SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

3/2023
ročník/volume 33

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

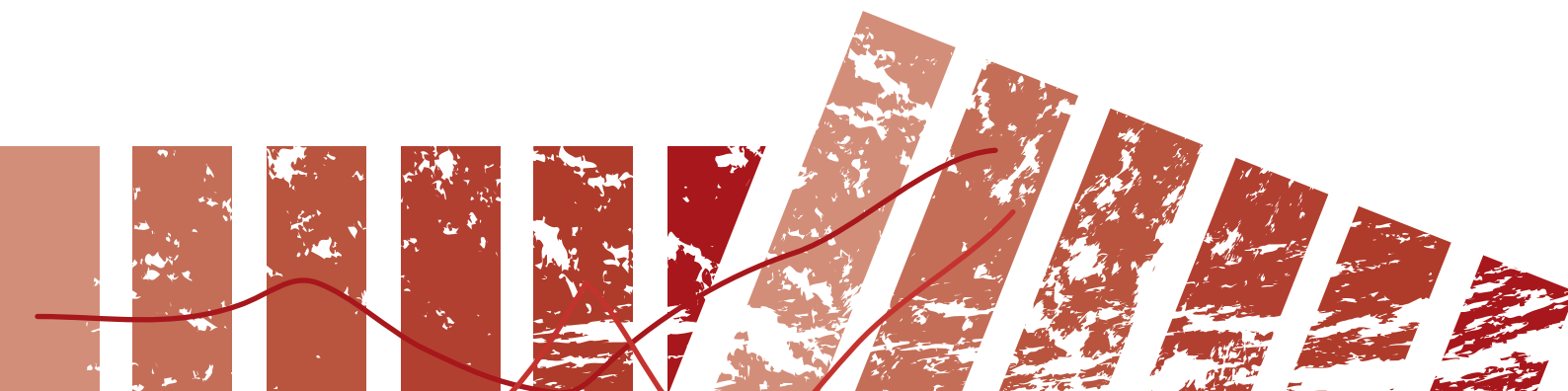
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 1

Typ článku/Type of article: editoriál/editorial

Strany/Pages: 3 – 6

Dátum vydania/Publication date: 15. júl 2023/July 15, 2023



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.