SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

1/2023
ročník/volume 33

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

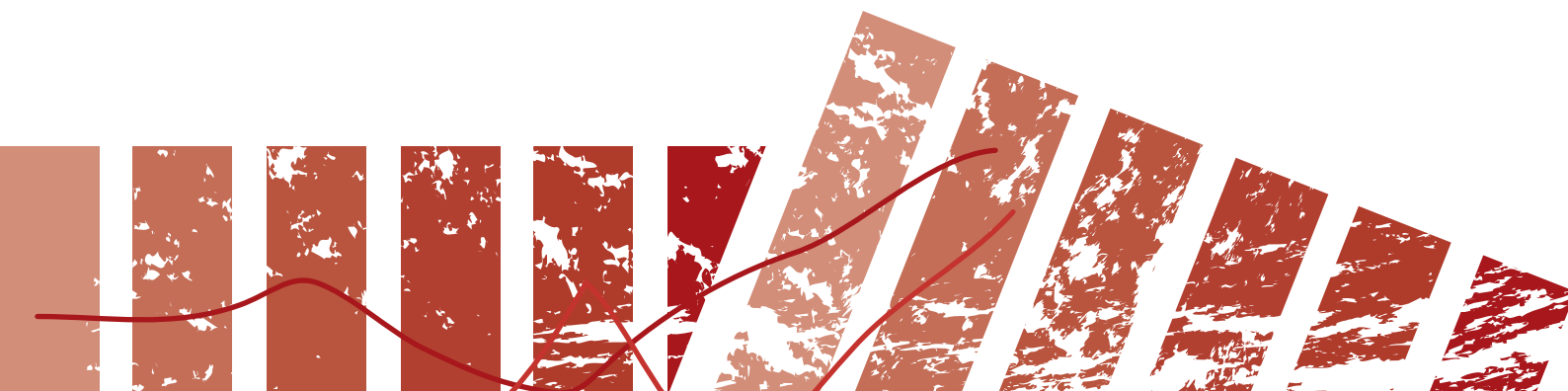
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 3

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 37 – 49

Dátum vydania/Publication date: 15. január 2023/January 15, 2023



Peter KNÍŽAT, Helena GLASER-OPITZOVÁ
Štatistický úrad Slovenskej republiky, Ekonomická univerzita v Bratislave

INDEX SPOTREBITEĽSKÝCH CIEN Z WEBSCRAPOVANÝCH ÚDAJOV: ANALÝZA VYBRANEJ PRODUKTOVEJ SKUPINY

CONSUMER PRICE INDEX FROM WEB-SCRAPED DATA: ANALYSIS OF SPECIFIC PRODUCT CATEGORY

ABSTRAKT
V dôsledku zmien spotrebiteľského správania sa spotrebiteľ orientuje viac na nákup cez internet. Štatistické inštitúcie zodpovedné za zber cien tovarov a služieb sú nútené prehodnotiť tradičný zber cien pre oblasť cenovej štatistiky a v niektorých prípadoch ho potenciálne nahradiť automatickým zberom cien cez internet, tzv. webscrapingom. Implementácia takéhoto zdroja údajov prináša so sebou rôzne výzvy, od otázok v metodickej oblasti až po významnú zmenu procesov spracovania údajov. Ide o spracovanie veľkého množstvo údajov, vrátane hodnotenia ich kvality, výberu reprezentantov a určenie cien jednotlivých tovarov, ktoré sú obvykle scrapované na dennej báze. Ďalšou výzvou je výber metódy na výpočet indexu spotrebiteľských cien, ktorá sa môže zásadne odlišovať od metódy výpočtu indexu spotrebiteľských cien použitej pri tradičnom zbere údajov. Cieľom tohto článku je predstaviť teoretický rámec na implementáciu webscrapovaných údajov do produkcie cenových štatistík. V prípadovej štúdii sme použili údaje o cenách pre produktovú skupinu chladničky, ktoré boli scrapované z webového porovnávacieho portálu https://www.heureka.sk/ .

ABSTRACT
As a consequence of changes in the consumer behaviour, a consumer prefers to shop online. Statistical institutions responsible for the collection of prices of goods and services for the area of price statistics are obligated to reconsider the traditional collection of prices and in some cases potentially replace it with automated collection of prices through internet, also called web-scraping. The implementation of this type of data sources entails various challenges, from questions in the methodological field to a significant changes of data processing. This involves the processing of big data including the evaluation of their quality, the selection of representatives and the determination of prices of individual goods, which are usually scraped on a daily basis. Another challenge is the selection of methodology for estimating the consumer price index (CPI) that can be fundamentally different from CPI estimation methodology used in the traditional data collection. The aim of this study is to present a theoretical framework for the implementation of web-scraped data in the production for price statistics. In the case study, we used data for the product category of refrigerators, scraped from the comparison website heureka.sk.

KĽÚČOVÉ SLOVÁ
webscrapované údaje, index spotrebiteľských cien, Jevonsov index, multilaterálne indexy

KEY WORDS
web-scrapped data, consumer price index, Jevons index, multilateral indexes

1. ÚVOD

Ceny tovarov, ktoré vstupujú do výpočtu indexu spotrebiteľských cien sa tradične zberajú v kamenných predajniach, vyhľadávaním na internetových stránkach alebo štatistickým zisťovaním. Tovary v spotrebnom koši sa každoročne aktualizujú na základe spotrebiteľského správania spoločnosti a vývoj ich cien sa sleduje a zaznamenáva mesačne. Index spotrebiteľských cien sa počíta na mesačnej báze.

V posledných rokoch dochádza k zmene spotrebiteľského správania. Spotrebiteľ sa viac orientuje na nákup cez internet, ktorý ponúka väčší sortiment niektorých tovarov, rýchle porovnanie cien a parametrov kvality jednotlivých tovarov a v niektorých prípadoch nižšiu cenu v porovnaní s cenami v kamenných predajniach.

V dôsledku tejto zmeny spotrebiteľského správania sú štatistické inštitúcie zodpovedné za zisťovanie cien tovarov a služieb a výpočet indexov spotrebiteľských cien nútené k prehodnoteniu a revízii tradičného zberu cien a jeho potenciálnemu nahradeniu automatickým zberom cien cez internet, tzv. webscrapovaním údajov.

Webscrapovanie údajov je sťahovanie informácií z internetu pomocou automatických robotov vytvorených v programovacích jazykoch, ktoré po naprogramovaní nepotrebujú žiaden ľudský zásah – intervencia je potrebná len pri nutnom preprogramovaní, napríklad pri zmene štruktúry webovej stránky, blokovaní robota majiteľom webovej stránky atď.

Štatistický úrad SR inicioval projekt zberu údajov webscrapovaním v spolupráci s Infostatom v roku 2020. Produkty, ktoré sa scrapujú, patria najmä do skupiny bielej a čiernej techniky a stiahnuté údaje obsahujú identifikátor tovaru, cenu a charakteristické vlastnosti jednotlivých tovarov. Sťahovanie prebieha každý tretí deň v nočných hodinách, aby sme neprimerane nezaťažovali webovú stránku https://www.heureka.sk/ .

Dôsledkom takéhoto spôsobu zberu údajov je potreba omnoho náročnejšieho spôsobu spracovania, čo zahŕňa analýzu kvality údajov, ich očistenie, výber tovarov do spotrebného koša a ich agregáciu na mesačnú úroveň predtým, než vôbec vstúpia do výpočtu indexu spotrebiteľských cien. Aj napriek z výšeným nárokom na IT infraštruktúru a tiež IT zručnosti zamestnancov môže mať podľa nášho názoru tento zdroj údajov potenciál na jeho využitie v praxi.

Autori vo vedeckom článku [11] testovali implementáciu webscrapingu údajov o cenách spotrebnej elektroniky a leteniek do výpočtu talianskeho harmonizovaného indexu spotrebiteľských cien. Ich štúdia sa zaoberá samotným scrapingovým nástrojom a IT infraštruktúrou potrebnou pre webscraping. Autori konštatujú, že scrapovanie údajov z webu má veľký potenciál nahradiť tradičné postupy zberu údajov, ale pre štatistické inštitúcie predstavuje veľkú výzvu z hľadiska spracovania veľkého množstva údajov (big data) pre oficiálnu štatistiku.

Ďalšou otázkou je výpočet samotného indexu spotrebiteľských cien. Vzorec na výpočet indexu spotrebiteľských cien sa môže zásadne odlišovať od vzorca, ktorý sa používa pri tradičnom zbere údajov, keďže pri webscrapingu je zozbierané veľké množstvo údajov pre všetky dostupné tovary v online predaji.

V online predaji pri niektorých druhoch tovarov dochádza k ich častej obmene. Napríklad, na trh vstupujú nové tovary a vystupujú z neho staré, v niektorých prípadoch ide o opätovný vstup toho istého tovaru na trh s viac alebo menej vylepšenými vlastnosťami (parametrami). Autori v článku [5] nazývajú tento jav zmenou kvality tovaru, ktorú je potrebné pri výpočte cenovej zmeny zohľadniť. Na meranie zmeny kvality medzi novým a starým tovarom sú vhodné a v akademickej literatúre odporúčané hedonické a tzv. time-dummy regresné modely. Pre podrobný popis a použitie pozri vedecká štúdia [11], alebo obsiahla príručka [16] pre hedonické regresie a úpravu kvality v cenových indexoch.

Ďalej môže dochádzať k dočasnej nedostupnosti cien niektorých tovarov v niektorých mesiacoch, ktoré by pri bilaterálnych indexoch spotrebiteľských cien založených na fixnom spotrebnom koši neboli zohľadnené vo výpočte. Fixný spotrebný kôš sa môže pri využití scrapovaných údajov nahradiť dynamickým spotrebným košom, keď výpočet indexu spotrebiteľských cien je medzimesačný s tzv. reťazením na bázické obdobie, čo však môže viesť k vychýleniu indexu, tzv. „driftu“ [5].

Autori vo vedeckej štúdii [10] vyriešili problém vychýlenia reťazeného indexu spotrebiteľských cien, ktoré nastáva pri mesačnom reťazení, rozšírením metódy GEKS ([8], [7] a [14]), ktorá sa toho času používala na porovnávanie priemernej zmeny cien medzi krajinami, na porovnávanie priemernej zmeny cien v čase. GEKS patrí do skupiny multilaterálnych indexov.

Pri použití multilaterálnych indexov, keď sa porovná priemerná zmena cien tovarov medzi viac ako dvomi časovými obdobiami súčasne, je možné zobrať do úvahy aj tovary, ktorých ceny v niektorých mesiacoch chýbajú. V uvedenej metóde GEKS je potrebná zhoda tovarov len pre každé dva porovnávané mesiace. Pri použití regresných typov indexov nie je potrebná žiadna zhoda tovarov medzi porovnávanými mesiacmi.

Vo vedeckej štúdii [5] autori vypočítali indexy spotrebiteľských cien pomocou metódy GEKS s hedonickými imputáciami a bez imputácií, t. j. tovary, ktorých ceny sa nenachádzali v oboch porovnávaných mesiacoch boli vylúčené. Taktiež vypočítali index spotrebiteľských cien pomocou regresnej metódy s časovou umelou premennou, tzv. time-product dummy (TPD) index spotrebiteľských cien. V prípadovej štúdii použili transakčné, tzv. scanner, údaje cien elektroniky predávanej na Novom Zélande, ktoré sa získavajú priamo od obchodných reťazcov, Tieto údaje obsahujú informácie o predaji, tržbách a predanom množstve jednotlivých tovarov, takže autori mohli vypočítať vážené indexy spotrebiteľských cien. V článku konštatovali, že najvhodnejšia metóda na výpočet indexu spotrebiteľských cien je GEKS s hedonickými imputáciami.

Autori v článku [6] použili na výpočet indexov spotrebiteľských cien údaje scrapované z internetu holandským štatistickým úradom. Článok konštatuje, že ceny tovarov získavané online by mohli čiastočne nahradiť ceny zberané tradičným spôsobom, keďže online zber cien je omnoho efektívnejším spôsobom zberu. Autori vypočítali indexy spotrebiteľských cien v dennej a týždennej periodicite pre 3 produktové kategórie: dámske tričká, mužské hodinky a kuchynské potreby. V článku sa konštatuje, že TPD index spotrebiteľských cien má praktickú výhodu, najmä ak je naším cieľom vypočítať vysokofrekvenčný index spotrebiteľských cien

využitím online údajov o cenách tovarov. Keďže štatistické inštitúcie počítajú najčastejšie indexy spotrebiteľských cien v mesačnej alebo štvrťročnej periodicite, táto prípadová štúdia je skôr vhodná na akademické účely.

Ďalej je nutné doplniť, že jeden z indexov spotrebiteľských cien, harmonizovaný index spotrebiteľských cien, sa zostavuje na základe európskeho nariadenia[1]. Eurostat postupne vydáva oficiálne usmernenia týkajúce sa implementácie nových zdrojov údajov do cenovej štatistiky[2]. Usmernenie týkajúce sa implementácie údajov získaných webscrapingom však v súčasnosti neexistuje.

V prípadovej štúdii sme analyzovali produktovú skupinu chladničky[3], ktorá zahŕňa kompletnú 5-miestnu kategóriu Európskej klasifikácie individuálnej spotreby podľa účelu (ďalej ako „ECOICOP“) [12], na ktorej úrovni sa vypočítava základný index spotrebiteľských cien. Spôsob výpočtu základného indexu spotrebiteľských cien nie je stanovený európskym nariadením a je zvolený na základe analýzy danej produktovej skupiny. Agregácia základných indexov spotrebiteľských cien do vyšších úrovní kategórií ECOICOP je už stanovená európskym nariadením, pozri poznámka č. 1 pod čiarou. Za produktovú skupinu chladničky máme k dispozícii scrapingom zozbieraný časový rad cien jednotlivých typov chladničiek dostupných v online predaji na webovom portáli https://www.heureka.sk/ od januára 2021 do januára 2022. Táto dĺžka časového radu cien už umožňuje aj výpočet multilaterálnych typov indexov spotrebiteľských cien.

Zámerom tohto článku je predstaviť teoretický rámec na agregovanie denných scrapovaných cien na mesačnú úroveň pre jednotlivé tovary, keďže Štatistický úrad SR počíta a publikuje index spotrebiteľských cien s mesačnou periodicitou. Ďalej predostrieme možnosti pre viacero spôsobov výpočtu indexu spotrebiteľských cien. Cieľom prípadovej štúdie je aplikovať tento teoretický rámec na pozorované webscrapované údaje a vyhodnotiť správanie rôznych typov indexov spotrebiteľských cien. Konečným cieľom nie je návrh najvhodnejšej metódy na výpočet indexu spotrebiteľských cien, ale prezentovanie súčasného stavu poznania v tejto oblasti v podmienkach Štatistického úradu SR.

Zlepšenie kvality výpočtu indexov spotrebiteľských cien implementáciou rôznych zdrojov údajov je kľúčovým cieľom viacerých európskych projektov [11]. Aj štúdia prezentovaná v tomto článku vznikla ako súčasť projektu dynamický cenový model (311071AA56), ktorý je spolufinancovaný Európskou úniou cez Európsky fond regionálneho rozvoja. Účelom realizácie projektu je otestovanie možností využitia iných, ako v súčasnosti využívaných údajov pre cenové štatistiky, ktoré realizuje Štatistický úrad SR. Ide hlavne o využitie moderných metód alternatívneho zberu údajov a porovnanie výsledkov týchto zisťovaní so štandardizovanými formami sledovania cien.

---

[1] Nariadenie Európskeho parlamentu a Rady (EÚ) 2016/792 z 11. mája 2016 o harmonizovaných indexoch spotrebiteľských cien a indexe cien nehnuteľností na bývanie a o zrušení nariadenia Rady (ES) č. 2494/95.

[2] Nedávno publikovaná príručka Eurostatu [2] na výpočet multilaterálnych indexov spotrebiteľských cien sa zaoberá spracovaním a použitím transakčných, tzv. scanner, údajov, ktoré sa získavajú priamo od maloobchodných predajcov a obsahujú informácie o predajoch jednotlivých tovarov. Transakčné údaje majú úplne inú štruktúru ako webscrapované údaje.

[3] Do produktovej skupiny chladničky sú tiež zahrnuté chladničky s mrazničkou a mrazničky.

2. TEORETICKÉ VÝCHODISKÁ

V tejto kapitole predstavíme teoretický rámec na počiatočnú analýzu webscrapovaných údajov a pre výpočet indexu spotrebiteľských cien.

V prvej časti predstavíme teoretický koncept, ktorý by mohol byť najvhodnejší na agregovanie denných cien na mesačné. V druhej časti predstavíme teoretický rámec na výpočet nevážených bilaterálnych a multilaterálnych indexov spotrebiteľských cien.

2.1 ANALÝZA WEBSCRAPOVANÝCH ÚDAJOV

Pri súčasnom tradičnom zbere údajov pre cenovú štatistiku sa ceny tovarov zisťujú jedenkrát mesačne. Tovary, ktoré boli vybrané do spotrebného koša sú fixne definované, revidované sú raz ročne a k zmene dochádza len pri výpadku tovaru z predaja u maloobchodného predajcu.

V online prostredí sa zber cien tovarov aplikuje na všetky dostupné tovary v predaji v daný deň. Webscrapovanie normálne prebieha na dennej báze, niekedy sa môže zúžiť na určitý počet dní v týždni. Jednotlivé tovary ponúkajú rôzni online predajcovia na tej istej webovej platforme za rozdielne ceny, takže spotrebiteľ má možnosť výberu na základe preferencie ceny alebo predajcu. Týmto sa akumuluje veľké množstvo údajov („big data“).

Z dôvodu volatility cien určíme dennú cenu tovaru ako geometrický priemer jednotlivých cien, za ktoré sa tovary ponúkajú.

Na popis použijeme nasledujúcu formuláciu. Máme denné ceny tovarov $\left[c_i^d\right]_p$, kde $i = 1, \ldots, N$ sú jednotlivé tovary ponúkané v daný deň $d = 0, \ldots, D$, jednotlivými online predajcami $p = 1, \ldots, P$. Prvý deň zberu definujeme ako bázické obdobie $d = 0$.

Geometrický priemer na výpočet dennej ceny jednotlivých tovarov je definovaný takto:

$$c_i^d = \prod_{p=1}^{P} \left(\left[c_i^d\right]_p\right)^{\frac{1}{P}} \tag{1}$$

Keďže index spotrebiteľských cien sa zostavuje na mesačnej báze, v ďalšom kroku musíme ceny jednotlivých tovarov agregovať na mesačnú úroveň. Podobne ako v predchádzajúcom kroku použijeme geometrický priemer denných cien jednotlivých tovarov ponúkaných v daný mesiac.

Geometrický priemer na výpočet mesačnej ceny c_i^m jednotlivých tovarov je definovaný takto:

$$c_i^m = \prod_{d=1}^{D} \left(c_i^d\right)^{\frac{1}{D}} \tag{2}$$

kde $m = 0, \ldots, M$ sú jednotlivé mesiace a $m = 0$ je bázické obdobie.

2.2 VÝPOČET INDEXU SPOTREBITEĽSKÝCH CIEN

V tejto časti predstavíme rôzne možnosti výpočtu indexov spotrebiteľských cien, ktoré sú vhodné pre webscrapované údaje. Keďže webscrapované údaje neobsahujú tržby a množstvo predaného tovaru za daný mesiac, môžeme použiť len vzorce na výpočet indexov, ktoré sú definované ako nevážené.

Jedným zo spôsobov ako sa rozhodnúť pre vhodný vzorec indexu je požadovať, aby spĺňal určité špecifické axiómy alebo testy. Tieto testy objasňujú vlastnosti indexov, ktoré nemusia byť na prvý pohľad zrejmé. Medzi štyri základné testy, ktorými môžeme ilustrovať axiomatický prístup patria [3]: test proporcionality, test súmernosti, časovo reverzný test a test tranzitivity. Jevonsov index na rozdiel od iných indexov spotrebiteľských cien spĺňa všetky uvedené testy a je teda z axiomatického hľadiska jednoznačne index s najlepšími vlastnosťami.

Jevonsov index možno tiež interpretovať ako geometrický priemer zmeny cien a je definovaný takto [3]:

$$I_{Jevons}^{0,m} = \prod_{i=1}^{N} \left(\frac{c_i^m}{c_i^0} \right)^{\frac{1}{N}} \tag{3}$$

kde *0* je bázické časové obdobie, ku ktorému porovnávame priemernú zmenu cien tovarov aktuálneho časového obdobia *m*. Spotrebný kôš tovarov je fixný a je definovaný v bázickom období *0*.

Ďalej pokračujeme predstavením multilaterálnych indexov spotrebiteľských cien v kontexte webscrapovaných údajov. Multilaterálne indexy spotrebiteľských cien porovnávajú priemernú zmenu cien medzi viac ako dvoma časovými obdobiami súčasne, v tzv. časovom okne *M*. Výhodou multilaterálnych indexov spotrebiteľských cien je, že dokážu zachytiť dynamiku predaja tovarov v časovom okne, napríklad vstupy a výstupy jednotlivých tovarov, chýbajúce ceny v niektorých mesiacoch. Ich nevýhodou je komplexnosť výpočtu a tým aj nejednoznačná interpretácia, ktorá môže byť podstatne rozdielna v závislosti od použitého typu multilaterálneho indexu.

Multilaterálna metóda GEKS používa bilaterálne indexy spotrebiteľských cien vypočítané medzi dvoma časovými obdobiami v definovanom časovom okne, ako tzv. základné elementy, ktoré vstupujú do výpočtu. Metóda GEKS vyžaduje výber vhodného bilaterálneho indexu spotrebiteľských cien. Pri použití Jevonsovho indexu, multilaterálny GEKS-Jevonsov index je definovaný nasledovne [2]:

$$I_{GEKS-Jevons}^{0,m} = \prod_{\substack{l=0 \\ l \neq m}}^{M} \left(I_{Jevons}^{0,l} \times I_{Jevons}^{l,m} \right)^{\frac{1}{M+1}} \tag{4}$$

kde *m = 0, …, M* sú časové obdobia.

Bilaterálne Jevonsove indexy sú vypočítané medzi časovými obdobiami *l* a *m* a spojené metódou GEKS, geometrickým priemerom, s *m = 0* ako bázickým časovým obdobím. Spotrebný kôš tovarov medzi porovnávanými časovými obdobiami sa musí zhodovať. Detaily výpočtu metódy GEKS sú popísané v [2]. Nevýhodou metódy GEKS

je, že chýbajúce ceny tovarov porovnávaných v jednotlivých časových obdobiach, je nutné imputovať, ak chceme, aby ich zmena bola reflektovaná vo výpočte indexu spotrebiteľských cien.

Priamej imputácii pri chýbajúcich cenách tovarov za niektoré časové obdobia je možné predísť použitím multilaterálnej metódy, ktorá je definovaná vo forme štatistickej regresie, tzv. modelu fixných efektov. Log-lineárny štatistický regresný model, tzv. time-product dummy (TPD) model, je definovaný takto [6]:

$$ln\ c_i^m = \partial^0 + \sum_{m=1}^{M} \partial^m F_i^m + \sum_{i=1}^{N-1} \gamma_i F_i + \varepsilon_i^t \qquad (5)$$

kde F_i je tzv. binárna premenná, ktorá nadobúda hodnotu *1* ak ide o cenu tovaru *i* a *0* v opačnom prípade, ε_i^t sú náhodné chyby regresie, ktoré majú normálne rozdelenie so strednou hodnotou *0* a konštantnou štandardnou odchýlkou *σ* a F_i^m je tzv. binárna časová premenná, ktorá nadobúda hodnotu *1* ak cena tovaru spadá do časového obdobia *m* a *0* v opačnom prípade. Regresné parametre $\partial^0, \partial^m, \gamma_i$ sa vypočítajú odhadom použitím metódy najmenších štvorcov.

Rovnica (5) patrí do tzv. regresných modelov s fixnými efektami, kde fixné efekty sú definované v parametri γ_i pre každú vzorku *i* v regresii, v našom prípade tovar. Podrobný popis týchto typov modelov a odhadov ich parametrov je uvedený v [1].

F_i^m sa vynechá pre *m = 0*, takže máme výpočtom odhadnuté parametre $\hat{\partial}^m$ za časové obdobia *m = 1, …,M*. TPD index spotrebiteľských cien sa vypočíta takto [5]:

$$I_{TPD}^{0,m} = exp\left(\hat{\partial}^m\right) = \frac{\prod_{i \epsilon S^m} (c_i^m)^{\frac{1}{N^m}}}{\prod_{i \epsilon S^0} \left(c_i^0\right)^{\frac{1}{N^0}}} exp\left[\sum_{i=1}^{N} \left(\bar{\hat{\gamma}}_i^0 - \bar{\hat{\gamma}}_i^m\right)\right] \qquad (6)$$

kde *m = 1, …,M*. sú časové obdobia a S^m a S^0 sú skupiny tovarov v časových obdobiach *0* a *m*. Ďalej $\bar{\hat{\gamma}}_i^0 = \sum_{i \epsilon S^0} \hat{\gamma}_i / N^0$ a $\bar{\hat{\gamma}}_i^m = \sum_{i \epsilon S^t} \hat{\gamma}_i / N^m$ sú priemery odhadovaných parametrov fixných efektov $\hat{\gamma}_i$ pre časové obdobie *0* a *m*. Ak sa skupiny tovarov medzi S^m a S^0 zhodujú, rovnica (6) sa zjednoduší na bilaterálny Jevonsov index definovaný v rovnici (3). Podrobné odvodenie rovnice (6) je uvedené vo vedeckej štúdii [6].

Ďalšou možnou regresnou metódou na výpočet indexu spotrebiteľských cien je tzv. hedonická regresia, pri ktorej sa použijú charakteristické vlastnosti jednotlivých tovarov ako nezávislé premenné vysvetľujúce logaritmus ceny daného tovaru. Hedonický regresný model definujeme takto [5]:

$$ln\ c_i^m = \partial^0 + \sum_{k=1}^{K} \beta_k Z_{ik} + \varepsilon_i^t \qquad (7)$$

kde Z_{ik} je matica charakteristických premenných súvisiacich s tovarom *i* a *k = 1, …, K* je počet charakteristických premenných s korešpondujúcim parametrami β_k. Ostatné premenné a parametre vrátane ich odhadov sú definované ako v rovnici (5).

Hedonická regresia v rovnici (7) má využitie aj na priame imputácie chýbajúcich cien tovarov. Ak cena daného tovaru chýba v aktuálnom mesiaci, ale bola pozorovaná v predchádzajúcom mesiaci, odhadnutý hedonický model sa môže použiť na predikciu ceny v aktuálnom mesiaci.

Pridaním binárnej časovej premennej F_i^m do rovnice (7) sa hedonický index spotrebiteľských cien odvodí podobne ako v prípade rovníc (5) a (6).

Nevýhodou multilaterálnych indexov spotrebiteľských cien je, že pri výpočte indexu za nadchádzajúce časové obdobie *M + 1* sa predchádzajúce indexy v časovom okne *M = 1, …, M + 1* znova prepočítajú, čím sa zmenia. Zmena indexov za predchádzajúce časové obdobia, v dôsledku prepočtu, je nežiaduca. Revízii predchádzajúcich indexov spotrebiteľských cien je možné predísť použitím tzv. splicingovej (spájacej) metódy, keď sa index spotrebiteľských cien vypočítaný za nové časové obdobie *M + 1* naviaže na predchádzajúce indexy spotrebiteľských cien. Keďže téma ohľadom aplikácie rôznych splicingových metód je veľmi rozsiahla, nie je súčasťou tohto článku.

3. VÝSLEDKY

V tejto kapitole aplikujeme uvedený teoretický rámec na webscrapované údaje za produktovú skupinu chladničky[4], kde typy chladničiek reprezentujú jednotlivé tovary v skupine, ktoré boli scrapované každý tretí deň za časové obdobie od 1. januára 2021 do 31. januára 2022 z porovnávacieho webového portálu https://www.heureka.sk/.

V nasledovnej tabuľke uvádzame sumárne informácie webscrapovaných údajov za uvedené časové obdobie.

Tabuľka č. 1: Sumárne informácie webscrapovaných údajov – produktová skupina chladničky

Produktová skupina	Počet cien tovarov	Počet tovarov (jednotlivých typov chladničiek)	Počet online predajcov	Počet charakteristických premenných
chladničky	1 468 279	3 332	121	10

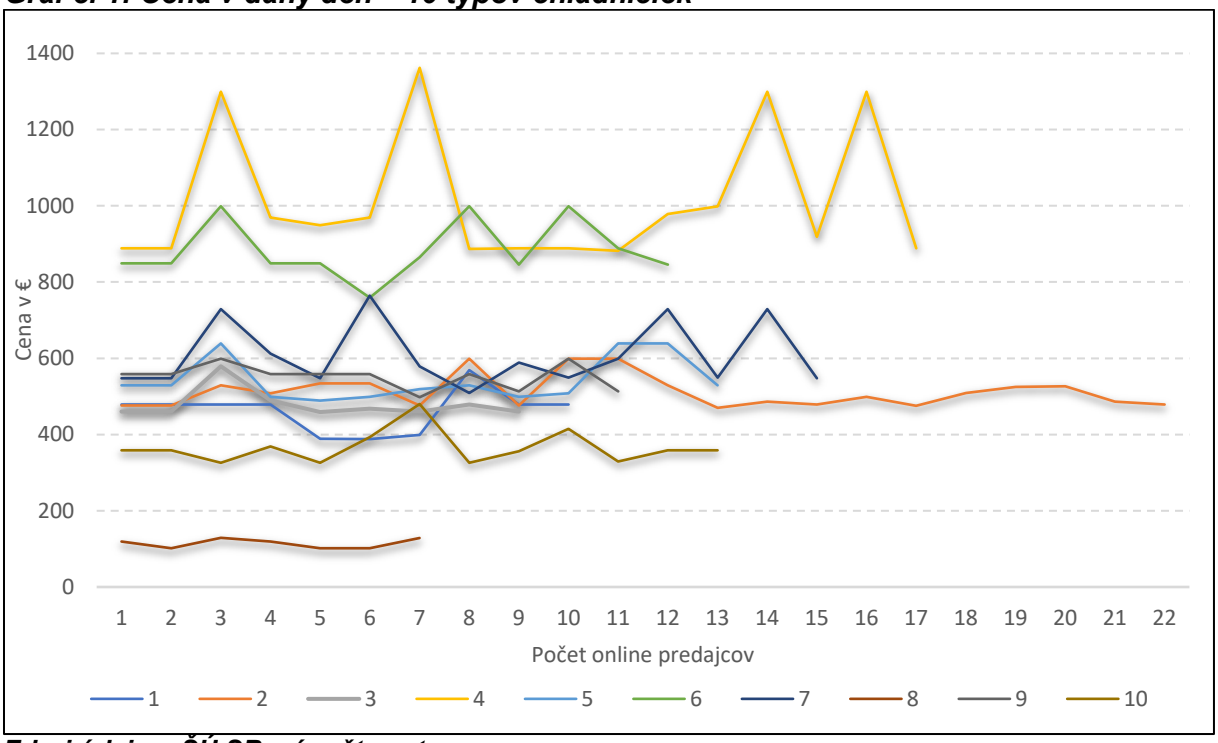
Zdroj údajov: ŠÚ SR, výpočty autorov

V tabuľke č. 1 vidíme, že webscraping umožnil za dané sledované obdobie na dennej báze získavať ceny pre 3 332 jednotlivých typov chladničiek, ktoré ponúkalo 121 online predajcov. Týmto bol výrazne rozšírený spotrebný kôš pre túto produktovú skupinu a na rozdiel od tradičného spôsobu zberu, pri ktorom je zaznamenávaná cena tovaru zvyčajne len raz mesačne, v údajoch z webscrapingu pozorujeme ceny tovarov s dennou alebo viacdennou periodicitou.

V nasledujúcom kroku aplikujeme agregovanie cien jednotlivých tovarov na mesačnú úroveň použitím rovníc (1) a (2). Na ukážku, že použitie rôznych priemerov pri agregovaní cien môže viesť k rozdielnym mesačným cenám vyberieme vzorku 10 tovarov, t. j. typov chladničiek, a zobrazíme ich ceny, ktoré v daný deň ponúkali rozdielni online predajcovia.

[4] *Do produktovej skupiny chladničky sú zahrnuté aj chladničky s mrazničkou a mrazničky.*

Graf č. 1: Cena v daný deň – 10 typov chladničiek

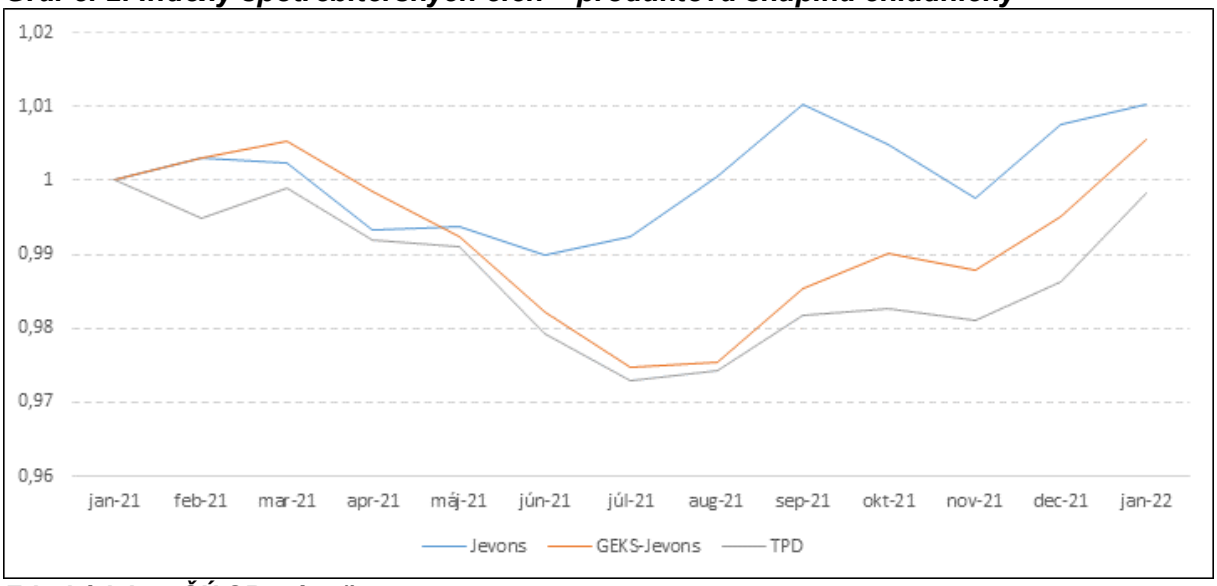


Zdroj údajov: ŠÚ SR, výpočty autorov

Graf č. 1 ukazuje, že cena tovaru od rôznych predajcov, v daný špecifický deň, môže byť podstatne rozdielna. Napríklad, tovar 2 ponúka 22 predajcov a tovar 4 má najväčší rozdiel v cenách v daný deň. V našej štúdii pri agregovaní cien používame geometrické priemery, rovnice (1) a (2), keďže tieto využijú všetky dostupné denné ceny vo výpočte a zredukujú efekt extrémnych hodnôt na výslednú dennú a mesačnú cenu.

V ďalšom kroku vypočítame Jevonsove, GEKS-Jevonsove a TPD indexy. Graf č. 2 zobrazuje vývoj týchto indexov spotrebiteľských cien za produktovú skupinu chladničky a za obdobie od januára 2021 do januára 2022, v ktorom január 2021 predstavuje bázické obdobie.

Graf č. 2: Indexy spotrebiteľských cien – produktová skupina chladničky



Zdroj údajov: ŠÚ SR, výpočty autorov

Graf č. 2 ilustruje, že multilaterálne indexy spotrebiteľských cien za produktovú skupinu chladničky majú podobný vývoj v čase na rozdiel od bilaterálneho indexu spotrebiteľských cien. Hodnota multilaterálneho indexu spotrebiteľských cien v danom období je ovplyvnená vývojom cien vo všetkých obdobiach definovaného časového okna na rozdiel od bilaterálneho indexu spotrebiteľských cien, keď sa porovnávajú len ceny dvoch časových období.

Výpočet hedonického indexu si vyžaduje dôkladnú analýzu charakteristických parametrov jednotlivých tovarov. Je potrebné tzv. čistenie údajov, t. j. odstránenie nesprávnych alebo doplnenie chýbajúcich údajov, štatistická analýza významnosti jednotlivých charakteristických premenných v súvislosti s cenou tovaru, korelačná analýza a ďalšie štatistické overenia správnosti výpočtu hedonickej regresie. Táto analýza bude súčasťou inej štúdie, ktorá sa bude zaoberať výpočtom hedonického indexu.

V rámci prezentovanej štúdie, autori dôkladne analyzovali webscrapované údaje cien v produktovej skupine chladničky. Štúdia bude slúžiť ako základ analýzy ďalších produktových skupín spotrebného koša. Multilaterálne metódy na výpočet indexov spotrebiteľských cien sú vo všeobecnosti vhodnejšie pre rýchloobrátkové tovary, keďže vo výpočte zahŕňajú viac časových období súčasne. V prípadovej štúdii, pri porovnaní s bilaterálnym indexom spotrebiteľských cien, ktorý sa štandardne používa, ak sa ceny tovarov získavajú tradičným spôsobom, hodnota oboch prezentovaných multilaterálnych indexov spotrebiteľských cien mala tendenciu byť pod jeho úrovňou. Rozhodnúť, ktorá z indexových metód je najvhodnejšia pre webscrapované údaje je náročná úloha, ktorá si vyžaduje množstvo ďalších analýz týkajúcich sa napríklad rozšírenia časového okna[5], spájania indexov tzv. splicingovými metódami a tiež analýzy ďalších produktových skupín spotrebného koša.

4. ZÁVER

Cieľom tohto článku bolo predstaviť teoretický rámec na implementáciu webscrapovaných údajov do produkcie cenových štatistík.

Prípadová štúdia prezentuje postup spracovania denných webscrapovaných údajov na mesačnú úroveň. Volatilita v denných cenách tovarov je významná, čo môže viesť k rozdielnym mesačným cenám, ak sa použijú rôzne typy priemerov pri agregovaní. V našej štúdii sme použili geometrický priemer, ktorý zabezpečí, že všetky ceny ponúkaných tovarov sú súčasťou výpočtu a vplyv extrémnych hodnôt je zredukovaný.

Porovnanie rôznych metód na výpočet indexov spotrebiteľských cien v prípadovej štúdii na produktovej skupine chladničky ukázalo, že výber metódy má podstatný vplyv na jeho vývoj v čase. Multilaterálne indexy spotrebiteľských cien zachytia dynamiku v zmene cien tovarov, keďže porovnávajú zmenu cien tovarov medzi viacerými časovými obdobiami súčasne a do výpočtu zahrnú ceny tovarov, ktoré sa nevyskytujú v každom časovom období. Ich nevýhodou je komplexnosť výpočtu a nejednoznačná interpretácia pre bežných používateľov.

Z prípadovej štúdie pre ďalšie pokračovanie projektu vyplynulo niekoľko dôležitých záverov. Napríklad, prvotný spôsob spracovania údajov a ich následné agregovanie

[5] V [2], strana 42, je navrhnuté zobrať do úvahy časové okno s 25 časovými obdobiami.

môže viesť k významne rozdielnym mesačným cenám jednotlivých tovarov. Taktiež nie je jednoznačné, na základe akých kritérií je možné vybrať najvhodnejšiu multilaterálnu metódu na výpočet indexu spotrebiteľských cien.

Projekt bude pokračovať dôkladnou analýzou údajov týkajúcich sa charakteristických parametrov jednotlivých tovarov, ktorá je potrebná na výpočet hedonického indexu. Hedonický regresný model bude v budúcnosti možné využiť aj pri priamej imputácii chýbajúcich cien tovarov v niektorých mesiacoch v prípade multilaterálnej metódy GEKS.

Ďalšou dôležitou problematikou skúmania je analýza spôsobov viazania indexov za nové časové obdobie T+1, tzv. splicingovej metódy.

LITERATÚRA

[1] ALLISON, P. D.: Fixed Effects Regression Methods for Longitudinal Data Using SAS, 2009, SAS Institute Inc., Cary, NC, USA. ISBN 978-1-59047-568-3.
[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] ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004). Consumer Price Index Manual: Theory and Practice. ILO Publications, Geneva.
[4] DE HAAN J.: Hedonic Prices Indexes: A Comparison of Imputation, Time Dummy and 'Re-Pricing' Methods, Jahrbücher f. Nationalökonomie u. Statistik, Lucius & Lucius, Stuttgart, 2010, Bd. (Vol.) 230/6, 772-791.
[5] DE HAAN J. – KRSINICH F.: Scanner Data and the Treatment of Quality Change in Rolling Year GEKS Price Indexes, 2012, Paper presented at the eleventh Economic Measurement Group Workshop, 21-23 November 2012, Sydney, Australia
[6] DE HAAN J. – HENDRIKS R.: Online data, fixed effects the construction of high-frequency price indexes, Paper presented at the Economic Measurement Group Workshop, 2013, 28-29 November 2013, Sydney, Australia.
[7] ELTETÖ, O. – KÖVES, P.: On a Problem of Index Number Computation Relating to International Comparisons, 1964, Statisztikai Szemle 42, s. 507 – 518 (originál v maďarčine).
[8] GINI, C.: On the Circular Test of Index Numbers, 1931, Metron 9, s. 3 – 24.
[9] 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.
[10] IVANCIC L. – DIEWERT W. E. – FOX K. J.: Scanner Data, Time Aggregation and the Construction of Price Indexes, 2011, In: Journal of Econometrics 161, s. 24 – 35.
[11] POLIDORO F. – GIONNINI R. – LO CONTE R. – MOSCA S. – ROSSETTI F.: Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation, 2015, In: Statistical Journal of the IAOS 31 (2015) 165-176, IOS Press.
[12] RAMON - Reference and Management of Nomenclatures: Europa - RAMON - Classification Detail List [cit. 2022-12-05].
[13] SILVER, M. – HERAVI, S.: The Measurement of Quality-Adjusted Price Changes. 2003, Pp. 277 – 317. In: Shapiro, M. – Feenstra, R. (eds.): Scanner Data and Price Indexes, NBER, Studies in Income and Wealth, vol. 61. Chicago: University of Chicago Press.

[14] SZULC, B.: Indices for Multiregional Comparisons, 1964, In: Przeglad Statystycny 3, s. 239 – 254 (originál v poľštine).
[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. 2006, Organization for Economic Co-operation and Development, Paris.

RESUMÉ
V súčasnosti sa ceny tovarov zberajú tradičným spôsobom, t. j. v kamenných predajniach, vyhľadávaním na internetových stránkach, alebo štatistickým zisťovaním. Webscraping je zber cien tovarov automatickým spôsobom, t. j. sťahovaním údajov z internetu, pomocou tzv. robota, ktorý je vyvinutý v niektorom z programovacích jazykov. Webscrapované údaje umožnia porovnať vývoj cien v časovom období pre všetky dostupné tovary v online predaji. Príspevok je zameraný na analýzu a spracovanie webscrapovaných údajov, ktoré sa získavajú v dennej periodicite. Štatistické inštitúcie zostavujú indexy spotrebiteľských cien s mesačnou periodicitou takže je potrebná agregácia denných webscrapovaných cien na mesačné. Tento krok je veľmi dôležitý, keďže použitím rôznych typov priemerov môže dôjsť k určeniu rozdielnych mesačných cien tovarov. Ďalej sa článok zameriava na prezentovanie vybraných metód na výpočet indexov spotrebiteľských cien. Webscrapované údaje neobsahujú informácie o predaji jednotlivých tovarov, ktoré by zahŕňali ich predané množstvá a tržby v určitom časovom období, čo umožňuje aplikovať len nevážené typy indexov. Prípadová štúdia prezentuje aplikáciu vybraných bilaterálnych a multilaterálnych metód na výpočet indexov spotrebiteľských cien v produktovej skupine chladničky. Multilaterálne indexy spotrebiteľských cien na rozdiel od bilaterálnych dokážu zachytiť dynamiku predaja jednotlivých tovarov tým, že do výpočtu zahrnú viac ako dve časové obdobia súčasne do výpočtu, čo je dôležité pre rýchloobrátkové tovary.
Prípadová štúdia, ktorá bola vypracovaná autormi tohto článku je súčasťou projektu. Dynamický cenový model. Tento projekt sa zaoberá modernizáciou cenových štatistík z pohľadu zdrojov v Štatistickom úrade SR. Webscrapovanie je technologicky náročný zber informácií priamo z internetu, ktorý si vyžaduje kvalifikované ľudské zdroje a pokročilé informačno-technologické systémy. Štatistický úrad SR momentálne uskutočňuje teoretickú a empirickú štúdiu použitia webscrapovaných údajov v cenových štatistikách.

RESUME
Currently, the prices of goods are collected in a traditional way, i.e. in the brick and mortar stores, by searching on the internet, or through statistical surveys. Web-scraping is a collection of prices of goods in an automated way, i.e. by downloading data from the internet using the so-called robot that is developed in a one of the programming languages. The web-scraped data enables a comparison of price changes over time periods for all available goods in online sales. The paper focuses on the analysis and processing of web-scraped data, which are obtained on a daily basis. Statistical institutes compile consumer price indexes on a monthly basis so the aggregation of daily web-scraped prices is necessary on a monthly level. This step is very important, since using different types of averages can result in a determination of different monthly prices of goods. Furthermore, the paper focuses on a presentation of different methods for estimating the consumer price index. The web-scraped data does not contain information on the sale of individual goods, which would include their

sold quantities and turnover in a given time period, which enables to apply only unweighted consumer price indexes. The case study presents the application of selected bilateral and multilateral methods for the calculation of consumer price indexes for the product category of refrigerators. Multilateral consumer price indexes, unlike bilateral ones are able to capture the sales dynamics of individual products by including more than two time periods simultaneously in its calculation, which is important for fast-moving goods.
This study which was developed by the authors of this paper is part of the dynamic pricing model project. This project deals with the modernization of price statistics from the perspective of different data sources in the Statistical Office of the Slovak Republic. Web-scraping is a technologically demanding collection of information directly from the internet, requiring highly qualified human resources and advanced information technology systems. The Statistical Office of the Slovak Republic is currently conducting a theoretical and empirical study of the use of web-scraped data in price statistics.

PROFESIJNÝ ŽIVOTOPIS
Peter Knížat, MSc, *je externým študentom doktorandského štúdia na Fakulte hospodárskej informatiky Ekonomickej univerzity v Bratislave. Pracuje ako dátový analytik v sekcii všeobecnej metodiky, registrov a koordinácie národného štatistického systému Štatistického úradu SR, kde je zodpovedný za návrh štatistickej metodiky v cenových štatistikách s využitím webscrapovaných údajov.*
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 riadi interný projekt úradu zameraný na modernizáciu cenových štatistík.*

KONTAKT
peter.knizat@statistics.sk
helena.glaser-opitzova@statistics.sk