Centrālā statistikas pārvalde

Alternatīvo datu vākšanas rīku izmantošana cenu statistikā

Patēriņa cenu indeksu daļa
Dāvis Liberts, Nataļja Dubkova

15. Maijs 2019

Centrālā statistikas pārvalde

Pieaugošs preču sortiments

Tehnoloģiju attīstība

Pirkumi interneta veikalos

Pakalpojumu digitalizācija

Dinamiskās cenas

Centrālā statistikas pārvalde

Šajā prezentācijā:

- Automātiskā datu vākšana internetā (web-scraping)
- Web-scraping datu pielietojums cenu statistikā
- Citu valstu pieredze
- CSP paveiktais
- Izaicinājumi šo datu iegūšanā un pielietošanā

Centrālā statistikas pārvalde

Kā tas sākās

Lai arī cenu statistikā ar tādiem jēdzieniem kā **Web-scraping** (Big Data) saskaramies ikdienā arvien biežāk tikai salīdzinoši nesen, to vēsture ir daudz senāka – un to pirmsākumi meklējami līdz ar pirmajām interneta meklētājprogrammām.

Cilvēkiem radās nepieciešamība iegūt informāciju vienlaicīgi no vairākām tīmekļa vietnēm – savukārt ne katrā no tām bija pieejama lejupielādēšanas iespējas, kā arī manuāla kopēšana ar roku bija ”garlaicīga” un neefektīva.

Tas bija aizsākums pirmajiem interneta datu vākšanas robotiem.

Centrālā statistikas pārvalde

Dati mums visapkārt

Mūsdienās šāda datu vākšana ir gājusi tikai plašumā, ko veic neskaitāmi uzņēmumi un organizācijas dažādu pētījumu, mārketinga vai tirgus analīzes nolūkiem.

Var iegūt praktiski jebkurus datus, ja vien tie ir pieejami internetā.

To nodrošina neskaitāmi gan maksas vai bezmaksas rīki, gan individuāli izstrādātas programmas.

Centrālā statistikas pārvalde

Web-scraping darbības princips

Divi galvenie procesi

Interneta lapas ielāde

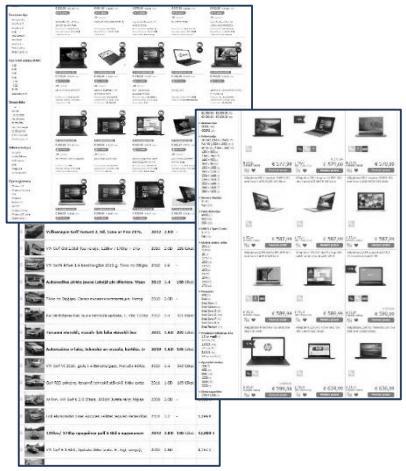
Datu izgūšana no lapas

Centrālā statistikas pārvalde

Automātiskā cenu vākšana

WEB-SCRAPING

DATU GLABĀŠANA



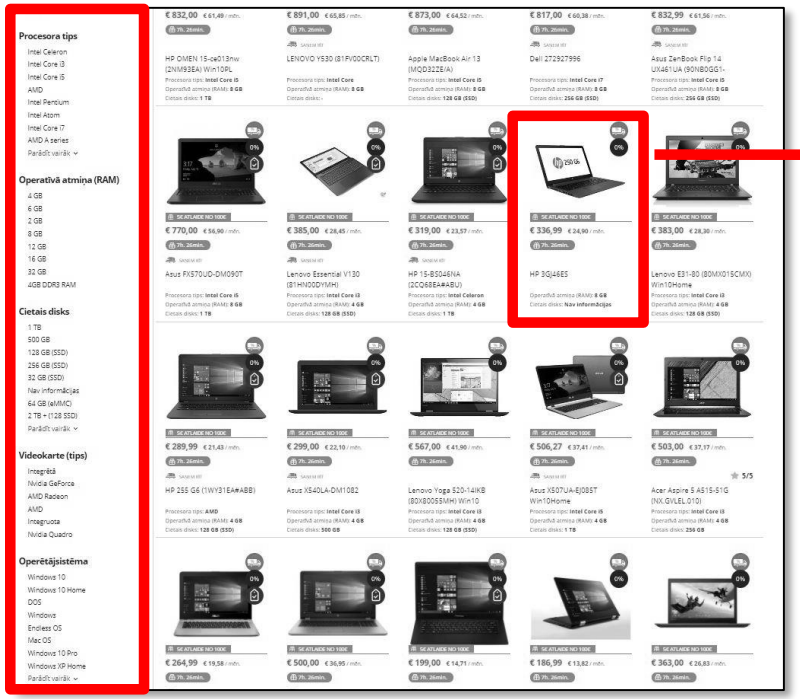
DATU VALIDĀCIJA

DATU ANALĪZE

Centrālā statistikas pārvalde

Web-scraping darbības princips

Automātiski tiek ievadīti nepieciešamie parametri

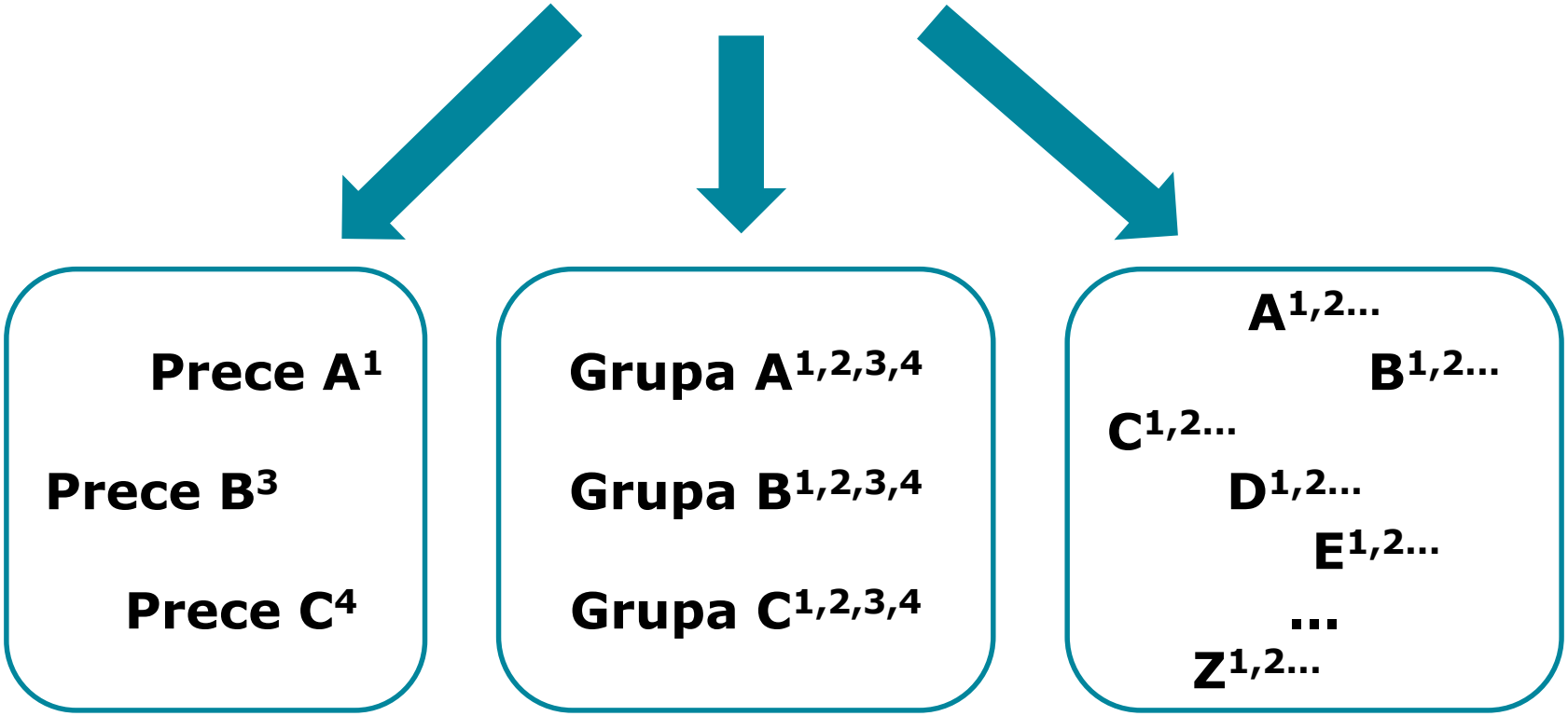


Tiek atlasīta kāda konkrēta prece vai preču grupa

Atlasītie dati tiek strukturizēti un saglabāti tālākai datu analīzei

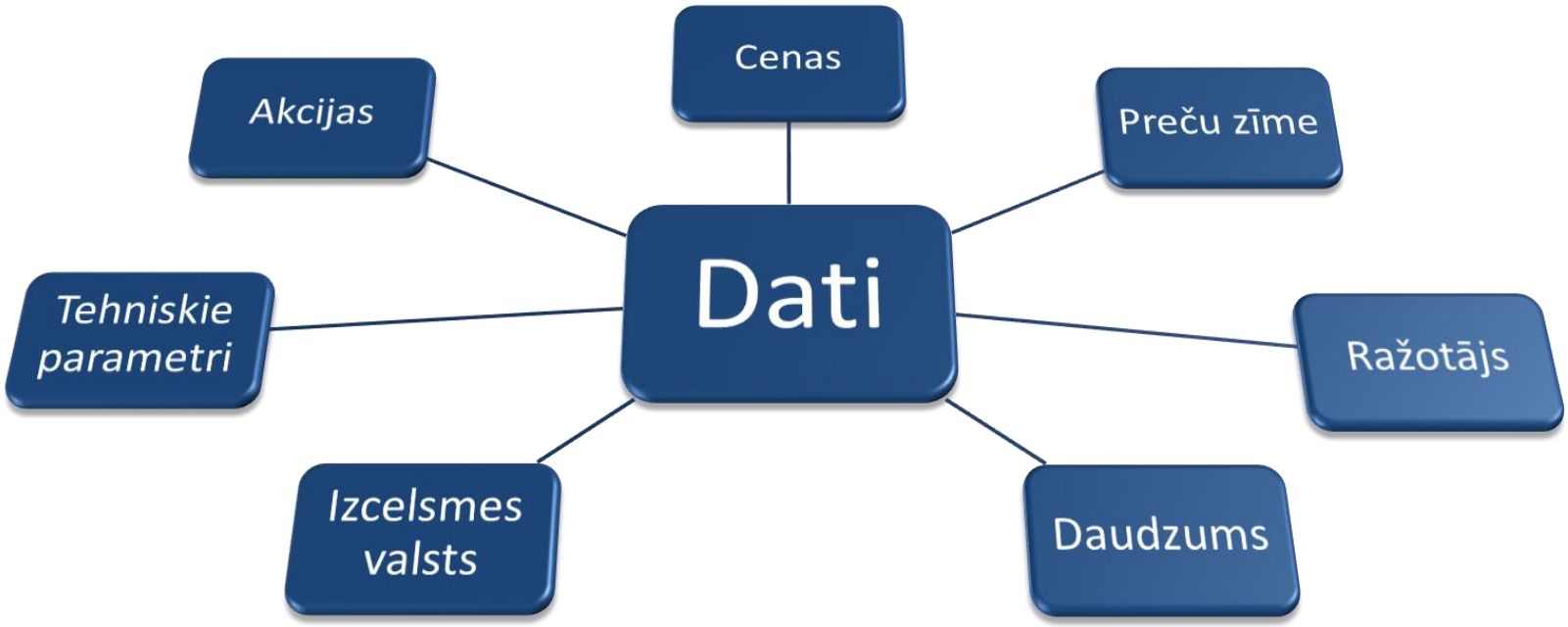
Centrālā statistikas pārvalde

Web-scraping metodes cenu statistikā



Centrālā statistikas pārvalde

Informācija cenu statistikai



Centrālā statistikas pārvalde

Pielietojums cenu statistikā

- Automātiskā datu vākšana internetā dod iespēju veikt cenu mērījumus ievērojami biežāk un plašākā apjomā
- Iespēja optimizēt datu vākšanas procesu
- Samazina respondentu noslodzi
- Būtiski vākt cenas tur, kur tiek izdarīti pirkumi
- Risinājums jauniem izaicinājumi cenu statistikā saistībā ar pieaugošu preču klāstu, tehnoloģiju attīstību un digitalizāciju

Centrālā statistikas pārvalde

Eurostat skatījums

- Eurostat saskata potenciālu web-scraping izmantošanā cenu statistikā
- Cenu statistika pāriet no «pildspalvas un papīra» uz inovatīvākām metodēm
- Digitalizācija arvien vairāk ienāk cilvēku dzīvē un maina pirkumu paradumus - interneta veikali, digitālie pakalpojumi
- Jauni datu avoti prasa arī jaunas metodes
- Izaicinājums nākotnē - jaunu datu avotu un metožu interpretācija un komunikācija ar datu lietotājiem

Centrālā statistikas pārvalde

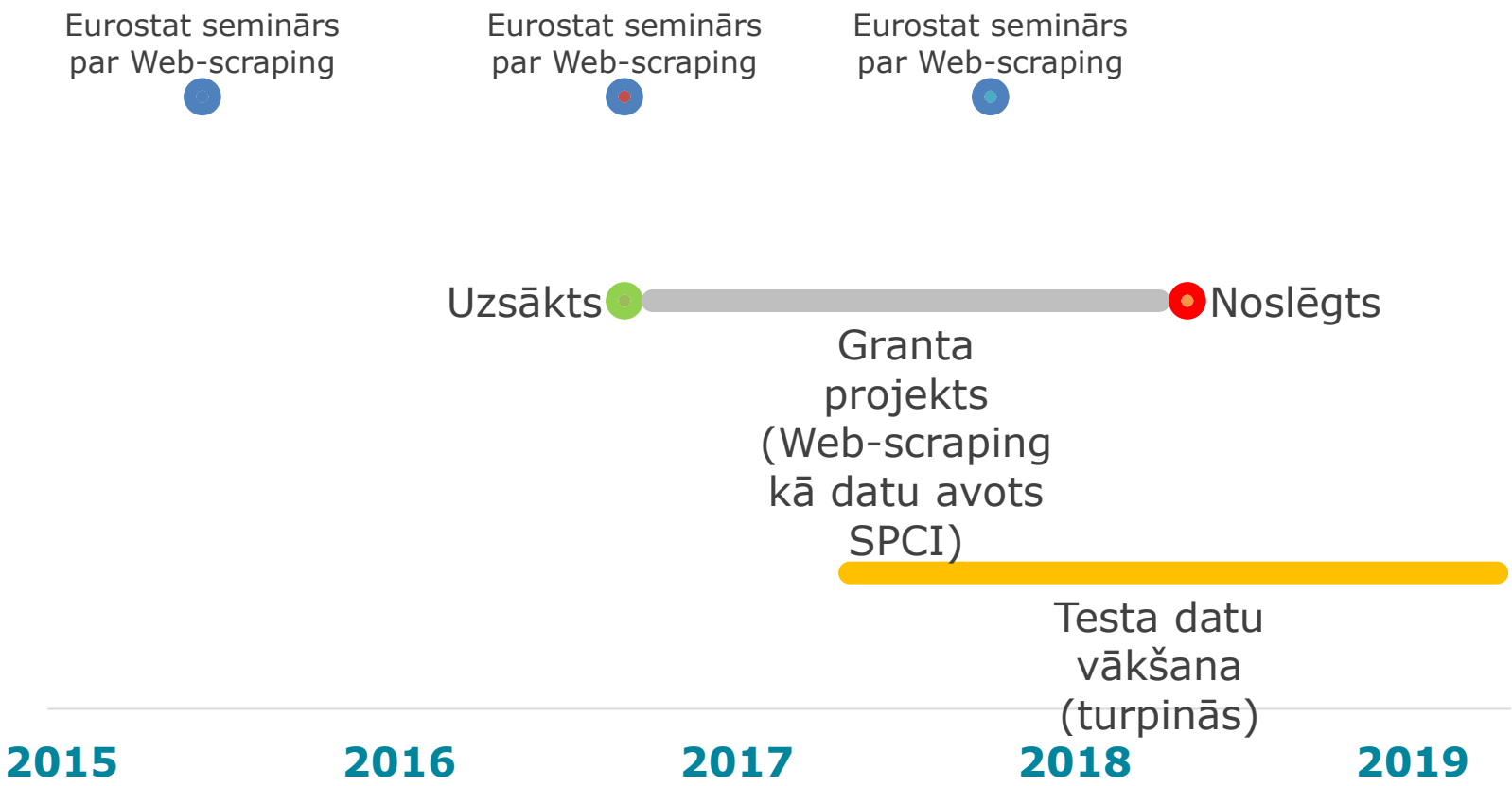
Centrālā statistikas pārvalde

Citu valstu pieredze

- Ir valstis, kas vāc liela apjoma datus par visām preču grupām. Ir valstis, kas sāk ar datu vākšanas automatizāciju, atbilstoši iepriekš noteiktai izlasei
- Vācot liela apjoma datus par visām pieejamām preču grupām un precēm, liels izaicinājums ir to klasificēšana – tādam nolūkam pēta **machine learning** iespējas
- Vācijas kolēģi veic pētījumu par dinamiskajām cenām un to izaicinājumiem cenu statistikā – ap 2700 precēm cenu vākšana tiek veikta reizi stundā

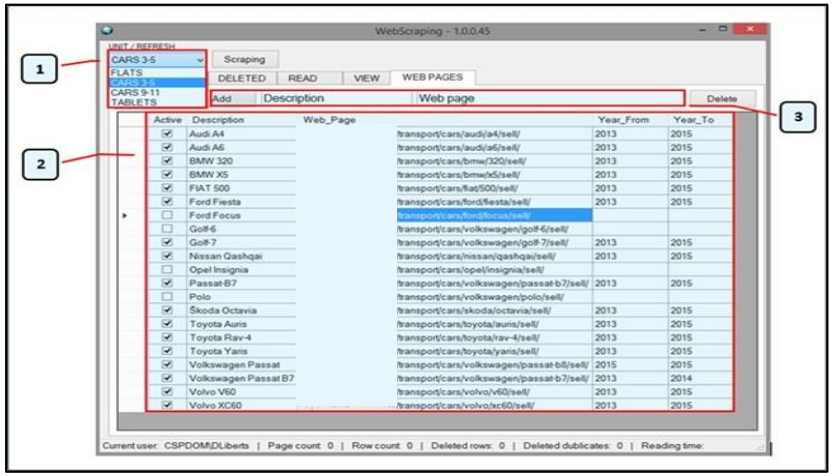
Centrālā statistikas pārvalde

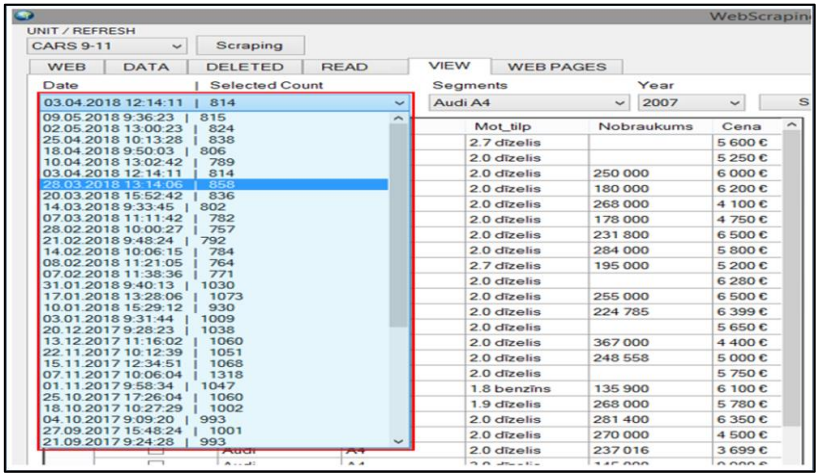
CSP paveiktais saistībā ar web-scraping

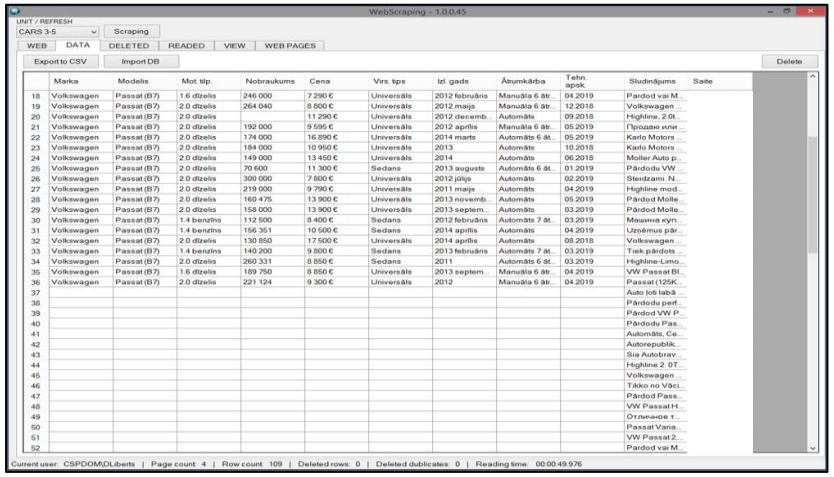


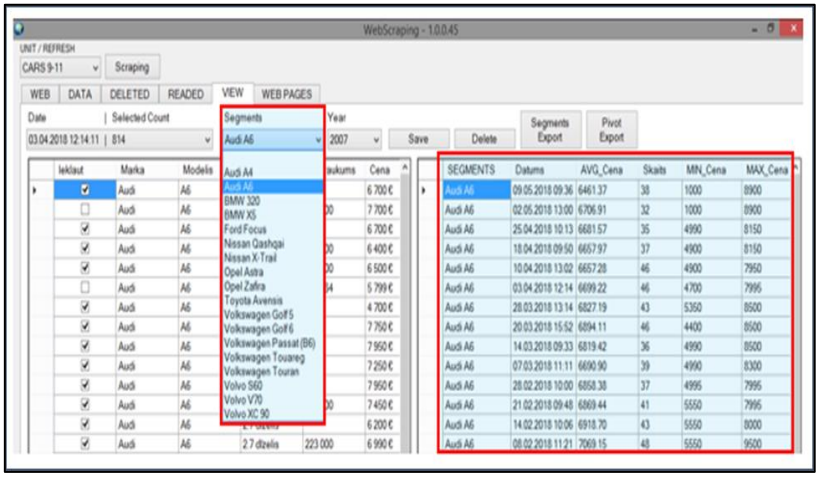
Centrālā statistikas pārvalde

Datu vākšanas rīks









Centrālā statistikas pārvalde

Priekšrocības

- Iespēja iegūt vairāk datu nekā tas iespējams manuāli
- Iespēja iegūt datus pat pa stundām, dienām vai nedēļām
- Cenu vākšanas laiks:

Preču grupa	Manuāli vācot cenas	Web-scraping
3-5 g lietotas automašīnas	~ 3 stundas	~ 10 minūtes
9-11 g lietotas automašīnas	~ 2 stundas	~ 21 minūte
Dzīvokļu īres cenas	~ 5 stundas	~ 30 sekundes

- Web-scraping viennozīmīgi negarantē labāku datu kvalitāti, bet noteikti daudz lielāku datu apjomu
- Papildus informācija par tirgus tendencēm

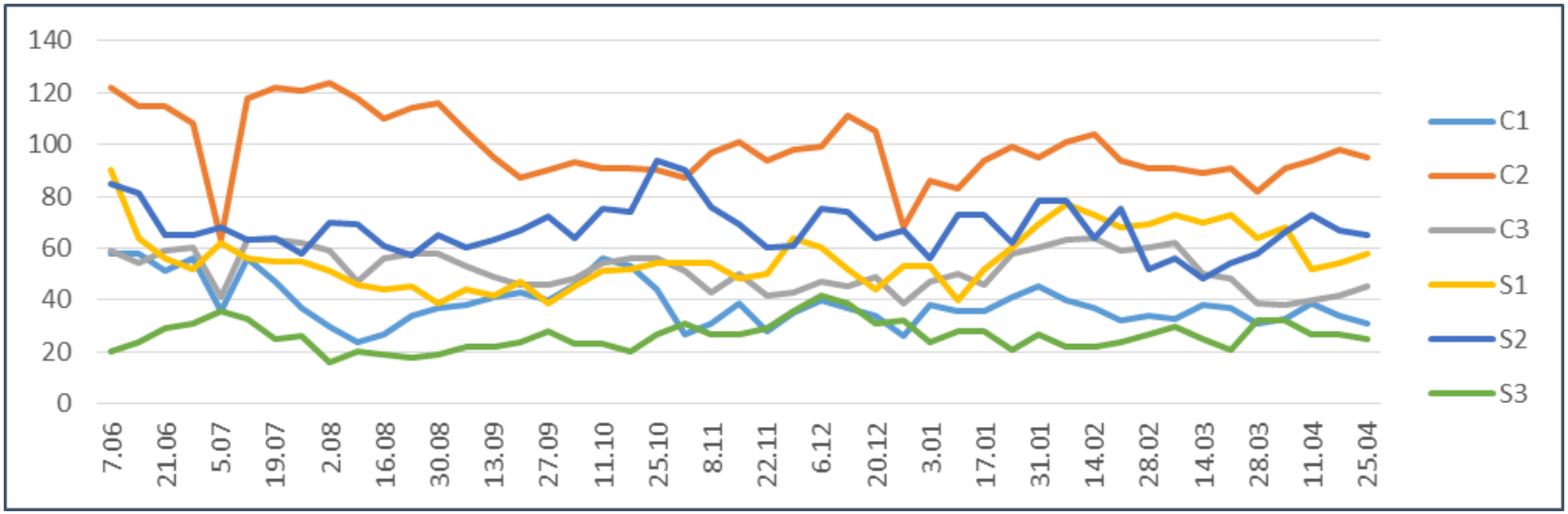
Centrālā statistikas pārvalde

Datu analīzes iespēju piemēri

Planšetdatoru pieejamība tirdzniecības vietās

Centrālā statistikas pārvalde

Datu analīzes iespēju piemēri



Dzīvokļu īres sludinājumu skaits sadalījumā pa segmentiem

Centrālā statistikas pārvalde

Izaicinājumi

- Tehnoloģiski sarežģīts datu vākšanas process
- Web-scraping rīks jāpielāgo katram veikalam
- Nepieciešama pastāvīga procesa uzraudzība un IT atbalsts
  (tehniskas problēmas, izmaiņas mājas lapās)
- Metodoloģiski izaicinājumi
  (izlases apjoms, cenu vākšanas biežums, datu segmentēšana, datu integrēšana patēriņa cenu indeksu aprēķinā)
- Cenu indeksu izstrāde kļūst arvien sarežģītāka - interesantāka

Centrālā statistikas pārvalde

Pārmaiņas cenu reģistrēšanas procesā

CENU VĀKŠANA MANUĀLI

Cenu reģistratori

Datu analītiķi

CENU VĀKŠANA AR WEB-SCRAPING

Pastāvīgs IT atbalsts

Centrālā statistikas pārvalde

Paldies par Jūsu uzmanību!

Davis.Liberts@csb.gov.lv
Natalja.Dubkova@csb.gov.lv

www.csb.gov.lv