ŠTATISTICKÝ ÚRAD SLOVENSKEJ REPUBLIKY

Dynamický cenový model

Last update:07.05.2025

Úvod

Štatistický úrad v súčasnosti využíva pre tvorbu cenových štatistík vo veľkej miere údaje o spotrebiteľských cenách zisťované priamo v teréne – vo vybranej sieti predajní a prevádzok služieb. Tieto údaje sú doplnené o údaje získané z pokladničných systémov obchodných reťazcov (transakčné dáta - scanner data).

S rozvojom informačných systémov a výraznej zmeny spotrebiteľského správania, a to najmä presunu časti realizovaných nákupov zo strany jednotlivcov a domácností do digitálneho priestoru (e-commerce), vzniká v tomto priestore množstvo údajov, ktoré by mohli byť využité ako nové dátové zdroje s cieľom presnejšie reflektovať spotrebiteľské správanie vo výpočte cenového indexu.

Štatistický úrad SR sa v nadväznosti na tieto skutočnosti, ako aj s ohľadom na vývoj nových alternatívnych metód výpočtu cenových indexov v iných krajinách rozhodol overiť možnosti využitia iných, ako v súčasnosti využívaných údajov pre cenové štatistiky. Pre tento účel realizuje projekt **Dynamickýcenový model,** ktorý skúma alternatívne možnosti formou výskumnej a experimentálnej štatistiky.

Cieľ projektu

Cieľom projektu bolo overenie možností využitia iných, ako v súčasnosti využívaných údajov pre cenové štatistiky, ktoré Štatistický úrad SR realizuje. Jedná sa o **využitie moderných metód alternatívnehozberu údajov o cenách** a porovnávanie výsledkov týchto zisťovaní s tradičnými formami sledovania cien.

Prínosom pre cenovú štatistiku je najmä veľké množstvo dát dostupných v digitálnom priestore, ktoré v súčasnosti nie sú obsiahnuté vo výpočte cenových indexov.

Počas projektu sa testovali a overovali nasledovné alternatívne a inovatívne metódy zberu dát:

- webscraping
- dátové feedovanie
- API

V rámci procesu následného spracovania zozbieraných dát bol vytvorený informačný systém **Dynamický cenový model (IS DCM)**, ktorý vyhodnocuje vývoj spotrebiteľských cien vybraných produktov a produktových skupín na základe dát, zbieraných v prostredí ecommerce.

Cieľom vytvorenia IS DCM je zároveň aj snaha využívať údaje pre prípravu dátových analýz,

ktoré slúžia ako podklad pre lepšie rozhodovanie a aplikovanie presnejších predikcií a modelov, ako aj aplikovať vhodné riešenie a postupy pre maximálne využitie dát v definovanej problémovej oblasti a overiť definované spôsoby založené na dátovej vede a analytických prístupoch priamo v rozhodovaní v oblasti cenovej štatistiky.

V neposlednom rade, realizácia projektu umožňuje zabezpečiť a posilniť možnosť experimentovať a zlepšovať rozsah analýz a šíriť skúsenosti v rámci národného a Európskeho štatistického systému a vytvára dynamické nástroje pre tvorbu cenových štatistík.

Prístup k modelovaniu

Dynamický cenový model využíva dáta zozbierané modernými metódami zberu z prostredia e – commerce. Ide najmä o metódu webscrapingu, feedovania dát a API.

Webscraping umožňuje automatizovaný zber veľkého množstva údajov pri vyššej frekvencii oproti štandardným metódam zberu údajov. Táto metóda zavádza automatizované procesy implementované pomocou robota alebo webového prehľadávača.

Charakteristickým faktorom scrapovania je častá zmena produktov (produkty sú nahradzované inými so zmenenými parametrami), ako aj potenciálna zmena štruktúry sledovaného portálu, na ktorú je potrebné promptne reagovať úpravou nástroja na scrapovanie.

Feedovanie dát je štandardným spôsobom zdieľania dát medzi dvoma protistranami v online prostredí. Služby fungujúce online tento spôsob využívajú na výmenu dát a realizáciu spoločných projektov a funkcionalít.

Pre účely štatistického zisťovania existujú určité limity (štruktúra dát), avšak z dlhodobého hľadiska je možné iniciovať komunikáciu smerom na poskytovateľov služby, ako aj vlastníkov dát, aby pripravili dátové feedy v štruktúre a kvalite reflektujúcej požiadavky Štatistickým úradom SR.

Z pohľadu typu získavaných dát je zber dát pomocou **API** využiteľný pre zber cenových dát, ako aj zber datasetov obsahujúcich produktové parametre.

Prístup k API (Application Programming Interface) konkrétneho zdroja údajov je štandardne realizovaný prihlásením sa na adresu API pomocou unikátneho kľúča užívateľa alebo Oauth pri partnerskom API. Pri otvorenom API postačí HTTP request bez identifikácie užívateľa. Tieto requesty je možné realizovať pomocou ľubovoľného nástroja, ktorý umožňuje zasielať takéto žiadosti. Ako príklad je možné uviesť jazyk Python, rozhranie Azure Data Factory z prostredia Microsoft Azure alebo dokonca aj novšie verzie známeho software Microsoft Excel.

Ako DCM funguje?

Dynamický cenový model (DCM) predstavuje technologické riešenie, ktoré je možné rozdeliť na tri na seba nadväzujúce časti – zber dát, ich čistenie a spracovanie a výpočet cenových indexov.

Fáza prvá – zber dát

Úvodným krokom systému DCM je zber dát.

Pre vybrané produktové skupiny boli namapované identifikované dátové zdroje. Zvážené boli dátové zdroje niekoľkých typov:

- Weby predajcov e-commerce – zamerané zväčša na jednu produktovú skupinu, ako napríklad web zalando.sk s odevmi a obuvou alebo hornbach.sk s materiálmi na údržbu. Preferované sú webové zdroje, ktoré majú vysoký počet produktov, ako aj vysoký počet dostupných produktových parametrov.
- Agregátory cenových ponúk – analyzované agregátory, ktoré sprostredkovávajú ponuky predajcov relevantných pre zvolené produktové skupiny. Záujem bol o favi.sk, glami.sk a heureka.sk, ktorý sa neskôr stal kľúčovým partnerom projektu
- Marketingové systémy vyhľadávačov – prezentujú ponuky predajcov v rámci webového vyhľadávania. Relevantné na území Slovenska je bing.com, zoznam.sk a google.com.

Vyšší počet produktov zabezpečí širšie pokrytie trhu danej produktovej skupiny, čo je v kontexte riešenia žiadúce. Produktové parametre umožnia automatizované spájanie produktov do menších celkov, na základe ktorých sa budú počítať cenové štatistiky.

Ideálnym stavom je možnosť získavať nielen dáta o cenách, ale aj dáta o predaných množstvách jednotlivých produktov, nakoľko predané množstvá umožňujú aplikáciu vážených štatistických metód ako pri tvorbe produktových skupín, tak aj pri výpočte cenových indexov. Táto možnosť je dostupná práve cez portál heureka.sk, ktorý zbiera anonymizované dáta o predajoch predajcov, ktorí na webe heureka.sk prezentujú svoje produkty. Spolupráca so spoločnosťou Heureka umožnila získavať na dennej úrovni dáta o predajoch a realizovaných cenách. K jednotlivým produktom sú získavané aj produktové parametre vo veľkom rozsahu, niektoré produkty mali asociovaných viac ako 90 parametrov. Dáta sú získavané formou dátového feedovania, na strane Heureky sú feedy vo formáte parquet generované systémom Keboola a automatizovane zasielané do prostredia projektu DCM.

Okrem feedov z Heureky získava DCM dáta aj z reklamného systému Google Ads napojením na externú službu API, ako aj z niektorých vybraných webov, ktoré sú webscrapované na týždennej báze. Dáta z Google Ads, ako aj zo scrapingu neobsahujú predané množstvá, vážené metódy tak v ich prípade nie je možné použiť.

Fáza druhá – čistenie dát a tvorba produktových skupín

Získané dáta vstupujú do dátového modelu DCM. Ten má za cieľ automatizovanou formou získané dáta očistiť a pripraviť na proces tvorby produktových skupín. Čistenie prebieha napríklad nad produktovými parametrami – pokiaľ získané produktové parametre neobsahujú vo veľkej miere hodnoty, nevstupujú do ďalšieho spracovania.

Produktové skupiny sú tvorené zo samostatných produktov a slúžia ako vstup pre výpočet cenových indexov. Jednotlivú skupinu je možné si predstaviť ako súbor 5 až 20 produktov, ktoré

sú si navzájom na základe ich parametrov veľmi podobné a ich cena v čase sa pohybuje po veľmi podobnej trajektórii. Pre účely výpočtu cenových indexov sa definujú a zafixujú produktové skupiny a následne sa na základe hodnôt produktových skupín v čase počítajú samotné cenové indexy. Hodnota skupiny je veľmi dôležitá, v prípade dostupných predajných množstiev je počítaná ako jednotková cena (obrat / predané množstvo) produktov v skupine za daný mesiac, ak predané množstvá dostupné nie sú, tak je počítaná ako aritmetický priemer cien produktov v skupine za daný mesiac.

Pri rýchlo získavaných dátach vo veľkých množstvách z digitálneho prostredia je zrejmé, že proces tvorby skupín a nápočet ich hodnôt musí byť plne automatizovaný. Túto automatizáciu riešenie DCM poskytuje. Pre tvorbu skupín produktov (tzv. clustrov) je dostupných niekoľko metód ako napríklad K-means, Mean Shift Kernel Clustering alebo hierarchické zhlukovanie. Celá logika systému je naprogramovaná v jazyku Python, niektoré štatistické metódy využívajú dostupné knižnice, predovšetkým knižnicu sklearn.

Fáza tretia – výpočet cenových indexov

Finálny krok systému DCM je výpočet cenových indexov nad pripravenými produktovými skupinami a ich mesačnými hodnotami. Dostupné sú vážené cenové indexy, ako napríklad bilaterálne indexy Tornquist alebo Fischer, ako aj multilaterálne indexy GEKS alebo Geary-Khamis. Tieto sú aplikovateľné iba na produktové skupiny, ktoré obsahujú predajné kvantity. Pre skupiny bez predajných kvantít sú dostupné neväzené cenové indexy ako Jevons, Carli alebo Dudot.

Cenové indexy sú počítané na úroveň produktových skupín definovaných EUROSTATom. Jedná sa o úroveň ECOICOP 5. Výpočet je dostupný jednak pomocou algoritmu vyvinutého v rámci DCM v Pythone, ako aj pomocou knižnice PriceIndices v jazyku R. Vypočítané indexy sú dostupné pre integráciu s čiastkovými indexami ostatných produktových skupín. Riešenie DCM aktuálne končí výpočtom cenových indexov na úrovni ECOICOP 5.

Zhrnutie kľúčových zistení

Alternatívne spôsoby zberu údajov o spotrebiteľských cenách umožňujú významne **vyššiufrekvenciu zberu dát** a **vyšší objem dát** oproti štandardným metódam zberu údajov, z ároveň sú získavané podrobné údaje o funkčných a kvalitatívnych parametroch, ktoré ovplyvňujú cenu produktu.

Významne sa tak zvyšuje počet údajov vstupujúcich do cenových štatistík, čím sa zvyšuje kvalita a presnosť výpočtu cenových indexov. Poskytované dáta sú sprístupnené okamžite, resp. na dennej báze, ide teda o **aktuálnedáta.**

Metódy zberu – zhodnotenie výsledkov

Z testovaných alternatívnych metód zberu údajov sa javí dátové feedovanie ako najvhodnejšia metóda zberu dát pre využitie v oblasti cenových štatistík. Jej ďalšie využitie bude závisieť od dohody s vlastníkmi, resp. poskytovateľmi dát.

V rámci projektu bola nadviazaná spolupráca so spoločnosťou Heureka, ktorá prostredníctvom dátových feedov vedela okrem cenových a produktových dát poskytnúť aj údaje o predaných kusoch (za vybrané druhy tovaru), t.j. transakčné dáta, ktoré významným spôsobom prispievajú k presnosti dát.

Pre ďalšie využitie webscrapingu pre potreby cenových štatistík bude potrebné ďalšie testovanie a monitorovanie zbieraných dát, ich kvalita a stabilita, ako aj stály príjem produktových a cenových dát zo zvolených zdrojov. Taktiež bude musieť byť zabezpečená údržba a kontrola jednotlivých webscraperov externým poskytovateľom.

Služba API bola v rámci projektu DCM skôr doplnková, keďže ide o platenú službu. Jej ďalšie využitie, resp. využitie iných API služieb od ďalších poskytovateľov bude závisieť jednak od ponuky a možností vhodných dátových zdrojov (vyhovujúca štruktúra dát a existencia koncových bodov), ako aj od nákladov spojených s touto službou.

Ďalšie kroky ŠÚ SR

- *Pred prijatím rozhodnutia o zaradení dát zbieraných alternatívnymimetódami do výpočtu publikovaných cenových indexov bude potrebné stanoviť minimálne časové obdobie, počas ktorého bude možnékvalitu dát a spoľahlivosť výpočtu cenových indexov otestovať (pričom totoobdobie môže byť stanovené na niekoľko rokov).*
- *Okrem potvrdenia validity týchto dát a ichkvality pre cenové štatistiky, bude potrebné zo strany Štatistického úradu SR vyhodnotiť ich prínos aj s ohľadom navynaložené a potrebné prostriedky (finančné, odborné, technologické), ako ajmožné riziká (napr. závislosť na tretích stranách, bezpečnosť údajov,dostatočné kapacity dátového úložiska).*
- *V neposlednom rade bude potrebné sledovať ajsmerovanie v oblasti vývoja metód a metodík výpočtu cenových štatistík vregióne a nadviazať na skúsenosti ostatných štatistických úradov EÚ, akoaj spolupracovať s Eurostatom prizavádzaní nových metód do praxe.*

Referencie a výstupy

Ďalším dôležitým cieľom projektu bola publikácia výstupov vo vedeckých časopisoch a ich prezentácia na medzinárodných vedeckých konferenciách. Štatistický úrad SR prezentoval výstupy vedeckej činnosti na konferencii Nové Techniky a Technológie v Štatistike (*NewTechniques and Technolgies for Statistics*) v Bruseli, ktorá bola organizovaná Európskym štatistickým úradom (Eurostatom). Vedecké skúmanie a výsledky boli následne publikované v indexovanom vedeckom časopise.

Ďalšia konferencia zameraná na prezentovanie výsledkov skúmania nových dátových zdrojov v cenových štatistikách bola organizovaná Štatistický úradom SR v rámci predsedníctva V4, na ktorej sa zúčastnili reprezentanti štatistických úradov Slovenska, Česka, Maďarska a Poľska. Výstupom tejto konferencie boli publikácie vo vedeckom časopise Slovenská štatistika a demografia.