table 3.

Table 1 Sub-models of the GG distribution

a	b	c	Sub-model
-	1	1	Exponential
-	1	-	Gamma
-	1	$c \in N$	Erlang
-	-	1	Weibull
2	1	$0.5n,\ n \in N$	Chi-square
$\sqrt{2}$	2	$0.5n,\ n \in N$	Chi
$\sigma\sqrt{2}$	2	1	Rayleigh
$\sigma\sqrt{2}$	2	1.5	Maxwell–Boltzmann

Source: Own construction based on Stacy and Mihram (1965)

Table 2 Sub-models of the GB2 distribution

a	b	c	d	Sub-model
-	1	-	-	Singh–Maddala (Burr XII)
-	-	1	-	Dagum (Burr III)
-	-	-	1	Beta type II
1	1	-	-	Standard Burr XII
1	-	1	-	Standard Burr III
1	-	-	1	Standard Beta type II
-	1	1	-	Fisk (log-logistic)
-	1	-	1	Lomax (Pareto type II)
-	1	-	h	Paralogistic
-	-	1	1	Inverse Lomax
-	-	-	α	Inverse paralogistic

Source: Mead et al. (2018)

Table 3 The PDFs used for data modeling

Distribution	PDF
BPr	$f_{BPr}(x; a, b) = \frac{x^{a-1}(1+x)^{-a-b}}{B(a,b)}$ $(x \geq 0; a, b > 0)$.
Gom	$f_{Gom}(x; a, b) = abexp(a + bx - ae^{bx})$ $(x \geq 0; a, b > 0)$.
InvN	$f_{InvN}(x; a, b) = \sqrt{\frac{b}{2\pi x^3}}\ exp\left[-\frac{b(x-a)^2}{2a^2x}\right](x > 0; a, b > 0)$.
Log	$f_{Log}(x; a, b) = \frac{1}{\sqrt{2\pi}bx}\ exp\left[-\frac{(lnx-a)^2}{2b^2}\right](x > 0; a \geq 0, b > 0)$.
LLap	$f_{LLap}(x; a, b) = \frac{1}{2bx}\ exp\left[-\frac{\lvert ln(x)-a\rvert}{b}\right](x > 0; a, b > 0)$.
Nak	$f_{Nak}(x; a, b) = \frac{2a^a}{\Gamma(a)b^a}\ x^{2a-1}exp\left[-\frac{a}{b}x^2\right](x \geq 0; a \geq 0.5, b > 0)$.
SGom	$f_{SGom}(x; a, b) = ae^{-ax}exp\ (-be^{-ax})\ [1 + b(1 - e^{-ax})]$ $(x \geq 0; a, b \geq 0)$.
InvW	$f_{InvW}(x; a, b, c) = \frac{c}{b}\left(\frac{x-a}{b}\right)^{-1-c}\ exp\left[-\left(\frac{x-a}{b}\right)^{-c}\right](x > 0; b, c > 0; a \in R)$.
GG	$f_{GG}(x; a, b, c) = \frac{b}{a\Gamma(c)}\left(\frac{x}{a}\right)^{bc-1}\ exp\left[-\left(\frac{x}{a}\right)^{b}\right](x \geq 0; a, b, c > 0)$.
GB2	$f_{GB2}(x; a, b, c, d) = \frac{d}{aB(\mathrm{b}, c)}\left(\frac{x}{a}\right)^{bd-1}\left[1+\left(\frac{x}{a}\right)^{d}\right]^{-(b+c)}$ $(x \geq 0; a, b, c, d > 0)$.

Source: Own construction

2.2 The used goodness-of-fit tests

Let $M(\mathbf{\Theta})$ be the model with the vector of parameters $\mathbf{\Theta}$ and $f_M\ (x; \mathbf{\Theta})$ be the PDF of this model. Let $x_1^*, x_2^*, \ldots, x_n^*$ be a random sample of size n from the $M(\mathbf{\Theta})$. Our target is to estimate the unknown parameters $\mathbf{\Theta}$ by using the maximum likelihood estimation (MLE) method. The likelihood function is given by:

$$L(\mathbf{\Theta}) = \prod_{i=1}^{n} f_M(x_i^*; \mathbf{\Theta}), \tag{1}$$

then the log-likelihood function is defined as:

$$l(\mathbf{\Theta}) = lnL(\mathbf{\Theta}) = \sum_{i=1}^{n} ln[f_M(x_i^*; \mathbf{\Theta})]. \tag{2}$$

Formulas $\frac{dl}{d\mathbf{\Theta}}$ have complex forms. In practice, the calculation of these derivatives is not necessary. We had better maximize the log-likelihood function using a mathematical software instead of struggling with a system of complicated nonlinear equations that may have extraneous roots.

To avoid local maxima of the log-likelihood function, the optimization routine was run repeatedly each time from different starting values that are widely scattered in the parameter space. The maximum

likelihood estimates of parameters **Θ** were calculated in R software (R Core Team, 2014) using the *fitdistr()* function (package *MASS*).

The K-S, AD and CVM tests were used for model fitting, while the information criteria such as AIC, BIC and HQIC were used for comparisons of models. Let us remind the reader that:

$$AIC = -2l + 2p,\ BIC = -2l + pln(n),\ HQIC = -2l + 2pln(ln(n)), \tag{3}$$

where l is the log-likelihood function (2), n is the sample size and p is the number of model parameters.

3 EMPIRICAL STUDY

3.1 Description of the used scanner data sets

In the following empirical study we use scanner data from one retail chain in Poland, i.e. monthly data on natural yoghurt (subgroup of COICOP 5 group: 011441), drinking yoghurt (subgroup of COICOP 5 group: 011441), long grain rice (subgroup of COICOP 5 group: 011111) and coffee powder (subgroup of COICOP 5 group: 012111) sold in 212 outlets in January and February 2022 (52 618 records, which means 42 MB of data). These groups will be designated in our study as **Cases 1–4**, respectively. We defined a homogeneous product at the most detailed level, i.e. at the EAN bar code level. We detected the following number of different EANs with respect to analyzed product groups: 59 (natural yoghurt), 106 (drinking yoghurt), 28 (long grain rice) and 98 (coffee powder). For each EAN the monthly price was calculated as the so called unit value, i.e. the monthly product price was determined as the quotient of the total value of sales of a given product by the number of units of the product sold. For each analyzed **Case**, the following variants for the price samples were considered: prices from the beginning of the research period (denoted by "**B**"), prices from the end of the research period (denoted by "**E**") and the variant with partial price indices (variant "**I**" with relative prices, i.e. ratios of February prices to January prices).

3.2 Scanner data processing applied

Before fitting probability distributions, the data sets (mentioned in Section 3.1) were carefully prepared. First, after deleting records with the missing data and performing the deduplication process, the products were classified first into the relevant elementary groups (COICOP level 5) and then into their subgroups (local COICOP level 6). Product classification was performed using the *data_selecting()* and *data_classification()* functions from the *PriceIndices* R package (Białek, 2021). The first function required manual preparation of dictionaries of keywords and phrases that identified individual product groups. The second function was used for problematic, previously unclassified products and required manual preparation of learning samples based on historical data. The classification itself was based on machine learning using random trees and the *XGBoost* algorithm (Tianqi and Carlo, 2016). Next, the product matching was carried out based on the available GTIN bar codes, internal retail chain codes and product labels. To match products we used the *data_matching()* function from the *PriceIndices* package. To be more precise: products with two identical codes or one of the codes identical and an identical description were automatically matched. Products were also matched if they had identical one of the codes and the Jaro-Winkler (1989) distance of their descriptions was smaller than the fixed precision value: 0.02. In the last step before calculating indices, two data filters were applied to remove unrepresentative products from the database, i.e. the *data_filtering()* function from the cited package was used. The extreme price filter (Białek and Beręsewicz, 2021) was applied to eliminate products with more than three-fold price increase or more than double price drop from month to month. The low sale filter (van Loon and Roels, 2018) was used to eliminate products with relatively low sales from the sample (almost 35% of products were removed).

3.3 Main results

Figures 1–4 show the estimated PDF and CDF for the selected models in relation to Cases I–IV, respectively. With very similar shapes of the estimated PDFs (see e.g. Figure 4; B, E data), additional numerical measures are necessary.

The first group of considered numerical measures consists of the values of information criteria: AIC, BIC and HQIC. Tables 4, 6, 8, 10 display values of the MLEs and the information criteria for Cases I–IV, respectively. The lowest values of the information criteria are marked in bold.

The second group of numerical measures includes values of all considered test statistics and the corresponding p-values. Tables 5, 7, 9, 11 present test statistic values and p-values calculated for the K-S, AD and CVM tests. The lowest statistics values (the highest p-values) are noted in bold.

The p-values for a given model were calculated as follows. Let $\boldsymbol{\Theta}$ be the vector of model parameters. Having estimated parameters vector $\hat{\boldsymbol{\Theta}}$ for a given sample of size n, we calculated test statistics $T(\hat{\boldsymbol{\Theta}}, n)$. Next, we generated 10^5 samples of size n for the given model with the estimated parameters vector $\hat{\boldsymbol{\Theta}}$ For each obtained sample s, we calculated the value of $T_i^s(\hat{\boldsymbol{\Theta}}, n)$. Finally, the p-value can be approximated as follows:

$$p \approx \#\{i\colon T_i^s(\hat{\boldsymbol{\Theta}}, n) > T(\hat{\boldsymbol{\Theta}}, n)\}10^{-5}. \tag{4}$$

As it is shown in Table 4, the GB2 model is the best in terms of AIC values for B, E, I data and in terms of BIC, HQIC values for E, I data. The Nak model is the best in terms of BIC, HQIC values for B data. The GB2 model (see Table 5) is definitely highlighted by the test statistic values and p values. The p-value ranking for the K-S test is the same as the p-value rankings for the AD and CvM tests. Please note, that the ranking on the basis of the information criteria differs from the analogical ranking based on p-values.

Figure 1 PDFs and CDFs of distributions for (B), (E), (I), respectively. Case I

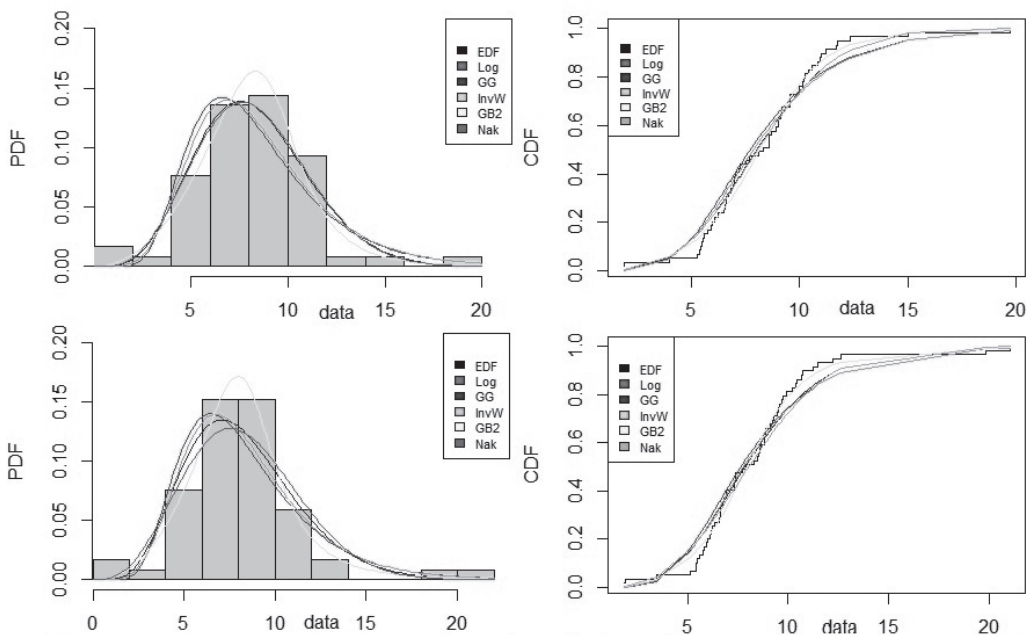
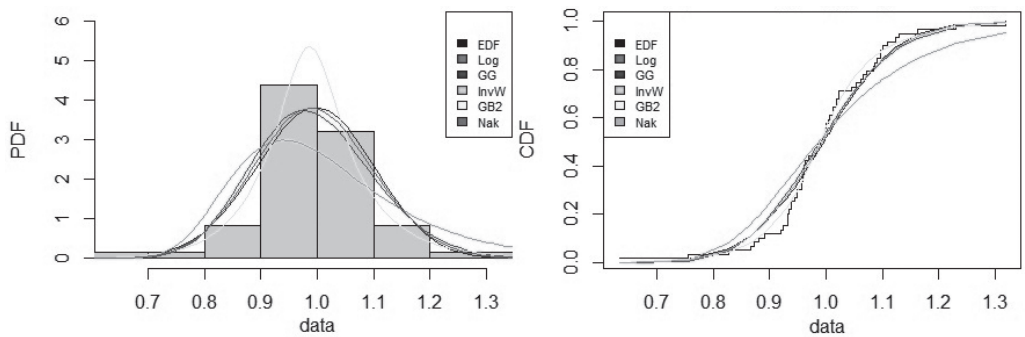


Figure 1 (continuation)



Source: Own construction in R

Table 4 Values of MLEs and information criteria. Case I

Model	Data	MLEs	AIC	BIC	HQIC
Log	B	$\hat{a} = 2.0451, \hat{b} = 0.3916$	302.118	306.273	303.74
	E	$\hat{a} = 2.0343, \hat{b} = 0.4039$	304.503	308.658	306.125
	I	$\hat{a} = -0.0109, \hat{b} = 0.1084$	–91.993	–87.838	–90.371
InvW	B	$\hat{c} = -329.72, \hat{b} = 336.6, \hat{a} = 128.27$	298.478	304.711	300.911
	E	$\hat{c} = -190.29, \hat{b} = 197.11, \hat{a} = 74.80$	301.191	307.424	303.624
	I	$\hat{c} = -28.02, \hat{b} = 28.966, \hat{a} = 236.07$	–72.178	–65.945	–69.745
Nak	B	$\hat{b} = 2.17389, \hat{a} = 76.4807$	294.080	**298.235**	**295.702**
	E	$\hat{b} = 1.8883, \hat{a} = 78.0220$	302.415	306.570	304.037
	I	$\hat{b} = 22.6211, \hat{a} = 1.0005$	–94.792	–90.637	–93.170
GG	B	$\hat{a} = 4.8563, \hat{b} = 1.7551, \hat{c} = 2.7431$	295.954	302.186	298.387
	E	$\hat{a} = 1.7696, \hat{b} = 1.1166, \hat{c} = 5.6167$	302.121	308.354	304.554
	I	$\hat{a} = 0.5932, \hat{b} = 3.6700, \hat{c} = 7.0233$	–93.469	–87.236	–91.036
GB2	B	$\hat{b} = 9.70, \hat{a} = 9.52, \hat{d} = 0.33, \hat{c} = 0.70$	**292.778**	301.088	296.022
	E	$\hat{b} = 11.38, \hat{a} = 8.7, \hat{d} = 0.29, \hat{c} = 0.46$	**294.889**	**303.199**	**298.133**
	I	$\hat{b} = 51.79, \hat{a} = 0.99, \hat{d} = \hat{c} = 0.27$	**–99.038**	**–90.728**	**–95.794**

Source: Own calculations in R

Table 5 Goodness-of-fit tests. Case I

Model	Data	KS		AD		CVM	
		statistic	p-value	statistic	p-value	statistic	p-value
Log	B	0.1093	0.4492	1.3810	0.2085	0.1698	0.3354
	E	0.1284	0.2621	1.588	0.1570	0.2044	0.2603
	I	0.1387	0.1860	1.0977	0.3087	0.1668	0.3427

Table 5 (continuation)

Model	Data	KS		AD		CVM	
		statistic	p-value	statistic	p-value	statistic	p-value
InvW	B	0.0995	0.5696	1.0789	0.3197	0.1313	0.4553
	E	0.1135	0.4002	1.2654	0.2411	0.1617	0.3533
	I	0.1884	0.0260	3.0886	0.0239	0.5037	0.0378
Nak	B	0.0961	0.6105	0.7224	0.5374	0.0795	0.6948
	E	0.1167	0.3663	1.3943	0.2033	0.1767	0.3174
	I	0.1262	0.2776	0.9383	0.3898	0.1447	0.4055
GG	B	0.0954	0.6233	0.727	0.5367	0.0798	0.6956
	E	0.1158	0.3808	1.2141	0.2650	0.1456	0.4067
	I	0.1200	0.3370	0.9207	0.4014	0.1458	0.4040
GB2	B	**0.0893**	**0.7005**	**0.4409**	**0.8060**	**0.0622**	**0.8016**
	E	**0.0819**	**0.7936**	**0.4946**	**0.7523**	**0.0637**	**0.7939**
	I	**0.0807**	**0.8083**	**0.3494**	**0.8969**	**0.0491**	**0.8837**

Source: Own calculations in R

As shown in Table 4, the GB2 model is the best in terms of AIC values for B, E, I data and in terms of BIC, HQIC values for E, I data. The Nak model is the best in terms of BIC, HQIC values for B data.

Figure 2 PDFs and CDFs of distributions for (B), (E), (I), respectively. Case II

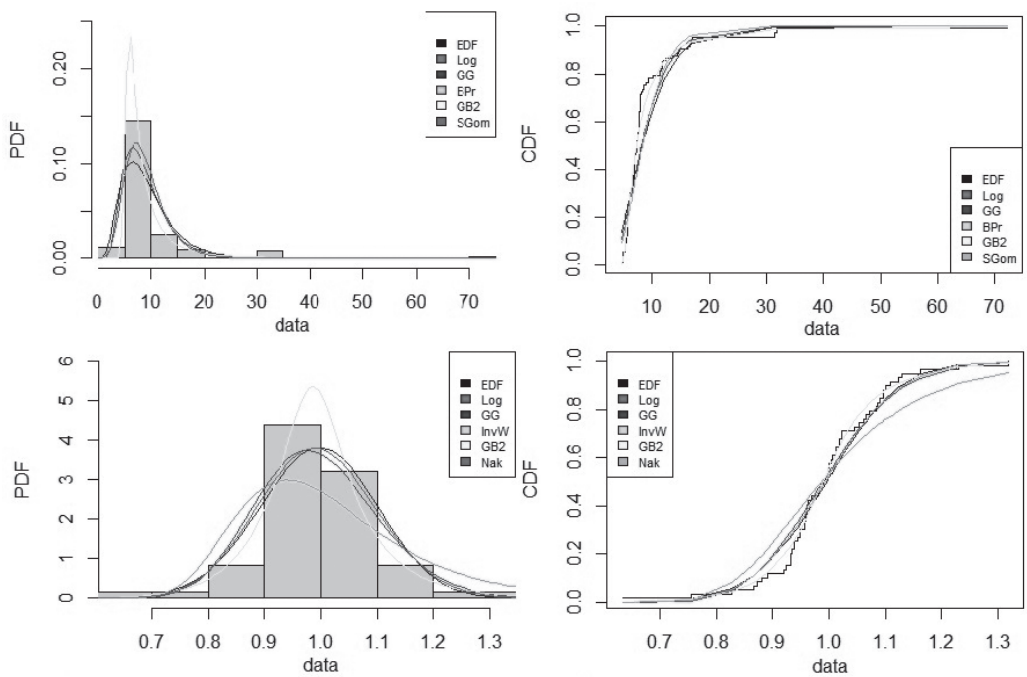
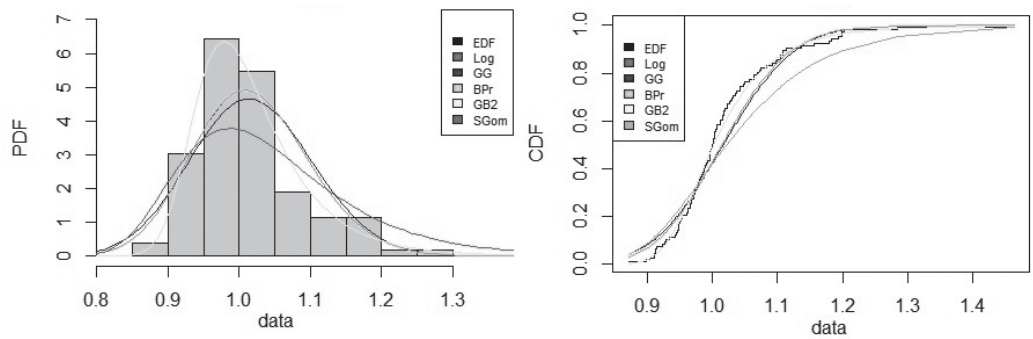


Figure 2 (continuation)



Source: Own construction in R

The GB2 model (see Table 5) is definitely highlighted by goodness-of-fit tests. The p-value ranking for the K-S test is the same as the p-value rankings for the AD and CvM tests. Based on the graphical and the numerical results, the GB2 and Nak models are considered as ones of the best models for the analyzed data set.

The GB2 model (see Table 6) is the best in terms of information criteria values. The GB2 model (see Table 7) is definitely distinguished by goodness-of-fit tests. The p-value ranking for the K-S test is the same as the p-value rankings for the AD and CvM tests. Based on the graphical and the numerical results, the GB2 model is considered as one of the best models for the Case II.

Table 6 Values of MLEs and information criteria. Case II

Model	Data	MLEs	AIC	BIC	HQIC
Log	B	$\hat{a} = 2.0977, \hat{b} = 0.4639$	302.118	306.273	303.74
	E	$\hat{a} = 2.1125, \hat{b} = 0.4897$	304.503	308.658	306.125
	I	$\hat{a} = 0.0148, \hat{b} = 0.0803$	–91.993	–87.838	–90.371
BPr	B	$\hat{a} = 51.7207, \hat{b} = 6.7802$	298.478	304.711	300.911
	E	$\hat{a} = 46.6895, \hat{b} = 6.0785$	301.191	307.424	303.624
	I	$\hat{a} = 313.0666, \hat{b} = 308.4696$	–72.178	–65.945	–69.745
SGom	B	$\hat{a} = 0.3309, \hat{b} = 9.7901$	294.080	298.235	295.702
	E	$\hat{a} = 0.2962, \hat{b} = 7.5927$	302.415	306.570	304.037
	I	$\hat{a} = 10.2637, \hat{b} = 24996.2643$	–94.792	–90.637	–93.170
GG	B	$\hat{a} = 0.0147, \hat{b} = 0.4545, \hat{c} = 18.1429$	295.954	302.186	298.387
	E	$\hat{a} = 0.0142, \hat{b} = 0.4431, \hat{c} = 17.2612$	302.121	308.354	304.554
	I	$\hat{a} = 0.1128, \hat{b} = 1.7436, \hat{c} = 46.5808$	–93.469	–87.236	–91.036
GB2	B	$\hat{b} = 8.60, \hat{a} = 3.38, \hat{d} = 46.76, \hat{c} = 0.28$	**292.778**	**301.088**	**296.022**
	E	$\hat{b} = 10.95, \hat{a} = 5.2, \hat{d} = 1.41, \hat{c} = 0.21$	**294.889**	**303.199**	**298.133**
	I	$\hat{b} = 22.54, \hat{a} = 0.91, \hat{d} = 4.23, \hat{c} = 0.705$	**–99.038**	**–90.728**	**–95.794**

Source: Own calculations in R

Table 7 Goodness-of-fit tests. Case II

Model	Data	KS		AD		CVM	
		statistic	p-value	statistic	p-value	statistic	p-value
Log	B	0.2309	0	5.9991	0.0011	1.1011	0.0015
	E	0.1817	0.0015	4.9616	0.0030	0.8780	0.0046
	I	0.1307	0.0490	2.3762	0.0576	0.4320	0.0596
BPr	B	0.21107	0.0001	4.4519	0.0053	0.8237	0.0063
	E	0.1625	0.0067	3.4589	0.0165	0.6074	0.0213
	I	0.1307	0.0491	2.3736	0.0588	0.4319	0.0604
SGom	B	0.2507	0.00000	7.1156	0.0003	1.3038	0.0004
	E	0.2094	0.0001	6.4618	0.0007	1.1511	0.0011
	I	0.1772	0.0023	5.6305	0.0015	1.0541	0.0017
GG	B	0.2436	0	7.4065	0.0003	1.3697	0.0003
	E	0.1911	0.0008	6.2738	0.0008	1.1272	0.0012
	I	0.1418	0.0259	2.9891	0.0284	0.5430	0.0318
GB2	B	**0.1176**	**0.0993**	**0.8511**	**0.4517**	**0.1615**	**0.3580**
	E	**0.0704**	**0.6438**	**0.4582**	**0.7999**	**0.0645**	**0.7866**
	I	**0.0577**	**0.8513**	**0.3730**	**0.8759**	**0.0663**	**0.7766**

Source: Own calculations in R

Figure 3 PDFs and CDFs of distributions for (B), (E), (I), respectively. Case III

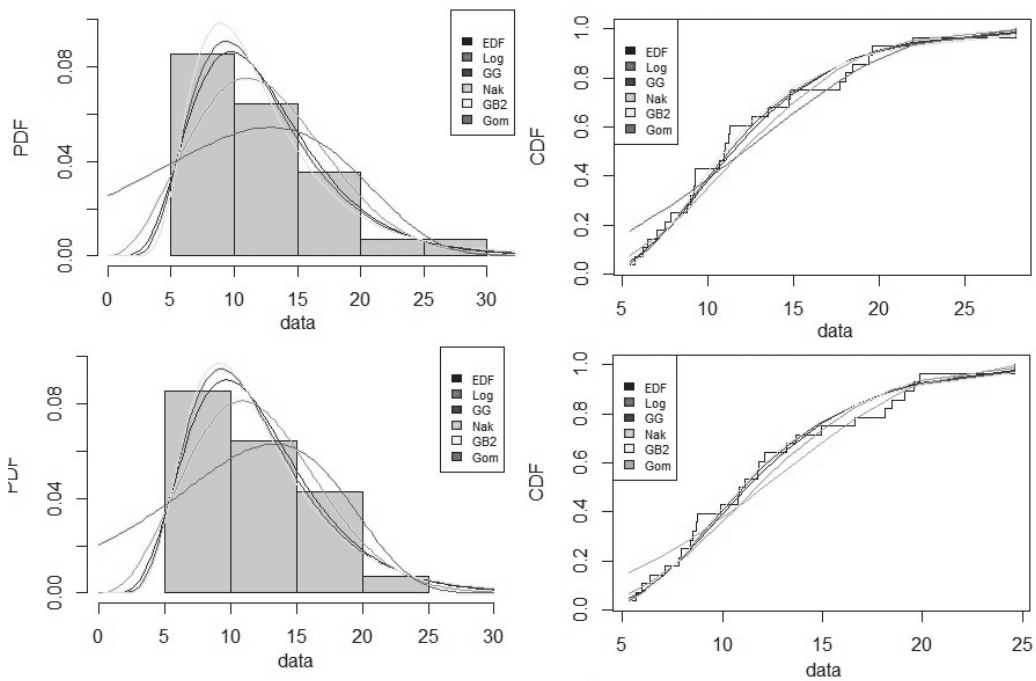
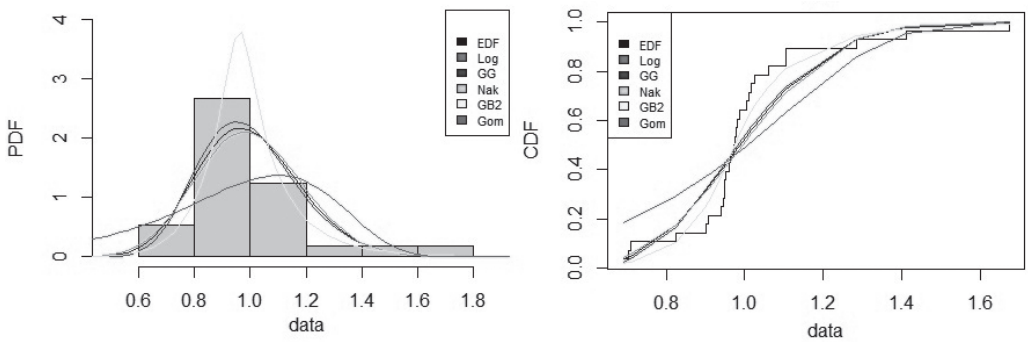


Figure 3 (continuation)



Source: Own construction in R

The Log model (see Table 8) is the best in terms of information criteria values for B, E data and the GB2 model is the best in terms of information criteria value for I data. The GB2 model (see Table 9) is the best in terms of goodness-of-fit tests. The p-value ranking for the K-S test is the same as the p-value rankings for the AD and CvM tests. Based on the graphical and the numerical results, the Log and GB2 models are considered to be best models in the case of the third analyzed data set.

Table 8 Values of MLEs and information criteria. Case III

Model	Data	MLEs	AIC	BIC	HQIC
Log	B	$\hat{a} = 2.4195, \hat{b} = 0.4286$	**171.507**	**174.171**	**172.321**
	E	$\hat{a} = 2.4019, \hat{b} = 0.4157$	**168.808**	**171.472**	**169.622**
	I	$\hat{a} = -0.0176, \hat{b} = 0.1809$	−13.281	−10.616	−12.466
Nak	B	$\hat{b} = 1.5029, \hat{a} = 182.3891$	174.781	177.445	175.595
	E	$\hat{b} = .6492, \hat{a} = 170.1581$	170.679	173.343	171.493
	I	$\hat{b} = 7.1613, \hat{a} = 1.0369$	−10.1622	−7.4978	−9.3477
Gom	B	$\hat{a} = 0.02556, \hat{b} = 0.1196$	182.754	185.418	183.568
	E	$\hat{a} = 0.0204, \hat{b} = 0.1497$	177.189	179.854	178.004
	I	$\hat{a} = 0.0642, \hat{b} = 3.6824$	7.689	10.353	8.503
GG	B	$\hat{a} = 0.0303, \hat{b} = 0.5097, \hat{c} = 20.9009$	174.037	178.034	175.259
	E	$\hat{a} = 0.0614, \hat{b} = 0.5585, \hat{c} = 18.6752$	171.062	175.059	172.284
	I	$\hat{a} = 0.0597, \hat{b} = 1.1157, \hat{c} = 23.2369$	−9.8031	−5.8065	−8.5813
GB2	B	$\hat{b} = 1.10, \hat{a} = 1.75, \hat{d} = 39.14, \hat{c} = 5.44$	175.265	180.594	176.893
	E	$\hat{b} = 0.63, \hat{a} = 1.92, \hat{d} = 58.11, \hat{c} = 19.44$	172.836	178.165	174.465
	I	$\hat{b} = 62.04, \hat{a} = 0.96, \hat{d} = 0.15, \hat{c} = 0.13$	**−18.987**	**−13.658**	**−17.358**

Source: Own calculations in R

Table 9 Goodness-of-fit tests. Case III

Model	Data	KS		AD		CVM	
		statistic	p-value	statistic	p-value	statistic	p-value
Log	B	0.1057	0.8800	0.2854	0.9492	0.04485	0.9103
	E	0.1043	0.8896	0.2833	0.9496	0.0382	0.9444
	I	0.1944	0.2102	1.5174	0.1710	0.2930	0.1406
Nak	B	0.1602	0.4254	0.5675	0.6791	0.1006	0.5870
	E	0.1279	0.7028	0.4825	0.7628	0.0777	0.7082
	I	0.2232	0.1040	1.7599	0.1242	0.3435	0.1001
Gom	B	0.1773	0.3070	1.0727	0.3227	0.1609	0.3613
	E	0.1521	0.4890	0.8627	0.4365	0.1291	0.4611
	I	0.2624	0.0342	3.2897	0.0201	0.6357	0.0178
GG	B	0.1184	0.785	0.3341	0.910	0.0547	0.851
	E	0.1126	0.8305	0.3150	0.92525	0.0446	0.9113
	I	0.2089	0.151	1.6257	0.150	0.3171	0.120
GB2	B	**0.1055**	**0.8822**	**0.2504**	**0.970**	**0.0366**	**0.952**
	E	**0.0994**	**0.9185**	**0.2746**	**0.9552**	**0.0362**	**0.9539**
	I	**0.1545**	**0.4685**	**0.7612**	**0.5570**	**0.1249**	**0.4783**

Source: Own calculations in R

Figure 4 PDFs and CDFs of distributions for (B), (E), (I), respectively. Case IV

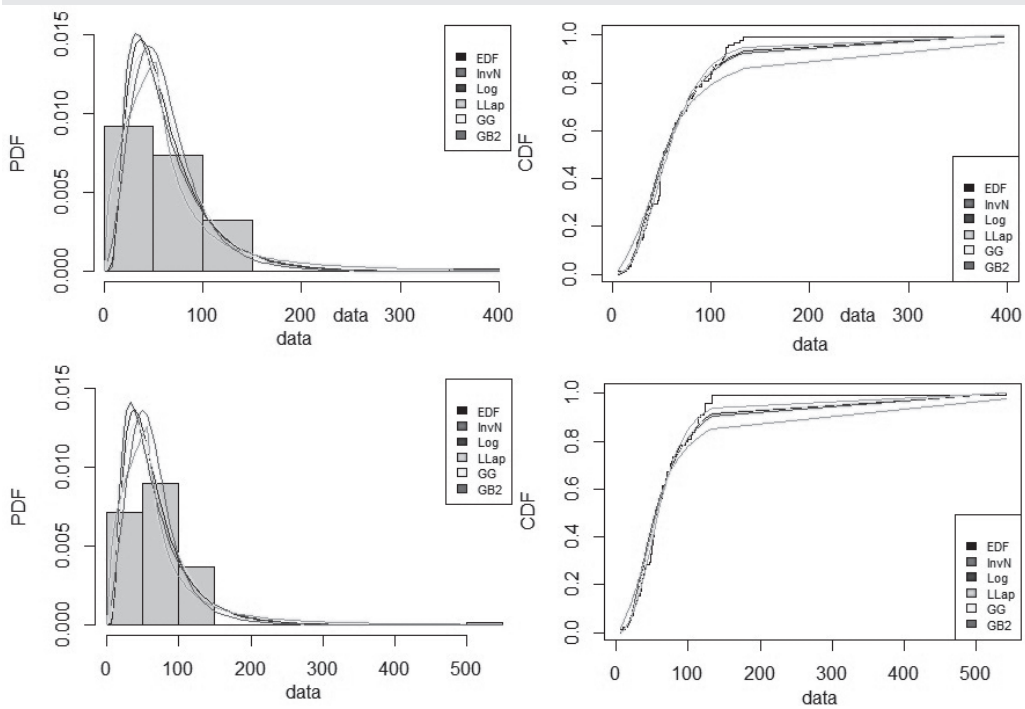
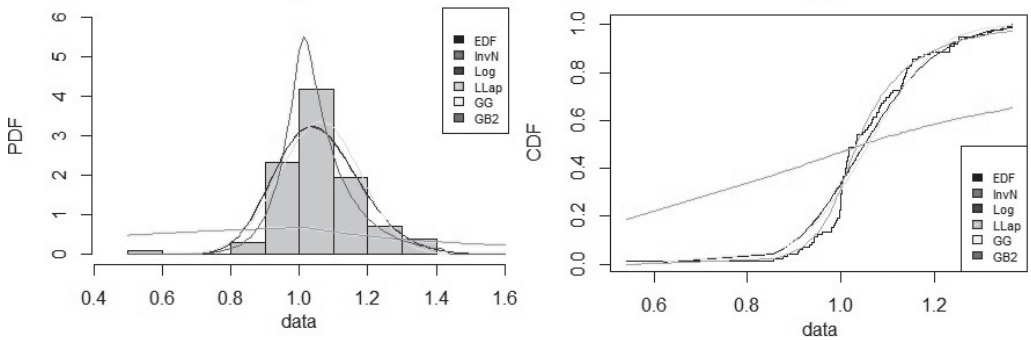


Figure 4 (continuation)



Source: Own construction in R

Table 10 Values of MLEs and information criteria. Case IV

Model	Data	MLEs	AIC	BIC	HQIC
InvN	B	$\hat{a} = 3.9891, \hat{b} = 135.4686$	971.116	976.286	973.207
	E	$\hat{a} = 68.6115, \hat{b} = 127.9516$	991.298	996.468	993.390
	I	$\hat{a} = 1.0579, \hat{b} = 75.4539$	−126.964	−121.794	−124.873
Log	B	$\hat{a} = 3.9796, \hat{b} = 0.6072$	964.353	**969.523**	**966.445**
	E	$\hat{a} = 4.0294, \hat{b} = 0.6312$	981.689	**986.859**	983.780
	I	$\hat{a} = 0.0498, \hat{b} = 0.1174$	−128.024	−122.854	−125.933
LLap	B	$\hat{a} = 3.9500, \hat{b} = 0.9748$	979.515	984.685	981.606
	E	$\hat{a} = 4.0239, \hat{b} = 0.9958$	990.352	995.522	992.443
	I	$\hat{a} = 0.0179, \hat{b} = 0.9604$	104.051	109.221	106.142
GG	B	$\hat{a} = 0.0827, \hat{b} = 0.4213, \hat{c} = 15.7802$	965.043	972.798	968.179
	E	$\hat{a} = 0.0515, \hat{b} = 0.3959, \hat{c} = 16.4507$	982.693	990.448	985.829
	I	$\hat{a} = 0.6722, \hat{b} = 3.7761, \hat{c} = 5.8984$	−134.951	−127.196	−131.814
GB2	B	$\hat{b} = 2.79, \hat{a} = 70.66, \hat{d} = 0.90, \hat{c} = 1.49$	**963.290**	973.630	967.472
	E	$\hat{b} = 3.79, \hat{a} = 69.99, \hat{d} = 0.59, \hat{c} = 0.91$	**976.721**	987.061	**980.904**
	I	$\hat{b} = 105.66, \hat{a} = 1.01, \hat{d} = 0.16, \hat{c} = 0.10$	**−149.997**	**−139.657**	**−145.815**

Source: Own calculations in R

Table 11 Goodness-of-fit tests. Case IV

Model	Data	KS		AD		CVM	
		statistic	p-value	statistic	p-value	statistic	p-value
InvN	B	0.1312	0.0618	1.1101	0.3020	0.1775	0.3132
	E	0.1301	0.0661	1.6027	0.1530	0.2493	0.1884
	I	0.1420	0.0347	2.2576	0.0679	0.3776	0.0834

Table 11 (continuation)

Model	Data	KS		AD		CVM	
		statistic	p-value	statistic	p-value	statistic	p-value
Log	B	0.1092	0.1786	0.6996	0.5569	0.1010	0.5797
	E	0.1041	0.2228	0.9124	0.4067	0.1265	0.4692
	I	0.1393	0.0406	2.2054	0.0713	0.3707	0.0862
LLap	B	0.1305	0.0751	2.5158	0.0584	0.3203	0.1354
	E	0.13844	0.0418	2.4393	0.0533	0.3188	0.1203
	I	0.3647	0.0000	25.2491	0.0000	5.0748	0.0000
GG	B	0.0932	0.3421	0.5533	0.6939	0.0763	0.7158
	E	0.0888	0.3965	0.7447	0.5220	0.0977	0.5954
	I	0.1244	0.0899	2.3132	0.0635	0.4049	0.0715
GB2	B	**0.0734**	**0.6386**	**0.4695**	**0.7805**	**0.0681**	**0.7654**
	E	**0.0592**	**0.8600**	**0.4195**	**0.8279**	**0.0529**	**0.8595**
	I	**0.1039**	**0.2242**	**1.0053**	**0.8787**	**0.1700**	**0.3343**

Source: Own calculations in R

The Log model (see Table 10) is the best in terms of BIC, HQIC values for B data and in terms of BIC values for E data. The GB2 model (see Table 10) is the best in terms of AIC values for B data, in terms of AIC and HQIC values for E data, in terms of information criteria values for I data. The GB2 model (see Table 11) is the best in terms of goodness-of-fit tests. The p-value ranking for the K-S test provides the same hierarchy of models as p-value rankings based on the AD and CvM tests. On the basis of graphical and numerical results, the Log and GB2 models are considered to be best models for the Case IV.

CONCLUSIONS

We used a distribution family with a non-negative domain to model scanner prices and relative scanner prices of natural yoghurt, drinking yoghurt, long grain rice and coffee. For the ranking of selected models, we used the values of the information criteria and p-values calculated for the goodness-of-fit tests. Interestingly, the ranking of models according to the AIC criterion is the same as according to the BIC and HQIC criteria (see Section 3.3). The ranking of the models according to the p-values determined for the K-S test is the same as according to the p-values obtained for the AD and CVM tests.

The article shows that the greater the number of model parameters, the more special cases a given model has (see Tables 1 and 2). One might expect that as the number of model parameters increases, the model will fit the data better. This rule does not apply to prices in Case 3 (see Table 8; data B, E), as the values of the information criteria taking into account the number of model parameters are smaller for the Log model (with two parameters) than for the GB2 model (with four parameters). In general, however, models with more parameters allow for more flexibility in the manipulation of normal and central moments of the distribution, which may be important in organizing simulation studies on price indices.

In summary, the generalized beta of the second type and the lognormal model are best suited for modeling scanner prices and relative scanner prices. Good results for the lognormal distribution obtained for the analyzed food products are consistent with the common opinion that this distribution characterizes product prices well (Silver and Heravi, 2007). This model was implemented in the PriceIndices package in the generate() function, which is used to generate artificial scanner data sets (Białek, 2021). Several of the remaining models also seem to be of good quality in price modeling, with the final selection

of the model probably depending on the product segment and the definition of a homogeneous product (the lower the aggregation level, the greater the price fluctuations we observe).

Potential directions for further work include an attempt to model the amount of purchased products (and thus consumption distribution) and, consequently, possibly also weighted indices. From the theoretical point of view, it would also be interesting to investigate whether the expected values determined on the basis of the theoretical distributions of weighted indices correspond to their sample values.

References

BEMMAOR, A. C. (1994). Modeling the diffusion of new durable goods: Word-of-mouth effect versus consumer heterogeneity [online]. In: LAURENT, G., LILIEN, G. L., PRAS, B. (eds.) *Research Traditions in Marketing*. Boston: Kluwer, 201–229. <https://doi.org/10.1007/978-94-011-1402-8_6>.

BIAŁEK, J. (2015). Construction of confidence intervals for the Laspeyres price index [online]. *Journal of Statistical Computation and Simulation*, 85(14): 2962–2973. <https://doi.org/10.1080/00949655.2014.946416>.

BIAŁEK, J., BOBEL, A. (2019). *Comparison of Price Index Methods for CPI Measurement using Scanner Data.* Paper presented at the 16th Meeting of the Ottawa Group on Price Indices, Rio de Janeiro, Brazil.

BIAŁEK, J. (2020). Remarks on Price Index Methods for the CPI Measurement Using Scanner Data [online]. *Statistika: Statistics and Economy Journal*, 100(1): 54–69, Prague: Czech Statistical Office. <https://www.czso.cz/documents/10180/125507867/32019720q1_54_bialek.pdf/f4ee19a0-75fd-41bf-b1fd-b192d177e125?version=1.2>.

BIAŁEK, J. (2021). PriceIndices – a New R Package for Bilateral and Multilateral Price Index Calculations [online]. *Statistika: Statistics and Economy Journal*, 101(2): 122-141, Prague: Czech Statistical Office. <https://www.czso.cz/documents/10180/143550797/32019721q2_bialek.pdf/3cd5bf11-22f4-4ee5-b294-1d7d5909e4b4?version=1.1>.

BIAŁEK, J., BERĘSEWICZ, M. (2021). Scanner data in inflation measurement: from raw data to price indices [online]. *The Statistical Journal of the IAOS*, 37: 1315–1336. <https://doi.org/10.3233/sji-210816>.

BIAŁEK, J. (2022). Elementary price indices under the GBM price model [online]. *Communications in Statistics – Theory and Methods*, 51(5): 1232–1251. <https://doi.org/10.1080/03610926.2021.1938127>.

CARLI, G. (1804). Del valore e della proporzione de'metalli monetati. In: *Scrittori Classici Italiani di Economia Politica*, 13: 297–336.

CHESSA, A. (2016). A new methodology for processing scanner data in the Dutch CPI. *Eurostat review of National Accounts and Macroeconomic Indicators*, 1: 49–69.

CHHIKARA, R. S., FOLKS, J. L. (1989). *The Inverse Gaussian Distribution: Theory, Methodology and Applications*. New York, USA: Marcel Dekker.

DRAPELLA, A. (1993). The complementary Weibull distribution: unknown or just forgotten? [online]. *Quality and reliability engineering international*, 9(4): 383–385. <https://doi.org/10.1002/qre.4680090426>.

DUTOT C. F. (1738). *Reflexions Politiques sur les Finances et le Commerce.* The Hague: Les Freres.

GADDUM, J. H. (1945). Lognormal distributions. *Nature*, 156(3964): 463–466.

JARO, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida [online]. *Journal of the American Statistical Association* 84(406): 414–420. <https://doi.org/10.1080/01621459.1989.10478785>.

JEVONS, W. S. (1865). The variation of prices and the value of the currency since 1782. *Journal of the Statistical Society of London*, 28: 294–320.

JOHNSON, N. L., KOTZ, S., BALAKRISHNAN, N. (1995). *Continuous Univariate Distributions* [online]. 2nd Ed., Vol. 2, Wiley. <https://doi.org/10.2307/2348907>.

LINDSEY, J. K. (2004). *Statistical analysis of stochastic processes in time* [online]. Cambridge University Press, Vol. 14. <https://doi.org/10.1017/cbo9780511617164>.

MEAD, M., NASSAR, M. M., DEY, S. (2018). A generalization of generalized gamma distributions [online]. *Pakistan Journal of Statistics and Operation Research*, 121–138. <https://doi.org/10.18187/pjsor.v14i1.1692>.

MCDONALD, J. B. (1984). Some generalized functions for the size distribution of income [online]. *Econometrica*, 52: 647–663. <https://doi.org/10.2307/1913469>.

NAKAGAM, M. (1960). The m-Distribution – a General Formula of Intensity Distribution of Rapid Fading [online]. In: HOFFMAN, W. C. (eds.) *Statistical Methods in Radio Wave Propagation*, Pergamon, 3–36. <https://doi.org/10.1016/b978-0-08-009306-2.50005-4>.

R CORE TEAM. (2014). R: *A language and environment for statistical computing* [online]. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org>.

SILVER, H., HERAVI, S. (2007). Why elementary price index number formulas differ: Evidence on price dispersion [online]. *Journal of Econometrics*, 140: 874–883. <https://doi.org/10.1016/j.jeconom.2006.07.017>.

STACY, E. W. (1962). A generalization of the gamma distribution [online]. *The Annals of mathematical statistics*, 33: 1187–1192. <https://doi.org/10.1214/aoms/1177704481>.

STACY, E. W., MIHRAM, G. A. (1965). Parameter estimation for a generalized gamma distribution [online]. *Technometrics*, 7(3): 349–358. <https://doi.org/10.1080/00401706.1965.10490268>.

TIANQI, C., CARLO, G. (2016). Xgboost: A scalable tree boosting system [online]. Proceedings of the *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, August 13–17, 785–794. <https://doi.org/10.1145/2939672.2939785>.

VAN LOON, K. V., ROELS, D. (2018). *Integrating big data in the Belgian CPI.* Paper presented at the Meeting of the group of experts on consumer price indices, 8–9 May, Geneva, Switzerland.