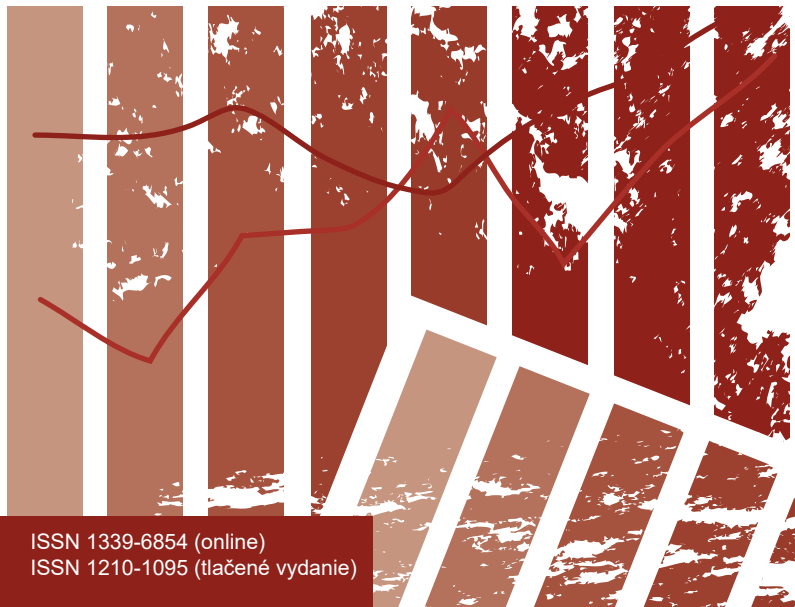
Štatistický úrad Slovenskej republiky
The Statistical Office of the Slovak Republic

SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

vedecký časopis/scientific journal

3/2023
ročník 33



ŠTATISTICKÝ ÚRAD SLOVENSKEJ REPUBLIKY

ISSN 1339-6854 (online)
ISSN 1210-1095 (tlačené vydanie)

SLOVENSKÁ ŠTATISTIKA A DEMOGRAFIA

Recenzovaný vedecký časopis založený v roku 1991. Jednotlivé čísla časopisu zverejňujeme aj v elektronickej podobe na ssad.statistics.sk a na slovak.statistics.sk. Názory autorov článkov sa nemusia zhodovať s názormi vydavateľa.

SLOVAK STATISTICS AND DEMOGRAPHY

The scientific peer-reviewed journal founded in 1991. Individual copies of the journal are available to readers in electronic form at the websites ssad.statistics.sk and slovak.statistics.sk. The opinions of the authors do not necessarily correlate with the opinions of the publisher.

Adresa redakcie/Address of Editorial Office
Slovenská štatistika a demografia
Štatistický úrad SR
Lamačská cesta 3/C, 840 05 Bratislava 45
Slovenská republika

E-mailová adresa/E-mail adress
SSaD@statistics.sk

ssad.statistics.sk
www.statistics.sk

OBSAH/CONTENTS

EDITORIÁL

Vážení čitatelia,

Ing. Helena Glaser-Opitzová

monotematické číslo časopisu Slovenská štatistika a demografia je venované téme, ktorá priťahuje čoraz väčší záujem štatistických úradov – **novým štatistikám a novým zdrojom údajov**.

Rozsah potenciálnych zdrojov údajov relevantných pre oficiálnu štatistiku zahŕňa údaje zozbierané priamo štatistickým úradom prostredníctvom štatistických zisťovaní, administratívne zdroje údajov a iné externé údaje, ktoré sú väčšinou vo vlastníctve súkromných spoločností.

Na rozdiel od minulosti, štatistický produkt patriaci do portfólia oficiálnej štatistiky si dnes môže vyžadovať kombináciu alebo integráciu rôznych vstupných zdrojov údajov, vtedy hovoríme o multizdrojovej štatistike alebo sa jeden zdroj údajov môže opätovne použiť na viacero štatistických produktov. Kombináciou zdrojov údajov môžu štatistické úrady vytvárať podrobnejšie a včasnejšie štatistiky a rýchlejšie reagovať na udalosti v spoločnosti. Kombináciou údajov zo štatistických zisťovaní s už dostupnými administratívnymi údajmi a inými externými údajmi vrátane „big data“ je možné ušetriť náklady na zber a spracovanie údajov a znížiť zaťaženie respondentov štatistických zisťovaní. Štatistiky zostavované z viacerých zdrojov však prinášajú množstvo nových problémov, ktoré je potrebné prekonať, kým bude výsledná kvalita výstupov dostatočná a kým sa tieto štatistiky budú môcť vytvárať efektívne. Tvorbu štatistík z viacerých zdrojov komplikuje aj skutočnosť, že sa môžu vyskytovať v rôznych variantoch, keďže súbory údajov možno kombinovať rôznymi spôsobmi.

Cieľom monotematického čísla je prezentovať využitie nových zdrojov údajov konkrétne pre oblasť cenovej štatistiky. Zdrojom údajov pre výpočet elementárnych cenových indexov, ktoré predstavujú základné stavebné prvky pre zostavenie Indexu spotrebiteľských cien, môžu byť okrem údajov zo štatistických zisťovaní a administratívnych zdrojov predovšetkým transakčné údaje obchodných reťazcov, napríklad pre oblasť potravín a nealkoholických nápojov alebo údaje získané web-scrapingom napríklad z internetových stránok online predajcov čiernej a bielej techniky. V článkoch týkajúcich sa využitia týchto nových zdrojov údajov autori poukazujú aj na riešenie parciálnych metodologických problémov, ktoré súvisia s ich implementáciou do produkcie cenovej štatistiky.

Čitateľa môže zaujať aj informácia o konaní medzinárodného workshopu krajín V4 týkajúceho sa rovnako modernizácie cenových štatistík alebo informácia o iniciatívach Eurostatu v oblasti využívania Big Data v oficiálnych štatistikách a transformácii týchto iniciatív do podmienok Štatistického úradu Slovenskej republiky.

Nové zdroje údajov prinášajú na scénu oficiálnej štatistiky aj nové informačné technológie. V oficiálnej štatistike sa prejavil nový trend, revolúcia v oblasti softvéru s otvoreným zdrojovým kódom sa dostala aj do sveta oficiálnej štatistiky. Objavili sa dve softvérové prostredia, ktoré sú vhodné na úlohy oficiálnej štatistiky, a to R a Python. Zatiaľ čo Python sa považuje za výpočtovo efektívnejší, R sa považuje za vhodnejší na štatistické účely nakoľko v R existujú balíky pre takmer všetky štatistické operácie, od výberu vzoriek až po vizualizáciu údajov. Väčšina národných štatistických organizácií (štatistických úradov) v rámci EŠS prechádza zo „starých softvérových balíkov“ väčšinou založených na komerčných riešeniach práve na prostredie R. Touto témou sa zaoberá aj jeden z vedeckých článkov, ktorý predstavuje balík R, ktorý autor vyvinul pre spracovanie transakčných údajov obchodných reťazcov, tzv. scaner data a pre následný výpočet rôznych cenových indexov, bilaterálnych aj multilaterálnych. Jeden z informatívnych článkov zasa opisuje návrhy riešenia na sťahovanie údajov z internetu v jazyku Python a moduly, v ktorých možno tento proces realizovať.

Uvedeným číslom chce redakcia časopisu Slovenská štatistika a demografia prispieť k informáciám o inovačných projektoch v rámci štatistiky zameraných na používanie nových zdrojov údajov a s tým spojené zmeny v metodike, procesoch a v IT nástrojoch.

Ing. Helena GLASER-OPITZOVÁ

Autorka je generálnou riaditeľkou Sekcie všeobecnej metodiky, registrov a koordinácie národného štatistického systému a zároveň hlavným štatistikom národného štatistického systému.

EDITORIAL

Dear readers,

The monothematic issue of the journal Slovak Statistics and Demography is dedicated to the topic that is attracting increasing interest of statistical offices - **new statistics and new data sources**.

The range of potential data sources relevant for official statistics includes data collected directly by the statistical office through statistical surveys, administrative data sources and other external data, mostly owned by private companies.

In contrast to the past, a statistical product belonging to the portfolio of official statistics may today require a combination or integration of different input data sources, then we are dealing with multi-source statistics or one data source can be reused for several statistical products. By combining data sources, statistical offices can create more detailed and timely statistics and respond more quickly to the events in society. By combining data from statistical surveys with already available administrative data and other external data, including "big data", it is possible to save costs for data collection and processing and reduce respondent burden when carrying out statistical surveys. However, statistics compiled from several sources bring a multitude of new problems that need to be overcome before the resulting quality of outputs is sufficient and before these statistics can be created efficiently. The production of statistics from multiple sources is also complicated by the fact that they can appear in different variants, as data sets can be combined in different ways.

The aim of the monothematic issue is to present the use of new data sources specifically for the area of price statistics. In addition to data from statistical surveys and administrative sources, the source of data for the calculation of elementary price indices, representing the basic building blocks for the compilation of the Consumer Price Index, can primarily be transaction data of retail chains, for example in the area of food and non-alcoholic beverages, or data obtained by web-scraping, for example from the Internet websites of online sellers of black and white technology. In the articles related to the use of these new data sources, the authors also point to the solution of partial methodological problems related to their implementation in the production of price statistics.

The reader may also be interested in information on holding of an international workshop of the V4 countries, also related to the modernization of price statistics, or information on Eurostat's initiatives in the field of using Big Data in official statistics and the transformation of these initiatives into the conditions of the Statistical Office of the SR.

New data sources also bring new information technologies to the main focus of official statistics. A new trend has emerged in official statistics, the open source software revolution has entered the world of official statistics. Two software environments have emerged that are suitable for the tasks of official statistics, namely R and Python. While Python is considered more computationally efficient, R is considered more suitable for statistical purposes as there are packages in R for almost all statistical operations, from sample selection to data visualization. The majority

of national statistical organizations (statistical offices) within the ESS are switching from "old software packages" mostly based on commercial solutions to the R environment. This topic is also dealt with one of the authors of the scientific articles, presenting the R package developed by the author for the processing of transactional data of business chains, the so-called scanner data and for the subsequent calculation of various price indices, both bilateral and multilateral. One of the informative articles, in turn, describes proposals for a solution for downloading data from the Internet in Python and modules in which this process can be implemented.

With this issue, the Editorial Board of the journal Slovak Statistics and Demography wishes to contribute to information about the innovative projects within the field of statistics focused on the use of new data sources and the related changes in methodology, processes and IT tools.

Ing. Helena GLASER-OPITZOVÁ

The author is the Director of the General Methodology, Registers and Coordination of National Statistical System Directorate and at the same time the Chief Statistician of the National Statistical System.

Jacek BIAŁEK
Statistics Poland, Department of Trade and Services
University of Lodz, Department of Statistical Methods

SCANNER DATA PROCESSING AND PRICE INDEX CALCULATIONS IN THE PRICEINDICES R PACKAGE

SPRACOVANIE ÚDAJOV ZO SKENEROV A VÝPOČET CENOVÝCH INDEXOV V R BALÍKU CENOVÝCH INDEXOV

ABSTRACT
The main purpose of the work is to present the utility of the *PriceIndices* R-package in the field of analysing the dynamics of scanner prices. The package can be a useful IT tool for National Statistical Offices for at least several reasons. First, it enables a comprehensive transition from raw scanner data to price indices, taking into account all intermediate steps (e.g., product classification or product matching). Secondly, the package's functions, through their parameterization, allow experimental work to be carried out, such as setting thresholds for data filters or the length of the time window for multilateral indices. Third, the package is written in the free R environment, which does not generate any additional costs for its implementation.
The presentation of this R package is divided into the following areas: scanner data preparing, data set characteristics, bilateral index calculations, multilateral index calculations, extensions of multilateral indices, aggregation of index results, and comparison of price indices. The paper presents examples concerning main package features on the basis of data sets which are available in the *PriceIndices* package.

ABSTRAKT
Hlavným cieľom práce je predstaviť využitie balíka R cenových indexov v oblasti analýzy dynamiky scanner data (transakčné dáta nazývané aj údaje zo skenerov). Balík môže byť užitočným IT nástrojom pre národné štatistické úrady minimálne z niekoľkých dôvodov. Po prvé, umožňuje komplexný prechod od nespracovaných údajov zo skenerov k cenovým indexom, pričom zohľadňuje všetky čiastkové etapy (napr. klasifikáciu produktov alebo párovanie produktov). Po druhé, funkcie balíka prostredníctvom ich parametrizácie umožňujú vykonávať experimentálne práce, ako je nastavenie prahov pre dátové filtre alebo dĺžky intervalu (časového okna) pre multilaterálne indexy. Po tretie, balík je vytvorený vo voľnom prostredí R, čo nevytvára žiadne dodatočné náklady na jeho implementáciu.
Prezentácia balíka R je rozdelená do nasledujúcich oblastí: príprava skenerových dát, charakteristiky dátových súborov, výpočty bilaterálnych indexov, výpočty multilaterálnych indexov, rozšírenia multilaterálnych indexov, agregácia výsledkov indexov a porovnávanie cenových indexov. Článok prezentuje príklady týkajúce sa hlavných funkcií balíka na základe dátových súborov, ktoré sú dostupné v balíku cenových indexov.

KEY WORDS
scanner data, PriceIndices package, price indices, multilateral indices

KĽÚČOVÉ SLOVÁ
údaje zo skenerov, balík Cenových indexov, cenové indexy, multilaterálne indexy

1. INTRODUCTION

The term “scanner data” refers to transaction data that specify turnover and numbers of items sold by Global Trade Article Number (GTIN) code, European Article Number (EAN) code or other barcodes [10]. Scanner data have numerous advantages compared to traditional survey data: such data sets are much larger than the traditional ones and contain complete transaction records, i.e. information about prices and quantities along with the additional information about products, including the grammage, unit, label with description, VAT, etc. In other words, scanner data contain full information at the most detailed item level. The methodology for inflation measurement by using scanner data has strongly evolved over the last few years (see for instance: [1, 3, 5, 6, 7, 12, 13, 14, 20, 21, 23]).

Despite many benefits of using scanner data, their use is still associated with many methodological challenges and problems. The main challenge on the IT side is to build the right software that efficiently and quickly enables the transition from raw scanner data to price indices. An example of such software is the *PriceIndices* package written in the R environment [2] and available on CRAN and GitHub. Although some packages dedicated to scanner data and price indices are available in the R environment (e.g. *IndexNumR* by Graham White [10], *IndexNumber* package by Alejandro Saavedra-Nieves and Paula Saavedra-Nieves [15] or *multilateral* by Matthew Stansfield [16]), their functionalities are much lower compared to the *PriceIndices* package and they do not operate on a time variable in the sense of year-month-day. The PriceIndices package contains a rich set of functions for the proceding scanner data and price index functions (over 200 functions), which sets it apart from the other R packages in this area. As it was mentioned above, the package can be a useful IT tool for National Statistical Offices because: (1) it enables a comprehensive transition from raw scanner data to price indices, taking into account all intermediate steps (e.g., product classification or product matching); (2) the package's functions, through their parameterization, allow experimental work to be carried out, such as setting thresholds for data filters or the length of the time window for multilateral indices; (3) the package is written in the free R environment, which does not generate any additional costs for its implementation. The package has an open-source code so it can be modified and extended by the Statistical Office.

The main purpose of the paper is to present the utility of the *PriceIndices* R package, which can be divided into the following areas: scanner data preparing, data set characteristics, price index calculations, extensions of multilateral indices, aggregation of index results, as well as comparison of price indices [9]. The paper presents examples concerning main package features on the basis of data sets which are available in the *PriceIndices* package (e.g. *milk*, *sugar* or *coffee* datasets). As a result, the reader will have the opportunity to perform all the examples presented on their own.

The structure of the paper is as follows: Section 2 presents the main package features for scanner data preparing, including functions for classifying and matching products, data filtering and imputing missing prices; Section 3 shows how to obtain some statistical characteristics of the analysed scanner products, Section 4 describes price index methods which are available in *PriceIndices*, Section 5 discusses the aggregation methods; Section 6 shows some ways of presenting and comparing price indices and Section 7 provides the main conclusions. The attached Appendix contains

reproducible R-language codes using the PriceIndices package and its included scanner data sets.

2. SCANNER DATA PREPARING

In general, the procedure of scanner data processing consists of the following steps: data cleaning ► extracting information from the product description ► classification of products into COICOP (Classification of individual consumption by purpose) groups ► product matching ► product filtering ► data standardization ► imputing missing prices (optional) ► price index calculations. All these procedures can be performed in the *PriceIndices* package. Initial cleaning of the data set, i.e. removal of data gaps, removal of zero prices and quantities and standardization of variable types, can be performed using the **data_preparing** function. Sometimes a retail chain supplies scanner data frames with additional columns specifying the basis weight (grammage) and the sales unit of the products. This information, however, can be sometimes hidden in the product label and then needs to be extracted. The **data_unit** function returns the user's data frame (provided as a data frame by *data* parameter) with two additional columns: *grammage* and *unit*. The values of these columns are extracted from product descriptions on the basis of provided units. Please note, that the function takes into consideration a sign of the multiplication, e.g. if the product description contains: '2x50g', we will obtain: *grammage* = 100 and *unit* = g for that product (for the *multiplication* parameter set to 'x'). To get a good understanding of the function for extracting information from a product label, please run **Example 1** from the **Appendix.**

2.1 PRODUCT CLASSIFICATION

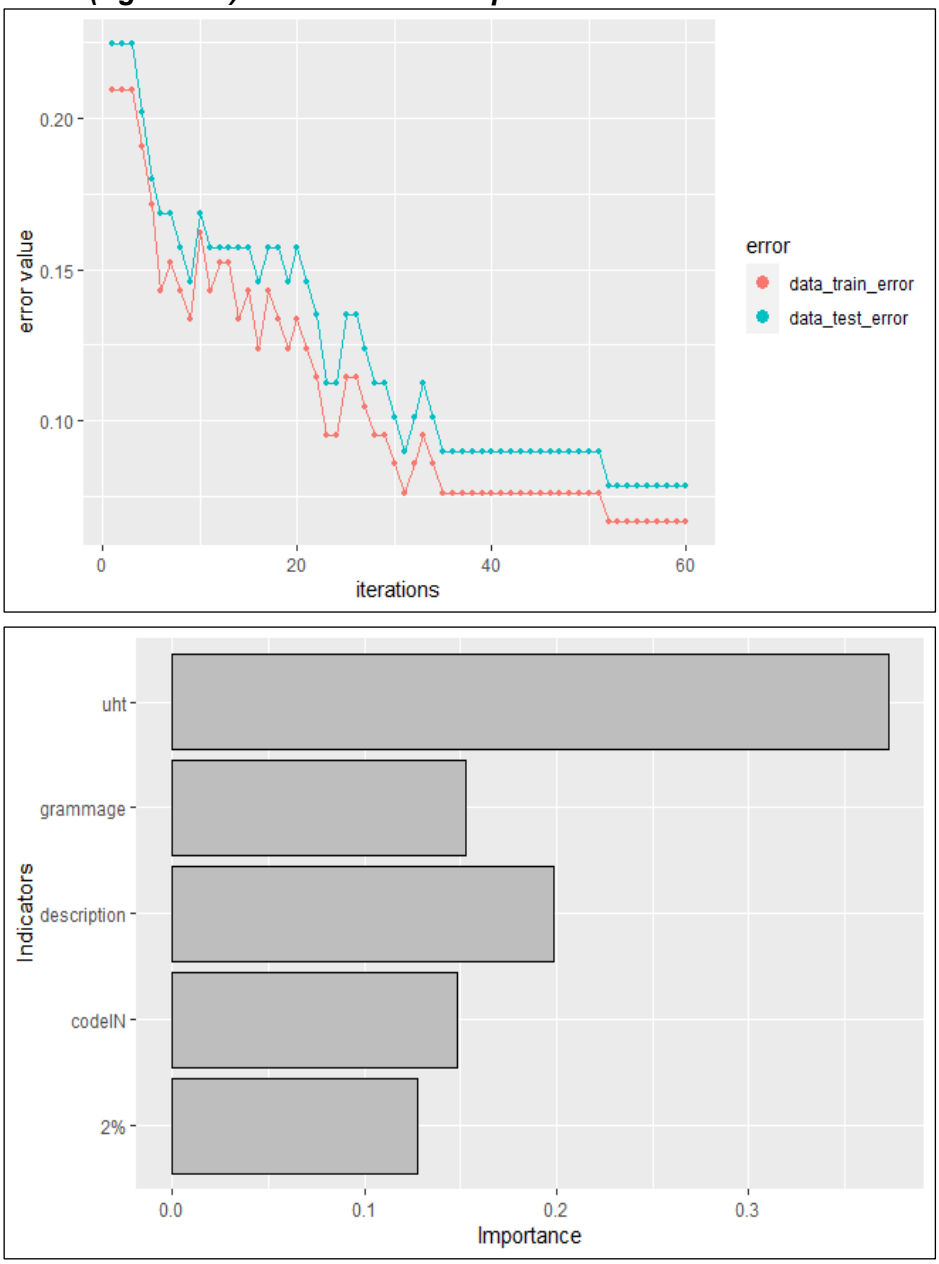
We can use the **data_selecting** or **data_classifying** function to classify products into COICOP (Classification of individual consumption by purpose) groups.

The **data_selecting** function returns a subset of the user's data set obtained by selection based on keywords and phrases defined by parameters: *include*, *must* and *exclude* (an additional *Coicop* column is optional and provides an a priori known product group). These parameters define those sets of phrases or keywords that can, must or cannot appear in the product label, respectively. The function is based on text analysis provided by *stringr* package [24] and is very effective when the product labels (*description* column) are of good quality. **Example 2** (Appendix) presents a selection of UHT milks excluding low-fat milks from the milk collection included in the *PriceIndices* package.

Alternatively, the **data_classifying** function can be used to classify the products into homogeneous groups. This function predicts product classification by COICOP levels using the selected, previously trained model for gradient-boosted decision trees (Chen et al., 2021). Thus, first we need to build a model for the product classification using the **model_classification** function, which is based on XGBoost algorithm from *xgboost* package by [4]. The algorithm is modified to take into account not only numeric columns (which is the case in the standard XGBoost algorithm) but also non-numeric columns (such as product labels) or artificially created binary columns based on the presence or absence of user-set keywords in the product description. We can save the built model to disk (**save_model**) and import it at any time **(load_model**). To get a good understanding of the function for product ML classification, please run **Example 3** (Appendix). In the above-mentioned example the *data COICOP* data set with correctly classified milk products to local COICOP 6 product groups is used. The user

may control the *accuracy* value obtained on the basis of training and testing data sets (*data_train* and *data_test*, respectively) and observe the importance of the used indicators (see Graph No. 1).

Graph No. 1: The level of errors while ML model training (left side) and importance of the used indicators (right side) obtained in Example 3



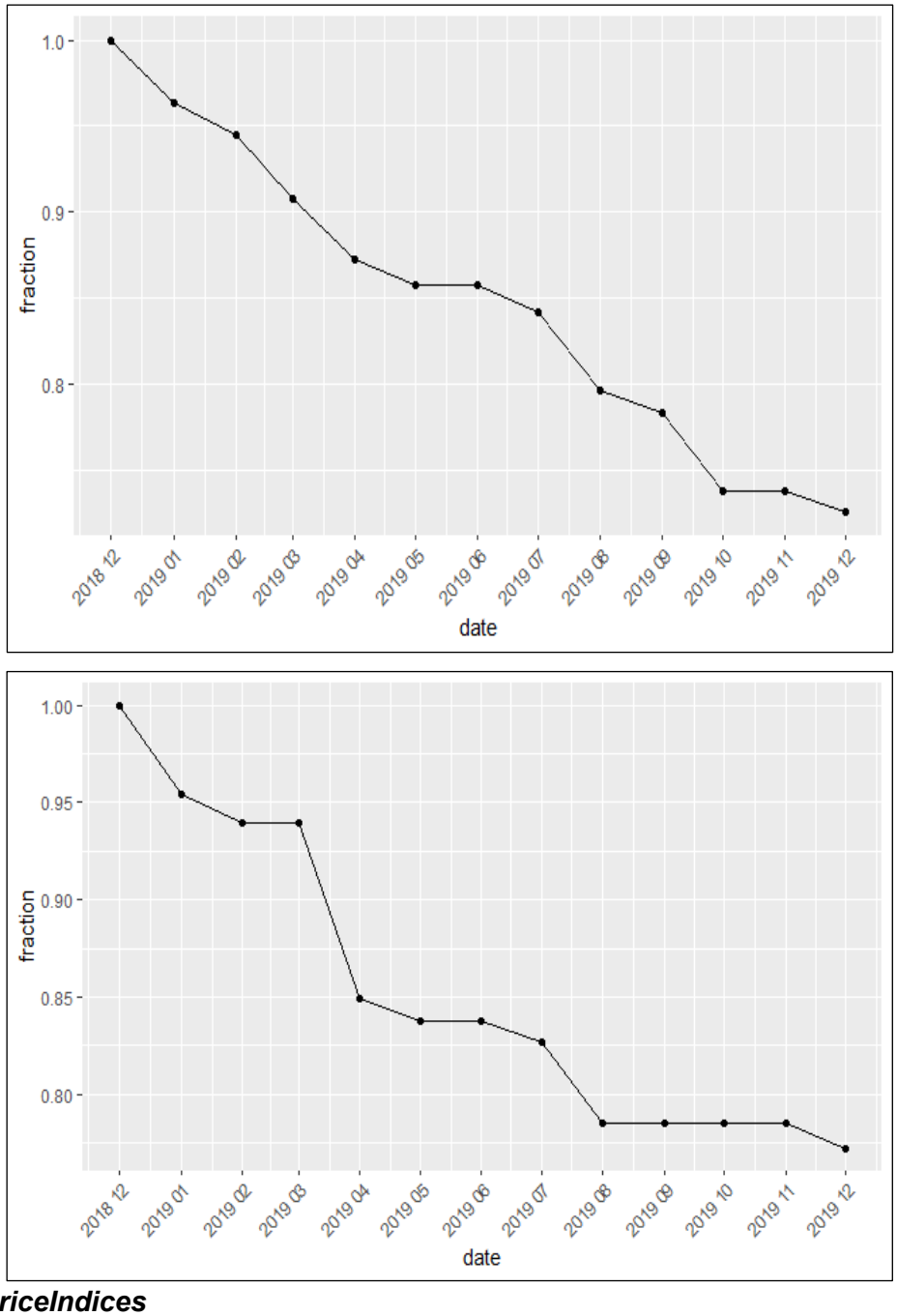
Source: PriceIndices

2.2 PRODUCT MATCHING

The next step in the procedure is to match products over time. This is an important step, especially when considering the lowest level of data aggregation (barcode level). This is because it can happen that a product changes the colour of its packaging and consequently its description and the EAN code, while in terms of quality it is the same product. Therefore, it is important to make sure that we follow the price evolution of the same product. The function for product matching (**data_matching**) is based on *recling* R package by van der Laan (2018) and it depends on the set of chosen

columns for matching. Product matching can be done only on the basis of its label (in which case the *onlydescription* parameter has the value TRUE) or with the use of the retailer codes (*codeIN*) and/or external codes (EAN, SKU - see the *codeOUT* parameter). These two approaches can lead to different numbers of the matched products (see **Example 4**). Generally, in most cases of data matching, we compare descriptions of products, and *PriceIndices* uses the Jaro-Winkler string similarity measure for this purpose. One should also be aware that the ratio of the matched products to all the available products is generally a decreasing function of time (see Graph No. 2 with results for *milk* and *coffee* products obtained after using the **matched_fig** function).

Graph No. 2: The ratio of the matched and available milk (left side) and coffee (right side) products obtained via the matched_fig function



Source: PriceIndices

2.3 DATA FILTERING

In the literature there is an ongoing discussion about whether or not to use data filters for scanner data, and if so, what kind of filters to apply. As a rule, scanner data indices are calculated using a dynamic approach, with most countries opting for the monthly chain Jevons index. This method is commonly referred to as the dynamic method [8]. The dynamic basket is determined using turnover figures of individual products in two adjacent months, i.e. the product is included in the sample if its turnover is above a fixed threshold determined by the number of products in a given product group (the *low sales filter*). A filter for removing products with extreme price changes (*extreme price filter*) and a filter for eliminating recalled products from sale (*dump price filter*) are also often considered [19]. These filters can be used separately or independently together using the **data_filtering** function (see **Example 5** for coffee products). Sometimes filtering is even applied to the weighted multilateral indices, since the reduction of the data set always results in savings in terms of time needed for index computations.

2.4 DATA STANDARDIZATION

If we know the product's grammage and unit of sale, standardizing prices and quantities to a fuller unit seems reasonable. After all, manufacturers often use the "trick" of lowering the grammage of a product with an unchanged price, which de facto means an increase in the price of the product. In such a case, a qualitative adjustment to normalize price and quantity is needed. In the *PriceIndices* package, this can be achieved by using the **data_norm** function (see **Example 6**).

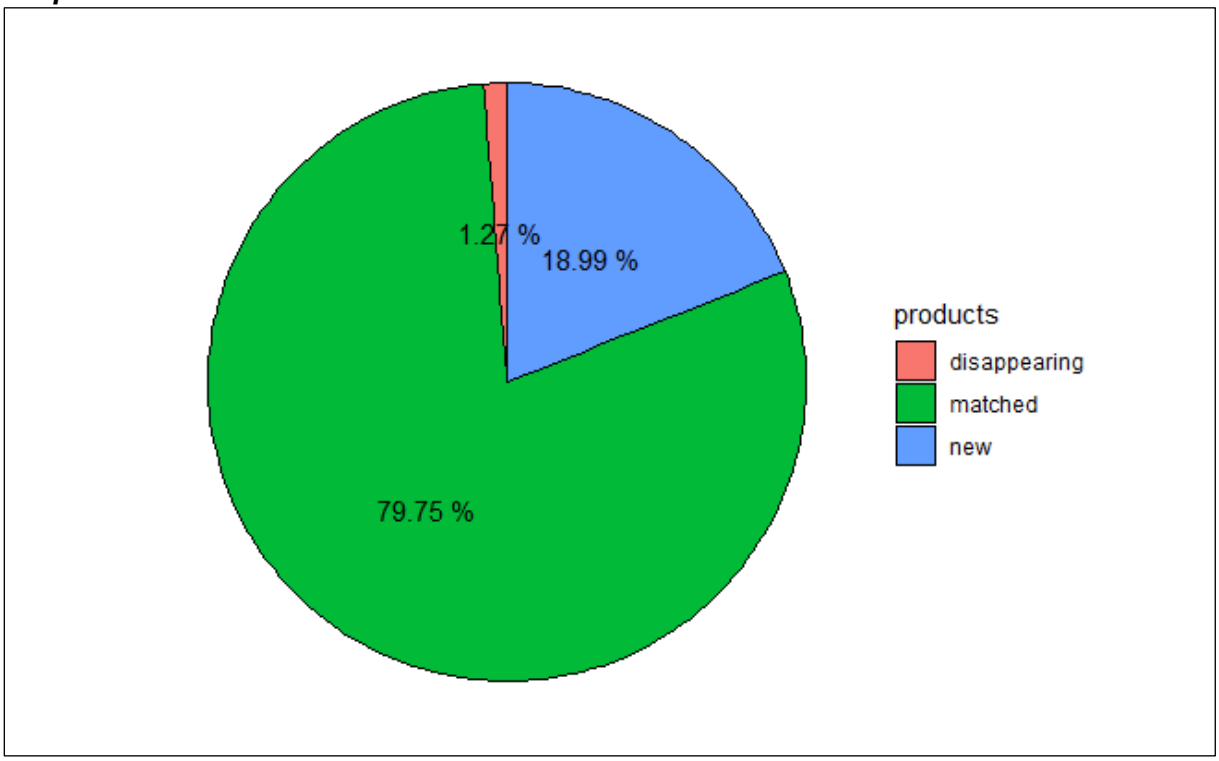
2.5 IMPUTING MISSING OR ZERO PRICES

The **data_imputing** function imputes missing prices (unit values) and (optionally) zero prices by using *carry forward / backward* prices. The imputation can be done for each outlet separately or for aggregated data (see the *outlets* parameter). If a missing product has a previous price then that previous price is carried forward until the next real observation. If there is no previous price then the next real observation is found and carried backward. The quantities for the imputed prices are set to zeros. The function returns a data frame (monthly aggregated) which is ready for price index calculations**.**

3. DATA SET CHARACTERISTICS

The *PriceIndices* package includes a number of functions for determining the characteristics of the analyzed scanner data sets (see [2]). In particular, we can analyze the differences in product sales levels (**sales_group**), correlations between product the prices and the quantities (**pqcor_fig**) or the level of product matching (**matched_fig**). One useful feature is certainly the **product** function, which analyzes the relationship the between matched products, new products, disappearing products and all available products in the given period (see **Example 7** and Graph No.3).

Graph No. 3: The matched, new and disappering coffee products – results obtained via the products function



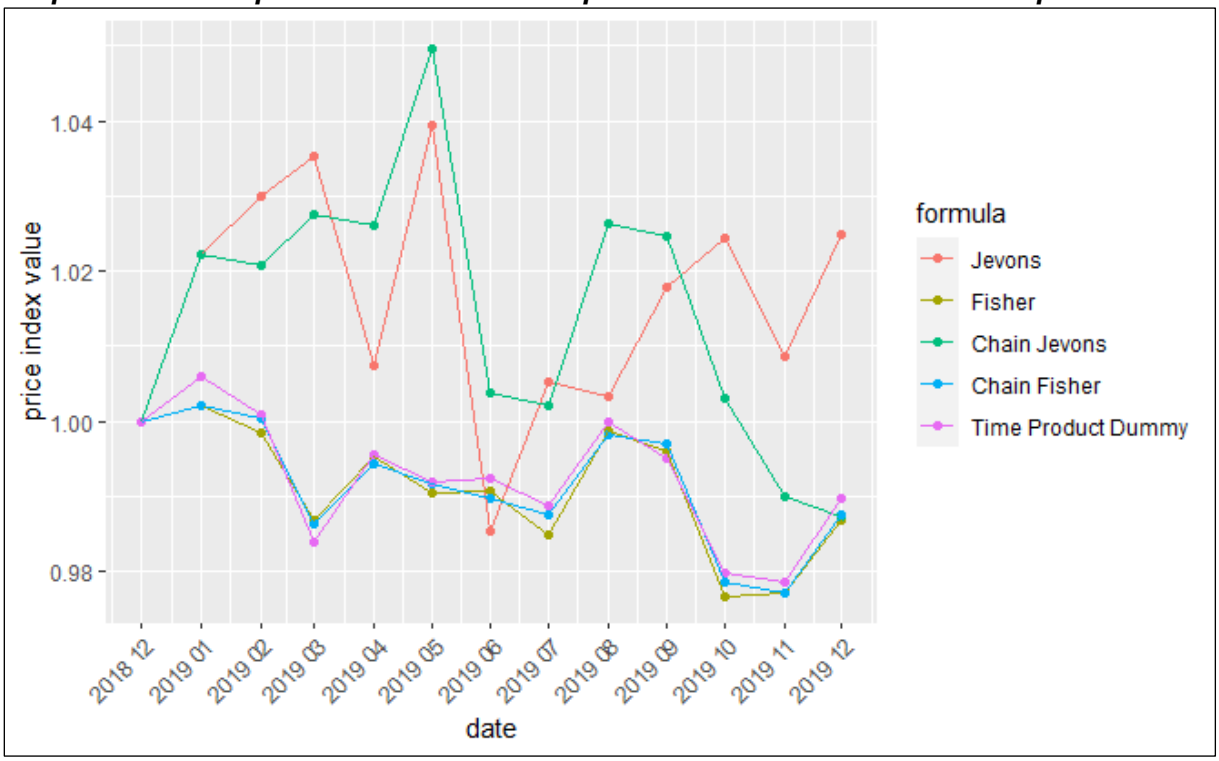
Source: PriceIndices

4. PRICE INDEX CALCULATIONS

The current version of the *PriceIndices* package (ver. 0.1.7) includes: 6 functions for calculating unweighted price indices, 30 functions for determining weighted bilateral indices, 35 functions dedicated to chain indices, and 22 functions for determining multilateral indices (unweighted and weighted – see CPI Manual (2004)). The **price_indices** function is a general function that allows the User to calculate price indices on a given dataset and for different parameters (such as the time window length in case of multilateral indices or the elasticity of substitution in case of the CES index). The above-mentioned function provides a data frame with index results. To compare the price indices determined in this way, we can use the **compare_indices_df** function (see **Example 8** and Graph No. 4).

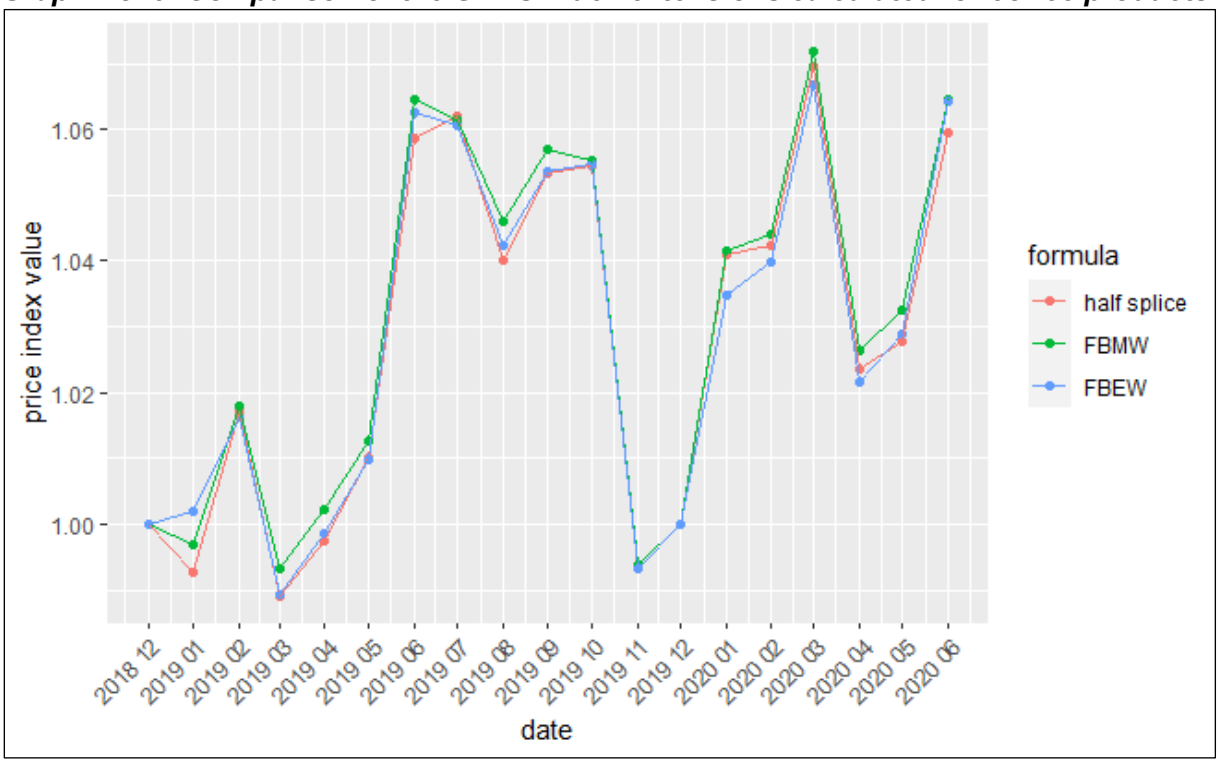
The package also includes extensions known from the literature for multilateral indices [9], which allow the inclusion of data from the following month without having to revise the price indices determined for the previous time window. Among the above-mentioned methods there are not only splicing methods (combining the old and the new time windows), but also FBEW (Fixed Base Expanding Window) or FBMW (Fixed Base Moving Window) methods are included - cf. [2, 3]. **Example 9** and Graph n. 5 present the multilateral GEKS index calculated for coffee products and extended by using a half splice method together with the FBEW and the FBMW methods.

Graph NO. 4: Comparison of the selected price indices calculated for milk products



Source: PriceIndices

Graph No. 5: Comparison of the GEKS index extensions calculated for coffee products



Source: PriceIndices

5. AGGREGATION OF PARTIAL INDEX RESULTS

Aggregation of partial price indices over outlets within a retail chain is advisable, as long as the retail chain follows a regional pricing policy. It should be borne in mind that if the number of outlets of a given retailer is large (e.g., it is several hundred), the time required to determine the final price index may increase substantially.

In the *PriceIndices* package, the **final_index** function allows the User to calculate any price index with optional consideration of aggregation of sub-indices calculated for the subgroups of products (then the User needs to indicate the grouping variable) and/or outlets. The user can choose from several possible functions that aggregate partial results, such as an arithmetic mean, geometric mean, the Laspeyres, Paasche or Fisher formula. **Example 10** demonstrates the aggregation of the Fisher indices over subgroups of milk products (*groups* = TRUE) and the over outlets (defined in the *retID* column with *outlets* = TRUE) where the Laspeyres formula is used for that aggregation (*aggr* = "laspeyres"). The obtained results, with the January 2019 as the fixed base month, are presented in Tab. No. 1.

Table No. 1: The final (fixed base) Fisher price index obtained after using the Laspeyres aggregation over subgroups of milk products and over outlets

Period	The final Fisher index
2019-01	1.0000000
2019-02	1.0002910
2019-03	0.9803976
2019-04	0.9929258
2019-05	0.9895008
2019-06	0.9872751

Source: PriceIndices

6. COMPARISON OF INDICES

The *PriceIndices* package includes two functions for a simple graphical comparison of price indices and two functions for calculating distances between indices. The first one, i.e. **compare_indices_df**, is based on the syntax of the **price_indices** function and thus it allows us to compare price indices calculated on the same data set. The second function, i.e. **compare_indices_list**, has a general character since its first argument is a list of data frames which contain results obtained by using the **price_indices** or **final_index** functions. The third one, i.e. **compare_distances**, calculates (average) distances between price indices, i.e. the mean absolute distance or root mean square distance is calculated. The next function, **compare_to_target**, allows to compute distances between indices from the selected index group and the indicated target price index. The last function, **compare_indices_jk**, presents a comparison of selected indices obtained by using the jackknife method.

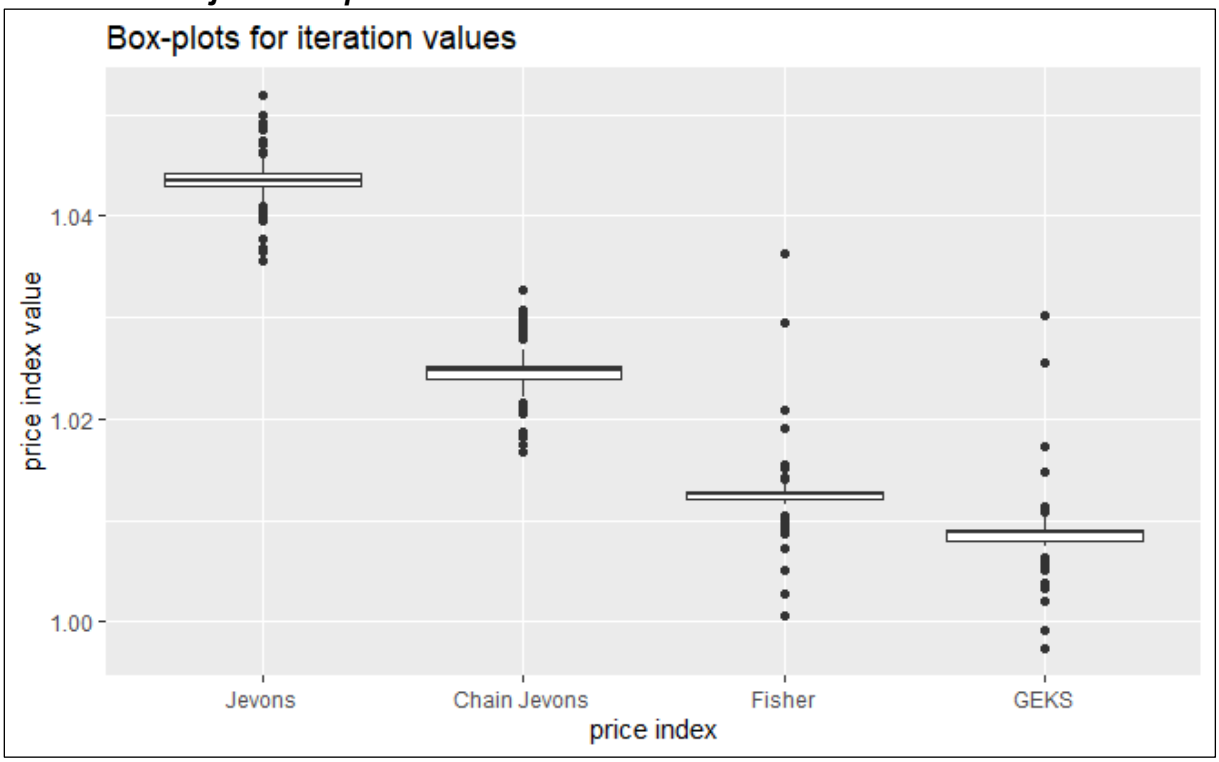
Example 11 and Table No. 2 presents a comparison of the selected monthly price indices calculated for the coffee products sold in 2019 in one of Polish retail chains and **Example 12** and Graph No 6 presents sensitivity of these indices on the product rotation in terms of the jackknife procedure (i.e. in each iteration one product is removed from the sample).

Table No 2: **Mean absolute distances (p.p.) between the selected monthly price indices calculated for the coffee products**

Formula	Jevons	Chain Jevons	Fisher	GEKS
Jevons	0.000	0.696	1.797	1.785
Chain Jevons	0.696	0.000	1.634	1.466
Fisher	1.797	1.634	0.000	0.812
GEKS	1.785	1.466	0.812	0.000

Source: PriceIndices

Graph No. 6: Sensitivity of the selected yearly price indices on the product rotation in terms of the jackknife procedure



Source: PriceIndices

7. CONCLUSIONS

The PriceIndices package is a full-featured IT tool for the procedure of scanner data. Its main advantages are: (a) no maintenance costs (R environment); (b) working with a time variable of the year-month-day type and operating on unit values, which makes the package very close to the practice of statistical offices; (c) wide functionality allowing to carry out the whole path from raw scanner data to the calculation of price indices; (d) wide range of available price index formulas.

The PriceIndices package can be useful for both the academics conducting research on the theory and practice of price indices and for the employees of statistical offices. Since many important features of the package are not discussed in this article (e.g., the author's price indices: the Białek index, the hybrid index, the GEKS-L, GEKS-GL, GEKS-AQU, GEKS-AQI indices, and the Bennet and Montgomery price and quantity indicators as well as many others) we refer the interested reader to the package's documentation available at:
https://cran.r-project.org/web/packages/PriceIndices/PriceIndices.pdf.

The author hopes that the added value of this article is the presentation of the PriceIndices package based on real scanner data sets which are available in the package. Some contribution of the paper is also the demonstration of the various stages of scanner data proceeding supported by reproducible R language codes included in the Appendix.

ACKNOWLEDGEMENTS

The author would like to thank the organizers of the workshop *Modernization of the consumer price statistics* held under the auspices of the Statistical Office of the Slovak Republic in Bratislava on March 23-24, 2023, for the invitation and the opportunity to present the *PriceIndices* package. The presentation delivered at this workshop formed the basis for this article.

BIBLIOGRAPHY

[1] BIAŁEK, J.: The general class of multilateral indices and its two special cases, Paper presented at the 17th Meeting of the Ottawa Group on Price Indices, Rome, Italy, 2022

[2] BIAŁEK, J.: Scanner data processing in a newest version of the PriceIndices package, Statistical Journal of the IAOS, 38 (4), p. 1369 – 1397, 2022.

[3] BIAŁEK, J. – ROSZKO-WÓJTOWICZ, E.: Potential reasons for CPI chain drift bias while using electronic transaction data, Technological and Economic Development of Economy, 29(2), 2023, p. 564-590.

[4] CHEN, T., HE, T. – BENESTY, M. – KHOTILOVICH, V. – TANG, Y – CHO, H. - CHEN, K. – MITCHELL, R. – CANO, I. – ZHOU, T. – LI, M. – XIE, J. – LIN, M. - GENG, Y. – LI, A.: Xgboost: Extreme Gradient Boosting. R package version 1.3.2.1, 2021.

[5] CHESSA, A. – GRIFFIOEN. R.: Comparing scanner data and web scraped data for consumer price indices. Report, Statistics Netherlands, 2016.

[6] CHESSA, A.: Comparisons of QU-GK indices for different lengths of the time window and updating methods. In: Second meeting on multilateral methods organised by Eurostat, 2017.

[7] DIEWERT, W. E. – FOX, K.J.: Substitution bias in multilateral methods for CPI construction using scanner data. UNSW Business School Research Paper (2018-13), 2018.

[8] EUROSTAT: Practical guide for processing supermarket scanner data. Harmonised Index of Consumer Prices. Luxembourg: Publications Office of the European Union, 2018, ISBN 978-92-79-76861-3.

[9] EUROSTAT: Guide on Multilateral Methods in the Harmonised Index of Consumer Prices. Luxembourg: Publications Office of the European Union, 2022, ISBN 978-92-76-44354-4.

[10] GRAHAM, W. : IndexNumR: A Package for Index Number Calculation. R package version 0.5.0, 2022.

[11] International Labour Office: Consumer price index manual: Theory and practice. Geneva, 2004.

[12] IVANCIC, L. – DIEWERT, W. E. – FOX, K. J.: Scanner data, time aggregation and the construction of price indices. Journal of Econometrics 161(1), 2011, p. 24 – 35.

[13] KRSINICH, F.: The FEWS index: Fixed effects with a window splice–non-revisable quality-adjusted price indices with no characteristic information. In meeting of the group of experts on consumer price indices, 2014, p. 26 – 28.

[14] MEHRHOFF, J.: Towards a new paradigm for scanner data price indices: applying big data techniques to big data. In Paper presented at the 16th Meeting of the Ottawa Group on Price Indices, Rio de Janeiro, Brazil, 2019.
[15] SAAVEDRA-NIEVES, A. – SAAVEDRA-NIEVES, P.: IndexNumber: Index Numbers in Social Sciences. R package version 1.3.1, 2021.
[16] STANSFIELD, M.: multilateral: Generalised Function to Calculate a Variety of Multilateral Price Index Methods. R package version 1.0.0, 2022.
[17] QUENOUILLE, M.H.: Notes on bias in estimation. Biometrika, 43 (3–4), 1956, p.353 – 360.
[18] VAN DER LAAN, J.: reclin: Record Linkage Toolkit. R package version 0.1.1, 2018.
[19] VAN LOON, K. V. – ROELS, D.: Integrating big data in the Belgian CPI. In Paper presented at the meeting of the group of experts on consumer price indices, Geneva, Switzerland, 2018.
[20] VON AUER, L.: The nature of chain drift. In Paper presented at the 17th Meeting of the Ottawa Group on Price Indices, Rio de Janerio, Brasil, 2019.
[21] WEBSTER, M. - TARNOW-MORDI, R. C.: Decomposing multilateral price indices into the contribution of individual commodities. Journal of Official Statistics (2), 2019, p. 461–486.
[22] WHITE, G.: IndexNumR: Index Number Calculation. R package version 0.5.0, 2022.
[23] ZHANG, L. – JOHANSEN, I. – NYAGAARD, R.: Tests for price indices in a dynamic item universe. Journal of Official Statistics 35(3), 2019, p. 683 – 697.
[24] WICKHAM, H.: stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4, 2019.

RESUME
The article presents the basic functionality of the PriceIndices package, which was developed for both academics and practitioners involved in implementing scanner data to measure inflation. Specifically, the paper demonstrates how to prepare scanner data sets, classify products into the appropriate COICOP groups, match products over time and filter out the irrelevant sales. Of course the demonstration of the package also covers the determination of price indices, their comparison and the aggregation of the partial results over the retailer outlets and the product subgroups. The whole presentation is supported by reproducible codes written in R language, which can be found in the Appendix.

RESUMÉ
Článok predstavuje základnú funkcionalitu balíka Cenových indexov, ktorý bol vyvinutý pre akademikov aj odborníkov z praxe, ktorí sa podieľajú na implementácii údajov zo skenerov slúžiacich na meranie inflácie. Dokument konkrétne ukazuje, ako vytvoriť súbory údajov zo skenerov, klasifikovať produkty do vhodných skupín klasifikácie COICOP, priraďovať produkty v priebehu rokov a vytriediť irelevantné predaje. Ukážka balíka zahŕňa aj zisťovanie cenových indexov, ich porovnávanie a agregáciu čiastkových výsledkov maloobchodných predajní a podskupiny produktov. Súčasťou prezentácie sú reprodukovateľné kódy vytvorené v jazyku R, ktoré nájdete v prílohe.

CURRICULUM VITAE
Dr hab. Jacek Białek *has been an employee of Statistics Poland since 2018 where he is an expert in the Department of Trade and Services. Working in this position, he is involved in the analysis of scanner data and their application in the measurement of inflation. At the same*

time, Jacek Białek is an associate professor at the University of Lodz in Poland, where he works in the Department of Statistical Methods. Jacek Bialek's research interests revolve around the theory and practice of price indices. He is the author of more than 90 papers on price index theory, financial mathematics and open pension funds.

CONTACT
J.Bialek@stat.gov.pl,
jacek.bialek@uni.lodz.pl

APPENDIX

Example 1
```
> library('PriceIndices')
> data_unit(dataU, units = c("g", "ml", "kg", "l"), multiplication = "x")
```

Example 2
```
> milkUHT<-data_selecting(milk, must="uht", exclude="low")
> head(milkUHT)
```

Example 3
```
#Building the model
> my.grid=list(eta=c(0.01,0.02,0.05),subsample=c(0.5,0.8))
> data_train<-dplyr::filter(dataCOICOP,dataCOICOP$time<=as.Date("2021-10-01"))
> data_test<-dplyr::filter(dataCOICOP,dataCOICOP$time==as.Date("2021-11-01"))
> ML<-model_classification(data_train,data_test,coicop="coicop6",grid=my.grid,
> indicators=c("description","codeIN","grammage"),
> key_words=c("uht","2%"),rounds=60)
#Accuracy (while training)
> ML$figure_training
#Importance of the indicators:
> ML$figure_importance
#Data classification
> data_classifying(ML, data_test)  #please compare the last two columns!
```

Example 4
```
> df1<-data_matching(dataMATCH, start="2018-12",end="2019-02",
> onlydescription=TRUE, interval=TRUE)
> df2<-data_matching(dataMATCH, start="2018-12",end="2019-02",
> precision=0.98, interval=TRUE)
> length(unique(df1$prodID))
> length(unique(df2$prodID))
```

Example 5
```
> data_filtering(coffee, start="2018-12",end="2019-03",
> filters=c("extremeprices","lowsales","dumpprices"),
> plimits=c(0.25,2), lambda=1.25, dplimits=c(0.7,0.7))
```

Example 6
```
> data<-data_unit(dataU,units=c("g","ml","kg","l"), multiplication="x")
# Normalization of grammage units
> data_norm(data, rules=list(c("ml","l",1000),c("g","kg",1000)))
```

Example 7
```
> list<-products(coffee, "2018-12","2019-12")
> list$details
> list$statistics
> list$figure
```

Example 8
```
> df_indices<-price_indices(milk, start = "2018-12", end = "2019-12",
> formula=c("jevons", "fisher", "chjevons","chfisher", "tpd"),
> names=c("Jevons", "Fisher", "Chain Jevons", "Chain Fisher",
> "Time Product Dummy"), window=c(13), interval = TRUE)
> compare_indices_df(df_indices)
```

Example 9
```
> df_geks<-price_indices(coffee, start = "2018-12", end = "2020-06",
> formula=c("geks_splice", "geks_fbmw", "geks_fbew"),
> splice=c("half"), names=c("half splice", "FBMW","FBEW"),
> window=c(13), interval = TRUE)
> compare_indices_df(df_geks)
```

Example 10
```
> final_index(milk, start = "2019-01", end = "2019-06",
> formula = "fisher", groups = TRUE, outlets = TRUE,
> aggr = "laspeyres", by = "description", interval = TRUE)
```

Example 11
```
#Creating a data frame with index values
> DF<-price_indices(coffee,
> formula=c("jevons", "chjevons", "fisher", "geks"),
> start="2019-01", end="2019-12",
> window=c(13), interval=TRUE)
#Calculating average distances between indices (in p.p)
> compare_distances(DF)
```

Example 12
```
#This is a time-consuming procedure:
> comparison<-compare_indices_jk(coffee,
> formula=c("jevons", "chjevons", "fisher", "geks"),
> start="2019-01", end="2019-12", window=c(13),
> names=c("Jevons", "Chain Jevons", "Fisher", "GEKS"),
> title_iterations="Box-plots for iteration values")
> comparison$figure_iterations
```

Peter KNÍŽAT
Statistical Office of the Slovak Republic, University of Economics in Bratislava

HEDONIC CONSUMER PRICE INDEX: DIAGNOSTICS AND ANALYSIS OF VARIANCE

HEDONICKÝ INDEX SPOTREBITEĽSKÝCH CIEN: DIAGNOSTIKA A ANALÝZA ROZPTYLU

ABSTRACT
Web scraping provides a new innovative data source that can be utilised in price statistics by National Statistical Institutes. The prices of product-offers are automatically downloaded from the internet, which allows a wider selection of representative products in the consumer basket. The other advantage of scraping online prices is that the user can obtain characteristic parameters of individual products. These parameters are used in calculating a hedonic price index that is more preferable for products with a high replacement rate. The hedonic regression assumes that the natural logarithm of the product's price can be explained by its characteristic parameters. In many previous researches, the authors estimate hedonic price indices for various products without any statistical verification of the fitted regression model. In this paper, we carry out a thorough analysis of the hedonic regression model that encompasses checking the model's diagnostics and its initial assumptions. Moreover, we use the analysis of variance to test contrasts between categories of individual characteristic parameters. The categories without any contrast are merged together and the hedonic price index is re-estimated. In the empirical study, we estimate and compare the hedonic price indices for different scenarios using observed web scraped prices from the Slovak market.

ABSTRAKT
Web scrapovanie poskytuje nový inovatívny zdroj údajov, ktorý môžu národné štatistické úrady využiť v cenovej štatistike. Ceny produktov sa automaticky sťahujú z internetu, čo umožňuje širší výber reprezentatívnych produktov v spotrebnom koši. Ďalšou výhodou scrapovania online cien je, že používateľ môže získať charakteristické parametre jednotlivých produktov. Tieto parametre sa využívajú pri výpočte hedonického cenového indexu, ktorý je vhodnejší pre produkty s vysokou mierou fluktuácie. Hedonická regresia predpokladá, že prirodzený logaritmus ceny produktu možno vysvetliť jeho charakteristickými parametrami. V mnohých predchádzajúcich výskumoch autori odhadnú hedonické cenové indexy rôznych produktov bez toho aby štatisticky vyhodnotili správnosť regresného modelu. V tomto článku vykonáme dôkladnú analýzu hedonického regresného modelu, ktorá zahŕňa kontrolu diagnostiky modelu a jeho počiatočných predpokladov. Okrem toho použijeme analýzu rozptylu na testovanie kontrastov medzi kategóriami pre jednotlivé charakteristické parametre. Kategórie bez kontrastu sa potom zlúčia a hedonický cenový index sa prepočíta. V empirickej štúdii vypočítame a porovnávame hedonické cenové indexy pre rôzne prípady, využitím web scrapovaných cien produktu zo slovenského trhu.

KEY WORDS
web scraping, online prices, consumer price index, hedonic regression, time-product dummy regression

KĽÚČOVÉ SLOVÁ
web scrapovanie, online ceny, index spotrebiteľských cien, hedonická regresia, regresia s časovo umelou premennou

1. INTRODUCTION

In the past decade, National Statistical Institutes (NSIs) consider various alternative data sources to supplement the traditional data collection of product prices. The online collection, also called web scraping, of prices is one of these alternatives. Web scraping can be automated and NSIs are able to gather all available product-offers, which can be used to select a representative sample of product items for the consumer basket. Noting that the traditional price collection only covers a limited number of product offers due to extensive requirement for human resources. The research paper [12] considers web scraping to modernise a data collection for Italian harmonised consumer price index, in the context of the European project, and evaluates its effectiveness. It states that a web scraping has a big potential as a new innovative data source for price statistics.

Through web scraping, we usually obtain daily prices of products that are collected from comparison website platforms, where multiple online retailers offer the same product. This leads to a collection of multiple daily prices for the same product, which need to be aggregated on a monthly level. In the paper [11], the author demonstrates various methods for aggregating daily prices. The empirical analysis shows that the geometric average(s) could be the most appropriate aggregation since it includes all prices of selected products and it reduces the effect of extreme price values.

Moreover, web scarping allows obtaining characteristic parameters of individual products that are assumed a significant determinant of the price of the product. A well- known method for estimating a price index, which considers characteristic parameters, is a hedonic regression. The advantage of the hedonic regression is that it can include all products, without any requirement of its price observed in every month. Additionally, new and disappearing products can also be included in the estimation of the hedonic price index that is particularly important when the replacement rate of products is substantially high. The guidelines [4] states that hedonic price indices are more appropriate when the scale of replacement rate is high.

A comprehensive guide on hedonic regression models and corresponding price indices is provided in [16]. Many other research papers, for example [5], [6] and [7] among others, are dedicated to study hedonic regression models in the context of the estimation of price indices. For example, the paper [5] concludes that a hedonic price index is more suitable, when the characteristic parameters for products are observable, than the time-product dummy price index. However, none of these shows an examination of the model's goodness of fit, statistical tests of the model's assumptions, or any detailed analysis of individual characteristic parameters.

The objective of this paper is to carry out a thorough analysis of the hedonic regression model in the context of the estimation of consumer price indices. The individual categories of characteristic parameters are evaluated by using analysis of variance, i.e., a contrast between a pairwise comparison of categories is tested by different hypothesis tests. The categories with contrast that is not statistically significant from zero, a test under the null hypothesis, are merged together and the

hedonic price index is re-estimated. In the empirical study, we estimate and compare hedonic price indices using observed web scraped prices from the Slovak market.

The paper is organised as follows. Section 2 shows a theoretical framework of the hedonic regression and its corresponding price index. It also outlines a methodology that can be used for the evaluation of the model goodness of fit and the analysis of variance for categorical characteristic parameters. Section 3 shows the empirical application of the discussed theory on the product category of refrigerators. Conclusion discusses the results, issues with hedonic regression models and a potential further research.

2. THEORETICAL FRAMEWORK

In this section, in the first part, we introduce a theoretical framework for estimating a hedonic price index that is based on defining a hedonic regression model. In the recently published guidelines by Eurostat [2], an estimation of the hedonic price index is briefly outlined but no further details of testing the model diagnostics, or a statistical significance of model parameters are provided.

In the second part, we discuss a theory of the analysis of variance (ANOVA) in the context of the hedonic regression model. Moreover, we show the diagnostic tests that are used for testing the model goodness of fit and its initial assumptions.

2.1 REGRESSION-TYPE MODELS FOR CONSUMER PRICE INDICES

We assume that the logarithm of the price of the product item i, i = 1, …, n, can be explained by its characteristic parameters, where a number of characteristic parameters z_{ik} with k = 1,…,K for each product item i is observable at the time period t, t = 0, …, T. The time window T is defined by the user and it usually covers at least one year, T = 12. The log-linear hedonic regression model can be defined as [2]:

$$ln\, p_i^t = \partial^0 + \sum_{t=1}^{T} \partial^t D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t \qquad (1)$$

where D_i^t is a dummy variable that is assigned a value of 1 if the observed price relates to the product item i at the time period t, and 0 otherwise. The parameters ∂^0, ∂^t, and β_k are estimated through the ordinary least squares method, which yields $\hat{\partial}^0$, $\hat{\partial}^{\mathrm{t}}$, and $\hat{\beta}_k$. The regression model in Eq. (1) assumes that the random errors ε_i^t are normally distributed with mean 0 and constant variance σ^2. Noting that in order for the regression model to be estimable, we need to omit one dummy variable due to the singularity issue of the covariate matrix. Thus, the time-dummy variable D_i^t for the base period t = 0 is omitted.

The exclusion of the base period from the estimation, and the logarithmic form of the regression model, leads to a plausible interpretation of the estimated $\hat{\partial}^{\mathrm{t}}$s in the context of consumer price indices. The estimated $\hat{\partial}^{\mathrm{t}}$s are the percentage changes in average prices between the time period 0 and t, where the characteristic parameters are controlled for.

Taking the exponential of the estimated $\hat{\partial}^{\mathrm{t}}$ leads to the following expression for the hedonic price index [5]:

$$I_{Hedonic}^{0,t} = exp\left(\hat{\partial}^{t}\right) = \frac{\prod_{i\epsilon S^{t}}(p_i^t)^{\frac{1}{N^t}}}{\prod_{i\epsilon S^{0}}\left(p_i^0\right)^{\frac{1}{N^0}}} exp\left[\sum_{k=1}^{K}\hat{\beta}_k(\bar{z}_k^0 - \bar{z}_k^t)\right] \tag{2}$$

where $\hat{\beta}_k\bar{z}_k^0$ and $\hat{\beta}_k\bar{z}_k^t$ are the mean of estimated characteristic parameters effects at the time period 0 and t, respectively, where $\bar{\mathrm{z}}_{\mathrm{k}}^{0} = \sum_{\mathrm{i}\epsilon\mathrm{S}^{0}} z_{ik}/N^0$ and $\bar{\mathrm{z}}_{\mathrm{k}}^{\mathrm{t}} = \sum_{\mathrm{i}\epsilon\mathrm{S}^{\mathrm{t}}} z_{ik}/N^t$ are the sample averages of characteristic parameters. A detailed derivation of Eq. (2) can be found in [5]. Eq. (2) represents a hedonic price index that can be interpreted as a geometric average of price changes, between the base period 0 and the current period t, weighted by the mean effect of characteristic parameters. For the matched sample of product items between the time period 0 and t, Eq. (2) simplifies to the bilateral Jevons index; refer to [10] or [11] for further definitions.

Moreover, the hedonic regression model in Eq. (1) can be used for the missing price imputation, that is, when a product falls out of offer in a particular month but it reappears again next month, the estimated model can be utilised for the missing price prediction.

If we assume that the characteristic parameters of individual product items are unobservable, a different version of the log-linear regression model can be used. The time-product dummy (TPD) regression model is defined as [5]:

$$ln\, p_i^t = \partial^0 + \sum_{t=1}^{T}\partial^t D_i^t + \sum_{i=1}^{N-1}\gamma_i D_i + \varepsilon_i^t \tag{3}$$

where D_i is a dummy variable that is assigned a value 1 if the observed price relates to the product item i, and 0 otherwise. Similarly as for the time dummy variable, a dummy variable for the arbitrary product item i is not included in order to estimate the regression model.

Similarly, the time-product dummy (TPD) price index can be expressed as [5]:

$$I_{TPD}^{0,t} = exp\left(\hat{\partial}^{t}\right) = \frac{\prod_{i\epsilon S^{t}}(p_i^t)^{\frac{1}{N^t}}}{\prod_{i\epsilon S^{0}}\left(p_i^0\right)^{\frac{1}{N^0}}} exp\left[\sum_{i=1}^{N}\left(\bar{\hat{\gamma}}_i^0 - \bar{\hat{\gamma}}_i^t\right)\right] \tag{4}$$

where $\bar{\hat{\gamma}}_i^0$ and $\bar{\hat{\gamma}}_i^t$ are the mean of estimated product items' effects at the time period 0 and t, respectively, where $\bar{\hat{\gamma}}_i^0 = \sum_{\mathrm{i}\epsilon\mathrm{S}^{0}}\hat{\gamma}_i/N^0$ and $\bar{\hat{\gamma}}_i^{\mathrm{t}} = \sum_{\mathrm{i}\epsilon\mathrm{S}^{\mathrm{t}}}\hat{\gamma}_i/N^t$ are the sample averages of the estimated parameters $\hat{\gamma}$. A detailed derivation of Eq. (4) can be found in [5]. Again, for the matched sample of product items between the time period 0 and t, Eq. (4) simplifies to the bilateral Jevons index; refer to [10] or [11] for further definitions.

2.2 MODEL DIAGNOSTICS AND ANALYSIS OF VARIANCE

Both the hedonic regression model in Eq. (1) and the time-product dummy regression model in Eq. (3) are assumed to be of the linear form. This linearity in the

functional must be verified since its misspecification leads to incorrectly defined random errors ε_i^t that further invalidates hypothesis tests for the fitted model(s). Moreover, it can also indicate that the hedonic regression model has missing or omitted characteristic parameters, which is a phenomenon that can cause erroneous hedonic price indices.

The estimated random errors, or residuals, $\hat{\varepsilon}_i^t$ for Eq. (1) can be expressed as:

$$\hat{\varepsilon}_i^t = \ln p_i^t - \ln \hat{p}_i^t = \ln p_i^t - \left(\hat{\partial}^0 + \sum_{t=1}^{T} \hat{\partial}^t D_i^t + \sum_{k=1}^{K} \hat{\beta}_k z_{ik} \right) \tag{5}$$

A similar expression can be shown for Eq. (3). To verify the aforementioned assumptions of the fitted model, i.e., a normal distribution of the estimated residuals and the assumption for a constant variance, we can use standard diagnostic tests that are estimable by any statistical software in the regression procedure. We use the SAS software, where the *proc glm* procedure displays, in a graphical format, the following diagnostic tests:

Table No. 1: Diagnostic tests

Diagnostic test (plot)	Correctly specified model	Mis-specified model
Predicted vs observed responses	The values are close to the diagonal line.	The values that are far from the diagonal line → the predicted response is far from the observed response, i.e., the residual is large.
Residuals and predicted residual	The values show no pattern → a set of independent and identically distributed random variables.	The values follow a curve → missing characteristic parameters. The values are 'fan shaped' → a non-constant variance. The values are not randomly scattered → a correlation between residuals (autocorrelation in responses).
Studentized residuals versus the leverage	The values are in the range ±2.	The values that exceed ±2 can be considered outliers.
Residuals vs quantiles	The values are close to the diagonal line.	The few values fall on a diagonal line → outliers. The left end of is below / above the diagonal line → long / short tail in the left. Similarly, for the right end.

Source: SAS, Author's construction

The manuscript [16] provides a comprehensive study of hedonic regression models when the functional form of Eq. (1) is misspecified, particularly, due to the missing characteristic parameters in the model.

Furthermore, we introduce a concept of the analysis of variance (ANOVA) that is a general method for studying contrasts, or differences, between treatments of the

categorical variable for the given response variable. In general, the ANOVA functional form can be expressed as the general linear model. To formulate the terminology, we have a number of observed prices, N_c, in each category (treatment) c = 1,…, C, for the given characteristic (of the categorical format) parameter, and the linear ANOVA model for the i^{th} logarithmic price in the c^{th} category can be defined as [1]:

$$ln\, p_{ic}^{t} = \mu + \alpha_c + \varepsilon_{ic}^{t} \tag{6}$$

where μ is the overall mean across all N observed prices and $N = \sum N_c$. The parameters α_c represent the specific effects which are departures from the overall mean specific to each category c. The errors ε_{ic}^{t} are the unexplained variation specific to the i^{th} observation within category c and are assumed to be normally distributed with zero mean and a constant variance σ^2. To be able to identify individual categories uniquely, the constraint $\sum_c \alpha_c = 0$ has to be satisfied.

It follows that Eq. (6) can be rewritten in the form of the general linear model, where the design matrix X_{ij} is defined as containing p scalar independent variables coded 0's or 1's for the response category memberships. It implies that Eq. (6) can be re-expressed by setting $\beta_0 = \mu$, $\beta_2 = \alpha_1$, and so on to $\beta_c = \alpha_c$. Hence, it follows that Eq. (6) has the equivalent formulation as:

$$ln\, p_{ic}^{t} = \sum_{j=0}^{c+1} X_{(ic)j}\beta_j + \varepsilon_{ic}^{t} \tag{7}$$

The design matrix $X_{(ic)j}$ in Eq. (7) is equivalent to Eq. (1) for each characteristic parameter z_{ik} that is of the categorical format. Eq. (7) includes all independent variables in the model, including the time dummy variables, which serve as control variables in the estimation. Noting that the parameter β_j, usually for the last category c, is not defined, also called a controlled category, to maintain a non-singularity of the design matrix. It follows that the null hypothesis test for testing the statistical significance for each category is defined as:

$$H_0: \{\beta_1 - \beta_c = 0\ and\ \beta_2 - \beta_c = 0\ and\ \ldots and\ \beta_{c-1} - \beta_c = 0\} \tag{8}$$

The F statistics is used for testing the hypothesis; a detailed derivation of the hypothesis test is provided in [1]. Noting that the linear difference of β_js to the overall mean of the characteristic parameter, instead of the controlled category, can also be tested.

Moreover, we are interested in comparing the contrast between each category, not only for the controlled category. The main reason is that the categories for which the response is shown statistically indifferentiable can be merged into one category. This could lead to a computational efficiency when dealing with a large number of characteristic parameters that have multiple categories.

The null hypothesis test for the pairwise comparison of contrasts between categories can be defined as follows:

$$H_0: \{\beta_i - \beta_k = 0\} \quad for\ i \neq k \tag{9}$$

The SAS procedure *proc glm*, through the statement *LSMeans*, carries out the pairwise comparison; details can be found in [14], with a theory of the hypothesis test discussed in [1]. Similarly, the contrast between a specific category versus multiple other categories[1] can also be tested, which, in our case, is needed when we require to merge more categories into another specific category.

Various other multiple comparison tests for the pairwise comparison were proposed in the context of ANOVA. These tests are relevant for different types of experimental design. In the case of the hedonic model, an experimental design is completely randomized, i.e., we assume that prices are assigned to the categories completely at random, with unequal sample sizes of product items within each category.

The Tukey method, first proposed in 1953 by Tukey in [17], which is considered more conservative among other multiple comparison tests, is the most suitable in our case. The Tukey's confidence interval for the difference between categories i and k at the α% significance level is defined as [1]:

$$\beta_i - \beta_k \pm \left(\frac{q_{v,n-v,\alpha}}{\sqrt{2}}\right)\sqrt{MSE\left(\frac{1}{r_i} + \frac{1}{r_k}\right)} \tag{10}$$

where $q_{v,n-v,\alpha}$ is a critical point at the α% significance level from the *Studentised range distribution* and r_i and r_k are sample sizes of category i and k, respectively; a detailed derivation is shown in [1]. If the confidence interval in Eq. (10) contains zero, the difference between categories β_i and β_k is not significant at the α% significance level. The original Tukey's method was developed for the equal sample sizes in categories. Hayter in [9] shows that the same form of interval in Eq. (10) can be used for unequal sample sizes in individual categories, and that the overall confidence level is then at least 100(1 − α)%.

3. RESULTS

In the empirical study, we use web scraped data to demonstrate an application of the theoretical framework outlined in the previous section. Daily web scraped prices of the product category refrigerators (ECOICOP 5-digit: 05.3.1.1, refer to [13]) are available from 01 December 2019 until 31 December 2020. The prices of product items, including its corresponding characteristics parameters, were scraped from the website https://www.heureka.sk/. The aggregation of daily prices is carried out as in [10], i.e., daily prices of product items are geometrically averaged on a monthly level.

Table No. 2 shows a list of the characteristic parameters that are used in the estimation of Eq. (1).

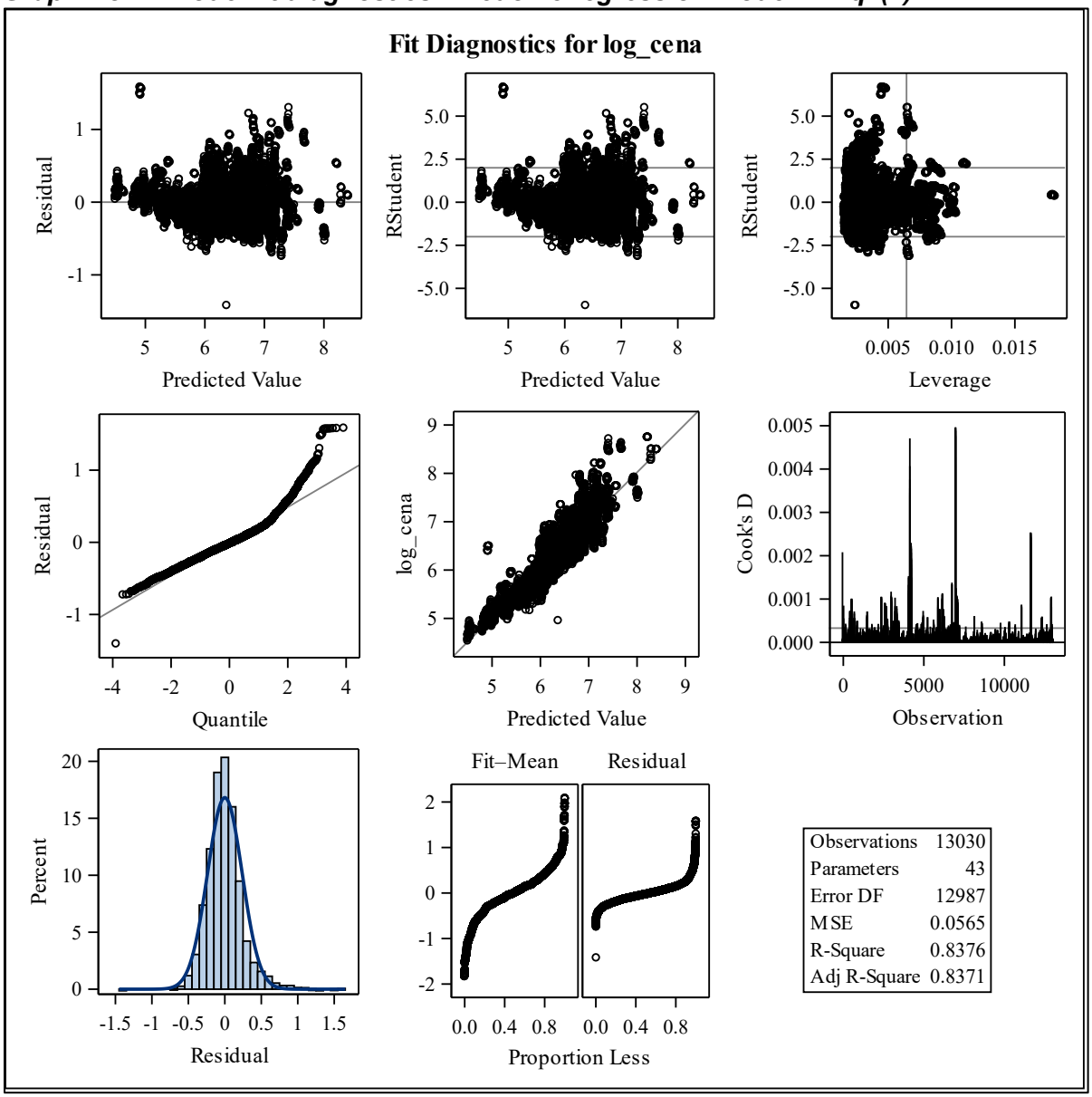
[1] It has to be correctly defined, which depends on the number categories tested, in the contrast statement in the SAS GLM procedure.

Table No. 2: The characteristic parameters

Parameter	Type (no. of categories)
Producer (výrobca)	Categorical (20)
Placement type (spôsob umiestnenia)	Categorical (2)
Noise (hlučnosť)	Numerical
Design (prevedenie)	Categorical (4)
Electricity consumption per year (spotreba energie za rok)	Numerical
Net volume of the fridge (čistý objem chladničky)	Numerical
Net volume of the freezer (čistý objem mrazničky)	Numerical
Height (výška)	Numerical
Width (šírka)	Numerical
Depth (hĺbka)	Numerical

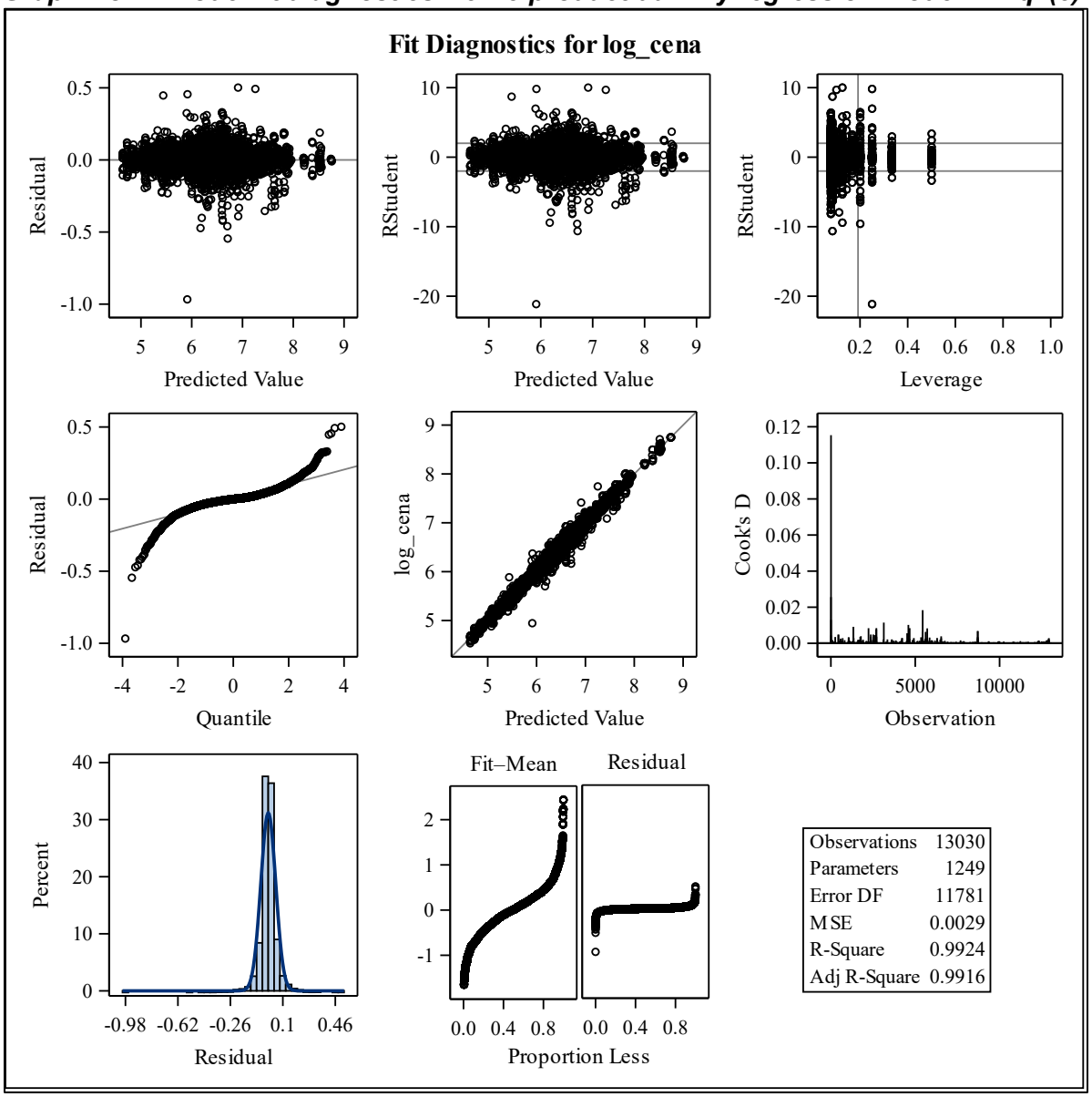
Source: SO SR, Author's construction

Graph No. 1: Model fit diagnostics – hedonic regression model in Eq. (1)



Source: SAS, Author's construction (Note: log_cena means log_price)

Graph No. 2: Model fit diagnostics – time-product dummy regression model in Eq. (3)



Source: SAS, Author's construction (Note: log_cena means log_price)

The estimation of Eqs. (1) and (3) is carried out in the SAS software, using the *proc GLM* procedure. The overall hypothesis test, using F statistics, show that both regression models are statistically significant at the 5% level. Additionally, all parameters, displayed in Table No. 2, in Eq. (1) are statistically significant at the 5% level. The full results are available from the author on request.

The graphs show the diagnostic tests, displayed in Table No. 1, for the fitted regression models in Eqs. (1) and (3), respectively.

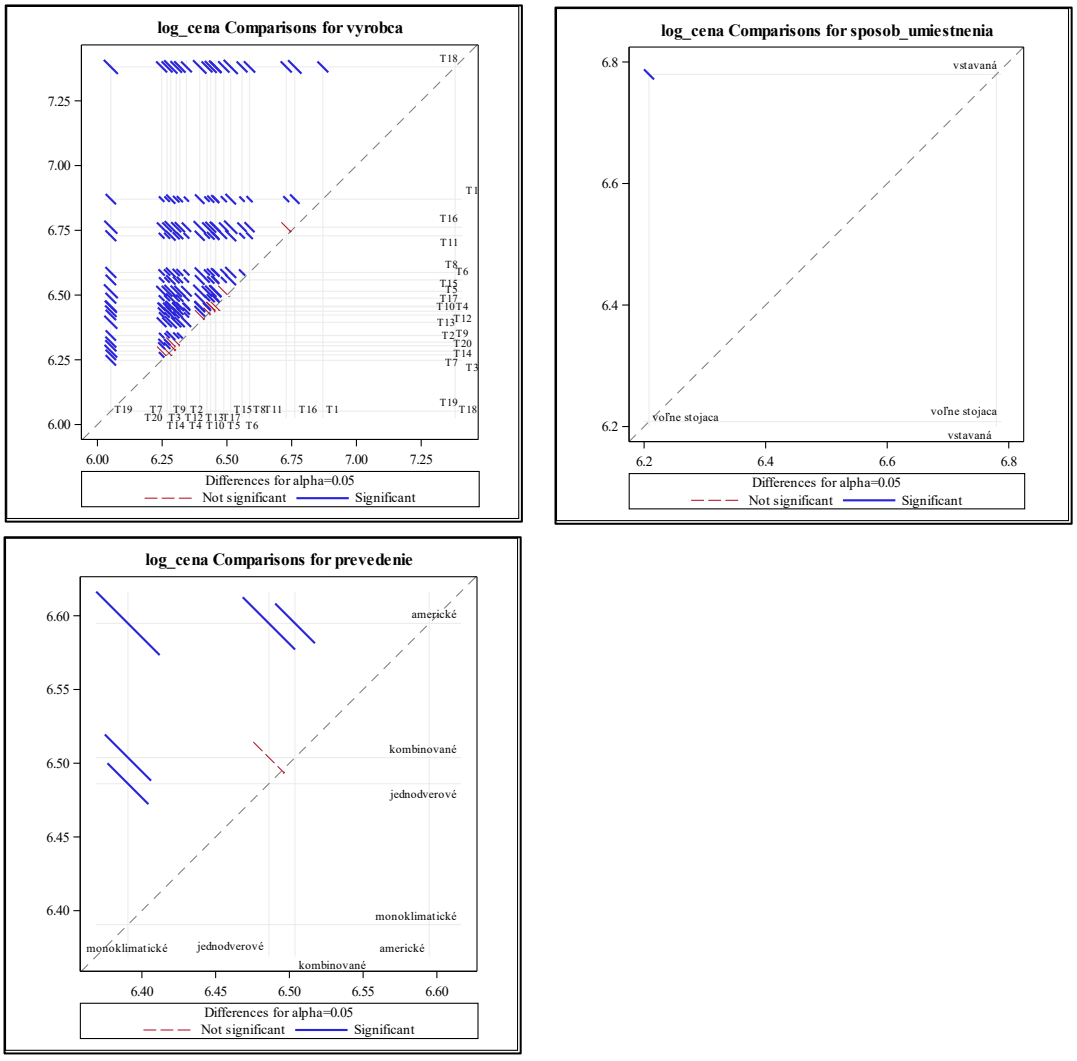
Graph No. 1 shows that the functional form of Eq. (1) is mis-specified. The plot of the residual vs predicted values shows a non-random distribution and a 'megaphone-like' shape that indicates a correlation in random errors. This leads to a conclusion that the random errors are not normally distributed and do not have a constant variance. Moreover, the graph of RStudent vs predicted values has many points beyond the limit(s) of ±2 that shows a presence of outliers in the observed prices that can be

inadvertently affecting the fitted model. From the second row of Graph n. 1, we can infer that the large prices cause a violation of the distributional properties of the random errors, which is seen in both the residual vs the quantile plot, the right end is far off the diagonal line, and the observed (log_cena) vs the predicted values plot.

Similar conclusions, for the presence of outliers in the observed prices, can be drawn for the fitted regression model in Eq. (3). However, the plot of residuals vs predicted values shows somewhat random distributions of values, which does not indicate a correlation among random errors, or a violation of the normal distribution assumption. In the third row, a distribution of residuals also resembles a normal curve, albeit with long tails caused by many outliers. Hence, we can infer that the fitted regression model is specified correctly, but the observed prices should be cleaned from outliers on both sides of the distribution.

We proceed to carry out a pairwise comparison of individual categories for each characteristic parameter. Noting that we have three categorical parameters, refer to Table No. 2. Graph No. 3 shows a significance level (at α = 5%) for the pairwise comparison.

Graph No. 3: Hypothesis test for the pairwise comparison of categories



Source: SAS, Author's construction (Note: log_cena means log_price)

The red line(s), in individual plots, indicate that the difference between categories[2] for the given characteristic parameter is not significantly different from zero, refer to Eq. (9). For categories that show non-significant differences, we create a single category. However, we consider the significance level for the difference at α = 10%. The following table shows the categories that are merged into one category.

Table No. 3: Merged categories – pairwise comparison

Control (main) category	Merged category
T12	T17
T12	T4
T14	T20
T14	T9
T15	T5

Source: Author's construction

Noting that for the merger of multiple categories, we carry out an additional contrast hypothesis test for in-between differences, for example, under the null hypothesis $\beta_{T12} = \beta_{T17} = \beta_{T4}$; refer to the SAS code in Annex B, particularly, the statement contrast.

In the follow-up analysis, we proceed to test the hypothesis of the pairwise comparison using the Tukey' multiple comparison method in Eq. (10). The results, which we display only for pairwise categories that are not significantly different from zero, are shown in Table No. 4:

Table No. 4: Confidence intervals – Tukey method

Comparisons significant at the 0.05 level are indicated by ***.			
vyrobca Comparison	Difference Between Means	Simultaneous 95% Confidence Limits	
T16 - T18	-0.085394	-0.177286	0.006497
T5 - T10	-0.006769	-0.048698	0.035159
T1 - T10	-0.022936	-0.062997	0.017125
T1 - T5	-0.016167	-0.054711	0.022377
T8 - T17	-0.056017	-0.113843	0.001808
T8 - T13	0.007288	-0.059478	0.074055
T13 - T17	-0.063306	-0.137796	0.011185
T14 - T15	-0.029089	-0.104795	0.046616
T3 - T15	-0.055178	-0.127308	0.016951
T3 - T14	-0.026089	-0.068946	0.016767
T3 - T4	0.025384	-0.010586	0.061355
T12 - T19	0.012626	-0.065001	0.090253

Source: SAS, Author's construction

Similarly based on the Tukey's method, we merged the categories with non-significant differences. Noting that the confidence interval that contains zero for the pairwise comparison indicates that the difference between these categories is not significant from zero. Table No. 5 shows the merged categories.

[2] *For the translation of individual categories refer to Annex A.*

Table No. 5: Merged categories – Tukey method

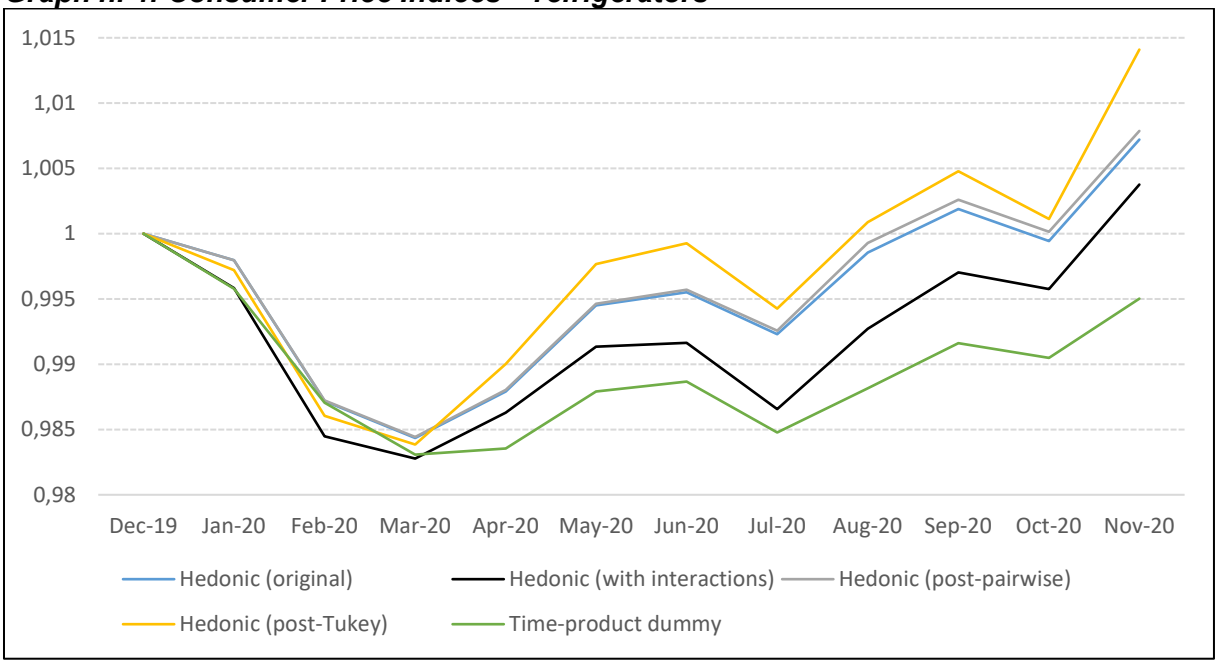
Control (main) category	Merged category
T16	T18
T1	T10
T1	T5
T8	T17
T8	T13
T3	T15
T3	T4
T3	T14
T12	T19

Source: Author's construction

Noting that after the merge of the categories in Table No. 5, the null hypothesis for the Tukey's test is rejected, at α = 5%, for all pairwise comparisons.

In the final stage of the analysis, we show the estimation of consumer price indices for all scenarios. The hedonic consumer price index is estimated for four different scenarios: original data, original data with interactions, post-pairwise and post-Tukey's merge of categories, respectively, along with the time-product dummy (TPD) consumer price index. The following graph shows the results.

Graph n. 4: Consumer Price Indices – refrigerators



Source: SOSR, Author's construction

Graph No 4 shows that all indices have the same trend. The TPD index is the lower bound in most of time periods. Adding the effect of interactions between characteristic parameters in Eq. (1) shifts the hedonic index down, towards the TPD index. This could indicate that missing parameters in the model might cause upward drift in the hedonic index. The hedonic index, after merging categories with non-significant differences, moves upwards. Noting that relatively big differences between individual price indices, for example, almost 2% between the lowest and largest index value, is solely due to

the specification of models, i.e., the number of parameters, explanatory variables, in the model.

All results have been generated in SAS software and are available from the author on request.

In general, based on the results, we can conclude that the number of characteristic parameters in the model plays an important role in the context of the estimated price indices. Therefore, it is important that all the relevant parameters of the particular product are collected and included in the estimation of the model. Moreover, the diagnostic tests show that extreme price values can cause a violation of the initial model's assumptions that can result in the erroneous price indices. The exclusion of extreme prices should be considered.

4. CONCLUSION

At the outset of the paper, we demonstrate a theoretical framework of hedonic and time-product dummy regression models. Additionally, a concept for diagnostic tests of the fitted regression model is outlined, including the analysis of variance that, in general, tests for the (average price) difference between categories of each characteristic (categorical) parameter. The diagnostic tests are crucial for testing the assumptions and the model's goodness of fit. An incorrect functional form of the regression model leads to errors in the estimated parameters, particularly, in hedonic consumer price indices. Moreover, the ANOVA might indicate that some categories of the characteristic parameters can be merged, which might lead to a minimisation of the computational time in practice.

It is important to consider online prices in the calculation of consumer price indices since the economic activity shows that consumers purchase some products online. In the empirical study, we use web scraping, which is considered a new data source, for collecting prices of products. The online daily prices are aggregated by using a geometric average on a monthly level.

Moreover, in the empirical analysis, using online prices for the product category of refrigerators from the Slovak market, we show that the functional form of the fitted hedonic regression model is misspecified. This is shown in the inspection of the diagnostic tests for the fitted model. The outliers play a role in the violation of the model assumptions, in particular, the assumptions of the normal distribution of random errors and its constant variance are not met. Hence, we are unable to confirm a suitability of the hedonic consumer price index for this case.

To improve the model's performance, we might consider excluding the extreme prices on both end of the distribution. On the other hand, these excluded product items, if economically important, would not be reflected in the estimation of the hedonic consumer price index. This shows that a disadvantage of web scraped data is that they do not contain any information on sales.

We show that characteristic parameters, and its corresponding categories, are statistically significant in determining the log-price of the product. Since some product categories, with a high replacement rate, and availability of big pool of product items require more complex formulas to capture the average price changes across time more

accurately, we recommend that hedonic price indices are considered by National Statistical Institutes for price statistics. Nevertheless, as demonstrated in the paper, NSIs must conduct a thorough analysis of hedonic regression models before its implementation in the production environment, either for estimating consumer price indices or for missing price imputations.

Further research, which can be part of the ongoing project of the modernisation of data sources for price statistics at the Statistical Office of the Slovak Republic, should concern the analysis of splicing methods that are used for avoiding the revision problem in consumer price indices when using multilateral methods.

ACKNOWLEDGEMENTS

This work was conducted as part of the European project *Dynamic Pricing Model* (311071AA56) that is supported by the European Union and co-financed through the European Regional Development Fund. The author would like to acknowledge a support of the grant by the Grant Agency of the Slovak Republic VEGA 1/0047/23 „The importance of spatial spillover effects in the context of the EU's greener and carbon-free Europe priority“. The author would also like to express an appreciation to the Heureka management for allowing the Statistical Office of the Slovak Republic to scrap daily product prices from their website platform.

BIBLIOGRAPHY

[1] DEAN, A., VOSS, D.: Design and Analysis of Experiments, Springer Texts in Statistics, 1999, Springer-Verlag New York Inc.
[2] EUROSTAT, Guide on Multilateral Methods in the Harmonised Index of Consumer Prices, Manuals and Guidelines, 2022, Luxembourg: Publication Office of the European Union. ISBN 978-92-76-44354-4.
[3] EUROPEAN COMISSION, EUROSTAT: Practical guidelines on web scraping for the HICP, Harmonised Indices of Consumer Prices, Directorate C: Macro-economic statistics, Unit C-4: Price statistics, Purchasing Power Parities, Housing statistics, November 2020.
[4] ILO/IMF/OECD/UNECE/EUROSTAT/THE WORLD BANK. Consumer Price Index Manual: Theory and Practice, 2004. ILO Publications, Geneva.
[5] DE HAAN, J., HENDRIKS, R.: Online data, fixed effects the construction of high-frequency price indexes. In: Paper presented at the Economic Measurement Group Workshop, 2013, 28-29 November 2013, Sydney, Australia.
[6] DE HAAN, J., KRSINICH, F.: Scanner Data and the Treatment of Quality Change in Nonrevisable Price Indexes. In: Journal of Business & Economic Statistics 32, 2014, pp. 341-358.
[7] DE HAAN, J.: Hedonic Prices Indexes: A Comparison of Imputation, Time Dummy and ‘Re-Pricing’ Methods. In: Jahrbücher f. Nationalökonomie u. Statistik, Lucius & Lucius, Stuttgart, 2010, Bd. (Vol.) 230/6, pp. 772-791.
[8] GLASER-OPITZOVÁ, H.: Nové zdroje údajov pre cenovú štatistiku a metódy ich spracovania. In: Slovenská štatistika a demografia, 2019, roč. 29, č.4, str. 49 – 66.
[9] HAYTER, A. J.: A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. In: The Annals of Statistics, 1984, vol. 12, no. 1, pp. 61-75.
[10] KNÍŽAT, P., GLASER-OPITZOVÁ, H.: Index spotrebiteľských cien z webscrapovaných údajov: analýza vybranej produktovej skupiny. In: Slovenská štatistika a demografia, 2023, roč. 33, č.1, pp. 37 – 49.

[11] KNÍŽAT, P.: Web scraped data in consumer price indices. In: Statistical Journal of the IAOS, 2023, vol. 39, no.1, pp. 203-212.
[12] POLIDORO, F., GIONNINI, R., LO CONTE, R., MOSCA, S. and ROSSETTI, F.: Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. In: Statistical Journal of the IAOS, 2015, vol. 31, no. 2, pp. 165-176.
[13] RAMON - Reference and Management of Nomenclatures: Europa - RAMON - Classification Detail List [cit. 2022-12-05].
[14] SAS INSTITUTE INC.: SAS/STAT® 13.1 User's Guide, The GLM Procedure, 2013, Cary, NC: SAS Institute Inc.
[15] Tovarová skupina chladničky: https://lednice.heureka.sk/ [cit. 2022-12-05].
[16] TRIPLETT, J. E.: Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes. Organization for Economic Co-operation and Development, 2006, Paris.
[17] TUKEY, J. W.: The problem of multiple comparisons. Dittoed manuscript of 396 pages, 1953, Department of Statistics, Princeton University.

RESUME
The Statistical Office of the Slovak Republic examines the use of new data sources for official price statistics. One of the alternative data sources for collecting prices of products is the online environment. Web scraping is an automated way of online price collection. The advantage of web scraping is that it does not require vast human resources that are needed in the traditional data collection from physical stores. Moreover, it also enables collecting characteristic parameters of individual products that can be used in the estimation of the hedonic regression model. The hedonic regression model can be used for the estimation of the hedonic price index, or for the imputation of missing prices. The advantage of the hedonic price index is that it captures the changes of all available products in the time window, including the new and disappearing products. In this paper, we show the estimation of the hedonic regression model and carry out various statistical tests to verify the model's specifications. In addition, we analyse the statistical significance of individual characteristic parameters, including a detailed analysis of variance for the categorical variables. Based on the analysis, we propose adjusting the categories for individual characteristic parameters that can reduce a calculation time when estimating the hedonic consumer price index for big data in practice. In the empirical study, we show the application, and compare hedonic consumer price indices for different scenarios, for the product category of refrigerators. In paper is part of the project that deals with the modernization of price statistics from the perspective of different data sources at the Statistical Office of the Slovak Republic.

RESUMÉ
Štatistický úrad SR skúma využitie nových zdrojov údajov pre oficiálnu cenovú štatistiku. Jedným z alternatívnych zdrojov dát na zber cien produktov je online prostredie. Web scraping je automatizovaný spôsob zberania cien z internetu. Výhodou web scrapingu je, že nevyžaduje veľa ľudských zdrojov, ktoré sú potrebné pri tradičnom zbere dát z obchodov. Okrem toho umožňuje aj zber charakteristických parametrov jednotlivých produktov, ktoré je možné použiť pri odhade hedonického regresného modelu. Model hedonickej regresie možno použiť na výpočet hedonického cenového indexu alebo na imputáciu chýbajúcich cien. Výhodou hedonického cenového indexu je, že zachytáva zmeny cien všetkých dostupných produktov v časovom okne, vrátane nových a tzv. odchádzajúcich produktov. V článku ukážeme

odhad hedonického regresného modelu a vykonáme rôzne štatistické testy na overenie správnosti modelu. Okrem toho analyzujeme štatistickú významnosť jednotlivých charakteristických parametrov, vrátane podrobnej analýzy rozptylu pre kategorické premenné. Na základe analýzy navrhneme upraviť kategórie jednotlivých charakteristických parametrov, čoho dôsledkom môže byť napr. skrátenie výpočtového času pri odhade hedonického indexu spotrebiteľských cien pre veľké dáta. V empirickej štúdii ukážeme aplikáciu výpočtu hedonických indexov spotrebiteľských cien a ich výsledky porovnáme pre rôzne prípady, pri výpočte použijeme ceny pre produktovú kategóriu chladničky. Príspevok je súčasťou projektu, ktorý sa zaoberá modernizáciou cenových štatistík z pohľadu využitia nových zdrojov údajov v Štatistickom úrade SR.

CURRICULUM VITAE

Peter Knížat MSc. *is an external PhD student at the Faculty of Economic Informatics, University of Economics in Bratislava. His main research interests are in spatial regressions and functional data; his publications can be found at: https://orcid.org/0000-0001-5100-1319. He teaches practical exercises for: Support of decision-making processes and Multiple attribute decision-making, including its application in the statistical software R. He works as a statistician at the Directorate of General Methodology, Registers and Coordination of the National Statistical System, Statistical Office of the Slovak Republic (SO SR), where he is responsible for proposing a statistical methodology for big data analysis. Prior to working at SO SR, he worked in an international bank as a senior risk and portfolio manager, where he lead the development of Basel internal risk-based models used in internal credit risk assessment processes and in the calculation of the bank's regulatory capital, and expected credit loss models used in the International Financial Regulatory Standards (IFRS) 9 reporting.*

CONTACT

peter.knizat@statistics.sk
peter.knizat@euba.sk

ANNEX A

Acronym	Producer (vyrobca)
T1	LIEBHERR
T2	GORENJE
T3	WHIRLPOOL
T4	ELECTROLUX
T5	SAMSUNG
T6	BOSCH
T7	BEKO
T8	AEG
T9	CANDY
T10	LG
T11	SIEMENS
T12	AMICA
T13	PHILCO
T14	HISENSE
T15	LORD
T16	HAIER
T17	CONCEPT
T18	MIELE
T19	INDESIT
T20	ZANUSSI

Acronym	Placement Type (sposob_umiestnenia)
vstavana	built-in
volne stojaca	free-standing

Acronym	Design (prevedenie)
americke	american
kombinovane	combined
jednodverove	one-door
monoklimaticke	monoclimatic

ANNEX B

```
/* Hedonic Regression index - original data */
ODS GRAPHICS ON;
TITLE; TITLE1 "Hedonic Model - ANOVA";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL)
on %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) at %TRIM(%SYSFUNC(TIME(),
TIMEAMPM12.))";

ods output
parameterestimates=FE_parameters_h;
proc glm data=unmatched_index_2
          PLOTS(maxpoints=none)=(DIAGNOSTICS RESIDUALS);
class vyrobca sposob_umiestnenia prevedenie cum_time (REF=FIRST);
model log_cena = cum_time vyrobca sposob_umiestnenia prevedenie
```

```
spotr_energie_rok objem_chladnicka vyska sirka hlbka objem_mraznicka hlucnost /
SS3 solution;
        LSMEANS vyrobca / PDIFF;
         LSMEANS sposob_umiestnenia / PDIFF;
         LSMEANS prevedenie / PDIFF;

         MEANS vyrobca / TUKEY CLDIFF ALPHA=0.05;
         MEANS sposob_umiestnenia / TUKEY CLDIFF ALPHA=0.05;
         MEANS prevedenie / TUKEY CLDIFF ALPHA=0.05;

         /* contrasts after pairwise */
        contrast "AMICA vs CONCEPT/ELECTROLUX"
                    vyrobca  0 0 0 -1 0 0 0 0 1,
                    vyrobca  0 0 0 -1 0 0 0 0 0.5 0 0 0 0 0 0.5;
         contrast "HISENSE vs ZANUSSI/CANDY"
                    vyrobca  0 0 0 0 0 -1 0 0 0 0 0 0 1,
                    vyrobca  0 0 0 0 0 -1 0 0 0 0 0 0 0.5 0 0 0 0 0 0 0.5;
run;

TITLE; FOOTNOTE;
ODS GRAPHICS OFF;

/* TPD Regression index - original data */
ODS GRAPHICS ON;
TITLE; TITLE1 "TPD Model";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (&_SASSERVERNAME, &SYSSCPL)
on %TRIM(%QSYSFUNC(DATE(), NLDATE20.)) at %TRIM(%SYSFUNC(TIME(),
TIMEAMPM12.))";

ods output
parameterestimates=FE_parameters_tpd;
proc glm data=unmatched_index_2
          PLOTS(maxpoints=none)=(DIAGNOSTICS RESIDUALS);
class  cum_time (REF=FIRST) id;
model log_cena = cum_time id / SS3 solution;
run;

TITLE; FOOTNOTE;
ODS GRAPHICS OFF;
```

Helena GLASER-OPITZOVÁ, Petra MAZUREKOVÁ
Štatistický úrad Slovenskej republiky

VPLYV PARCIÁLNEHO ČASOVÉHO POKRYTIA ÚDAJOV ZO SKENEROV NA PRESNOSŤ CENOVÝCH INDEXOV

EFFECT OF PARTIAL TIME COVERAGE OF SCANNER DATA ON THE ACCURANCY OF PRICE INDICES

ABSTRAKT
V procese modernizácie a skvalitňovania cenových štatistík sa Štatistický úrad Slovenskej republiky snaží o implementáciu nových zdrojov údajov. Jedným z možných nových zdrojov sú transakčné údaje obchodných reťazcov, nazývané aj údaje zo skenerov. Využívanie takýchto zdrojov však predstavuje významnú metodologickú zmenu. Jednou zo zásadných zmien je zmena cenového konceptu. Ceny získavané tradičným spôsobom ako tzv. pultové ceny, ktoré sa zisťujú v konkrétnom čase, sú nahradené priemernými cenami za jednotku tovaru. Na presnosť jednotkovej ceny má vplyv obdobie, ktoré berieme do úvahy pri jej stanovení. Cieľom tohto článku je zistiť, či a aký vplyv má navrhované parciálne časové pokrytie sledovaného obdobia (2 týždne vs. mesiac) na hodnoty mesačných cenových indexov.

ABSTRACT
In the process of modernizing and improving price statistics, the Statistical Office of the Slovak Republic is attempting to implement new data sources. One of the possible new sources is transaction data from retail chains, also known as scanner data. However, the use of such sources represents a significant methodological change. One of the fundamental changes is the shift in the price concept. Prices obtained in the traditional way, known as shelf prices, which are determined at a specific time, are replaced by average prices per unit of goods. The accuracy of the unit price is influenced by the period taken into account when determining it. The aim of this article is to determine whether and what impact the proposed partial time coverage of the observed period (2 weeks vs a month) has on the values of monthly price indices.

KĽÚČOVÉ SLOVÁ
cenové indexy, transakčné údaje, bilaterálne indexy, multilaterálne indexy

KEY WORDS
price indices, transaction data, bilateral indices, multilateral indices

1. ÚVOD
Používané zdroje údajov vo všeobecnosti významným spôsobom ovplyvňujú kvalitu štatistického produktu a efektivitu štatistického procesu.

V oblasti cenovej štatistiky sú dátové zdroje o cenách tovarov a služieb v podmienkach Štatistického úradu SR stále získavané prostredníctvom terénneho zberu údajov. Zber údajov sa realizuje priamo v prevádzkach a obchodoch na celom území SR, kde obyvatelia obvykle nakupujú, a ceny, ktoré sa zisťujú, sú tzv. pultové ceny. Ceny sa zisťujú počas prvých 20 dní sledovaného mesiaca a pri tradičnom zbere ide o ceny zisťované v konkrétnom čase [22].

V súčasnosti sa vo svete tradičný zber cien v prevádzkach a obchodoch postupne nahrádza, najmä v oblasti cien potravín, nealkoholických nápojov a tovarov bežnej spotreby údajmi zo skenerov alebo údajmi získanými formou web scrapingu [20]. Meranie cien na základe údajov zo skenerov je už niekoľko rokov aktívnou oblasťou výskumu aj v Štatistickom úrade SR. Ich použitie okrem iných významných metodologických zmien prináša aj zmenu cenového konceptu. Ceny získavané tradičným spôsobom sú nahradené cenami za jednotku tovaru.

Podľa [1] jednotková cena presnejšie odráža ceny, ktoré platia spotrebitelia počas celého sledovaného obdobia, ako zistenie ceny v konkrétnom čase pri tradičnom zbere. Jednotkové ceny zohľadňujú zľavy a ich vplyv na množstvo predaného tovaru. Na presnosť jednotkovej ceny má vplyv obdobie, ktoré berieme do úvahy pri jej stanovení (napr. mesiac vs len 2 alebo 3 týždne referenčného mesiaca). [4] tvrdia, že jednotkové ceny používané na konštrukciu CPI[1] by mali byť vypočítané za rovnaké obdobie ako je obdobie, za ktoré sa zostavuje index (napr. mesiac), a nie za čiastkové obdobie. V praxi však štatistické úrady zodpovedné za zostavenie HICP[2]/CPI štandardne používajú čiastkové obdobie referenčného obdobia z dôvodu včasnosti[3] a časovej presnosti[4] poskytovania štatistických produktov. Vzhľadom na termíny poskytovania transakčných údajov zo strany obchodných reťazcov a publikačnú prax Štatistického úradu SR predpokladáme využívať na stanovenie jednotkovej ceny prvé dva kompletné týždne referenčného mesiaca. Vzhľadom na uvedené skutočnosti je nanajvýš žiaduce vplyv nedostatočného časového pokrytia kvantifikovať a vyhodnotiť.

2. VLASTNOSTI TRANSAKČNÝCH ÚDAJOV

Transakčné údaje nazývané aj údaje zo skenerov sú pre štatistické úrady relatívne novým zdrojom údajov, ich dostupnosť sa však v posledných rokoch zvyšuje. Sú to údaje, ktoré zaznamenávajú maloobchodníci pri nákupoch spotrebiteľov skenovaním čiarových kódov. Tieto údaje obsahujú pre každý predaný tovar v obchode v danom období predané množstvo a predajnú/realizačnú cenu na úrovni kódu položky.

Transakčné údaje za konkrétneho maloobchodníka a časové obdobie predstavujú tak vyčerpávajúci zoznam všetkých kódov položiek, ktoré boli predané, ich tržby a predané množstvá. Umožňujú zostaviť index zo všetkých transakcií maloobchodníka alebo obchodu, pričom do výpočtu vstupujú realizačné ceny produktov a umožňujú zahrnúť do CPI/HICP oveľa viac položiek v porovnaní s tradičným zberom cien. Na porovnanie môžeme uviesť, že pri tradičnom zbere údajov v súvislosti so spotrebiteľskými cenami tovarov a služieb za oblasť potravín a nealkoholických nápojov je každá z vybraných predajní navštívená jedenkrát v mesiaci a za každého reprezentanta sa v predajni vyberie len jeden produkt (spolu cca 142 konkrétnych cien) [22] Oproti tomu pri údajoch zo skenerov získame informáciu za celý predaj daného

[1] CPI (Consumer Price Index) – Index spotrebiteľských cien meria celkovú zmenu spotrebiteľských cien na základe reprezentatívneho spotrebného koša tovarov a služieb v čase [22].

[2] HICP (Harmonised Index of Consumer Prices) – Harmonizovaný index spotrebiteľských cien bol vytvorený s cieľom poskytnúť vysokokvalitný, porovnateľný ukazovateľ inflácie spotrebiteľských cien [22].

[3] Včasnosť – vzťahuje sa na obdobie medzi dostupnosťou informácie a udalosťou alebo javom, ktorý opisuje.

[4] Časová presnosť – súvisí s časovým rozdielom medzi termínom zverejnenia údajov a cieľovým termínom, keď mali byť údaje dodané napríklad v súvislosti s termínmi zverejnenými v niektorom oficiálnom kalendári, ustanovenom predpismi alebo predchádzajúcou dohodou medzi partnermi (odchýlka od harmonogramu zverejnenia).

produktu v týždennom agregáte. To však znamená, že pri použití týchto údajov na účely zostavenia HICP/CPI musí dôjsť k zmene cenového konceptu. Do výpočtu cenových indexov tak nebude vstupovať cena zistená v konkrétnom čase, ale priemerná cena za jednotku tovaru za dané sledované obdobie určená takto:

priemerná cena = tržby / počet predaných kusov

Znamená to tiež, že ak sú k dispozícii informácie o tržbách, resp. predaných množstvách, môžeme každej položke priradiť váhu, čo otvára možnosti použiť na zostavenie cenových indexov na úrovni elementárneho agregátu[5] aj indexové vzorce pre superlatívne indexy, ktoré sú najpreferovanejšími indexmi na účely merania CPI [3]. Superlatívne indexy využívajú ceny a množstvá (t. j. výdavkové váhy) v oboch porovnávaných obdobiach (referenčné obdobie a bežné sledované obdobie).

V doterajšej praxi váhy výdavkov bežného obdobia, ale ani referenčného obdobia nie sú známe, takže v praxi sa štatistici pri zostavení CPI spoliehajú na elementárnej úrovni na nevážené indexy a pri agregácii na vyššiu úroveň klasifikácie ECOICOP[6] využívajú fixné váhy, ktoré sa vzťahujú na predchádzajúci rok.

3. ZDROJE ÚDAJOV A PRVOTNÉ SPRACOVANIE

V súčasnosti Štatistický úrad SR preberá transakčné údaje za oblasť potravín a nealkoholických nápojov od piatich obchodných reťazcov na týždennej báze. Dátové súbory obsahujú tržby a predané množstvá jednotlivých tovarov vo forme týždenného agregátu (od pondelka do nedele). Údaje sa na základe dohody medzi Štatistickým úradom SR a jednotlivými obchodnými reťazcami zasielajú s oneskorením spravidla 3 dni po ukončení referenčného obdobia.

Na vstupe do informačného systému prebieha štrukturálna validácia súboru a pri identifikovaní závažných chýb je poskytovateľ údajov kontaktovaný a požiadaný o poskytnutie opravného súboru.

Následne sa jednotlivé produkty zatrieďujú do klasifikácie ECOICOP, ktorá rozdeľuje kôš tovarov a služieb do odborov (2-miestna), skupín (3-miestna), tried (4-miestna) a podtried (5-miestna). Pre divíziu 01 – Potraviny a nealkoholické nápoje klasifikácie ECOICOP bola na účely spracovania údajov zo skenerov, definovaná národná, podrobnejšia 6-miestna úroveň klasifikácie - ECOICOP6, ktorá je spoločná pre údaje všetkých obchodných reťazcov, ktoré v súčasnosti spolupracujú so Štatistickým úradom SR. 6-miestna úroveň bola definovaná tak, aby vznikli homogénne skupiny produktov.

Každý produkt by mal vstupovať do výpočtu indexu na elementárnej úrovni na základe svojej dôležitosti. V usmernení Eurostatu [7] týkajúceho sa spracovania transakčných údajov zo supermarketov, sú odporúčané dva možné prístupy k výberu položiek, ktoré by mali byť zahrnuté do výpočtov, statický a dynamický prístup.

---

[5] Elementárny agregát je najmenší a relatívne homogénny súbor tovarov alebo služieb, pre ktorý sa určuje indexové číslo bez explicitných výdavkových váh. Môže byť definovaný nielen na základe charakteristických vlastností tovarov a služieb, ale aj s ohľadom na región a typ odbytiska, v ktorom sa nachádzajú a predávajú.

[6] ECOICOP je Európska klasifikácia individuálnej spotreby podľa účelu.

Statický prístup napodobňuje tradičný fixný výber (spotrebiteľský kôš) s tým rozdielom, že dochádza k zmene cenového konceptu. Ceny získané tradičným spôsobom sú nahradené cenami za jednotku tovaru z údajov zo skenerov a množina produktov je výrazne väčšia.

V prípade dynamického prístupu sa automaticky vyberajú produkty, ktoré sú v predaji súčasne v oboch po sebe nasledujúcich obdobiach/mesiacoch (t a t + 1, t + 1 a t + 2, t + 2 a t + 3 atď.) s tržbami nad určitou hranicou. Každý mesiac sa teda súbor jednotlivých výrobkov, ktoré vstupujú do zostavovania indexu, vyberá nanovo. Index elementárneho agregátu sa vypočíta na základe súboru spárovaných reprezentatívnych tovarových položiek, ktoré sa skutočne predávajú v dvoch nasledujúcich obdobiach.

Štatistický úrad SR sa po mnohých analýzach a sledovaní diskusií v rámci európskeho štatistického systému rozhodol ísť cestou dynamického prístupu, ktorý lepšie zohľadňuje aktuálnosť predávaných produktov a je jednoduchší z pohľadu automatizácie procesu spracovania.

Keďže cenové indexy sú publikované s mesačnou periodicitou, je potrebné týždenné súbory agregovať na mesačné. Tieto mesačné súbory obsahujú agregované hodnoty tržieb a predaného množstva jednotlivých tovarových položiek za príslušné vybrané týždne daného mesiaca a následne je vypočítaná priemerná cena produktu/tovarovej položky.

Na takto pripravené údaje sú aplikované filtre v závislosti od použitej metódy na zostavenie indexu, ktoré z výpočtu vylučujú niektoré produkty. Filtrovanie odstraňuje produkty s extrémnou zmenou ceny voči predchádzajúcemu obdobiu a ďalej sú odstránené dopredajové produkty (pokles v cene a súčasne výrazný pokles v tržbách). V prípade využitia bilaterálnych, resp. superlatívnych indexov sa aplikuje aj filter na málo predávané produkty. Viac o použití filtrov v súvislosti s údajmi zo skenerov sa uvádza v [11].

4. VYBRANÉ TYPY INDEXOV

Na výpočet cenových indexov na základe transakčných údajov je možné použiť viaceré indexové metódy (bilaterálne, multilaterálne, vážené, nevážené) a k nim prislúchajúce indexové vzorce. Pri výbere vhodného indexu, ktorý bude použitý na elementárnej úrovni v procese kompilácie CPI/HICP, možno použiť dva hlavné prístupy, axiomatický a ekonomický prístup.

Axiomatický prístup znamená, že index spĺňa určité špecifické axiómy alebo testy, na základe ktorých je možné vybrať vhodný index. Najpreferovanejšími testami sú test proporcionality, test súmernosti, časovo reverzný test a test tranzitivity. [15]. Z axiomatického hľadiska je Jevonsov index jednoznačne indexom s najlepšími vlastnosťami. Jeho nevýhodou je, že je to nevážený typ indexu, do výpočtu nevstupujú informácie o predaných množstvách resp. tržbách a teda pri použití Jevonsovho indexu plnohodnotne nevyužijeme výhody transakčných údajov.

Jevonsov index možno interpretovať ako geometrický priemer zmien cien takto [7]:

$$I_{Jevons}^{m-1,m} = \prod_{i \in N} \left(\frac{p_i^m}{p_i^{m-1}} \right)^{\frac{1}{N}} \quad (1)$$

kde *m* je aktuálne obdobie, *m-1* je predchádzajúce obdobie a p_i je cena *i*-tého produktu v danom období. Spotrebný kôš je dynamický a tvoria ho produkty, ktoré sú súčasne v aktuálnom a predchádzajúcom mesiaci v predaji a neboli vylúčené filtrami.

Okrem axiomatického prístupu na výber vhodného indexu je možné použiť aj prístup ekonomický, založený na ekonomickej teórii správania sa spotrebiteľa, čomu zodpovedá použitie vážených indexov, kde použitá váha je odrazom predaného množstva jednotlivých tovarových položiek.

Z indexov využívajúcich váhy sú najviac preferované superlatívne indexy Fischera a Törnqvista. Törnqvistov index je definovaný takto [15]:

$$I_{Tornqvist}^{m-1,m} = \prod_{i \in N} \left(\frac{p_i^m}{p_i^{m-1}} \right)^{\frac{s_i^{m-1} + s_i^m}{2}} \quad (2)$$

kde $s_i^{m-1} = p_i^{m-1} q_i^{m-1} / \sum_{i \in N} p_i^{m-1} q_i^{m-1}$ a $s_i^m = p_i^m q_i^m / \sum_{i \in N} p_i^m q_i^m$ sú podiely výdavkov v období *m-1* a *m* v danej skupine produktov, pričom q_i sú predané množstvá v danom mesiaci.

Tieto indexy porovnávajú zmenu ceny medzi dvoma po sebe idúcimi obdobiami, avšak pre publikačnú prax Štatistického úradu SR v oblasti cenovej štatistiky sú dôležité indexy k základnému obdobiu *0*, t. j. k decembru predchádzajúceho roku. Preto je potrebné medzimesačné indexy reťaziť.

$$I^{0,m} = I^{0,1} . I^{1,2} \dots I^{m-1,m} \quad (3)$$

Ak sa v praxi HICP/CPI zostavujú z transakčných údajov vo všeobecnosti sa odporúča na elementárnej úrovni používať reťazené superlatívne indexy z dôvodu vyššieho stupňa zhody jednotlivých kódov položiek medzi dvoma nasledujúcimi obdobiami a predpokladu menších rozdielov v cenách a množstve. Tento predpoklad však nezohľadňuje existenciu výpredajov a zliav, ktoré môžu výrazne zvýšiť množstvo predaného tovaru, a to až niekoľkonásobne. Po skončení obdobia zľavy sa cena vráti na pôvodnú hodnotu. Obyvateľstvo však môže byť na dostatočne dlhý čas zásobené a bude trvať určitú dobu, kým sa množstvo predaného tovaru vráti na pôvodnú hodnotu. Za týchto podmienok majú reťazené superlatívne indexy v porovnaní s neváženými indexmi tendenciu klesať a len postupne sa vracajú na pôvodnú úroveň. Tento jav sa v odbornej literatúre nazýva *chain drift*. Jedným zo spôsobov, ako sa v dôsledku predzásobenia tomuto javu vyhnúť, je na výpočet cenových indexov

na elementárnej úrovni použiť neväzené indexy, najmä medzimesačný Jevonsov a ten reťaziť, hoci aj reťazený Jevonsov index môže za určitých okolností driftovať.[7]

V [18] navrhli postup, ktorý poskytuje indexy superlatívneho typu bez driftu, prostredníctvom adaptácie teórie multilaterálnych indexov, ktoré merajú agregovanú zmenu cien medzi dvoma obdobiami na základe pozorovaných cien vo viacerých obdobiach vrátane dvoch porovnávaných období. Multilaterálne indexy boli vyvinuté na porovnanie cien medzi krajinami (parita kúpnej sily) a boli prispôsobené na porovnanie cien v čase. Boli navrhnuté najmä na použitie údajov zo skenerov.

Jedným z predstaviteľov tejto skupiny indexov je GEKS-Törnqvistov index. Multilaterálna metóda GEKS používa bilaterálne indexy (Törnqvistov), ktoré sú vypočítané medzi každou dvojicou časových období v definovanom časovom okne.

Pri aplikácii metódy je najprv potrebné určiť dĺžku intervalu (okna), na ktorý sa má metóda aplikovať. Metóda sa spravidla uplatňuje na dĺžku okna jedného alebo dvoch rokov plus jedno obdobie (v prípade mesačných údajov je to 13, resp. 25 mesiacov), aby sa zohľadnili aj produkty zaťažené sezónnosťou. Dĺžku okna budeme označovať *W*. V rámci okna sa vyberie základné obdobie označené ako *l* a vypočíta sa bilaterálny index medzi obdobím *l* a každým nasledujúcim obdobím v okne. Postup sa opakuje pre všetky možné alternatívy *l*. Týmto spôsobom je možné získať maticu bilaterálnych indexov veľkosti *w x w* pre všetky možné dvojice jednotlivých období v rámci časového okna. Multilaterálny GEKS-Törnqvistov index je definovaný takto [8]:

$$I^{0,m}_{GEKS-Törnqvist} = \prod_{l \in W} \left(I^{0,l}_{\text{Törnqvis}} \times I^{l,m}_{\text{Törnqvist}} \right)^{\frac{1}{|W|}} \tag{4}$$

Hore uvedené indexy sú aplikované na elementárnej úrovni, čo v našom prípade znamená úroveň ECOICOP6 pre každý obchodný reťazec. Na výpočet cenovej zmeny na vyššej úrovni agregácie klasifikácie ECOICOP sa používa index Laspeyresovho typu s fixnými váhami, ktoré sa vzťahujú na predchádzajúci rok ako referenčné obdobie (*y-1*) [12]:

$$P^{y,m/y-1}_{A} = \frac{\sum_{a \in A} w_a^{y-1} I_a^{y,m/y-1}}{\sum_{a \in A} w_a^{y-1}} \tag{5}$$

kde w_a^{y-1} sú váhy založené na ročných výdavkoch za všetky položky patriace do elementárneho agregátu a bez ohľadu na to, či boli zahrnuté do výpočtu po aplikovaní filtrov. Následne sú takto vypočítané indexy reťazené k základné obdobiu.

[7] *O reťazovom indexe sa hovorí, že driftuje, ak sa nevráti k jednotke, keď sa ceny v bežnom období vrátia na úroveň, ktorú dosiahli v základnom období. Reťazové indexy sú náchylné na drift, keď ceny počas období, ktoré pokrývajú, kolíšu.*

5. ANALÝZA VPLYVU ČASOVÉHO POKRYTIA

Ako už bolo uvedené, jednotkové ceny používané na konštrukciu CPI by mali byť vypočítané za rovnaké obdobie ako je obdobie, za ktoré sa zostavuje index (napr. v našom prípade mesiac), a nie za limitované parciálne obdobie. V praxi sa však štandardne používajú parciálne obdobia z dôvodu včasnosti a časovej presnosti poskytovania štatistických produktov. Vzhľadom na termíny poskytovania transakčných údajov zo strany obchodných reťazcov a publikačnú prax Štatistického úradu SR predpokladáme využívať na stanovenie jednotkovej ceny prvé dva kompletné týždne referenčného mesiaca.

Naším cieľom bolo formou referenčného porovnávania analyzovať vývoj časových radov cenových indexov a kvantifikovať vplyv parciálneho časového pokrytia na 6-miestnej, národnej úrovni klasifikácie ECOICOP, pretože cenové indexy na tejto úrovni sú základné stavebné prvky na zostavenie HICP/CPI. Porovnanie sme vykonali aj na 3-miestnej úrovni klasifikácie, pre skupinu tovarov Potraviny. Výstupy na tejto úrovni sú pravidelne publikované a vždy budú stredobodom záujmu pre svoj významný vplyv na životnú úroveň obyvateľstva.

Štatistický úrad SR preberá transakčné údaje od poskytovateľov na základe dohody, na týždennej báze vo forme týždenných a nie denných agregátov, čo má svoje konzekvencie. Pri zostavovaní mesačných súborov agregujeme tržby a predané množstvá za vybrané týždne. Pri zostavovaní mesačných súborov z údajov za celé obdobie mesiaca, bolo potrebné stanoviť určité pravidlá. Napríklad, v prípade prelomového týždňa, t. j. časť dní v danom týždni patrilo do jedného mesiaca a časť dní do druhého, bolo zavedené pravidlo, že daný týždeň bude zaradený do mesačného súboru k mesiacu, ku ktorému prislúchal väčší počet dní daného týždňa.

Nasledujúca tabuľka č. 1 porovnáva jednotkové ceny v eurách vypočítané z údajov za kompletné obdobie daného mesiaca a parciálne obdobie 2 týždňov na úrovni vybraných produktov homogénnej skupiny cukor kryštálový.

Modrou a žltou farbou je označená protichodná medzimesačná zmena ceny daného produktu a zelenou farbou takmer 19-percentný pokles priemernej mesačnej jednotkovej ceny produktu spôsobený parciálnym časovým pokrytím.

Tabuľka č. 1 : Porovnanie jednotkových cien tovarov – homogénna skupina cukor kryštálový

Kód produktu	Opis produktu	Mesiac	Jednotková cena v eurách	
			Pokrytie	
			Parciálne – 2T	Kompletné – 4T
4008671013004	Cukor kryštálový 1kg	01	0,7826	0,7864
4008671013004	Cukor kryštálový 1kg	02	0,7814	0,7789
4008671013004	Cukor kryštálový 1kg	03	0,6574	0,7133
4008671013004	Cukor kryštálový 1kg	04	0,7774	0,7793
4008671013004	Cukor kryštálový 1kg	05	0,7813	0,7813
8588000178391	Cukor kryštálový korunný 1kg	01	0,7057	0,7139
8588000178391	Cukor kryštálový korunný 1kg	02	0,7178	0,7065
8588000178391	Cukor kryštálový korunný 1kg	03	0,5643	0,6490
8588000178391	Cukor kryštálový korunný 1kg	04	0,7096	0,7059
8588000178391	Cukor kryštálový korunný 1kg	05	0,7139	0,7139
8588000178391	Cukor kryštálový korunný 1kg	06	0,6903	0,6943
8594003781131	Cukor korunný kryštálový 1kg	01	0,7347	0,7334
8594003781131	Cukor korunný kryštálový 1kg	02	0,7395	0,7365
8594003781131	Cukor korunný kryštálový 1kg	03	0,5762	0,7094
8594003781131	Cukor korunný kryštálový 1kg	04	0,7670	0,7644

Zdroj údajov: Štatistický úrad SR, výpočty autoriek

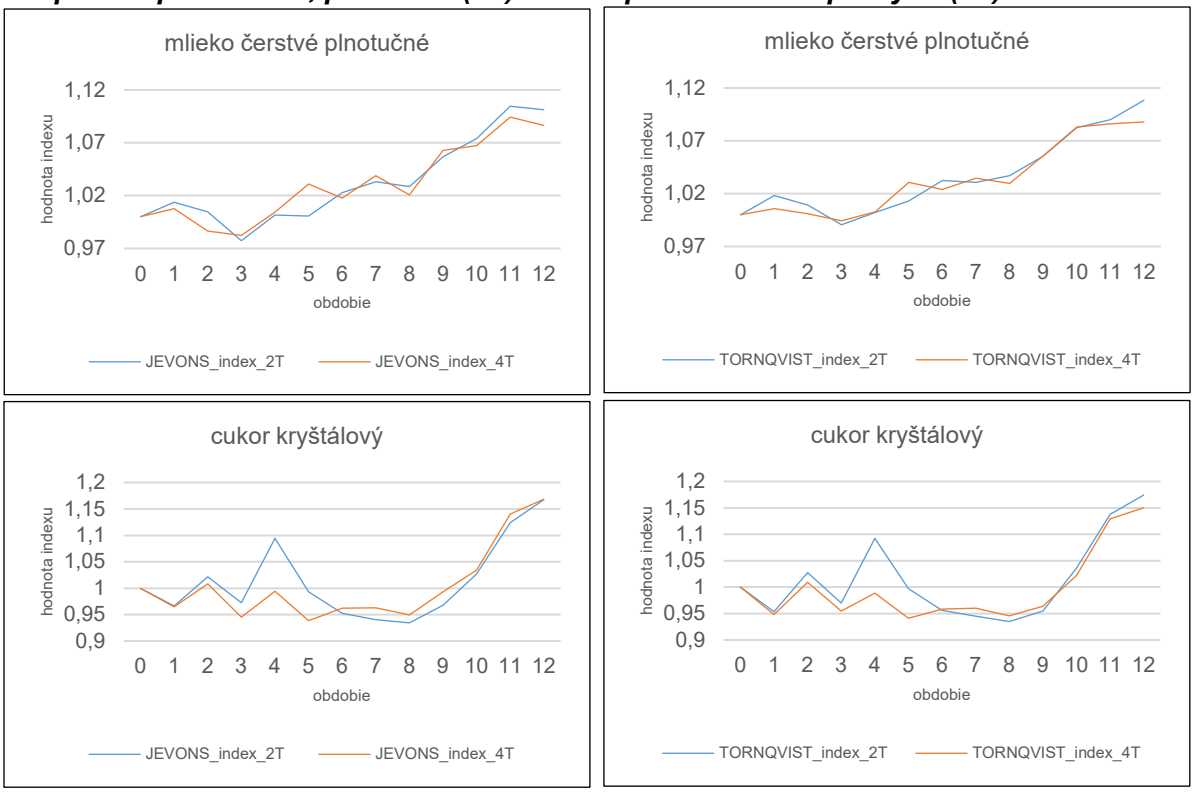
5.1 BILATERÁLNE INDEXY

Najprv sme vplyv časového pokrytia na vývoj cenových indexov analyzovali pre dynamický prístup k výberu produktov a následne sme zmenu ceny na elementárnej úrovni (ECOICOP6 a obchodný reťazec) vypočítali ako bilaterálny nevážený Jevonsov index (1) a vážený Törnqvistov index (2).

Na mesačné súbory údajov tržieb a predaného množstvá za všetky tovary predané v jednotlivých obchodných reťazcoch boli aplikované už spomenuté tri typy filtrov na očistenie údajov a to filter na extrémne zmeny ceny, na odstránenie dopredajových a málo predávaných produktov. Indexy vypočítané na elementárnej úrovni boli agregované spolu za všetky dostupné obchodné reťazce použitím vzorca pre cenový index Laspeyresovho typu (5) na úroveň ECOICOP6 alebo ECOICOP3 a následne reťazené k základnému obdobiu.

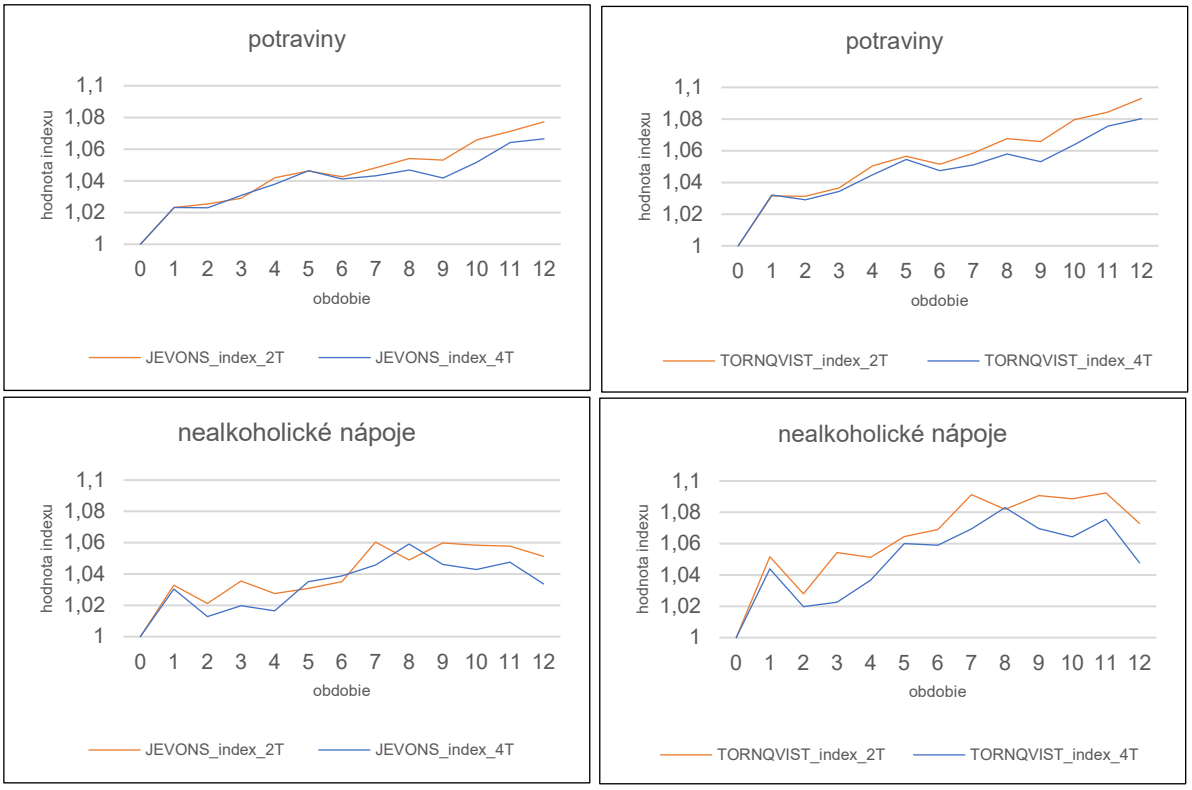
Vplyv parciálneho časového pokrytia na hodnoty a vývoj Jevonsovho a Törnqvistovho cenového indexu, ktoré sú vypočítané na úrovni vybraných homogénnych skupín produktov – mlieko čerstvé plnotučné a cukor kryštálový ilustrujú grafy č. 1 až 4. Modrou čiarou je zobrazený časový rad indexov, ktoré boli vypočítané z mesačných súborov údajov s parciálnym časovým pokrytím 2 týždňov, a oranžová čiara zobrazuje časový rad indexov, ktoré boli vypočítané nad kompletnou databázou údajov, ktorá pokrýva celé časové obdobie.

Grafy č. 1 – 4: Porovnanie vývoja mesačných cenových indexov na homogénnych skupinách produktov , parciálne (2T) vs kompletné časové pokrytie (4T)



Zdroj údajov: Štatistický úrad SR, výpočty autoriek

Grafy č. 5 – 8 : Porovnanie vývoja mesačných cenových indexov na skupine produktov – potraviny a nealkoholické nápoje, parciálne (2T) vs kompletné časové pokrytie (4T)



Zdroj údajov: Štatistický úrad SR, výpočty autoriek

Na úrovni homogénnych skupín produktov bol rozdiel hodnôt indexov úplného a parciálneho pokrytia v prípade niektorých produktov výrazný (cukor kryštálový) a v iných prípadoch minimálny. Hodnoty indexov na takejto nízkej úrovni klasifikácie nie sú štandardne publikované, preto sme sa zaoberali porovnávaním aj na vyššej 3-miestnej úrovni klasifikácie ECOICOP, a to konkrétne pre skupinu 01.1 - Potraviny a 01.2 Nealkoholické nápoje. Grafy č. 5 až 8 zobrazujú vplyv parciálneho časového pokrytia na hodnoty a vývoj Jevonsovho a Törnqvistovho cenového indexu.

Možno konštatovať, že parciálne časové pokrytie ovplyvňuje hodnoty jednotkových cien a v niektorých momentoch je tento vplyv aj protichodný, ale nespôsobuje systematické vychýlenie cenovej úrovne (tabuľka č. 1).

Jednotkové ceny na rozdiel od „pultových“ cien, ktoré sa využívajú pri tradičnom prístupe k výpočtu cenových indexov, zohľadňujú zľavy a ich vplyv na množstvo predaného tovaru, a preto časové pokrytie zohľadnené vo výpočte jednotkovej ceny logicky do určitej miery ovplyvňuje hodnoty cenových indexov. Veľkosť vplyvu závisí od skutočnosti, či zľava predstavovala významnú cenovú zmenu, či sa týkala viacerých produktov patriacich do spoločnej homogénnej skupiny a či boli produkty v zľave v jednom alebo vo viacerých obchodných reťazcoch v približne rovnakom čase. Nemenej dôležitá je aj skutočnosť, či bol tovar ponúkaný v zľave len v priebehu parciálneho časového obdobia alebo práve len mimo neho.

Chybu, akej by sme sa dopustili pri parciálnom časovom pokrytí, t. j. použitím len 2 úplných týždňov na stanovenie jednotkovej ceny tovarov do výpočtu cenových indexov k úplnému časovému pokrytiu, sme kvantifikovali prostredníctvom priemernej absolútnej percentuálnej chyby na dĺžke intervalu 12 mesiacov definovanej ako:

$$MAPE = \frac{1}{12}\sum_{i=1}^{12}\frac{|I_i^{4t} - I_i^{2t}|}{I_i^{4t}} * 100 \qquad (6)$$

kde I_i^{4t} je index vypočítaný v *i*-tý mesiac z kompletných údajov zvyčajne zostavených za obdobie štyroch týždňov a I_i^{2t} je index vypočítaný z údajov za obdobie dvoch kompletných týždňov *i-tého* mesiaca. Priemerná absolútna percentuálna chyba je miera relatívnej chyby, ktorá používa absolútne hodnoty, aby sa kladné a záporné chyby navzájom nerušili, a možno ju interpretovať ako priemerný percentuálny rozdiel medzi hodnotami z modelu, ktorý využíva parciálne časové pokrytie sledovaného obdobia a hodnotami z modelu, ktorý využíva úplné časové pokrytie a ktorého hodnoty považujeme za referenčné. Výsledky výpočtov sú uvedené v tabuľke č. 2.

Tabuľka č. 2: Porovnanie priemernej ročnej absolútnej percentuálnej chyby

Produktová skupina klasifikácie ECOICOP	Priemerná ročná absolútna percentuálna chyba (MAPE)	
	Jevonsov index	Törnqvistov index
Mlieko čerstvé plnotučné	0,9574%	0,7069%
Cukor kryštálový	2,4894%	2,3794%
Potraviny	0,5189%	0,6626%
Nealkoholické nápoje	1,0272%	1,4786%

Zdroj údajov: Štatistický úrad SR, výpočty autoriek

Na základe výsledkov výpočtov uvedených v tabuľke č. 2 môžeme predpokladať, že ak do štatistickej praxe zavedieme v súvislosti s implementáciou údajov zo skenerov výpočet bilaterálnych indexov, či už vážených, alebo neválených a parciálne časové pokrytie sledovaného obdobia, dopustíme sa na 3-miestnej úrovni klasifikácie, t. j. na úrovni Potraviny chyby menšej ako 1 % a v prípade Nealkoholických nápojov menej ako 1,5%.

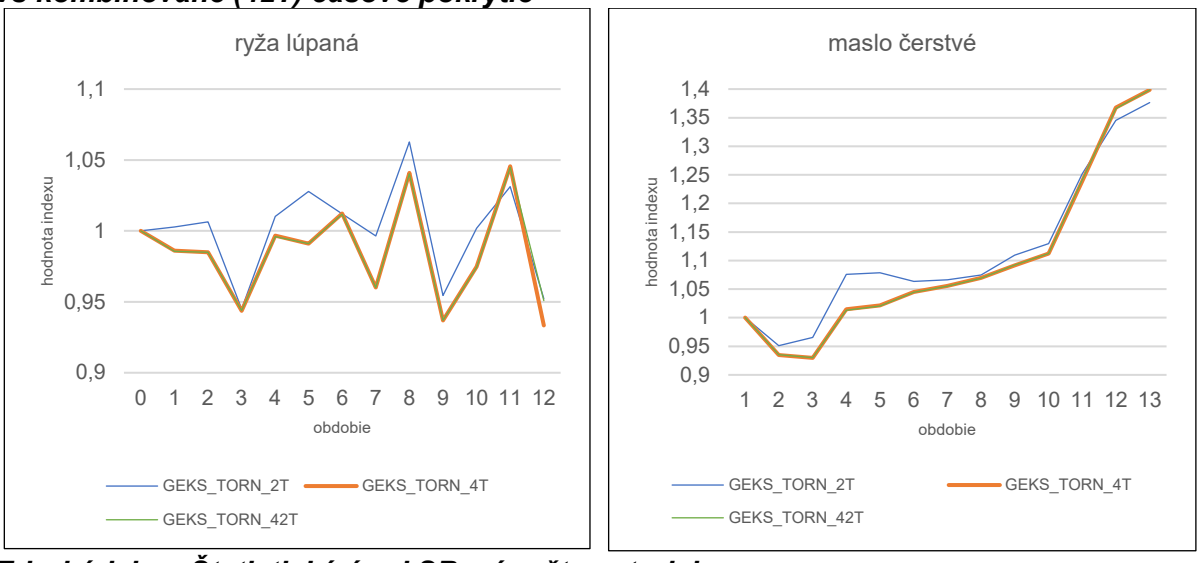
5.2 MULTILATERÁLNY PRÍSTUP

Ako už bolo uvedené, v prípade implementácie transakčných údajov obchodných reťazcov ako dátového zdroja pre výpočet cenových indexov, môžu byť cenové indexy na elementárnej úrovni vypočítané ako multilaterálne indexy, t. j. indexy, ktoré merajú agregovanú zmenu cien medzi dvoma obdobiami na základe pozorovaných cien vo viacerých obdobiach. V prípade multilaterálnych indexov sme vplyv parciálneho časového pokrytia analyzovali prostredníctvom GEKS-Törnqvistovho indexu (4) s dĺžkou časového okna 13 mesiacov.

K referenčným hodnotám, ktoré boli vypočítané nad mesačnými súbormi údajov s úplným časovým pokrytím sme porovnávali hodnoty indexov vypočítané nad parciálnym časovým pokrytím jednotlivých mesiacov, ktoré vstupovali do výpočtu, ako aj hodnoty indexov, ktoré boli vypočítané nad údajmi s kombinovaným pokrytím. V prípade kombinovaného pokrytia, ktoré by sa tiež mohlo použiť v praxi, aktuálny sledovaný mesiac bol pokrytý parciálne (2 týždne) a predchádzajúcich 12 mesiacov bolo pokrytých kompletne.

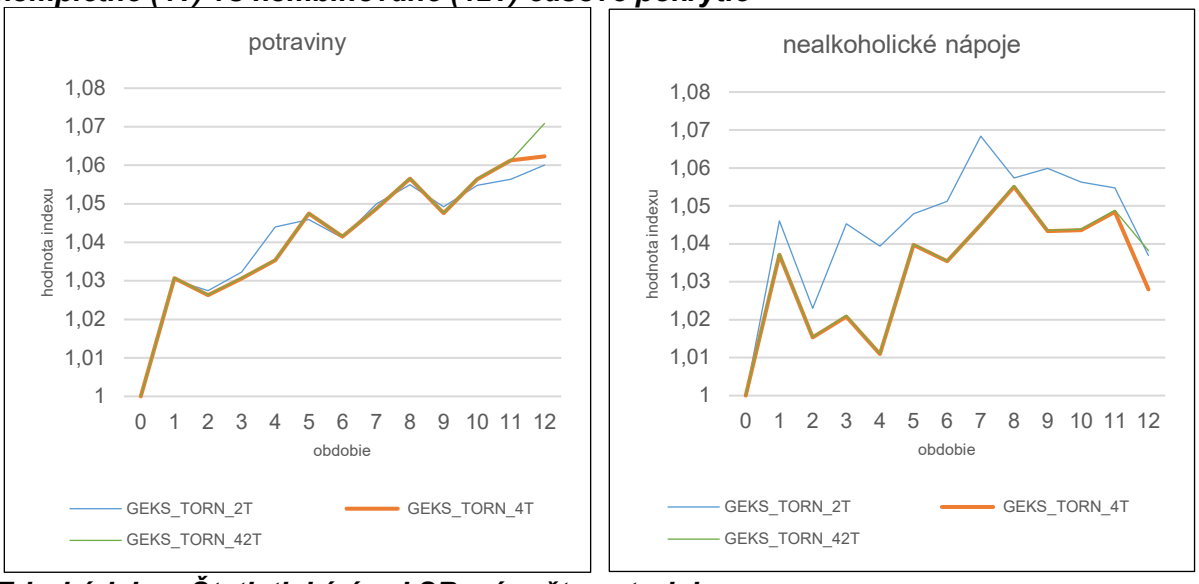
Grafické znázornenie vplyvu časového pokrytia na hodnoty cenových indexov pri použití multilaterálneho prístupu sú uvedené v nasledujúcich grafoch.

Grafy č. 9 a 10: Porovnanie vývoja mesačných cenových indexov (GEKS-Törnqvist) na skupine produktov – ryža lúpaná a maslo čerstvé, parciálne (2T) vs kompletné (4T) vs kombinované (42T) časové pokrytie



Zdroj údajov: Štatistický úrad SR, výpočty autoriek

Grafy č. 11 a 12: Porovnanie vývoja mesačných cenových indexov (GEKS-Törnqvist) na skupine produktov – 01.1. Potraviny a 01.2 Nealkoholické nápoje, parciálne (2T) vs kompletné (4T) vs kombinované (42T) časové pokrytie



Zdroj údajov: Štatistický úrad SR, výpočty autoriek

Podobne ako pri bilaterálnych indexoch grafy č. 9 až 12 ilustrujú vplyv parciálneho časového pokrytia na hodnoty a vývoj GEKS-Törnqvistových multilaterálnych cenových indexov, ktoré sú vypočítané na úrovni vybraných homogénnych skupín produktov – ryža lúpaná, maslo čerstvé a na 3-miestnej úrovni klasifikácie ECOICOP - 01.1-Potraviny a 01.2 Nealkoholické nápoje. Modrá čiara predstavuje hodnoty indexu vypočítaného nad mesačnými údajmi s parciálnym časovým pokrytím 2 týždňov, oranžová čiara predstavuje hodnoty indexov vypočítané nad databázou údajov s úplným časovým pokrytím a zelená čiara predstavuje multilaterálne indexy vypočítané nad kombinovaným časovým pokrytím.

Taktiež pri použití multilaterálneho prístupu sa vypočítala priemerná ročná absolútna percentuálna chyba (6). Hodnoty výpočtov sú uvedené v nasledujúcej tabuľke č. 3.

Tabuľka č. 3: Porovnanie priemernej ročnej absolútnej percentuálnej chyby

Produktová skupina klasifikácie ECOICOP	Priemerná ročná absolútna percentuálna chyba (MAPE)	
	Parciálne pokrytie	Kombinované pokrytie
Ryža lúpaná	1,9177%	0,1902%
Maslo čerstvé	2,3286%	0,0695%
Potraviny	0,2166%	0,0749%
Nealkoholické nápoje	1,3214%	0,1002%

Zdroj údajov: Štatistický úrad SR, výpočty autoriek

Z porovnania voči referenčným hodnotám indexov, ktoré boli vypočítané nad databázou údajov s plným časovým pokrytím, vyplynulo, že aj pri multilaterálnych indexoch sa na úrovni homogénnych produktových skupín pri parciálnom časovom pokrytí dopúšťame porovnateľnej chyby. Na úrovni produktovej skupiny Potraviny (3-miestna úroveň ECOICOP) sú však výsledky priaznivejšie a v prípade multilaterálnych

indexov je chyba spôsobená parciálnym pokrytím nižšia (0,2%) oproti chybe viac ako 0,5 % pri bilaterálnych indexoch.

Či už z grafov alebo tabuľky č. 3 je zrejmé, že ideálnym kandidátom na riešenie problému s parciálnym časovým pokrytím je kombinovaný prístup, ktorý by bol aplikovateľný v praxi. Problémom sú však samotné multilaterálne indexy, ktorých výpočet je náročný tak z pohľadu samotných postupov ako aj času. Ďalšou ich nevýhodou je, že princíp a metodiku ich zostavenia je veľmi náročné vysvetliť bežnej populácii používateľov.

6. ZÁVER

Implementácia transakčných údajov do výpočtu cenových indexov predstavuje významnú metodologickú zmenu jednak v spracovaní zdrojových údajov, ale aj v procese ich kompilácie. Eurostat ako náš partner v európskom štatistickom systéme poskytuje v tomto kontexte len odporúčania a je na národných štatistických úradoch, aby si zvolili čo najvhodnejší postup a metódu spracovania vzhľadom na podmienky v tej ktorej krajine. Pri výbere vhodných pracovných postupov treba zvažovať nielen zabezpečenie kvality štatistického produktu z hľadiska presnosti, ale aj včasnosť a časovú presnosť poskytovania štatistických produktov. V reálnej štatistickej praxi teda hľadáme kompromisné riešenie, ktoré zabezpečí dostatočnú kvalitu produktu a nenaruší publikačnú prax požadovanú používateľmi týchto produktov.

Cieľom článku bolo zistiť, či a aký vplyv na presnosť má kompromisné riešenie týkajúce sa parciálneho časového pokrytia sledovaného obdobia (v našom prípade mesiac vs 2 týždne), ktoré má zabezpečiť včasnosť a časovú presnosť poskytovania cenových indexov. Analýza bola vykonaná pre bilaterálne typy indexov, vážený (Törnqvistov) aj nevážený (Jevonsov) a tiež pre multilaterálne typy indexov, ktoré reprezentoval GEKS-Törnqvist.

Z výsledkov vyplynulo, že vo všetkých analyzovaných prípadoch je na úrovni produktovej skupiny Potraviny priemerný absolútny percentuálny rozdiel medzi hodnotami cenových indexov vypočítaných nad databázou údajov s parciálnym pokrytím oproti výsledkom nad databázou s plným pokrytím menší ako 1 %, čo je viac ako vyhovujúci výsledok. Okrem iného, nárast alebo pokles cien z druhej polovice mesiaca, ktoré nevstupovali do výpočtu cenového indexu, budú zohľadnené v nasledujúcom mesiaci.

Navyše tak ako vyplynulo z analýzy týkajúcej sa multilaterálnych indexov, najlepšie výsledky sme dosiahli kombinovaným prístupom k riešeniu parciálneho pokrytia. Jeho prípadná aplikácia v praxi si však vzhľadom na mnohé komplikácie súvisiace jednak s aplikáciou samotných indexov ako aj špeciálnou prípravou zdrojových údajov bude vyžadovať ďalší výskum v oblasti metodológie a automatizácie spracovania.

LITERATÚRA

[1] BALK, B. M.: On the Use of Unit Value Indices as Consumer Price Subindices. Paper presented at the Fourth Meeting of the International Working Group on Price Indices, Washington, DC, April 22–24. 1998. http://www.ottawagroup.org.

[2] DE HAAN, J. – KRSINICH, F.: Scanner Data and the Treatment of Quality Change in Rolling Year GEKS Price Indexes. Paper presented at the eleventh Economic Measurement Group Workshop, 21-23 November 2012, Sydney, Australia

[3] DIEWERT, W.E.: Exact and Superlative Index Numbers. In: Journal of Econometrics, 1976, 4, s. 114 – 145.
[4] DIEWERT, W. E. – FOX,K. J. – DE HAAN, J. : A Newly Identified Source of Potential CPI Bias: Weekly versus Monthly Unit Value Price Indexes, Economics Letters, 2016, 141, s. 169 – 172.
[5] ELTETÖ, O. – KÖVES, P.: On a Problem of Index Number Computation Relating to International Comparisons, Statisztikai Szemle 1964, 42, s. 507 – 518 (originál v maďarčine).
[6] EUROPEAN COMMISSION, EUROSTAT: Harmonised Index of Consumer Prices (HICP): Methodological Manual. Luxembourg: Publications Office of the European Union, 2018. 354 s. ISBN 978-92-79-76861-3.
[7] EUROPEAN COMMISSION, Eurostat (September 2017): HICP, Practical Guide for Processing Supermarket Scanner Data [cit. 2023-05-02] Dostupné na: https://circabc.europa.eu/ui/group/7b031f10-ac19-4da3-a36f-58708a70133d/library/8e1333df-ca16-40fc-bc6a-1ce1be37247c/details
[8] EUROSTAT: Guide on Multilateral Methods in the Harmonised Index of Consumer Prices, Manuals and Guidelines, Luxembourg: Publication Office of the European Union, 2022. ISBN 978-92-76-44354-4.
[9] GINI, C.: On the Circular Test of Index Numbers, 1931, Metron 9, s. 3 – 24.
[10] G LASER-OPITZOVÁ, H.: Nové zdroje údajov pre cenovú štatistiku a metódy ich spracovania. In: Slovenská štatistika a demografia, 2019, roč. 29, č. 4, s. 49 – 66. [cit. 2023-05-02] Dostupné na: https://ssad.statistics.sk/SSaD/wp-content/files/4_2019/4_2019_komplet_cislo.pdf
[11] GLASER-OPITZOVÁ, H.: Use of Scanner Data in Measuring the Consumer Price Index in the Conditions of the Slovak Republic, 2022. Proceedings from the EDAMBA 2021 conference, s. 92 – 102 [cit. 2023-05-02]. Dostupné na: https://doi.org/10.53465/EDAMBA.2021.9788022549301.92-102
[12] VAN DER GRIET,H. – DE HAAN, J.: The use of supermarket scanner data in the Dutch CPI. [cit. 2023-05-02] Dostupné na: 2010-scanner-data-dutch-cpi
[13] CHESSA, A.: A new methodology for processing scanner data in the Dutch CPI. In.: Eurostat review of National Accounts and Macroeconomic Indicators, 2016, č. 1 s. 49 – 69.
[14] CHESSA, A.: The QU-method: A new methodology for processing scanner data. Proceedings of Statistics Canada 2016 International Methodology Symposium.
[15] ILO/IMF/OECD/UNECE/Eurostat/The World Bank. Consumer Price Index Manual: Concepts and Methodts. IMF Publications, Washington, DC. 2020. ISBN 978-1-51354-298-0 (PDF).
[16] ILO/IMF/OECD/UNECE/Eurostat/The World Bank. Consumer Price Index Manual: Theory and Practice. ILO Publications, Geneva. 2004. ISBN 92-2-113699-X.
[17] IVANCIC L. – DIEWERT W. E. – FOX K. J.: Scanner Data, Time Aggregation and the Construction of Price Indexes, , In: Journal of Econometrics. 2011. vol. 161, no. 1, s. 24 – 35.
[18] IVANCIC, L. – DIEWERT, W. E., – FOX, K. J. *Scanner data, time aggregation and the construction of price indexes.* Discussion paper 09-09, University of British Columbia, Vancouver, Canada. 2009.
[19] KNÍŽAT, P. – GLASER-OPITZOVÁ, H.: Index spotrebiteľských cien z webscrapovaných údajov: analýza vybranej produktovej skupiny. In: Slovenská štatistika a demografia, 2023, roč. 33, č.1, s. 37 – 49. [cit. 2023-05-02]. Dostupné na: https://ssad.statistics.sk/SSaD/wp-content/files/1_2023/SSaD_1_2023_kompletne_cislo.pdf

[20] KNÍŽAT, P.: Web scraped data in consumer price indices. In: Statistical Journal of the IAOS, 2023, vol. 39, no.1, s. 203 – 2012.
[21] RAMON - Reference and Management of Nomenclatures: Europa - RAMON -Classification Detail List [cit. 2022-12-05].
[22] Spotrebiteľské ceny a ceny produkčných štatistík [cit. 2023-05-02].

RESUMÉ

Súčasná doba ponúka čoraz viac nových zdrojov údajov, ktoré môžu prispieť ku skvalitneniu a modernizácii štatistík. Jedným z takýchto nových zdrojov údajov sú transakčné údaje obchodných reťazcov, nazývané aj údaje zo skenerov. Tieto údaje môžu veľkou mierou prispieť k modernizácii štatistiky spotrebiteľských cien, ktorá patrí medzi významné makroekonomické ukazovatele danej krajiny. Údaje obsahujú úplné informácie o predaných tovaroch a to okrem kódu a stručného popisu hlavne predané množstvo a tržby za dané obdobie. Naproti tomu, ceny zbierané tradičným spôsobom sú ceny vybranej, relatívne malej množiny tovarov, za ktoré je tovar ponúkaný. Informácie o predaných množstvách nie sú dostupné. Implementácia transakčných údajov do produkcie cenovej štatistiky znamená implementáciu mnohých metodologických zmien. Jednou z nich je aj zmena cenového konceptu. Ceny získavané tradičným spôsobom sú nahradené cenami za jednotku tovaru. Jednotkové ceny používané na konštrukciu cenových indexov by mali byť teoreticky vypočítané za rovnaké obdobie ako je obdobie, za ktoré sa zostavuje index (napr. mesiac). Vzhľadom na termíny poskytovania transakčných údajov zo strany obchodných reťazcov, publikačnú prax Štatistického úradu SR a termíny poskytovania štatistických produktov ich používateľom je nevyhnutné nájsť kompromis medzi presnosťou cenových indexov a ich včasnosťou a časovou presnosťou. Takýto kompromis predstavuje parciálne časové pokrytie, t. j. výpočet indexov sa realizuje nad databázou údajov, ktoré nepokrývajú obdobie celého sledovaného mesiaca, ale len obdobie napríklad 2 týždňov. Cieľom príspevku je zistiť či má takéto kompromisné riešenie zásadný vplyv na úroveň a vývoj časových radov indexov a tento vplyv kvantifikovať. Analýza bola realizovaná pre bilaterálne, vážené aj nevážené indexy ako aj pre multilaterálne indexy. Z výsledkov analýzy vyplynulo, že vo všetkých analyzovaných prípadoch bol na úrovni produktovej skupiny Potraviny, t. j. na 3-miestnej úrovni klasifikácie ECOICOP, priemerný absolútny percentuálny rozdiel medzi hodnotami cenových indexov vypočítaných nad databázou údajov s parciálnym pokrytím oproti výsledkom nad databázou s plným pokrytím menší ako 1 %, čo môžeme považovať za viac ako vyhovujúci výsledok.

RESUME

The present time offers an increasing number of new data sources that can contribute to the improvement and modernization of statistics. One such new data source is transaction data from retail chains, also known as scanner data. These data can greatly contribute to the modernization of consumer price statistics, which are among the significant macroeconomic indicators of a country. These data contain comprehensive information about the sold goods, including not only the code and a brief description but also the quantity sold and the turnover for a given period. In contrast, prices collected in the traditional way are prices of selected, relatively small set of goods, for which the goods are offered, and information about the quantities sold is not available. The implementation of transaction data into price statistics production entails several methodological changes, including a change in the price concept. Prices obtained in the traditional way are replaced by prices per unit of goods. Unit prices used to

construct price indices should ideally be calculated for the corresponding period as the period for which the index is compiled (e.g., a month). Given the timing of providing transaction data by retail chains, the publication practices of the Statistical Office of the Slovak Republic, and the deadlines for providing statistical products to the users, it is necessary to find a compromise between the accuracy of price indices and their timeliness and temporal accuracy. Such a compromise is represented by a partial time coverage, which means that the calculation of indices is performed using a database that does not cover the entire observed month but only a period, i.e., the first two weeks of the given months. The aim of this contribution is to determine whether such a compromise solution has a significant impact on the level and trend of index time series and to quantify this impact. The analysis was conducted for bilateral, weighted and unweighted indices, as well as for multilateral indices. The results of the analysis showed that in all the analyzed cases, at the product group level "Food," i.e., at the 3-digit level of the ECOICOP classification the average absolute percentage difference between the values of price indices calculated using a database with partial coverage (2 weeks) compared to the results using a database with full coverage was less than 1%, which can be considered as a highly satisfactory outcome.

PROFESIJNÝ ŽIVOTOPIS

Ing. Helena Glaser-Opitzová *je generálna riaditeľka sekcie všeobecnej metodiky, registrov a koordinácie národného štatistického systému Štatistického úradu SR a členka riaditeľskej skupiny Eurostatu pre metodológiu (DIME), ktorá poskytuje poradenstvo Európskemu štatistickému výboru (ESSC) v strategických otázkach. Riadila a podieľala sa na mnohých modernizačných aktivitách úradu. V súčasnosti vedie interný projekt úradu zameraný na modernizáciu cenových štatistík.*

RNDr. Petra Mazureková PhD*. pracuje ako metodička v sekcii všeobecnej metodiky, registrov a koordinácie národného štatistického systému na odbore metód štatistických zisťovaní Štatistického úradu SR ako členka tímu zodpovedného za efektívne štatistické metódy, analýzy a výpočty s dôrazom na štandardizáciu štatistických procesov, systém monitorovania, vykazovania a hodnotenia kvality štatistických zisťovaní a ich produktov, a využívanie administratívnych zdrojov údajov na štatistické účely. Od mája 2018 je projektovou manažérkou projektu Scanner data.*

KONTAKT

helena.glaser-opitzova@statistics.sk
petra.mazurekova@statistics.sk

Informatívny článok/Informative article

Silvia KOMARA
Katedra štatistiky, Fakulta hospodárskej informatiky
Ekonomickej univerzity v Bratislave

Michal PÁLEŠ
Katedra matematiky a aktuárstva, Fakulta hospodárskej informatiky
Ekonomickej univerzity v Bratislave

VYUŽITIE JAZYKA PYTHON V OBLASTI WEB SCRAPINGU

THE USE OF THE PYTHON LANGUAGE IN WEB SCRAPING

ABSTRAKT
Príspevok sa zameriava na predstavenie základných atribútov web scrapingu v kontexte v súčasnosti tak skloňovaných pojmov, ako sú nové zdroje štatistiky veľké dáta, strojové učenie, umelá inteligencia, Business Intelligence a pod. Opisuje návrhy riešenia sťahovania údajov z internetu v jazyku Python a moduly, v ktorých možno tento proces realizovať. Špecificky sa venuje aj prepojeniu oblasti strojového učenia s web scrapingom. V praktickej ukážke predstavujeme funkcionalitu jazyka Python na získanie údajov z PDF dokumentov.

ABSTRACT
The paper focuses on presenting the basic attributes of web scraping in the context of currently used terms such as new sources of statistics, big data, machine learning, artificial intelligence, Business Intelligence, etc. It describes the Python language's options for downloading data from the Internet and modules in which this process can be executed. It is also specifically dedicated to connecting the field of machine learning with web scraping. In a practical demonstration, we present the functionality of the Python language for scraping data from the PDF documents.

KĽÚČOVÉ SLOVÁ
jazyk Python, web scraping, strojové učenie, PDF dokument

KEY WORDS
Python language, web scraping, machine learning, PDF dokument

1. ÚVOD
Dáta sú jednou z kľúčových hodnôt v dnešnej dobe a ich objem neustále rastie. Nové technológie prinášajú nové dáta, nové dáta generujú ďalšie dáta, vznikajú nové technológie a tento cyklus sa stále opakuje. Rozsiahla digitalizácia tento rast len urýchľuje. Ale s rastom objemu dát vzniká aj problém, a to zhoršenie ich kvality. Dáta sú cenné aktívum, ktoré je potenciálne schopné priniesť úžitok aj stratu. Kvalitné dáta s väčšou pravdepodobnosťou prinesú úžitok. Nekvalitné dáta v najlepšom prípade neprinesú nič. Aby teda dáta priniesli úžitok, je potrebné vedieť s nimi správne manipulovať a vedieť transformovať nekvalitné dáta na kvalitné. Rast objemu dát vo svete prispel k vytvoreniu konceptu Data Drive Decision Making, čo znamená rozhodovanie skôr na základe dát, faktov a metrík ako na základe intuície. Tento prístup umožňuje prijímať najlepšie rozhodnutie pre firmu. Na využitie daného

konceptu je však potrebné mať obrovské množstvo kvalitných dát. A pokiaľ firmy väčšinou nemajú problémy s objemom dát, tak s kvalitou je to čoraz zložitejšie. Získať znalosti z dát firmám pomáha proces, ktorý je známy ako Knowledge Discovery in Databases (KDD). KDD sa skladá z rôznych krokov,jedným je data mining (dolovanie dát) [12]. Jednou z oblastí data miningu je práve web scraping.

Automatizované zhromažďovanie údajov z internetu je takmer také staré ako samotný internet. Hoci web scraping nie je nový pojem, v minulých rokoch sa táto prax častejšie označovala ako screen scraping, data mining, web harvesting alebo podobne. Zdá sa, že všeobecný konsenzus dnes uprednostňuje web scraping [7]. Web scraping je technologické riešenie, ktoré extrahuje údaje z webových stránok rýchlym, efektívnym a automatizovaným spôsobom a ponúka údaje vo formáte, ktorý je štruktúrovaný a ľahko použiteľný. [3] Teoreticky je web scraping zhromažďovanie údajov akýmkoľvek iným spôsobom ako prostredníctvom priamej komunikácie s rozhraním webového servera (alebo samozrejme prostredníctvom človeka používajúceho webový prehliadač). Najčastejšie sa to dosahuje napísaním automatizovaného programu, ktorý webovému serveru odosiela žiadosti o určité údaje (zvyčajne vo forme HTML a iných súborov, z ktorých sa skladajú webové stránky), následne tieto údaje analyzuje s cieľom získať potrebné informácie a dáta ukladá v určitom formáte napríklad do dátového skladu. V praxi web scraping zahŕňa širokú škálu programovacích techník a technológií, ako je analýza údajov, rozbor prirodzeného jazyka a zabezpečenie informácií [7].

Štúdia v [6] definuje web scraping ako prvý krok v procese dolovania dát. Samotné dolovanie dát sa považuje za súčasť Business Intelligence (BI), prvýkrát rozpracované Howardom Dresnerom z Gartner Group v roku 1989. Podľa jeho názoru je Business Intelligence súbor konceptov a metód na zlepšenie procesu rozhodovania v manažmente pomocou informačných systémov, využívajúcich obchodné dáta. Podľa Lifecycle Software Ltd. existujú dva prvky, ktoré odlišujú BI systémy od iných, a to integrácia dát, teda zlučovanie údajov z rôznych zdrojov a v rôznych formátoch, a poskytovanie koherentného prístupu k nim: poskytovanie techník na analýzu a vizualizáciu informácií novým spôsobom, zrozumiteľným pre používateľov [11].

Internet obsahuje nespočetné množstvo informácií rôzneho druhu. Či už ide o informácie o počasí, marketingové dáta, rešerše a kvalitatívne dáta, dáta týkajúce sa sociálnych sietí a mnoho ďalších, ich analýza môže napomôcť k hlbšiemu pochopeniu konkrétnej problematiky. Tieto informácie však málokedy majú formu, ktorú je možné využiť na analýzu dát. Každá webová stránka má svoju štruktúru inú, avšak spája ich značkovací jazyk HTML. Špecifická forma tohto jazyka umožňuje pomocou nástrojov vyhľadať konkrétne informácie a uložiť ich do čitateľnejšieho formátu. Spôsob, akým nástroj vykonáva tento úkon, je veľmi podobný spôsobu, aký by použil bežný používateľ. Na internetovej stránke sa vyberú potrebné dáta, tie sa následne skopírujú a vložia do tabuľky. Takýto proces je možný v prípade malého počtu dát, avšak ak sa jedná o počet presahujúci tisíce jednotiek, je tento proces časovo veľmi náročný. V tomto prípade je možná náhrada používateľa za robota, ktorý opakovane vykonáva ten istý úkon [4].

Klasickým príkladom je napríklad knižnica Selenium v programovacom jazyku Python alebo Puppeteer v programovacom jazyku Javascript. Obe tieto metódy

fungujú na princípe skrytého prehliadača. Robot je nastavený na určitú webovú stránku a z tej potom vyberie dôležité dáta. Ak nie je potrebná žiadna interakcia s webovou stránkou, metódy so skrytým prehliadačom sú zbytočne komplikované. Dáta je možné získať priamo z konkrétnej webovej stránky pomocou príkazu „request“. Manipulácia s dátami je následne uľahčená použitím knižnice Beautiful Soup programovacieho jazyka Python [10].

Programovať web scraping od samého začiatku je však často náročná cesta. Dnes už existuje niekoľko webových aplikácií a rozšírení, ktoré scraping dát zvládnu bez programovania. Konkrétnym príkladom je portál import.io [9].

Web scraping môže byť využitý na zhromaždenie dátového setu obsahujúceho informácie o online cenách na vytvorenie denného prehľadu cien. Web scrapingom možno preskúmať a vyhodnotiť cenové praktiky využívané spoločnosťami pôsobiacimi v elektronickom obchode, resp. na realitnom trhu. Banky a iné finančné inštitúcie používajú web scraping na analýzu svojej konkurencie. Banky frekventovane sťahujú dáta konkurentov, napríklad o tom, kde sa novo otvorili či zavreli pobočky alebo tiež napríklad na sledovanie aktuálnej úrokovej sadzby pri pôžičkách. Tieto informácie potom zakomponujú do svojich interných modelov a predpokladov. Niektoré spoločnosti sa tiež špecializujú na predaj pracovných profilov pomocou zberu a analýzy verejne dostupných dát napríklad z LinkedIn [13]. V oblasti aktuárstva to môže byť napríklad prehľadávanie cenových kalkulačiek v rámci havarijného poistenia.

2. PROCES WEB SCRAPING

Proces sťahovania dát z internetu je možné prvotne rozdeliť do dvoch po sebe idúcich krokov (pozri obrázok č. 1):

1. **získanie webových zdrojov (webové stránky zdrojov),**
2. **extrahovanie požadovaných informácií zo získaných zdrojov.**

Konkrétne sa proces web scrapingu začína odosielaním požiadavky HTTP na získanie zdrojov cielenej webovej stránky. Táto požiadavka môže byť naformátovaná buď ako adresa URL obsahujúca dotaz GET, alebo ako časť správy HTTP obsahujúca dotaz POST.

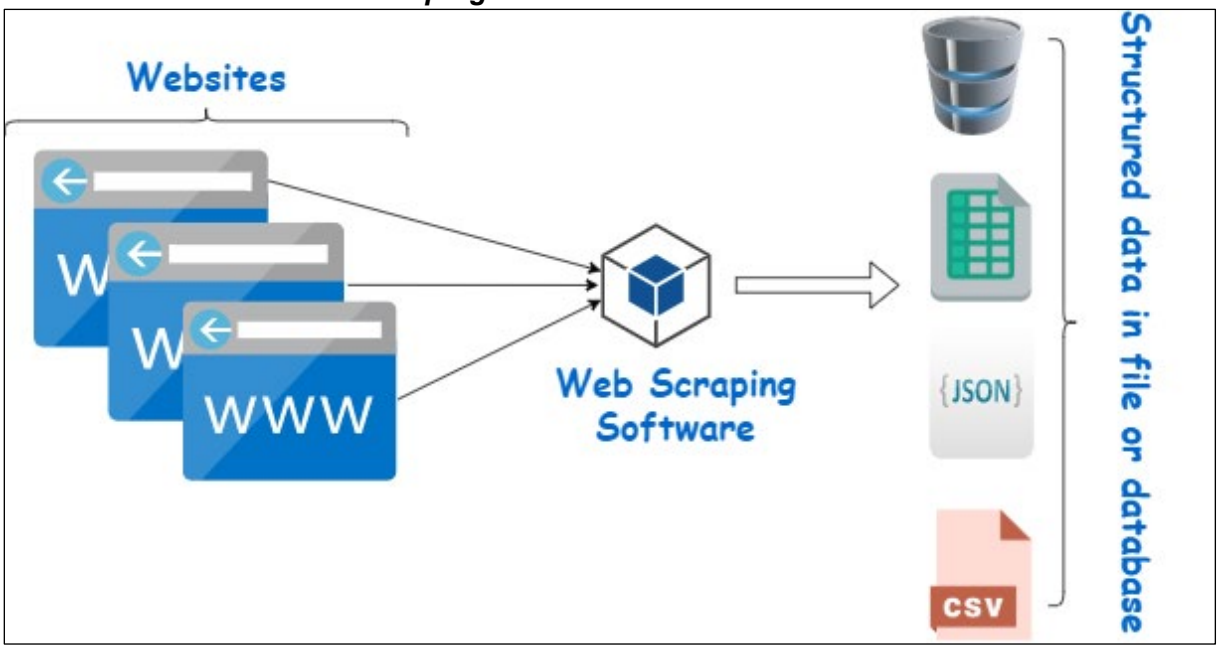
Len čo je požiadavka úspešne prijatá a spracovaná cielenou webovou stránkou, požadovaný zdroj bude z webovej stránky načítaný a potom odoslaný späť. Zdroj môže byť v niekoľkých formátoch, napríklad vo formáte HTML, XML alebo JSON. Zdroj môže tiež obsahovať multimediálne dáta, ako sú obrázky, audio alebo video súbory. Teda web scraping transformuje neštruktúrované dáta z webových stránok na štruktúrované.

Po získaní webových zdrojov proces extrakcie pokračuje v analýze, preformátovaní a usporiadaní dát štruktúrovaným spôsobom. Dôležitou súčasťou web scraperu sú dátové lokátory (čiže selektory) slúžiace na vyhľadanie dát, ktoré je potrebné z HTML súboru extrahovať – obvykle sa používa XPath, selektory CSS, regulárne výrazy alebo ich kombinácia.

Zjednodušený proces web scrapingu možno popísať nasledujúcimi krokmi (podľa [12], pozri tiež [1]):

1. **Identifikácia cieľovej webovej stránky.**
2. **Odosielanie požiadavky na URL adresu cieľovej webovej stránky.**
3. **Získanie HTML kódu stránky.**
4. **Vyhľadanie informácie v HTML kóde.**
5. **Uloženie dát do štruktúrovaného formátu (JSON, CSV a pod.).**

Obrázok č. 1: Proces web scrapingu



Zdroj: WebHarvy

Extrahované neštruktúrované údaje z webových lokalít je ďalej potrebné transformovať do potrebnej podoby hodnôt a štruktúry. Následne sa transformované a štruktúrované dáta uložia do databázy, resp. vyexportujú v príslušnom formáte. Transformácia údajov teda môže pozostávať z rôznych fáz, napríklad z čistenia údajov, prevodu kódovaných hodnôt, výpočtu nových hodnôt a pod. Významnú úlohu tu zohráva využitie rôznych techník machine learningu, databázových jazykov a štatistických metód (exploratívna analýza údajov). Pozri tiež kapitolu č. 4.

3. KNIŽNICE JAZYKA PYTHON PRE WEB SCRAPING

Jazyk Python umožňuje používanie širokej škály štatistických a grafických techník, ako napríklad regresná analýza, analýza časových radov, štatistické testy, zhluková analýza a i. Všetky tieto skutočnosti z neho robia ideálnu voľbu pre Data Science, analýzu Big Data a Machine Learning. Iinformácie o jazyku Python pozri v [8], resp. v inej zodpovedajúcej literatúre (tieto informácie tu neuvádzame).

Na rozšírenie možností jazyka sa v programovacích jazykoch používajú knižnice. Knižnica je súbor funkcií, tried, metód atď., zhromaždených na jednom mieste, ktoré sú potom využívané inými programami. Napríklad pre web scraping v jazyku Python sú to tieto knižnice (podľa [12]):

Requests (https://docs.python-requests.org)

Requests je jednoduchá knižnica, ktorá umožňuje odosielať HTTP požiadavky pomocou Pythonu. HTTP požiadavka potom vráti objekt odpovede so všetkými dátami (obsah, kódovanie, stav atď.). Nie je to knižnica výlučne určená pre web scraping, ale predstavuje pre neho osnovu.

BeautifulSoup (www.crummy.com/software/BeautifulSoup)

BeautifulSoup je jedným z najjednoduchších nástrojov pre web scrapingu v Pythone. Táto knižnica vyvinutá v roku 2004 poskytuje niekoľko jednoduchých metód na vyhľadávanie a extrahovanie potrebných dát. Niekedy funkcionality tejto knižnice úplne postačujú na vyriešenie problému, a zároveň výsledný skript nebude obsahovať mnoho kódov.

Tu však stojí za zmienku, že moderné webové stránky je možné rozdeliť na 2 typy: stránky so statickým obsahom a stránky s dynamickým obsahom. S druhým typom stránok sa BeautifulSoup nevysporiada. To znamená, že ak webová stránka obsahuje JavaScript alebo JQuery elementy, BeautifulSoup jednoducho nedokáže vyexportovať obsah vo vnútri nej.

Na druhú stranu má BeautifulSoup výhodu oproti iným nástrojom a tou je jeho schopnosť automaticky detegovať kódovanie, čo umožňuje spracúvať HTML dokumenty so špeciálnymi znakmi. Vie tak prevádzať prichádzajúce dokumenty na Unicode (medzinárodný štandard, ktorého cieľom je definovať kódovaciu schému schopnú reprezentovať väčšinu znakov používaných v písaných jazykoch spolu s inými symbolmi [15]) a odchádzajúce dokumenty na UTF-8 (8-bitový Unicode Transformation Format, je bezstratové kódovanie s variabilnou dĺžkou určené pre Unicode znaky [15]).

Selenium (https://www.selenium.dev)

Selenium je pôvodne automatizovaný testovací rámec používaný na overovanie webových aplikácií naprieč rôznymi prehliadačmi a platformami. Selenium umožňuje automatizovať webové prehliadače a má knižnice pre rôzne programovacie jazyky, vrátane Pythonu.

Selenium používa WebDriver na ovládanie webových prehliadačov, ako sú Chrome, Firefox alebo Safari. Postupom času sa však začala táto knižnica využívať nielen na testovanie aplikácií, ale aj na web scraping, a to vďaka svojej funkcionalite a kompatibilite s JavaScriptom.

Selenium je užitočný, keď je potrebné vykonať nejakú akciu na webe, napríklad na vypĺňanie polí alebo formulárov, rolovanie stránky, kliknutie na tlačidlá, vytvorenie snímky obrazovky. Ďalšou výhodou Selenia je možnosť fungovania s JavaScriptom. Vie napríklad načítať obsah vnorený do prvkov JavaScriptu. Selenium tiež podporuje tzv. *headless* prehliadače, čo sú prehliadače bez GUI, ktoré sa spúšťajú v príkazovom riadku. K výhodám podobných prehliadačov patrí väčšia rýchlosť a menšia spotreba pamäte.

Scrapy (https://www.scrapy.org)

Scrapy je open source a kolaboratívny rámec na extrahovanie dát z webových stránok rýchlym, jednoduchým a pritom rozšíriteľným spôsobom. V podstate ide

o najkomplexnejšie riešenie pre web scraping, ktoré poskytuje nástroje na prehliadanie webových stránok, sťahovanie dát, ich analýzu a ukladanie. Scrapy podporuje rozšírenie, čo prináša možnosť pridania proxy, spracovania súborov s cookies a ovládania hĺbky prehliadania.

Ďalšou vlastnosťou Scrapy je jeho asynchrónny spôsob spracovania požiadaviek. To umožňuje extrahovať dáta rýchlo aj z viacerých stránok naraz. Je však zrejmé, že keďže tento nástroj umožňuje viac, je tiež ťažšie ho nastaviť.

4. VYUŽTIE TECHNÍK STROJOVÉHO UČENIA

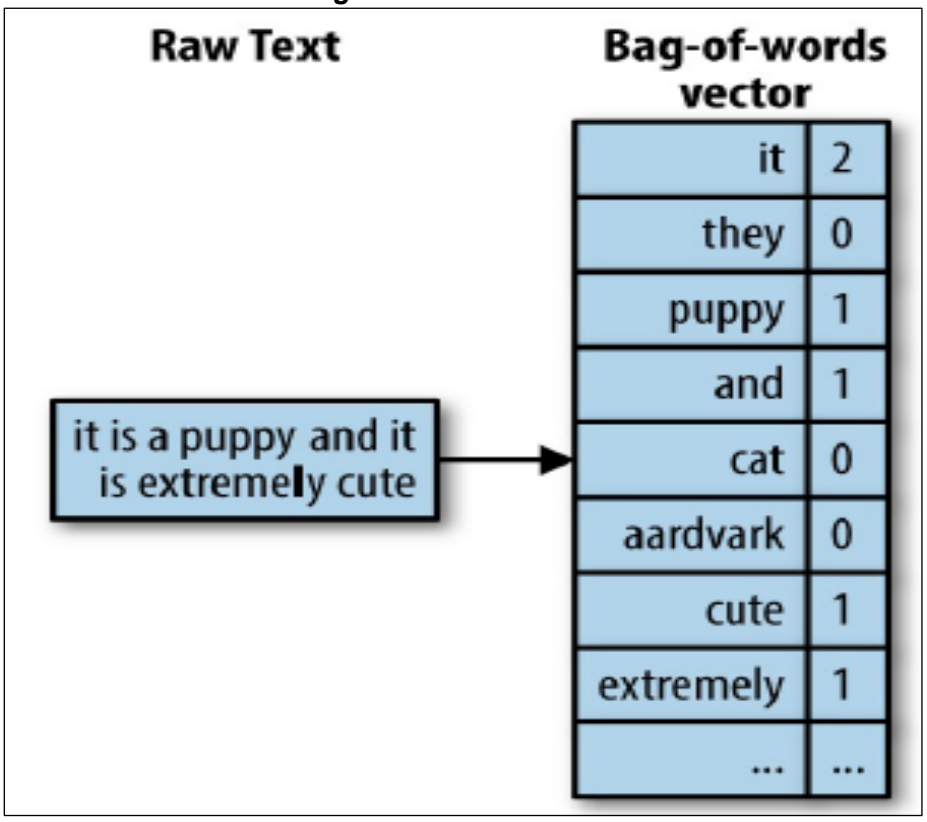
Autor v [2] skúma možnosti vytvorenia robustného algoritmu web scrapingu, ktorý je navrhnutý tak, aby neustále scrapoval konkrétnu webovú stránku, aj keď je HTML kód zmenený. Algoritmus je určený na použitie na webových stránkach, ktoré majú opakujúcu sa štruktúru HTML obsahujúcu údaje, pre ktoré je možné použiť web scraping. Opakujúca sa štruktúra HTML sa nachádza napríklad v spravodajských článkoch, videách, knihách a pod. V HTML tak vzniká kód, ktorý sa mnohokrát opakuje a líši sa tak napríklad len v nadpisoch. Dobrým príkladom môže byť napríklad Youtube. Scraper funguje pomocou klasifikácie textu slov v kóde HTML a trénuje podporný vektorový stroj na rozpoznávanie slov alebo názvov premenných. Klasifikácia slov obklopujúcich hľadané údaje sa vykonáva s predpokladom, že budúce HTML webovej stránky budú podobné súčasnej HTML, čo zase umožňuje vykonávať robustné scrapovanie. Na vyhodnotenie jeho výkonu sa používa webový archív, v ktorom sa výkon algoritmu spätne testuje na minulých verziách stránky, aby sme získali predstavu o tom, ako by mohol tento výkon v budúcnosti vyzerať. Algoritmus dosahuje rôzne výsledky v závislosti od veľkého množstva premenných v rámci samotných webových stránok, ako aj od minulých verzií webových stránok. Najlepšia výkonnosť bola napríklad dosiahnutá na Yahoo news s presnosťou 90 % z obdobia troch mesiacov od zastavenia scrapera.

Štandardný algoritmus scrapovania teda využíva statické cesty na navigáciu programu cez HTML k hľadaným údajom. Použitie statických ciest v nestatickom prostredí vedie k pokazeniu scrapera pri aktualizácii HTML vyžadujúcej aktualizáciu kódu, ktorý scrapovanie vykonáva. Frekvencia, s akou sa to musí robiť, do značnej miery závisí od webovej lokality a nemusí to byť príliš často. Ak sa firma spolieha na webový scraper pracujúci vo dne aj v noci, jeho pokazenie by mohlo spôsobiť narušenie procesov firmy. Keďže HTML je kód v textovom formáte, údaje musia byť pre model strojového učenia preformátované do číselnej formy. Predspracovanie údajov sa vykonáva pomocou extrakcie textových prvkov a využíva *bag-of-words model* s ďalšími úpravami na vytvorenie optimálneho výkonu. Predspracované údaje sa vložia do stroja Support Vector Machine (SVM, podporný vektorový stroj) na klasifikáciu údajov a extrahovanie údajov, ktoré sa pôvodne požadovali.

SVM sú trénované prostredníctvom učenia s učiteľom, čo znamená, že údaje, na ktorých je model trénovaný, musia byť označené. SVM môže byť použitý na klasifikáciu alebo regresiu. SVM na klasifikáciu v *N*-rozmernom kontexte klasifikuje dáta rozdelením *N*-rozmerného priestoru nadroviny tak, aby bolo správne klasifikovaných čo najviac bodov. [2] Pre optimálnu nadrovinu platí, že musí byť umiestnená v čo najväčšom odstupe (angl. Maximal margin) od krajných bodov, nazývaných podporné vektory (angl. Support vectors). Lineárny variant tohto algoritmu sa používa, keď sú dáta lineárne oddeliteľné, čiže oddeliteľné lineárnou nadrovinou. Algoritmus pracuje

v pôvodnej dvojrozmernej rovine dát. Nelineárna SVM sa používa, keď rovinu dát nie je možné rozdeliť lineárne. Tu sa uplatňuje funkcia nazývaná jadrová transformácia (angl. kernel transformation).

Obrázok č. 2: Príklad bag-of-words vektora

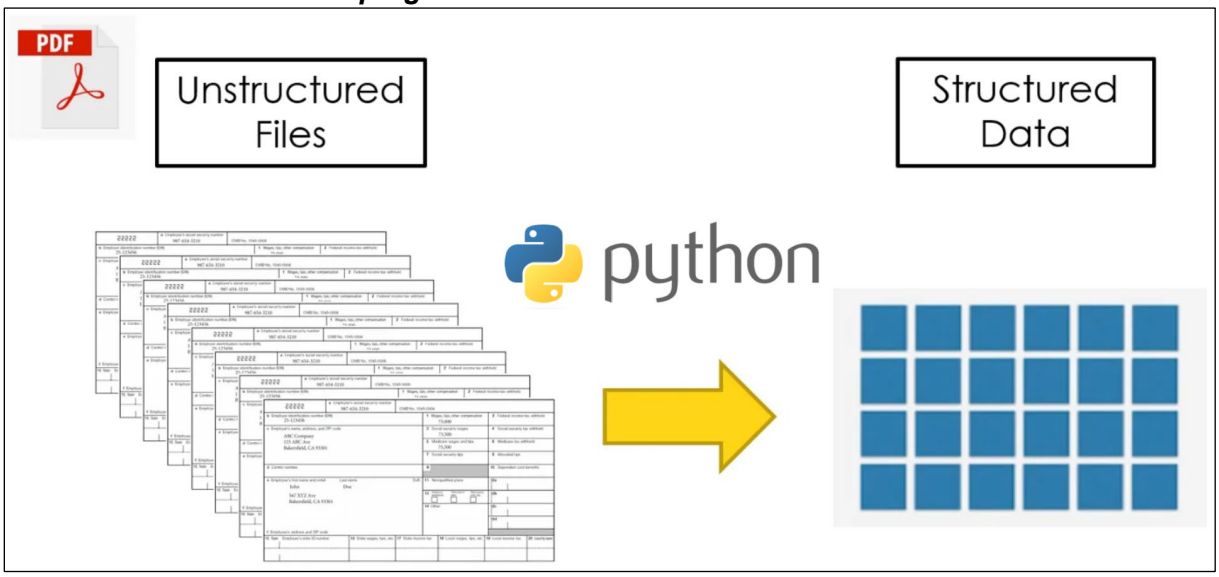


Zdroj: [2]

5. PRAKTICKÁ UKÁŽKA ZÍSKAVANIA ÚDAJOV Z PDF DOKUMENTOV

Väčšinou je web scraping založený na analýze štruktúry webovej stránky, a preto je dôležité pochopiť jej princíp. Existujú tzv. značkovacie jazyky, ktoré pomocou špeciálnych značiek vysvetľujú význam (sémantiku) rôznych častí textu alebo určujú vzhľad (formát) jednotlivých častí textu. K najvýznamnejším značkovacím jazykom patrí HTML a XML. HTML a XML sa používajú na ukladanie štruktúrovaných dát, prípadne na tvorbu vizuálneho pohľadu webových stránok. Tieto formáty definujú obsah celej webovej stránky, zatiaľ čo jedným z krokov web scrapingu je vyhľadanie konkrétnej informácie v zdrojovom kóde. Na tento krok je možné použiť XML Path Language (angl. skratka XPath). XPath je dotazovací jazyk, ktorý je užitočný pre identifikáciu a extrahovanie častí z dokumentov HTML/XML. Každý HTML alebo XML dokument si možno predstaviť ako strom. XPath potom umožňuje vyhľadávanie v podobných dokumentoch pomocou dotazu. [12] Keďže ide o širokú problematiku na ukážku jednoduchého cenového web scrapera v jazyku Python pozri napr. [5].

Obrázok č. 3: Proces scrapingu z PDF dokumentov



Zdroj: [14], upravené autormi

Niekedy však môžu byť údaje uložené aj v nekonvenčnom formáte, ako je napríklad PDF (obrázok č. 3). Ďalej teda budeme prezentovať prístup scrapingu z PDF pomocou Python modulu tabula-py. Pre viac informácií odporúčame dokumentáciu na webovej stránke: https://tabula-py.readthedocs.io/en/latest/.

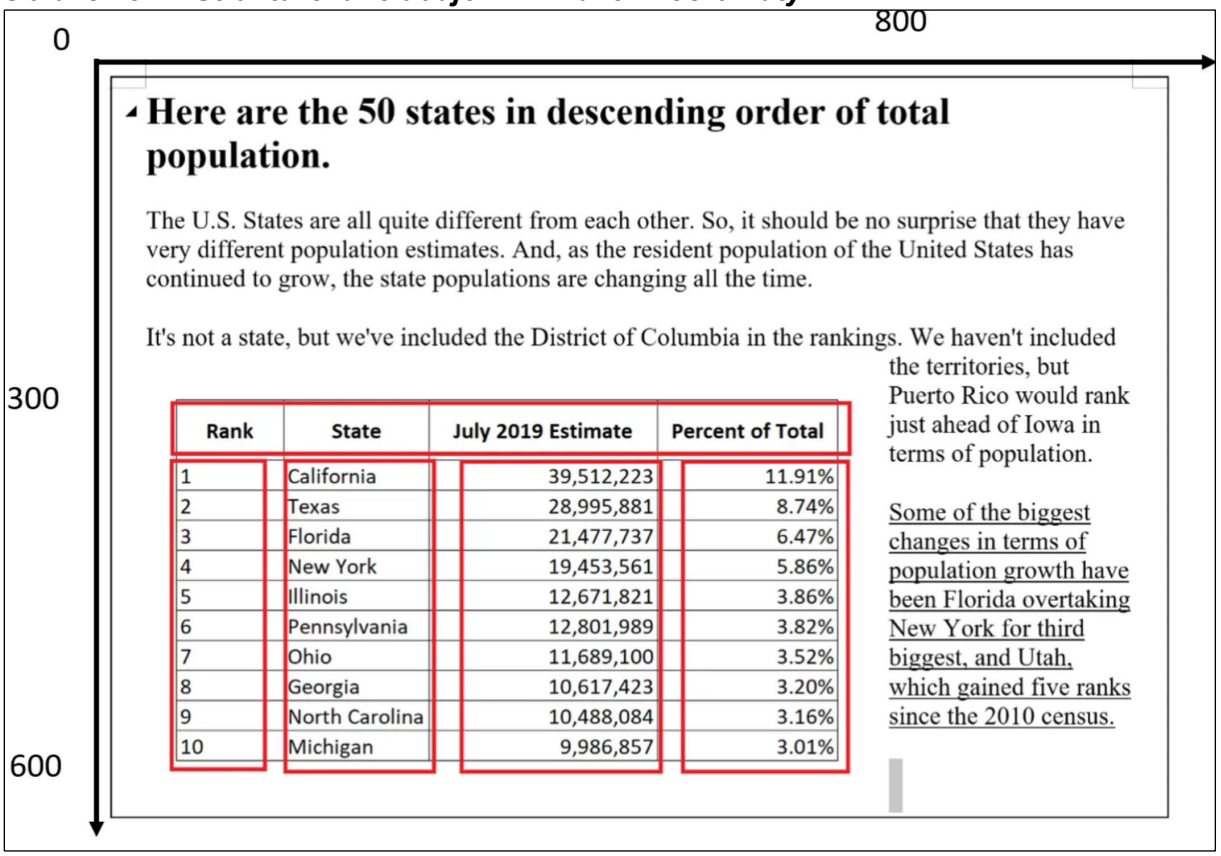
Jednoduchší prípad je, ak extrahujeme údaje z PDF v štruktúrovanom formáte. Ak napríklad chceme prehľadať tabuľku zvýraznenú na obrázku č. 4 – štruktúrované tabuľkové údaje, v ktorých sú prehľadne definované riadky a stĺpce. Extrahovanie takýchto údajov z PDF v štruktúrovanej forme je jednoduché využitím práve Python modulu tabula-py. Potrebujeme len zadať umiestnenie tabuľkových údajov (koordináty) na stránke PDF zadaním (hore, vľavo, dole, vpravo) súradníc danej oblasti. Ak stránka PDF obsahuje iba cieľovú tabuľku, potom ani nemusíme špecifikovať oblasť, funkcia tabula-py by mala byť schopná automaticky rozpoznať riadky a stĺpce danej tabuľky. Nižšie uvádzame kód v jazyku Python na extrahovanie údajov z PDF v štruktúrovanej forme:

```
pip install tabula-py
pip install pandas

import tabula as tb
import pandas as pd
import re

file = 'moje_pdf_1.pdf'
data = tb.read_pdf(file, area = (300, 0, 600, 800), pages = '1')
```

Obrázok č. 4: Štruktúrované údaje v PDF a ich koordináty



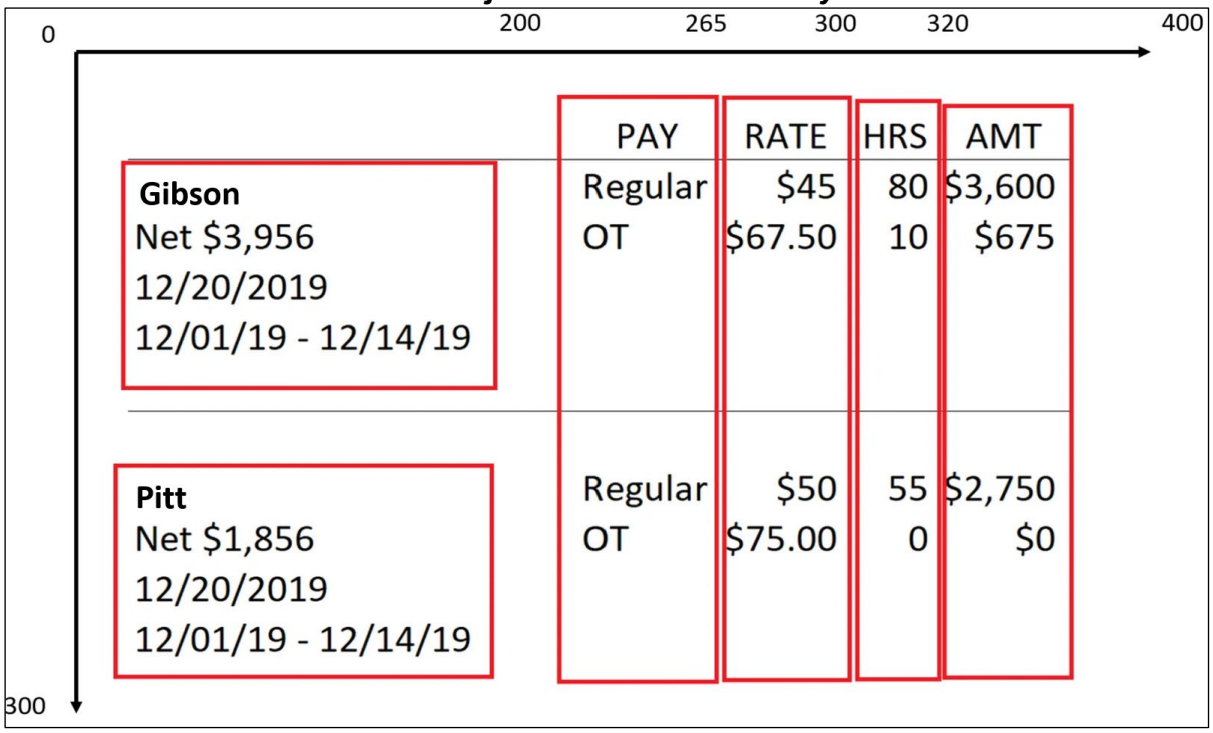
Zdroj: [14]

Na implementáciu štatistických analýz, vizualizáciu údajov a aplikáciu modelov strojového učenia potrebuje analytik údaje obvykle v tabuľkovej forme (panelové údaje). Mnohé údaje sú však dostupné iba v neštruktúrovanom formáte (pozri obrázok č. 5).

Najskôr importujeme údaje rovnakým spôsobom ako údaje v štruktúrovanom formáte, pričom musíme špecifikovať ďalšie atribúty na správny import údajov:

```
file = 'moje_pdf_2.pdf'
df= tb.read_pdf(file, pages = '1', area = (0, 0, 300, 400), columns = [200, 265, 300, 320], pandas_options={'header': None}, stream=True)[0]
```

Obrázok č. 5: Neštruktúrované údaje v PDF a ich koordináty



Zdroj: [14], upravené autormi

Tu používame aj nastavenie na identifikáciu umiestnení všetkých relevantných stĺpcov. Niekedy sa využíva metóda *pokus-omyl*. Ak existujú vo vstupe mriežkové čiary, ktoré oddeľujú bunky, môžeme použiť lattice = True na automatickú identifikáciu každej bunky. Ak nie, môžeme použiť stream = True a columns na manuálne určenie každej bunky. Režim stream bude hľadať medzery medzi stĺpcami. Tieto nastavenia sú často veľmi relevantné a môžu výrazne vylepšiť celkový scraping.

Obrázok č. 6: Získané neštruktúrované údaje z PDF

0	nan	PAY	RATE	HRS	AMT
1	Gibson	Regular	$45	80	$3,600
2	Net $3,956	OT	$67.50	10	$675
3	12/20/2019	nan	nan	nan	nan
4	12/01/19 - 12/14/19	nan	nan	nan	nan
5	Pitt	Regular	$50	55	$2,750
6	Net $1,856	OT	$75.00	0	$0
7	12/20/2019	nan	nan	nan	nan
8	12/01/19 - 12/14/19	nan	nan	nan	nan

Zdroj: vlastný, podľa údajov [14]

Následne získame údaje (obrázok č. 6), s ktorými môžeme ďalej pracovať. Na manipuláciu s dátovým rámcom môžeme využiť v jazyku Python napríklad knižnicu Pandas a naprogramovať kód (tento postup tu pre rozsah a náročnosť neuvádzame), ktorý údaje upraví do finálnej podoby (obrázok č. 7).

Obrázok č. 7: Získané štruktúrované údaje z neštruktúrovaných údajov z PDF

Index	row	Surname	net_amount	pay_date	pay_period
0	1	Gibson	$3,956	12/20/2019	12/01/19 - 12/14/19
1	2	Pitt	$1,856	12/20/2019	12/01/19 - 12/14/19

Index	row	OT_Rate	Regular_Rate	OT_Hours	Regular_Hours	OT_Amt	Regular_Amt
0	1	$67.50	$45	10	80	$675	$3,600
1	2	$75.00	$50	0	55	$0	$2,750

Zdroj: vlastný, podľa údajov [14]

6. ZÁVER

Web scraping je populárny spôsob získavania dát, ktorý má však má aj slabé stránky. Niektoré weby sa snažia zbaviť web scraperov a používajú tak rôzne techniky, ako sa vyrovnať s nežiaducim scrapovaním webu. Moderné riešenia umožňujú takéto nástroje detegovať a blokovať. Sem napríklad patria: CAPTCHA, Honeypot, dynamický obsah, zmeny štruktúry webový stránok, mnoho častých HTTP požiadaviek, IP blokovanie.

Niekedy vlastníci webových stránok píšu o tom, či je možné kopírovať obsah ich webu priamo na stránkach, a to spravidla dole, v tzv. footeri stránky. Ak tam nie je nič napísané, ďalším vhodným krokom je otvoriť súbor robots.txt.

Keďže web scraping je extrakcia dát z verejných webových stránok, tieto stránky niekedy obsahujú dáta, ktoré patria práve ich vlastníkom, na základe čoho vznikla diskusia, či je web scraping legálny. Na túto otázku stále neexistuje definitívna odpoveď. Na jednej strane sú údaje zverejnené na stránkach verejne dostupné, na druhej strane to, že sú tieto dáta zverejnené na internete, neznamená, že ich môže ktokoľvek používať. To platí najmä pre citlivé dáta alebo osobné údaje, ktoré podliehajú GDPR (General Data Protection Regulation) [12].

Práve prezentácie ukážky extrakcie údajov z PDF sa nemusia významne dotýkať týchto problémov pretože používateľ môže využiť scraping na získanie údajov, ktoré sa môžu týkať vlastnej firmy, resp. organizácie (faktúry, vlastné zoznamy a dokumenty, sprístupnené a schválené zdroje a pod.).

V súčasnosti veľa spoločností stále manuálne spracúva údaje z PDF. Pomocou predstavenej procedúry môžu ušetriť čas aj zdroje automatizáciou tohto procesu získavania údajov zo súborov PDF a konverziou neštruktúrovaných údajov na údaje panelové. Uveďme však, že aj tu je potrebné poznať zmluvné podmienky, resp. povolenia autora PDF dokumentu na takéto operácie.

Príspevok bol vytvorený v rámci projektov VEGA č. 1/0431/22; 1/0561/21.

LITERATÚRA

[1] BIHÁNY, D.: Cenové stratégie za využitia analýzy veľkých dát. Bakalářská práce: FPH, VŠE v Praze, 2022.
[2] CARLE, V.: Web Scraping using MachineLearning. 2020. [online]. [cit. 19. 4. 2023]. Dostupné na: https://www.diva-portal.org/smash/get/diva2:1468583/FULLTEXT01.pdf

[3] CASTRILLO-FERNÁNDEZ, O.: Web Scraping: Applications and Tools. 2015. [online]. [cit. 11. 4. 2022]. Dostupné na: https://data.europa.eu/sites/default/files/report/2015_web_scraping_applications_and_tools.pdf
[4] DENSMORE, J.: Ethics in Web Scraping. 2017. [online]. [cit. 21. 2. 2021]. Dostupné na: https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01
[5] KOMARA, S. – PÁLEŠ, M.: Web scraping v súčasnom konkurenčnom prostredí. In: Sborník příspěvků z 15. ročníku Mezinárodní vědecké konference KONKURENCE, Jihlava, Česká republika, 2023.
[6] MILEV, P.: Conceptual Approach for Development of Web Scraping Application for Tracking Information. In: Economic Alternatives, 2017, (3), s. 475 – 485.
[7] MITCHELL, R.: Web Scraping with Python. O'Reilly, 2018.
[8] PÁLEŠ, M.: Jazyk Python pre aktuárov. Bratislava: Letra Edu, 2022.
[9] PAPCO, L.: Efekt krátkodobého pronajímání nemovitostí na cenu nemovitostí v Praze. Diplomová práce: FPH, VŠE v Praze, 2021.
[10] SIRISURIYA, D. S.: A comparative study on web scraping. 2015. [online]. [cit. 19. 2. 2021]. Dostupné na: http://ir.kdu.ac.lk/bitstream/handle/345/1051/com-059.pdf?sequence=1&isAllowed=y
[11] STEFANOVA, K.: Factors and Main Directions for Business Intelligence Systems Design and Development. 2008. [online]. [cit. 14. 3. 2022]. Dostupné na: http://unweyearbook.org/uploads/Yearbook/Yearbook_2008_No6_K%20Stefanova.pdf
[12] TSAKUNOV, I.: Využití data miningu pro analýzu českého realitního trhu. Bakalářská práce: FIS, VŠE v Praze, 2022.
[13] ZANIKOV, M.: Analýza a vizualizace dat z webových portálů nabídek práce. Diplomová práce: FIS, VŠE v Praze, 2020.
[14] ZHU, A.: How to Scrape and Extract Data from PDFs Using Python and tabula-py. 2021. [online]. [cit. 19. 4. 2023]. Dostupné na: https://towardsdatascience. com/scrape-data-from-pdf-files-using-python-fe2dc96b1e68
[15] https://en.wikipedia.org/wiki/Unicode [cit. 19. 5. 2022].
[16] https://en.wikipedia.org/wiki/Web_scraping [cit. 19. 5. 2022]

RESUMÉ

V tomto príspevku sa zameriavame na špecifickú problematiku scrapingu, ktorou je extrakcia údajov z PDF dokumentov s využitím modulu tabula-py jazyka Python. Prezentujeme aj transformáciu neštruktúrovaných údajov na štruktúrované, čo môže byť využité aj pri klasickom web scrapingu.
Web scraping, web harvesting alebo extrakcia údajov z webu zahŕňa metódy na extrakciu údajov z webových stránok. Softvér web scrapingu môže priamo pristupovať na webovú stránku pomocou HTP protokolu alebo pomocou webového prehliadača. V súčasnosti tento pojem zahŕňa automatizované procesy implementované pomocou robota alebo webového prehľadávača na zhromažďovanie a kopírovanie špecifických údajov z webu, zvyčajne do centrálnej lokálnej databázy alebo tabuľky, na neskoršie vyhľadávanie alebo analýzu. Scraping webovej stránky zahŕňa jej načítanie a extrakciu údajov z nej. Obsah stránky možno analyzovať, prehľadávať a preformátovať a jej údaje skopírovať do tabuľky alebo načítať do databázy. Príkladom môže byť vyhľadanie a skopírovanie mien a telefónnych čísel, názov spoločnosti a ich URL alebo e-mailových adries do zoznamu (scrapingových kontaktov). Okrem toho sa web scraping používa ako súčasť aplikácií na indexovanie webu, ťažbu z webu a dolovanie údajov, online sledovanie zmeny cien a porovnávanie cien, scrapovanie recenzií produktov (na sledovanie konkurencie), zhromažďovanie

zoznamov nehnuteľností, údajov o počasí, detekciu zmien webových stránok a pod. Webové stránky sú vytvorené pomocou značkovacích jazykov založených na texte (HTML a XHTML) a často obsahujú množstvo užitočných údajov v textovej forme. Väčšina webových stránok je však navrhnutá pre koncových používateľov a nie na automatizované používanie. V dôsledku toho boli vyvinuté špecializované nástroje a softvér na uľahčenie scrapovania webových stránok.
Existuje mnoho dostupných softvérových nástrojov, ktoré možno použiť na web scraping. Takýto softvér sa môže pokúsiť automaticky rozpoznať dátovú štruktúru stránky alebo poskytnúť nahrávacie rozhranie, ktoré odstraňuje nutnosť manuálneho zápisu kódu na scrapovanie webu, alebo niektoré skriptovacie funkcie, ktoré možno použiť na extrahovanie a transformáciu obsahu, a databázové rozhrania, ktoré môžu ukladať skopírované údaje v lokálnych databázach. Niektorý softvér možno použiť aj na priamu extrakciu údajov z rozhrania API (Application Programming Interface) [16]. Jedným často využívaným softvérom je jazyk Python (napríklad s využitím modulu BeautifulSoup). Strojové učenie môže byť využité na vytvorenie robustného algoritmu web scrapingu, ktorý je navrhnutý tak, aby neustále scrapoval konkrétnu webovú stránku, aj keď je HTML kód zmenený. Algoritmus je určený na použitie na webových stránkach, ktoré majú opakujúcu sa štruktúru HTML obsahujúcu údaje, pre ktoré je možné použiť web scraping. Štandardný algoritmus scrapovania teda využíva statické cesty na navigáciu programu cez HTML k hľadaným údajom. Použitie statických ciest v nestatickom prostredí má napríklad za následok pokazenie scrapera pri aktualizácii HTML. Techniky strojového učenia sa využívajú vtedy, keď údaje musia byť pre model strojového učenia preformátované do číselnej formy. Predspracovanie údajov sa realizuje pomocou extrakcie textových prvkov a využíva *bag-of-words model* s ďalšími úpravami na vytvorenie optimálneho výkonu. Predspracované údaje sa vložia do stroja Support Vector Machine (SVM, podporný vektorový stroj) na klasifikáciu údajov a extrahovanie údajov, ktoré boli pôvodne požadované.

RESUME
In this paper, we focus on the specific issue of scraping, which is the extraction of data from PDF documents using the tabula-py module of the Python language. We also present the transformation of unstructured data into structured data, which can also be used in classic web scraping.
Web scraping, web harvesting, or web data extraction involves methods for extracting data from web pages. Web scraping software can directly access a website using the HTP protocol or using a web browser. Currently, the term includes auto-mated processes implemented using a robot or web crawler to collect and copy specific data from the web, usually into a central local database or table, for later retrieval or analysis. Scraping a website involves its loading and data extracttion data from it. The content of the page can be analyzed, searched and reformatted, and its data can be copied into a table or loaded into a database. An example could be searching and copying names and phone numbers, names of companies and their URLs or email addresses into a list (scraping contacts). In addition, web scraping is used as part of applications used for web indexing, web mining and data mining, online price change tracking and price comparison, scraping product reviews (to track competitors), collecting real estate listings, weather data, website change detection etc. Web pages are created using text-based markup languages (HTML and XHTML) and often contain a lot of useful data in text form. However, the most websites are designed for end users and not for an automated use. As a result, specialized tools and software have been developed to facilitate website scraping.

There are many software tools available that can be used for web scraping. Such software may attempt to automatically recognize the data structure of a page or provide an upload interface that eliminates the need to manually write web scraping code, or some scripting functions that can be used to extract and transform content, and database interfaces that can store scraped data in local databases. Some software can also be used to directly extract data from an Application Programming Interface (API) [16]. One frequently used software is the Python language (for example, using the BeautifulSoup module). Machine learning can be used to create a robust web scraping algorithm that is designed to continuously scrape a specific web page even if the HTML code is changed. The algorithm is intended for use on web pages that have a repeating HTML structure containing data that can be web scraped. Thus, the standard scraping algorithm uses static paths to navigate the program through HTML to the searched data. For example, using static paths in a non-static environment results in the scraper breaking when updating the HTML. Machine learning techniques are used when data needs to be reformatted into numerical form for a machine learning model. Data pre-processing is performed using text feature extraction and uses a "bag-of-words model" with additional adjustments to create optimal performance. The pre-processed data is fed into a Support Vector Machine (SVM) to classify the data and extract the data that was originally requested.

PROFESIJNÝ ŽIVOTOPIS

Ing. Silvia Komara, PhD., *pôsobí na Katedre štatistiky Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave. Jej vedecká činnosť je orientovaná v prvom rade na modelovanie a analýzu ekonomických časových radov a finančných časových radov so zameraním na kvalitu krátkodobej prognózy. Okrem toho sa venuje metódam strojového učenia (machine learning). Vyučuje predmety analýza časových radov, Machine Learning na 2. stupni štúdia v študijnom programe Data Science v ekonómii a predmety štatistika a štatistika v anglickom jazyku na 1. stupni štúdia. Absolvovala viaceré zahraničné pobyty na univerzitách v Madride, Sydney, Ostrave, Girone a i.*

Doc. Ing. Michal Páleš, PhD., *pôsobí ako vedúci Katedry matematiky a aktuárstva Fakulty hospodárskej informatiky Ekonomickej univerzity v Bratislave. V rámci pedagogickej činnosti vyučuje predmety matematika, matematika pre ekonómov, teória pravdepodobnosti, softvérové aplikácie pre aktuárov, teória rizika v poistení, úvod do aktuárstva a vybrané kapitoly z matematiky pre ekonómov. Vo svojej vedeckej práci sa orientuje na aktuársku vedu, využitie kvantitatívnych metód v ekonómii a softvérovú podporu riadenia rizík (najmä jazyk R). Je členom Slovenskej spoločnosti aktuárov a autorom viacerých vedeckých monografií, medzinárodne ocenených vysokoškolských učebníc a článkov z oblasti aktuárstva.*

KONTAKT

silvia.komara@euba.sk
michal.pales@euba.sk

Informatívny článok/Informative article

POTENCIÁL VYUŽÍVANIA BIG DATA V ŠTATISTIKE (NESMIE NÁM UJSŤ VLAK)

THE POTENTIAL OF USING BIG DATA IN STATISTICS (WE CAN´T MISS THE TRAIN)

ABSTRAKT

Príspevok sa zaoberá iniciatívami Eurostatu v oblasti využívania Big Data v oficiálnych štatistikách a transformáciou týchto iniciatív na podmienky Štatistického úradu Slovenskej republiky. Tieto iniciatívy sa v súčasnosti realizujú prostredníctvom dvoch projektov financovaných zo zdrojov Európskej únie, pričom ich cieľom je preukázať vhodnosť využívania Big Data pri štatistickej produkcii v rôznych sférach štatistiky.

ABSTRACT

The paper deals with Eurostat's initiatives in the field of using Big Data in official statistics and the transformation of these initiatives into the conditions of the Statistical Office of the Slovak Republic. These initiatives are currently implemented through two projects funded by the European Union, while their goal is to demonstrate the appropriateness of using Big Data in statistical production in various spheres of statistics.

KĽÚČOVÉ SLOVÁ

Big Data, oficiálna štatistika, cenové indexy, index sociálneho napätia, pohyb obyvateľstva, výkonnosť ekonomiky

KEY WORDS

Big Data, official statistics, price indices, social tension index, population movement, performance of the economy

1. ÚVOD

Big Data prinášajú nesmierne dôležitý rozmer do všetkých sfér života. Britský matematik Clive Humby už v roku 2006 slávne povedal, že „údaje sú nová ropa“. Odvtedy za posledných 16 rokov spoločnosti všetkých veľkostí a vo všetkých odvetviach vyhľadávali a ukladali čoraz viac údajov o svojich zákazníkoch, obchodných operáciách, aby vedeli podporiť svoj výkon a dosiahli stanovené ciele [1].

Treba povedať, že: „*Potenciál Big Data pre oficiálnu štatistiku spočíva práve v obrovskom množstve informácií, ktoré obsahujú. Môžu tiež pokrývať oblasti spoločnosti, pre ktoré ešte neexistujú oficiálne štatistiky. Ich veľkosť a objem môže viesť k vyššej presnosti a získaniu väčších detailov na tvorbu štatistík. Vysoká rýchlosť ich produkcie môže viesť k častejším a včasnejším štatistickým odhadom a ich veľká rozmanitosť zároveň k tvorbe štatistík v nových oblastiach.“* [2]

Z toho vyplýva, že ak Štatistický úrad SR nevyužije v dobe Big Data ich potenciál, môže prísť komerčný sektor a efektívne tieto údaje využiť a vytvoriť tak suplement k štatistickej produkcii, ktorý sa časom môže stať presnejším, efektívnejším a zrozumiteľnejším. A preto nie je na čo čakať a je potrebné, aby sa Štatistický úrad SR vydal smerom k Big Data.

2. VÝCHODISKÁ NA VYUŽITIE BIG DATA V ŠTATISTIKE V PODMIENKACH SR

Štatistický úrad SR má povinnosť a zodpovednosť za hľadanie a využívanie vhodných údajov na štatistickú tvorbu.

Štatistický úrad Slovenskej republiky je v zmysle §5 ods. 1 zákona č. 540/2001 Z. z. zákona o štátnej štatistike ústredným orgánom štátnej správy pre oblasť štátnej štatistiky, je kontaktným miestom pre Európsku komisiu v rámci Európskeho štatistického systému a koordinátorom Národného štatistického systému.

Jednou zo základných úloh Štatistického úradu SR je získavanie štatistických údajov a štatistických informácií potrebných na posudzovanie sociálno-ekonomického vývoja spoločnosti, poskytovanie a zverejňovanie štatistických informácií, koordinácia Národného štatistického systému Slovenskej republiky a plnenie záväzkov vyplývajúcich z medzinárodných zmlúv v oblasti štátnej štatistiky, ktorými je Slovenská republika viazaná.

Realizácie projektov v súlade s iniciatívami Európskej komisie v oblasti Big Data pomôžu zviditeľniť a spropagovať Štatistický úrad SR.

Práve v súvislosti s vyššie uvedeným a s možnosťou získavania dátových zdrojov sa dlhodobo rozvíja iniciatíva Európskej komisie prostredníctvom Eurostatu na využívanie Big Data v kontexte oficiálnych štatistík. Táto iniciatíva sa transformovala do projektu ESSnet Big Data (Big data: from exploration to exploitation). Cieľom projektu je integrácia Big Data do pravidelnej tvorby oficiálnych štatistík prostredníctvom pilotných prieskumov potenciálu vybraných Big Data a budovania konkrétnych aplikácií. Slovensko sa do tejto iniciatívy zapojilo ako jeden z 28 partnerov (*https://cros-legacy.ec.europa.eu/content/essnet-big-data-1_en).*

Závery iniciatívy podľa Eurostatu sú:

- údaje obsiahnuté na webových stránkach môžu byť prínosné pre oficiálnu štatistiku,
- tradičné spôsoby zberu môžu byť obohatené týmito údajmi,
- údaje môžu pomôcť získavať ďalšie indikátory, ktoré nie je možné produkovať z tradičných zdrojov údajov,
- **nevyužívaním alternatívnych zdrojov údajov v oblasti oficiálnych štatistík sa môže oslabiť postavenie národných štatistických úradov v prospech podnikateľského komerčného sektora.**

Preto sa aj Štatistický úrad SR rozhodol prispieť k vytváraniu pilotných riešení na transformáciu Big Data do štatistických produktov, či už vo forme experimentálnej štatistiky, alebo aj do oficiálnych štatistík po overení stability a vhodnosti využitia príslušných údajov.

3. TRANSFORMÁCIA INICIATÍVY EURÓPSKEJ KOMISIE V PODMIENKACH ŠTATISTICKÉHO ÚRADU SR

Realizácia experimentálnej štatistiky nie je jednoduchá a vyžaduje si relatívne vysoké finančné a odborné kapacity. V období rozvíjania myšlienok o využití Big Data v experimentálnej štatistike Ministerstvo investícií, regionálneho rozvoja a informatizácie SR vyhlásilo výzvu, zameranú na zlepšenie využívania údajov vo verejnej správe (https://www.mirri.gov.sk/projekty/projekty-esif/operacny-program-

integrovana-infrastruktura/prioritna-os-7-informacna-spolocnost/vyzvania-a-vyzvy/vyzva-c-opii-2019-7-10-dop/index.html).

Cieľom tejto dopytovo orientovanej výzvy je realizovať projekty zamerané na analytické využitie údajov v definovanej problémovej oblasti s cieľom poskytnúť riešenia na zlepšenie rozhodovania a zabezpečenia fungovania verejnej správy.

Štatistický úrad sa zapojil do tejto výzvy predložením dvoch projektov, na ktorých realizáciu prostredníctvom žiadosti o nenávratný finančný príspevok získal zdroje práve prostredníctvom predmetnej výzvy. Ide o nasledovné projekty:

- Dynamický cenový model – https://www.itms2014.sk/schvalena-zonfp?id=51997001-4fc1-42d9-8191-c2b15abbaa74
- Socioekonomické aspekty Big Data v štatistike – https://www.itms2014.sk/schvalena-zonfp?id=c13d0406-9036-44bb-8b76-51f5f9b53193

V nasledujúcom texte uvádzame rámcové zameranie projektov, ktoré bolo definované na začiatku kreovania projektov:

- Socioekonomické aspekty Big Data v štatistike – účelom realizácie projektu je spracovanie veľkých objemov dát od mobilných operátov a zo sociálnych sietí na vytvorenie modelov pohybu populácie a modelu nálad populácie s dôrazom na:
  - spracovanie údajov a prepojenie do GIS na identifikovanie sociálnych vplyvov pohybu populácie,
  - spracovanie analýz na vypracovanie odhadov tzv. dennej populácie, ktoré poskytujú prehľad o počte ľudí umiestnených v konkrétnych oblastiach v konkrétnom čase,
  - spracovanie informácií na sociálnych sieťach a definovanie závislosti medzi náladou v spoločnosti a súčasnou socioekonomickou problematikou v jednotlivých regiónoch,
  - spracovanie modelu dopravy z údajov z Národnej diaľničnej spoločnosti NDS a definovanie odhadov HDP, resp. indexu priemyselnej produkcie.
- Dynamický cenový model – účelom realizácie projektu je otestovanie možností využitia iných ako v súčasnosti využívaných údajov pre cenové štatistiky, ktoré Štatistický úrad SR realizuje. Ide o využitie moderných metód alternatívneho zberu údajov ako napr. web scrapingu, data feeding (alebo obdobných) na získavanie údajov o cenách a porovnávanie výsledkov s tradičnými formami sledovania cien.
- Výstupmi projektu budú cenové modely, ktoré budú testované v porovnaní s existujúcim štandardným postupom v rámci štatistickej produkcie.

4. PROCES NASTAVENIA REALIZÁCIE PROJEKTOV A ICH REALIZÁCIA

Realizácia projektov nie je samoúčelná a na ich obhajobu bolo potrebné definovať hypotézy a benefity tak, aby mohli získať podporu

V procese prípravy projektov bola vypracovaná štandardná projektová dokumentácia v zmysle výzvy. Súčasťou dokumentácie bolo aj definovanie prínosov, ktoré projekty spoločnosti prinesú. Tak ako uvádza projektová dokumentácia, dosah

realizácie oboch projektov bol vysoko pozitívny už v čase ich prípravy, na čo poukazuje aj analýza prínosov a nákladov (CBA analýza) projektov.

Medzi najvýznamnejšie benefity patria:

- Socioekonomické aspekty Big Data v štatistike:
  - zvyšovanie transparentnosti a adresnosti rozhodovania, najmä v oblasti investícií,
  - znižovanie prevádzkových nákladov dopravných spoločností s ohľadom na poznatky o pohybe obyvateľstva,
  - adresné rozhodovanie pri socioekonomických otázkach (napr. sociálne balíčky) v závislosti od správania populácie,
  - zvýšenie využitia potenciálu dátového trhu poskytnutím otvorených údajov.
- Dynamický cenový model:
  - znižovanie záťaže podnikateľov zaradených do procesu zisťovania spotrebiteľských cien,
  - znižovanie nákladov na spracovanie cenových štatistík na úrade,
  - zefektívnenie rozhodovania v oblasti investícií Slovenskej republiky,
  - spresnenie údajov o inflačnom vývoji Slovenskej republiky.

Počas realizácie oboch projektov sa neustále prehodnocujú benefity a objavujú nové potenciálne prínosy, ktoré sa ich realizáciou postupne overujú.

V súčasnosti prebieha realizácia projektov, ktorá iteratívne poukazuje na dôležitosť hľadania spôsobov, ako využívať Big Data v oblasti štatistiky.

Oba projekty sa dostali do realizačnej fázy, ktorá prebieha práve v súčasnosti. Na projektoch pracujú projektové tímy zložené z interných expertov, ako aj externých expertov, ktorých úlohou je práve tvorba dátových modelov, na ktorých sú oba projekty založené.

Vzhľadom na fakt, že ide do veľkej miery o experimentálne projekty, boli spresnené očakávané výstupy projektov, a to takto:

- Dynamický cenový model:
  - Dátový model na výpočet indexov položiek spotrebného koša, ktorý je založený na využití online transakčných údajov spoločnosti Heureka. Spolupráca bola aj medializovaná.
  - Potenciálom tohto modelu je nahradiť tradičné procesy získavania a spracovania údajov o cenách jednotlivých položiek spotrebného koša.
- Socioekonomické aspekty Big Data v štatistike – tento projekt rozvinul svoje zadanie na 3 základné modely: model pohybu obyvateľstva; model indexu sociálneho napätia a model výkonu ekonomiky na základe transakčných údajov z mýtneho systému:
  - Model pohybu obyvateľstva – cieľom tohto modelu je vytvoriť mapu dennej a nočnej populácie založenej na údajoch od poskytovateľov mobilných telekomunikačných služieb. Údaje budú spracovávané anonymizovane. Potenciál modelu do budúcnosti je predovšetkým v oblasti tvorby funkčných regiónov, resp. intervenčných opatrení, ktoré budú mať vyššiu adresnosť.

- Model indexu sociálneho napätia – tento model sa realizuje podľa vzoru Holandského štatistického úradu. Cieľom modelu je hľadať korelácie medzi významnými udalosťami, ktoré sa v spoločnosti udiali a zmenou nálad, ktorá bude vyhodnocovaná na základe algoritmov umelej inteligencie.
- Model výkonu ekonomiky – cieľom tohto modelu je na základe transakčných údajov z mýtneho systému identifikovať tzv. flash odhad ekonomických indikátorov krajiny. Potenciálom modelu je disponovať časovým predstihom informácie o budúcom ekonomickom vývoji krajiny, čo je potrebné na makroekonomické rozhodovanie a tvorbu politík.

5. ZÁVER

Realizáciou týchto projektov Štatistický úrad SR potvrdzuje, že záujem o využívanie Big Data v štatistických procesoch nie je len platonický, ale predstavuje reálnu a dôležitú súčasť budúcnosti štatistiky. Výstupy projektov budú prezentované po ukončení ich realizácie (cca september – október 2023), avšak už dnes sa ukazuje, že to je ten správny smer, ktorým sa Štatistický úrad SR dostáva viac do pozitívneho povedomia spoločnosti a využije potenciál, ktorý doba údajov prináša.

Skúmanie a experimentovanie v oblasti využívania takéhoto typu údajov je kontinuálny proces a jedným alebo dvoma projektmi nie je možné preskúmať všetky možnosti a otestovať ich v praxi. Týmito projektmi sa otvorilo veko truhlice pokladov, ktoré Big Data pre štatistiku prinášajú.

Už v súčasnosti sa hľadajú možnosti, ako pokračovať v začatých iniciatívach a prípadne ich rozvinúť o nové možností a trendy. Pevne veríme, že sa tento smer zachová a výstupy z projektov budú zavedené do praxe v rámci produkcie oficiálnych štatistík.

LITERATÚRA

[1] LAROCK, T.: „Data is the new oil“, But that also means it can be risky. [online]. [cit.20-02-2023] Dostupné na: https://www.dbta.com/Columns/Next-Gen-Data-Management/Data-is-the-New-Oil-But-That-Also-Means-it-Can-be-Risky-155275.aspx

[2] BRAAKSMA, B. – ZEELENBERG, Z.: Big data in official Statistics. [online], [cit. 30-05-2023, s. 4] Dostupné na: https://www.cbs.nl/-/media/_pdf/2020/04/dp-big-data-in-official-statistics.pdf

RESUMÉ

Štatistika je oblasť, pre ktorú sú údaje nesmierne dôležité a bez nich by nebolo možné realizovať štatistické výstupy. Projekty, ako sú dva dopytové projekty opísané v texte, sú nevyhnutné na overovanie možností využívania aj alternatívnych zdrojov údajov pre oficiálne štatistiky. Vzhľadom na dôležitosť a striktnosť štatistického procesu je relatívne dlhá cesta od identifikovania potenciálneho nového zdroja údajov do jeho zavedenia do štatistickej produkcie a procesu diseminácie. Experimentálne štatistické projekty preto musia byť súčasťou agendy úradu, aby bolo možné zefektívňovať, modernizovať a skvalitňovať výstupy hodnototvorného štatistického procesu. Preto je potrebné neustále hľadať zdroje na podporu takýchto iniciatív, ktoré z dlhodobého hľadiska prinesú rozvoj a pokrok do sféry štatistiky. Bez investícií nie je pokrok.

RESUME
Statistics is a field for which data is extremely important and without them it would be impossible to carry out statistical outputs. Projects, such as the two projects described in the text, are necessary to verify the possibilities of using alternative data sources for official statistics. Given the importance and stringency of the statistical process, there is a relatively long way from identifying a potential new source of data to introducing this source into the statistical production and dissemination process. Experimental statistical projects must therefore be part of the Office's agenda in order to enhance, modernize and improve the quality of the outputs of the value-creating statistical process. Therefore, it is necessary to constantly look for resources to support such initiatives, which in the long term will bring development and progress to the field of statistics. There is no progress without investment.

PROFESIJNÝ ŽIVOTOPIS
Peter Ďuriš *je absolventom Ekonomickej univerzity v Bratislave. Profesionálne začínal ako procesný analytik v konzultačnej spoločnosti Centire, kde mal na starosti okrem riadenia analytických tímov aj riadenie projektov. Od roku 2012 je konateľom spoločnosti Go SMART, s. r. o., v ktorej okrem iného pôsobí aj ako konzultant v oblasti využívania nových zdrojov údajov a metód v oficiálnej štatistike. Rovnako sa zaoberá prípravou projektov a ich realizáciou v štátnej, verejnej ale aj súkromnej sfére.*

KONTAKT
peter.duris@gosmart.consulting

Informácia/Information

12. KONFERENCIA: NOVÉ TECHNIKY A TECHNOLÓGIE V ŠTATISTIKE

12TH CONFERENCE: NEW TECHNIQUES AND TECHNOLOGIES FOR STATISTICS

V dňoch 7. – 9. marca 2023 sa v Bruseli v priestoroch budovy Charlemagne Európskej komisie, konal 12. ročník medzinárodnej vedeckej konferencie Nové techniky a technológie v štatistike (**New Techniques and Technologies for Statistics – NTTS**), ktorú organizoval Eurostat.

Zdroj: Eurostat

Konferencia sa pravidelne koná každé dva roky a je zameraná na nové metódy a technológie pre oficiálnu štatistiku a ich vplyv na zber údajov, jej tvorbu a šírenie. Podujatie sa vysiela prostredníctvom internetu. Účelom konferencie bolo umožniť prezentáciu výsledkov aktuálne prebiehajúcich výskumných projektov, podnietiť a uľahčiť prípravu nových inovatívnych projektov podporou spolupráce a výmeny názorov medzi štatistikmi z verejného, ale aj súkromného sektora a akademikmi s cieľom zvýšiť kvalitu a užitočnosť oficiálnej štatistiky. Zúčastnilo sa na nej viac ako 500 hostí a viac ako 150 prednášajúcich vrátane posterových prezentácií.

Konferenciu otvorila generálna riaditeľka Eurostatu Mariana Kotzeva, ktorá privítala účastníkov a vyjadrila potešenie, že tento ročník mohol byť organizovaný s osobnou účasťou, keďže predchádzajúci ročník sa konal len online formou z dôvodu pandémie COVID-19. Zdôraznila, že konferencia NTTS je jedna z popredných vedeckých štatistických konferencií vo svete a jedna z hlavných konferencií organizovaných Eurostatom. Ďalej uviedla, že bude prezentovaných množstvo tém z rôznych štatistických oblastí napr., inovácie v štatistike, umelá inteligencia v štatistike, revolúcia v analýze údajov, dátové ekosystémy a pod. V súčasnosti sa dostávajú do popredia nové typy technológií, ktoré slúžia na zber veľkého množstva údajov (napr. web scrapovanie) a ich spracovanie (napr. prostredníctvom strojového učenia). Vo vystúpení boli zdôraznené aj etické princípy využívania týchto nových zdrojov údajov pre oficiálnu štatistiku, ktorých rešpektovanie zabezpečí dôveryhodnosť štatistických inštitúcií v budúcnosti.

Súčasťou konferencie boli sprievodné aktivity, ktoré sa konali v dňoch 6. a 10. marca. Prvý deň bola prezentovaná platforma zberu údajov (**MOTUS**), ktorú možno využiť napr. na zisťovanie o rodinných účtoch. V záverečný deň bol predstavený projekt **ESSnet Web Intelligence Network**, ktorý skúma nové netradičné zdroje údajov s cieľom vytvoriť centrálnu databázu, ku ktorej by mali prístup členovia

Európskeho štatistického systému. Ďalšou aktivitou bola diskusia o získavaní údajov z pozorovania zeme, ako sú satelitné a letecké snímky, ktoré by sa mohli využiť v projektoch týkajúcich sa Európskej zelenej dohody (**European Green Deal**).

Počas celej konferencie prebiehali paralelné sekcie s rôznymi tematickými okruhmi.

1. deň konferencie (7. 3. 2023)

V prvom dni boli príspevky zamerané na témy výberov, nové zdroje údajov a ich integráciu, využívanie údajov vlastnených súkromným sektorom (privately-held data) a inovácie použitím umelej inteligencie v štatistike.

Témou hlavnej interaktívnej diskusie bolo využitie umelej inteligencie v štatistike. Hlavný prednášajúci Mark Grobelnik je výskumník v oblasti umelej inteligencie a riaditeľ laboratória umelej inteligencie v Inštitúte Jožefa Stefana v Ľubľane v Slovinsku. Hlavnou témou bolo predstavenie aplikácie **ChatGPT**, ktorá je verejne dostupným zdrojom umelej inteligencie a ukážka ako ju využiť v oficiálnych štatistikách. Na záver p. Grobelnik hovoril o nadchádzajúcej legislatíve týkajúcej sa využívania umelej inteligencie, ktorá sa pripravuje v Európskej únii.

V rámci sekcie web scrapovanie, Štatistický úrad SR prezentoval čiastkové výsledky výskumu týkajúceho sa využitia webscrapovaných údajov v cenových štatistikách, ktorý je súčasťou Európskeho projektu spolufinancovaného cez Európske investičné štrukturálne fondy. Prezentácia bola zameraná na aplikovanie rôznych typov cenových indexov v podmienkach SR a predostrela námety na ďalšiu odbornú diskusiu.

2. deň konferencie (8. 3. 2023)

V druhý deň konferencie sa väčšina tematických okruhov týkala modelov strojového učenia, spracovania veľkého množstva údajov, nowcastingu ekonomických ukazovateľov a využívania údajov od mobilných operátorov.

V úvode dňa vystúpil profesor Stefan Thurner s príspevkom, ktorý rozoberal komplexné systémy z pohľadu klasickej štatistickej teórie. Ide o sieťové dynamické systémy, ktorých jednotlivé komponenty majú nelineárnu štruktúru, podliehajú viacerým zdrojom náhodných stochastických javov a sú od seba vzájomne závislé. Ako príklad prof. Thurner uviedol javy vyskytujúce sa v prírodnom prostredí, napr. zosuvy pôdy, lesné požiare a škody spôsobené hurikánmi, a fenomény, ktoré je možné zaznamenávať v socioekonomických oblastiach, napr. distribúcia veľkosti miest vo svete, náklady na zdravotnú starostlivosť a systémové riziká v dodávateľských reťazcoch.

3. deň konferencie (9. 3. 2023)

V tretí deň konferencie sa paralelné prezentácie zamerali na podnikové a cenové štatistiky, analýzu časových radov a štatistickú analýzu geo-priestorových údajov.

Popoludní vystúpil bývalý generálny riaditeľ Eurostatu profesor Walter Radermacher s príspevkom na tému štatistika ako verejné dobro v ekosystéme digitálnej spoločnosti. Po vystúpení nasledovala panelová diskusia so zástupcami medzinárodných štatistických organizácií. V diskusii rezonovala myšlienka, že štatistická profesia sa stáva čoraz zodpovednejšou za výsledky, ktoré môže byť

vyvodené z publikovania a šírenia oficiálnych štatistík pre verejnosť. Globálne problémy a kríza, napr. klimatické zmeny a finančné krízy, si vyžadujú primerané formy poskytovania štatistických údajov a výsledkov založených na spoľahlivých dôkazoch a ešte užšiu spoluprácu medzi národnými štatistickými inštitúciami.

Hlavné závery konferencie možno zhrnúť do týchto bodov:

- Umelá inteligencia sa dostáva do pozornosti štatistikov, keďže dokáže automatizovať rutinné procesy, čoho výsledkom je optimalizácia pracovných postupov a ľudských zdrojov.
- Štatistické inštitúcie sa snažia doplniť tradičné zdroje údajov o údaje získavané web scrapovaním alebo o údaje súkromného sektora, aby tak zvýšili kvalitu a rozšírili portfólio oficiálnych štatistík. Pri rozširovaní zdrojov údajov je potrebná ich integrácia a prepájanie, ktoré by v ideálnom prípade mali byť zabezpečované automaticky.
- Pri vývoji a využívaní nových technológií je veľmi dôležitá spolupráca medzi súkromným a verejným sektorom.
- Štatistické inštitúcie čelia problémom, ktoré sa týkajú náboru kvalifikovaných ľudských zdrojov. V tejto oblasti je potrebná spolupráca s akademickými inštitúciami, ktoré by mali zabezpečiť výchovu dostatočného počtu absolventov pre trh práce v budúcnosti.

Na konferencii sa zúčastnili aj zástupcovia Štatistického úradu SR: generálna riaditeľka sekcie všeobecnej metodiky, registrov a koordinácie národného štatistického systému p. Helena Glaser-Opitzová, a zástupcovia tejto sekcie p. Petra Mazureková a p. Peter Knížat.

Zdroj: vlastná fotografia

Všetky príspevky a prezentácie (vo forme web prenosov), ako aj podrobný program konferencie je dostupný na stránke https://cros-legacy.ec.europa.eu/content/NTTS2023_en.

Ing. Helena GLASER-OPITZOVÁ
RNDr. Petra MAZUREKOVÁ, PhD.
Peter KNÍŽAT, MSc.

Autori pracujú v sekcii všeobecnej metodiky, registrov a koordinácie národného štatistického systému Štatistického úradu SR.

Informácia/Information

MEDZINÁRODNÝ WORKSHOP KRAJÍN VYŠEHRADSKEJ ŠTVORKY NA TÉMU MODERNIZÁCIA ŠTATISTIKY SPOTREBITEĽSKÝCH CIEN

INTERNATIONAL WORKSHOP OF THE VISEGRAD COUNTRIES ON MODERNIZATION OF THE CONSUMER PRICE STATISTICS

V dňoch 23. a 24. marca 2023 sa na pôde Štatistického úradu SR konal medzinárodný workshop na tému ***Modernizácia štatistiky spotrebiteľských cien***, na ktorom sa zúčastnili experti z Českého štatistického úradu, Štatistického úradu Poľska a Maďarského centrálneho štatistického úradu. Workshop sa konal na základe dodatku V k Memorandu o porozumení a spolupráci medzi štatistickými úradmi krajín Vyšehradskej skupiny. Bol zameraný na výmenu skúseností národných štatistických úradov krajín V4 v oblasti modernizácie cenových štatistík. Z vystúpení účastníkov jednoznačne vyplývala snaha o modernizáciu cenových štatistík minimálne v dvoch rovinách: cez zabezpečenie nových zdrojov údajov a aplikovaním nových metód ich štatistického spracovania.

Význam worshopu zdôraznil aj predseda Štatistického úradu SR pán Peter Peťko, ktorý po privítaní zúčastnených, otvoril podujatie.

Zdroj: vlastná fotografia

Prvou nosnou témou workshopu bola spolupráca s obchodnými reťazcami pri získavaní transakčných údajov o predaji, tzv. skener data a ich využitie na skvalitnenie výpočtu indexu spotrebiteľských cien. Ako uviedli v prezentáciách Petra Mazureková (*Spôsob využitia skener dát na Štatistickom úrade SR*) a kolegovia z Českého štatistického úradu pod vedením jeho podpredsedu Jaroslava Sixtu (*Implementácia skener dát do oficiálnej českej štatistiky*), využívanie transakčných údajov obchodných reťazcov je trend, ktorý sme zachytili aj v našom regióne. Využívanie týchto údajov umožní významnú úsporu osobných nákladov elimináciou ich zberu v teréne a súčasne významným spôsobom skvalitní údajovú základňu na výpočet indexu spotrebiteľských cien.

Jacek Białek zo Štatistického úradu Poľska vo svojej prezentácii *Spracovanie skenerových dát a výpočty cenového indexu* ukázal rôzne spôsoby výpočtu indexu spotrebiteľských cien vrátane zhodnotenia ich využiteľnosti na účely cenovej štatistiky. V príspevku riešil nielen využiteľnosť bilaterálnych, ale aj multilaterálnych metód výpočtu indexov vrátane ich interpretácie, pričom predstavil aj ním vytvorenú knižnicu indexov v programovacom prostredí R.

Druhou nosnou témou workshopu bolo využitie webscrapingu na účely cenových štatistík. Túto tému prvý deň uzavrela prezentácia Petra Knížata zo Štatistického úradu SR s názvom *Web scraping dát v indexoch spotrebiteľských cien*. Prezentácia Petra Knížata ukázala, že potenciál na modernizáciu štatistiky spotrebiteľských cien má aj využitie tzv. web scrapovaných dát. Prezentácia vychádzala z dát, ktoré boli scrapované z portálu cenového porovnávača Heureka. Peter Knížat odprezentoval aj obmedzenia tohto zdroja údajov, nakoľko obsahuje len predajné ceny bez údajov o predaných množstvách. Tieto nevýhody web scrapovaných údajov znamenali, že Štatistický úrad SR zahájil komunikáciu so spoločnosťami podnikajúcimi v oblasti e-commerce s cieľom využívať tzv. dátové feedy.

Spolupráca Štatistického úradu SR so spoločnosťou Heureka Group vyriešila nielen etický problém scrapovania dát z ich webových stránok, ale zabezpečila poskytovanie tzv. dátových feedov, ktoré obsahujú nielen údaje o predajných cenách, parametre predávaných tovarov zo širokej škály e-shopov, ktoré zastrešuje spoločnosť Heureka Group, ale aj údaje o predaných množstvách, čím tieto údaje už majú charakter tzv. on line transakčných dát. O ich využití na účely cenovej štatistiky hovoril vo svojom príspevku *Digitálne riadený zber dát z elektronického obchodu* na úvod druhého dňa workshopu Martin Faith. Tento príspevok vznikol v rámci projektu Dynamický cenový model, ktorý sa realizuje v Štatistickom úrade SR a ktorý je financovaný z Európskych štrukturálnych fondov.

Na tému webscrapingu nadviazal vo svojej prezentácii *Nové zdroje údajov (CPI/HICP) – úspechy grantového projektu 2020/2022* Péter Quittner z Maďarského centrálneho štatistického úradu. Projekt si stanovil podobné ciele ako prezentácie predchádzajúcich účastníkov: identifikovať možnosti nových zdrojov údajov a začať ich implementovať do výpočtu indexu spotrebiteľských cien. Zaujímavá bola široká škála zozbieraných a spracovaných údajov, pričom zaujali napr. výstupy z údajov o predaji ojazdených áut, údajov z čerpacích staníc a údajov o dovolenkovom ubytovaní. Bolo vidieť veľkú podobnosť medzi prístupmi jednotlivých prezentujúcich, ktorá spočívala aj vo využití umelej inteligencie pri klasifikácii produktov, ako aj podobné prístupy pri analýze a interpretácii získaných výstupov.

Posledný príspevok druhého dňa workshopu bol trochu iného charakteru, ale ukázal, že aj administratívne zdroje údajov majú ešte potenciál na modernizáciu cenových štatistík, čo potvrdil vo svojej prezentácii *Nový zdroj údajov v štatistike cien nehnuteľností* Martin Jankovič zo Štatistického úradu SR. Preberanie údajov z Úradu geodézie kartografie a katastra zreálnilo index cien nehnuteľností, pretože sa začalo pracovať nie s ponukovými cenami ale realizačnými, ktoré vytvárajú reálnejší obraz o dianí na realitnom trhu. Takisto zvýšený počet zaznamenaných transakcií zvýšil aj kvalitu poskytovaných výstupov.

Na záver workshopu zúčastnení konštatovali, že toto stretnutie a výmena skúseností naštartovali ďalšiu a intenzívnejšiu spoluprácu medzi štatistickými úradmi krajín V4. Ukázalo sa, že v mnohých prípadoch experti zúčastnených krajín riešia rovnaké problémy, čelia podobným výzvam, majú podobné ciele, a ako bolo povedané v závere stretnutia, je pred nami ešte dlhá cesta.

Ing. Peter MIŽIK

Autor pracuje v Štatistickom úrade SR ako riaditeľ odboru administratívnych zdrojov údajov.

PRIPRAVUJEME/COMING SOON

Boris VAŇO
OBYVATEĽSTVO SLOVENSKEJ REPUBLIKY PODĽA RODINNÉHO STAVU
POPULATION OF THE SLOVAK REPUBLIC BY MARITAL STATUS

Milan TEREK
CHARAKTERISTIKY VÝKONNOSTI PROCESU V METODOLÓGII SIX SIGMA
PROCESS PERFORMANCE MEASURES IN SIX SIGMA METHODOLOGY

Boris FRANKOVIČ
VÝPOČET RÝCHLYCH ODHADOV V ŠTATISTIKE CESTOVNÉHO RUCHU
CALCULATION OF FLASH ESTIMATES IN TOURISM STATISTICS

* * *

ONLINE VERZIA ČÍSLA 3/2023 SLOVENSKEJ ŠTATISTIKY A DEMOGRAFIE JE VEREJNE DOSTUPNÁ *na webových stránkach slovak.statistics.sk a ssad.statistics.sk od* ***15. JÚLA 2023****.*

THE ONLINE VERSION OF THE JOURNAL SLOVAK STATISTICS AND DEMOGRAPHY No 3 (2023) IS PUBLICLY BE AVAILABLE *at the websites slovak.statistics.sk and ssad.statistics.sk from* ***JULY 15, 2023.***

INFORMÁCIE PRE PRISPIEVATEĽOV

Príspevky prijímame v slovenskom, v českom a v anglickom jazyku. Musia rešpektovať odborné zameranie časopisu a jeho vedecký charakter. Zaslaný príspevok nesmie byť v recenznom konaní v inom časopise, ani uverejnený v odbornej a inej tlači.

Príspevky zasielajte v elektronickej forme vo formáte MS Word alebo Open Office, typ písma Arial, veľkosť 12, riadkovanie 1. Nad titulkom treba uviesť meno autora a jeho pracovisko.

Súčasťou príspevku je abstrakt (základný popis cieľa a spôsobu spracovania faktov v rozsahu do 100 slov), kľúčové slová (maximálne 5), resumé (stručné zhrnutie obsahu článku s dôrazom na jeho prínos a najvýznamnejšie závery v rozsahu do 500 slov), profesijný životopis (v rozsahu do 120 slov) a kontakt (e-mailová adresa autora). Názov článku, abstrakt, kľúčové slová a resumé poskytne autor aj v anglickom jazyku. Zoznam použitej literatúry v abecednom poradí s úplnými bibliografickými údajmi sa uvádza na konci článku. Odkazy na literatúru sa uvádzajú v texte číslami v hranatých zátvorkách. Poznámky s poradovým číslom sú umiestnené pod čiarou na príslušnej strane textu, ku ktorému sa vzťahujú. Podrobnejšie pokyny nájdu autori na ssad.statistics.sk.

Maximálny rozsah vedeckých článkov je 15 normostrán, informatívnych článkov 6 normostrán, recenzie, rozhovory a informácie publikujeme v rozsahu maximálne 3 normostrany. Tabuľky, mapy, grafy a obrázky musia mať názov a uvedený zdroj údajov; odporúčame, aby kopírovali šírku textu. Skratky sa používajú len minimálne, pri prvom použití je potrebné skratku v zátvorke rozpísať. Redakcia zabezpečuje jazykovú úpravu textu.

Príspevky sú recenzované. Oponentské konanie je obojstranne anonymné. Konečné rozhodnutie o publikovaní článku vydáva redakčná rada.

Redakcia si vyhradzuje právo zverejniť články schválené redakčnou radou v tlačenej a elektronickej podobe na ssad.statistics.sk.

INFORMATION FOR AUTHORS

Articles are accepted in Slovak, Czech and English languages and must comply with the journal´s professional specialisation and scientific nature as well. The submitted articles should not be reviewed by another journal and should not have already been published in any specialised or other press.

Please submit your articles in electronic form, in MS Word or Open Office format, Arial font, size 12 and typed in single spacing. The author´s name and workplace should be indicated above the title.

Articles should contain an abstract (general description of the objective and the processing methods used up to 100 words), key words (max. 5), resume (brief summary of the article´s content emphasizing its contribution and the most important conclusions up to 500 words), curriculum vitae of the author (no more than 120 words) and the author´s contact (e-mail address). The author should submit the article´s title, abstract, key words and resume in English language. List of the literature used with full bibliographic data should be given in alphabetical order at the end of an article. Bibliographic citations should be given in square brackets. References are indicated by numbers in a text in square brackets. Footnotes should be numbered in the order of the corresponding page of a text. Authors can find more details at the website ssad.statistics.sk.

Maximum scope of a scientific article is up to 15 standard pages, informative articles should be up to 6 standard pages in length, reviews, discussions and information not more than 3 standard pages. Tables, maps, graphs and pictures should have a title and the data source indicated, it is also advised to copy the width of a text. Abbreviations should be used only rarely and should be appropriately explained in parentheses when first used. Language text revisions are provided by the editorial office.

Articles are reviewed. The opponent procedure is mutually anonymous. The final decision on the article´s publication is made by the editorial board.

The editorial office reserves the right to publish articles approved by the editorial board in printed and electronic form at the website ssad.statistics.sk.

SLOVENSKÁ ŠTATISTIKA A DEMOGRAFIA

je jediný recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov. Propagujeme miesto a význam slovenskej štatistiky v Európskom štatistickom systéme, spoluprácu Eurostatu a národných štatistických úradov pri harmonizácii zisťovaní a multidimenzionálny rozmer štatistiky. Podporujeme rozvoj štatistickej teórie a jej prepojenie s praxou. Naším cieľom je prispievať k využiteľnosti štatistických výstupov v rôznych oblastiach a k zvyšovaniu ich kvality a efektivity.

Publikujeme analytické články, prognózy, názory, diskusné príspevky, recenzie, rozhovory, informácie a oznamy z rôznych oblastí štatistiky (národné účty, produkčné štatistiky, sociálne štatistiky, štatistika životného prostredia a pod.) a demografie (demografická štatistika, teoreticko-metodologické východiská demografie, historická demografia a pod.), vrátane sčítania obyvateľov, domov a bytov ako neodmysliteľnej súčasti demografickej štatistiky.

Vydáva:
Štatistický úrad SR

Identifikačné číslo vydavateľa:
IČO 00166197

Vychádza:
Štyrikrát ročne

Dátum vydania:
15. júl 2023

Tlač:
Reprografické stredisko
Štatistického úradu SR

Predplatné:
20 € (na rok)
5 € (za jeden výtlačok)

Objednávky prijíma:
Informačný servis
Štatistického úradu SR
Tel.: +4212/502 36 339
+4212/502 36 335
E-mail: info@statistics.sk

Evidenčné číslo/Evidence number 272/08

SLOVAK STATISTICS AND DEMOGRAPHY

is the only scientific reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures. Our aim is to promote the position and importance of Slovak statistics in the European Statistical System, cooperation between the Eurostat and the national statistical offices in the field of survey harmonisation and the multidimensional character of statistics as well. We support the development of statistical theory and its connection with practice. We aim to contribute to the utility of statistical outputs in various fields and to the improvement of quality and efficiency.

We publish analytic articles, prognoses, views, discussion contributions, reviews, discussions, information and announcements from various statistical fields (national accounts, production statistics, social statistics, environmental statistics etc.) and demography (demographic statistics, theoretical and methodological bases of demography, historical demography etc.) including the population and housing census as an essential part of demographic statistics.

Issued by:
Statistical Office of the SR

Company registration number:
00166197

Published:
Four times a year

Date of issue:
15th July 2023

Press:
Reprographic centre of the
Statistical Office of the SR

Subscription:
€20 (per year)
€5 (for one copy)

Orders are to be addressed to:
Information Service of the
Statistical Office of the SR
Tel.: +4212/502 36 339
+4212/502 36 335
E-mail: info@statistics.sk