STATISTIKA

STATISTICS
AND ECONOMY
JOURNAL

VOL. **102** (3) 2022

CZECH
STATISTICAL
OFFICE

EDITOR-IN-CHIEF

EDITORIAL BOARD

EXECUTIVE BOARD

MANAGING EDITOR

Dear Readers,

It has been three years since the publication of the special joint Czech-Slovak Statistika issue (No. 4/2019) and two years since *Statistika: Statistics and Economy Journal* celebrated its 100th Volume (following the previous journals of the state statistical service issued since 1920).

We would like follow up the international cooperation through publication of this new joint Czech-Polish special issue of our professional quarterly: to commemorate and remind events, developments and current state and quality of research in official statistics, to strengthen the further developing and very successful cooperation between our national statistical offices and our two countries. Therefore, this issue include articles from Czech and Polish authors only.

We believe that papers published in this special issue will be interesting and beneficial for all its readers. We are looking for further cooperation (not only) with authors (and reviewers) from our two countries and wish all our colleagues, partners, and collaborators plenty of creative thoughts, professional success, and satisfaction.

Marek Rojíček
President, Czech Statistical Office
Prague

Dominik Rozkrut
President, Statistics Poland
Warsaw

CONTENTS

ANALYSES

CONSULTATION

INFORMATION

About Statistika

The journal of Statistika has been published by the Czech Statistical Office since 1964. Its aim is to create a platform enabling national statistical and research institutions to present the progress and results of complex analyses in the economic, environmental, and social spheres. Its mission is to promote the official statistics as a tool supporting the decision making at the level of international organizations, central and local authorities, as well as businesses. We contribute to the world debate and efforts in strengthening the bridge between theory and practice of the official statistics. Statistika is professional double-blind peer reviewed open access journal included in the citation database of peer-reviewed literature **Scopus** (since 2015), in the **Web of Science** *Emerging Sources Citation Index* (since 2016), and also in other international databases of scientific journals. Since 2011, Statistika has been published quarterly in English only.

Publisher

The Czech Statistical Office is an official national statistical institution of the Czech Republic. The Office´s main goal, as the coordinator of the State Statistical Service, consists in the acquisition of data and the subsequent production of statistical information on social, economic, demographic, and environmental development of the state. Based on the data acquired, the Czech Statistical Office produces a reliable and consistent image of the current society and its developments satisfying various needs of potential users.

Contact us

Journal of Statistika | Czech Statistical Office | Na padesátém 81 | 100 82 Prague 10 | Czech Republic
e-mail: statistika.journal@czso.cz | web: www.czso.cz/statistika_journal

Methodological Aspects of Measuring Preferences Using the Rank and Thurstone Scale

Joanna Dębicka | *Wroclaw University of Economics and Business, Wroclaw, Poland*
Edyta Mazurek[1] | *Wroclaw University of Economics and Business, Wroclaw, Poland*
Katarzyna Ostasiewicz | *Wroclaw University of Economics and Business, Wroclaw, Poland*

Received 31.1.2022 (revision received 1.4.2022), **Accepted** (reviewed) 4.4.2022, **Published** 16.9.2022

Abstract

The fundamental problem with the measurement of preferences is that it not only attempts to measure something that is, by its nature, "unmeasurable", but also hidden from a direct observation. In addition, a person's current emotional, material and social situation influences the measurement of preferences resulting from the person's system of values. The paper is a study on the methodology of preference measurement, a comparison and evaluation of two methods of scale construction. Among various techniques we investigate the two methods: Thurstone procedure for finding scale separations developed by Thurstone and the simplest rank method of scaling. This study examines the relative merits of Thurstone and rank techniques of scale construction.

Keywords	*DOI*	*JEL code*
Preference ranking, Case V, Thurstone model, structure similarity test, input matrix	*https://doi.org/10.54694/stat.2022.5*	*C18, C83, C46*

INTRODUCTION

Questionnaire survey research became very popular in scientific research. In social and psychological research it is often used to describe and explore human behavior (Singleton and Straits, 2018). Moreover, this kind of learning about social preferences is frequently used by policymakers. The presidential ballot is a kind of "survey", in which each citizen is asked about his preferences for the person to be the head of the country. In more common situations citizens are asked to vote for the projects to be addressed by civic funds. While willing to eliminate the barriers that keep disabled persons from full participation in social and civic life, it is worth knowing, which barriers are most burdensome for most of them.

The shift from authoritative decision making to public consultations has been included in the EU's research and innovation program *Horizon Europe 2021–2027* (*Horizon Europe, European Commission*).

[1] Wroclaw University of Economics and Business, Department of Statistics, Komandorska 118/120, Wroclaw, Poland. Corresponding author: e-mail: edyta.mazurek@ue.wroc.pl.

This research agenda has been defined in terms of five missions to be carried out. One of those missions concerns is the adaptation to social transformation consisting directly to involving all actors, including citizens, civil society organizations, and public authorities, in research, innovation and change. One such change/transformation concerns the reduction of socio-economic inequalities. Co-creation is the preferred method of analyzing changes (as well as a key aspect of research and implementation of solutions) in *Horizon Europe 2021–2027*. In this context, the issues of identifying the most necessary changes (ranking individual preferences and building scales) and defining specific actions for priority areas of change (questionnaire research on specific solutions, including the impact of the method of asking a question (type of question) on the answer) become important.

Although most of the researchers agree on the importance of survey research, there are many doubts and controversies concerning the methods of asking questions and, after collecting data, of aggregating answers into something that might be regarded as the collective preferences and choices (Holbrook, Cho and Johnson, 2006). The problem has two most important components. The first is a mainly psychological matter of how to ask questions to make people reveal their true attitudes. Still, some psychologists claim, that surveyed persons do not reveal but rather construct their attitudes in the process of being surveyed. It is known, that websites and online survey software are on the one hand useful to assist in the design and delivery of questionnaires, but, on the other hand, they can also introduce sources of bias (Ball, 2019).

The second crucial issue is the aggregation of preferences, and this is the particular branch of survey studies to which our paper is contributing. It focuses on the methodology of aggregated preference measurement, a comparison and evaluation of two methods of scale construction. Among various techniques, we investigate the two methods: Thurstone procedure for finding scale separations developed by Thurstone and the simplest rank method of scaling.

Stepping back to XVIII century for the discussion between Borda and Condorcet about the best method of aggregating preferences in voting systems, which by now has not found the conclusive solution, one may say that the issue, which method of aggregation is the very best, is a vague and to some degree unscientific but also an axiological question. Still, it is worth comparing different kinds of aggregation methods, to be conscious of potential differences and characteristics. For example, it is worth knowing if different methods give qualitatively different results. If so, it is of crucial importance to consider very thoroughly which method is better for a given aim and why. If the results are similar, it might be of use to investigate more technical properties – stability of the results, sensitivity for individual observations and so on.

Thurstone scaling is the well-known tool for the estimation of preferences among objects by the observed frequencies of their paired comparisons (Thurstone, 1927a; Thurstone and Chave, 1929; Thurstone and Jones, 1957). The positioning of items on this scale can be found by averaging the percentiles of the standard normal distribution corresponding to the proportions of the respondents preferring one item over each of the others. This scaling is widely used in applied psychology, particularly in marketing and advertising research (Edwards and Kenney, 1946; Escher, 2010). Statistical approaches to the Thurstone scaling were considered by Mosteller (Mosteller, 1951), and various modifications of this model were developed by Lipovetsky (Lipovetsky, 2007), and Saffir (Saffir, 1937). The authors made comparison of the methods of attitude scale construction of Thurstone, Likert, and Guttman and Bradley-Terry model (Edwards and Kenney, 1946; Edwards and Kilpatrick, 1948; Lipovetsky and Conklin, 2004; Drasgow, Chernyshenko and Stark, 2010; Tsukida and Gupta, 2011; Stadthagen-González et al., 2018; Edwards and Kilpatrick, 1948). When the number of stimuli is large, the number of pairs to be compared becomes very large, and the similarity task is inefficient. Tsogo, Masson and Bardot reviewed the main similarity task methods suitable for large sets of objects (Tsogo, Masson and Bardot, 2000). They point out the advantages and disadvantages of such methods as: incomplete similarity tasks, binary dissimilarities, hierarchical sorting tasks, conditional rank-order. Among the comparisons, there was no comparison with the classical approach based on the sum of ranks. We decided to compare Thurstone scale and direct

rank method of scaling. In the paper, the words project, object or stimulus, are used interchangeably and symbolically denote a ranking object.

The paper is organized as follows. In Section 1 we give information about the classification of scaling techniques and describe the rank scale and Thurston method in detail. Section 2 consists of results obtained from the survey, which was constructed based on the original Thurstone study that measured social values, specifically the seriousness of different types of crimes or offences. In Section 2.1. we describe the survey, dataset and give rankings of offences, from the worst to the lightest. Section 2.2. offers a comparative analysis of the crime severity scales obtained by ranking methods and the Thurstone scale. In Section 2.3 we check the assumption of the independence of alternatives for subsets of crime and offences. Finally, last Section sums up results obtained in the paper indicating also the direction of future research.

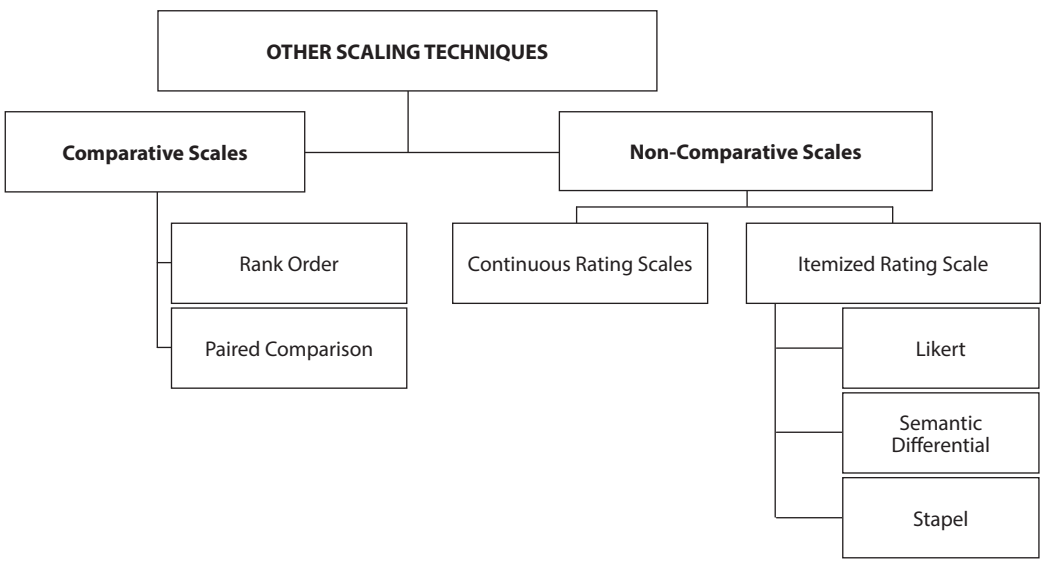
1 SCALING TECHNIQUES

Scaling objects can be used for a comparative study of more than one object. For a long time marketing and some other kinds of researches have been highly dependent on those techniques. Scaling emerged from the social sciences in an attempt to order attributes with respect to quantitative attributes. Scaling provides a mechanism for measuring abstract concepts.

1.1 Classification

In general, scaling techniques can be divided into two categories: non-comparative scales and comparative scales (see Figure 1). A non-comparative scale is used to analyze an individual product or object's performance on different parameters and is most frequently used in marketing research. In this approach each object is scaled independently on the others, e.g., respondents may be asked how they are satisfied with product A, product B, etc., without comparing the product. Contrary, within comparative scales, respondents are asked to place one object regarding other objects.

Figure 1 Classification of scaling techniques



Source: Own elaboration, following standard textbook presentation

Our research focuses on ranking, so on ordinal tasks. Ordinal tasks involve ranking objects in some way to produce dominance data; that is, one stimulus dominates another, so only judgments of greater than or less than are required. Ranking can be accomplished directly or derived from pairing the objects. A paired comparison symbolizes two objects from which the respondent needs to select one according to their preference. The direct ranking consists of assigning integers to objects, indicating the order of preferences. These two methods of creating ranking are most commonly used in practice. For that reason, the paper refers to comparing two methods of attitude scale construction. Both rank scale and Thurstone method belong to comparative scales. The difference in classification is, that the rank scale can be accomplished only by rank order, while Thurstone method can be applied in the case of both pairwise comparisons and by rank order (from which pairwise comparisons can be obtained).

1.2 Rank scale

The method of determining the scale based on the assigned ranks will be presented in an example (see Table 1). Let us assume that we have five respondents (so-called judges). Each respondent orders crimes A-E from the worst to the lightest. Then for each offence, we set the sum of ranks. In the end, we re-scale it to the range 0–1 using min-max normalization. According to this example, crime E is the worst and crime A is the lightest. Eventually, we assigned a rank and a value from the range [0.1] for each crime.

Table 1 Example of the scale based on the assigned ranks

Judge	Offences				
	A	B	C	D	E
1	5	3	4	2	1
2	3	4	2	5	1
3	4	5	3	2	1
4	5	3	4	1	2
5	5	2	3	4	1
Sum of ranks	22	17	16	14	6
Min-max scaling	1	0.69	0.63	0.5	0

Source: Own elaboration

The disadvantage of this approach is decreasing efficiency while increasing the number of evaluated objects. Moreover, we assume that the distances between objects, considered as the validity of one object over the second one, is equal. On the other hand, the great advantage of this approach is its simplicity and lack of assumptions.

1.3 Thurstone scaling (Case V)

Thurstone pair comparison model is considered a probabilistic choices model with the following assumptions:

1. Distribution of the hidden preferences in the preferences had a normal distribution.
2. Preferences are independent of each other, and they have one source of variance (the assumption that there is zero correlation might be softened to the assumption that there is a correlation between pairs).
3. The probability of the intransitive preferences is different from zero.
4. Measurements errors are non-correlated, and they have a normal distribution.

Thurstone (1927b) assumed and provided a rationale for ordering objects on a continuum. Although we may have more or less favourable reactions to a particular object, Thurstone suggested that there was a most frequent or typical reaction to any object. Because the normal curve is symmetrical, the most frequent reaction occupies the same scale position as the mean. Thus, the mean can also represent the scale value for the particular object. So, in his simplest Thurstone model (so-called Case V), he assumed that reactions to various stimuli were normally distributed. He also assumed that the variance of the reactions around each mean would be the same. The means of normal distribution of each object are interpreted as scale values.

In the method of pairs comparison: for n objects, we get $\frac{n(n-1)}{2}$ pairs. Let X_i (i = 1, 2, ..., n) be the characteristic of an object. We assume that $X_i \sim N(\mu_i, \sigma_i)$.

Parameter μ_i is an expected value of the ith object and is the main topic of interest in the current context, as we want to compare the relative positions of the objects, i.e. their central tendencies. Estimation μ_i, as an item on a scale, is based on observation of the difference $X_i - X_j$. Note that a random variable $Y = Y_{ij} = X_i - X_j$ has a normal distribution with the following density function:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} \exp\left(-\frac{(y-(\mu_i-\mu_j))^2}{\sqrt{2\pi\sigma_{ij}^2}}\right), \tag{1}$$

where: $\sigma_{ij} = \sqrt{\sigma_i^2 + \sigma_j^2 - 2\sigma_i\sigma_j\rho_{ij}}$.

Through the comparison of projects in pairs, (X_i, X_j)it is possible to determine the probability estimator:

$$p_{ij} = P(X_i - X_j > 0) = \Phi\left(\frac{\mu_i - \mu_j}{\sigma_{ij}}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(\mu_i-\mu_j)/\sigma_{ij}} e^{-y^2/2} dy\,. \tag{2}$$

Knowing p_{ij} it is possible to determine $z_{ij} = \Phi^{-1}(p_{ij})$. Then by using the least-squares approximation $(\sum_{i\neq j}^{m} (z_{ij} - (\mu_i - \mu_j))^2 \rightarrow \min)$ we determine $\mu_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij} = \frac{1}{n}\sum_{i=1}^{n} \Phi^{-1}(p_{ij})$.

After re-scaling μ_j to the range 0-1 using min-max normalization:

$$\mu_j' = \frac{\mu_j - \min_j(\mu_j)}{\max_j(\mu_j) - \min_j(\mu_j)}, \tag{3}$$

the average values μ create a Thurston preferences scale that is most commonly scale for range [0,1].

As it could be seen Thurstone scale is based on some particular assumptions. It seems legitimate to inquire whether it works better than the simpler scale that may be used and whether it is possible to construct equally reliable scales without making unnecessary statistical assumptions. To this aim in the next section simple rank scale will be introduced.

2 SURVEY RESEARCH – EXPERIMENT

2.1 Description survey, dataset and scales

The survey was constructed based on the original Thurstone study that measured social values, specifically the seriousness of different types of crimes or offences. The following types of crimes/offenses were considered in the survey:

P1. Violent rape.
P2. Assault with a severe body injury.
P3. Paedophile acts.
P4. Domestic violence.
P5. Threats.
P6. Murder.
P7. Defamation (slander).
P8. Harassment.
P9. Kidnapping for ransom.
P10. Identity theft.

We shall assume the seriousness of an offence to be the seriousness as judged rather than as measured in terms of objective consequences or in some normative way. The main aim of the study was to obtain data for comparative analysis. The intermediate aim was to perform some kind of pilot study as a preliminary step to learn about societal preferences regarding the strength of the crime. In Poland, public dissatisfaction resulting from inadequate punishment for crime is often heard in discussions. It could be useful to have some knowledge of societal judgments of sentences for given crimes. The pilot study could serve as a useful tool to project the actual survey, especially the final set of crimes to be included. Of course, the final survey would have to be carried out on a representative sample of the population.

The respondents were 219 students (individuals aged 19–23) in the conducted study. Students responded in two ways. The first way was that the offences were arranged in pairs so that they were paired with every other one. The total number of pairs of offences presented was $10(10 - 1)/2 = 45$. A student had to choose a more severe crime from each pair. This method excludes the draw situation. Hence, if a student considered crimes equally serious, they have to choose one of them as worse. The input matrix **P** (matrix observed proportion of times that object i was chosen over object j) obtained based on the data is presented in the form of Table 2, where e.g. $p_{21} = P(P2 \succ P1)$ means that 18% of respondents considered that the P2 offense (assault with a serious body injury) is more serious than the P1 offence (violent rape) and $p_{12} = P(P1 \succ P2)$ means that 82% of respondents answered the opposite, that the P1

Table 2 The input matrix and Thurstone scale for survey data

P	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P1	0	0.82	0.33	0.82	0.99	0.33	0.99	0.97	0.94	0.92
P2	0.18	0	0.31	0.64	1.00	0.13	0.99	0.96	0.86	0.89
P3	0.67	0.69	0	0.84	0.98	0.37	0.98	0.98	0.83	0.93
P4	0.18	0.36	0.16	0	0.96	0.11	0.95	0.92	0.50	0.80
P5	0.01	0.00	0.02	0.04	0	0.02	0.60	0.27	0.05	0.21
P6	0.67	0.87	0.63	0.89	0.98	0	1.00	0.97	0.95	0.95
P7	0.01	0.01	0.02	0.05	0.40	0.00	0	0.22	0.06	0.19
P8	0.03	0.04	0.02	0.08	0.73	0.03	0.78	0	0.11	0.42
P9	0.06	0.14	0.17	0.50	0.95	0.05	0.94	0.89	0	0.77
P10	0.08	0.11	0.07	0.20	0.79	0.05	0.81	0.58	0.23	0
					↓					
Thurstone scaling	*0.09*	*0.23*	*0.10*	*0.40*	*0.98*	*0*	*1*	*0.81*	*0.48*	*0.68*

Source: Own elaboration

crime is more severe than the P2 crime. At the end of Table 2 is the *Thurston scale* (developed as intended and the technique presented in Section 1.2).

Another way of scaling crimes was based on students' ranking of offences, from the worst to the lightest. In that case, the draw situation was not possible, either. The results are given in Table 3.

Table 3 Rankings for survey data

Rank	Offences									
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**
1	39	0	31	1	0	146	0	0	0	2
2	102	27	58	4	0	21	1	0	3	3
3	61	43	59	13	0	34	1	0	6	2
4	13	76	36	47	1	10	0	3	30	3
5	2	45	21	67	3	5	1	3	61	11
6	1	19	12	59	9	0	6	13	73	27
7	0	7	1	21	23	0	10	57	26	74
8	1	2	0	3	59	0	27	79	11	37
9	0	0	1	3	75	2	40	51	4	43
10	0	0	0	1	49	1	133	13	5	17
Sum of ranks	*502*	*891*	*661*	*1 133*	*1 871*	*383*	*2 022*	*1 725*	*1 243*	*1 614*
Rank	*0.1*	*0.3*	*0.2*	*0.4*	*0.9*	*0*	*1*	*0.7*	*0.5*	*0.6*
Rank scale Min-max scaling	*0.07*	*0.31*	*0.17*	*0.46*	*0.91*	*0*	*1*	*0.82*	*0.52*	*0.75*

Source: Own elaboration

The penultimate row of Table 3 contains a scale, or rather no scale, called a *rank*. It shows the situation where we rank the crimes and do not perform the scaling. In that situation, the distances, considered differences between subsequent offences' validity, are the same. The last row of Table 3 contains the ranking combined with the re-scaling the sum of ranks to the range 0–1 using min-max normalization, so-called *rank scale (*cf. Section 1.1).

Moreover, students were asked which way of crime assessment was easier and more comfortable for them: ranking or pair comparison. 67% of responders said it was easier to rank the crimes instead of pair comparison. However, 33% of them preferred to evaluate pairs. It can be assumed that the predominance of respondents preferring to rank crimes will increase with the increase in the number of crimes (or objects to compare).

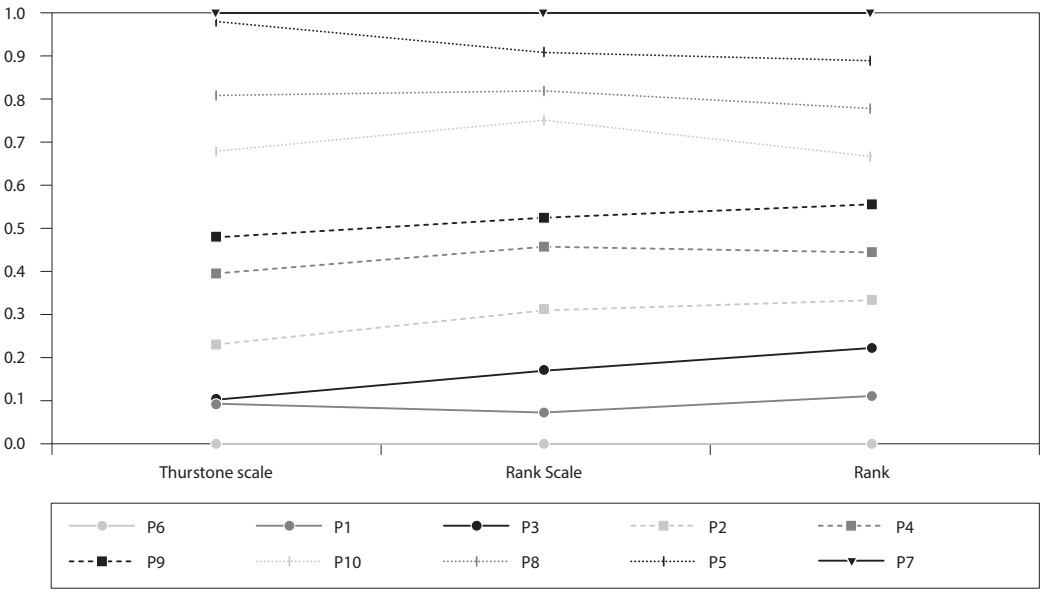
2.2 Comparative analysis of scales

First, let's compare the crimes' scale obtained by the ranking methods and the Thurstone scale. Hence, we have to ask ourselves: *How do the offences arrange themselves in a quantitative continuum from those that seem to be most serious to those that seem relatively least objectionable?* Figure 2 is a graphic illustration of the Thurston scale and rank scale, and the simplest ranking is also included for comparison.

As we can see in the graph in Figure 2, each method shows the same hierarchy of crimes. Here, the results are consistent. In contrast, the differences could be seen in the Thurstone and rank scale results. The greatest difference refers to the crimes: P1 and P3 and P5 and P7. According to the Thurstone scale, crimes in both pairs are equally important. While, on the contrary, the rank scale differs the importance

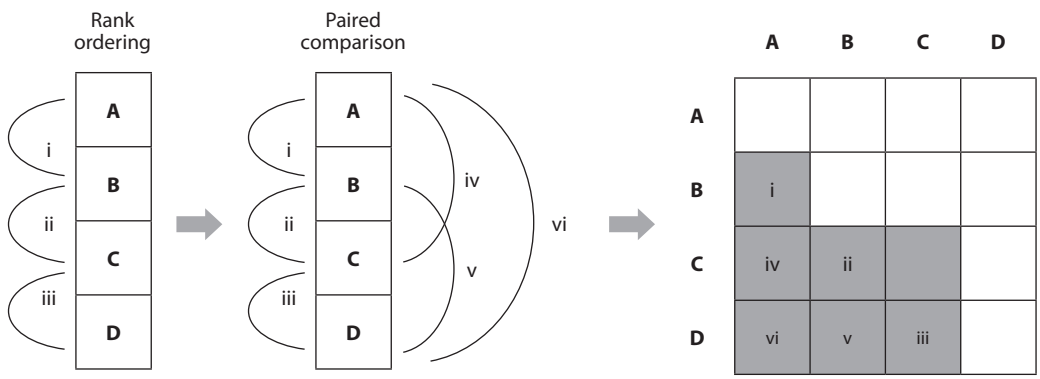
of crimes in these pairs. It means that the choice of method affects the final results. The question arises: *Are the differences statistically significant, and whether the way respondents are asked about the importance of crimes affects the scaling results?*

Figure 2 Comparison of rankings



Source: Own elaboration

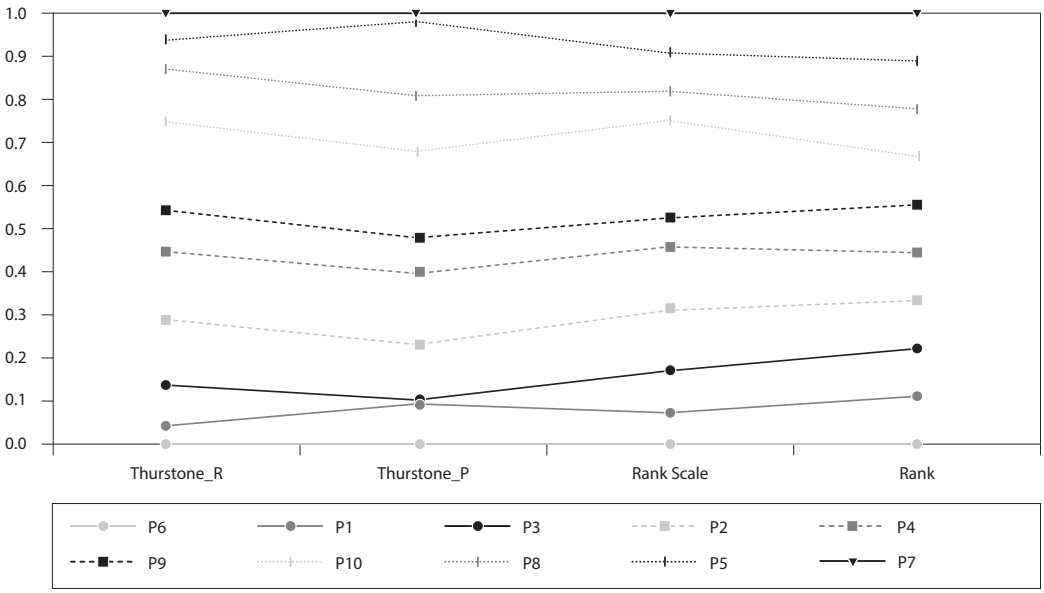
Figure 3 Schematic representation of deriving paired comparison data based on rank data



Source: Own elaboration

The input data for the Thurstone scale is the matrix P, where symmetric cells sum to unity. We could obtain matrix P from the rank order as well as from the paired comparison (see Figure 3). Because in the experiment, the respondents ranked crimes using both methods, so it is possible to compare the results of the Thurstone scale obtained from the rank ordering and paired comparison.

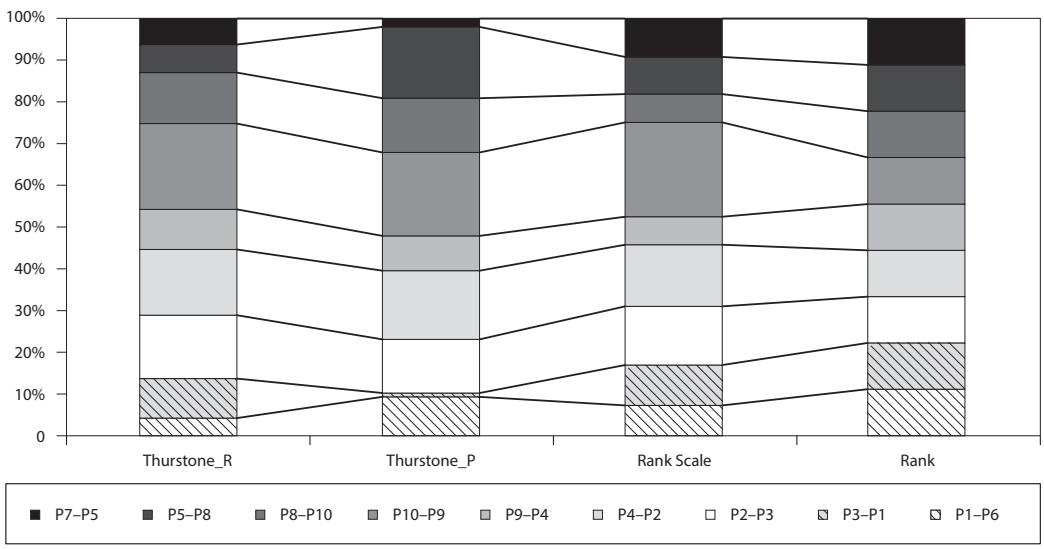
Figure 4 Schematic representation of deriving paired comparison data based on rank data



Source: Own elaboration

Figure 4 shows the influence of the used way of asking about preferences on the scale values. Thurstone_P denotes the Thurstone scale obtained by the paired comparison. Whereas Thurstone_R denotes the Thurstone scale obtained from the rank ordering. For comparison, the classification of crimes has

Figure 5 The structures of the differences between the values for scales



Source: Own elaboration

been included, without their scaling, i.e. rank. Again, the difference between scales could be seen. The scale Thurstone_R, similarly to the rank scale, do not state pairs of crimes: P1, P3 and P5, P7 as of equal importance. It means that the way of asking the respondents about the projects' hierarchy has an impact on the determined scale. However, whether the differences between the obtained results are statistically significant is still valid. We used the test for Similarity of Structures proposed by (Sokołowski, 1993) to answer this question. There are two test hypotheses:

H_0: The structures are dissimilar,

H_1: The structures are similar.

The test statistic is the structural similarity coefficient based on the Bray-Curtis distance, and for a given level of significance, we have a right-tailed rejection region. For each considered scale, the structure was made by the differences between the values on a scale (as in Figure 5).

Table 4 The test similarity of structures of the differences between the values for scales

Compared structures for scales	The value of the test statistic	Significance level	Critical value
Thurstone_P : Thurstone_R	0.83	0.01	0.7705
Thurstone_P : Rank Scale	0.90	0.05	0.7115
Thurstone_R : Rank Scale	0.80	0.10	0.6747
Thurstone_R : Rank	0.81	0.16	0.80
Thurstone_P : Rank	0.76	0.02	0.21
Rank Scale : Rank	0.82	0.63	0.95

Source: Own elaboration

For almost all pairs of the structures compared, the test rejected the null hypothesis, which means that each method showed a similar scale of crimes (cf. Table 4). The exception is comparing Thurstone_P and the rank on significance level equal 0.1. Whereas the test result, it turned out that the distance structures between successive distances of the Thurstone_P scale differ significantly from those obtained without any scaling (rank).

That result shows another direction for further research. Searching for conditions concerning matrix P leading to equality of the Thurstone scale and the rank scale. Moreover, the conclusion can be stated that the differences between obtained results for scaled orders (Thurston_P, Thurston_R and the rank scale) in the set of crimes and offences are not statistically significant.

2.3 The independence of alternatives

Finally, we consider *the problem of the independence on alternatives.* Independence on alternatives means that if P1 is preferred to P2 out of the choice set {P1, P2}, then introducing a third alternative P (thus expanding the choice set to {P1, P2, P}) should not make P2 preferred to P1. Hence, the independence of alternatives assumes that ordering a given pair of items does not depend on the other options available. Changes in individuals' rankings of irrelevant alternatives (ones outside a certain subset) should have no impact on the societal ranking of the subset.

As both the ranking scale and Thurstone method are special kinds of aggregation of preferences, it is known – as proven by Arrow's Impossibility Theorem – that dependence on irrelevant alternatives cannot be avoided in a general case. Still, we may follow the attempts of researchers who investigate the frequency of occurring the Condorcet paradox in the real data (although, again, this paradox cannot be avoided in principle), and to check the frequency of occurring the dependence on irrelevant

alternatives in scaling tasks. Not much effort has been devoted by now to this aim, so our investigation is a contribution and a trigger to such discussion. The very seminal paper of Thurstone (1929) has not discussed that topic, although investigating the data available in this paper one may note the existence of this undesirable property of the data.

We have checked the dependence on alternatives for our data for rank method. Namely, we have investigated the relative positions of all combinations of subsets (starting from 2 ending with 9 elements) to check if their relative positions may change depending on the other objects in the subset under consideration. We conclude, that there is not a single interchange of the relative position of any couple of objects.

Contrary, the graph in Figure 6 shows the counterexamples for Thurstone scale for the three examples of crimes' subsets consisting of:

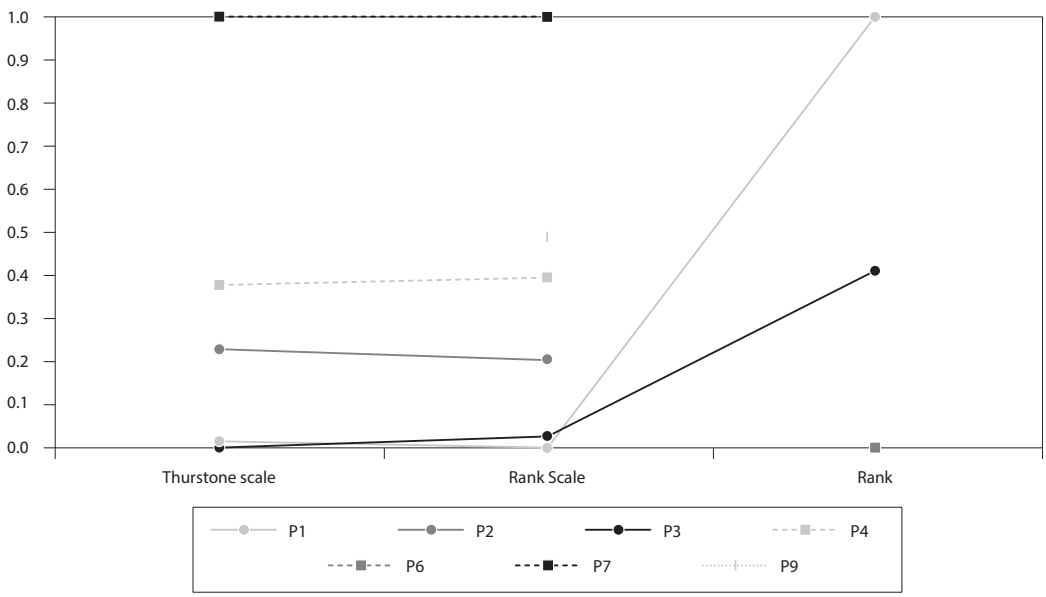
Case 1: {P1, P2, P3, P4, P7},

Case 2: {P1, P2, P3, P4, P7, P9},

Case 3: {P1, P3, P6}.

Note that the second case extends the first case by the crime P9. The third case consists of 2 questions from the previous two cases and additionally P6. The importance of crimes P1 and P2 is influenced by the other crimes considered in the comparison. The inclusion of the offence P9 made the offence P3 lighter than the crime P1.

Figure 6 Thurstone scales for the subgroups of crimes



Source: Own elaboration

We do not infer anything conclusive with this demonstration but rather pose a question. Does it seem sensible to reject those results of finding the aggregated scales in which the dependence on irrelevant alternatives is observed (or at least to reject those objects, which relative positions are sensitive for alternatives)? The lack of this effect of course does not ensure that adding yet another alternative that has not been included in the survey will not change the situation and one is not able to protect from this

possibility. Still, the lack of the undesirable effect – while it is not a proof – is at least some corroboration of stability of the results, given the method used to obtain them. Again relating to the Condorcet Paradox – some authors suggest that if the data reveals this undesirable property for one method of revealing the winner, one should switch into another method. Thus, in the case when the data reveal sensitivity for irrelevant alternatives, should one switch to a more robust (but more coarse) method or is it enough to reject the troublesome object? But what in the case when they are crucial for the task at hand and shouldn't be removed without the loss of the usefulness of the results?

While not daring to suggest the definite answer, we think, that investigating the existence and frequency of the occurring of undesirable properties of Thurstone method as compared with more simple ones is a valuable deepening the understanding of this scaling method.

CONCLUSIONS

Among various scaling techniques, the simplest one is probably the rank scaling. The simplicity may be however regarded as oversimplicity. On one hand, an individual is asked to rank objects in an ordinal way, but as a final result, by averaging ranks, we obtain an interval scale, which seems to be a kind of inconsistency. One may argue that this way of obtaining the scale is supported by the assumption of normality (or at least symmetry) of distribution of ranks – that is, the same fraction of respondents ranking object as the second would place it in the vicinity of the first one (rather than in the middle between the first and the third) as the fraction of individuals who would place it in the vicinity of the third object rather than in the equal distance from the first and the third. Thurstone scale is based on the explicit and detailed model of objects, which perception is distributed normal. The estimation of the expected value of each distribution is intermediated by the relative positions of all couples of two objects. Although both methods are prone to some disadvantages – in this paper we have examined the dependence on irrelevant alternatives – it seems that Thurstone method, as a more sophisticated one, is also more susceptible to such effects, at least in the particular case of our empirical study.

We have shown that both methods gave qualitatively similar results in the statistical sense. However, as for precise results, it is still not obvious which scale should be treated as appropriate. One question is the justifiability of the assumption underlying the Thurstone model. The other is the undesirable property of this particular results – dependence on the irrelevant alternatives. It seems that if we want to adopt Thurstone scale as a valid one, we should be conscious of the problem with "unstable" results (specifically, objects P1 and P3), and either remove them from the analysis or treat them as undistinguishable within the given method.

Our study does not propose any definite answer to the question of which method to be used but rather identifies the problem with the supposed unreliability of some results in the case of dependence on irrelevant alternatives – the problem that was bypassed in silence by both Thurstone and the followers.

References

BALL, H. L. (2019). Conducting Online Surveys [online]. *Journal of Human Lactation*, 35(3): 413–417. <https://doi.org/10.1177/0890334419848734>.

DRASGOW, F., CHERNYSHENKO, O. S., STARK, S. (2010). 75 Years After Likert: Thurstone Was Right! [online]. *Industrial and Organizational Psychology*, 3(4): 465–476. <https://doi.org/10.1111/J.1754-9434.2010.01273.X>.

EDWARDS, A. L. KENNEY, K. C. (1946). A comparison of the Thurstone and Likert techniques of attitude scale construction [online]. *Journal of Applied Psychology*, 30(1): 72–83. <https://doi.org/10.1037/H0062418>.

EDWARDS, A. L., KILPATRICK, F. P. (1948). Scale analysis and the measurement of social attitudes [online]. *Psychometrika*, 13(2): 99–114. <https://doi.org/10.1007/BF02289081>.

ESCHER, I. (2010). Pomiar kierunku i siły marketingowej postawy pracownika – kompromis pomiędzy teorią a praktyką marketingową [online]. *Acta Universitatis Nicolai Copernici Ekonomia*, 397(0): 159–174. <https://doi.org/10.12775/AUNC_ECON.2010.012>.

HOLBROOK, A., CHO, Y. I., JOHNSON, T. (2006). The impact of question and respondent characteristics on comprehension and mapping difficulties [online]. *Public Opinion Quarterly*, 70(4): 565–595. <https://doi.org/10.1093/poq/nfl027>.

Horizon Europe [online]. European Commission. [cit. 15.1.2022]. <https://ec.europa.eu/info/funding-tenders/find-funding/eu-funding-programmes/horizon-europe_en>.

LIPOVETSKY, S. (2007). Thurstone scaling in order statistics [online]. *Mathematical and Computer Modelling*, 45(7–8): 917–926. <https://doi.org/10.1016/J.MCM.2006.09.009>.

LIPOVETSKY, S., CONKLIN, W. M. (2004). Thurstone scaling via binary response regression [online]. *Statistical Methodology*, 1(1–2): 93–104. <https://doi.org/10.1016/J.STATMET.2004.04.001>.

MOSTELLER, F. (1951). Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed [online]. *Psychometrika*, 16(2): 207–218. <https://doi.org/10.1007/BF02289116>.

SAFFIR, M. A. (1937). A comparative study of scales constructed by three psychophysical methods [online]. *Psychometrika*, 2(3): 179–198. <https://doi.org/10.1007/BF02288395>.

SINGLETON, R. A., STRAITS, B. C. (2018). *Approaches to Social Science Research.* Oxford UP, p. 635.

SOKOŁOWSKI, A. (1993). Propozycja testu podobieństwa struktur. *Przegląd Statystyczny*, 40(3–4): 295–301.

STADTHAGEN-GONZÁLEZ, H. et al. (2018). Using two-alternative forced choice tasks and Thurstone's law of comparative judgments for code-switching research [online]. *Linguistic Approaches to Bilingualism*, 8(1): 67–97. <https://doi.org/10.1075/LAB.16030.STA/CITE/REFWORKS>.

THURSTONE, L. L. (1927a). Psychological Analysis [online]. *The American Journal of Psychoanalysis*, 38(3): 368–389. <https://doi.org/10.3917/cohe.174.0156>.

THURSTONE, L. L. (1927b). The method of paired comparisons for social values [online]. *Journal of Abnormal and Social Psychology*, 21(4): 384–400. <https://doi.org/10.1037/H0065439>.

THURSTONE, L. L., CHAVE, E. J. (1929). *The measurement of attitude.* Chicago: The University of Chicago Press.

THURSTONE, L. L., JONES, L. V. (1957). The Rational Origin for Measuring Subjective Values [online]. *Journal of the American Statistical Association*, 52(280): 458–471. <https://doi.org/10.1080/01621459.1957.10501401>.

TSOGO, L., MASSON, M. H., BARDOT, A. (2000). Multidimensional Scaling Methods for Many-Object Sets: a Review [online]. *Multivariate Behavioral Research*, 35(3): 307–319. <https://doi.org/10.22004/ag.econ.135504>.

TSUKIDA, K., GUPTA, M. R. (2011). *How to Analyze Paired Comparison Data.* UWEE Technical Report, (206): 18.

Missing Data Imputation for Categorical Variables

Jaroslav Horníček[1] | *Prague University of Economics and Business, Prague, Czech Republic*
Hana Řezanková[2] | *Prague University of Economics and Business, Prague, Czech Republic*

Received 20.1.2022, **Accepted** (reviewed) 9.2.2022, **Published** 16.9.2022

Abstract

Dealing with missing data is a crucial part of everyday data analysis. The IMIC algorithm is a missing data imputation method that can handle mixed numerical and categorical datasets. However, the categorical data are crucial for this work. This paper proposes the new improvement of the IMIC algorithm. The two proposed modifications consider the number of categories in each categorical variable. Based on this information, the factor, which modifies the original measure, is computed. The factor equation is inspired by the Eskin similarity measure that is known in the hierarchical clustering of categorical data. The results show that as the missing value ratio in the dataset grows, better results are achieved using the second modification. The paper also shortly analyzes the advantages and disadvantages of using the IMIC algorithm.

Keywords	***DOI***	***JEL code***
IMIC algorithm, missing value imputation, categorical variables	*https://doi.org/10.54694/stat.2022.3*	*C38, C40, C80*

INTRODUCTION

The missing value imputation problem can be frequently encountered in natural and social sciences and technology. NASA uses missing value imputation when reconstructing images sent from outer space because it is not technologically possible to transfer every image pixel without information loss. On the other hand, social scientists may use this method in a survey to compensate for the reluctance of the respondents to answer questions. The correct imputation of missing values in such cases is crucial and affects the quality of the final research. Statistical analysis methods often require the information inherent in data to be complete (no missing values). Otherwise, the methods fail.

Before the basic methods for working with missing values are introduced, the basic terminology will be mentioned. There are three natural mechanisms that can cause incomplete data to occur. Namely, it is MCAR, MAR, and MNAR, described by Rubin (1976), Rubin and Little (2002), or Baraldi and Enders (2010). MCAR (missing completely at random) occurs when the data are missing randomly without any observable pattern. MAR (missing at random) happens when the missing values of one variable are dependent on another variable, e.g. with decreasing attained level of respondent education, there are more

[1] Prague University of Economics and Business, Faculty of Informatics and Statistics, Department of Statistics and Probability, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. Corresponding author: e-mail: horj31@vse.cz.
[2] Prague University of Economics and Business, Faculty of Informatics and Statistics, Department of Statistics and Probability, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: hana.rezankova@vse.cz.

missing values in the variable for respondent income. Finally, MNAR (missing not at random) occurs when the observed values depend on each other, e.g., the respondents who have problems with alcohol might be less willing to answer in a survey on alcoholism (Petrúšek, 2015).

The simplest solution to dealing with missing values is simply removing the whole incomplete multidimensional observations. However, this reduces the number of observations and affects the randomness of the sample selection, which is mentioned by Azar (2002) or de Leeuw et al. (2003). In the case of pairwise statistical methods (e.g. correlation analysis), we do not have to remove every incomplete multidimensional observation. Pairwise methods can, in some cases, avoid the problem of deleting the incomplete observations from the sample, but it does not solve it.

One of the simplest methods of imputing the missing values is the replacement of the missing values with their mean. This approach may not change the mean value of the variable but can significantly affect the variability of the result. This method can safely be used only in cases where the mechanism of missing values is MCAR, as Baraldi and Enders (2010) explain.

A more advanced method of missing value imputation uses a regression function. However, even this method can significantly affect the variability of the sample. In practice, stochastic regression is often used. A random error term is added to the individual predicted values in such a case. This ensures that the imputed values do not strictly copy the given regression function and artificially create variability of the result, which is desirable in most cases, as Baraldi and Enders (2010) point out.

Apart from regression, there are many simple methods for imputing the missing values. These methods are based on simple linear models and other prediction algorithms of machine learning. Significant improvement was only achieved by introducing the multiple imputation method (Rubin, 1987) and the method based on the maximal likelihood estimation (Allison, 2012). However, even with these methods, we cannot avoid some inaccuracy in the estimation of the true values in case the missing data were created by the MNAR mechanism. Nevertheless, the estimation is demonstratively better than in the case of the simple methods, as Schafer and Graham (2002) say.

In their simplest form, the multiple imputation methods randomly select a subset of the original dataset and conduct a regression analysis on it. From each (stochastic) regression function obtained this way, missing values can be predicted. The final value is then a result of calculating the mean of these values. The procedure can be modified by selecting several subsequent stochastic regressions, where the correlation estimation and the mean from the previous step are used to calculate the new regression coefficients of each new regression, as Baraldi and Enders (2010) explain.

Methods based on the maximal likelihood are built on a complex mathematical background, which is out of the scope of this article. Both advanced methods (multiple imputation and maximal likelihood estimation) are currently recommended for handling in missing values. Unfortunately, they can mostly work for quantitative data only.

Regardless of the multiple imputation method for categorical data introduced by Akande et al. (2017), there is generally no missing value imputation method for categorical data, which would be demonstratively better than any other. However, several papers (Sulis and Porcu, 2008; Ferrari et al., 2011; Wu et al., 2012; Pecáková, 2014; Stavseth et al., 2019) study this topic. Worthy of mention is the method based on the rough sets theory. It is a hierarchical method, which looks for the nearest pairs within a set of observations with the help of a specially defined metric. If a pair of observations contains one missing and one non-missing value in a certain variable, the missing value is replaced by a non-missing value in this pair. An important concept based on the rough sets theory is a so-called extended tolerance relation, explained by Nguyen et al. (2013), which can measure the similarity between a complete and an incomplete observation. As a sufficient explanation, we can say that any two identical sets would be in the equivalence relation. If a value from a set is deleted and called missing, the equivalence relation would no longer exist, but a tolerance relation still exists.

Feng et al. (2011) introduced the IMIC algorithm. The IMIC is a missing data imputation method that can handle both categorical and numerical variables. This algorithm does not need a set of predictor variables without missing values for the prediction of missing values in another variable. Due to the hierarchical clustering, every single variable in the dataset can contain missing values, and the IMIC algorithm fills in all unknown values in one run of the iteration process.

The IMIC algorithm can easily handle missing values in multiple variables in an incomplete dataset. It can be easily used by an inexperienced user. These advantages make the algorithm very promising. On the other side, it is very time-consuming because the algorithm computes similarity measures between each pair of observations in hierarchical clustering. Hence efficient implementation is crucial.

This paper proposes the new improvement of the IMIC algorithm. The two proposed modifications consider the number of categories in each categorical variable. Based on this information, the factor, which modifies the original measure, is computed. The factor equation is inspired by the Eskin similarity measure (Eskin et al., 2002) that is known in the hierarchical clustering of categorical data (Šulc and Řezanková, 2014; Cibulková et al., 2021). The results show that as the missing value ratio in the dataset grows, better results are achieved using the modification.

1 THE MAIN PRINCIPLE OF THE IMIC ALGORITHM

The IMIC algorithm utilizes hierarchical clustering. At the beginning of the process, each observation X_i (the vector of the variable values) is an isolated cluster; $X_i \subset X$, $i \in \{1, 2, ..., n\}$, where n is the number of observations in the dataset X. The cluster is a name either for two or more observations joined together or for one single observation (one element cluster).

In the case of categorical variables only, the algorithm computes $ISMD_C$ (Incomplete Set Mixed Dissimilarity in Categorical attributes) between two clusters in the rth step of the algorithm as

$$ISMD_C(X_i^r, Y_j^r) = \frac{\left|\{s_k(X_i^r, Y_j^r) \mid s_k(X_i^r, Y_j^r) = \emptyset\}\right|}{\sqrt{\left|X_i^r\right| + \left|Y_j^r\right|} \cdot \left|\{s_k(X_i^r, Y_j^r) \mid s_k(X_i^r, Y_j^r) \neq \emptyset\}\right|}, \tag{1}$$

where X_i^r, Y_j^r are two different clusters $X_i^r \subset X$ and $Y_j^r \subset X$, $k \in \{1, 2, ..., q\}$, q is the number of categorical variables, the operator $|\cdot| : A \to N$ returns the number of elements in the set A and $|X_i^r|$ returns the number of observations in the cluster X_i^r, symbol $\emptyset$ represents empty set, and $s_k(X_i^r, Y_j^r)$ is defined as

$$s_k(X_i^r, Y_j^r) = \begin{cases} s_k(Y_j^r), & s_k(X_i^r) = * \wedge s_k(Y_j^r) \neq * \\ s_k(X_i^r), & s_k(X_i^r) \neq * \wedge s_k(Y_j^r) = * \\ s_k(X_i^r) \cap s_k(Y_j^r), & s_k(X_i^r) \neq \emptyset \wedge s_k(Y_j^r) \neq \emptyset \end{cases}, \tag{2}$$

where the symbol * represents a missing value, symbol $\cap$ represents the set intersection operator, $\wedge$ represents logical "and", and $s_k(X_j^r)$ is defined as

$$s_k(X_i^r, Y_j^r) = \begin{cases} v_{a_{kp}}, & (\exists x_i \in X_j^r(a_k(x_i) = v_{a_{kp}}) \wedge (\forall x_i \in X_j^r)(a_k(x_i) = v_{a_{kp}} \vee a_k(x_i) = *) \\ *, & (\forall x_i \in X_j^r)(a_k(x_i) = *) \\ \emptyset, & (\forall p)(\exists x_i \in X_j^r)(a_k(x_i) \neq v_{a_{kp}} \vee a_k(x_i) \neq *) \end{cases}, \tag{3}$$

where $\vee$ represents logical "or", and $v_{a_{kp}}$ is the value of the kth variable and the pth value of all c_k unique values of the kth variable ($p \in \{1, 2, ..., c_k\}$), $a_k(x_i)$ represents the value of the kth variable and the ith cluster. Based on the algorithm of Feng et al. (2011), the set of s_k for one cluster is denoted as CS feature.

In the first step ($r = 0$), every value of $s_k(X_j^r)$ is set equal $v_{a_{kp}}$ or * based on the true value in the kth variable and the ith cluster trivially (the cluster is one single observation in the case of $r = 0$). After that, the $ISMD_C(X_i^r, Y_j^r)$ is computed between every two clusters. In the case of categorical variables only, every pair X_i^r, Y_j^r with minimal $ISMD_C(X_i, Y_i)$ are added together, so these two clusters X_i^r, Y_j^r rare joined to the new cluster $X_{i,j}{}^r = X_f^{(r+1)}$, where $f \in \{1, 2, ..., q - t\}$, where t means number of cluster pairs added together in the rth step.

When the one step of clustering based $ISMD_C$ is finished, the algorithm tries to replace missing values in the cluster with one or more missing value * as

$$a_k(X_i^r) = \begin{cases} v_k, & v_k \in s_k(X_i^r) \wedge v_k \neq * \\ *, & v_k \in s_k(X_i^r) \wedge v_k = * \\ Mode_k(X_i^r), & s_k(X_i^r) = \emptyset \end{cases}, \tag{4}$$

After that, we set $r = r + 1$ and proceed new iteration until no missing values are present or $r = n$.

The $ISMD_C$ is increasing as the clusters are growing. The $ISMD_C$ in Formula (1) can be understood as a ratio of different values to same values in each variable of the new potential cluster. In other words, two clusters will be joined more likely if the values in compared variables are the same.

For a better understanding of the algorithm, there is a small example. Assume that the small set of binary data is given

$$U_0 = (\{X_1^0\}, \{X_2^0\}, \{X_3^0\}, \{X_4^0\}) = (\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}) = \begin{pmatrix} \{0 & 1 & 1 & 1 & 0\} \\ \{0 & * & 1 & 1 & 0\} \\ \{1 & 0 & 0 & * & 1\} \\ \{1 & 0 & 1 & 0 & 1\} \end{pmatrix},$$

where U_0 represents the initial set of multidimensional observations of the four observations (clusters) ($\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$). As can be seen, each of these clusters consists of the values of five variables. It is evident that the second observation in the second variable and the third observation in the fourth variable contain a missing value.

Firstly, the algorithm computes the set of s_k based on Formula (3). For the first cluster $X_1{}^0$, the CS feature will be equal $CS(x_1) = \{\{0\},\{1\},\{1\},\{1\},\{0\}\}$, for the second cluster $CS(x_2) = \{\{0\},\{*\},\{1\},\{1\},\{0\}\}$, for the third cluster $CS(x_3) = \{\{1\},\{0\},\{0\},\{*\},\{1\}\}$, and for the fourth cluster $CS(x_4) = \{\{1\},\{0\},\{1\},\{0\},\{1\}\}$.

After that, the algorithm can recompute CS feature (2) and $ISMD_C$ for each pair of clusters according to the Formula (1). Therefore, the CS feature and $ISMD_C$ are the following (the symbol ∪ denotes the pair of clusters):

$CS(x_1 \cup x_2) = \{\{0\},\{1\},\{1\},\{1\},\{0\}\}$,
$CS(x_1 \cup x_3) = \{\{\emptyset\},\{\emptyset\},\{\emptyset\},\{1\},\{\emptyset\}\}$,
$CS(x_1 \cup x_4) = \{\{\emptyset\},\{\emptyset\},\{1\},\{\emptyset\},\{\emptyset\}\}$,
$CS(x_2 \cup x_3) = \{\{\emptyset\},\{0\},\{\emptyset\},\{1\},\{\emptyset\}\}$,
$CS(x_2 \cup x_4) = \{\{\emptyset\},\{0\},\{1\},\{\emptyset\},\{\emptyset\}\}$,
$CS(x_3 \cup x_4) = \{\{1\},\{0\},\{\emptyset\},\{0\},\{1\}\}$,

$ISMD_C(x_1 \cup x_2) = 0$,

$$ISMD_C(x_1 \cup x_4) = \frac{4}{\sqrt{2}},$$
$$ISMD_C(x_2 \cup x_3) = \frac{3}{2\sqrt{2}},$$
$$ISMD_C(x_2 \cup x_4) = \frac{3}{2\sqrt{2}},$$
$$ISMD_C(x_3 \cup x_4) = \frac{3}{4\sqrt{2}}.$$

The minimum of $ISMD_C$ for each pair of clusters is $ISMD_C(x_1 \cup x_2)$, which is equal to zero. Following Formula (4), the missing value $a_2(x_2)$= * can be replaced as $a_2(x_2) = 1$. After this replacement

$$\{X_1^1, X_2^1, X_3^1\} = \{y_1, y_2, y_3\} = \begin{pmatrix} \{0 & 1 & 1 & 1 & 0\} \\ \{0 & 1 & 1 & 1 & 0\} \\ \{1 & 0 & 0 & * & 1\} \\ \{1 & 0 & 1 & 0 & 1\} \end{pmatrix},$$

where $X_1^1 = X_1^0 \cup X_2^0 = y_1$. The CS feature for y_1 is equal to $CS(y_1)$={{0},{1},{1},{1},{0}}.

The CS feature for clusters y_2 and y_3 remains the same as in the first step, specifically $CS(y_2)$ = {{1},{0},{0},{*},{1}}, and $CS(y_3)$ = {{1},{0},{1},{0},{1}}.

Based Formula (2):

$$CS(y_1 \cup y_2) = \{\{\emptyset\},\{\emptyset\},\{\emptyset\},\{1\},\{\emptyset\}\},$$
$$CS(y_1 \cup y_3) = \{\{\emptyset\},\{\emptyset\},\{1\},\{\emptyset\},\{\emptyset\}\},$$
$$CS(y_2 \cup y_3) = \{\{1\},\{0\},\{\emptyset\},\{0\},\{1\}\},$$

and $ISMD_C$ then

$$ISMD_C(y_1 \cup y_2) = \frac{4}{\sqrt{3}},$$
$$ISMD_C(y_1 \cup y_3) = \frac{4}{\sqrt{3}},$$
$$ISMD_C(y_2 \cup y_3) = \frac{1}{4\sqrt{2}}.$$

For this step, the minimum of $ISMD_C$ for each pair of clusters is $ISMD_C(y_2 \cup y_3)$. Following Formula (4), the missing value $a_4(y_2)$ = * can be replaced as $a_4(y_2) = 0$. After this replacement, the algorithm will be stopped, because no missing value remains. The final imputed dataset is equal to

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Feng et al. (2011) introduced the theoretical principles of the IMIC method but did not include an objective summary of its advantages and disadvantages. One of the advantages of this method is that the algorithm is easy to use. There is no parameter that needs to be set up, so no additional knowledge or experience with statistical modeling is needed. Another advantage of the method is that it can be used on a dataset with mixed categorical and numerical values.

However, the advantages mentioned above can also be seen as disadvantages. There are not enough possibilities to improve the accuracy of the result. The main problem of the IMIC method is that it is time-consuming. The time complexity is $O(n^3)$ (Murtagh, 1983), where n is the number of data points.

2 THEORETICAL PRINCIPLES OF THE PROPOSED MODIFICATIONS

Formula (1) does not consider the actual number of different categories. Based on the results of Šulc and Řezanková (2014) it is reasonable to try modifying it like this

$$ISMD_C(X_i^r, Y_j^r) = \frac{\sum_k \frac{n_k^2}{n_k^2+2}\, h(s_k(X_i^r, Y_j^r))}{\sqrt{|X_i^r| + |Y_j^r|} \cdot |\{s_k(X_i^r, Y_j^r) \mid s_k(X_i^r, Y_j^r) \neq \emptyset\}|}, \tag{5}$$

where:

$$h(s_k(X_i^r, Y_j^r)) = \begin{cases} 1, & s_k(X_i^r, Y_j^r) = \emptyset \\ 0, & s_k(X_i^r, Y_j^r) \neq \emptyset \end{cases}, \tag{6}$$

where the factor $\frac{n_k^2}{n_k^2+2}$ is inspired by the Eskin similarity measure proposed by Eskin et al. (2002), where n_k is the number of categories in the kth variable. If the original algorithm encounters different categories in the cluster in the kth variable, it increases the numerator by one. The possible advantage of our modification is that the $ISMD_C$ can consider the actual number of different categories in the cluster.

The possible problem with this approach is that the situation when the categories are different can occur rarely. In such case, the impact of this improvement can be hardly detected. Given this fact, the equation can be changed as follow:

$$ISMD_C(X_i^r, Y_j^r) = \frac{\sum_k \frac{n_k^2}{n_k^2+2}\, h(s_k(X_i^r, Y_j^r)) + 1)}{\sqrt{|X_i^r| + |Y_j^r|} \cdot (|\{s_k(X_i^r, Y_j^r) \mid s_k(X_i^r, Y_j^r) \neq \emptyset\}| + 1)}. \tag{7}$$

The difference from previous improvement, the numerator is increasing for each kth variable by factor $\frac{n_k^2}{n_k^2+2}$ and multiply by the number of categories plus one. Adding one to the $h(s_k(X_i^r, Y_j^r))$ ensures that the numerator is always non-zero. Therefore, the factor $\frac{n_k^2}{n_k^2+2}$ is not negligible even if the $h(s_k(X_i^r, Y_j^r))$ equals zero.

3 DATA SOURCE AND APPLICATIONS OF THE PROPOSED MODIFICATIONS

For our experiment, we use the data about students during the 2005–2006 school year, collected by Cortez and Silva (2008). We chose the subset of 395 observations (students who attended the mathematical class) with the following 17 categorical (binary or nominal) variables (of 33 variables overall):

- Sex – student's sex (binary: female or male),

- School – student's school (binary: Gabriel Pereira or Mousinho da Silveira),
- Address – student's home address type (binary: urban or rural),
- Pstatus – parent's cohabitation status (binary: living together or apart),
- Mjob – mother's job (nominal, teacher, health care related, civil services (e.g. administrative or police), at home or other),
- Fjob – father's job (nominal, teacher, health care related, civil services (e.g. administrative or police), at home or other),
- Guardian – student's guardian (nominal: mother, father or other),
- Famsize – family size (binary: less than or equal to 3 or greater than 3),
- Reason – reason to choose this school (nominal: close to home, school reputation, course preference or other),
- Schoolsup – extra educational school support (binary: yes or no),
- Famsup – family educational support (binary: yes or no),
- Activities – extra-curricular activities (binary: yes or no),
- Paidclass – extra paid classes (binary: yes or no),
- Internet – Internet access at home (binary: yes or no),
- Nursery – attended nursery school (binary: yes or no),
- Higher – wants to take higher education (binary: yes or no),
- Romantic – with a romantic relationship (binary: yes or no),
- PassXfail – created variable based on true student's score which shows if students pass the exam or not (binary: yes or no).

The initial dataset is complete without any missing data. In our experiment, the missing values were created in each of the 17 variables separately in five different ratios (5%, 15%, 25%, 35%, and 45%). In each of these configurations, the missing values were created randomly (as MCAR). It is also possible to create the missing values in the whole dataset at once, but there should be no difference due to twenty replications of the experiment. The three methods of the missing imputation were used – the original IMIC algorithm, Modification 1 based on Formula (5), and Modification 2 based on Formula (7). These methods were implemented in the R environment (R Core Team, 2020) and the package RCPP (Eddelbuettel and François, 2011; Eddelbuettel, 2013; Eddelbuettel and Balamuta, 2018) was used in crucial parts for better performance.

The algorithm was executed twenty times for each of the five ratios and three versions of the IMIC algorithm for better result stability. In each of these twenty steps, the missing values were generated independently. Therefore, there were 300 runs of the algorithm in summary. After that, the results were averaged based on the specific missing values ratio and the algorithm version.

This setting allows comparing the accuracy of the algorithm based on the specific method and the missing value ratio. For binary variable (with values "0" and "1"), the accuracy can be defined as

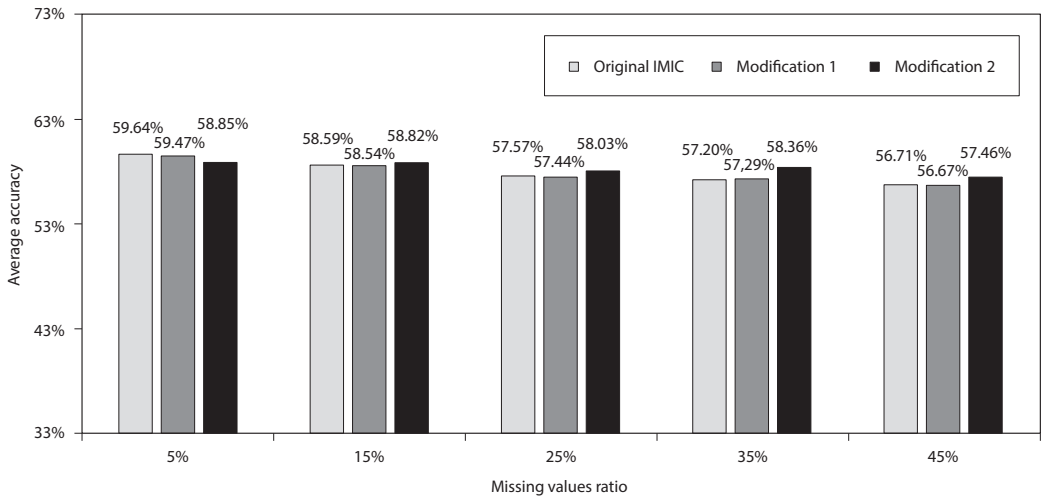
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{8}$$

where *TP* stands for true positive (the missing value is imputed by "1" correctly), *TN* for true negative (the missing value is imputed by "0" correctly), *FP* for false positive (the missing value is imputed by "1" falsely), and *FN* for false negative (the missing values is imputed by "0" falsely). This formula is defined for binary classification only, but the mean accuracy can be obtained in multiple classification cases. In this paper, the final overall mean accuracy is computed as mean accuracy over all variables and all twenty repetitions.

4 EXPERIMENTAL RESULTS

This section is focused on the simulation evaluations obtained on the dataset collected by Cortez and Silva (2008). The results in Figure 1 show that the Modification 1 works as well as the original IMIC. The Modification 2 works worse when the missing ratio is low, but the mean accuracy improves as the ratio of the missing values grows. The difference in overall average accuracy among these methods is not that large in absolute value, but as presented below, the pairwise t-test shows that the Modification 2 works significantly better than the original IMIC on the dataset used.

Figure 1 Average accuracy for different ratios of missing values (dataset with all categorical variables)



Source: Data collected by Cortez and Silva (2008), own calculation

As illustrated in Figure 1, when the ratio of the missing values hits 15%, the Modification 2 starts to be slightly better than the other two implementations. Moreover, if the vector of twenty accuracies of original IMIC from each of twenty replications (average over all used variables) is compared to the same vector evaluated using Modification 2, the difference is noticeable. For measuring this difference, the one-sided pairwise t-test was used, see Table 1. When the ratio of the missing values is equal to 35%, the Modification 2 becomes significantly better than the original IMIC algorithm on the dataset used.

Table 1 Comparison of the Modification 2 with the original IMIC algorithm (percent of the missing values and p-values for the t-test)

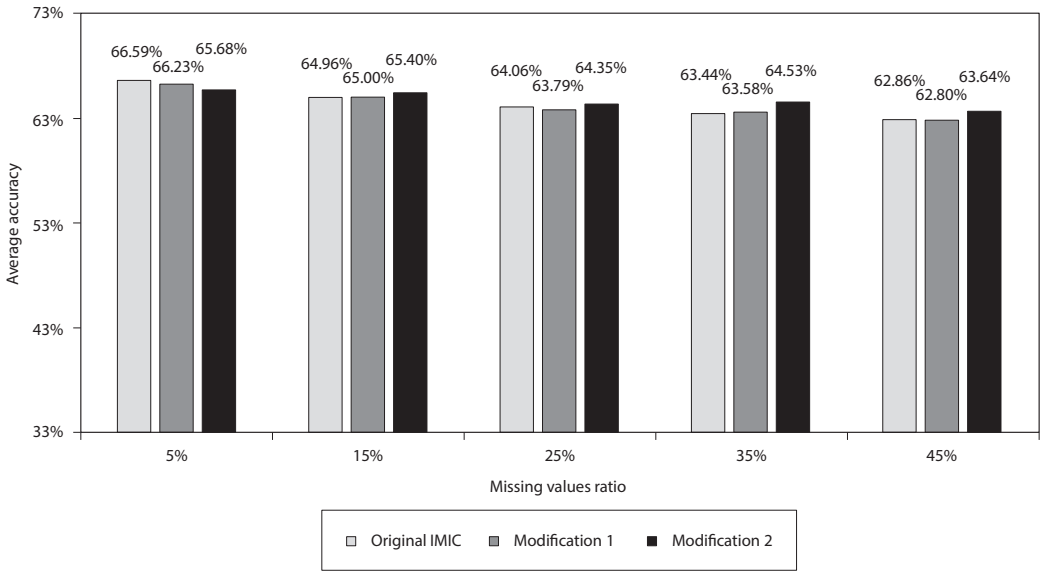
Share of missing values	P-value
5%	0.859
15%	0.266
25%	0.069
35%	< 0.001
45%	0.001

Source: Data collected by Cortez and Silva (2008), own calculation

The dataset can be investigated more deeply. When the variables are split into binary and nominal subsets, the accuracy for binary variables is about 65% (Figure 2) despite the 35% accuracy for nominal variables (Figure 3). The Modification 2 scores better in both cases regardless of the absolute values.

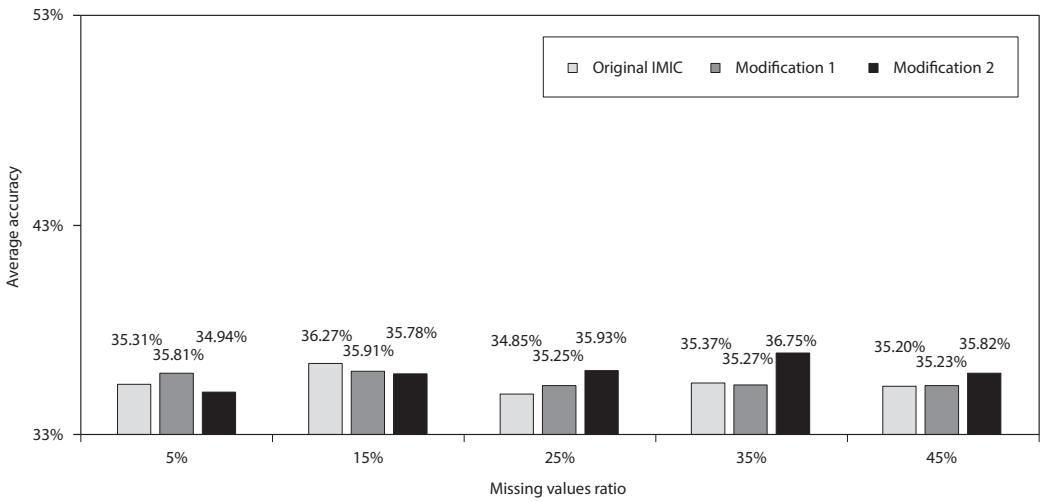
Figure 5 illustrates that the Modification 2 scores better in almost every twenty replications with a 35% missing values ratio compared to Figure 4, which illustrates the same situation with a 5% missing values

Figure 2 Average accuracy for different ratios of missing values (dataset with binary variables)



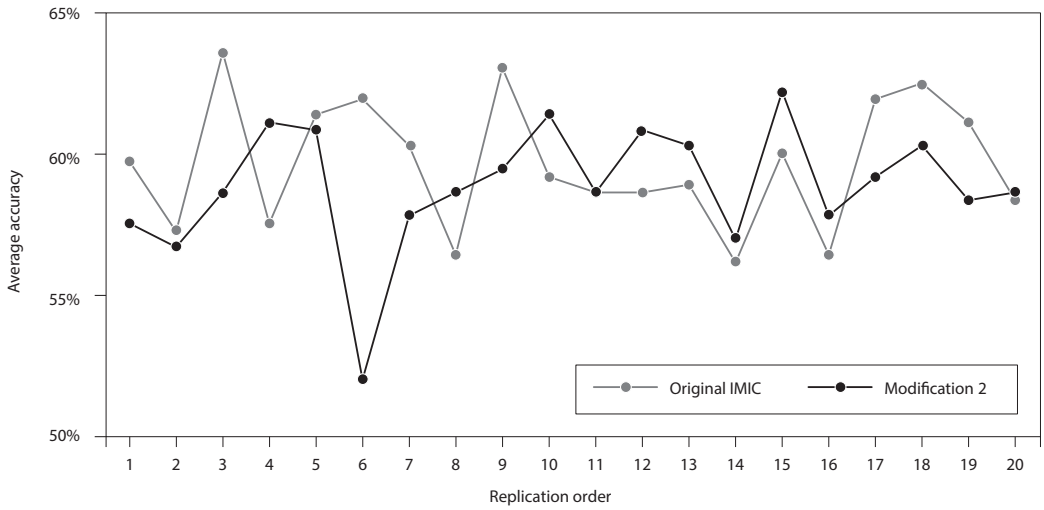
Source: Data collected by Cortez and Silva (2008), own calculation

Figure 3 Average accuracy for different ratios of missing values (dataset with nominal variables)



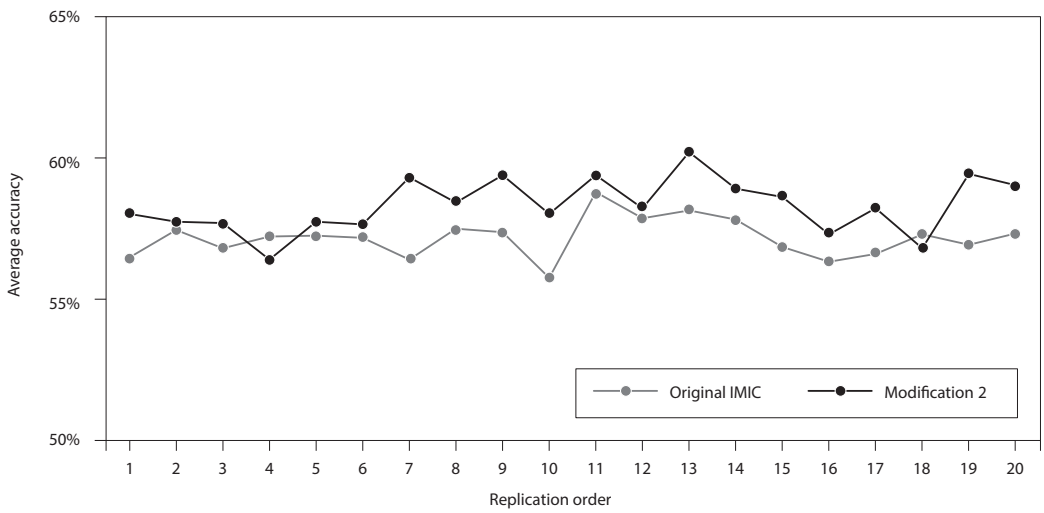
Source: Data collected by Cortez and Silva (2008), own calculation

Figure 4 Average accuracy over variables for 20 replications in 5% missing values ratio setting (dataset with all categorical variables)



Source: Data collected by Cortez and Silva (2008), own calculation

Figure 5 Average accuracy over variables for 20 replications in 35% missing values ratio setting (dataset with all categorical variables)



Source: Data collected by Cortez and Silva (2008), own calculation

ratio. It seems that, in the 35% setting, the algorithm is more stable. The coefficient of variation, which is defined as the standard deviation divided by mean, is lower in the 35% missing values ratio setting; concretely, the coefficient of variation for the Modification 2 equals about 0.0161 compared to the 5% missing values ratio setting where the coefficient of variation equals 0.0376.

CONCLUSION

For the purpose of this work, the IMIC algorithm was implemented. This IMIC is easy to use and does not require any additional assumptions on the dataset's properties. It can deal with categorical as well as numerical variables. The main disadvantage lies in time complexity, which is a problem of hierarchical clustering methods in general. Unluckily, this problem makes the simulations very CPU time demanding.

In this paper, two modifications of the IMIC algorithm were proposed and studied on the dataset collected by Cortez and Silva (2008). The first modification, which counts different categories in mismatched observations, was less successful than the second, which considers the overall frequency of categories in each categorical variable in the whole dataset. The differences in accuracy were not too large in absolute values, but the Modification 2 works stably better based on the one-sided pairwise t-test results. These results show the notable difference between accuracy for binary and nominal variables. However, the second modification works better in both cases.

Thanks to the full implementation of the IMIC algorithm, there are many ways for future research. Based on metrics known from hierarchical clustering, the IMIC algorithm can be modified in many different ways. The algorithm, unlike many others, considers the imputed variable itself. Due to this property, it can have some advantages when dealing with MNAR type of missing data. It could be examined in future work. Last but not least, the algorithm should be rewritten for its better efficiency.

ACKNOWLEDGMENTS

This work was supported by the Prague University of Economics and Business under the IGA project No. F4/22/2021.

References

AKANDE, O., LI, F., REITER, J. (2017). An Empirical Comparison of Multiple Imputation Methods for Categorical Data [online]. *American Statistician*, 71(2): 162–170. <https://doi.org/10.1080/00031305.2016.1277158>.

ALLISON, P. D. (2012). Handling Missing Data by Maximum Likelihood [online]. In: *SAS Global Forum*, paper 312. <https://statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf>.

AZAR, B. (2002). Finding a Solution for Missing Data [online]. *Monitor on Psychology*, 33(7). <http://www.apa.org/monitor/julaug02/missingdata>.

BARALDI, A. N., ENDERS, C. K. (2010). An Introduction to Modern Missing Data Analyses [online]. *Journal of School Psychology*, 48(1): 5–37. <https://doi.org/10.1016/j.jsp.2009.10.001>.

CIBULKOVÁ, J., NOVÁKOVÁ, L., HORNÍČEK, J. (2021). Imputation Methods for Missing Categorical Data in Cluster Analysis [online]. In: *The 15th International Days of Statistics and Economics*, Prague: Prague University of Economics and Business, 378–388. <https://msed.vse.cz/msed_2021/article/471-Cibulkova-Jana-paper.pdf>.

CORTEZ, P., SILVA, A. M. G. (2008). Using Data Mining to Predict Secondary School Student Performance. In: BRITO, A., TEIXEIRA, J. (eds.) *Proceedings of the 5th Annual Future Business Technology Conference*, Porto, 5–12.

EDDELBUETTEL D. (2013). *Seamless R and C++ Integration with Rcpp* [online]. New York: Springer. <https://doi.org/10.1007/978-1-4614-6868-4>.

EDDELBUETTEL D., BALAMUTA J. (2018). Extending extitR with extitC++: a Brief Introduction to extitRcpp [online]. *The American Statistician*, 72(1): 28–36. <https://doi.org/10.1080/00031305.2017.1375990>.

EDDELBUETTEL D., FRANÇOIS R. (2011). Rcpp: Seamless R and C++ Integration [online]. *Journal of Statistical Software*, 40(8): 1–18. <https://doi.org/10.18637/jss.v040.i08>.

DE LEEUW, E. D., HOX, J., HUSMAN, M. (2003). Prevention and Treatment of Item Nonresponse. *Journal of Official Statistics*, 19(2): 153–176.

ESKIN, E., ARNOLD, A., PRERAU, M., PORTNOY, L., STOLFO, S. V. (2002). A Geometric Framework for Unsupervised Anomaly Detection. In: *Applications of Data Mining in Computer Security*, Boston: Springer, 78–100.

FENG, X., WU, S., LIU, Y. (2011). Imputing Missing Values for Mixed Numeric and Categorical Attributes Based on Incomplete Data Hierarchical Clustering. In: *International Conference on Knowledge Science, Engineering and Management*, Berlin, Heidelberg: Springer, 414–424.

FERRARI, P. A., ANNONI, P., BARBIERO, A., MANZI, G. (2011). An Imputation Method for Categorical Variables with Application to Nonlinear Principal Component Analysis. *Computational Statistics & Data Analysis*, 55(7): 2410–2420.

MURTAGH, F. (1983). A Survey of Recent Advances in Hierarchical Clustering Algorithms [online]. *The Computer Journal*, 26(4): 354–359. <https://doi.org/10.1093/comjnl/26.4.354>

NGUYEN, D. V., YAMADA, K., UNEHARA, M. (2013). Extended Tolerance Relation to Define a New Rough Set Model in Incomplete Information Systems [online]. *Advances in Fuzzy Systems.* <https://doi.org/10.1155/2013/372091>.

PECÁKOVÁ, I. (2014). Problém chybějících dat v dotazníkových šetřeních [online]. *Acta Oeconomica Pragensia*, 22(6): 66–78. <http://www.vse.cz/aop/459>.

PETRÚŠEK, I. (2015). *Analýza chybějících hodnot.* Prague: Sociologický ústav AV ČR.

R CORE TEAM (2020). *A Language and Environment for Statistical Computing* [online]. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>.

RUBIN, D. B. (1976). Inference and Missing Data [online]. *Biometrika*, 63(3): 581–592. <https://doi.org/10.1093/biomet/63.3.581>.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Hoboken, New Jersey: John Wiley & Sons.

RUBIN, D. B., LITTLE, R. J. A. (2002). *Statistical Analysis with Missing Data.* 2nd Ed. Hoboken, New Jersey: John Wiley & Sons, Inc., Wiley Series in Probability and Statistics.

SCHAFER, J. L., GRAHAM, J. W. (2002). Missing Data: Our View of the State of the Art [online]. *Psychological Methods*, 7(2): 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>.

SULIS, I., PORCU, M. (2008). *Assessing the Effectiveness of a Stochastic Regression Imputation Method for Ordered Categorical Data.* Working Paper, Centro Ricerche Econoiche Nord Sud.

STAVSETH, M. R., CLAUSEN, T., RØISLIEN, J. (2019). How Handling Missing Data May Impact Conclusions: a Comparison of Six Different Imputation Methods for Categorical Questionnaire Data [online]. *SAGE Open Medicine.* <https://doi.org/10.1177/2050312118822912>.

ŠULC, Z., ŘEZANKOVÁ, H. (2014). Evaluation of recent similarity measures for categorical data [online]. In: *Proceedings of the 17th International Conference Applications of Mathematics and Statistics in Economics*, Wroclaw: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, 249–258. <https://doi.org/10.15611/amse.2014.17.27>.

WU, S., FENG, X., HAN, Y. et al. (2012). Missing Categorical Data Imputation Approach Based on Similarity. In: *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2827–2832.

Modelling Marital Reverse Annuity Contract in a Stochastic Economic Environment

Joanna Dębicka | *Wroclaw University of Economics and Business, Wrocław, Poland*
Stanisław Heilpern | *Wroclaw University of Economics and Business, Wroclaw, Poland*
Agnieszka Marciniuk[1] | *Wroclaw University of Economics and Business, Wroclaw, Poland*

Received 15.1.2022 (revision received 1.3.2022), **Accepted** (reviewed) 3.3.2022, **Published** 16.9.2022

Abstract

In the paper, we present the methodology of calculating the benefit of a marriage reverse annuity using the multiple state model for marriage life insurance. We model the probabilistic structure and cash flows arising from marriage reverse annuity contracts in the case of the joint-life status and the last surviving status assuming that the spouses' future lifetimes are independent. Usually, it is assumed that the interest rate is constant and the same through the years. It is not a realistic assumption. Therefore, this article's purpose is to calculate benefits under the assumption that the interest rate is a stochastic process or a fuzzy number model of the constant interest rate. We conduct a comparative analysis of the amount of benefit (taking into account the different frequency of their payment) for the different models of interest rates.

Keywords	***DOI***	***JEL code***
Stochastic interest rate, fuzzy interest rate, reverse annuity contract, the joint-life status, the last surviving status	*https://doi.org/10.54694/stat.2022.2*	*C69, E47, G17, G22*

INTRODUCTION

This article aims to apply multiple state models for marriage insurances to model the marriage reverse annuity contract. A reverse annuity contract is one type of equity release contract. It is a sales model. The real estate owner receives a monthly whole life annuity in exchange for selling the real estate to a company (usually a mortgage fund) interested in buying it. The beneficiary has the right to live in the property until his death. In Poland, the reverse annuity contract has only been sold in individual form since 2005. This paper aims to analyze the benefits of a marriage reverse annuity contract in the case of last survivor status and joint living status.

Usually, in the actuarial literature, the interest rate is assumed constant and the same through the years. It is not a realistic assumption. Therefore, this article's purpose is to calculate reverse annuity

[1] Faculty of Economics and Finance, Department of Statistics, Wroclaw University of Economics and Business, Komandorska 118/120, 53-345 Wrocław, Poland. Corresponding author: e-mail: agnieszka.marciniuk@ue.wroc.pl. ORCID-ID 0000-0002-9039-196X

benefits under the assumption that the interest rate is a stochastic process or a fuzzy number model of the constant interest rate. Many different interest rate models exist. The purpose of the article is to calculate net benefits (net cash flows). Hence the technical interest rate is applied. The expected future net present value benefits determined under the equivalence principle must cover the expected present value contributions at the time of entering into the contract. Therefore, inflation is not taken into account in this case. However, with rising inflation, the technical interest rate must be estimated appropriately since a reverse annuity contract is long-term.

We distinguish two ways of interest rate modelling, i.e., actuarial (the technical interest rate) and financial (the short-term rate). The following models of the stochastic interest rate are considered: the first-order autoregression, the Wiener process, the Vasicek and Cox-Ingersoll-Ross model. The third possibility considers the constant interest rate modelled by a fuzzy number. The fuzzy rate has not yet been applied to the determination of reverse annuity benefits. Therefore, the research hypothesis is that interest rates and their models (types) have a significant impact on the mortgage annuity benefit. In addition, the benefit is influenced by the frequency of payments. We show this effect.

The first section presents a literature review. Section 2 focuses on a discrete-time model, where the annuity is paid at the beginning of particular time units. We assume that the time-nonhomogeneous Markov chain describes the evolution of the contracted risk. Moreover, actuarial values are considered under the assumption of different types of interest rates. We propose to employ a matrix notation that makes calculations easier and provides a compact form for reverse annuity benefit formulas. This section is also dedicated to the characteristics of the interest rates models. Section 3 presents numerical examples and conclusions. The introduced matrix notation efficiently analyzes the influence of interest rate on the annuity installment. The numerical results are based on simulated data and the Polish Life Table, assuming that the spouses' future lifetimes are independent random variables. Section 4 presents discussion.

1 LITERATURE SURVEY

Multiple state models (MSM) have a wide application for describing a different kind of problems in finance and insurance, in particular analyzing cash flows arising from different kind of contracts. There is a vast literature on theoretical aspects and applications of MSM. Of particular note are the monographs e.g. (Cook and Lawless, 2018; Haberman and Pitacco, 2018; Hougaard, 2000; Huzurbazar, 2019) which show the potential of using this type of modelling. The general methodology of modern life insurance mathematics in the framework of a MSM can be found among others in (Bowers et al., 1986; Dębicka, 2012; Dickson et al., 2019; Norberg, 2002; Spierdijk and Koning, 2011). In particular MSM was used for analysis mortgage and reverse loan contracts e.g. (Dębicka et al., 2020; Dębicka and Marciniuk, 2014; Marciniuk, 2017; Marciniuk et al., 2020a; Zmyślona and Marciniuk, 2020).

The financial institutions in different countries propose the so-called equity release products for the retired (Hanewald et al., 2016), which provide an additional income to surrender their real estate. Various such products are available in the biggest world in the United States of America and Australia. The United Kingdom market is the largest European market for equity release contracts (Shao et al., 2015). Still, these contracts exist also in many other European countries (e.g. Spain, Ireland, France, Germany, Italy and Poland). There are two main types of equity release products the loan model and the sale model. According to (Dębicka and Marciniuk, 2014; Marciniuk, 2017; Marciniuk et al., 2020a), both products are available to individuals in Poland, i.e. a reverse mortgage (the loan model) and reverse annuity contract (the sale model). (Dębicka et al., 2020), (Dębicka and Marciniuk, 2014), (Marciniuk, 2017), (Marciniuk et al., 2020) and (Zmyślona and Marciniuk, 2020) consider also contracts for marriage. (Dębicka et al., 2020) and (Marciniuk, 2016) distinguished the dependence between the future lifetimes of spouses. In (Marciniuk, 2017) the reverse annuity is applied to derive two lemmas used to determine marriage benefits payable more than once a year inter alia in the case of Last Surviving Status and Joint Life Status. In the above papers,

constant or depending on time, the interest rate is applied. (Marciniuk, 2021) consider an interest rate that varies from year to year. Different interest rate models are widely discussed in the literature. (Kellison, 2009) and (Boyle, 1976) used random variables as interest rates. Autoregressive processes are applied to actuarial science as technical interest rate models (Bellhouse and Panjer, 1981; Marciniuk, 2004; Panjer and Bellhouse, 1980). (Beekman and Fuelling, 1990, 1993; Dębicka, 2003; Garrido, 1988; Marciniuk, 2004; Parker, 1994a, 1994b) described the Wiener and Ornstein-Uhlenbeck process as an interest rate model. An extensive application of the short-term interest rate and description of the models can be found in the works of (Carriere, 1999, 2004; Jakubowski et al., 2003; Musiela and Rutkowski, 1988). (Dhaene, 2000) applies the CIR model to life insurance. (Carriere, 1999, 2004) considers stochastic interest rate models to determine actuarial values of life annuities and net premiums in life insurance. In actuarial science, the fuzzy set theory has been used to model problems connected with subjective judgment and situations when information available is imprecise and incomplete. We can meet articles on general actuarial issues using fuzzy sets in life and non-life insurance (Derrig and Ostaszewski, 2004; Lamaire, 1990; Ostaszewski, 1993; Shapiro, 2004). Life insurance issues such as calculation price life insurance policies, life insurance portfolios, life contingencies, life actuarial liabilities, life annuities have been covered in (Andres-Sanchez, Gonzalez-Vila Puchades and Gonzalez-Vila Puchades L., 2012; Andres-Sanchez and Gonzalez-Vila Puchades, 2017a; Andres-Sanchez and Gonzalez-Vila Puchades, 2017b) articles. (Derrig and Ostaszewski, 1997) dealt with issues related to property-liability insurance and (Shapiro, 2013) dealt with modelling future lifetime. The problems connected with risk models as ruin theory and claims aggregation were investigated in (Heilpern, 2018; Huang et al., 2009). The fuzzy interest rate was dealt in (Andres and Terceno, 2003; Betzuen et al., 1997; Lamaire, 1990).

2 METHODS

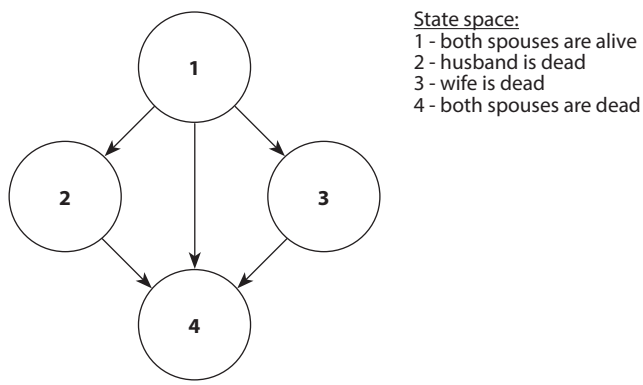
2.1 Matrix representation of benefits

2.1.1 Multiple state model

Multiple state modelling is a classical tool for designing and implementing insurance contracts. Among others it is used to calculate premiums for marriage insurances. We used the multistate model for marriage insurances to determine the benefits resulting from the equity release contract.

Generally the pair is called a multiple state model (MSM). S is the set space, where each state corresponds to an event which determines the cash flows (premiums and/or benefits). T denotes the set of direct transitions between states of S. For marriage reverse annuity contract, MSM has the following form:

Figure 1 Scheme of the multiple state model for marriage reverse annuity contract



Source: Adopted from (Denuit et al., 2001)

$$(S,T) = (\{1,2,3,4\},\{(1,2),(1,3),(1,4),(2,4),(3,4)\}), \tag{1}$$

and describes all possible contracted risk events up to the end of the contract. MSM given by Formula (1): is graphically presented in Figure 1, where states are marked with circles, and arrows indicate possible direct transitions between them.

In order to determine benefits it is necessary to formalize the description of probabilistic structure of the multiple state model, and all cash flows arising from the contract. Because the cash flows are realized at a certain time, it is necessary to choose an appropriate interest rate model. In this section we briefly discuss all three of these elements necessary for the valuation of benefits and finally we present the formulas for the annuity installment.

2.1.2 Probabilistic structure

The value of benefit is determine on the basis of spouses future lifetime, which depends on age at entry and the period of the contract. By x_w we denote the wife's age at entry and by x_m the husband's age at entry. We will assume that *the future life time of husband and wife are independent* (ASSUMPTION 1). In this paper, we consider a contract issued at time 0 and terminating according to the plan at a later time n, which is called the term of contract or the contract period. The length of the contract depends on its type. We will consider two types of contract, the last surviving status (LSS) when the annuity is paid as long as one spouse is alive and the joint-life status (JLS) when benefits are paid only while both spouses are alive. Thus the period of the contract n is defined as follows:

$n = \omega - min\{x_w, x_m\}$ for LSS,
$n = min\{\omega - x_w, \omega - x_m\}$ for JLS,

where ω is the maximum life expectancy which varies between 100 and 110 years depending on the population.

For a given contract, the function $Y(k)$ means that at time $k = \{0, 1, 2, ..., n\}$ (meaning the time that has elapsed from the beginning of the contract) the marriage is in one of four life situations described by the multiple state model shown in Figure 1. The life cases covered by the contract are random in nature. So it is natural to assume that $\{X(k); k = 0, 1, 2, ..., n\}$ is a discrete-time stochastic process taking values from a set space $S = \{1, 2, 3, 4\}$ and used to describe the evolution of the insured risk. Consistent with the actuarial literature, we assume that $\{X(k)\}$ is modelled by a nonhomogeneous Markov chain cf. (Hoem, 1969 and 1988; Norberg, 2002; Wolthuis, 1994). Because we focus on discrete-time model, where cash flows are made at the ends of time interval, the probabilistic structure of the model can be described in the matrix form:

$$\mathbf{D} = \begin{pmatrix} pr_1(0) & pr_2(0) & pr_3(0) & pr_4(0) \\ pr_1(1) & pr_2(1) & pr_3(1) & pr_4(1) \\ \vdots & \vdots & \vdots & \vdots \\ pr_1(n) & pr_2(n) & pr_3(n) & pr_4(n) \end{pmatrix}, \tag{2}$$

where $pr_i(k) = P(X(k) = i)$ and $i \in S$. Notice that each row of the matrix $\mathbf{D} \in \mathbb{R}^{(n+1)\times 4}$ is one-dimensional distribution for particular moment of the contract's period.

2.1.3 Cash flows

The individual's presence in a given state or movement (transition) from one state to another may have some financial impact. So, as a result of the agreement arise financial streams, consisting of cash flows of the agreement between the parties at the time. We consider two types of cash flow:

- a single payment (inflow from the company's mortgage fund point of view):

$$\pi_j(k) = \begin{Bmatrix} \alpha W & \text{for } j = 1 \text{ i } k = 0 \\ 0 & \text{besides} \end{Bmatrix}, \tag{3}$$

where W is the value of property (benefits are calculated for a percentage α of the value of property, where $0 < \alpha \leq 0{,}5$ cf. (Dębicka and Marciniuk, 2014)),

- an annuity benefit i.e. a fixed annuity installment payable in advance which is determined depending on the type of contract. (outflow from the company's mortgage fund point of view). Depending on a status, annuity benefits $\ddot{b}$ are paid in more than one state (LSS):

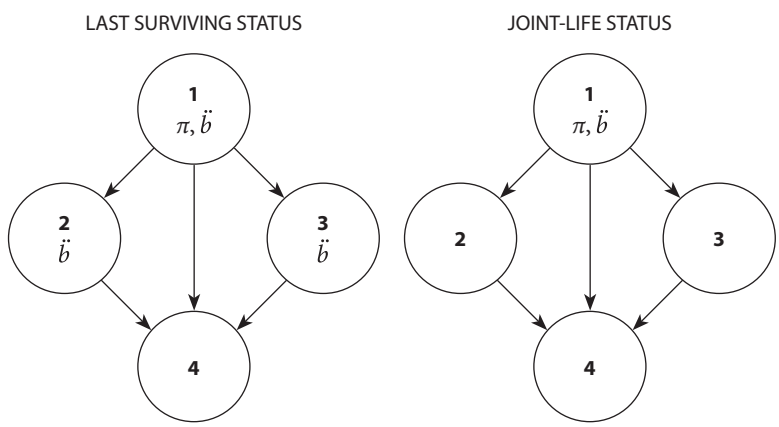
$$\ddot{b}_j(k) = \begin{cases} \ddot{b} & \text{for } j = 1 \text{ and } k = 0, 1, ..., n-1 \\ & \text{for } j = 2, 3 \text{ and } k = 1, 2, ..., n-1 \\ 0 & \text{besides} \end{cases}, \tag{4}$$

or only in one state (JLS):

$$\ddot{b}_j(k) = \begin{cases} \ddot{b} & \text{for } j = 1 \text{ and } k = 0, 1, ..., n-1 \\ 0 & \text{besides} \end{cases}. \tag{5}$$

Figure 2 illustrates the cash flows for the different statuses of the marriage reverse annuity contract.

Figure 2 MSM and cash flows for statuses of the marriage reverse annuity contract



Source: Based on (Dębicka et al., 2020)

Depending on a status, annuity benefits are paid in different. Therefore, we need to define the cash flow matrices for each status separately. Note that for the LSS contract, the annuity is paid when the process $\{X(k)\}$ is in states 1, 2 and 3, and then the cash flow matrix $\mathbf{C} \in \mathbb{R}^{(n+1)\times 4}$ has the following form:

$$\mathbf{C} = \begin{pmatrix} \alpha W - \ddot{b} & 0 & 0 & 0 \\ -\ddot{b} & -\ddot{b} & -\ddot{b} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ -\ddot{b} & -\ddot{b} & -\ddot{b} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \alpha W & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} -\ddot{b} & 0 & 0 & 0 \\ -\ddot{b} & -\ddot{b} & -\ddot{b} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ -\ddot{b} & -\ddot{b} & -\ddot{b} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{C}_{in} + \mathbf{C}_{out} . \tag{6}$$

Whereas, for the JLS contract, the annuity is paid when the process {X(k)} is in state 1, and then the cash flow matrix $\mathbf{C} \in \mathbb{R}^{(n+1)\times 4}$ is as follows:

$$\mathbf{C} = \begin{pmatrix} \alpha W - \ddot{b} & 0 & 0 & 0 \\ -\ddot{b} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ -\ddot{b} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \alpha W & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} -\ddot{b} & 0 & 0 & 0 \\ -\ddot{b} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ -\ddot{b} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{C}_{in} + \mathbf{C}_{out} . \tag{7}$$

Let us note that in Formulas (6) and (7) matrices $\mathbf{C}_{in}$ contain only cash flows which are inflow to the company's mortgage fund and $\mathbf{C}_{out}$ matrices are defined based on cash flows that are outflow from the company's mortgage fund.

2.1.4 Discount function

Interest rate may change with time, so it can be a function depending on time or can be modelled by stochastic process. Let $Y(t)$ denote the rate of interest in time interval [0, t]. Then $\{Y(t); t \geq 0\}$ is the interest rate accumulation process (stochastic process with stationary increments). From a technical point of view we assume *all moments of the random discounting function* $e^{-Y(t)}$ *are finite* (ASSUMPTION 2) cf. (Dębicka, 2013). Moreover, the future life time of spouses are independent on the interest rate and that means *the random variables* $X(t)$ *and* $Y(t)$ *are independent* (ASSUMPTION 3) cf. (Frees, 1990). For the discrete-time model, we define matrix $\mathbf{\Lambda} = [\lambda_{k_1 k_2}]_{k_1,k_2=0}^{n} \in \mathbb{R}^{(n+1)\times(n+1)}$ based on the discount function, where:

$$\lambda_{k_1 k_2} = E(e^{-(Y(k_1)-Y(k_2))}) = \begin{cases} E(v(k_2, k_1)) & \text{dla} \quad k_1 > k_2 \\ 1 & \text{dla} \quad k_1 = k_2 \\ E(r(k_1, k_2)) & \text{dla} \quad k_1 < k_2 \end{cases} . \tag{8}$$

For more on modelling the process $Y(t)$, and hence determining the elements of matrix $\mathbf{\Lambda}$, see Section 2.

2.1.5 Benefits

In Theorem 1 we present formulas for the annuity payable in advance for both statuses.

Theorem 1. Assume that the principle of equivalence and ASSUMPTIONS 1–3 are satisfied. For n-year marriage reverse annuity contract the cash flow matrix is determined for company's mortgage fund and αW is the capital for which the value of benefit is calculated. Let $\ddot{b}$ denote the annuity payable in advance if:

- both spouses are alive (JLS), then:

$$\ddot{b} = \frac{\mathbf{I}_1^T \mathbf{C}_{in} \mathbf{J}_1}{\mathbf{I}_1^T \mathbf{\Lambda}^T (\mathbf{I} - \mathbf{I}_{n+1} \mathbf{I}_{n+1}^T) \mathbf{D} \mathbf{J}_1} , \tag{9}$$

- at least one of the spouses is alive (LSS), then:

$$\ddot{b} = \frac{\mathbf{I}_1^T \mathbf{C}_{in} \mathbf{J}_1}{\mathbf{I}_1^T \mathbf{\Lambda}^T (\mathbf{I} - \mathbf{I}_{n+1} \mathbf{I}_{n+1}^T - \mathbf{I}_1 \mathbf{I}_1^T) \mathbf{D}(\mathbf{S} - \mathbf{J}_4) + 1}, \tag{10}$$

where the vectors in (9) and (10) are defined as follows: $\mathbf{S} = (1, 1, 1, 1)^T$, $\mathbf{J}_1 = (1, 0, 0, 0)^T$, $\mathbf{J}_4 = (0, 0, 0, 1)^T$, $\mathbf{I}_1 = (1, 0, ..., 0)^T \in \mathbb{R}^{(n+1)}$ and $\mathbf{I}_{n+1} = (0, ..., 0, 1)^T \in \mathbb{R}^{(n+1)}$.

We drop the proof of the theorem on the grounds that it is analogous to the proofs of the theorems presented in (Dębicka, 2013; Dębicka et al., 2020).

Note that, since the matrix $\mathbf{C}_{in}$ (which depends only on cash flow arising from the contract) and matrix $\mathbf{D}$ (which depends on the probabilistic structure of the MSM) are constant for a given contract, the amount of the annuity instalment depends on the choice of the interest rate model (forms of the matrix $\mathbf{\Lambda}$).

It is easy to observe that the numerators in Formulas (9) and (10) are equal to the corresponding percentage of the value of property, that is $\mathbf{I}_1^T \mathbf{C}_{in} \mathbf{J}_1 = \alpha W$. Moreover, the denominators correspond to the unit annuities payable in advance (temporary life annuity due) $\ddot{a}_{\overline{x_u x_m:n|}}$:

- for JLS (the unit annuity instalment is paid only if both spouses are alive – in state 1) we have $\ddot{a}_{\overline{x_u x_m:n|}} = \mathbf{I}_1^T \mathbf{\Lambda}^T (\mathbf{I} - \mathbf{I}_{n+1} \mathbf{I}_{n+1}^T) \mathbf{D} \mathbf{J}_1$,

- for LSS (the annuity instalment is paid only if at least one spouse is alive – in states 1, 2, 3) we have $\ddot{a}_{\overline{x_u x_m:n|}} = \mathbf{I}_1^T \mathbf{\Lambda}^T (\mathbf{I} - \mathbf{I}_{n+1} \mathbf{I}_{n+1}^T - \mathbf{I}_1 \mathbf{I}_1^T) \mathbf{D}(\mathbf{S} - \mathbf{J}_4) + 1$.

The formulas in Theorem 1 can also be used when annuities are paid more frequently than annually. Let m be the frequency of annuity payments, e.g. $m = 2$ (half-yearly annuity), $m = 4$ (quarterly annuity), $m = 12$ (monthly annuity). The annual annuity installment payable m times a year is thus given by the formula:

$$\ddot{b}^{(m)}_{\overline{x_u x_m:n|}} = \frac{\alpha W}{\ddot{a}_{\overline{x_u x_m:n|}}} = m \cdot \frac{\alpha W}{\ddot{a}_{\overline{x_u x_m:n \cdot m|}}}, \tag{11}$$

where $\ddot{a}_{\overline{x_u x_m:n|}}$ is the unit annual annuity due paid m times a year in the amount of $\frac{1}{m}$:

$$\ddot{a}^{(m)}_{\overline{x_u x_m:n|}} = \begin{cases} \frac{1}{m} \sum_{k=1}^{nm-1} \mathrm{E}(v(0, \frac{k}{m})) pr_1(k/m) & \text{for JLS} \\ \frac{1}{m}\, 1 + \sum_{k=1}^{nm-1} \mathrm{E}(v(0, \frac{k}{m}))(pr_1(k/m) + pr_2(k/m) + pr_3(k/m))) & \text{for LSS}, \end{cases} \tag{12}$$

where $pr_i(k/m) = pr(X(k/m) = i)$ for $k = 0, 1, 2, ..., nm - 1$ and $i = 1, 2, 3$. Note that from (12) we have $\ddot{a}_{\overline{x_u x_m:n|}} = \frac{1}{m} \ddot{a}_{\overline{x_u x_m:n \cdot m|}}$ where $\ddot{a}_{\overline{x_u x_m:n \cdot m|}}$ is the unit annuity payable in advance for the period and can be determined in a matrix manner as in Theorem 1:

$$\ddot{a}_{\overline{x_u x_m:n \cdot m|}} = \begin{cases} \mathbf{I}_1^T \mathbf{\Lambda}^T (\mathbf{I} - \mathbf{I}_{n \cdot m+1} \mathbf{I}_{n \cdot m+1}^T) \mathbf{D} \mathbf{J}_1, & \text{for JLS} \\ 1 + \mathbf{I}_1^T \mathbf{\Lambda}^T (\mathbf{I} - \mathbf{I}_{n \cdot m+1} \mathbf{I}_{n \cdot m+1}^T - \mathbf{I}_1 \mathbf{I}_1^T) \mathbf{D}(\mathbf{S} - \mathbf{J}_4) & \text{for LSS}. \end{cases} \tag{13}$$

Note that, increasing the frequency of annuity payments will increase the dimension of the matrices that depend on the insurance period n which are used in (13). The dimensions of these matrices will grow up into $n{\cdot}m{+}1$. In particular matrix $\mathbf{D} \in \mathbb{R}^{(n\cdot m+1)\times 4}$ and its elements will be determined under the assumption that the probability of death within a year is uniform. Similarly, the matrix $\mathbf{\Lambda} \in \mathbb{R}^{(n\cdot m+1)\times(n\cdot m+1)}$ and its elements will be determined for a correspondingly shorter interest rate.

Additionally, let us note that in (11) the factor $\frac{\alpha W}{\ddot{a}_{\overline{x_h x_m:n\cdot m|}}}$ is the amount of the monthly annuity due paid $m \cdot n$ times.

2.2 Interest rate modelling

There are three possibilities for modelling the interest rate: the actuarial, financial and fuzzy. The first two ways are described in detail in (Marciniuk, 2009). These ways assume that the interest rate is a stochastic process and require knowledge about the technical and short-term rates. The third possibility considers the constant interest rate modelled by a fuzzy number. The necessary concepts will be introduced below and the interest rate models used in the empirical section.

2.2.1 Actuarial and financial modelling of interest rate

At the beginning, it is necessary to explain the relationship between actuarial and financial approaches to the interest rate modelling.

The actuarial way requires an understanding of a technical discounting function. Let us assume that capital K_t is invested at the moment t. Its value after T ($0 \le t \le T$) years is K_T. The discounting function from time T on time t is defined as follows (Marciniuk et al., 2020):

$$v_{t,T} = \frac{K_T}{K_t},\ 0 \le t \le T. \tag{14}$$

The denotation in the financial literature of the discounting function $v_{t,T}$ is equivalent to that adopted in Section 2.1, i.e. $v(t,T)$ (c.f. Formula (8)).

The discounting function has the following relationship with the force of interest rate function $\delta_{t,T}$ as follows (Bellhouse and Panjer, 1981):

$$v_{t,T} = exp(-\delta_{t,T}), \quad 0 \le t \le T, \tag{15}$$

therefore in the actuarial approach the discount function or the force of interest rate can be modelled.

To determine financial models of an interest rate it is necessary to introduce the concept of pricing a zero-coupon bond (Marciniuk et al., 2020). A zero-coupon bond is a stock, which is sold at a discount. The customer's profit is the difference between its nominal and selling price (Musiela and Rutkowski, 1988). By convention, the nominal price is one financial unit. It means that the zero-coupon's holder will receive one unit of cash at moment T. The price of a zero-coupon bond of maturity T at any instant t ($0 \le t \le T$) is denoted by $P_{t,T}$. It is obvious that $P_{t,T} = v_{t,T}$, because:

$$v_{t,T} = \frac{K_t}{K_T} = \frac{P_{t,T}}{P_{T,T}} = \frac{P_{t,T}}{1} = P_{t,T}. \tag{16}$$

Therefore, the discounting function could be described as the price of a zero-coupon bond (Marciniuk, 2009) and could be modelled by the use of $P_{t,T}$.

In the actuarial approach, the models of the force interest rate function $\delta_{t,T}$ from time t to time T, $0 \le t \le T$, are applied. The force of interest δ_t at the moment $t \ge 0$ is described using a stochastic process with discrete or continuous time. Hence:

$$\delta_{t,T} = \begin{cases} \sum_{t}^{T} \delta_t, & when \quad s = 0, 1, 2, ... \\ \int_{t}^{T} \delta_t ds, & when \quad s \ge 0 \ . \end{cases} \tag{17}$$

The most popular stochastic model of the force of interest is a process with a discrete-time called an autoregressive process of order one (AR(1)). This process is defined by the following recursive relation (Brockwell and Davis, 1996):

$$\delta_t = \mu + \phi(\delta_{t-1} - \mu) + \varepsilon_t, \quad t = 1, 2, ..., \tag{18}$$

where $\delta_0 \in \mathbb{R}, \mu \in \mathbb{R}, |\phi| < 1$. Moreover $\varepsilon_t \sim N(0,\sigma^2)$ and variables δ_t and ε_s are independent for $s < t$. This process is stationary and has a normal distribution with mean μ and variance $\frac{\sigma^2}{1-\phi^2}$.

To calculate the benefit of reverse marriage annuity, it is necessary to know the value $E(v_{0,t}^k)$. It can be concluded from the fact that the process $\{\delta_t\}_{t\ge0}$ has a normal boundary distribution, that the interest rate function $\delta_{t,T}$ also has a normal distribution. Hence and from Formula (16), it follows that (Panjer and Bellhouse, 1980):

$$E(v_{0,t}^k) = M_{\delta_{0,t}}(-k) = exp(-k\mu t + 0{,}5k^2\, V(\delta_{0,t})) \ , \tag{19}$$

where:

$$V(\delta_{0,t}) = \frac{t\sigma^2}{1-\phi} + 2\,\frac{\sigma^2}{1-\phi^2}\,\frac{\phi}{1-\phi}\,(t - 1 - \phi\,\frac{1-\phi^{t-1}}{1-\phi}) \ . \tag{20}$$

$M_X(k)$ means the function generating the moments of the variable X in point k.

The autoregressive process of order one is a discrete version of the continuous Ornstein-Uhlenbeck process.

The process, which is applied as a continuous force of interest $\{\delta_t\}_{t\ge0}$, is the following stochastic Wiener process (Dhaene, 2000):

$$d\delta_t = \sigma dB_t, \qquad t \ge 0, \tag{21}$$

where $\delta_0 \in \mathbb{R}$, $\sigma \ge 0$ and $\{B_t\}_{t\ge0}$ mean the standard Brownian motion.

The solution of the above stochastic differential equation is a process with the following form:

$$\delta_t = \delta_0 + \sigma B_t, \quad t \ge 0. \tag{22}$$

Hence it is easy to prove that (Musiela and Rutkowski, 1988):

$$E(\delta_t) = \delta_0, \quad V(\delta_t) = \sigma^2 t, \quad C(\delta_s,\delta_t) = \sigma^2 C(B_s,B_t) = \sigma^2\, min(s, t).$$

Because process $\{\delta_t\}_{t\geq 0}$ is also a Gaussian process, the following equity is accurate:

$$E(v_{0,t}^k) = M_{\delta_{0,t}}(-k) . \tag{23}$$

In this case $M_{\delta_{0,t}}(k)$ is given as follows (Marciniuk, 2004):

$$M_{\delta_{0,t}}(k) = exp(kt\delta_0 + k_2\frac{\sigma^2 t^3}{6}) . \tag{24}$$

The price of a zero-coupon bond can be determined using the short-term interest rate process $\{r_t\}_{t\geq 0}$ (Carriere, 2004; Musiela and Rutkowski, 1988). If r_t is a stochastic process adaptive with the filtering F_t, $\int_0^T |r_s| ds < \infty$ and a martingale measure **Q** equivalent to the measure **P** exists on the probabilistic space $(\Omega, F, \mathbf{P})$, then:

$$P_{t,T} = E^{Q}(exp(-\int_t^T r_s ds)|F_t) . \tag{25}$$

Consider the case when the short-term rate is determined by the following Ornstein-Uhlenbeck process (Vasicek, 1999):

$$dr_t = -\alpha(r_t - \mu)dt + \sigma dB_t , \tag{26}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$, $\alpha > 0$.

Process $\{B_t\}_{t\geq 0}$ is a standard Brownian motion under the **Q** measure. From the Girsanov theorem, it is known that $B_t = B_t^* + B_t$, where $\{B_t^*\}_{t\geq 0}$ is the standard Brownian motion under the **P** measure, where β ($\beta > 0$) means the price of the risk (Carriere, 2004; Jakubowski et al., 2003; Vasicek, 1999). This model of the short-term rate is known as a Vasicek model.

The following formula gives the price of a zero-coupon bond at the moment 0 with the maturity t ($t \geq 0$):

$$P_{0,t} = exp(-r_0 \cdot n_{0,t} + m_{0,t}), \tag{27}$$

where:

$$n_{0,t} = \frac{1}{\alpha}\, n_{0,t}(1 - e^{-\alpha t}), \tag{28}$$

$$m_{0,t} = -\mu t + \mu \cdot n_{0,t} + 0.5\sigma^2(\frac{t}{\alpha^2} + \frac{2}{\alpha^3}(1 - e^{-\alpha t}) - 0.5\frac{1}{\alpha^3}(1 - e^{-2\alpha t})). \tag{29}$$

Consider the Cox-Ingersoll-Ross (CIR) model, when the short-rate is determined by the use of the stochastic differential equation (Dhaene, 2000; Jakubowski et al., 2003; Musiela and Rutkowski, 1988):

$$dr_t = \alpha(\mu - \alpha r_t)dt + \sigma\sqrt{r_t} dB_t , \tag{30}$$

where $\mu \in \mathbb{R}$, $0 < \sigma < \alpha$, $B_t = B_t^* + \int_0^t \sqrt{r_u}\, du$.

Formula (27) gives the price of a zero-coupon bond, where (Denuit et al., 2001; Jakubowski et al., 2003; Marciniuk, 2009):

$$n_{0,t} = \frac{2(exp(2\gamma t) - 1)}{2\gamma - \alpha + (2\gamma + \alpha)\, exp(2\gamma t)}, \tag{31}$$

$$m_{0,t} = \frac{2\mu}{\sigma^2} ln(\frac{4\gamma\, exp(0.5(2\gamma + \alpha)t)}{2\gamma - \alpha + (2\gamma + \alpha)\, exp(2\gamma t)}), \tag{32}$$

and $\gamma = 0.5\sqrt{\alpha^2 + 2\sigma^2}$.

2.2.2 Fuzzy interest rate

First, we recall some notions connected with fuzzy sets. The fuzzy subset **A** of the space X is described by its membership function $\mu\mathbf{A}$: $X \to [0, 1]$ (Dubois and Prade, 1980; Zadeh, 1965). The value $\mu_{\mathbf{A}}(x)$ represents the degree of membership of the element x to the fuzzy set **A**. If $\mu_{\mathbf{A}}(x) = 0$ then we have non-membership of x and for $\mu_{\mathbf{A}}(x) = 1$ we obtain absolute membership. The crisp sets $A\alpha = \{x \in X \mid \mu_{\mathbf{A}}(x) \geq \alpha\}$ for any $0 < \alpha \leq 1$ are called the α-cuts and the support $A0$ of the fuzzy set **A** is the closure of set $\{x \in X \mid \mu_{\mathbf{A}}(x) > 0\}$. The set A_1 is called the core of **A**. The α-cuts $A\alpha$ unambiguously define the fuzzy set **A**. We will denote the fuzzy sets by the bold letters, e.g. **A**, and the crisp sets by the italic, non-bold letters, e.g. $A\alpha$.

In our paper, we will use the fuzzy subsets **A** of real line $\mathbb{R}$, i.e. $X = \mathbb{R}$. In addition, we assume that $A_0 = [a, d]$, $A_1 = [b, c]$ and the membership function $\mu_{\mathbf{A}}$ is continuous. Moreover, this function is strictly increasing on $[a, b]$ and strictly decreasing on $[c, d]$. Every α-cuts of such fuzzy set is the compact interval, i.e. $A_\alpha = [A_\alpha{}^L, A_\alpha{}^U]$ and this fuzzy set is called the fuzzy number (Dubois and Prade, 1980). We obtain the trapezoidal fuzzy number when the membership function on the intervals $[a, b]$ and $[c, d]$ are linear. We denote it as $\mathbf{A} = (a, b, c, d)$. If $a = b$ we have the triangular fuzzy number $\mathbf{A} = (a, b, d)$. The actual number a can be treated as the degenerate triangular fuzzy number (a, a, a).

In our paper, we will use the arithmetic operation on fuzzy numbers. Moreover, the fuzzy numbers used in this paper are positive, i.e. $a > 0$. These arithmetic operations are based on the α-cuts and they are defined in the following way:

$$(A + B)_\alpha{}^L = A_\alpha{}^L + B_\alpha{}^L, (A + B)_\alpha{}^U = A_\alpha{}^U + B_\alpha{}^U, \tag{33}$$

$$(A \cdot B)_\alpha{}^L = A_\alpha{}^L \cdot B_\alpha{}^L, (A \cdot B)_\alpha{}^U = A_\alpha{}^U \cdot B_\alpha{}^U, \tag{34}$$

$$(A / B)_\alpha{}^L = A_\alpha{}^L / B_\alpha{}^U, (A / B)_\alpha{}^U = A_\alpha{}^U / B_\alpha{}^L, \tag{35}$$

$$(\lambda A)_\alpha{}^L = \begin{cases} \lambda A_\alpha{}^L \text{ for } \lambda \geq 0 \\ \lambda A_\alpha{}^U \text{ for } \lambda < 0 \end{cases}, (\lambda A)_\alpha{}^U = \begin{cases} \lambda A_\alpha{}^U \text{ for } \lambda \geq 0 \\ \lambda A_\alpha{}^L \text{ for } \lambda < 0 \end{cases}, \tag{36}$$

$$(A^\alpha)_\alpha{}^L = (A_\alpha{}^L)^\alpha \quad (A^\alpha)_\alpha{}^U = (A_\alpha{}^U)^\alpha. \tag{37}$$

The sum of the fuzzy triangular numbers **A** + **B** is the triangular fuzzy number, too. This property also holds for the product λ**A**. The results of other arithmetic operations are the fuzzy number, but not fuzzy triangular numbers, the membership functions are not linear. We can define the difference of the fuzzy numbers **A** – **B** using α-cuts or the formula:

$$\mathbf{A} - \mathbf{B} = \mathbf{A} + (-1)\mathbf{B}. \tag{38}$$

The area below the graph of membership function of the fuzzy number can be treated as the degree of imprecision and it is denoted as Imp(**A**). For the triangular fuzzy number **A** = (*a*, *b*, *d*) we obtain:

$$Im\ \mathrm{p}(\mathrm{A}) = \frac{d-a}{2}. \tag{39}$$

The mean value of the fuzzy number **A** is defined as follows:

$$M(\mathbf{A}) = \frac{1}{2}\int_0^1 (A_\alpha^{\ L} + B_\alpha^{\ U})\, d\alpha. \tag{40}$$

We have:

$$M(\mathbf{A}) = \frac{a + 2b + d}{4} \tag{41}$$

for the triangular fuzzy number **A**.

The triangular fuzzy number **A** = (*a*, *b*, *d*) has a linear membership function. So, the parameters *a*, *b* and *d* univocally define it.

We can define the order between triangular fuzzy numbers in the following, natural way. Let $\mathbf{A_1} = (a_1, b_1, d_1)$ and $\mathbf{A_2} = (a_2, b_2, d_2)$ than:

$$\mathbf{A_1} \le \mathbf{B_2} \Leftrightarrow (a_1 \le a_2, b_1 \le b_2, d_1 \le d_2). \tag{42}$$

Now we investigate the case when the interest rate has imprecision form. For instance, we obtain the imprecision information, that it is "about 0.055". We can model such interest rate as the triangular fuzzy number **i** = (*a*, *b*, *d*). The parameters of this fuzzy number are determined based on additional information, experts' assessments, and own intuition. Let us assume that fuzzy interest rate takes the form:

$$\mathbf{i} = (0.05, 0.055, 0.065). \tag{43}$$

The mean value of this fuzzy number is equal $M(\mathbf{i}) = 0.05625$ and imprecision $\mathrm{Imp}(\mathbf{i}) = 0.0075$.

The graph of the membership function of triangular fuzzy number **i**, the sample α-cut I0.4 and the mean value $M(\mathbf{i})$ are included in Figure 7a.

The fuzzy number 1 + **i**, where 1 is treated as a degenerate triangular fuzzy number (1, 1, 1), is triangular too. We have 1 + **i** = (1.05, 1.055; 1.065). The fuzzy discounting factor **v** = 1/(1 + **i**) is not a triangular fuzzy number. Every α-cuts of it take a form:

$$v_\alpha = [(1.065 - 0.01\alpha)^{-1}, (1.05 + 0.005\alpha)^{-1}]. \tag{44}$$

So, it is the fuzzy number with the following membership function:

$$\mu_{\mathrm{V}}(x) = \begin{cases} (1.065x - 1) / 0.01x & \text{for } 1 / 1.065 \le x < 1 / 1.055 \\ (1 - 1.05x) / 0.005x & \text{for } 1 / 1.055 \le x \le 1 / 1.05. \\ 0 & \text{otherwise} \end{cases} \tag{45}$$

The graph of such membership function is presented in Figure 7b. We see that it is almost linear, so we can treat it as a linear fuzzy number (0.939, 0.948, 0.952).

We can approximate the fuzzy discounting factor **v** as the triangular fuzzy number (v_a, v_b, v_d). The α-cut of fuzzy power $\mathbf{v}^c$, where $c > 0$, takes the form:

$$(v^c)_\alpha = [(v_a + (v_b - v_a)\alpha)^c, (v_d - (v_d - v_b)\alpha)^c]. \tag{46}$$

This fuzzy number is almost linear, too.

If we consider the joint-life status (JLS), then based on (12), we can treat the fuzzy joint-life annuity $\ddot{\mathbf{a}}^{(m)}_{\overline{x_w x_m:n|}}$ as the triangular fuzzy number.

$$(\frac{1}{m}\sum_{k=0}^{nm-1} v_a^{k/m} pr_1(\frac{k}{m}), (\frac{1}{m})\sum_{k=0}^{nm-1} v_b^{k/m} pr_1(\frac{k}{m}), (\frac{1}{m})\sum_{k=0}^{nm-1} v_d^{k/m} pr_1(\frac{k}{m}). \tag{47}$$

Considering (47) and taking into account (35), we derive the fuzzy annuity benefit paid m times a year:

$$\ddot{\mathbf{b}}^{(m)} = (\ddot{b}_a^{(m)}, \ddot{b}_b^{(m)}, \ddot{b}_d^{(m)}) = (\frac{\alpha W}{(\ddot{a}^{(m)}_{\overline{x_w x_m:n|}})_d}, \frac{\alpha W}{(\ddot{a}^{(m)}_{\overline{x_w x_m:n|}})_b}, \frac{\alpha W}{(\ddot{a}^{(m)}_{\overline{x_w x_m:n|}})_a}). \tag{48}$$

3 RESULTS

The short-term rate is not directly observed on the financial market, therefore to calculate the benefit of marriage reverse annuity contract, we assume that we know the simulated data of the short-term rate. The data was generated from the following distribution (James and Webber, 2000; Marciniuk, 2009):

$$r_{t_{i+1}} \mid r_{t_i} \sim N(\mu + (r_{t_i} - \mu)\, exp(-\alpha(t_{i+1} - t_i)), \sigma\sqrt{\frac{1 - exp(-2\alpha(t_{i+1} - t_i))}{2\alpha}}, \tag{49}$$

for $\alpha = 8$, $\mu = 0.055$, $\sigma = 0.04$, $r_0 = 0.05$.

We assume that the weekly data was observed throughout 20 years. The parameters of the interest rate models were estimated based on these data using the maximum likelihood method in the case of the Wiener process, AR(1) process and the Vasicek model. The general method of moments was applied in the case of the CIR model. In this aim, we use the packet Solver in Microsoft Excel. The results of the estimation are as follows (Marciniuk, 2009):

- AR(1) process:
  $d\delta_t = 0.05524 + 0.84598(\delta_{t-1} - 0.05524) + \varepsilon_t$,
  $\varepsilon_t \sim N(0, 0.009375)$, $\delta_0 = 0.04845$,
- Wiener process:

$d\delta_t = 0.0052 dB_t$, $\delta_0 = 0.04845$,

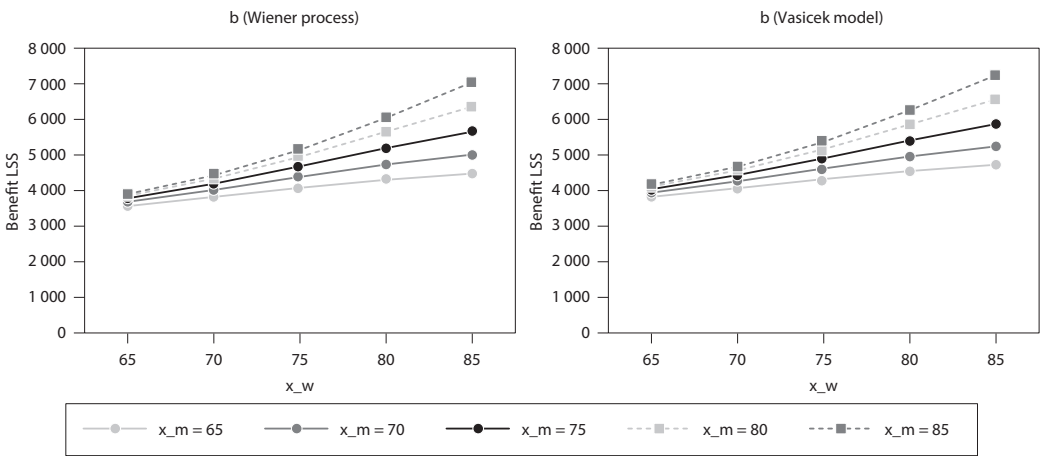
– Vasicek model:
$dr_t = -8.67(r_t - 0.055)\, dt + 0.04 dB_t$,

– CIR model:
$dr_t = (0.06218 - 1.1254 r_t)\, dt + 0.32 \sqrt{r_t}\, dB_t$.

Moreover, we assume that the property's actual value W is 100 000 euros and a reduction factor α of 50%. We use Polish Life Tables from 2012. We take into account the uniform distribution within a year. The spouses' future lifetimes are independent random variables. The constant technical interest rate is equal to the long-term interest rate in the Vasicek model and is almost 5.5%. Hence the fuzzy interest rate is also approximated at about 5.5%.

At the beginning, we discuss actuarial and financial models. We distinguish between two statuses: a joint-life status (JLS), when the benefit is paid only until the death of the first spouse, and a last surviving status (LSS) contract by which the benefit is paid until the death of the other spouse (Dębicka et al., 2020). The benefits are calculated from Formulas (9) and (10) and their modification, described at the end of Section 2.1. The graph below shows the benefit amounts of the marital reversionary annuity for the Vasicek model and the Wiener process.

Figure 3 The LSS benefits in the case of the Wiener process and the Vasicek model

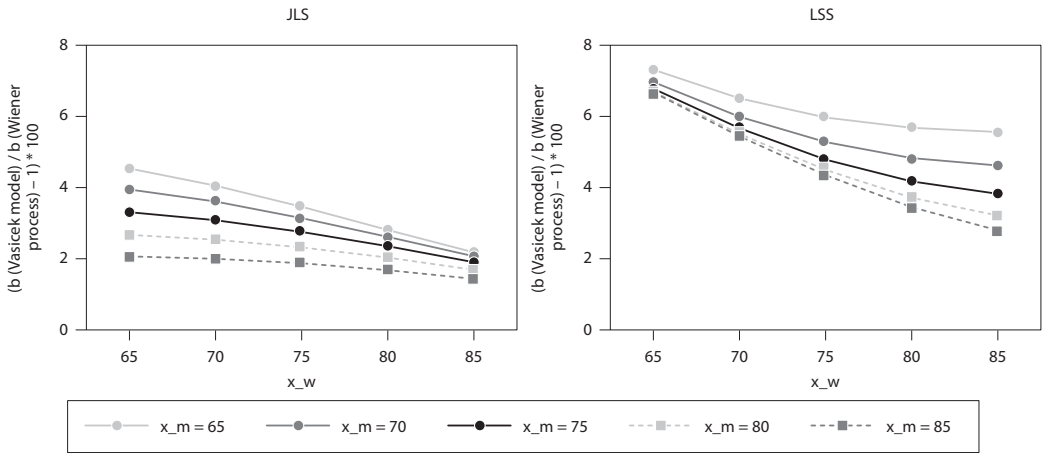


Source: Own elaboration

The Wiener process obtains the lowest benefits and the highest for the Vasicek model. Therefore in the following graphs, we can see the relative percentage differences between these benefits for the man at different ages x_m depending on woman's age x_w, calculated as follows:

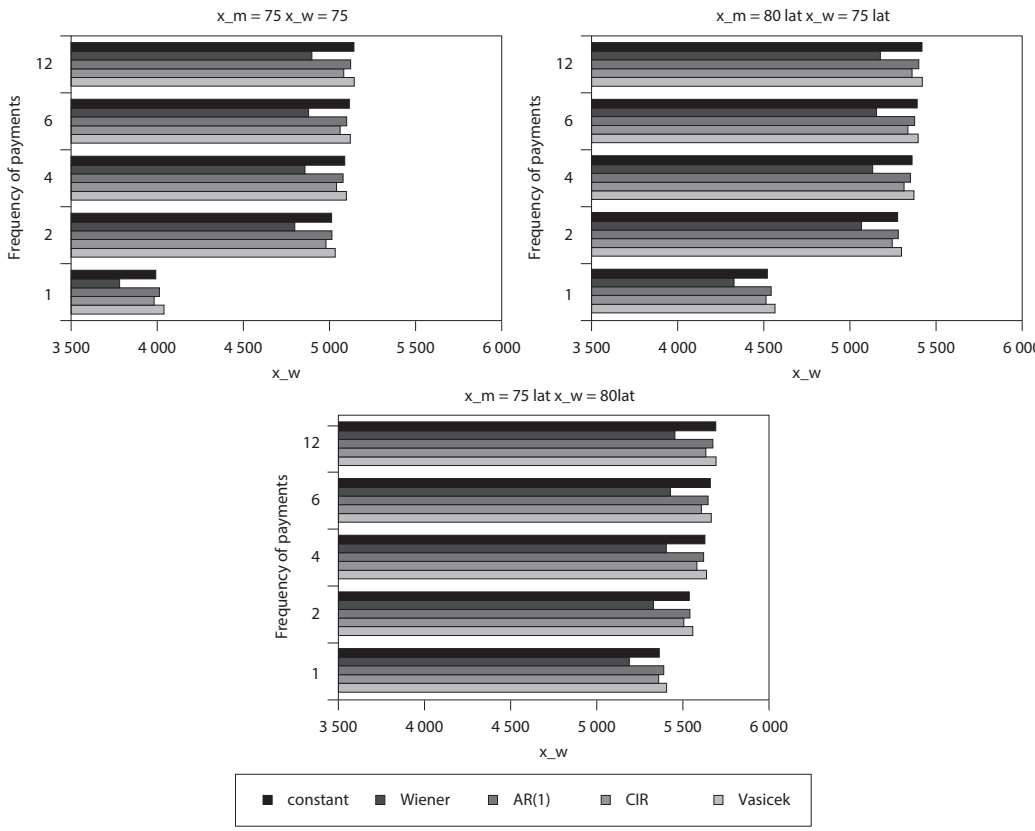
$$relativedifferences = \left(\frac{b(Vasicekmodel)}{b(Wienerprocess)} - 1 \right) \cdot 100\% \,. \tag{50}$$

Figure 4 The relative percentage differences between benefits in the case of the Wiener process and the Vasicek model



Source: Own elaboration

Figure 5 Annual annuity benefits for all models

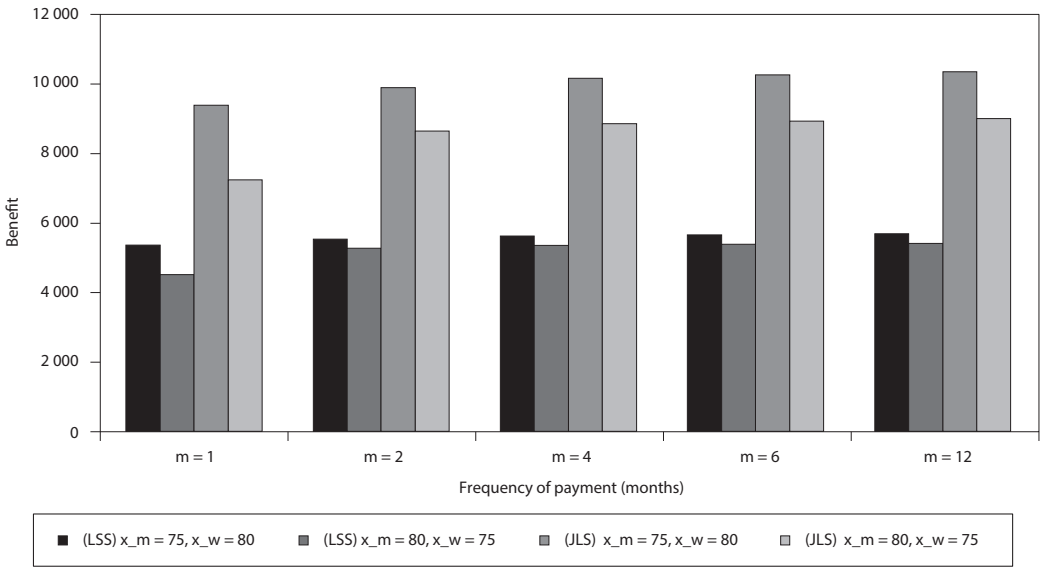


Source: Own elaboration

Figure 5 shows the differences in annuities for spouses of different ages in the case of LSS, when benefits are paid annually and more than once a year (m = 1, 2, 4, 6, 12).

For the joint-live status, the situation is similar, but the differences in benefit amounts are lower. It can be seen in Figure 6 for a fixed technical interest rate. Fogure 6 shows benefits for spouses of different ages for both statuses.

Figure 6 Annual annuity benefits for constant rate model



Source: Own elaboration

The differences in benefits between more frequent and annual payments are shown in Table 1.

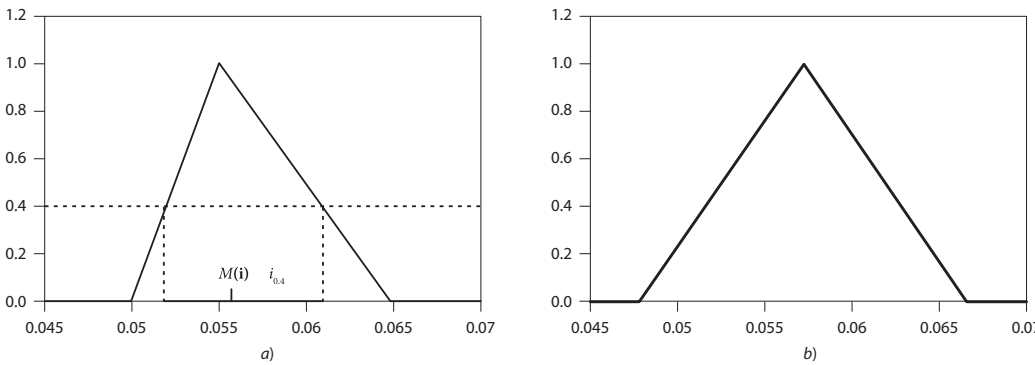
Table 1 The differences in benefits between more frequent and annual payments

	$m = 2$	$m = 4$	$m = 6$	$m = 12$
(LSS) $x\_m = 75, x\_w = 80$	3.25%	4.96%	5.54%	6.13%
(LSS) $x\_m = 80, x\_w = 75$	19.49%	22.38%	23.37%	24.37%
(JLS) $x\_m = 75, x\_w = 80$	5.42%	8.32%	9.30%	10.30%
(JLS) $x\_m = 80, x\_w = 75$	14.43%	17.57%	18.64%	19.72%

Source: Own elaboration

For LSS the differences range from 3.25% (m = 2) to 6.13% and for JLS (m = 12) from 5.42% (m = 2) to 10.3% (m = 12) when $x_m = 75$ and $x_w = 80$. When $x_w = 80$ and $x_m = 75$ the differences are significantly higher than for younger male and older female, but for JLS they are lower than for LSS (JLS: from 14.43% (m = 2) to 19.72% (m = 12); LSS: from 19.49% (m = 2) to 24.37% (m = 12)). The differences in benefits paid six or twelve times a year compared to benefits paid respectively four or six times a year are 0.56%

Figure 7 The membership functions of fuzzy interest rate **i** (*a*) and fuzzy discounting factor **v** (*b*)



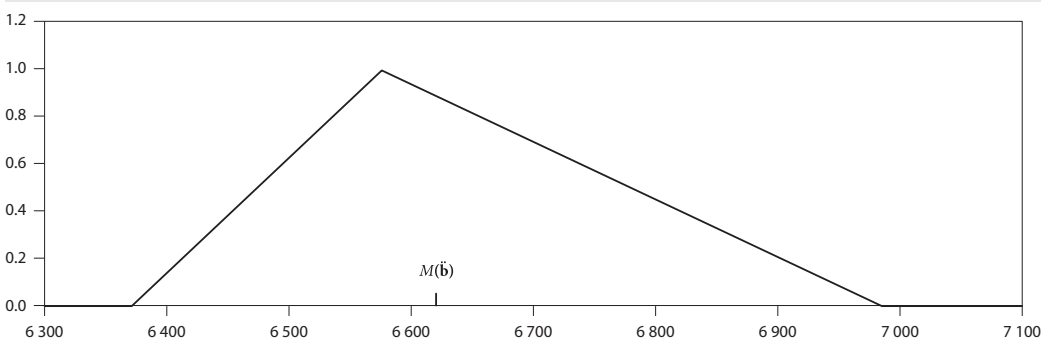
Source: Own construction

(LSS) and 0.91% (JLS) for younger men. The difference is slightly lower for younger women (0.54% and 0.81% respectively).

In Section 2.2.2 we computed the fuzzy interest rate **I** and the fuzzy discounting factor **v**. The graphs of such fuzzy numbers are presented in Figure 7.

Now, we derive the fuzzy annuity benefit $\ddot{\mathbf{b}}^{(m)}$. We assume that, man and woman are 65 years old, i.e. $x_m = x_w = 65$ and the joint-life annuity is paid monthly, i.e. $m = 12$ (compare Section 2.2). The fuzzy annual annuity benefit is the almost triangular fuzzy number (6 373.66, 6 576.82, 6 986.77) with mean value $M(\ddot{\mathbf{b}}^{(m)}) = 6\,628.52$ and imprecision $\mathrm{Imp}(\ddot{\mathbf{b}}^{(m)}) = 306.55$. The graph of the membership function of this fuzzy number and its mean value are included in Figure 8.

Figure 8 The membership function of fuzzy annual annuity benefit $\ddot{\boldsymbol{b}}^{(m)}$



Source: Own construction

Table 2 contains the parameters of the fuzzy annual annuity benefits, the mean values and imprecisions of such fuzzy numbers, when the spouses are the same age and when the husband is 75 years old and the wife is of different ages.

Table 2 The parameters of the fuzzy annual annuity benefits for JLS

x_m	*x_w*	*a*	*b*	*c*	Mean	Imp
The same age of spouses						
65	65	6 373.66	6 576.82	6 986.77	6 628.52	306.55
70	70	7 653.35	7 856.37	8 264.86	7 907.73	305.76
75	75	9 670.28	9 876.48	10 290.06	9 928.32	309.89
80	80	12 831.30	13 044.39	13 470.50	13 097.64	319.60
85	85	17 553.49	17 774.78	18 216.27	17 829.83	331.39
The different age of spouses						
75	65	8 352.05	8 558.78	8 974.11	8 610.93	311.03
75	70	8 805.96	9 011.84	9 425.23	9 063.72	309.64
75	75	9 670.28	9 876.48	10 290.06	9 928.32	309.89
75	80	11 175.06	11 384.13	11 802.79	11 436.53	313.87
75	85	13 533.87	13 747.27	14 173.91	13 800.58	320.02

Source: Own elaboration

4 DISCUSSION

In the first part of Section 3, we discussed benefits under assumptions that actuarial and financial models model interest rates. Regardless of the status of the contract, it turned out that the benefit amounts of the marital reversionary annuity for the Wiener process obtain the lowest benefits and the highest for the Vasicek model. Figure 3 and Figure 4 illustrate the fact that the benefits increase and the differences decrease with the rise of spouses' age. However, the payment structure in the two statuses is various. The differences between benefits are higher for the younger spouses and lower for the older spouses in the last surviving status and vice versa in the joint-life status.

Apart from the interest rate model, the frequency of benefit payments significantly impacts their amount. It turns out that there are differences in annuities for spouses of different age in the case of last surviving status when benefits are paid annually and more than once a year. It can be observed (com. Figure 5) that the benefits are lowest in the case of the Wiener process. In other cases, they are of a similar amount. The more often the benefit is paid, the higher the benefit is, but the differences are insignificant for the higher than two months frequency. The most significant differences are between an annual benefit and a benefit paid every two months. For the joint-live status, the situation is similar, but the differences in benefit amounts are lower (com. Figure 6 and Table 1).

The age difference between the spouses also affects the amount of benefits. For both statuses, the differences are significantly higher for an older husband and younger wife than for a younger male and older female, but for JLS they are lower than for LSS.

In the second part of Section 3, we derived the fuzzy annuity benefit. Firstly, we considered the situation that the spouses are the same age. It turned out that as the age of spouses increases, the value of the fuzzy annual annuity benefit increases, too. For older spouses, the increase is more significant (com. Figure 8). We can also observe a slight increase in imprecision. Secondly, we analysed the fuzzy annual annuity benefits when the husband is 75 years old, and the wife is of different age (com. Table 2). We obtained similar precision as in the case that spouses are equal in age.

To summarize, the benefit amounts are highest for the Vasicek model, slightly lower benefits are obtained for the AR(1) process. The autoregressive process of order one is the discrete equivalent of the Ornstein-Uhlenbeck process (Vasicek model). Therefore, similar results are obtained for both models.

The lowest benefit is obtained for the Wiener process. As the frequency of benefits increases, the annual benefit increases. The benefit obtained at a constant interest rate is similar to the Vasicek model. The number of payments per year affects the annual benefit obtained, with withdrawals more frequent than 12 times per year no longer significantly improving the benefit. As the age of spouses increases, the value of the fuzzy annual annuity benefit increases, too. For older spouses, the increase is more significant. We can also observe a slight increase in imprecision. The fuzzy rate has not yet been applied to the determination of reverse annuity benefits.

CONCLUSION

Equity release contracts have long been widely discussed. These contracts are addressed to older people. Europe's ageing population, and therefore a rising dependency ratio of retirees on the working population, strongly suggests that a pensions funding gap will be a key social issue in the future. Equity release contracts that can fill this gap. Many older people own property and in return for a monthly benefit, they could access using equity release products.

The article sets out the benefits of a marriage reverse annuity contract, which exists in Poland only in individual form. Usually, the spouses own their property. That is why marriage contracts are a natural research issue. Various interest rate models were used for this purpose. It was shown what effect have the different models of interest rate on the amount taking into account the different frequency of their payment. Attention has been focused only on net benefits. The determination of the interest rate is of great importance for the correct estimation of the sum of benefits, especially facing the increasing inflation. Let us note that in practice, insurers use various forms of indexation to protect benefits against the effects of inflation. In this area, the use of the fuzzy interest rate in the indexation mechanism may be an interesting and important issue for further research.

References

ANDRES-SANCHEZ, J., GONZALEZ-VILA PUCHADES, GONZALEZ-VILA PUCHADES, L. (2012). Using fuzzy random variables in life annuities pricing [online]. *Fuzzy Sets and Systems*, 188: 27–44. <https://doi.org/10.1016/j.fss.2011.05.024>.

ANDRES-SANCHEZ, J., GONZALEZ-VILA PUCHADES, L. (2017a). Some computational results for the fuzzy random value of life actuarial liabilities. *Iranian Journal of Fuzzy Systems*, 14: 1–25.

ANDRES-SANCHEZ, J., GONZALEZ-VILA PUCHADES, L. (2017b). The Valuation of life contingencies: a symmetrical triangular fuzzy approximation [online]. *Insurance: Mathematics and Economics*, 72: 83–94. <https://doi.org/10.1016/j.insmatheco.2016.11.002>.

ANDRES, J., TERCENO, A. (2017). Application of fuzzy regression in actuarial analysis [online]. *Journal of Risk and Insurance*, 70: 665–699. <https://doi.org/10.1046/J.0022-4367.2003.00070.X>.

BEEKMAN, J. A., FUELLING, C. P. (1990). Interest and mortality randomness in some annuities [online]. *Insurance Mathematics and Economics*, 9(2–3): 185–196. <https://doi.org/10.1016/0167-6687(90)90033-A>.

BEEKMAN, J. A., FUELLING, C. P. (1993). One Approach to Dual Randomness in Life Insurance. *Scandinavian Actuarial Journal*, 2: 173–182.

BELLHOUSE, D. R., PANJER, H. H. (1981). Stochastic modelling of interest rates with applications to life contingencies – part II [online]. *Journal of Risk & Insurance*, 48: 628–637. <https://doi.org/10.2307/252824>.

BEUTZEN, A., JIMENEZ, M., RIVAS, J. A. (1997). Actuarial mathematics with fuzzy parameters. An application to collective pension plans [online]. *Fuzzy Economic Review*, 2: 47–66. <https://doi.org/10.25102/fer.1997.02.04>.

BOWERS, N., GERBER, H., HICHMANN, J., JONES, D., NESBITT, C. (1986). *Actuarial mathematics.* Society of Actuaries.

BOYLE, P. P. (1976). Rates return as random variable. *Journal of Risk & Insurance*, 43(4): 693–713.

BROCKWELL, P. J., DAVIS, R. A. (1996). *Introduction to Time Series and Forecasting.* Springer-Verlag.

CARRIERE, J. F. (1999). No arbitrage pricing for life insurance and annuities. *Economics Letters*, 64: 339–342.

CARRIERE, J. F. (2004). Martingale valuation of cash flows for insurance and interest models [online]. *North American Actuarial Journal*, 8(3): 1–16. <https://doi.org/10.1080/10920277.2004.10596150>.

COOK, R. J., LAWLESS, J. F. (2018). *Multistate Models for the Analysis of Life History Data.* Chapman & Hall/CRC Monographs on Statistics and Applied Probability, CRC Press.

DĘBICKA, J. (2003). Moments of the cash value of future payment streams arising from life insurance contracts [online]. *Insurance: Mathematics and Economics*, 33(3): 533–550. <https://doi.org/10.1016/j.insmatheco.2003.07.002>.
DĘBICKA, J. (2012). Modelowanie strumieni finansowych w ubezpieczeniach wielostanowych. *Monografie i Opracowania Uniwersytetu Ekonomicznego we Wrocławiu*, 1st Ed. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, No. 204.
DĘBICKA, J. (2013). An approach to the study of multistate insurance contracts [online]. *Applied Stochastic Models in Business and Industry*, 29(3): 224–240. <https://doi.org/10.1002/asmb.1912>.
DĘBICKA, J., HEILPERN, S., MARCINIUK, A. (2020). Application of copulas to modelling of marriage reverse annuity contract [online]. *Prague Economic Papers*, 29(4): 445–468. <https://doi.org/10.18267/j.pep.745>.
DĘBICKA, J., MARCINIUK, A. (2014). Comparison of Reverse Annuity Contract and Reverse Mortgage on the Polish Market [online]. *17th AMSE Applications of Mathematics and Statistics in Economics Conference Proceedings.*, 55–64. <https://doi.org/10.15611/amse.2014.17.06>.
DENUIT, M., DHAENE, J., LE BAILLY DE TILLEGHEM, C., TEGHEM, S. (2001). Measuring the impact of a dependence among insured life lengths. *Belgian Actuarial Bulletin*, 1(1): 18–39.
DERRIG, R. A., OSTASZEWSKI, K. (1997). Managing the tax liability of a property liability insurance company [online]. *Journal of Risk and Insurance*, 64: 695–711. <https://doi.org/10.2307/253892>.
DERRIG, R. A., OSTASZEWSKI, K. (2004). *Fuzzy sets. In Encyclopaedia of Actuarial Science.* John Wiley & Sons, 2: 745–75.
DHAENE, J. (2000). Stochastic interest rates and autoregressive integrated moving average process. *Astin Bulletin*, 30(1): 123–140.
DICKSON, D. C., HARDY, M. R., WATERS, H. R. (2019). *Actuarial Mathematics for Life Contingent Risks.* 3rd Ed. Cambridge University Press.
DUBOIS, D., PRADE, H. (1980). *Fuzzy Sets and Systems: Theory and Applications.* Academic Press.
FREES, E. W. (1990). Stochastic life contingencies with solvency considerations [online]. *Proceedings of the 2nd Conference in Actuarial Science and Finance on Samos*, Karlovassi, Greece, 91–129. <http://www.stat.ucl.ac.be/samos2002/proceedsibillo.pdf>.
GARRIDO, J. (1988). Diffusion premiums for claim severities subject to inflation [online]. *Insurance Mathematics and Economics*, 7(2): 123–129. <https://doi.org/10.1016/0167-6687(88)90105-9>.
HABERMAN, S., PITACCO, E. (2018). Actuarial Models for Disability Insurance. In: *Actuarial Models for Disability Insurance*, Routledge.
HANEWALD, K., POST, T., SHERRIS, M. (2016). Portfolio Choice in Retirement-What is The Optimal Home Equity Release Product? [online]. *Journal of Risk and Insurance*, 83(2): pp. 421–446. <https://doi.org/10.1111/jori.12068>.
HEILPERN, S. (2018). Risk process with uncertain claims amount. In: HRONOVÁ S., PTÁČKOVÁ, V., ŠAFR, K., VLTAVSKÁ K. (eds.) *Proceedings of 21st International Scientific Conference AMSE Applications of Mathematics and Statistics in Economics*, Prague: University of Economics, Oeconomica Publishing House, 144–153.
HOEM, J. M. (1969). Markov Chain Models in Life Insurance [online]. *Blätter Der DGVFM*, 9(2): 91–107. <https://doi.org/10.1007/BF02810082>.
HOEM, J. M. (1988). The Versality of the Markov Chain as a Tool in the Mathematics of Life Insurance. *Transactions of the 23rd International Congress of Actuaries*, Helsinki, 171–202.
HOUGAARD, P. (2000). *Analysis of Multivariate Survival Data* [online]. Series: Statistics for Biology and Health, Springer-Verlag. <https://doi.org/10.1007/978-1-4612-1304-8_1>.
HUANG, T., ZHAO R., TANG, W. (2009). Risk model with fuzzy random individual claim amount [online]. *European Journal of Operational Research*, 192: 879–890. <https://doi.org/10.1016/j.ejor.2007.10.035>.
HUZURBAZAR, A. V. (2019). Modeling Time-To-Event Data Using Flowgraph Models [online]. *Advances on Methodological and Applied Aspects of Probability and Statistics*, CRC Press, 561–572. <https://doi.org/10.1201/9780203493212-31>.
JAKUBOWSKI, J., PALCZEWSKI, A., RUTKOWSKI, M., STETTNER, Ł. (2003). *Matematyka finansowa. Instrumenty pochodne.* WNT.
JAMES, J., WEBBER, N. (2000). *Interest rate modelling.* John Willey & Sons Ltd.
KELLISON, S. G. (2009). *The theory of interest.* 3rd Ed. Homewood.
LAMAIRE, J. (1990). Fuzzy Insurance [online]. *Astin Bulletin*, 192: 33–55. <https://doi.org/10.2143/AST.20.1.2005482>.
MARCINIUK, A. (2004). Składki ubezpieczeń na życie ze świadczeniem płatnym na koniec podokresu roku śmierci ubezpieczonego. In: OSTASIEWICZ, W. (eds.) *Zastosowania statystyki i matematyki w ekonomii*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu.
MARCINIUK, A. (2009). *Modele stóp procentowych i ich zastosowania w ubezpieczeniach.* Master thesis, Uniwersytet Ekonomiczny we Wrocławiu.
MARCINIUK, A. (2016). Małżeńska renta hipoteczna uwzględniająca zależność między przyszłym czasem trwania życia małżonków. *Studia Ekonomiczne*, 114–132.
MARCINIUK, A. (2017). Marriage Reverse Annuity Contract and Reverse Mortgage – Application of a Generalized Model of Reversionary Annuity [online]. In: GARDOŃ, A., KOZYRA, C., MAZUREK, E. (eds.) *Applications of Mathematics and Statistics in Economics 2017 Conference Proceedings*, Wrocław: Wrocław University of Economics Press, 297–306. <https://doi.org/10.15611/amse.2017.20.24>. ISBN 978-83-7695-693-0

MARCINIUK, A. (2021). Equity Release Contracts with Varying Payments [online]. *Prague Economic Papers*, 30(5): 552–574. <https://doi.org/10.18267/j.pep.784>.
MARCINIUK, A., ZIMKOVÁ, E., FARKAŠOVSKÝ, V., LAWSON, C. W. (2020). Valuation of equity release contracts in Czech Republic, Republic of Poland and Slovak Republic [online]. *Prague Economic Papers*, 29(5): 505–521. <https://doi.org/10.18267/j.pep.743>.
MUSIELA, M., RUTKOWSKI M. (1988). *Martingale Methods in Financial Modelling.* Springer.
NORBERG, R. (2002). *Basic Life Insurance Mathematics* [online]. Manuscript of Book. <http://web.math.ku.dk/~mogens/lifebook.pdf>.
OSTASZEWSKI, K. (1993). *An Investigation Into Possible Applications of Fuzzy Sets Methods in Actuarial Science.* Actuarial Research Clearing House.
PANJER, H. H., BELLHOUSE, D. R. (1980). Stochastic modelling of interest rates with application to life contingencies [online]. *The Journal of Risk and Insurance*, 47(1): 91–110. <https://doi.org/10.2307/252684>.
PARKER, G. (1994a). Moments of the present value of a portfolio of policies [online]. *Scandinavian Actuarial Journal*, 1: 53–67. <https://doi.org/10.1080/03461238.1994.10413929>.
PARKER, G. (1994b). Two Stochastic Approaches for Discounting Actuarial Functions [online]. *Astin Bulletin*, 24(2): 167–181. <https://doi.org/10.2143/AST.24.2.2005063>.
SHAO, A. W., HANEWLAD, K., SHERRIS, M. (2015). Reverse mortgage pricing and risk analysis allowing for idiosyncratic house price risk and longevity risk [online]. *Insurance: Mathematics and Economics*, 63: 76–90. <https://doi.org/10.1016/j.insmatheco.2015.03.026>.
SHAPIRO, A. F. (2004). Fuzzy logic in insurance [online]. *Insurance: Mathematics and Economics*, 35: 399–424. <https://doi.org/10.1016/j.insmatheco.2004.07.010>.
SHAPRIO, A. F. (2013). Modeling future lifetime as a fuzzy random variable [online]. *Insurance: Mathematics and Economics*, 53: 864–870. <https://doi.org/10.1016/j.insmatheco.2013.10.007>.
SPIERDIJK, L., KONING, R. H. (2011). *Calculating Loss Reserves in a Multistate Model for Income Insurance* [online]. <https://ssrn.com/abstract=1871229> or <http://dx.doi.org/10.2139/ssrn.1871229>.
VASICEK, O. (1999). An equilibrium charakterization of the term structure [online]. *Journal of Financial Economics*, 3(3): 177–188. <http://doi.org/10.1016/0304-405X(77)90016-2>.
WOLTHUIS, H. (1994). *Life insurance mathematics (the Markovian model).* Bruxelles: CAIRE Education Series.
ZADEH, L. A. (1965). Fuzzy sets [online]. *Information and Control*, 8: 338–353. <https://doi.org/10.2307/2272014>.
ZMYŚLONA, B., MARCINIUK, A. (2020). Financial Protection for the Elderly – Contracts Based on Equity Release and Critical Health Insurance [online]. *European Research Studies Journal*, XXIII(Special Issue 1): 867–882. <https://doi.org/10.35808/ersj/1798>.

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

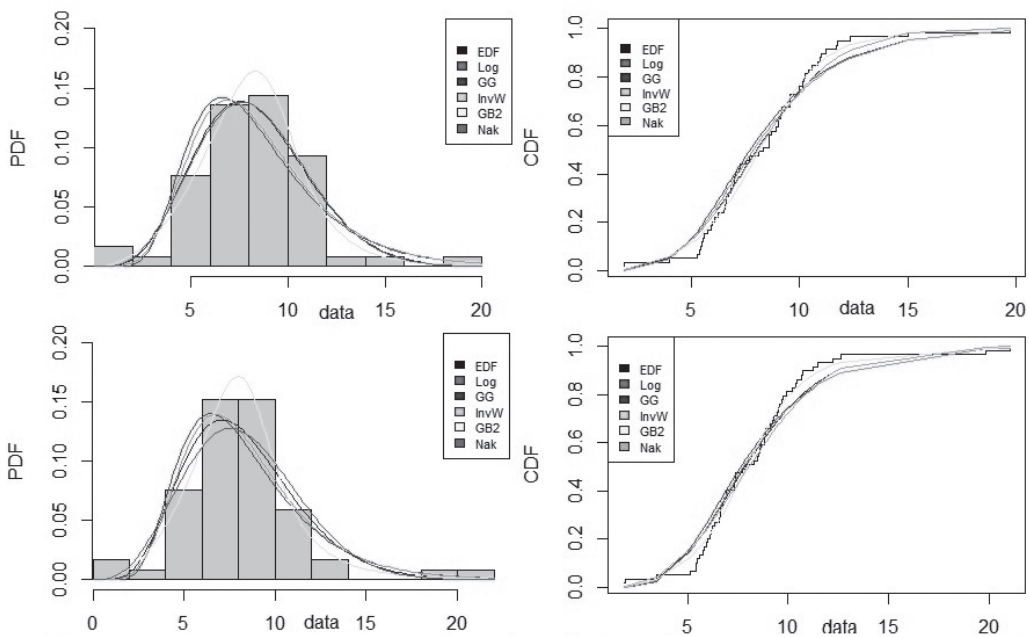
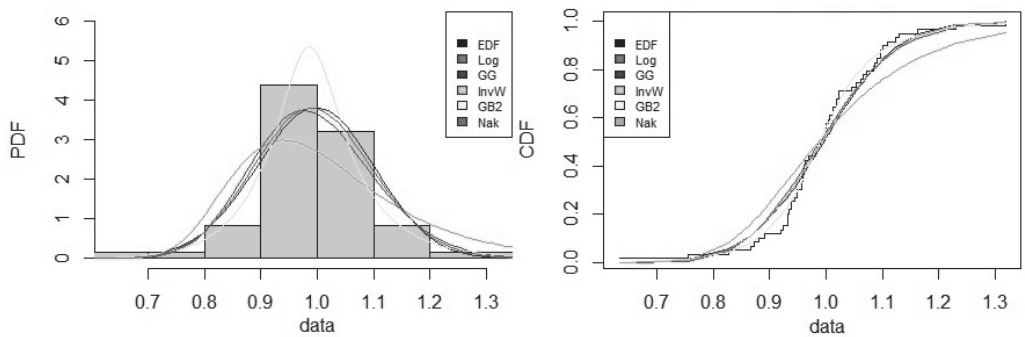


Figure 1 (continuation)



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

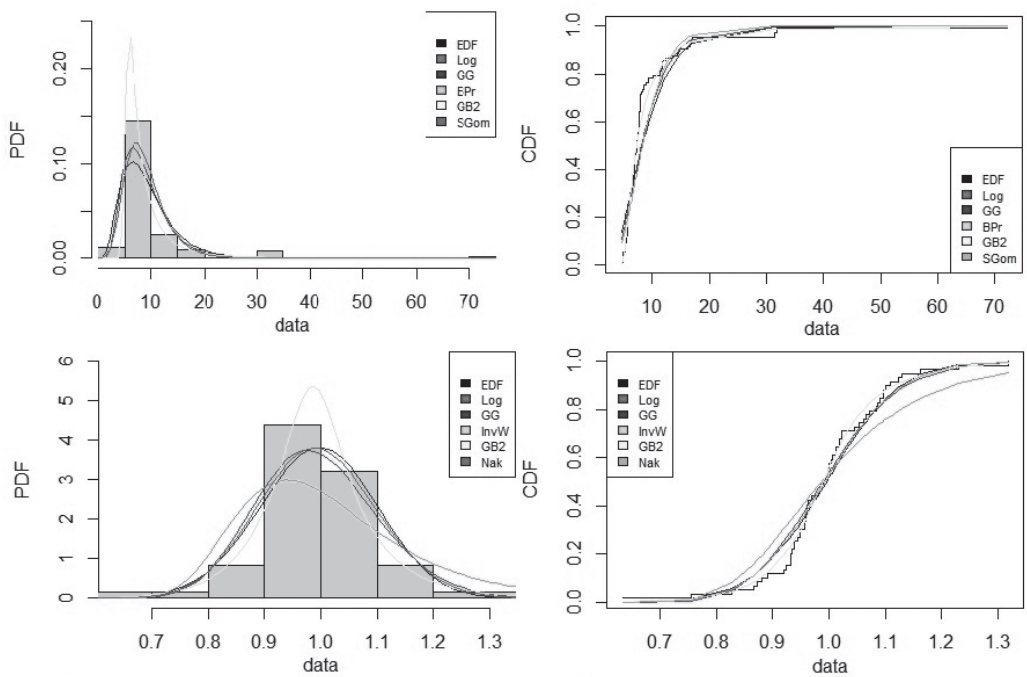
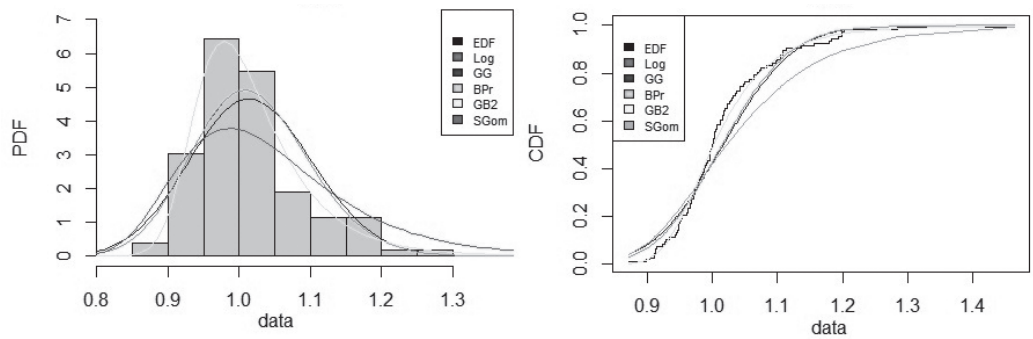


Figure 2 (continuation)



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

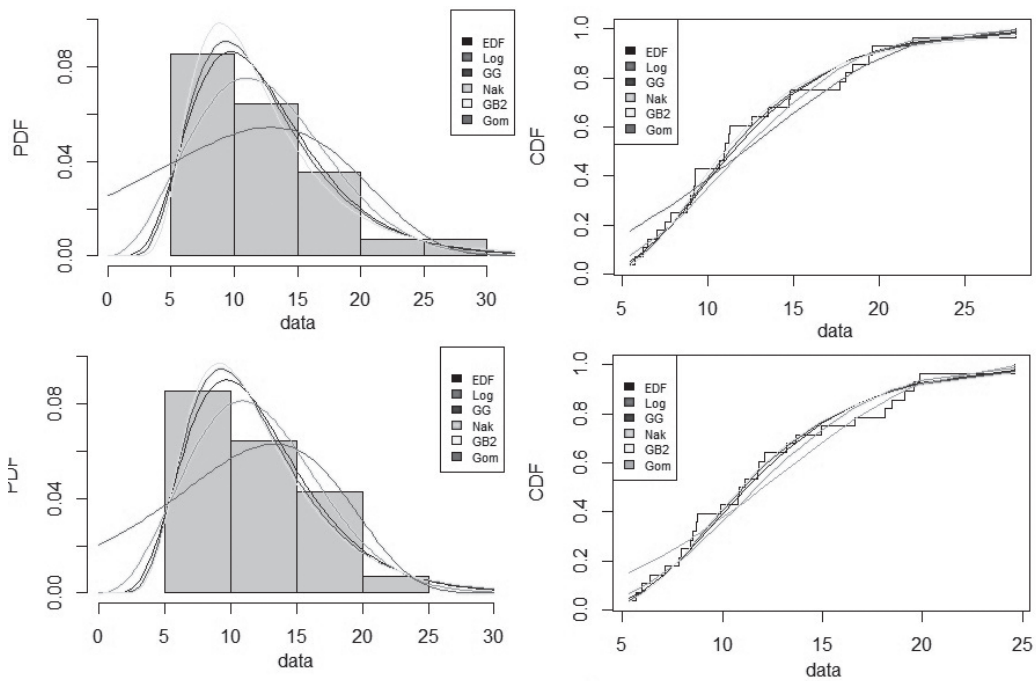
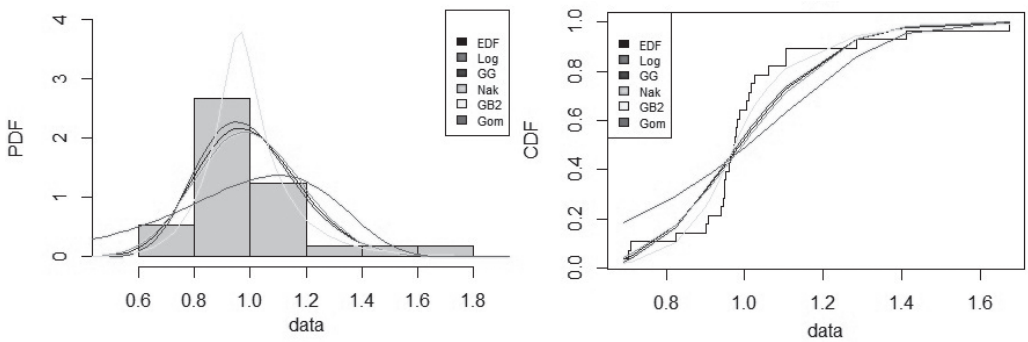


Figure 3 (continuation)



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

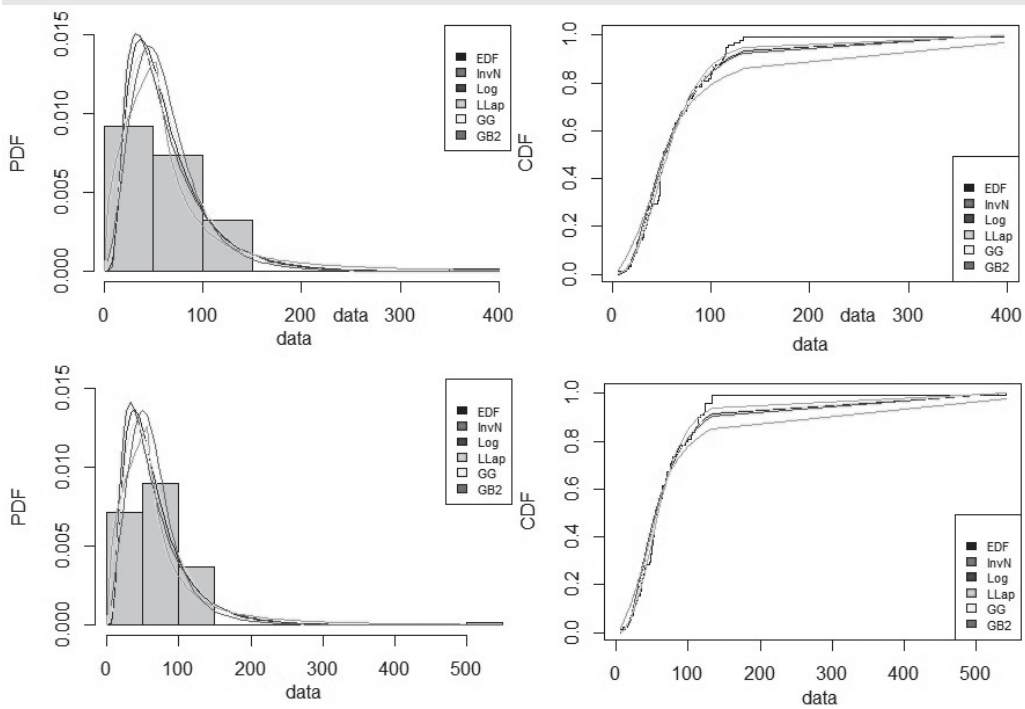
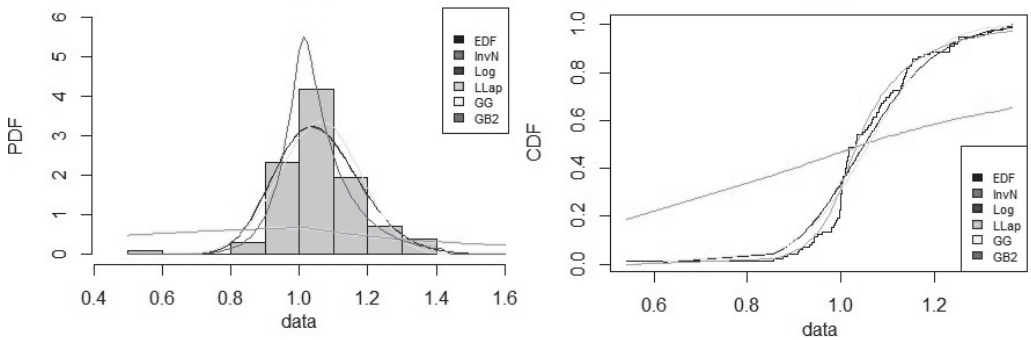


Figure 4 (continuation)



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.

Population Census Microdata Availability

Jiří Novák[1] | *Prague University of Economics and Business, Prague, Czech Republic*

Received 12.12.2021 (revision received 30.3.2022), **Accepted** (reviewed) 12.4.2022, **Published** 16.9.2022

Abstract

In this paper, the author aims to describe the dissemination of microdata from the population census by National Statistical Offices. This type of data is highly confidential, and approaches to protection vary across the world. National Statistical Offices mostly strive to publish their data as much as possible, but they are bounded by national and international laws to protect the personal data of respondents. The primary goal is mapping the differences between countries and their categorization. Different approaches to microdata availability are described, and various data access approaches are depicted. The information was obtained from publicly available documentation and a survey in which selected statistical offices were contacted. Discovered were that of the 223 countries (including dependent territories), 100 countries have made microdata available for the scientific community, with 30 countries also providing microdata access to the public. This paper presents a mapped overview and aggregated information on the publication of microdata of the population census from around the world.

Keywords	***DOI***	***JEL code***
Microdata, population census, statistical disclosure control, confidentiality, public use files, scientific use files	*https://doi.org/10.54694/stat.2021.44*	*C80*

INTRODUCTION

A growing pressure from the research community, policymakers and citizens has been observed for publishing more data in increasing detail. Producers of the statistic, specifically National Statistical Offices (NSO) or other statistical agencies, stand on the other side. They are bound by national and international laws protecting personal data and keeping the trust of their respondents. The confidentiality of the data is essential to all statistical offices and agencies because respondents and NSO's can trust each other only if the trust is maintained, which increases the quality of the data obtained from individual respondents.

However, the dissemination of census data has many positives, and their subsequent analysis has many benefits that aim to improve the social well-being and everyday lives of ordinary people, who mostly do not even know that it was the census data used to improve their surroundings. The population census is a unique statistical survey that makes it possible to obtain essential information that cannot usually be obtained in any other way. Census microdata can be used for research and planning in areas such as health, fertility, housing, transportation, education, employment, migration and regional development.

[1] Department of Economic Statistics, Prague University of Economics and Business, W. Churchill Sq. 1938/4, 130 67, Prague 3, Czech Republic. E-mail: jiri.novak.kest@vse.cz, phone: (+420)224095435.

For example, as stated in Černý (2021a), NSO's obtain information about population distribution, which in European Union censuses is based on usual residence (the place where the person actually resides, regardless of where he is domiciled). Not only that, NSO's further obtain information about education with other socio-demographic indicators, such as the place of residence, size of the dwelling, number of household members, etc.

In the following article by Černý (2021b), several projects were presented, which have used census data to improve the state of affairs. To give an idea of the benefits of census data for the city, one of the projects was the Transport model of the city of České Budějovice. In the project, the authors created a mathematical model of automobile transport in České Budějovice and its surroundings. The model then enabled them to plan municipal transit/transport, assess changes in the organization of transit/transport and forecast car transport or evaluate the impact of new investments. This model included data that cannot be obtained from a source other than the census, such as information concerning commuting to work or the actual number of people in the surrounding area.

There are many approaches how to make microdata accessible by NSO. It can be accessed via Data Laboratories, Remote Access Facilities or are provided as products for use outside the NSO. Disseminated microdata can be categorized into two groups based on intended. The first group is microdata that the general public can access, and the second one is microdata that can access only by approved researchers. These groups do not have a fixed nomenclature, as it varies by region. In the European region is used term Public Use Files (PUF) for the first group and the Scientific use files (SUF) for the second. Other regions are described further in the article.

In this research paper, the author aims to examine the approaches taken by the individual European National/Central Statistical Offices and other statistical agencies around the world in relation to the publication of microdata. The main goal is to compare specific approaches between the individual countries (European and non-European). The primary interest of this work is whether the statistical offices publish microdata at all and, if so, what approach and models they chose to protect the personal data of their respondents. The information has been obtained from publicly available documentation and from a survey sent to selected statistical offices.

1 LEGISLATION AND LITERATURE

Microdata are data that contain information about an individual person, household, business, or other entity (Templ et al., 2014). Usually, they are collected directly by NSO, or they can be obtained from the administrative sources or surveys. Raw data of this character are highly sensitive on disclosure of confidential information, and in case of release to the scientific community or sometimes even to public, there have to be special measures applied to secure data and protect personal information. In the case of a population census, microdata is the data that contains information about individual households and housing units (United Nations 2017), who are located in a specific territory.

According to United Nations (2007) every official statistical system should support any research based on microdata. The following benefits are defined, which result from more accessible publishing of micro-data:

i. *microdata permits policy makers to pose and analyse complex questions. In economics, for example, analysis of aggregate statistics does not give a sufficiently accurate view of the functioning of the economy to allow analysis of the components of productivity growth;*
ii. *access to microdata permits analysts to calculate marginal rather than just average effects. For example, microdata enable analysts to do multivariate regressions whereby the marginal impact of specific variables can be isolated;*
iii. *broadly speaking, widely available access to microdata enables replication of important research;*
iv. *access to microdata for research purposes, and the resulting feedback, can facilitate improvements in data quality. For example, the US Bureau of the Census has formalised the documentation it requires from researchers to assist it in improving the quality of its surveys;*

v. *it increases the range of outputs derived from statistical collections and hence the overall value for money obtained from these collections.*

International organizations such as United Nations (UN) and European Union (EU) support the publication of microdata in compliance with strict privacy of personal data.

United Nations defined this approach in their Fundamental Principles of Official Statistics (United Nations, 2015a) as a sixth principle – Confidentiality, which states that: individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes. This principle is followed by United Nations (2007), which sets out further principles for the handling of confidential microdata. The mentioned principles are as follow:

Principle 1: It is appropriate for microdata collected for official statistical purposes to be used for statistical analysis to support research as long as confidentiality is protected.

Principle 2: Microdata should only be made available for statistical purposes.

Principle 3: Provision of microdata should be consistent with legal and other necessary arrangements that ensure that confidentiality of the released microdata is protected.

Principle 4: The procedures for researcher access to microdata, as well as the uses and users of microdata, should be transparent and publicly available.

European Union defined this approach in their European Statistics Code of Practice (Eurostat, 2018) as a fifth principle – Statistical Confidentiality and Data Protection, which states that: *the privacy of data providers, the confidentiality of the information they provide, its use only for statistical purposes and the security of the data are absolutely guaranteed.* This is more specified in Eurostat (2018) by the following subpoints:

5.1 Statistical confidentiality is guaranteed in law.

5.2 Staff sign legal confidentiality commitments on appointment.

5.3 Penalties are prescribed for any wilful breaches of statistical confidentiality.

5.4 Guidelines and instructions are provided to staff on the protection of statistical confidentiality throughout the statistical processes. The confidentiality policy is made known to the public.

5.5 The necessary regulatory, administrative, technical and organisational measures are in place to protect the security and integrity of statistical data and their transmission, in accordance with best practices, international standards, as well as European and national legislation.

5.6 Strict protocols apply to external users accessing statistical microdata for research purposes.

Efforts to make microdata available are not in conflict with UN principles and the EU Code as long as the data are used for statistical purposes (what is the statistical purpose is described below) and it is prevented that individual data on respondents can be identified from the published data.

1.1 European legislation

Microdata publications have to be supported by corresponding legislation as is set out in the principles above. The protection of personal data is usually already covered by national law, and in this paper, when examining the approach of national offices to publishing microdata, the author also recorded the relevant laws that cover this issue. The relevant laws of individual countries can be found in the results section, if information about the laws was available.

In European Union, the national statistical confidentiality is provided for in the EU legislation. The principal regulations on the EU level that cover statistical confidentiality and protection of personal information are the following regulations.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data also known as General Data Protection Regulation (GDPR) represents an EU-wide

legal framework for the protection of personal data, which protects the rights of its citizens against the unauthorized treatment of their data and personal data.

In Recital 162 of GDPR (Regulation (EU) 2016/679), there are defined statistical purposes as *"any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results" and that "the statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person"*. This is a crucial sentence that guarantees to all persons whose data is used in the microdata analysis that the results of the analyses will not be used against them. The result of microdata analysis should not be information that could lead to measures against certain individuals but rather aggregated information from which can a population benefit as a whole. The researcher should not be interested in information about the persons contained in the data, but in the structure and links hidden in the data, which is information that may be lost by pre-processing of the data when the researcher has available only an aggregated dataset. In the recital is also stated that "where personal data are processed for statistical purposes, this Regulation should apply to that processing" and that "Union or member state law should, within the limits of this Regulation, determine statistical content, control of access, specifications for the processing of personal data for statistical purposes and appropriate measures to safeguard the rights and freedoms of the data subject and for ensuring statistical confidentiality".

GDPR Regulation (EU) 2016/679 lays down rules for the publication in Article 6 Lawfulness of processing, where is stated that *Processing shall be lawful only if and to the extent that at least one of the following applies:*

a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes;
b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;
c) processing is necessary for compliance with a legal obligation to which the controller is subject;
d) processing is necessary in order to protect the vital interests of the data subject or of another natural person;
e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;
f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

Sub-point c), which refer to compliance with a legal obligation, is a crucial legal instrument for publishing microdata. The legal obligation is represented by other legislation, that has specified the rights and obligations for the handling of personal data in the given case specified by law. The legislation that specifies this is a regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics as amended by Regulation (EU) No 2015/759, which then describes in detail the handling of data for statistical purposes. This is also confirmed in GDPR by Recital 163.

In Recital 26 of (Regulation (EC) No 223/2009) is acknowledged that *the research community should enjoy wider access to confidential data used for the development, production and dissemination of European statistics, for analysis in the interest of scientific progress in Europe. Access to confidential data by researchers for scientific purposes should therefore be improved without compromising the high level of protection that confidential statistical data require.* In Article 19 Public use files (PUFs) is then further specified that data on individual statistical units, disseminated in the form of a public use file consisting of anonymised records which have been prepared in such a way that the statistical unit cannot be identified,

either directly or indirectly, when account is taken of all relevant means that might reasonably be used by a third party. In Article 23 Access to confidential data for scientific purposes, which describes Scientific use files (SUFs) is stated that access to confidential data which only allow for indirect identification of the statistical units may be granted to researchers carrying out statistical analyses for scientific purposes by the Commission (Eurostat) or by the NSIs or other national authorities, within their respective spheres of competence. (Regulation (EC) No 223/2009).

EU legislation not only provides legal tools for publishing microdata, in its regulations, it also encourages data to be made available to the scientific community, while maintaining strict confidentiality in the data and under the conditions of examining the data as a whole. Each member country has its own approach to the protection of microdata. There is no regulation at the European level that prescribes protection procedures. However, there is inter-European coordination, where experiences are exchanged within a working group and expert group on statistical disclosure control.

1.2 Statistical disclosure control

The set of methods focusing on the protection of disseminated data are called Statistical Disclosure Control (SDC). SDC is the term most used in the European region, but in other regions are used slightly different terms like Statistical disclosure limitation or Disclosure avoidance. Every statistic released from NSO is controlled by these methods to protect statistical confidentiality.

There are many approaches to publishing data and according to United Nations (2007) they can be categorized as follows:

(i) Statistical products for use outside the NSO.

– Statistical Tables which can be standardized or specially generated on specific request of the researcher. Data Cubes which are flexible multi-dimensional matrices that allow to researcher to create required table on their own.

– Public Use Files (PUF) are microdata that can be accessed by general public, usually after registration, and because of that the highest level of security against disclosure of respondents' personal data is necessary. PUF is the most used expression in the European region; its definition is codified by Regulation EC No 223/2009 2015. In the USA, PUF's are called Anonymised Microdata Files.

– Scientific use files (SUF) are microdata that can access only approved researchers, usually is signed contract. This type of microdata needs lower level of security than PUFs, but the personal information of respondents is still preserved against disclosure. SUF is term most used in the European region, its definition is codified by Regulation EC No 223/2009 2015. In USA is SUF called Licensed Anonymised Microdata Files.

(ii) A service window through which researchers can submit data requests.

– Remote Access Facilities which means that researchers could create their statistical tasks remotely from microdata through computer networks.

(iii) Arrangements for allowing researchers to work on the premises of the National Statistical Office.

– Data Laboratories also known as Data Centers or Safe Centers, which allows researchers to access microdata in secure place at NSO and because of that can allow to researchers work with more detailed data.

The methods designed for protecting microdata are in detail described in Hundepool et al. (2010) and Hundepool (2012). Microdata protection can be categorized into three main groups: Non-perturbative microdata masking, Perturbative microdata masking, Synthetic and hybrid data. A general overview of these methods is as follows:

(i) Non-perturbative microdata masking is based on reduction of detail in data. Examples of t methods in this category are Sampling, Global recoding, Top and bottom coding, Local suppression. In Sampling is published only a sample of the original dataset. In Global recoding are generalized categories, which means creating new more general categories. Top and bottom coding is global recoding for ranked data.

Top values or bottom values are merged together, which means creating new more general top or bottom category. Local suppression means replacing risk values with missing values.

(ii) Perturbative microdata masking is based on perturbating data, which means distortion of the original data. Examples of the methods in this category are: Noise masking, Microaggregation, Data swapping and Rounding. Noise masking is based on adding noise into the data. It is a broad category with many approaches to adding noise. For example, normally distributed errors are added into the original data. Microaggregation is a broad category of methods for continuous microdata. Those methods are based on replacing individual values with values computed on small aggregates. This approach creates groups of k or more individuals, where no individual dominates the group and k is a threshold value. Data swapping is based on exchanging values of confidential variables between individual records. In Rounding are original values replaced with rounded values.

(iii) Synthetic and hybrid data are methods based on generating new data with the preservation of certain statistics or internal relationships of the original data set. Methods in this category are Fully synthetic data, Partially synthetic data and Hybrid data. In Fully synthetic data a completely new dataset which does not contain the original data is released. In Partially synthetic data only the most sensitive data are generated, and the rest of the original data are kept in the dataset. Hybrid data are the original data and synthetic data combined together.

1.3 Census recommendations

This paper is focused on publication of microdata outputs from the population census. A comprehensive description of the methodology of the population census is provided in the United Nations (2017) and United Nations (2015a). Papers describing the content, methodology and preparation of the Census 2021 in the Czech Republic are following: Čtrnáct (2016), Sudková (2016), Škrabal et al. (2016), Škrabal (2017), Báčová (2018), Moravec (2018), and Kozelek (2019).

The recommendations for confidentiality and security of census data are described in United Nations (2006) and United Nations (2015a). The general aim of the census is to collect information on each person, household, and dwelling. The purpose is to provide information about the population, which can then be used to improve life in the country. In the use of these data, NSO are not so interested in information about each individual, but they gain valuable information about the structure of the population. However, in order to obtain reliable results, the confidentiality of the information collected about the respondents must be guaranteed. The security rules around the census carry out the entire operation from the actual data collection to its processing and subsequent publication to both the scientific community and the general public. The census is a unique project that allows the state to obtain detailed data down to the level of very little geographical detail.

In order to release the data obtained from the census, SDC methods have to be applied to all outputs, which prevent the disclosure of sensitive information about specific respondents. The goal of statistical protection of data confidentiality is to ensure maximum data utility while minimizing data information loss and maximizing the protection of published data. In case of the publishing of census microdata, there must be always removed direct identifiers such as name, addresses and personal identification numbers. Any unique variables, which could cause re-identification of any potential respondent, have to be perturbed by SDC methods, in order to minimize the risk of their disclosure. In the case of the census, the public confidence in the security and confidentiality of the information is primary and all operations are carried out to respect this precondition (United Nations, 2006; United Nations, 2015a).

1.4 International organizations

Several international organizations are interested in the field of microdata dissemination. Mostly it is a recommendation of methods and procedures how microdata would be published. Important

documents of United Nations and European Union are described above. Another relevant organization is OECD, which created expert group for international collaboration on microdata access. In OECD (2014) is their summary of the methods, practices, and recommendations. Two important organizations are further described, which strive for the widest possible dissemination of microdata from population census, IPUMS and the Pacific Community.

1.4.1 IPUMS

IPUMS (Integrated Public Use Microdata Series) is an organization which is mainly focusing on inventory, preservation, harmonization, and disseminating of census microdata from countries around the world. IPUMS is a part of the Institute for Social Research and Data Innovation at the University of Minnesota. According to their website,[2] they collaborate with 105 national statistical agencies, 9 national archives, and 3 genealogical organizations, which according to them created the world's largest database of census microdata. This database contains U.S. censuses from 1790 to the present and other international censuses of over 100 countries. IPUMS claims that the library has over a billion records. They stored in their library microdata describing 1.4 billion individuals which come from over 450 censuses and surveys.

Countries listed in Table 1 to Table 5 are the countries that can be found in their library and for clarity they were categorized geographically by continent.

Table 1 Datasets of IPUMS in Europe

Austria	Belarus	Finland	France	Germany
Greece	Hungary	Ireland	Italy	Netherlands
Poland	Portugal	Romania	Russia	Slovakia
Slovenia	Spain	Switzerland	Ukraine	United Kingdom

Source: Own construction

Table 2 Datasets of IPUMS in America

Argentina	Bolivia	Brazil	Canada	Chile
Colombia	Costa Rica	Cuba	Dominican Republic	Ecuador
El Salvador	Guatemala	Haiti	Honduras	Jamaica
Mexico	Nicaragua	Panama	Paraguay	Peru
Puerto Rico	Saint Lucia	Suriname	Trinidad and Tobago	United States of America
Uruguay	Venezuela			

Source: Own construction

Table 3 Datasets of IPUMS in Asia

Armenia	Bangladesh	Cambodia	China	Indonesia
Iran	Iraq	Israel	Jordan	Kyrgyzstan
Laos	Malaysia	Mongolia	Myanmar	Nepal
Pakistan	Palestine	Philippines	Thailand	Turkey
Vietnam				

Source: Own construction

[2] *<https://ipums.org>*.

Table 4 Datasets of IPUMS in Africa

Benin	Botswana	Burkina Faso	Cameroon	Egypt
Ethiopia	Ghana	Guinea	Kenya	Lesotho
Liberia	Malawi	Mali	Mauritius	Morocco
Mozambique	Rwanda	Senegal	Sierra Leone	South Africa
South Sudan	Sudan	Tanzania	Togo	Uganda
Zambia	Zimbabwe			

Source: Own construction

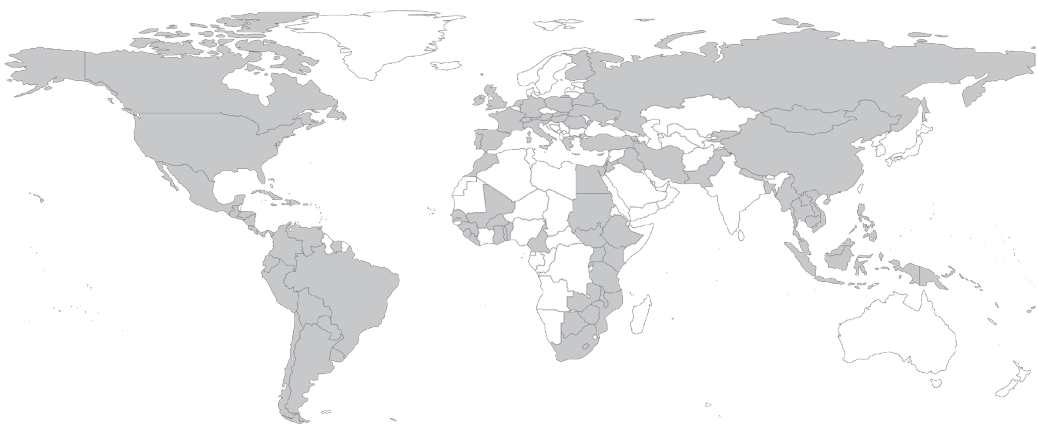
Table 5 Datasets of IPUMS in Oceania

Fiji	Papua New Guinea

Source: Own construction

Table 30, which was placed in the Annex due to its size, then shows the complete range of microdata available to IPUMS library[3] on 1/3/2022. Picture 1 shows on a world map the countries that are represented in their database. The full-size map can be downloaded from the author's GitHub.[4] The differences between the table and the map are the countries Iceland, Norway, Sweden and Denmark, because these countries provide only historical values from the years before 1960.

Picture 1 Map of member states of IPUMS



Source: Own construction

According to Sobek and Cleveland (2020) confidentiality of the data is ensured by application of suppression to very small categories and also application of top- and bottom-code thin tails of continuous variables. At geographical level they swap a small number of households to add an additional degree of uncertainty and for areas with fewer than 20 000 population in recent censuses are combined together with neighbouring units until the threshold is achieved.

[3] *<https://international.ipums.org/international-action/samples>*.
[4] *<https://github.com/Kyoshido/Population-census-microdata-availability/blob/main/map_IPUMS_03-2022.png>*.

Microdata are available at their website,[5] but the World Bank library of microdata offers a better search environment at their website.[6] The microdata are free of charge, but the researcher must submit an application to use restricted microdata via an electronic authorization form identifying the user by name, electronic address, and institutional affiliation.

1.4.2 The Pacific Community

The Pacific Community (SPC) is the scientific and technical organisation in the Pacific region. SPC division, which is responsible for the dissemination of the statistics data, is Statistics for Development Division[7] (SDD). SDD has provided the census support for its members since the 1990's and their technical support covers all aspects of the census cycle, from the questionnaire design and preparation of census cartography, to training of field staff, data processing, tabulation, analysis, and reporting and dissemination of results.

Member States of the pacific community. which are geographically located in Oceania are listed in Table 6.

Table 6 Member states of The Pacific Community

Australia	American Samoa	Cook Islands	Federated States of Micronesia	Fiji
French Polynesia	Guam	New Zealand	Kiribati	Mariana Islands
Marshall Islands	Nauru	New Caledonia	Niue	Palau
Papua New Guinea	Pitcairn Islands	Samoa	Solomon Islands	Tokelau
Tonga	Tuvalu	Vanuatu	Wallis and Futuna	

Source: Own construction

The microdata are available on their website only if the country signs a Memorandum of Understanding and gives SPC the right to publish their data sets. In the case of census microdata, all datasets in the microdata library are set as No Access. Because SPC helped with the collection of the microdata, they include them in the list in the library and also hold the dataset, but SPC don't disseminate them. Researchers have to contact the country National Statistics office to get those data.

2 SURVEY ON THE PRACTICES

There are plenty of approaches on how to make microdata available. A comprehensive list of recommended ways can be found at United Nations (2007). Generally, the publication of microdata can be divided based on the audience for which it is intended, into PUF and SUF microdata. Public Use Files (PUFs) in the USA called Anonymised microdata files, are microdata that can be accessed by the general public. That is the reason why the highest level of security against disclosure of individual respondent personal data is necessary. Scientific use files (SUF) in the USA called Licensed anonymised microdata files, are microdata that can be accessed only by approved researchers, usually with a signed contract. This type of microdata needs a lower level of security than PUFs, but the personal information of respondents is still preserved against disclosure.

In order to find out the approaches of individual statistical offices the survey and following research were made, in which author was focused on the current situation and the plans following the data

[5] *<https://international.ipums.org/international>*.
[6] *<https://microdata.worldbank.org/index.php/catalog/ipums>*.
[7] *<https://sdd.spc.int>*.

protection in the European and non-European countries. Mentioned information is based on author's research and the information from the official websites. A survey (questions are in the Annex) was created and then sent to the countries, which are shown in Table 7 to Table 10 and are categorized by continent. The following countries were chosen to represent the European approach with the addition of important world statistical offices.

Table 7 Surveyed countries of Europe

Albania	Austria	Belgium	Bulgaria	Croatia
Cyprus	Czech Republic	Denmark	Estonia	Finland
France	Germany	Greece	Hungary	Ireland
Israel	Italy	Latvia	Lithuania	Luxembourg
Malta	Montenegro	Netherlands	Poland	Portugal
North Macedonia	Romania	Russia	Serbia	Slovakia
Slovenia	Spain	Sweden	Switzerland	United Kingdom

Source: Own construction

Table 8 Surveyed countries of America

Brazil	Canada	Mexico	USA

Source: Own construction

Table 9 Surveyed countries of Asia

China	Indie	Japan	Turkey

Source: Own construction

Table 10 Surveyed countries of Oceania

Australia	New Zealand

Source: Own construction

Surveyed countries were grouped by their approach to publishing of microdata. Created clusters with the information obtained were: 1) Publishing SUF, 2) Publishing SUF and PUF, 3) Not publishing, and 4) No data.

Author purposely distinguishes between not publishing and no data, because he wanted to provide complete information and with this distinction, he purposefully differentiates situations where author was sure that a given country intentionally does not publish data and situations where author does not have the opportunity to make an exact decision.

Initially, 45 countries were addressed in the questionnaire; however, while gathering information from the survey and researching literature for the paper, author found that it was an initial mistake to consider only a limited range of countries as he found out that other countries also had something to offer in terms of microdata publication and that the original intention would capture only limited information about dissemination of microdata from the population census.

Thus, author decided to extend the research from the limited number of statistical offices to the whole world. Therefore, the results give general information about all the census microdata that can be obtained.

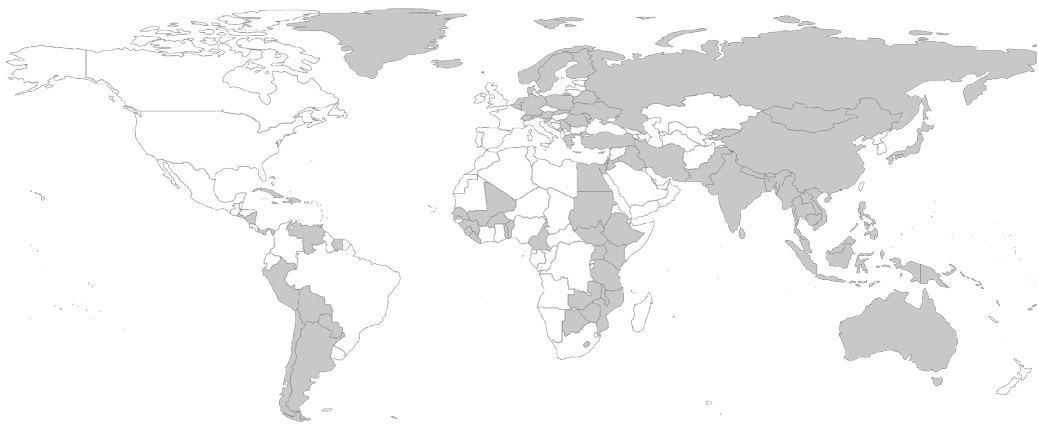
2.1 Publishing SUF

As the first one is presented countries that publish only census microdata for scientific use purposes (SUF). Countries that were categorized it this cluster are listed in Table 11 to Table 15 and are categorized by continent. Not all countries are described in the text below. This is because if a country provides its microdata via IPUMS and does not provide additional useful information, it has been omitted from the detailed description.

Further, the individual countries and their approaches to publishing SUF microdata are described. For clarity, a form was chosen where the links to individual offices and to microdata are listed in the Annex (Tables 31 to 33).

Picture 2 shows on a world map the countries that publish SUF microdata. Full size map can be downloaded from the author's GitHub.[8]

Picture 2 Map of countries publishing SUF



Source: Own construction

2.1.1 Countries of Europe publishing SUF microdata

The part further below describes countries only publishing SUF microdata in Europe. Not all countries in Table 11 are described in the text below. This is because if a country provides its microdata via IPUMS or does not provide additional useful information, it is omitted from the detailed description.

Table 11 Countries of Europe publishing SUF microdata

Austria	Belarus	Belgium	Denmark	Finland
France	Faroe Islands	Germany	Greece	Iceland
Luxembourg	Malta	Montenegro	Moldova	Netherlands
Norway	Poland	Romania	Russia	Serbia
Slovakia	Slovenia	Sweden	Switzerland	Ukraine

Source: Own construction

[8] *<https://github.com/Kyoshido/Population-census-microdata-availability/blob/main/map_SUF_03-2022.png>.*

Belgium
The scientific use of microdata is subject to the prior approval of the Commission for the Protection of Privacy. Researchers interested in census microdata have to submit their data request to the data protection officer. The data request must comply with the principles of finality and proportionality as set out in Belgian legislation. However, Belgium NSO provides only aggregated data with the lowest territorial level being a building. To ensure the confidentiality of the personal information of the enumerated persons, the Cell Suppression Method is applied.

Denmark
Since 1981 Denmark has had annual censuses based on their register data. Access to microdata can only be granted to researchers and analysts in Danish research environments on the research servers after approval from Denmark NSO.

Germany
The legislation that covers the confidentiality of data is Federal Statistics Law (BStatG). Germany NSO do not publish a sample of microdata from the census, instead they have a different approach when they publish the so-called Microcensus, which is a census executed on the representative sample (one percent) of the population, about 370 000 households with 810 000 household members. Microdata are available to researchers on site through Safe Centre (GWAP) or Remote Execution (KDFV) at Research Data Centre of the Federal Statistical Office. Available SUF is anonymised 70% subsample of the households from the Microcensus.

Luxembourg
To access microdata, researchers have to sign a formal convention between the researcher, the institution and Luxembourg NSO. After the convention is signed, Luxembourg NSO would send a sample (5% of the data) to the researcher to prepare his work, especially the syntax, which is send to Luxembourg NSO, who will then check the results if the confidentiality of the data is respected. Luxembourg NSO does not have the SDC model to protect the data confidentiality, but for detailed data they do not provide data with less than 3 observations.

Montenegro
Statistics to researchers are provided under the Law on Official Statistics and Official Statistical System. To obtain access to individual data without identifiers, researchers need to send the license/proof document that your institution is scientific, and the research institution issued by a licensed institution. After that a commission will decide on the provision of individual data.

Norway
Researchers who are affiliated with an approved research institution, or a public authority, can apply to access microdata. The census is not specifically mentioned in their available datasets, but Norway NSO provides data on population or families and households, immigration, and other demographic information, which are updated annually from registers.

Slovakia
Slovakia NSO disseminates microdata with 28 variables, and more detail can be found at their website, which is listed in Table 31, but only in the Slovakian language. An English version of their websites offers much less data, and the information about microdata access is not visible. The Slovakian version

informs that the data are available only for scientific and research purposes. Those interested can request the information service of the Statistical Office of the Slovak Republic for selected microdata.

Sweden

Sweden NSO's platform for access to microdata is called MONA (Microdata Online Access). Users can log in through Security card or Smart phone. Data are delivered in SQL format from STATA, SAS, SPSS or R.

2.1.2 Countries of Americas publishing SUF microdata

Below are listed countries only publishing SUF microdata in the Americas. Countries in Table 12 are not further described in the text below. This is because if a country provides its microdata via IPUMS or does not provide additional useful information, it is omitted from the detailed description.

Table 12 Countries of Americas publishing SUF microdata

Argentina	Bolivia	Brazil	Canada	Chile
Colombia	Costa Rica	Cuba	Dominican Republic	Ecuador
El Salvador	Greenland	Haiti	Jamaica	Nicaragua
Panama	Paraguay	Peru	Saint Lucia	Suriname
Trinidad and Tobago	Venezuela			

Source: Own construction

2.1.3 Countries of Asia publishing SUF microdata

In the part further below are described countries only publishing SUF microdata in Asia. Not all countries in Table 13 are described in the text below. This is because if a country provides its microdata via IPUMS or does not provide additional useful information, it is omitted from the detailed description.

Table 13 Countries of Asia publishing SUF microdata

Armenia	Bangladesh	Bhutan	Cambodia	China
India	Indonesia	Iran	Iraq	Israel
Japan	Jordan	Kyrgyzstan	Laos	Malaysia
Maldives	Mongolia	Myanmar	Nepal	Pakistan
Palestine	Philippines	Sri Lanka	Thailand	Turkey
Vietnam				

Source: Own construction

India

Researchers can access the data under secure environment of the Workstation for Research on Sample Micro-Data from Census. Confidentiality of microdata is secured by the law in The Census Act 1948. 1% and 5% sample of micro-data from the 2001 Census and 2011 Census are available for research to Universities or Institutes. Available are SPSS and Stata software.

Japan

For the dissemination of the microdata, they established the Micro Data Usage Portal Site (Miripo). Miripo website is listed in Table 32. Microdata from the years 1960, 1980 and 1990 are provided

on a magnetic medium and are in TXT format. Microdata from the years 2005, 2010 and 2015 are available only on-site.

Maldives

Request to access the Maldives census 2014 micro dataset can be found at their website, which is listed in Table 32. The dataset will be made available to users after submitting the form and once the application has been processed a Memorandum of Understanding (MoU) will be signed between NBS and the user.

Sri Lanka

Microdata are disseminated in compliance with the Statistics Law of Sri Lanka. Researchers must submit an application which can be found at their website, which is listed in Table 32.

2.1.4 Countries of Africa publishing SUF microdata

In following part are described countries only publishing SUF microdata in Africa. Countries in Table 14 are not further described in the text below. This is because if a country provides its microdata via IPUMS or does not provide additional useful information, it is omitted from the detailed description.

Table 14 Countries of Africa publishing SUF microdata

Benin	Botswana	Burkina Faso	Cameroon	Egypt
Ethiopia	Ghana	Guinea	Kenya	Lesotho
Liberia	Malawi	Mali	Mauritius	Mozambique
Senegal	Sierra Leone	South Sudan	Sudan	Tanzania
Togo	Uganda	Zambia	Zimbabwe	

Source: Own construction

2.1.5 Countries of Oceania publishing SUF microdata

In following part are described countries only publishing SUF microdata in Oceania. Not all countries in Table 15 are described in the text below. This is because if a country provides its microdata via IPUMS or does not provide additional useful information, it is omitted from the detailed description.

Table 15 Countries of Oceania publishing SUF microdata

Australia	Fiji	Papua New Guinea

Source: Own construction

Australia

Confidentiality of microdata is secured in Australia by their law under the Census and Statistics Act 1905, Privacy Act 1988 and Census and Statistics (Information Release and Access) Determination 2018. From the 2016 Census are available the following microdata: 1% sample Basic CURF (Confidentialised Unit Record File), 5% Expanded CURF. CURF is the regional name for SUF.

The 1% Basic CURF contains data on 87 798 dwellings, 93 002 families and 215 597 persons. It provides a sample of one private dwelling record in every hundred from the census, and the associated family and person records. Dwellings with more than six usual residents were removed from the sample to ensure confidentiality of large dwellings. For non-private dwellings the sampling is applied to persons present,

where one person in every hundred is selected and the associated dwelling records included on the file. Data are available on CD-ROM and through the Remote Access Data Laboratory or the Data Laboratory.

The 5% Expanded CURF contains data on 422 725 dwellings, 450 038 families and 1 083 585 persons. It provides a sample of one private dwelling in every twenty from the Census, and the associated family and person records. Dwellings with more than eight usual residents were also removed from the sample to ensure confidentiality of large dwellings. For non-private dwellings the sampling is applied to persons present, where five persons in every hundred are selected and the associated dwelling records included on the file. Data are available through the Remote Access Data Laboratory or the Data Laboratory. Both the 1% and 5% CURFs are available in SAS, SPSS and STATA formats. Both CURF files are not available on CD-ROM to overseas customers.

2.2 Publishing SUF and PUF

Here are presented countries that publish census microdata not only for scientific purpose (SUF) but also for public (PUF). Countries that were categorized it this cluster are listed in Table 16 to Table 20 and are categorized geographically by continent. Picture 3 shows on a world map the countries that publish SUF and PUF microdata. A full size map can be downloaded from the author's GitHub.[9] Picture 3 shows on a world map the countries that publish PUF microdata.

Picture 3 Map of countries publishing SUF and PUF

Source: Own construction

Further, the individual countries and their approaches to publishing SUF and PUF microdata are described. For clarity, a form was chosen where the links to individual offices and to microdata are listed in the Annex (Tables 34 to 37).

2.2.1 Countries of Europe publishing SUF and PUF microdata

In following part are described countries publishing SUF and PUF microdata in Europe. Countries in Table 16 are described in the text below.

[9] *<https://github.com/Kyoshido/Population-census-microdata-availability/blob/main/map_PUF_03-2022.png>*.

Table 16 Countries of Europe publishing SUF and PUF microdata

Albania	France	Hungary	Ireland	Italy
Portugal	Spain	United Kingdom		

Source: Own construction

Albania

The legislation that covers the confidentiality of data is Law No 9888, dated 1.3.2008 "On the Protection of Personal Data" and Law No 17/2018 "On official statistics". The PUF microdata from the Population and Housing Census 2011 can be found at their website, which is listed in Table 33. Data can be downloaded directly without the need for registration or any control.

In the explanatory note to the PUF micro data there is written that they provide 3% of households and the sample technique used to select the sample is a simple random sample. The sample from Population and Housing Census 2011, which was randomly selected, was 21 665 households from 722 226 households in the whole country. In their opinion this method ensures that the sample is representative not just at national level, but also at prefecture level. The data are provided in SPSS format. The methods used to protect confidentiality of the census microdata are resampling and perturbative methods.

France

The processing of personal data for statistical purposes complies with the law Informatique et Libertés. The PUF harmonized microdata from the Population and Housing Census from 1968 to 2017 can be found at their website, which is listed in Table 33. The file contains 18 variables and 52 448 313 records. This dataset has no explanations and documentation in English, it is only available in French. However, it is not the only one, they have 7 126 datasets with English explanations in their database, while they have 23 465 datasets in French. Data can be downloaded directly without the need for registration or any control. The data are provided in dBase and CSV format. The file is created as a file of cumulative individuals, so the observations having the same modalities for all the variables have been grouped, therefore it's essential to use a weighting (variable POND) to recreate population.

Hungary

Hungary NSO is providing PUF files in form of Test files. Test files have the same structure as the microdata sets, which are available internally, but do not reflect the relationships between the variables. That means the files are not determined for analysis. Their purpose is to provide preparatory support for the researchers in accessing data in the Safe Centre and remote execution. Thus, the researcher can pre-prepare his code in advance of the test data at rest and then apply this code in a secure environment to microdata already suitable for the given analyses. Each record is fictitious and the logical correlations between the variables are not fulfilled either. The Test files PUF microdata from the Population and Housing Census 2011 can be found at their website, which is listed in Table 33.

Ireland

The legislation that covers the confidentiality of data is the Statistics Act, 1993. Ireland NSU provides access to two types of microdata files: Anonymised Microdata Files (AMFs), which are PUFs and Research Microdata Files (RMFs), which are SUFs.

Access to AMFs must be approved in advance by Ireland NSU. 5% of anonymised samples of the population census from 1996, 2002, 2006 are available. More information here can be found at their website, which is listed in Table 33.

Access to RMF's is strictly controlled and can be accessed remotely via CSO Researcher Data Portal (RDP) or on-site in a Ireland NSU via the Researcher Data Portal (RDP). Recently in June 2021 Ireland NSU has allowed access to Researcher Microdata Files from home offices. They achieved it by adding an extra layer of security provided by two-factor authentication (2FA).

Italy

Data are protected by the Legislative Decree No. 322 of 6.9.1989. The PUF microdata from the Population and Housing Census 2011 can be found at their website, which is listed in Table 33. The data represent a 1% sample and are provided in TXT format. Data can be downloaded after User Authentication through registration to Italy NSO single sign-on system.

Portugal

The PUF microdata from the Population and Housing Census 1981, 1991, 2011, 2011 can be found at their website, which is listed in Table 33. The microdata represents a 5% sample on individuals and dwellings. The data are provided in Microsoft Access Database. Data can be downloaded directly without the need for registration or any control, just acceptance of use conditions is necessary.

Spain

The data are provided in ASCII format. Data can be downloaded directly without the need for registration or any control. The PUF microdata from the Population and Housing Census 1991, 2001 and 2011 can be found at their website, which is listed in Table 33. The following SDC methods were used to secure confidentiality. Variables of municipalities with fewer than 20 000 inhabitants have to be recorded: place of residence, place of birth, previous place of residence, place of residence 1 year ago, place of residence 10 years ago, place of second home, place of work/study. To protect people who work in occupations related to the Armed Forces, they were moved to other categories.

United Kingdom

Data confidentiality is protected by the law Statistics and Registration Service Act 2007 and Data Protection Act. ONS has a facility called the Secure Research Service which is providing access for approved researchers to restricted microdata. England and Wales microdata samples, two Northern Ireland microdata samples, and two Scottish microdata samples, which are created from 10% of people or households in the 2011 Census are available. From statistical disclosure control methods applied to 2011 Census data were targeted record swapping and restriction of detail.

United Kingdom NSO also publishes Microdata Teaching File, which is the PUF microdata from the Population and Housing Census 2011 and can be found at their website, which is listed in Table 33. The Microdata Teaching File contains a 1% sample of people with just a small number of characteristics. Its purpose is to serve as an educational tool to assist with the teaching.

2.2.2 Countries of Americas publishing SUF and PUF microdata

In following part are described countries publishing SUF and PUF microdata in America. Countries in Table 17 are described in the following text.

Brazil

The PUF microdata from the Population and Housing Census 2010 can be found at their website, which is listed in Table 34. The microdata are in ASCII format. Data can be downloaded directly without the need for registration or any control.

Table 17 Countries of Americas publishing SUF and PUF microdata

Brazil	Canada	Chile	Colombia	Costa Rica
Ecuador	Guatemala	Honduras	Mexico	Puerto Rico
Uruguay	USA			

Source: Own construction

Canada

Data are protected by laws – the Statistics Act and the Privacy Act. The PUF microdata from the Population and Housing Census 2016 can be found as Hierarchical File and as Individuals File at their website, which is listed in Table 34. Hierarchical File provides access to non-aggregated data covering a sample of 1% of the Canadian households, 140 705 household records, which are representing 343 330 persons that have been anonymized. The Individuals File contains a 2.7%, sample of anonymous responses, 930 421 individual records, to the 2016 Census questionnaire. The researcher must complete an order, where he describes his intentions, to get the data. The data are provided in ASCII format and SAS, SPSS or Stata format program source codes.

Chile

The PUF microdata from the Population and Housing Census 2017 can be found at their website, which is listed in Table 34. The data are provided in CSV format. Data can be downloaded directly without the need for registration or any control.

Colombia

Data are protected by Law 79 of 1993 Article 5: the PUF microdata from the Population and Housing Census 2018 can be found at their website, which is listed in Table 34. Data can be downloaded directly without the need for registration or any control.

Costa Rica

The PUF microdata from the Population and Housing Census 2011 can be found at their website, which is listed in Table 34. Microdata is covering a sample of 10%, 427 972 records. To get the microdata you need to register to their website and then just fill in the reason and the intended purpose of what you want to do with the microdata. After that, you can download the microdata without any control of your text admission. The data are provided in SPSS format.

Ecuador

The PUF microdata from the Population and Housing Census 2010 can be found at their website, which is listed in Table 34. Data can be downloaded directly without the need for registration or any control. The data are provided in SPSS format.

Guatemala

The PUF microdata from the Population and Housing Census 2018 can be found at their website, which is listed in Table 34. The microdata are in CSV and SPSS format. Data can be downloaded directly without the need for registration, only control is reCaptcha.

Honduras

The PUF microdata from the Population and Housing Census 2013 can be found at their website, which is listed in Table 34. Unfortunately, the link to download the data does not work because the data are stored on external storage. The link to these databases will transfer researcher to the main page.

Mexico

The PUF microdata from the Population and Housing Census 2017 can be found at their website, which is listed in Table 34. Mexico NSO provides various files of microdata: files of Census (basic questionnaire) and files of Sample (extended questionnaire). Census (basic questionnaire) files provides examples of the Basic Questionnaire database, while ensuring that any type of inference can't be made. Their task is to show characteristics and for users to test their syntax before sending it to be processed through the sections of: Microdata Laboratory, Remote Processing and Processing Service. Sample (extended questionnaire) files provides the results derived from the Extended Questionnaire on the characteristics of inhabited private housing units and their occupants. On these data researchers can perform their analyses.

The data are provided in CSV format. Data can be downloaded directly without the need for registration or any control.

United States of America

By United States of America (USA) law, more precisely by Title 13 of the United States Code, they are obligated to ensure that private information about any specific individual, household, or business is never published and revealed. The results of their census are of great political importance for the USA as they are used to determine the number of seats in their House of Representatives and further determine the size of the legislative districts from the congress to the city councils.

The USA NSO allows researchers to access microdata through a nationwide network of secure Research Data Centres which they created by partnership with various universities, non-profit research institutions, and government agencies. In these Data Centres, researchers can access microdata in the form of restricted use files. As the USA NSO states on its data centre website, research based on micro-data is crucial for them as it also provides them with feedback on the quality of the data and provides them with feedback on the strengths and weaknesses of the microdata records.

In dissemination of the Public Use Microdata (PUM) USA's NSO distinguish between Stateside and Island Areas. Island Areas are Guam and Virgin Islands. The PUM for Island Areas from the Population and Housing Census 2010 can be found at their website, which is listed in Table 35. Stateside areas are 51 states plus Puerto Rico. The PUM for Stateside areas from the Population and Housing Census 2010 can be found at their website, which is listed in Table 34. Each country has its own PUM file. Data can be downloaded directly without the need for registration or any control, The data are provided in TXT format. The disclosure avoidance methods which they used to protect PUM were Data swapping, Synthetic data, Top-coding and bottom-coding, Age perturbation, Reduced detail for categorical variables. Minimum population threshold is set on 100 000. They provide a 10% sample of the population.

In terms of future of statistical disclosure control applied on census data, their researchers are fully aware of the increasing possibilities of growing computer power in combination with progress which was done in the field of mathematic and its possible misuse to compromise and disclose private data. Because of this, they moved in Census 2020 from classical methods to the new concept called "differential privacy", which was firstly introduced in Dwork et al. (2006) and protect every record in their database with this new approach. The Census Bureau was the first one which used this method to disseminate its data. Differential privacy was also implemented by private industries such as Google in their browser Chrome, Uber in their app, Microsoft in their operating system Windows 10 and Apple in their product iPhone.

Uruguay

The PUF microdata from the Population and Housing Census 2011, 2004, 1996, 1985, 1975, 1963 can be found at their website, which is listed in Table 34. The data are provided in SPSS and DBF format. Data can be downloaded directly without the need for registration or any control.

2.2.3 Countries of Asia publishing SUF and PUF microdata

In following part are described countries publishing SUF and PUF microdata in Asia. Countries in Table 18 are described in the text below.

Table 18 Countries of Asia publishing SUF and PUF microdata

Armenia	South Korea

Source: Own construction

Armenia

The PUF microdata from the Population and Housing Census 2014 can be found at their website, which is listed in Table 35. Data can be downloaded directly without the need for registration or any control.

To maintain confidentiality, they made the following changes to the data. The sample from the census contains every tenth household, and they have chosen only households consisting of not more than 20 members and at least with one of them with a status of permanent inhabitant.

Because of small numbers in some variables, they had altered the data as follows. Variable Age had the maximum limit changed to 90 years, so the ages above 90 are equated to 90. In variable Country of Citizenship, they changed countries with less than 5 cases to the group "other". In variable Nationality, they changed the cases of less than 20 to the group "other". In variable Mother tongue, they changed the cases of less than 5 to the group "other". In variable Religion, they changed the cases of less than 10 to the group "other". Variable Born and alive children to mothers had changed the cases of more than 7 to 7. Variable Economic activity type and occupation are presented in two-digit and one-digit codes, and the occupation "servicemen" is recorded as "not applicable". For variable household living conditions, they had changed the cases of more than 5 rooms to 5, and the cases of more than 200 sq/m dwelling space were changed to 200. Their 2011 census sample data file is a text file, which is created by software CSPro.

South Korea

South Korea NSO run a service called Microdata Integrated Service (MSI), which allows Download, Remote Access service, Microdata Research Center Service, On-Demand Service. The English version of MSI page contains only general information. In order to download data or obtain information, it is necessary to switch to Korean. The PUF microdata from the Population and Housing Census 2015 can be found at their website, which is listed in Table 35. They provide a 2% sample of the population. To download data it's necessary to subscribe to Statistics Korea ONE-ID, which is only for Koreans.

2.2.4 Countries of Africa publishing SUF and PUF microdata

In following part are described countries publishing SUF and PUF microdata in Africa. Countries in Table 19 are described in the text below.

Table 19 Countries of Africa publishing SUF and PUF microdata

Angola	Burundi	Ghana	Morocco	Namibia
Rwanda	Somalia	South Africa		

Source: Own construction

Angola

The PUF microdata from the Population and Housing Census 2014 can be found at their website, which is listed in Table 36. Data can be downloaded directly without the need for registration or any control.

Burundi
The PUF microdata from the Population and Housing Census 2008 can be found at their website, which is listed in Table 36. To get the microdata you need to register to their website and then just fill in the reason and the intended purpose of what you want to do with the microdata. After that, you can download the microdata without any control of your text admission. The data are provided in SPSS format.

Ghana
The PUF microdata from the Population and Housing Census 2010 can be found at their website, which is listed in Table 36. To get the microdata researcher need to register to their website and then just fill in the reason and the intended purpose of what researcher want to do with the microdata. After that, anybody can download the microdata without any control of their text admission. The data are provided in SPSS format.

Morocco
The PUF microdata from the Population and Housing Census 2014 can be found at their website, which is listed in Table 36. The data are provided in STATA, SPSS and TXT format. Data can be downloaded directly without the need for registration or any control.

Namibia
The PUF microdata from the Population and Housing Census 2011 can be found at their website, which is listed in Table 36. The dataset is available to everyone after registering in the system. Microdata were sampled based on a stratified random sample. The stratum was at the constituency and urban/rural levels, the threshold for sampling within stratum in file is 250 households. In other cases a simple random sampling was used for each stratum at a 20% sample rate. The microdata includes 93 674 housing units and 418 362 people.

Rwanda
The PUF microdata from the Population and Housing Census 2012 can be found at their website, which is listed in Table 36. To download the data researchers, have to register at their website. Microdata are created as 10% equal probability sample from the population.

Somalia
In Somalia is available Population Estimation Survey which is a first milestone reached towards implementing a full and comprehensive population and housing census. The aim of the analysis was to count Somali Population to know size of population living with cities. To get the microdata researcher need to register to their website and then just fill in the reason and the intended purpose of what researcher want to do with the microdata. After that, anybody can download the microdata without any control of their text admission. The data are provided in SPSS format. Population Estimation Survey can be found at website, which is listed in Table 36.

South Africa
The PUF microdata from the Population and Housing Census 2011 can be found at their website, which is listed in Table 36. The microdata represents a 10% sample of the population.

2.2.5 Countries of Oceania publishing SUF and PUF microdata
In following part are described countries publishing SUF and PUF microdata in Oceania. Countries in Table 20 are described in the text below.

Table 20 Countries of Oceania publishing SUF and PUF microdata

		New Zealand		

Source: Own construction

New Zealand

New Zealand NSO have a large research database called Integrated Data Infrastructure (IDI) which contains microdata of people and households. Available are Census data from 2013 and 2018. Data can be accessed only in their secure virtual environment, in approved facilities (the Data Lab). Outside of the Data Lab environment researchers can apply for confidentialised unit record files (CURFs). The application is listed in a Table 37. CURFs are microdata that were created to protect confidentiality and maintain integrity of the data. They can therefore be classified as SUF microdata. SDC methods which they are using are top-coding, data swapping, and collapsing categorical variables to the unit records. Available are Census data from 2001 and 2013. After the approval, the microdata can be downloaded.

2.3 Not publishing

Countries that do not publish microdata from the population and housing census are listed here. These are countries that have openly stated in the survey that their country does not publish this data, it is clear from their website that they do not publish census microdata, or their datasets are not part of the IPUMS microdata library. These countries are listed in Tables 21 to 25.

Table 21 Countries of Europe not publishing microdata

Bosnia and Herzegovina	Bulgaria	Croatia	Cyprus	Czech Republic
Estonia	Gibraltar	Guernsey	Kosovo	Latvia
Liechtenstein	Monaco	North Macedonia	San Marino	Vatican

Source: Own construction

Table 22 Countries of Americas not publishing microdata

Anguilla	Antigua and Barbuda	Aruba	Bahamas	Barbados
Belize	Bermuda	Cayman Islands	Guyana	Montserrat
Saint Kitts and Nevis	Saint Martin	Saint Vincent and the Grenadines	Sint Maarten	

Source: Own construction

Table 23 Countries of Asia not publishing microdata

Afghanistan	Azerbaijan	Bahrain	Brunei	Georgia
Hong Kong	Lebanon	Macau	Nigeria	Oman
Qatar	Singapore	Syria	Taiwan	Tajikistan
Turkmenistan	United Arab Emirates	Uzbekistan	Yemen	

Source: Own construction

Table 24 Countries of Africa not publishing microdata

Algeria	Cape Verde	Comoros	Eritrea	Gambia
Guinea-Bissau	Libya	Mauritania	Mayotte	Saint Helena
Sao Tome and Principe	Swaziland			

Source: Own construction

Table 25 Countries of Oceania not publishing microdata

American Samoa	Cocos Islands	Christmas Island

Source: Own construction

2.4 No data

A special table was designed for those countries in respect of which author does not have enough information about their approach to access to statistical data. These are the countries that have not responded to the survey or the author could not find any information about their approach to the publication of microdata on their websites or their datasets are not part of the IPUMS microdata library. These countries are listed in Tables 26 to 29.

Table 26 Countries of Europe with no data

Andorra	Lithuania

Source: Own construction

Table 27 Countries of America with no data

Dominica	Falkland Islands	Grenada	Haiti

Source: Own construction

Table 28 Countries of Asia with no data

East Timor	Kazakhstan	Kuwait	North Korea	Saudi Arabia

Source: Own construction

Table 29 Countries of Africa with no data

Central African Republic	Chad	Democratic Republic of the Congo	Djibouti	Equatorial Guinea
Gabon	Ivory Coast	Madagascar	Niger	Republic of the Congo
Seychelles	Tunisia			

Source: Own construction

CONCLUSION

Aim of in this paper was to examine the approaches of individual European statistical offices and other statistical offices around the world to the publication of microdata. The main goal was to compare individual approaches between countries. The primary interest of this work was to summarize whether

NSO publish microdata at all and if so what approach and models did they choose to protect the personal data of their respondents.

At the beginning, author was focused only on a selected sample of statistical offices. During the collection of information on available microdata, author decided to expand the field of processing to the whole world and provide complete information on all possible published microdata from population censuses. The most common problem, which was encountered during processing, was non-response by statistical offices and also the language barrier. Fortunately, the language barrier was overcome with the help of Google Translator, which was an invaluable helper during the writing of this paper. Many NSO do not have their pages and documents translated into English. Complete translated webpages are a rare phenomenon and usually researchers will find only a part of websites in English.

Countries were divided according to their approach to microdata publishing into the following categories: 1) countries that publish only for the scientific community, 2) countries that also provide microdata to the public and 3) countries that do not publish microdata at all. The third group is further divided into two sub-groups. Author distinguishes here between countries for which he was sure that NSO do not publish microdata and NSO for which author did not have enough information to make a final decision on categorization of the approach. This sub-group can be considered as a non-publishing group, but for the sake of clarity, it was decided to create a separate subgroup.

International organizations play an important role in the dissemination of microdata. These international organizations help their members with the preparation of surveys and censuses, as well as with the processing. The most important organization is IPUMS from the United States, which dedicates its existence to the harmonization, collection, and dissemination of microdata to the scientific community. In their microdata library, they made available census data from 96 countries. Thanks to the IPUMS organization, there is a large number of countries that publishes SUF microdata, but they do not have all countries that publish in their library. This is because population census microdata is a very sensitive topic and not all countries are willing to pass this data to a foreign institution located in another country. In total, there are 100 countries that publish SUF microdata.

Countries that provide PUF microdata are much more uncommon. In total, there are 30 countries, but there are large differences between them. The differences consist in the degree of control of microdata access when on the one side there are completely free census microdata that can be downloaded by anyone without any control and on the other side, there is maximum protection, and microdata can be downloaded only if you are a citizen of the state and have the appropriate identifiers.

This paper furthermore provides three maps that were created for countries in IPUMS library, countries that publish SUF and countries that publish PUF. Author hopes that these maps will provide a better overview of the differences between countries. In future research, the aim is to learn more from the findings on the publication of microdata from other statistical offices and enable access to microdata from the population census for the scientific community in the Czech Republic.

ACKNOWLEDGMENTS

This paper has been prepared with the support of a project of the Prague University of Economics and Business – Internal Grant Agency, project No. F4/50/2021.

References

BÁČOVÁ, P. (2018) Příští sčítání bude provázet mnoho novinek [online]. *Statistika&My*, 8(6). [cit. 1.7.2021]. <https://www.statistikaamy.cz/wp-content/uploads/2018/06/18041806.pdf>.

ČERNÝ, P. (2021a). K čemu je dobré sčítání [online]. S*tatistika&My*, 11(1). [cit. 1.7.2021]. <https://www.statistikaamy.cz/wp-content/uploads/2021/02/01_21_SaM.pdf>.

ČERNÝ, P. (2021b). Kde pomohla data ze sčítání [online]. *Statistika&My*, 11(1). [cit. 1.7.2021]. <https://www.statistikaamy.cz/wp-content/uploads/2021/02/01_21_SaM.pdf>.

ČTRNÁCT, P. (2016). Příprava světových sčítání kolem roku 2020 pokračuje [online]. *Demografie*, 58(2). [cit. 1.7.2021]. <https://www.czso.cz/documents/10180/33199357/SLDB.pdf/6d6e4dca-fce8-4d1e-b09c-f464905bf4aa?version=1.0>.

DWORK, C., MCSHERRY, F., NISSIM, K., SMITH, A. (2006). *Calibrating Noise to Sensitivity in Private Data Analysis* [online]. Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg. [cit. 1.7.2021]. <https://doi.org/10.1007/11681878_14>.

EUROSTAT. (2018). *European Statistics Code of Practice: for the National Statistical Authorities and Eurostat (EU Statistical Authority)* [online]. Luxembourg: Publications Office of the European Union. [cit. 1.7.2021]. <https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7?t=1528447068000>.

HUNDEPOOL, A., DOMINGO-FERRER, J., FRANCONI, L., GIESSING, S., LENZ, R., LONGHURST, J., NORDHOLT, E. S., SERI, G., DE WOLF, P. P. (2010). *Handbook on Statistical Disclosure Control* [online]. ESSNet. [cit. 1.7.2021]. <https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf>.

HUNDEPOOL, A. (2012). *Statistical disclosure control.* Wiley series in survey methodology, Chichester, West Sussex, United Kingdom: Wiley.

KOZELEK, V. (2019). Územní příprava sčítání se neobejde bez spolupráce s obcemi [online]. *Statistika&My*, 9(9). [cit. 1.7.2021]. <https://www.statistikaamy.cz/wp-content/uploads/2019/03/18041903.pdf>.

OECD (2014). *Final report* [online]. OECD expert group for international collaboration on microdata access, Paris: OECD. [cit. 1.7.2021]. <https://www.oecd.org/sdd/microdata-access-final-report-OECD-2014.pdf>.

MORAVEC, Š. (2018). Dvoustranné pracovní jednání ČSÚ a ŠÚSR k přípravě sčítání lidu, domů a bytů v roce 2021 [online]. *Demografie*, 60(1). [cit. 1.7.2021]. <https://www.czso.cz/documents/10180/61449042/scitani+lidu.pdf/cd6f2eee-615a-47cb-bd2b-138a5faa8034?version=1.0>.

Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities [online]. European Parliament, Council of the European Union, OJ L 87, 31.3.2009: 164–173. [cit. 1.7.2021]. <http://data.europa.eu/eli/reg/2009/223/oj>.

Regulation (EU) No 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [online]. European Parliament, Council of the European Union, OJ L 119, 4.5.2016: 1–88. [cit. 1.7.2021]. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.

SOBEK, M., CLEVELAND, L. (2020). *Building Research Infrastructure for Harmonized International Census Microdata* [online]. IPUMS Working Paper Series. [cit. 1.7.2021]. <https://doi.org/10.18128/IPUMS2020-01>.

SUDKOVÁ, E. (2016). Konzultace s uživateli dat o obsahu Sčítání lidu, domů a bytů v roce 2021 [online]. *Demografie*, 58(3). [cit. 1.7.2021]. <https://www.czso.cz/documents/10180/33199355/scitani+lidu.pdf/88b2d15c-1387-4a20-b630-3ca726696d57?version=1.0>.

ŠKRABAL, J., ŠANDA, R., HABARTOVÁ, P. (2016). Současný stav přípravy sčítání lidu, domů a bytů v roce 2021 [online]. *Demografie*, 58(1). [cit. 1.7.2021]. <https://www.czso.cz/documents/10180/33199359/SLDB.pdf/d4061f0c-0123-4d88-896c-8e5677f12ba4?version=1.1>.

ŠKRABAL, J. (2017). Příprava sčítání lidu, domů a bytů v roce 2021 [online]. *Demografie*, 59(2). [cit. 1.7.2021]. <https://www.czso.cz/documents/10180/46203818/SLDB.pdf/77b9c078-cd8f-439f-b5b9-1367cd31cec1?version=1.0>.

TEMPL, M., MEINDL, B., KOWARIK, A., CHEN, S. (2014). *Introduction to Statistical Disclosure Control (SDC)* [online]. International Household Survey Network. [cit. 1.7.2021] <https://www.ihsn.org/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf>.

UNITED NATIONS. (2006). *Conference of european statisticians recommendations for the 2010 censuses of population and housing* [online]. United Nations Publication. [cit. 1.7.2021]. <https://unece.org/fileadmin/DAM/stats/publications/CES_2010_Census_Recommendations_English.pdf>.

UNITED NATIONS. (2007). *Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good Practice* [online]. New York and Geneva: UNECE. [cit. 1.7.2021]. <https://unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf>.

UNITED NATIONS. (2015). *Conference of european statisticians recommendations for the 2010 censuses of population and housing* [online]. United Nations Publication. [cit. 1.7.2021]. <https://unece.org/DAM/stats/publications/2015/ECECES41_EN.pdf>.

UNITED NATION. (2017). *Principles and Recommendations for Population and Housing Censuses* [online]. Economic & Social Affairs, New York: United Nations. [cit. 1.7.2021] <https://unstats.un.org/unsd/demographic-social/Standards-and-Methods/files/Principles_and_Recommendations/Population-and-Housing-Censuses/Series_M67rev3-E.pdf>.

ANNEX 1: SURVEY QUESTIONS

Author have asked the following questions:

(i) If you do not currently provide data to respondents, are you considering this? If so, in what form?
(ii) If you are already publishing data, in what form is the presentation taking place? What models do you use to ensure the protection of respondents' data?
(iii) If you are considering publishing microdata, what will be your next step, or what has hindered your implementation?
(iv) If you have considered the possibility of presenting microdata but have withdrawn from it, please send this message as well.

ANNEX 2: TABLES

Table 30 Library of IPUMS

Countries	2010s	2000s	1990s	1980s	1970s	1960s	Pre-1960
Argentina	2010	2001	1991	1980	1970		
Armenia	2011	2001					
Austria	2011	2001	1991	1981	1971		
Bangladesh	2011	2001	1991				
Belarus		2009	1999				
Benin	2013	2002	1992		1979		
Bolivia	2012	2001	1992		1976		
Botswana	2011	2001	1991	1981			
Brazil	2010	2000	1991	1980	1970	1960	
Burkina Faso		2006	1996	1985			
Cambodia	2013	2008	1998				
		2004					
Cameroon		2005		1987	1976		
Canada	2011	2001	1991	1981	1971		1911
							1901
							1891
							1881
							1871
							1852
Chile		2002	1992	1982	1970	1960	
China		2000	1990	1982			
Colombia		2005	1993	1985	1973	1964	
Costa Rica	2011	2000		1984	1973	1963	
Cuba	2012	2002					
Denmark							1801
							1787

Table 30 (continuation)

Countries	2010s	2000s	1990s	1980s	1970s	1960s	Pre-1960
Dominican Republic	2010	2002		1981	1970	1960	
Ecuador	2010	2001	1990	1982	1974	1962	
Egypt		2006	1996	1986			
El Salvador		2007	1992				
Ethiopia		2007	1994	1984			
Fiji	2014	2007	1996	1986	1976	1966	
Finland	2010						
France	2011	2006	1999	1982	1975	1968	
			1990			1962	
Germany				1987	1971		1819
				1981	1970		
Ghana	2010	2000		1984			
Greece	2011	2001	1991	1981	1971		
Guatemala		2002	1994	1981	1973	1964	
Guinea	2014		1996	1983			
Haiti		2003		1982	1971		
Honduras		2001		1988	1974	1961	
Hungary	2011	2001	1990	1980	1970		
Iceland							1910
							1901
							1801
							1729
							1703
Indonesia	2010	2005	1995	1985	1976		
		2000	1990	1980	1971		
Iran	2011	2006					
Iraq			1997				
Ireland	2016	2006	1996	1986	1979		1911
	2011	2002	1991	1981	1971		1901
Israel		2008	1995	1983	1972		
Italy	2011	2001					
Jamaica		2001	1991	1982			
Jordan		2004					
Kenya		2009	1999	1989	1979	1969	
Kyrgyz Republic		2009	1999				
Laos		2005					
Lesotho		2006	1996				

Table 30 (continuation)

Countries	2010s	2000s	1990s	1980s	1970s	1960s	Pre-1960
Liberia		2008			1974		
Malawi		2008	1998	1987			
Malaysia		2000	1991	1980	1970		
Mali		2009	1998	1987			
Mauritius	2011	2000	1990				
Mexico	2015	2005	1995		1970	1960	
	2010	2000	1990				
Mongolia		2000		1989			
Morocco	2014	2004	1994	1982			
Mozambique		2007	1997				
Myanmar	2014						
Nepal	2011	2001					
Netherlands	2011	2001			1971	1960	
Nicaragua		2005	1995		1971		
Norway							1910
							1900
							1875
							1865
							1801
Pakistan			1998	1981	1973		
Palestine	2017	2007	1997				
Panama	2010	2000	1990	1980	1970	1960	
Papua New Guinea		2000	1990	1980			
Paraguay		2002	1992	1982	1972	1962	
Peru		2007	1993				
Philippines	2010	2000	1995				
			1990				
Poland	2011	2002		1988	1978		
Portugal	2011	2001	1991	1981			
Puerto Rico	2010	2005	1990	1980	1970		
		2000					
Romania	2011	2002	1992		1977		
Russia	2010	2002					
Rwanda	2012	2002	1991				
Saint Lucia			1991	1980			
Senegal	2013	2002		1988			
Sierra Leone		2004					

Table 30 (continuation)

Countries	2010s	2000s	1990s	1980s	1970s	1960s	Pre-1960
Slovenia		2002					
South Africa	2016	2007	1996				
	2011	2001					
South Sudan		2008					
Spain	2011	2001	1991	1981			
Sudan		2008					
Suriname	2012	2004					
Sweden							1910
							1900
							1890
							1880
Switzerland		2000	1990	1980	1970		
Tanzania	2012	2002		1988			
Thailand		2000	1990	1980	1970		
Togo	2010				1970	1960	
Trinidad and Tobago	2011	2000	1990	1980	1970		
Turkey		2000	1990	1985			
Uganda	2014	2002	1991				
Ukraine		2001					
United Kingdom		2001	1991				1911
							1901b
							1901a
							1891b
							1891a
							1881b
							1881a
							1871b
							1861b
							1861a
							1851c
							1851b
							1851a
United States	2015	2005	1990	1980	1970	1960	1910
	2010	2000					1900
							1880b
							1880a
							1870

Table 30 (continuation)

Countries	2010s	2000s	1990s	1980s	1970s	1960s	Pre-1960
							1860
							1850b
							1850a
Uruguay	2011	2006	1996	1985	1975	1963	
Venezuela		2001	1990	1981	1971		
Vietnam		2009	1999	1989			
Zambia	2010	2000	1990				
Zimbabwe	2012						

Source: IPUMS Microdata library[10]

Table 31 Countries of Europe publishing SUF with links to microdata

Country	Web pages of the office	Web pages of the microdata
Belgium	<https://statbel.fgov.be>	-
Denmark	<https://www.dst.dk>	-
Faroe Islands	<https://hagstova.fo>	<https://hagstova.fo/fo/atgongd-til-avnevndar-mikrodatur>
Germany	<https://www.destatis.de>	-
Luxembourg	<https://statistiques.public.lu>	-
Malta	<https://nso.gov.mt>	<https://nso.gov.mt/en/Services/Microdata/Pages/Access-to-Microdata.aspx>
Moldova	<https://statistica.gov.md>	<https://statistica.gov.md/pageview.php?l=en&idc=636>
Montenegro	<http://www.monstat.org>	-
Norway	<https://www.ssb.no>	-
Serbia	<https://www.stat.gov.rs>	-
Slovakia	<https://slovak.statistics.sk>	<https://slovak.statistics.sk/wps/portal/ext/themes/demography/census/indicators>
Sweden	<https://www.scb.se>	-

Source: Own construction

Table 32 Countries of Asia publishing SUF with links to microdata

Country	Web pages of the office	Web pages of the microdata
Bhutan	<https://www.nsb.gov.bt>	<https://www.nsb.gov.bt/services/statistical-data-request>
India	<https://www.censusindia.gov.in>	<https://censusindia.gov.in/2011census/workstation.html>
Japan	<http://www.stat.go.jp>	<https://www.e-stat.go.jp/microdata>
Maldives	<http://statisticsmaldives.gov.mv>	<http://statisticsmaldives.gov.mv/census-dataset>
Sri Lanka	<http://www.statistics.gov.lk>	<http://www.statistics.gov.lk/Datadessimination>

Source: Own construction

[10] *<https://international.ipums.org/international-action/samples>.*

Table 33 Countries of Europe publishing PUF and SUF with links to microdata

Country	Web pages of the office	Web pages of the microdata
Albania	<http://www.instat.gov.al>	<http://www.instat.gov.al/en/figures/micro-data>
France	<https://www.insee.fr>	<https://www.insee.fr/fr/statistiques/4995124?sommaire=2414232>
Hungary	<http://www.ksh.hu>	<http://www.ksh.hu/nepszamlalas/tesztallomanyok>
Ireland	<https://www.cso.ie>	<https://www.cso.ie/en/census/censusreports1821-2006>
Italy	<https://www.istat.it>	<https://www.istat.it/en/archivio/196131>
Portugal	<https://www.ine.pt>	<https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_pufs&xlang=en>
Spain	<https://www.ine.es>	<https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176992&menu=resultados&idp=1254735572981#!tabs-1254736195714>
United Kingdom	<https://www.ons.gov.uk>	<https://www.ons.gov.uk/peoplepopulationandcommunity/educationandchildcare/datasets/2011censusteachingfile>

Source: Own construction

Table 34 Countries of Americas publishing PUF and SUF with links to microdata

Country	Web pages of the office	Web pages of the microdata
Brazil	<https://www.ibge.gov.br>	<https://www.ibge.gov.br/en/statistics/social/population/18521-2000-population-census.html?edicao=18531&t=microdados>
Canada	<https://www.statcan.gc.ca>	Hierarchical file <https://www150.statcan.gc.ca/n1/en/catalogue/98M0002X> Individuals file <https://www150.statcan.gc.ca/n1/en/catalogue/98M0001X>
Colombia	<https://www.dane.gov.co>	<http://microdatos.dane.gov.co/index.php/catalog/643/get_microdata>
Costa Rica	<https://www.inec.cr>	<http://sistemas.inec.cr/pad5/index.php/catalog/113/get-microdata>
Ecuador	<https://www.ecuadorencifras.gob.ec>	<https://anda.inec.gob.ec/anda/index.php/catalog/659/get_microdata>
Guatemala	<https://www.ine.gob.gt>	<https://www.censopoblacion.gt/descarga>
Honduras	<www.ine.gob.hn>	<http://170.238.108.229/index.php/catalog/69/get_microdata>
Chile	<https://www.ine.cl>	<http://www.censo2017.cl/microdatos>
Mexico	<https://www.inegi.org.mx>	<https://www.inegi.org.mx/programas/ccpv/2020/?ps=microdatos>
United States of America	<https://www.census.gov>	Stateside areas: <https://www2.census.gov/census_2010/12-Stateside_PUMS> Island areas: <https://www2.census.gov/census_2010/11-Island_Areas_PUMS>
Uruguay	<https://www.ine.gub.uy>	<https://www.ine.gub.uy/web/guest/censos1>

Source: Own construction

Table 35 Countries of Asia publishing PUF and SUF with links to microdata

Country	Web pages of the office	Web pages of the microdata
Armenia	<https://www.armstat.am>	<https://www.armstat.am/en/?nid=210>
South Korea	<http://kostat.go.kr>	<https://mdis.kostat.go.kr/extract/extYearsSurvSearchNew.do?curMenuNo=UI_POR_P9012>

Source: Own construction

Table 36 Countries of Africa publishing PUF and SUF with links to microdata

Country	Web pages of the office	Web pages of the microdata
Angola	<https://www.ine.gov.ao>	<https://andine.ine.gov.ao/nada/index.php/catalog/3>
Burundi	<http://www.isteebu.bi>	<http://www.isteebu.bi/nada/index.php/catalog/3>
Ghana	<https://statsghana.gov.gh>	<https://www2.statsghana.gov.gh/nada/index.php/catalog/51/get_microdata>
Morocco	<https://www.hcp.ma>	<https://www.hcp.ma/downloads/RGPH-2014-Microdonnees-anonymisees-Open-Data_t21400.html>
Namibia	<https://nsa.org.na>	<https://nsa.org.na/microdata1/index.php/catalog/19>
Rwanda	<https://www.ine.pt>	<https://microdata.statistics.gov.rw/index.php/catalog/65/related_materials>
South Africa	<http://www.statssa.gov.za>	<https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/485/get-microdata>
Somalia	<https://www.nbs.gov.so>	<http://microdata.nbs.gov.so/index.php/catalog/4/get_microdata>

Source: Own construction

Table 37 Countries of Oceania publishing PUF and SUF with links to microdata

Country	Web pages of the office	Web pages of the microdata
New Zealand	<https://www.stats.govt.nz>	<https://www.stats.govt.nz/integrated-data/apply-to-use-microdata-for-research> <https://www.stats.govt.nz/integrated-data/apply-to-use-microdata-for-research/confidentialised-unit-record-files-curfs>

Source: Own construction

Selected Coefficients of Demographic Old Age in Traditional and Potential Terms on the Example of Poland and Czechia

Joanna Adrianowska[1] | *University of Lodz, Lodz, Poland*

Received 15.5.2022, **Accepted** (reviewed) 15.6.2022, **Published** 16.9.2022

Abstract

The aging process of the population is a natural demographic process, which is gaining more and more intensity both in Poland, Czechia and other countries. This is a demographically important issue, as it is related to many aspects of life, such as the social care system, the healthcare system or the pension system. The article presents selected demographic coefficients in the traditional approach, in which the construction of measures is based on determining the participation of elderly people in the total population or reflecting the relationship between different age groups. The article also presents coefficients in potential (static) terms, in which not only the number of age groups is important, but also how many years a person or age group can still survive. The values of population ageing coefficients in terms of potential and traditional demography were calculated on the example of Poland and Czechia.

Keywords	***DOI***	***JEL code***
Demography, population ageing, traditional demographic, potential demography	*https://doi.org/10.54694/stat.2022.22*	*J11, J14*

INTRODUCTION

Population ageing is a process that can be considered from both an individual and a collective point of view. An individual's old age is the last stage of a person's life in which many vital functions decrease, physical fitness decreases and the level of health declines. From a community perspective, the phenomenon of population ageing is often defined as an increase in the proportion of older age groups in the total population, resulting from long-term changes in mortality and fertility. The changes are an increase in the older age groups and a decrease in the younger age groups. Factors generating the aging process of the population are also the increasing life expectancy and future life expectancy and low fertility rates

[1] Department of Statistical Methods, Uniwersity of Lodz, Rewolucji 1905, No 41, 90-214 Łódź, Poland. E-mail: jadriana@wp.pl.

(Okólski and Fihel, 2012). Population ageing is a demographic phenomenon in which the age structure of the population is changing from stagnant to regressive, with an increasing proportion of the population at old age (Kurkiewicz, 1992). Population ageing is a demographic process, which increase can be observed in recent years both in Poland and Czechia, as well as in other European societies. The phenomenon of population ageing is an important issue in demographic research, especially in European countries (Długosz, 1996 and 2002). The population ageing process is an intensifying phenomenon leading to significant changes in the demographic structure of the country. The decreasing number of people of working age and the increasing number of people of old age (requiring support and care) has a number of socio-economic consequences.

The ageing of the population in terms of research requires ongoing monitoring and analysis, in order to capture both the general trend and interregional differences. The article aims to illustrate the aging process of the population of Poland and Czechia against the background of 27 countries of the European Union (excluding the United Kingdom). The analyses presented in this article serve to show the growth of measures that define the aging process of the population. The analysis of demographic indicators is carried out on the basis of indicators of traditional demography and potential demography. In the traditional view, the measurement of the ageing of a population usually involves determining the relationship between the size of age groups, an example of which is, among others the determination of the proportion of people aged 65 and over in the total population, i.e. the demographic ageing coefficient.

However, it is important not only the very fact of reaching a certain age of 65 years, referred to as the beginning of old age (Rosset, 1959; Wierzchosławski, 1999), but also how many years after reaching it a person is likely to live. It is clear that, depending on the country, a person aged 65 may have a different number of years to live, which is reflected in life expectancy tables. Potential demography takes that fact into account. In terms of potential, an individual is considered through their life potential, i.e. the number of years they can still live according to current life expectancy tables. The process of ageing of a potential population is considered not through the prism of the number of people who have reached a predefined ageing threshold, but in terms of the years that individuals would live beyond the age considered to be the beginning of that period (Murkowski, 2018). The potential approach to population ageing complements the traditional methods, as it can take into account not only the life expectancy of people currently regarded as elderly but also the number of years that would be lived at old age by people who have not yet reached that age (Murkowski, 2018). The analysis conducted in this article compares selected indicators of population ageing for Poland and Czechia for 2010 and 2019. A comparison of measures from 2019 with measures from 2020 as the first year of the Sars-Covid-19 pandemic is also presented. The aim of the study is to show the phenomenon of population ageing in Poland and Czechia using two approaches – traditional and potential. The analysis involves comparing these measures. In the case of the traditional approach, the analysis consists in identifying the country with the highest value of the indicator and the country with the lowest value, as well as in assessing the difference of the selected indicator for Poland and Czechia in relation to the lowest and the highest value recorded in the given period in the group of 27 EU countries. In the case of the potential approach, it is a comparison of the values of potentials of given age groups as well as the values of potential population ageing rates defined analogically to traditional measures. For indicators in potential terms, an analysis using data from 2010 and 2019 is presented. The values of demographic coefficients in traditional and potential terms presented in the article are calculated on the basis of data from Eurostat, the Central Statistical Office (GUS) and the Czech Statistical Office (CZSO).

1 SELECTED MEASURES OF ADVANCED AGEING

1.1 Demographic coefficients in classical terms

The article contains an analysis of the level of the population ageing in 27 European Union countries using the values of selected demographic coefficients in classical terms (Długosz, Kurek, Kwiatek-Sołtys,

2011; Murkowski, 2018a and 2018b; Cieślak, 2004; Kot and Kurkiewicz, 2004; Holzer, 2003). Classical measures are based on established old-age thresholds and reflect the relationship between different age groups. In the analysis, age 65 was taken as the old age threshold.

The traditional demographic coefficients considered in the analysis are:

- Demographic old age ratio – share of population aged 65 and over in the total population:

$$\frac{L_{65+}}{L_{0-14}} \cdot 100\%, \tag{1}$$

where: L_{65+} – number of people aged 65 and over, L – total population;

- Demographic old age index defined as the quotient of the number of older people, e.g. aged 65+ and over, to the number of young people, e.g. aged up to 15 years:

$$\frac{L_{65+}}{L_{0-14}}, \tag{2}$$

where: L_{65+} – number of people aged 65 and over, $L_{0\text{-}14}$, – number of people up to 15 years of age;

- Ageing rate, defined as the quotient of the number of people aged 85 and over among older people, e.g. aged 65 and over (double ageing index):

$$\frac{L_{85+}}{L_{65+}}, \tag{3}$$

where: L_{85+} – number of people aged 65 and over L_{65+} – number of people aged 65 and over;

- Median age – the median value marks the age limit that half of the study population has already exceeded and the other half has not yet reached.

In order to determine the level of population ageing, a scale of a given measure must be used. The literature describes many proposals for an ageing scale, but none of them is universal.

The ageing of the population is a dynamic phenomenon and the scales are period-adapted and therefore modified over the years. In the case of the demographic old-age coefficient in the Polish literature, the scale proposed by E. Rosset (Rosset, 1959) was most often used, assuming the age of 60 as the threshold of demographic old age. The scale used in the analyses of this article is proposed by J. T. Kowaleski (Kowaleski and Majdzińska, 2012) which is a modification of the Rosset scale with age 65 as the value of old age (Table 1).

In the case of the demographic age index, the scale used in the analysis is presented in Table 2. According to this scale, the actual old age of a population starts when the 0–14 age group becomes less numerous than the 65+ age group, i.e. when the demographic old age index takes on a value greater than unity (Kowaleski, 2011; Kowaleski and Majdzińska, 2012).

To determine the level of old age on the basis of the median, a modified scale is used in the article, in which demographically old are those populations in which the median age of the population exceeds 30 years, and the percentage of the population aged 65+ is at least 15% in them (Maksimowicz, 1990; Kowaleski and Majdzińska, 2012). If the median value is greater than 35 years and the proportion of the population of 65+ is above 20%, then the population is defined as very old.

Table 1 Demographic old age scale based on the proportion of the population aged 65 and over

Share of population aged 65+ (in %)	Demographic study
Stages of population ageing	
Below 10	Demographic youth
10–12	Foreground of ageing
12–14	Ageing well
Over 14	Demographic ageing
Degrees of demographic ageing	
14–16	I
16–18	II
18–20	III
20 and more	IV

Source: Kowaleski and Majdzińska (2012)

Table 2 The scale of advancement of demographic old age based on the index of old age

Demographic age index	Demographic study
Stages of population ageing	
Up to 0.6	Demographic youth
0.6–0.8	Foreground of ageing
0.8–1.0	Ageing well
1.0 and over	Demographic ageing
Degrees of demographic ageing	
1.0–1.2	I
1.2–1.4	II
1.4 and over	III

Source: Kowaleski and Majdzińska (2012)

1.2 Demographic coefficients in potential – static terms

From the point of view of potential demography, it is not the fact of living to a certain age that is important, but how many years one will live after reaching that age. In potential demography, it is not people or events that count, but time – understood as the life potential of individuals. This potential is established on the basis of life expectancy tables. A full description of life tables can be found in the book by Kędelski and Paradysz (Kędelski and Paradysz, 2006).

The basic value of potential demography is the life potential of an individual, defined by the expected number of years (defined in life tables by the symbol ex) and will be calculated from the formula:

$$V(x) = \frac{e_x + e_{x+1}}{2}, \tag{4}$$

where: e_x – the average life expectancy of people at the exact age of x.

The second basic value in the theory of potential demography is the total life potential (Vielrose, 1958; Murkowski, 2013 and 2018), which we will denote by the symbol *PC*. It determines the expected number of years that the study population has to survive in total. It is calculated from the formula:

$$PC = V(0, \omega; 0, \omega) = \sum_{x=0}^{\omega-1} P_x \frac{e_x + e_{x+1}}{2}, \tag{5}$$

where:

ω – the upper age limit in the life expectancy table, in which the number of living people is equal to zero, in this analysis was assumed that ω = 100,

P_x – the average population for a given age group,

e_x – the average life expectancy of people at the exact age of x.

The analysis presented in the article contains the values of potential demography coefficients in static terms. In this approach, it is used the fact that the total life potential can be divided into partial potentials, i.e. the life potentials of people of a certain age for the whole of their further life. Thus, from the total life potential, we extract partial potentials, i.e. the expected number of remaining years of life among people in a fixed age group (Murkowski, 2018a). The potential of people aged from m to M years in relation to their further period of life will be marked with the symbol V(m, M; m, ω) and will be calculated from the formula:

$$V(m, M; m, \omega) = \sum_{x=m}^{M-1} P_x \frac{e_x + e_{x+1}}{2}. \tag{6}$$

Potential ageing rates are calculated by analogy with traditional coefficients, replacing the size of the age groups with the corresponding partial potentials, i.e. the number of years to live of a given population group included in the definition of the coefficient.

Demographic coefficients in potential terms included in the analysis are:

- Demographic old-age ratio:

$$W_{65+} = \frac{V(65, \omega; 65, \omega)}{PC} \cdot 100\%, \tag{7}$$

where: $V(65, \omega; 65, \omega)$ – determines the expected number of remaining years of life among people aged 65+,

- Demographic old age index defined:

$$W_{65+/0-15} = \frac{V(65, \omega; 65, \omega)}{V(0, 15; 0, \omega)} \cdot 100\%, \tag{8}$$

where: $V(0, 15; 0, \omega)$ – determines the expected number of remaining years of life among people aged 0–14 years,

- The advanced ageing index (double ageing index):

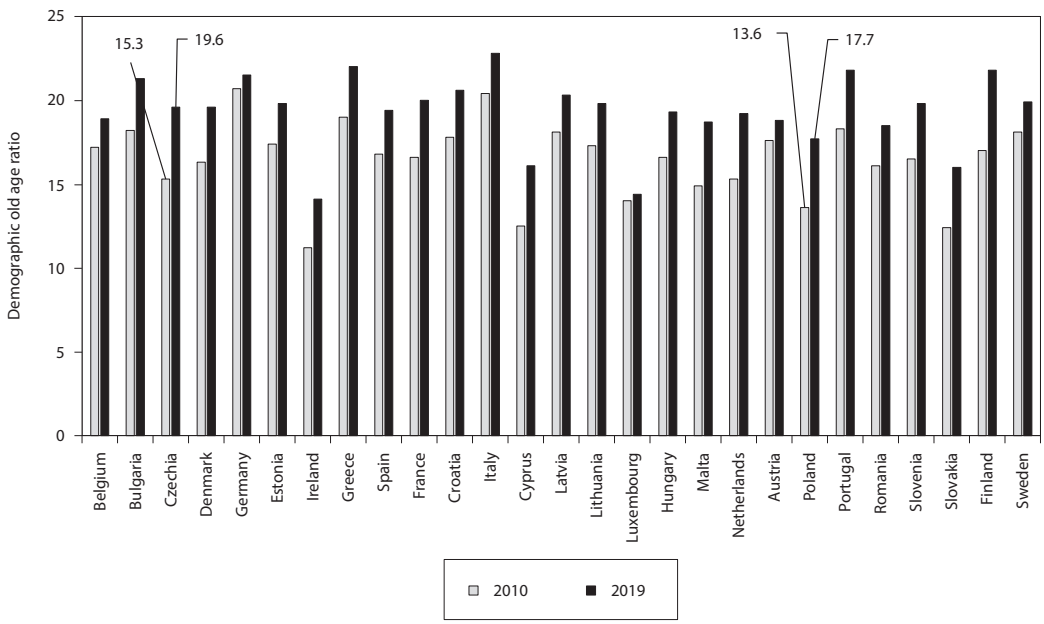
$$W_{85+/65+} = \frac{V(85, \omega; 85, \omega)}{V(65, \omega; 65, \omega)} \cdot 100\%. \tag{9}$$

where: $V(85, \omega; 85, \omega)$ – determines the expected number of remaining years of life among people aged 85+ years.

2 ANALYSIS OF SELECTED MEASURES OF POPULATION AGEING BASED ON CLASSICAL AND POTENTIAL MEASURES

The first demographic coefficient analysed is the demographic old age coefficient – defined as the share of the population aged 65 and over in the total population. The old age dependency ratio for all 27 countries has increased when comparing the 2010 values with those in 2020. For Czechia, the old-age dependency ratio value in 2010 is 15.3%, in 2019 it is 19.6% and in 2020 19.9%. For Poland the value of the coefficient in the following analysed years is 13.6%, 17.7% and 18.2%. The increase in the coefficient between 2010 and 2019 for the Czechia is 4.6 percentage points, for Poland 4.1 percentage points, which gives a similar value of growth for these countries. Comparing the year 2019 with the first year of the Sars-Covid-19 pandemic, i.e. 2020, for Czechia the increase in the demographic ageing coefficient is 0.3 pct %, for Poland the increase was slightly higher – 0.5 pct% (Figure 1 and Figure 2).

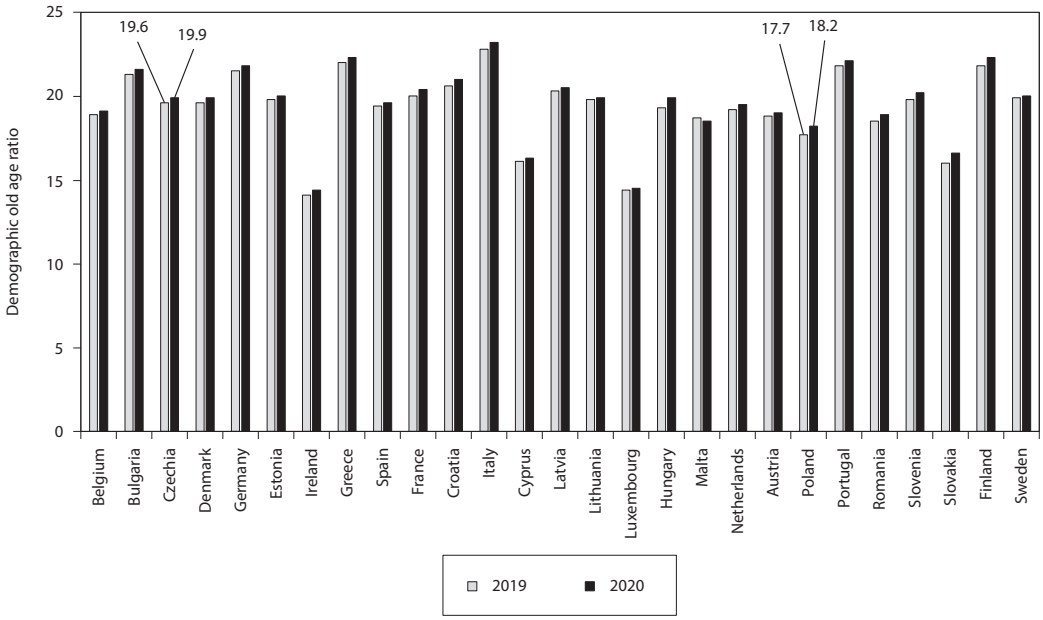
Figure 1 Demographic ageing index values compared to the 27 EU countries



Source: Own elaboration

The country with the highest demographic coefficient over the analysed period was Germany, in 2010 with a coefficient value of –20.7, followed by Italy in 2019 and 2020 with values of 22.8 and 23.2. The state with the lowest value in the three years analysed was Ireland with values of 11.2 – 14.1 – 14.4, respectively (Table 3). Czechia in the ascending ranking among EU-27 countries in 2010 was in the seventh position and Poland in the fourth (Table 4). In 2019 and then in 2020, Czechia moved up the ranking to the thirteenth then the twelfth place, bringing it closer in value to countries defined as demographically old. Poland occupied the fifth position in these two years.

Figure 2 Demographic ageing index values compared to the 27 EU countries



Source: Own elaboration

Table 3 Demographic demographic old age ratio compared to the 27 EU countries

Year	Czechia	Poland	Country with the highest coefficient value	Country with the lowest coefficient value
2010	15.30	13.57	Germany 20.7	Ireland 11.2
2019	19.60	17.70	Italy 22.8	Ireland 14.1
2020	19.90	18.20	Italy 23.2	Ireland 14.4

Source: Own elaboration

Table 4 Place of Czechia and Poland in the ascending ranking of the EU-27 countries

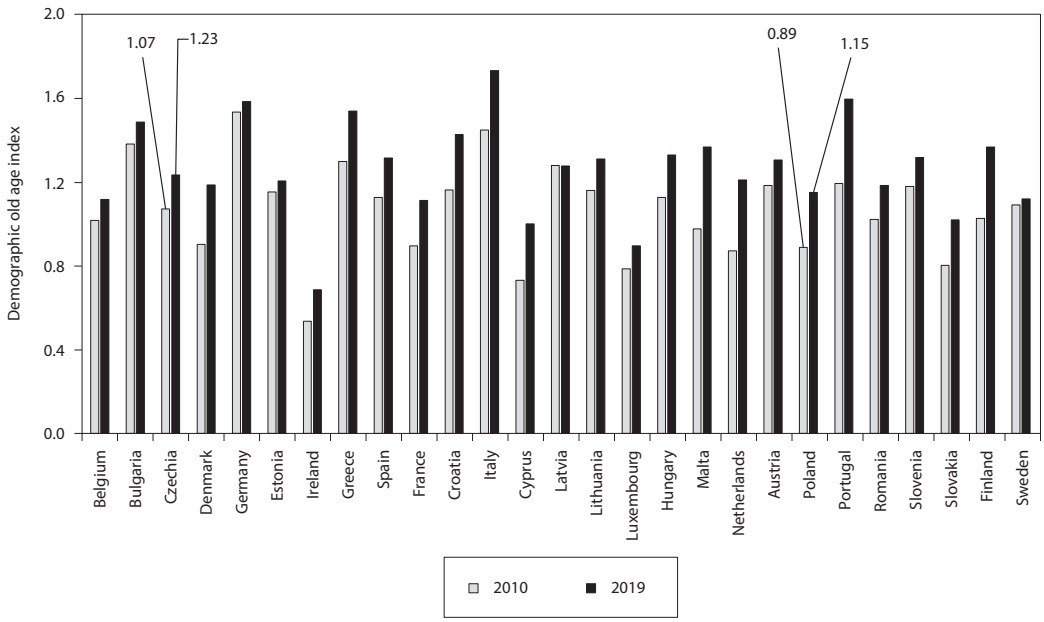
	2010	2019	2020
Czechia	7	13	12
Poland	4	5	5

Source: Own elaboration

According to the scale presented in Table 1, only Ireland was at the pre-ageing stage in 2010, while Slovakia, Cyprus and Poland were at the appropriate ageing stage. The other countries included in the analysis reached the stage of demographic old age, including Czechia of the first degree and Italy and Germany of the IV degree.

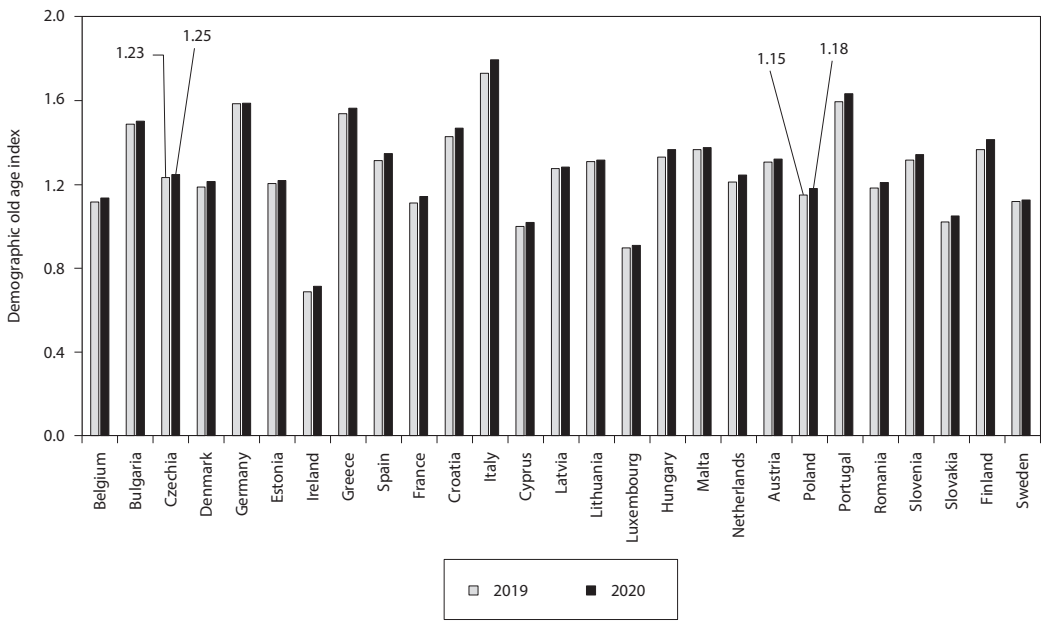
All 27 countries have already reached demographic old age in 2019 and 2020. Czechia has moved to level three of demographic old age, Poland from level two in 2019 to level three in 2020. The group

Figure 3 Demographic age index values compared to the 27 EU countries



Source: Own elaboration

Figure 4 Demographic age index values compared to the 27 EU countries



Source: Own elaboration

of the oldest countries in 2019 was joined by: France, Latvia, Croatia, Bulgaria, Portugal, Greece and Finland, in 2020 this group already consisted of 12 countries. Another factor is the demographic old age index. This ratio determines the quotient of the number of people aged 65 and over to the number of children aged 0–15 years. The literature states that the demographic old age of a population begins when the group of children becomes less numerous than the older population group, i.e. when the coefficient takes a value greater than unity (Kowaleski, 2011; Kowaleski and Majdzińska, 2012; Gavrilova and Gavrilov, 2009).

In the three analysed years, the country with the lowest value of the demographic old age index was Ireland with values of 0.54, 0.69 and 0.71, which, according to the scale (Table 2) placed the country into the group of demographically young countries. The country with the highest value of the measure in 2010 was Germany while in 2019 and 2020 it was Italy (Table 5). Poland in 2010 was at the stage of proper aging with the value of 0.89, Czechia with the value of 1.07 had already reached the demographic old age of the first degree. In 2019 and 2020, both countries Poland and Czechia have already reached index values that allow them to be classified as demographically old countries of the second degree. In 2019 the countries that did not exceed an index value greater than unity were only Ireland, Luxembourg, Cyprus, in 2020 only Ireland and Luxembourg. In the ascending ranking of the 27 countries analysed, Czechia maintained its 13th position in the indicated years, Poland moved from the 6th to the 8th position (Tables 5 and 6).

Table 5 Demographic age index values compared to the 27 EU countries

Year	Czechia	Poland	The country with the highest index value	The country with the lowest index value
2010	1.07	0.89	Germany 1.53	Ireland 0.54
2019	1.23	1.15	Italy 1.73	Ireland 0.69
2020	1.25	1.18	Italy 1.79	Ireland 0.71

Source: Own elaboration

Table 6 Place of Czechia and Poland in the ranking of the EU-27 countries

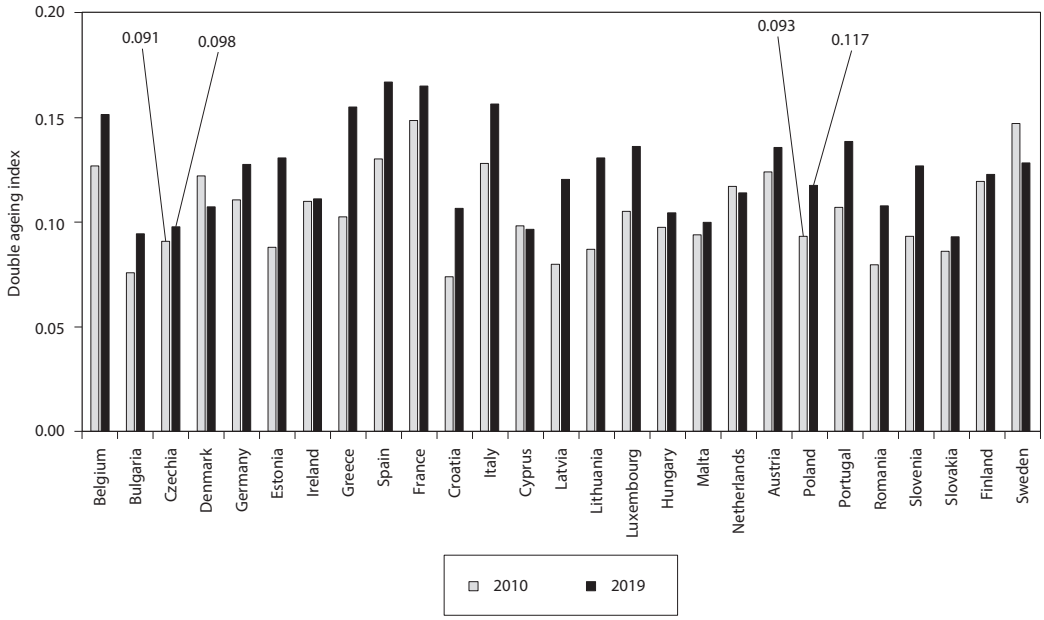
	2010	2019	2020
Czechia	13	13	13
Poland	6	8	8

Source: Own elaboration

Another coefficient determining the level of ageing of the population is the quotient of the number of people aged old to the number of people aged 65 and over, referred to as the double ageing index or the index of old age. Figures 5 and 6 present a picture of the situation described by the old age index for the group of countries being under consideration for the years 2010, 2019 and 2020. Comparing the values from 2010 to 2019, apart from Cyprus, Denmark and Sweden, the remaining countries had higher index values. Both for Poland and Czechia the value of the coefficient has also increased between 2010 and 2019 in the case of Poland by 0.024 and in the case of Czechia by a much smaller value of 0.007. This indicator decreased marginally in 2020 compared to 2019 in the case of Poland the value did not change.

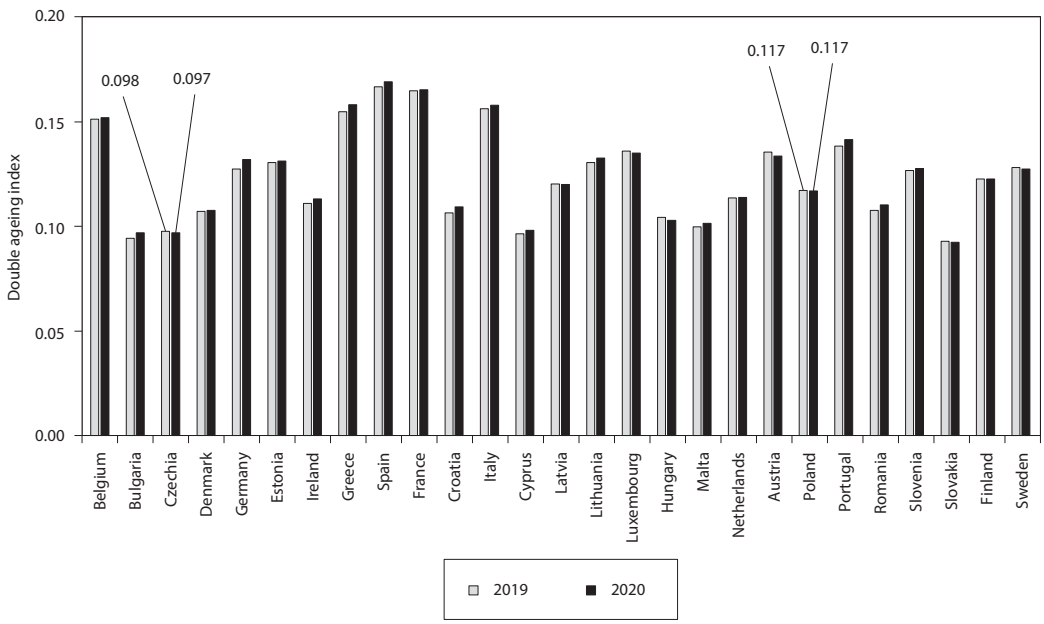
In the ranking of analysed countries (Table 8), Poland in 2010 occupied position 9th then moved towards countries previously described as old countries and occupied position 12th in 2019 and 2020. Czechia was in the 8th position in 2010 then moved up the ranking to be in the 3rd position in 2020 – where the top five countries with the lowest values are: Slovakia, Bulgaria, Czechia, Cyprus and Malta.

Figure 5 Double ageing index values compared to the 27 EU countries



Source: Own elaboration

Figure 6 Double ageing values compared to the 27 EU countries



Source: Own elaboration

Table 7 Double ageing index values compared to the 27 EU countries

Year	Czechia	Poland	The country with the highest index value	The country with the lowest index value
2010	0.091	0.093	France 0.148	Croatia 0.0738
2019	0.098	0.117	Spain 0.167	Slovakia 0.093
2020	0.097	0.117	Spain 0.169	Slovakia 0.092

Source: Own elaboration

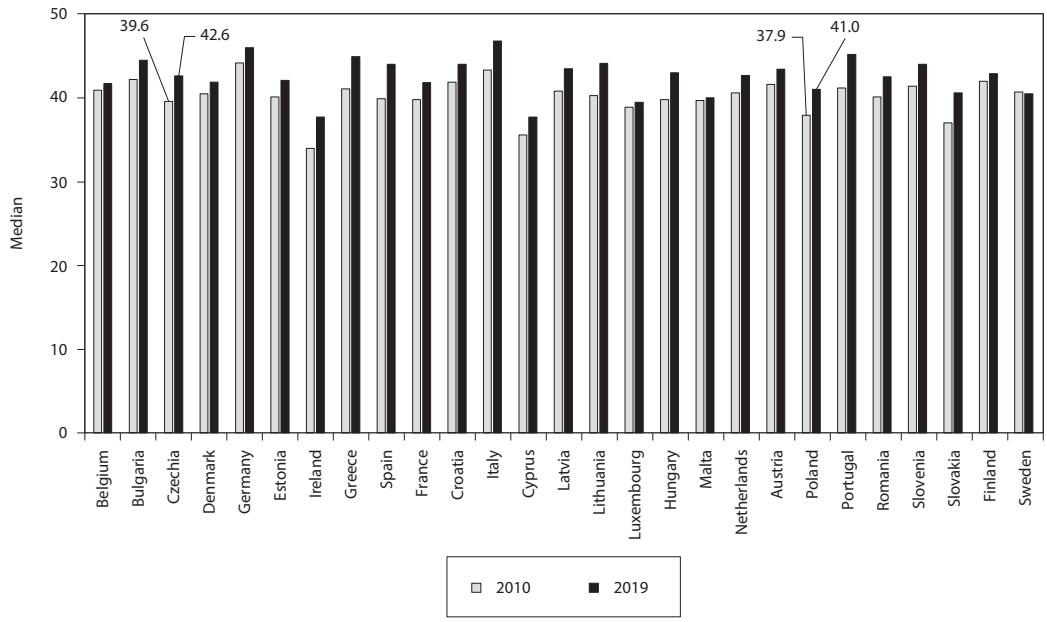
Table 8 Place of Czechia and Poland in the ascending ranking of the EU-27 countries

	2010	2019	2020
Czechia	8	4	3
Poland	9	12	12

Source: Own elaboration

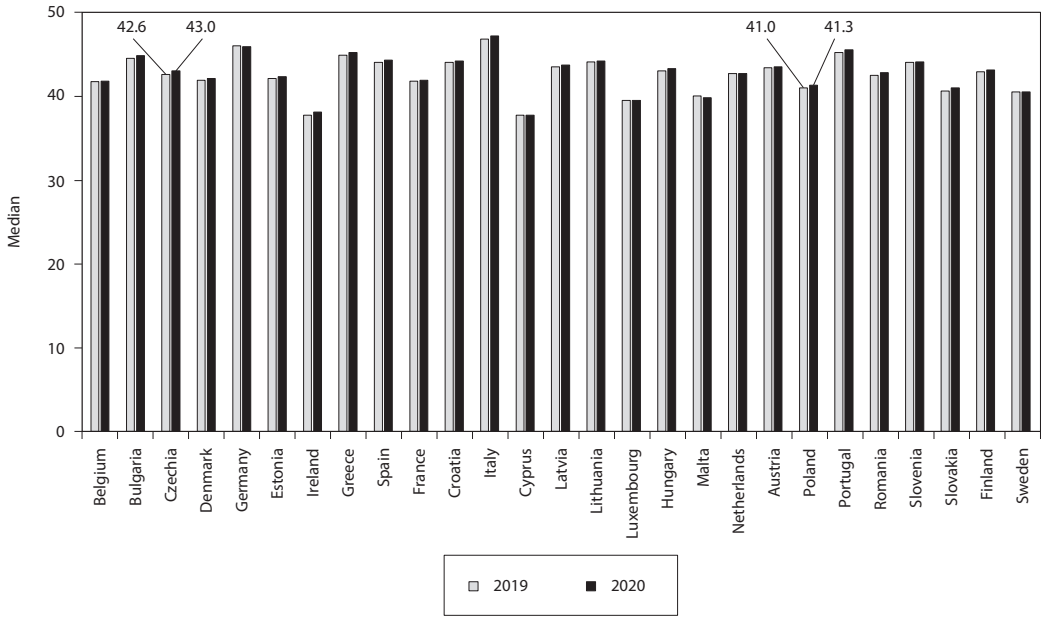
The last classical measure discussed is the median. In 2010 the country with the lowest median value of 34 was Ireland, in 2020 the lowest median was recorded in Cyprus 37.7. The country with the highest value at the beginning of the considered period was Germany with a value of 44.2, in 2019 and 2020 it was already Italy with a value of 46.8 and in 2020 with a value of 44.7 (Table 9). The median values (Figure 7) of 2019 for countries outside Sweden (median decreased by 0.2) were higher than in 2010. Between 2019 and 2020 (chart 8) there was still an increase in value apart from Germany and Malta where the decrease in value was 0.1 and 0.2, respectively.

Figure 7 Median values compared to the 27 EU countries



Source: Own elaboration

Figure 8 Median values compared to the 27 EU countries



Source: Own elaboration

For Poland the increase in the median value in 2019 compared to 2010 was 3.1, for Czechia in the same period was 3.0. In 2020 there was also an increase in this indicator respectively for Poland by 0.3 and Czechia by 0.4. In the increasing ranking presented in Table 10, Poland was ranked 4th out of 27 countries in 2010, and in the following years it moved up to 7th position. In contrast, Czechia has moved from the sixth place in 2010 towards demographically old countries, taking 13th place in 2019 and 14th place the following year.

Table 9 Median values compared to the 27 EU countries

Year	Czechia	Poland	The country with the lowest median value	The country with the highest median value
2010	39.6	37.9	Germany 44.2	Ireland, 34.0
2019	42.6	41.0	Italy 46.8	Ireland, Cyprus 37.7
2020	43.0	41.3	Italy 47.2	Cyprus 37.7

Source: Own elaboration

Table 10 Place of Czechia and Poland in the ranking of the EU-27 countries

	2010	2019	2020
Czechia	6	13	14
Poland	4	7	7

Source: Own elaboration

According to the ageing level scale of A. Maksimowicz (Maksimowicz 1990), Poland was an ageing population in 2010, while Czechia was in the phase of advanced ageing. In 2019 and 2020, both countries were advanced old countries.

The basic value of potential demography is the life potential of the unit V(x) (defined by Formula 4), in which the main unit is e_x – the average life expectancy of people at the exact age of x.

The data in Tables 11 and 12 give examples of average life expectancy for people over 0, 14 and 65 years of age, as well as potential values for people of over the indicated age. In 2010, the average life expectancy for newborn babies under one year of age is longer for people from Czechia than for people from Poland. For people aged 14, the life expectancy for boys from Czechia is longer than for men from Poland. Life expectancy tables show that life expectancy for men up to the age of 70 in Czechia is longer than for men of the same age in Poland. For men from Poland aged 71 and over, life expectancy is longer than for men of the same age from Czechia. However, in case of women from the first year of life onwards Polish women are characterized by a longer life expectancy. The values of individual life potentials in 2010 for Czech men aged 0–70 are higher than for Polish men, in case of Polish women the values of individual life potentials are higher.

In 2019, for Czech men aged 0–69, life expectancy was longer than for Polish men. In case of individual potential, it was higher for Czech men aged 0–70 than for Poles in the same age range. For women aged 0–59 in 2019, both life expectancy values and individual potential values were higher for the Czech women than for the Polish ones. At the age of 60 and over, women in Poland in 2019 were characterized by longer further age and higher values of individual potentials.

Table 11 Unit life potential values for Poland for 2010 and 2019

Sex	Age (x)	2010		2019	
		e_x	V(x)	e_x	V(x)
Male	0	72.10	71.60	74.07	73.71
	14	58.63	58.14	60.46	59.97
	65	15.06	14.76	15.95	15.65
Female	0	80.59	80.26	81.75	81.40
	14	67.09	66.60	68.14	67.65
	65	19.39	18.99	20.10	19.71

Source: Own elaboration

Table 12 Unit life potential values for Czechia for 2010 and 2019

Sex	Age (x)	2010		2019	
		e_x	V(x)	e_x	V(x)
Male	0	74.40	74.01	76.33	75.94
	14	60.74	60.25	62.66	62.16
	65	15.29	14.97	16.29	15.95
Female	0	80.63	80.23	82.10	81.69
	14	66.95	66.46	68.38	67.89
	65	18.75	18.35	19.94	19.53

Source: Own elaboration

According to Formula (5), the total potential of a given population (and partial potential) of the population is calculated by adding the average numbers of years of people from given age group multiplied by the average duration of life of people aged x corresponding to the different age groups according to the Formula (4).

Population ageing rates in potential (static) terms are compiled for Poland and Czechia for 2010 and 2019. The life potentials that form the basis for the calculation of the coefficient values in the potential demography theory were expressed in years and listed in Table 13.

Table 13 Static potential values for the populations of Poland and Czechia in 2010 and 2019

Description	2010		2019	
	Poland	Czechia	Poland	Czechia
Population	38 022 869	10 462 088	37 972 812	10 649 800
Total life potential *PC*	1 512 678 908.32	408 997 284.06	1 483 686 379.11	417 076 022.68
The number of years to live for people up to 15 years of age $V(0, 15; 0, \omega)$	402 320 042.03	105 482 444.56	412 190 272.64	121 824 764.80
Number of years to live for people over 60 years of age $V(65, \omega; 65, \omega)$	57 769 382.35	17 689 148.42	82 282 911.02	26 590 534.46
Number of years to live for people over 85 years of age $V(85, \omega; 85, \omega)$	2 315 979.12	599 849.67	3 843 824.52	890 176.58
Number of years (%) to live for the population in the period of 65+	3.82	5.55	4.33	4.24
Number of years (%) to live for the population in the period of 85+	0.15	0.26	0.15	0.14

Source: Own elaboration

The level of advancement of population ageing in potential (static) terms is much lower than in case of traditional measures. The population of Poland aged 65 and over in 2010 accounted for more than 13% of the country's total population had to live 3.82% years of the total years to be lived by the entire population (Table 12). In the case of Czechia, the traditional demographic ageing coefficient was 15.3%, in potential terms it is 5.55, i.e. the potential of the old age group represents more than 5% of the total potential of Czechia's population. In 2019, the values of the discussed coefficient for both countries increased – for Poland by 1.73 percentage points and for Czechia by 0.88 percentage points. The contributing factors are both the higher percentage of old people in the population of Poland and the fact that people aged 65 and over in Poland are characterised by longer life expectancy as defined by life expectancy tables. In the case of the potential old-age ratio for Poland, the value increased by 0.66 % percentage points in 2019 compared to 2010. For Czechia, the ratio decreased by 1.32 % percentage points (Table 14 and Figure 9).

Note that both coefficients analysed here - demographic old age and old age – in potential terms reach much lower values (in percentage terms) than the corresponding coefficients in traditional terms. This is due to the fact that the average life span of people included in the so-called older subpopulation is much shorter than the average life span of other members of the population. The last of the analysed coefficients - the demographic ageing index determines the share of the potential of people aged 65–100 years to the value of the potential of people aged 0–15 years. Based on the values from Table 12 and it can be seen that in 2010 in Poland the potential of young people was almost 7 times greater than the potential of people of post-productive age, in Czechia it was 6 times greater. In 2019, as a result of the change

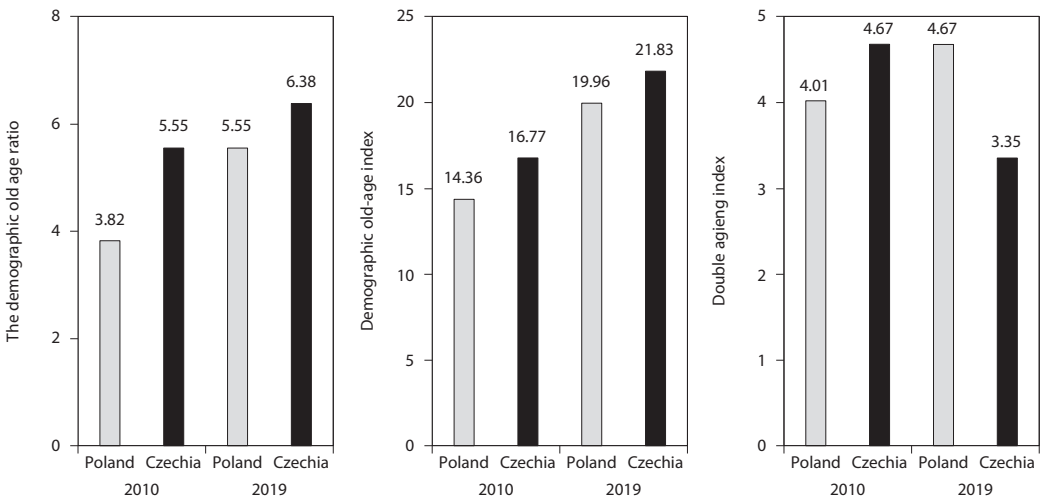
Table 14 The values of the rates of demographic old age potentially static terms, calculated for Poland and Czechia for 2010 and 2019

Potential coefficients – static approach	Value of the coefficient			
	2010	2010	2019	2019
	Poland	Czechia	Poland	Czechia
Demographic old-age ratio	3.82	5.55	5.55	6.38
Demographic old-age index	14.36	16.77	19.96	21.83
Double ageing index	4.01	4.67	4.67	3.35

Source: Own elaboration

in the number of the discussed age groups, the potential of young people in Poland was almost 6 times greater and in Czechia 4.5 times greater. The values of the demographic aging index in 2019 were higher than in 2010, increased for Poland by 5.6% percentage points for the other country by about 5% percentage points. Thus, the potential of people 65 years and older in the final analysed period accounted for almost 20% of the years for further living of the group of people up to 15 years for Poland, while for Czechia this potential accounted for almost 22% of the total potential of people up to 15 years. The increasing value of the demographic old age index is influenced by the fact that the dynamics of the growth of the potential of people 65+ is greater than the dynamics of the growth of the potential of the age group up to 15 years old. The values of particular coefficients and their dynamics are presented in Figure 9.

Figure 9 Values of demographic indicators in terms of potential for Poland and Czechia in comparison between 2010 and 2019



Source: Own elaboration

CONCLUSION

Traditional indicators for measuring the degree of ageing of the population are based on the number of people in individual age groups. In the article, the traditional approach to the study of the ageing process of the population has been extended by a potential approach in a static approach, i.e. using potentials

in calculating measures, which are the number of years that a given population group can still live based on life expectancy tables. The analysis was carried out for Poland and Czechia. Most of the metrics discussed in both potential and traditional terms for both countries reached values greater in 2019 than in 2010. The exception was the ratio defining the share of people 85+ and over to people 65+ and over, whose value for Czechia decreased in 2019 compared to 2010. When analyzing the ranking of the 27 EU countries, Poland and Czechia are placed at the beginning or below the middle position of the ranking. That is why Poland and Czechia are referred to as relatively "young" countries in comparison with the above mentioned group of EU countries. However, the increasing number of people of post-productive age and also extension of life expectancy, and thus the increasing potential of the countries, suggest that in the future both Poland and Czechia will catch up with the countries which are already achieving now high rates of population ageing coefficients.

References

DŁUGOSZ, Z. (1996). Zróżnicowanie struktury wieku ludności na świecie a metody jej klasyfikacji. *Przegląd Geograficzny*, T. LXVIII, 1–2.

DŁUGOSZ, Z. (2002). *Próba określenia stanu i tendencji procesu starzenia się ludności w Europy w świetle wybranych mierników.* Biuletyn Geograficzny, No 1, Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

DŁUGOSZ, Z., KUREK, S., KWIATEK-SOŁTYS, A. (2011). Stan i perspektywy starzenia się ludności w Polsce i Europie [online]. In: SOJA, M. ZBOROWSKI, A. (eds.) *Człowiek w przestrzeni zurbanizowanej*, Kraków: Instytut Geografii i Gospodarki Przestrzennej UJ, 11–26. <http://denali.geo.uj.edu.pl/publikacje,000161>.

GAVRILOVA, N. S., GAVRILOV, L. A. (2009). Rapidly Ageing Populations: Russia/Eastern Europe. In: UHLENBERG, P. (eds.) *International Handbook of Population Aging, International Handbooks of Population*, Springer, 1: 113–131.

HOLZER, J. (2003). *Demografia*. Warszaw: PWE.

KĘDELSKI, M., PARADYSZ, J. (2006). *Demografia.* Poznań: Wydawnictwo Akademii Ekonomicznej w Poznaniu.

KOT, M., KURKIEWICZ, J. (2004). The new measures of the population ageing. *Studia Demograficzne*, 2(146): 17–29.

KOWALESKI, J. T., MAJDZIŃSKA, A. (2012). Miary i skale zaawansowania starości demograficznej. In: ROSSA, A. (eds.) *Wprowadzenie do gerontometrii*, Wydawnictwo Uniwersytetu Łódzkiego.

KOWALESKI, J. T. (2011). *Struktura demograficzna starszego odłamu ludności w województwach (stan aktualny i prognozy do roku 2030)* [online]. University of Lodz. <https://dspace.uni.lodz.pl/xmlui/bitstream/handle/11089/5427/Przestrzenne%20Kowaleski.pdf?sequence=1&isAllowed=y>.

KURKIEWICZ, J. (1992). *Podstawowe metody analizy demograficznej.* Warszaw: Wydawnictwo Naukowe PWN.

MURKOWSKI, R. (2013). *Potencjał życiowy ludności Państw Unii Europejskiej w latach 1995–2009.* Uniwersytet Ekonomiczny w Poznaniu, Wydział Ekonomii Katedra Statystyki i Demografii, Poznań.

MURKOWSKI, R. (2018a). Metody pomiaru zaawansowania procesu starzenia się ludności. *Humanities and Social Sciences*, XXIII, 25(3): 213–229.

MURKOWSKI, R. (2018b). Zaawansowanie procesu starzenia się populacji Polski w latach 1990–2050. *Studia Oeconomica Posnaniensia*, 6(9): 59–77.

Przestrzenne zróżnicowanie starzenia się ludności Polski. Przyczyny, etapy, następstwa. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, 9–33.

MAKSIMOWICZ, A. (1990). Przemiany struktury ludności wg wieku. In: OKÓLSKI, M. (eds.) *Teoria przejścia demograficznego*, Warszaw: PWE.

OKÓLSKI, M., FIHEL, A. (2012). *Demografia Współczesne zjawiska i teorie.* Warszaw: Wydawnictwo Naukowe Scholar.

ROSSET, E. (1959). *Proces starzenia się ludności.* Warszaw: PWE.

VIELROSE, E. (1958). *Zarys demografii potencjalnej.* Warszaw: Państwowe Wydawnictwo Naukowe.

WIERZCHOSŁAWSKI, S. (1999). Demograficzne aspekty procesu starzenia się ludności Polski. *Ruch Prawniczy, Ekonomiczny i Socjologiczny*, 1: 19–56.

Fisim Methodology and Options of Its Estimation: the Case of the Czech Republic

Jakub Vincenc[1] | *Prague University of Economics and Business, Prague, Czech Republic*

Received 20.8.2021 (revision received 7.2.2022), **Accepted** (reviewed) 14.2.2022, **Published** 16.9.2022

Abstract

Financial intermediation services indirectly measured, or simply FISIM, is an adjustment made in national accounts which constitutes significant element in output of the financial institutions. Therefore, the methodological aspects of this adjustment are still broadly discussed issue.

In case of the Czech Republic, the institution responsible for the estimation is the Czech Statistical Office. The paper deeply analyses the approach of this institution and compare it with opinions of many authors. Based on this literature research, the aim of this paper is to propose improvements in the current estimation and find out other options how to estimate the most accurate value of FISIM.[2]

Keywords	***DOI***	***JEL code***
National accounts, FISIM, methodology, Czech Statistical Office, production, interests	*https://doi.org/10.54694/stat.2021.26*	*G21, E23, E40*

INTRODUCTION

The estimations of productivity are broadly discussed issue for a relatively long time because the productivity from an economic point of view is an extremely significant indicator which influences many other statistics. Nevertheless, it is essential to keep in mind that it is only an estimation affected by observational errors and methodological assumptions. These are the main reasons why valuating production more and more accurately is a challenge and many different approaches have been used, especially in the sector of financial institutions (S.12).

The valuation of financial services provided by the financial institution faces up to a few obstacles such as obstacles associated with the payment for the services, which may be done directly or indirectly. Direct payments such as fees and commissions are easily detectable in statistical surveys. However, this part does not include the whole production of financial institutions and as a treatment of differences

[1] Prague University of Economics and Business, Faculty of Informatics and Statistics, Department of Economic Statistics, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. Also Czech Statistical Office, Department of Financial Accounts, Na padesátém 81, 100 82 Prague 10, Czech Republic. E-mail: jakub.vincenc@czso.cz.

[2] Earlier version of this paper was published as final thesis of Postgraduate Statistical Study at Faculty of Informatics and Statistics, Prague University of Economics and Business.

between business and national accounts the estimations of indirect payments for the financial services are needed. The indirect payments arise from the financial operations such as acquisition and disposal of financial assets and liabilities in financial markets (for more details, see Kramulová, Houžvičková and Vincenc, 2019), insurance and pension schemes and from the financial services provided in association with interest charges on loans and deposits.

This paper focuses on the last-mentioned indirect payment which is part of the interest rates on loans and deposits. The adjustment estimating the value of this payments is called financial intermediation services indirectly measured (hereinafter FISIM). In this article I would like to follow up on the discussion of the current form of FISIM adjustment from a methodological point of view and point out possible changes in the current methodology or to find alternative methods of FISIM estimation. The above mentioned will be illustrated on the current methodology and data sources of the Czech Statistical Office (hereinafter CZSO).

The paper is organized as follows: Section 1 contains literature research to show the current level of knowledge. It briefly presents many authors' opinions on the methodological background. It mentions a few controversial aspects in comparison with the CZSO. Section 2 arises from findings made in the first section. It is dedicated to applying the findings into the formulas and shows the calculation and results of the alternative methods in case of the CZSO.

1 CURRENT STAGE OF KNOWLEDGE

1.1 FISIM as a portion of the interest

The FISIM represents the part of services charged by financial intermediator. Payments for these services are included in interest rates on loans and deposits. In case of loans it means, the client pays a higher interest than the reference rate. On the other hand, in case of deposits the client receives a lower interest rate than the reference rate and by that is the intermediation service paid.

The reference rate stays for pure costs of borrowed funds which basically means the cost of money without risk premia and without payment for the intermediation service. This rate is in general located somewhere between the interest rate on loans and the interest rate on deposits. The spread between these rates and the reference rate is FISIM from deposits or FISIM from loans. Their sum is the total FISIM which represents the volume of payments for the intermediation of the financial services related to the providing loans or taking deposits.

In the Czech Republic FISIM takes around 30% of output in the sector of financial institutions and the rest of the output (directly measured part) is estimated almost the same way in each of its subsectors.[3] Especially in the subsectors, which are supervised by the Czech National Bank (hereinafter CNB). It means that directly measured part of the output is mainly sum of two items: "Income from fees and commissions" and "Other operating income". Differences in the calculation within subsectors are non-significant in our case, more relevant is to focus on what specific kind of data these two items contain.

The item called "Other operating income" consists of returns on investments to property and other commodities, earnings from lease or received compensations such as fines. There would be nothing related to lending money or deposit-taking, even if these activities were at the very beginning of banks existence, which make an essential part of S.12.

The output coming from these activities is at least partly included in the item called "Income from fees and commissions", because commissions and fees captured here are linked with financial instrument operations in general. Nevertheless, the output of the financial institutions based on these two items

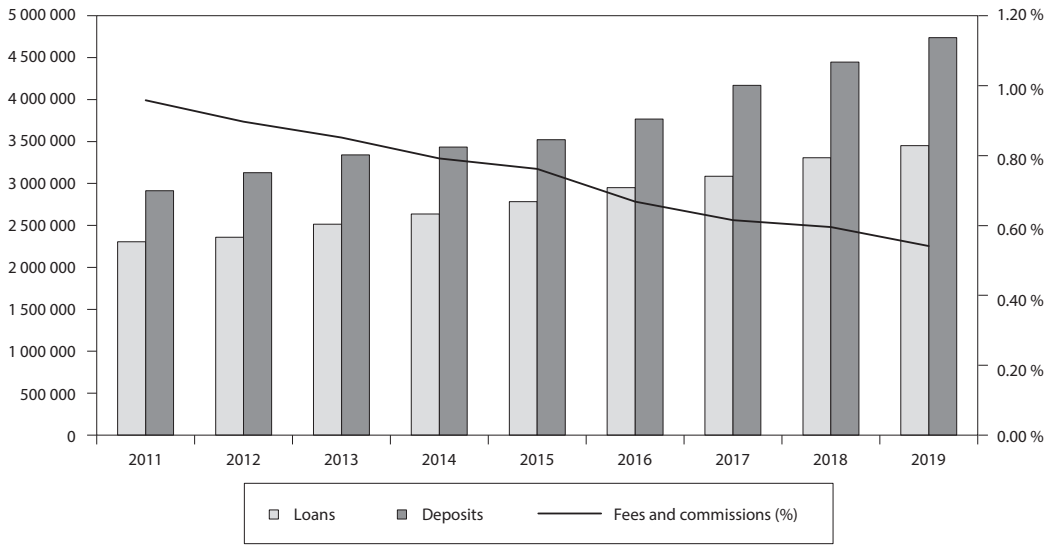
[3] The sector of financial institutions in national accounts has nine subsectors according to European System of Accounts ESA 2010 (Eurostat, 2013) – an accounting framework for the system of national accounts and regional accounts used by members of the European Union.

is not high enough. In some subsectors, the gross value added could be under zero, which is not consistent with the profits in the sector.

Great example in the Czech Republic is the banking sector which is in profit for such a long period. At the end of 2018, banks turned in a profit of CZK 82.1 billion. It was a rise of 8.9% on a year earlier. In 2019 the profit increased by 11.6%, year-on-year, and reached a historical level of CZK 91 billion. Nowadays, the profits of banks decrease mainly due to the coronavirus pandemic. Still, if their output was measured using just these two items mentioned above, it would not reflect the reality (Czech National Bank, 2019 and 2020).

In the field of national accounts was and still is a debate about this topic. The basic agreement is that there are missing parts of the output that must be measured indirectly, as mentioned in the introduction, because these services are not fully paid through fees and commissions. The output captured in item "Income from fees and commissions" decreases over time. The drop can be caused by increasing competitiveness or marketing strategies. However, the point is that the volume of services provided by banks is still rising and their profits as well. The rising volume of deposits and loans in comparison with decreasing fees illustrates Figure 1.

Figure 1 Volume of deposits and loans provided by deposit-taking corporations except for the central bank in the Czech Republic (in mil. CZK), fees and commissions as a ratio to sum of these deposits and loans (in %)



Source: ARAD

The issue is that the item "Income from fees and commissions" includes only fees and commissions paid directly according to the commission agreement and not the indirect part. In this case, this part is included in the interest payments, because according to the manuals for national accounts the interest earnings are not part of the output, although the interest is an earning related to money lending services (at least according to business accounting). The interest in national accounts is a property income only, but the part of it is moved to the output, because "interest on loans and deposits payable to and receivable from financial institutions include an adjustment for a margin that represents an indirect payment for the services provided by the financial institutions in providing loans and accepting deposits" (Eurostat, 2013, p. 98).

The adjustment is called FISIM and is used to solve the problem of different treatment on interest between business and national accounting as the national account's manuals explain that the payment for the intermediation service is included in the interest payment. This part of the interest marked as service must be moved to the output. At the same time, this adjustment also results in better international comparison of the statistics, as in each country may dominate a different charging policy of the financial institutions. Thus, if only direct payments were included in the output, there would be significant irregularities.

The national accounts compilers mostly agree that financial intermediators charge for their services by a portion of interests. This way of charging is typical, especially for banks, because their interest profit has long been the main profitability source (Czech National Bank, 2019).

Hood (2013) explains that banks compensate low profit from fees and commission by a portion of the interest they charge on loans or by a reduction in the interest rates they pay to depositors. The banks use this way of charging rather than charging by explicit fees.

Among others, the financial services occurring in the interest also confirms Akritidis (2007), who adds that commission, account charges and flat rate fees for overdrafts are significantly below the costs paid by the banking industry on wages and bonuses and intermediate costs such as rental, electricity and stationary purchases. It means that when you use conventional treatment of measuring output, which includes direct payments only, the threat of what the OECD described as "the paradox of a prosperous industry" can occur.

The paradox of prosperous industry indicates prosperous industries which have a negligible impact on the national product. The paradox is linked with neoclassical ideas about economy rooted in nineteenth century and makes the opposite view. Neoclassical economics saw banks generate a profit, despite providing activities inherently unproductive (according to the neoclassical economics), therefore the profit was considered accidental. Productivity in general was defined inherently just to some activities and labours; the others were inherently unproductive. The great political economists of the eighteenth and nineteenth centuries saw the productiveness as already determined at the moment when there is a input of labour and depending on whether such labour create material goods (Christophers, 2013).

Christophers (2013) explain that this problem or paradox occurred just because national accountants could not realize the existence of bank's intermediation services. Therefore, this kind of problematic neoclassical theory has long been overcome and there is no doubt about indirect payments for financial services, which must be estimated by some adjustments and added to the output such as FISIM. The approach of the CZSO is in line with that.

1.2 Producers of FISIM

Banks are not the only ones able to produce FISIM. There are other financial institutions to be involved in the estimation. The ESA 2010 says that FISIM is produced by "…deposit-taking corporations except for the central bank (hereinafter S.122); and other financial intermediaries, except insurance corporations and pension funds (hereinafter S.125)" (Eurostat, 2013, p. 331).

In the System of National Accounts SNA 2008[4] (United Nations, 2009), FISIM producer's delimitation is less strict. "These indirect charges in respect of interest apply only to loans and deposits and only when those loans and deposits are provided by, or deposited with, financial institutions" (United Nations, 2009, p. 116). It means that in a theoretical way, the producer of FISIM can be every financial institution that can deposit or lend money. It is not necessary to do both because the amount of money lent usually does not match with the amount of money deposited. Therefore, the indirect charging is imputed in all loans

[4] SNA 2008 is the international standard system of national accounts, which is the background to ESA 2010. In the case of this paper, which focuses on the national accounts compiled by the CZSO, the SNA 2008 provides different, usually looser, interpretation. Nevertheless, the ESA are obligatory for the CZSO.

and deposits offered by a financial institution irrespective of the funds' source and the volume of indirect charges can be different depending on the source of money and its costs.

The manuals (especially SNA 2018) suppose that the producers of FISIM are the financial corporations (S.12) only. It means that financial institutions as a non-market producers captured in general government (S.13) are excluded from the estimation. In case of the Czech Republic, it has a significant impact mainly due to the two institutions: the Czech Export Bank and the National Development Bank. Still, the possibility of negligible FISIM outside S.12 is also mentioned there, but the manual adds that providing financial services is typically under strict regulation and retailers usually do not provide them as secondary production. There are also discussions about the central bank as a FISIM producer, but it is usually also a non-market producer. Moreover, ESA 2010 defines its production as the sum of its costs and its interest rates are affected by monetary policy (United Nations, 2009).

Zieschang (2012) also says that FISIM might occur outside the sector of financial institutions. He respects that these kinds of loans are not included in FISIM estimation and agrees that the value of FISIM from these loans probably will not be quantitatively significant. However, he adds that the treatment of the SNA 2008 does not have to be the most accurate.

We can suppose that almost no consumer's loans are provided directly by retailers, but the CZSO does not even estimate FISIM from loans provided by Other financial intermediaries (hereinafter OFI) classified as a part of S.125. These financial intermediaries usually provide consumers loans instead of retailers, and from my perspective, there is no reason to exclude them from the estimation. ESA 2010 also confirms that OFI are involved in lending money, so the CZSO should enlarge its FISIM producers list by them. Nowadays, the list includes only S.122 and the financial lease provided by the financial intermediary as a part of S.125, but S.125 includes more suitable units such as OFI or Financial payment institutions. Moreover, leasing companies can provide consumers loans as well; at least as a smaller part of financial services which they provide. These loans should be added to the estimation as well.

All the proposed units belong to NACE 64 which refers to "the activities of obtaining and redistributing funds other than for the purpose of insurance or pension funding or compulsory social security" (European Commission and Eurostat, 2008, p. 257). Therefore, they are suitable FISIM producers instead of almost similar units captured in NACE 66 which do not do them-selves (directly) provide financial services.

The impact of this extension in case of the Czech Republic using data of the CZSO is calculated in Section 2.1.

1.3 Allocation of FISIM

On the other hand, the volume of FISIM produced, must be used. There are limited options how to allocate these services on the user side, which includes intermediate consumption, final consumption expenditure or export. In practice, it may be difficult to find the right method of allocating FISIM to the various recipients or users of the services. Therefore, in the past, the manual allowed to record the whole FISIM output as the intermediate consumption of a notional unit with zero output, the so-called "nominal sector". Nowadays, the allocation depends on the institutional sector of user:

a) FISIM used by non-financial corporation, other financial corporation, general government, households as owners of dwellings, households as owners of unincorporated enterprises and non-profit institutions serving households belong to the intermediate consumption;
b) FISIM used by households for individual consumption belong to the final consumption;
c) FISIM used by non-residents belong to the export.

To allocate FISIM, it is necessary to have data about the stock of loans and deposits as well as related value of interests broken by sector of depositor or borrower. Then, it is possible to identify who borrowed or deposited the money and, accordingly, allocate FISIM correctly. The correct allocation is important for many reasons, but the major one is linked with its impact on GDP.

ESA 2010 manual excludes loans and deposits provided between banks from the estimation of FISIM, because there is almost no FISIM occurrence, thus these transactions are used for calculation of internal reference rate (see Section 1.5). Deposits and loans provided by the central bank are also excluded, because it is non-market producer and its interest rates are affected by the monetary policy. Nevertheless, the CZSO excludes also loans and deposits, where the sector of user are Money market funds (S.123), Non-MMF investment funds (S.124) and Other financial intermediaries, except Insurance corporations and Pension funds (S.125). It means that FISIM is not estimated from these loans and deposits and not allocated in these subsectors as well.

Units captured in S.123 and S.124 are in general just funds issuing shares, etc. However, they can also invest on their own account and borrow money for this purpose. This idea is supported by the fact that the CNB reports data, which says that stock of loans and deposits used for FISIM estimation belongs among the other sectors to S.123 and S.124. It is possible that these funds can reach better interest rates with almost no FISIM occurrence, because they are often closely related to the banks, but it is not the reason for the exclusion. Moreover, there is also no reason to exclude S.125, because these units also arrange loans with banks or have deposits there.

In the ARAD[5] database are the volumes of loans and deposits provided by banks to S.123, S.124 and S.125 available. The estimation of FISIM including S.123, S.124 and S.125 on the user side is part of Section 2.2.

1.4 Financial assets and liabilities affected by FISIM

According to the current regulation,[6] the only financial instruments affected by FISIM adjustment are loans and deposits. It is due to suitable properties of its interest rates. Akritidis (2007) explains that these interest rates are under control of commercial banks unlike the interest rates on other financial instruments, such as bonds or securities. They are easily identifiable in division into interest rates on loans and interest rates on deposits, which has consequences in the current method of FISIM estimation. Extending the estimation by bonds may lead to a negative FISIM, it means to produce a negative service.

There is still a debate whether negative FISIM occurrence is explainable, but it is not in line with the current convention and that debate is not the aim of this article. In general, it is not appropriate to include bonds and especially the bonds with interest rates often lower than the reference rate to the estimation (Akritidis, 2007).

Reinsdorf (2011) mentions direct contact between bank and a customer as a key factor for providing implicit services. Thus, the bond purchased by bank on the open market does not produce services that are used by the actual bond issuer. Based on above mentioned, it is possible to also exclude securities. The interbank borrowings are also excluded. Even though SNA 2008 confirms impact of the exclusion on the FISIM estimation, in these transactions is a little if any FISIM. Banks usually borrow from and lend to each other at a risk-free rate.

Zieschang (2012) confirms mentioned above and says that deposits and loans only are affected by FISIM, thus there is nothing inconsistent in the approach of the CZSO.

1.5 Reference rate approach of FISIM estimation

The volume of FISIM is estimated using reference rate approach adopted from the theory of user cost of money which determines whether a financial product is an input or an output due to its net contribution to its revenue. This approach is applied to loans and deposits in the FISIM estimation (Abhiman and Ramesh, 2017).

---

[5] ARAD is a public database, forming part of the Czech National Bank's information service. The purpose of the database is to create a unified system for presenting time series of aggregated data for individual statistics and financial market areas.

[6] ESA 2010, paragraph 14.03.

FISIM from loans is calculated as a difference between interest rate on loans and the reference rate because the reference rate basically represents the average costs of the lender. Hence, FISIM in general should reach positive values. In case of deposit, FISIM is calculated conversely as a difference between reference rate and interest rate on deposits. This means that the interest rate applied to deposits is generally lower than the reference rate. In this case, the client receives a lower interest rate and thus essentially pays for the service. So, the main concept is to divide whole amount of interest by reference rate into two parts. First one which should remain classified as the interests (D.41) and second one which should be part of the output (FISIM (from loans and from deposits).

The SNA 2008 claims that: "The reference rate should contain no service element and reflect the risk and maturity structure of deposits and loans." This is because, after adjusting interest using the reference rate, the service payment should be the only component of FISIM. The interest rate used for inter-bank borrowing and lending may be a suitable choice for a reference rate. Because "for banks within the same economy, there is often little, if any, service provided in association with banks' lending to and borrowing from other banks" (United Nations, 2009, p. 583).

According to the ESA 2010 manual "the internal reference rate is calculated as the ratio of interest receivable on loans within and between subsectors S.122 and S.125 to stocks of loans within and between subsectors S.122 and S.125. When the deposits data is more reliable, the internal reference rate (hereinafter IRR) should be calculated on interbank deposits" (p. 331).

To use the IRR based on the inter-bank transactions has also some pitfalls, mainly because of the risk premia. The risk premia is another part of the interest and serves as a compensation for the possibility that the borrower will not repay the entire liability. To exclude it from FISIM it is necessary for the IRR to include a proportion of the risk, but the risk premia differ from loan to loan. The IRR based on the inter-bank transactions involves almost no risk premia.

The choice of IRR can significantly affect the resulting volume of FISIM. Thus, it is important to set it right based on financial instrument (loan or deposit) with the same maturity (term premia) and with the same risk premia as the instrument from which the FISIM is estimated. Otherwise, some fluctuations may occur, because as Zieschang (2012) presents, using a single IRR inherently allows maturity and risk premia to enter FISIM estimations which are the parts of the interest that should not be captured in the output.

When the stock of loans is multiplied by the IRR (containing an appropriate level of risk and term premia), the result is the volume of interests that represent the costs of the intermediator with the risk and term premia included. Then the amount of interest charged by the intermediary is by these costs. The result obtained is FISIM from loans including only the payment for the intermediation services. Therefore, there is an effort to use the IRR without any intermediation services, but with risk premia (default margin) and term premia corresponding to the affected loans and deposits.

Using more IRRs for loans and deposits differing in risk and term premia is one possibility. The other option is to use its average amount in the economy. However, according to the Advisory Expert Group on National Accounts (2013): "…excluding credit default risk from FISIM, in practice it does not seem feasible, at least in a way that can ensure reasonable comparability across most countries, and so the Task Force concluded that credit default risk should remain part of FISIM in order to facilitate international comparability, at least in the immediate future" (p. 5).

There are opinions that keeping the risk premium as part of FISIM is not just because we are unable to exclude the risk premium while maintaining international comparability. Based on these opinions the real reason is that the risk premia serves to cover the costs related to insurance activities in case of the intermediator does these activities to mitigate the risk. In my point of view, the risk premia should be excluded from FISIM anyway, because it is not the payment for the service of intermediation. However, it does not mean, that the risk premia should not be part of the output in general (Advisory Expert Group on National Accounts, 2013).

The Advisory Expert Group on National Accounts (2013) has also concluded that a term premium should be reflected in FISIM as well. This means that the IRR should be without any payments for services and should include the risk and term premia. "The Task Force stated that channelling funds from borrowers to lenders is a fundamental function of banks, and maturity transformation is inherent to Financial Intermediaries" (p. 41). Moreover, the possible exclusion of the term and risk premia from FISIM faces the very limited suitable data availability.

Except the term premia and risk premia is the currency mix of loans or deposits which can be significantly different from the inter-bank transactions influencing the IRR. Therefore, in the UK the IRR is at first computed separately for each currency and then their weighted average is made to get the overall IRR, but this issue is not of such importance in the Czech Republic (Akritidis, 2017).

Another reason for using more than one IRR may be the difference between the so-called "creditor" and "debtor approach". Debtor approach means that interest payments are predetermined and unchanged in the future. This is how nearly all commercial bank rates are set. On the other hand, the IRR is always based on current financial market conditions and is therefore different for each FISIM estimation (creditor approach). In order to calculate FISIM correctly, it would be necessary to have a unique IRR applied to each loan and deposit at the time of its origin and not to change it for their whole duration. In practice, this method of calculation is hardly implementable, especially due to the volumes of loans and deposits arranged, the limitations of data sources, etc.

Based on the motioned above, Section 2.3 shows some applications of different IRRs in the CZSO approach, which is currently based on interbank loans according to the ESA 2010.

2 VARIOUS METHODS OF FISIM ESTIMATION IN CASE OF THE CZECH REPUBLIC

The alternative methods of calculating the FISIM adjustment discussed above are summarized in the following chapters, which serve for a general overview of this issue. In each chapter of Section 2 I will try to apply the methodological findings made in Section 1 to the data of the CZSO in a time series from 2015 to 2019. Due to the limited possibilities of publishing some data needed for estimations I will focus on the results mainly and its impact on key macroeconomic indicators.

2.1 Enlargement of FISIM producers

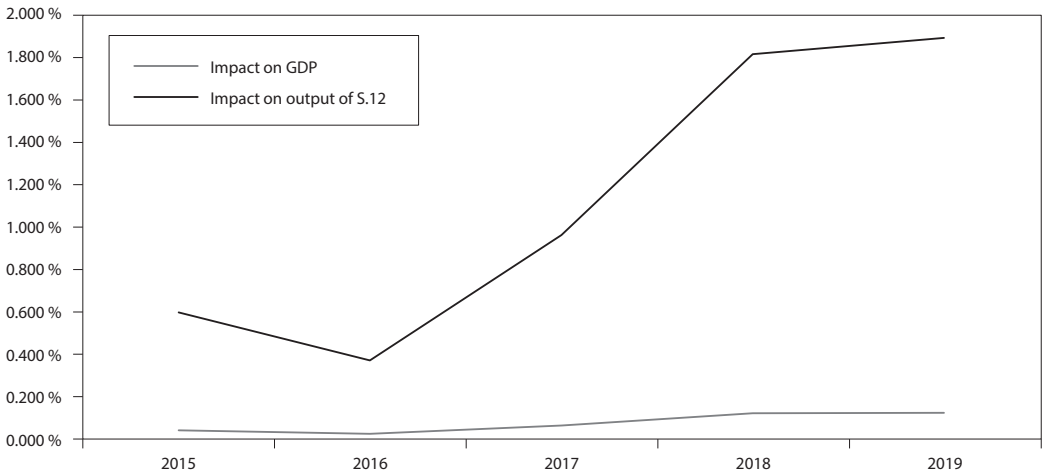
In case of the CZSO, S.125 is further subdivided into four divisions: Leasing companies, Securities dealers on own account, Other financial intermediaries, Financial payment institutions including electronic money institutions. And as I stated in Section 1.2, S.125 includes much more FISIM producers than just Leasing companies which are already included, but only in case of providing financial leasing. However, according to the Czech Leasing and Financial Association (CLFA, 2021) these companies also provide customer loans. Interests from these loans should be also affected by the estimation and when the rest of S.125 as a part of NACE 64 can also provide these loans, it should be included, too. Therefore, I enlarged the current CZSO approach by customer loans provided by units mentioned above and in the estimation I used the same IRRs used by the CZSO. It means that the results show only the impact of the enlargement.

In general, the data needed was taken from national accounts compiled by the CZSO. The volume of loans provided by S.125 was taken from the item Loans (AF.4) with exclusion of financial leasing which is already included. The volume of interests comes from the item Interest (D.41). The exclusion of interests from financial leasing was also made.

This enlargement results in an average increase of 3.2% in the current volume of FISIM produced by resident units between 2015 and 2019. It leads to the growth of output in the whole sector S.12 by 0.9% and by 7.2% in S.125. As you can see in Figure 2 the impact on the output of S.12 is still rising and is driven mainly by an increasing volume of interests in OFI (group of units as a part of S.125).

It is possible that the main part of this enlargement consists of consumer loans, whose borrowers are mainly households. Therefore, on the user side the output would be captured mainly in final consumption of households. Adding these values in the system of national accounts will affect the level of GDP. The increase will be on average 0.063% of GDP at current prices (see Figure 2).

Figure 2 The percentage increase in GDP and output of S.12 (both in current prices) caused by adding customer's loans provided by units classified in S.125 to the FISIM estimation – the Czech Republic (in %)



Source: Own computation from CZSO data

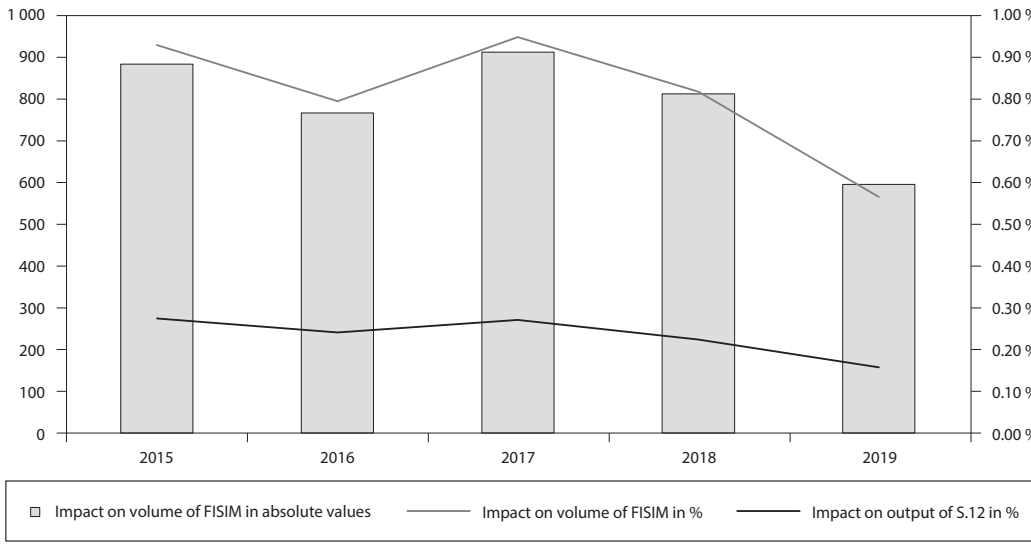
It is necessary to consider that units in S.125 probably have higher cost of money than banks, so the use of current IRR arising mainly from loans within S.122 may not be appropriate. In addition, the interest rates of loans provided by units in S.125 may contain significant default margin, because in my opinion they are often used by clients who would not be able to apply for cheaper loans. Therefore, it can be assumed that actual FISIM coming from this enlargement would be a little lower.

2.2 Enlargement of FISIM users

The FISIM is not estimated from deposits and loans where the bank is a borrower and depositor as well. In these interbank transactions can occur at least small FISIM, but the IRR is based on them, thus the estimation of FISIM in S.122 according to the current approach would be zero. The exclusion of central bank as a user of FISIM was already mentioned in Section 1.3, but the current FISIM estimation made by the CZSO is furthermore underestimated by excluded loans and deposits provided by banks to S.123, S.124 and S.125. According to my finding, these subsectors are suitable for the FISIM estimation and the following results shows the impact of adding them.

These loans and deposits are provided by commercial banks to S.123-5, so the banks are the producers in this case. S.123-5 are the enlargement on the user's side. I used the same IRRs used by the CZSO in 2015–2019 again, the data for loans based on ARAD database and the bank interests coming from the report provided by the CNB. The results approved the occurrence of FISIM it these subsectors even if it is low in comparison to the other subsectors. The average growth of the output makes about 0.23% impact on the output of whole S.12 and the average growth is 0.79%. For more details see Figure 3.

Figure 3 The absolute growth of FISIM produced in the Czech Republic due to the enlargement of the estimation on users' side (in mil. CZK) and the resulting increase in the output of S.12 and the volume of FISIM produced (both in %)



Source: Own computation from CZSO data

These three subsectors belong into financial institutions, which usually have a close relationship with banks. Especially the funds in S.123 and S.124. This is probably the reason why, despite high levels of deposits and loans, these units do not have as high FISIM as in other sectors. In addition, there is sometimes even negative FISIM on the deposit side. This means that these units are able to negotiate a higher interest rate than the IRR.

Based on the above mentioned occurrence of negative FISIM and the close links of some units with the banks, we can have a debate here about their non-market behaviour. This would result in the non-inclusion of these units in the FISIM estimation. Moreover, their small impact on the output would become part of the intermediate consumption with no effect on GDP. There is a possibility to exclude only some of the units closely linked with banks, which probably make the negative FISIM occurrence. However, in general, I assume that at least part of this FISIM is missing in the national accounts of the CZSO.

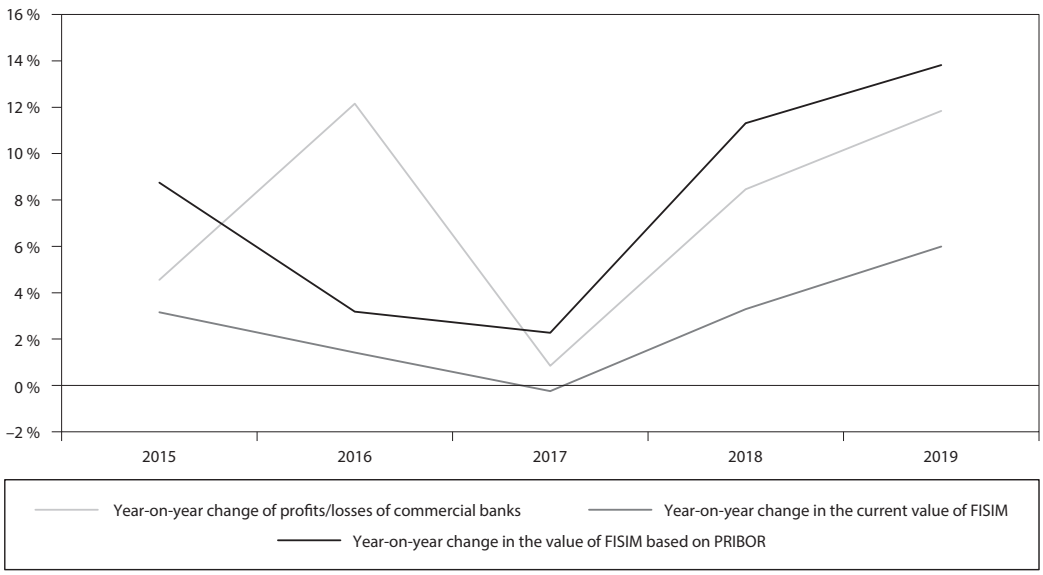
2.3 Alternative reference rate

The sectors included in the following estimations remain the same as they are in the current approach made by the CZSO. No enlargement is made. The focus is only on possible alternatives to the currently used IRRs and their impact on the value of FISIM, GDP or S.12 output.

The first option how to replace currently used IRR is PRIBOR. The interest rate which reflects the willing of bank to lend money on the Czech interbank money market. Several of its values exist and differ according to the length of the interbank loan. As based on ARAD database more than 88% of client loans are provided for a period longer than one year. Most of the deposits (about 75%) can be withdrawn on-demand, but their volume is still rising in time and usually much faster than inflation. It means that even if they are on-demand, they remain on accounts more than one year. Therefore, I decided to use the average annual PRIBOR rate, which is the longest one published by the CNB on its website.

The issue of using PRIBOR as an IRR is that its value is obviously more sensitive, and it is changing much more quickly, especially in 2018 when its value was more than double of 2017. It leads to significant year-on-year changes of total FISIM and huge year-on-year changes in the proportion between FISIM from loans and FISIM from deposits, which does not look reliable. On the other hand, if we accept the idea that the year-on-year change in the total FISIM (as one of the main indicators of banks' output) should be consistent with the year-on-year change in profits of the banking sector, then the use of annual PRIBOR does not seem to be a bad choice (see Figure 4). However, the current value of FISIM still has a better correlation coefficient.

Figure 4 Year-on-year changes of the current FISIM values, FISIM values according to PRIBOR with year-on-year change of profits/losses of commercial banks (in %)



Source: Own computation from CZSO data

The second option as a potential IRR should be interest rates published by the CNB. However, these rates are heavily affected by needs of monetary policy, especially in the couple of last years. For example, in 2016 the two-week repo rate hit the "technical zero" as an effort to prevent potential deflation. So, the usage of these interest rates as IRR is not appropriate.

CONCLUSION

Situation on the financial market reflected in a drop of fees and commissions led to the need of national accountants to develop the concept of indirectly measured services. The FISIM, as one of these services, is an important part of the S.12 output. However, its most appropriate estimation is still broadly discussed issue in the field of national accounts.

The estimation has been developing through the years, but most authors assume that FISIM is charged only as part of interest on loans and deposits. No more financial assets or liabilities are affected, even if they are linked with interest profits or losses. The purpose of FISIM is to replace the missing interest earnings in the output of financial institutions, because the interests are captured only as a property

income according to the manuals. This is where the different approach of national accounting from business accounting becomes visible.

In case of banks, the interest from loans and deposits is the main source of profitability. Therefore, the part of them must be marked as FISIM and moved to their output. The banks are not the only producers of FISIM. The CZSO estimation approach also includes the leasing companies. However, based on the manuals and the opinion of Zieschang (2012), it is possible to enlarge the current range of FISIM producers. Enlargement should focus primarily on those financial institutions that provide consumer's loans. It means that at least all the units captured in S.125 should be marked as FISIM producer.

Possibility to enlarge the estimation is also on the side of consumers (users) of FISIM, because the CZSO approach does not calculate the FISIM from loans and deposits provided to units captured in S.123, S.124 and S.125. The stock of borrowings and deposits provided by S.122 to them is available and I have not found a reason to exclude all of the unites in these three subsectors. The possible exclusion could only apply to those units that are able to negotiate interest rates on the financial market at better than market conditions.

Methodology of the estimation is based on the reference rate approach. Thus, the choice of IRR is crucial to get the most accurate results, but the IRR is usually badly affected by many factors such as risk and term premia, different maturity or currency mix and it is not easy to find the most accurate one. The IRR currently used by the CZSO is based on the inter-bank transactions, so PRIBOR seemed to be suitable alternative. However, it does not lead to better results. The second alternative, interest rates published by the CNB, are badly affected by the monetary policy needs. Therefore, I did not come up with an improvement of IRR.

The main outcome of my thesis is that the approach of the CZSO reflects the reality of financial markets, but could be enlarged on the producers and consumers side. However, as my calculations have shown, in the Czech Republic the impact of these enlargements is not so significant in the absolute values of FISIM.

In the case of the IRR, there remains room for further exploration, in particular with regard to the use more than one reference rate in the estimation. Then, the reference rate could by more closely aligned with maturity, risk and term premia or currency mix of loans and deposits provided, which could lead to more accurate results. Leaving aside the high workload needed, which may not result in the corresponding improvement in the FISIM values, this approach is likely to face a shortage of quality data sources.

References

ADVISORY EXPERT GROUP ON NATIONAL ACCOUNTS. (2013). *8th Meeting, Agenda item: 2, Topic: FISIM* [online]. Luxembourg, May 29–31. <https://unstats.un.org/unsd/nationalaccount/aeg/2013/M8b-2.pdf>.

ARAD. (2021). *Data series system* [online]. Prague: Czech National Bank. [cit. 11.8.2021]. <https://www.cnb.cz/docs/ARADY/HTML/index_en.htm>.

AKRITIDIS, L. (2017). *Financial intermediation services indirectly measured (FISIM) in the UK revisited.* Office for National Statistic in the UK.

AKRITIDIS, L. (2007). Improving the measurement of banking services in the UK National Accounts [online]. *Economic & Labour Market Review*, 1(5): 29–37. <http://dx.doi.org/10.1057/palgrave.elmr.1410073 >.

CHRISTOPHERS, B. (2013). *Placing Finance in Capitalism*. Banking Across Boundaries. John Wiley & Sons. ISBN: 978-1-444-33828-7.

CZECH LEASING AND FINANCIAL ASSOCIATION. (2021). *Report on the state and development of the non-bank leasing, credit and factoring market in the Czech Republic in 2020* [online]. Prague. [cit. 11.8.2021]. <https://www.clfa.cz/data/dokumenty/1164-rok2020zprava.pdf>.

CZECH NATIONAL BANK. (2020). *Financial stability report 2019/2020* [online]. Prague. [cit. 11.8.2021]. <https://www.cnb.cz/export/sites/cnb/en/financial-stability/.galleries/fs_reports/fsr_2019-2020/fsr_2019-2020.pdf>.

CZECH NATIONAL BANK. (2019). *Financial stability report 2018/2019* [online]. [cit. 11.8.2021]. <https://www.cnb.cz/export/sites/cnb/en/financial-stability/.galleries/fs_reports/fsr_2018-2019/fsr_2018-2019.pdf>.

DAS, A., JANGILI, R. (2017). Financial Intermediation Services Indirectly Measured (FISIM): The role of reference rate [online]. *Statistical Journal of the IAOS*, 33: 515–524. <http://dx.doi.org/10.3233/SJI-160280 >.

EUROSTAT. (2013). *European system of accounts ESA 2010* [online]. Luxembourg: Publications Office of the European Union. <http://doi.org/10.2785/16644>.

EUROSTAT. (2008). *Statistical classification of economic activities in the European Community.* Luxembourg: Publications Office of the European Union. ISBN 978-92-79-04741-1.

HOOD, K., K. (2013). *Measuring the Services of Commercial Banks in the National Income and Product Accounts: Changes in Concepts and Methods.* Bureau of Economic Analysis.

KIMBERLY, D., ZIESCHANG. (2016). *FISIM Accounting.* CEPA Working Papers Series WP012016, School of Economics, University of Queensland, Australia.

KRAMLULOVÁ, J., HOUŽVIČKOVÁ, H., VINCENC, J. (2019). Methodology of Estimating "Financial" Margins and their Capturing in the System of National Accounts [online]. *Statistika: Statistika: Statistics and Economy Journal,* 1: 6–23. <https://www.czso.cz/documents/10180/88506450/32019719q1_006.pdf/392aab01-6268-45a5-92df-7e303795506b?version=1.0>.

REINSDORF, M. (2011). *Measurement of Implicitly Priced Output of Commercial Banks in the U.S. National Accounts.* Paper presented at the Meeting of the Task Force on Financial Intermediation Services Indirectly Measured (FISIM), March 3–4, Washington.

UNITED NATIONS. (2009). *System of National Accounts 2008.* New York: United Nations. ISBN 978-92-1-161522-7.

Remembering Professor Petr Hebák (9.8.1940–12.6.2022)

Libor Svoboda | *Czech Statistical Office, Prague, Czech Republic*
Tomáš Karel | *Prague University of Economics and Business, Prague, Czech Republic*

On Sunday afternoon, June 12, we learned the sad news that Professor Petr Hebák has left us forever at the age of 81. He was the Nestor[1] of Czech statistics, but above all else a multifaceted personality, good man, colleague, and friend.

In his person, the statistical community has lost an important expert and a passionate promoter of statistics. He dealt mainly with the problems of regression analysis, multivariate statistical methods, and Bayesian statistics and their use in economic analyses. He was the author or co-author of 8 monographs in the field of statistics. We should mention *Multivariate Statistical Methods with Applications* (SNTL, 1987), which he wrote with his generational contemporaries Jiří Hustopecký, and a trilogy with a similar title, which he published together with other authors in 2004–2007. The monograph *Statistical Thinking and Tools for Data Analysis* (Informatorium, 2013) was the collective work in which he and his colleagues summarized the results of their research. In addition to books, he has authored dozens, perhaps hundreds, of articles in journals and conference proceedings. Among his popular works, let us recall the articles *Statistical Data and Their Meaning*, *Statistical Data and Their Telling Power*, *and Statistical Data and Their Use*s, which he published in the journal Statistics in 2002 and 2003.

Petr Hebák's rich teaching activity was mainly connected with the Department of Statistics at the University of Economics, where he worked since 1962 for more than 50 years (over 100 semesters). When shaping the profile and focus of the department in the early 1990s, he was at its head and actively participated in the establishment of the Faculty of Computer Science and Statistics. Through lectures, seminars, and supervision of theses and Ph.D. students, he was involved in the formation of several generations of future statisticians, including many current academics, staff of the National Statistical Office, researchers, and other professionals. He has written over 30 titles of university and high school textbooks and scripts. In this field, his collection of problems, *Probability Counts in Examples* (with co-author Jana Kahounová), which has been released in seven editions between 1978 and 2014, will probably remain unsurpassed.

However, Petr Hebák was not only a statistician and a great teacher; his range of activities was incredibly wide. Let's try to mention at least some of them. Petr was a long-time player at the national team level, a coach, and an official of the bridge association. In 1993 he published the first Czech modern textbook of this card game, *Bridge for Everyone*, which has also been published repeatedly, most recently in 2010. He also played chess at the master level. It is also worth mentioning his acting in a theatre company, and his interest in boating and cottaging in the Highlands.

Peter will always be remembered as a highly sociable, amusing, and entertaining person. When we meet with friends, we will certainly recall with a smile many stories in which he figured (some

[1] We think the designation is very appropriate in his case. Nestor was a figure of mythical Greece. According to Homer, he was the oldest, most righteous and also the most prudent of all the Achaean leaders fighting at Troy.

of which have long since taken on a life of their own). His extraordinarily developed sense of justice and social sensitivity in combination with his outspokenness caused some difficulties throughout his life

His beloved wife Olga, always kind and patient, gave him great support in all his activities. Given the above list of activities, the reader will understand that it was not always entirely easy. It must be said that Peter was very attached to his large family, his four daughters, their partners, grandchildren, and great-grandchildren.

While celebrating his 80th birthday, Peter told us that "we have come together to rejoice that we are still alive". It is very hard to accept the fact that this will no longer be the case.

Conferences

The **24th AMSE** Scientific Conference (**Applications of Mathematics and Statistics in Economics**) took place ***from 31st August to 4th September 2022*** in Velké Losiny, Czechia. More at: <*http://www.amse-conference.eu*>.

The **40th MME** International Conference (**Mathematical Methods in Economics**) was held ***during 7–9 September 2022*** in Jihlava, Czechia. More at: <*https://mme2022.vspj.cz*>.

The **2022 IDIMT** Conference (**Interdisciplinary Information Management Talks**) took place ***from 7th to 9th September 2022*** in Prague, Czechia. More at: <*https://idimt.org*>.

The **16th MSED** Conference (**International Days of Statistics and Economics**) was held during ***8–10 September 2022*** in Prague, Czechia. More at: <*https://msed.vse.cz*>.

Papers
We publish articles focused at theoretical and applied statistics, mathematical and statistical methods, conception of official (state) statistics, statistical education, applied economics and econometrics, economic, social and environmental analyses, economic indicators, social and environmental issues in terms of statistics or economics, and regional development issues.

The journal of Statistika has the following **sections**:

The ***Analyses*** section publishes complex and advanced analyses based on the official statistics data focused on economic, environmental, social and other topics. Papers shall have up to 12 000 words or up to 20 1.5-spaced pages.

Discussion brings the opportunity to openly discuss the current or more general statistical or economic issues, in short what the authors would like to contribute to the scientific debate. Contribution shall have up to 6 000 words or up to 10 1.5-spaced pages.

In the ***Methodology*** section we publish articles dealing with possible approaches and methods of researching and exploring social, economic, environmental and other phenomena or indicators. Articles shall have up to 12 000 words or up to 20 1.5-spaced pages.

Consultation contains papers focused primarily on new perspectives or innovative approaches in statistics or economics about which the authors would like to inform the professional public. Consultation shall have up to 6 000 words or up to 10 1.5-spaced pages.

Book Review evaluates selected titles of recent books from the official statistics field (published in the Czech Republic or abroad). Reviews shall have up to 600 words or 1–2 1.5-spaced pages.

The ***Information*** section includes informative (descriptive) texts, information on latest publications (issued not only by the Czech Statistical Office), or recent and upcoming scientific conferences. Recommended range of information is 6 000 words or up to 10 1.5-spaced pages.

Language
The submission language is English only. Authors are expected to refer to a native language speaker in case they are not sure of language quality of their papers.

Recommended Paper Structure
Title — Authors and Contacts — Abstract (max. 160 words) — Keywords (max. 6 words / phrases) — Introduction — 1 Literature survey — 2 Methods — 3 Results — 4 Discussion — Conclusion — Acknowledgments — References — Annex (Appendix) — Tables and Figures (for print at the end of the paper; for the review process shall be placed in the text).

Authors and Contacts
Rudolf Novak,1 Institution Name, City, Country
Jonathan Davis, Institution Name, City, Country
1 Address. Corresponding author: e-mail: rudolf.novak @domainname.cz, phone: (+420)111222333.

Main Text Format
Times 12 (main text), 1.5 spacing between lines. Page numbers in the lower right-hand corner. *Italics* can be used in the text if necessary. *Do not* use **bold** or underline in the text. Paper parts numbering: 1, 1.1, 1.2, etc.

Headings
1 FIRST-LEVEL HEADING (Times New Roman 12, bold)
1.1 Second-level heading (Times New Roman 12, bold)
1.1.1 Third-level heading (Times New Roman 12, bold italic)

Footnotes
Footnotes should be used sparingly. Do not use endnotes. Do not use footnotes for citing references.

References in the Text
Place references in the text enclosing authors' names and the year of the reference, e.g. „…White (2009) points out that…", „…recent literature (Atkinson and Black, 2010a, 2010b, 2011; Chase et al., 2011: 12–14) conclude…". Note the use of alphabetical order. Between the names of two authors please insert „and", for more authors we recommend to put „et al.". Include page numbers if appropriate.

List of References
Arrange list of references alphabetically. Use the following reference styles: [book] HICKS, J. (1939). *Value and Capital: An Inquiry into Some Fundamental Principles of Economic Theory.* 1st Ed. Oxford: Clarendon Press. [chapter in an edited book] DASGUPTA, P. et al. (1999). Intergenerational Equity, Social Discount Rates and Global Warming. In: PORTNEY, P., WEYANT, J. (eds.) *Discounting and Intergenerational Equity.* Washington, D.C.: Resources for the Future. [on-line source] CZECH COAL. (2008). *Annual Report and Financial Statement 2007* [online]. Prague: Czech Coal. [cit. 20.9.2008]. <http://www.czechcoal.cz/cs/ur/zprava/ur2007cz.pdf>. [article in a journal] HRONOVÁ, S., HINDLS, R., ČABLA, A. (2011). Conjunctural Evolution of the Czech Economy. *Statistika: Statistics and Economy Journal*, 91(3): 4–17. [article in a journal with DOI]: Stewart, M. B. (2004). The Employment Effects of the National Minimum Wage [online]. *The Economic Journal*, 114(494): 110–116. <http://doi.org/10.1111/j.0013-0133.2003.0020.x>.
Please **add DOI numbers** to all articles where appropriate (prescribed format = link, see above).

Tables
Provide each table on a separate page. Indicate position of the table by placing in the text "insert Table 1 about here". Number tables in the order of appearance Table 1, Table 2, etc. Each table should be titled (e.g. Table 1 Self-explanatory title). Refer to tables using their numbers (e.g. see Table 1, Table A1 in the Annex). Try to break one large table into several smaller tables, whenever possible. Separate thousands with a space (e.g. 1 528 000) and decimal points with a dot (e.g. 1.0). Specify the data source below the tables.

Figures
Figure is any graphical object other than table. Attach each figure as a separate file. Indicate position of the figure by placing in the text "insert Figure 1 about here". Number figures in the order of appearance Figure 1, Figure 2, etc. Each figure should be titled (e.g. Figure 1 Self-explanatory title). Refer to figures using their numbers (e.g. see Figure 1, Figure A1 in the Annex).

Figures should be accompanied by the *.xls, *.xlsx table with the source data. Please provide cartograms in the vector format. Other graphic objects should be provided in *.tif, *.jpg, *.eps formats. Do not supply low-resolution files optimized for the screen use. Specify the source below the figures.

Formulas
Formulas should be prepared in formula editor in the same text format (Times 12) as the main text and numbered.

Paper Submission
Please email your papers in *.doc, *.docx or *.pdf formats to statistika.journal@czso.cz. All papers are subject to double-blind peer review procedure. Articles for the review process are accepted continuously and may contain tables and figures in the text (for final graphical typesetting must be supplied separately as specified in the instructions above). Please be informed about our Publication Ethics rules (i.e. authors responsibilities) published at: http://www.czso.cz/statistika_journal.

Managing Editor: Jiří Novotný
Phone: (+420) 274 054 299 | **fax:** (+420) 274 052 133
E-mail: statistika.journal@czso.cz | **web:** www.czso.cz/statistika_journal
Address: Czech Statistical Office | Na padesátém 81 | 100 82 Prague 10 | Czech Republic

Subscription price (4 issues yearly)
CZK 66 per copy + postage.
Printed copies can be bought at the Publications Shop of the Czech Statistical Office (CZK 66 per copy).
Address: Na padesátém 81 | 100 82 Prague 10 | Czech Republic

Subscriptions and orders
Czech Statistical Office | Information Services
Na padesátém 81 | 100 82 Prague 10 | Czech Republic
Phone: (+420) 274 052 733, (+420) 274 052 783
E-mail: objednavky@czso.cz

Design: Toman Design
Layout: Ondřej Pazdera
Typesetting: Družstvo TISKOGRAF, David Hošek
Print: Czech Statistical Office

102nd year of the series of professional statistics and economy journals of the State Statistical Service in the Czech Republic: ***Statistika*** (since 1964), ***Statistika a kontrola*** (1962–1963), ***Statistický obzor*** (1931–1961) and ***Československý statistický věstník*** (1920–1930).

Published by the Czech Statistical Office
ISSN 1804-8765 (Online)
ISSN 0322-788X (Print)
Reg. MK CR E 4684