National bank of Ukraine

Combination of techniques for online price data collection: Case of Ukraine

National bank of Ukraine

Geneva, April 2023

Summary

Start

Online data collection project was launched in 2015. Initially, it was a set of scripts for price scrapping from online sources with semi-manual data processing in Excel after. Data was collected mainly from the websites of small regional supermarkets

Development

Update of online data collection system to a fully automated data collection/filtering/calculation system was started in 2018. From the very beginning, it was planned to receive information about food products and fuel prices, as well as certain services with regional breakdown, prices of which are available online. 5 large supermarkets were selected as the main source of information

Additional initiatives

The project also included sentiment analysis module development - collection of text information, processing, classification and determination of its sentiments using a CNN (confluence neural network) approach
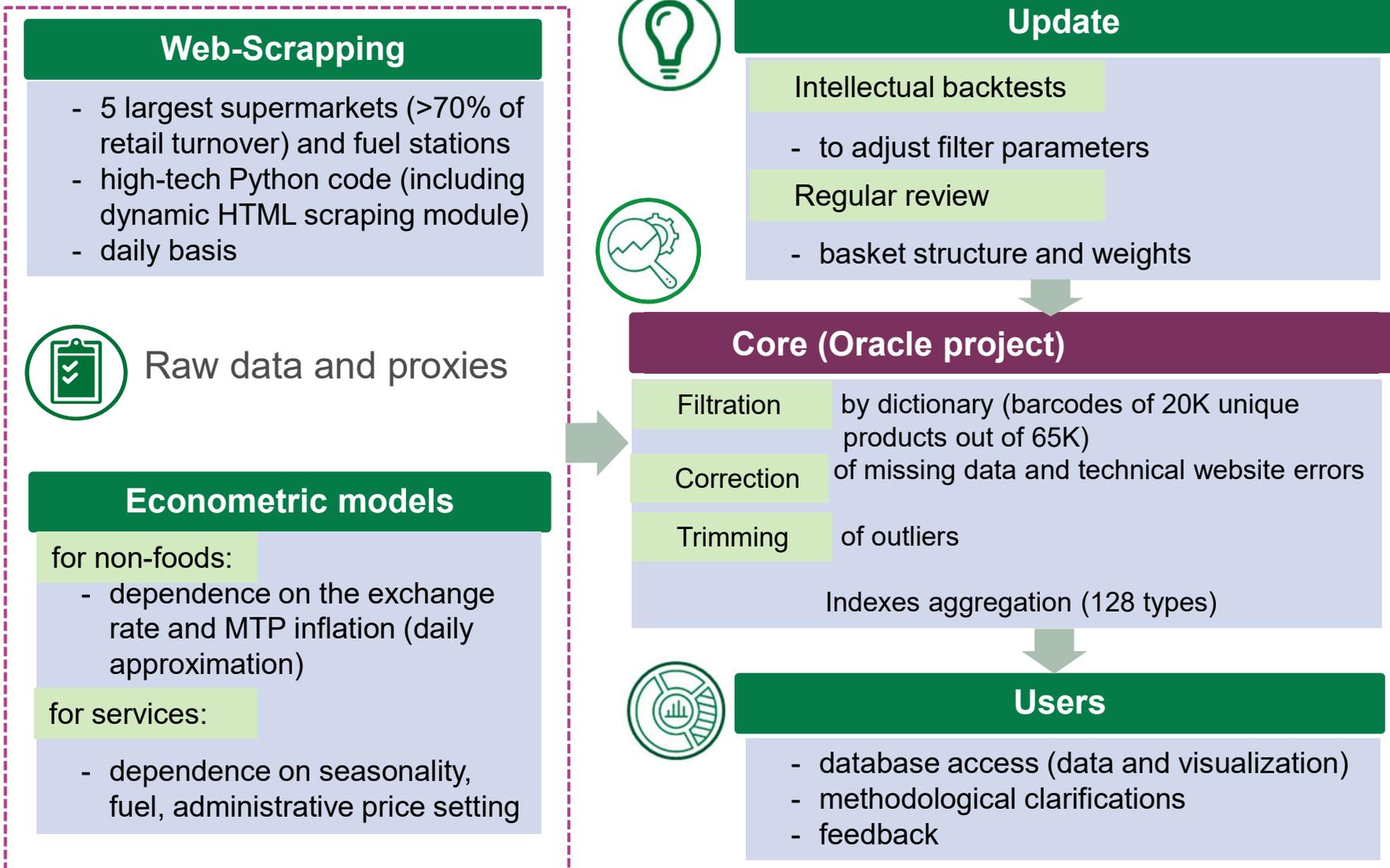
What are the goals of the project?

The goal of the project was to launch automated system of data collection, error correction, and price index calculation to:

- obtain recent data in uncertainty times and high inflation rates for consideration at the Monetary Policy Committee meetings

- track the effects of monetary transmission

- conduct research (on "price stickiness", regional price development and other microeconomic facts of pricing)
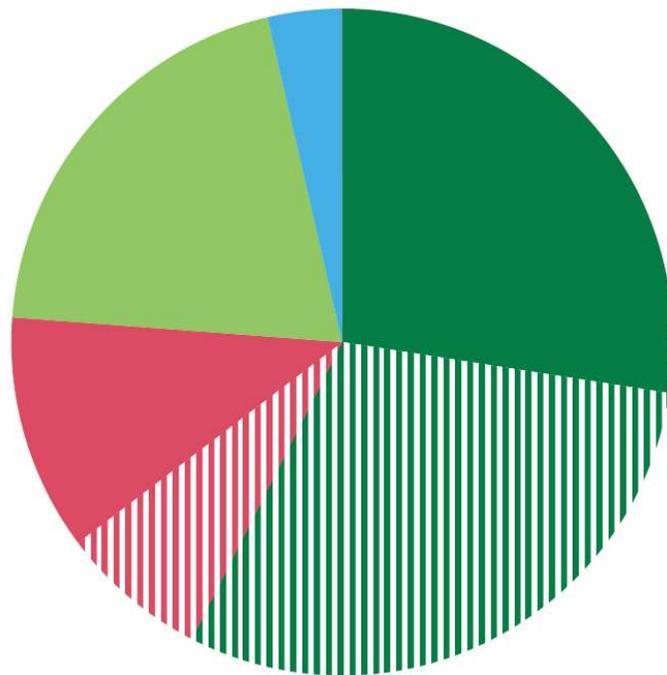
Architecture of the model has to be expandable for real estate prices, employment and output estimation

Online data collection system architecture

Web-Scrapping

- 5 largest supermarkets (>70% of retail turnover) and fuel stations
- high-tech Python code (including dynamic HTML scraping module)
- daily basis

Raw data and proxies

Econometric models

for non-foods:
- dependence on the exchange rate and MTP inflation (daily approximation)

for services:
- dependence on seasonality, fuel, administrative price setting

Update

Intellectual backtests
- to adjust filter parameters

Regular review
- basket structure and weights

Core (Oracle project)

Filtration   by dictionary (barcodes of 20K unique products out of 65K)

Correction   of missing data and technical website errors

Trimming   of outliers

Indexes aggregation (128 types)

Users

- database access (data and visualization)
- methodological clarifications
- feedback

Coverage of CPI basket by web-scraped items is more than 70%

Legend:
- Core CPI
- Regulated prices and tariffs
- Raw food
- Fuel

* Shaded field is the share of products for which data are not collected.
Source: State Statistics Committee, NBU calculations.

Items scrapped: appr. 65 000
Items selected for basket: appr. 20 000
Groups: 128

Risks for online data collection system

| Risk	Probability	Comment
Change of supermarket's site structure	🟡	Model needs an update. Can take 1-2 weeks
New site platforms	🟡	Model needs an update. Can take a week
Need to update the set of products and price filter settings every year	🟢	Update of dictionary can take two weeks or more
Non-food products prices are difficult to collect (goods are changed frequently)	🟡	Update of econometric models for non-food prices estimation. More frequent update of a dictionary

🔴 High
🟡 Average
🟢 Low

The difference of scrapped data with official data widened as the situation in Ukraine became unstable

Core CPI data comparision, pp

Legend: Core | Core (realtime) | Difference
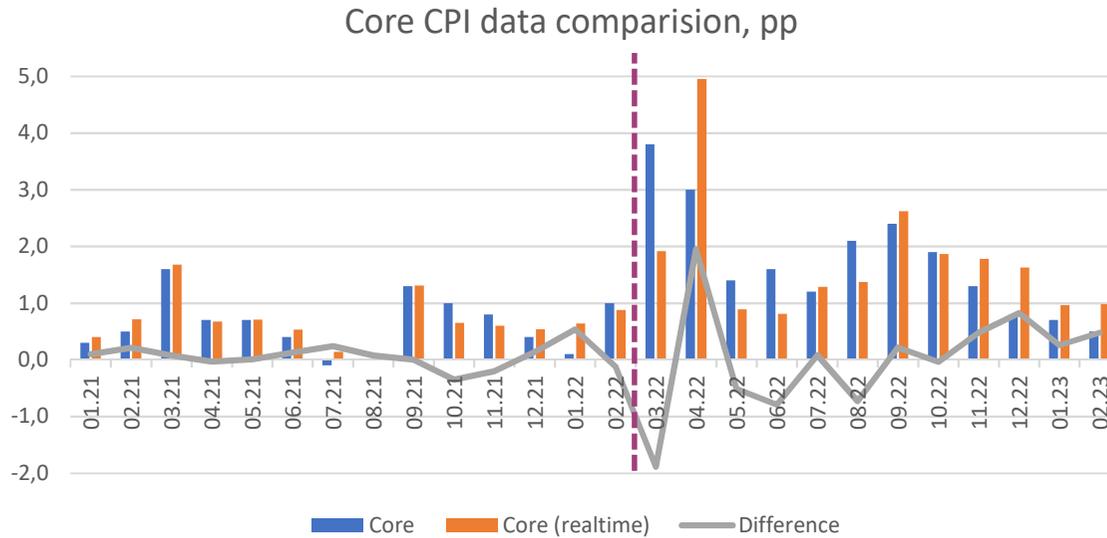
From March, 2022 prices fluctuations become more unpredictable. All risks were realized.

CPI data comparision, pp

Legend: CPI | CPI (realtime) | Difference
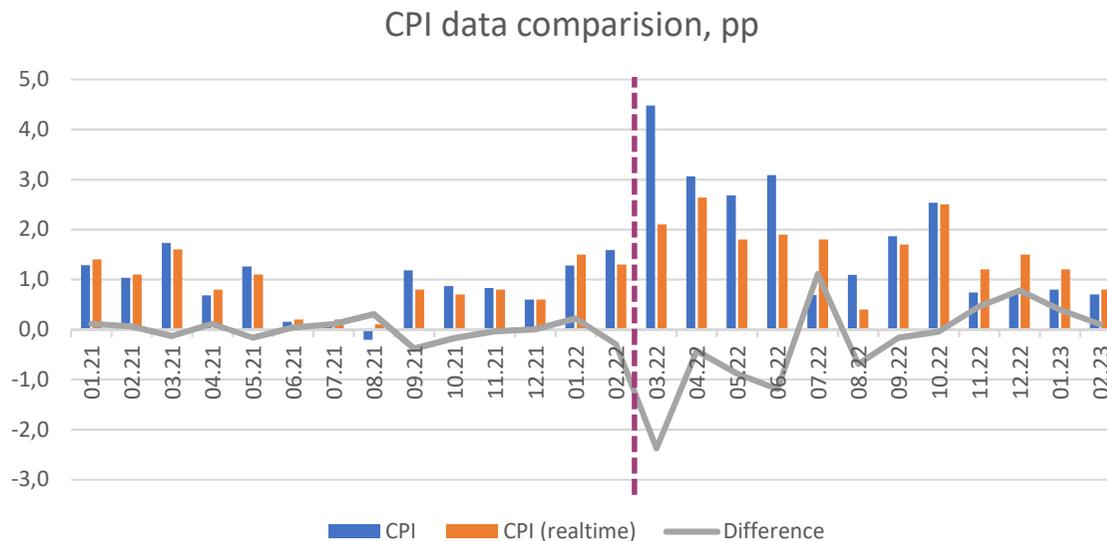
The higher inflation – the bigger difference between data collected in real-time mode and official data