Innovative uses of web scraped data in the Canadian Clothing and Footwear Consumer Price Index

Valéry Dongmo Jiongo, Statistics Canada

ESCoE Conference on Economic Measurement 2021

May 12, 2021

Delivering insight through data for a better Canada

Outline

- Background
- Web scraping
- Classification
- Index number formula
- Integration of web scraped data
- Future work

Background

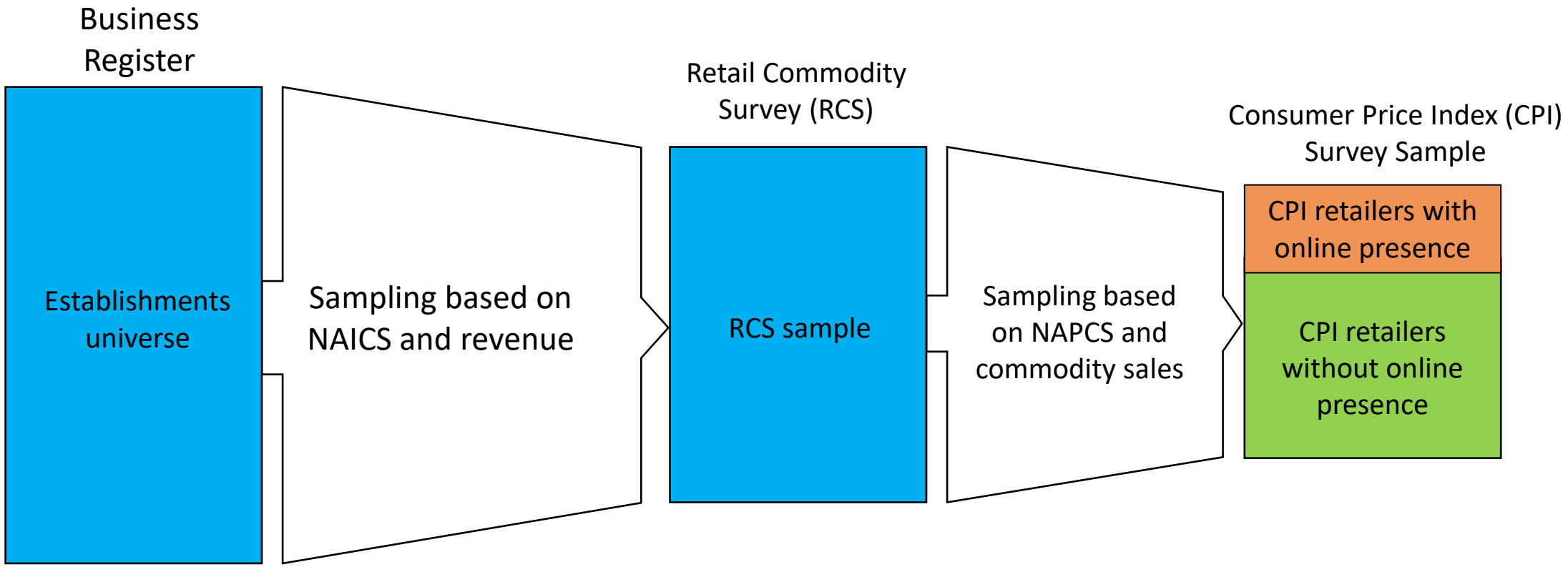
E-commerce expanded

Source: Table 20-10-0072-01 Retail e-commerce sales, unadjusted (x 1,000)

Source: Aston, Vipond, Virgin and Youssouf (2000) - *Statistics Canada* - *Catalogue no. 45280001*

Background

Traditional field collection may not provide sufficient coverage of online products



NAICS: North American Industry Classification System

NAPCS: North American Product Classification System

Background

Implementation strategy

- **Progressively replace field collected data with web scraped data:** For retailers included in the CPI sample having both physical store and online presence.

- **January 2020: First introduction of web scraped data in production**

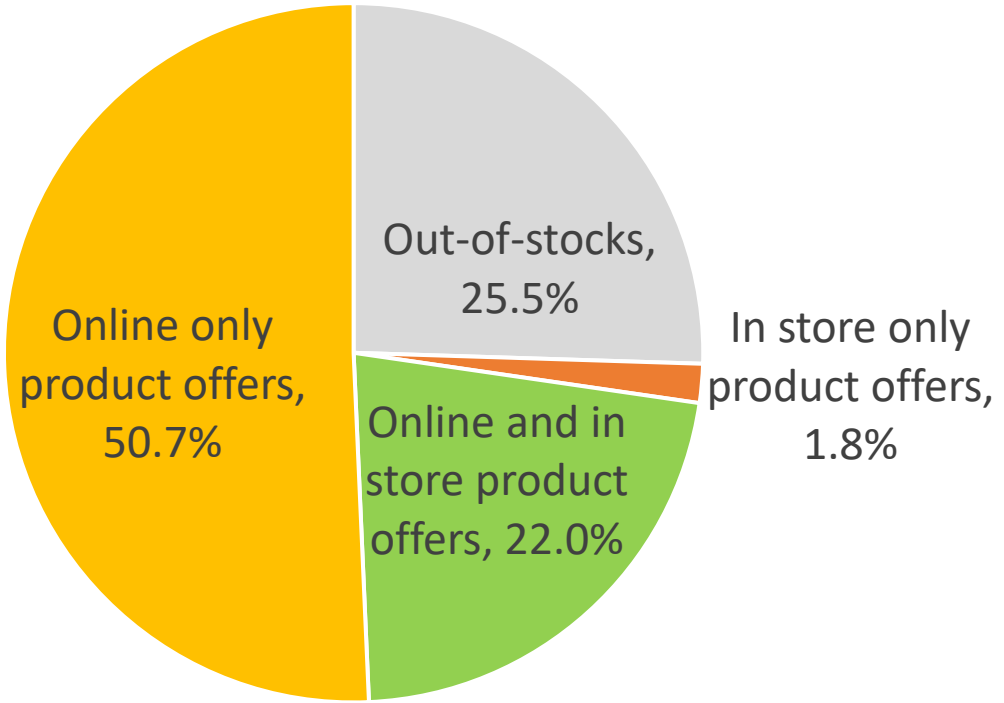
  - Three selected retailers were included in the web scraped collection and processing method.

  - They represented together 8% of the apparel market in 2019.

  - Statistics Canada interviewers are collecting approximately 10% fewer prices in the clothing and footwear retail stores every months.

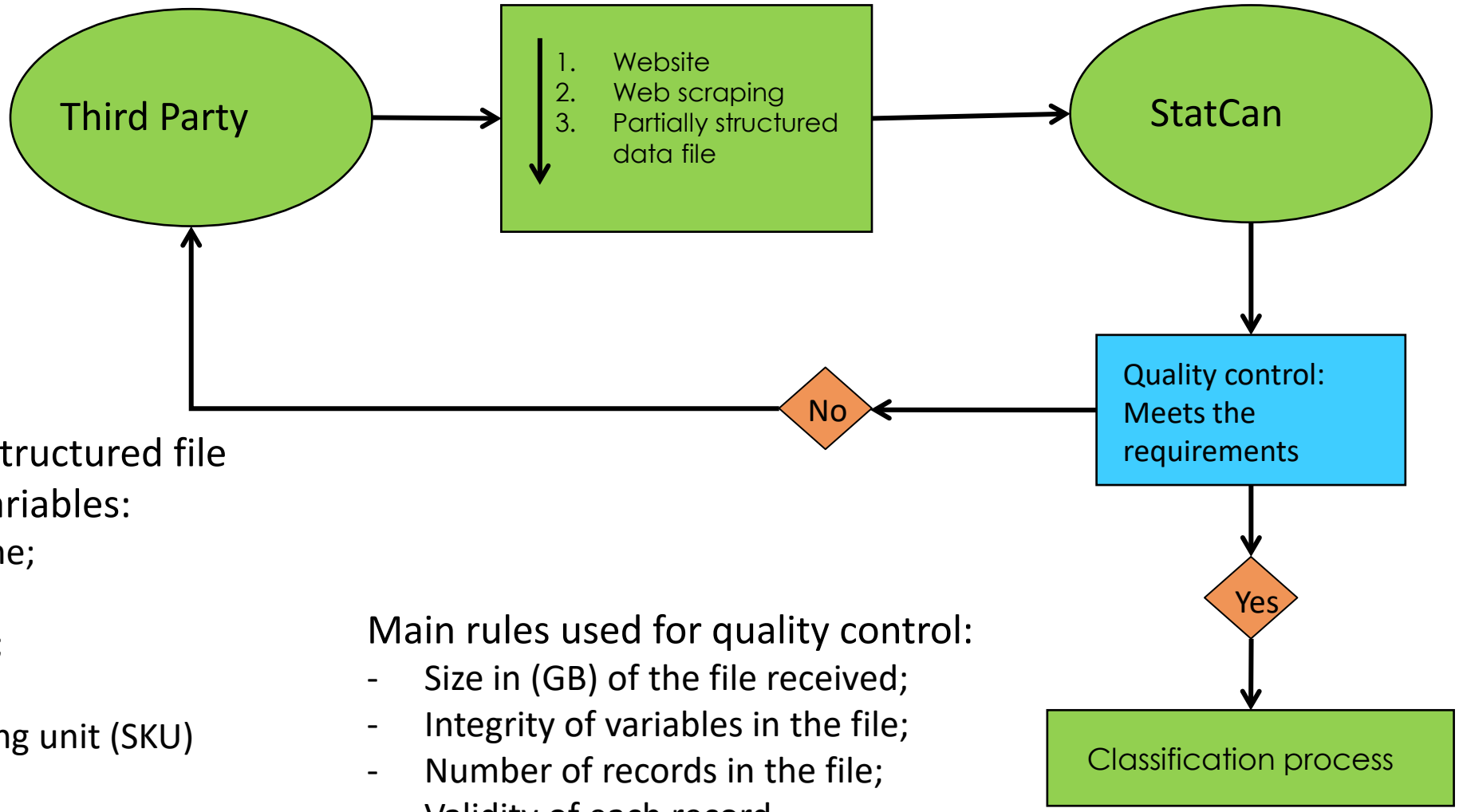
Web scraped data improves coverage

Three web scraped retailers: Distribution of products offered according to their availability in stores or online



Source: January 2020 web scraped data

Web scraping



The partially structured file contains 29 variables:

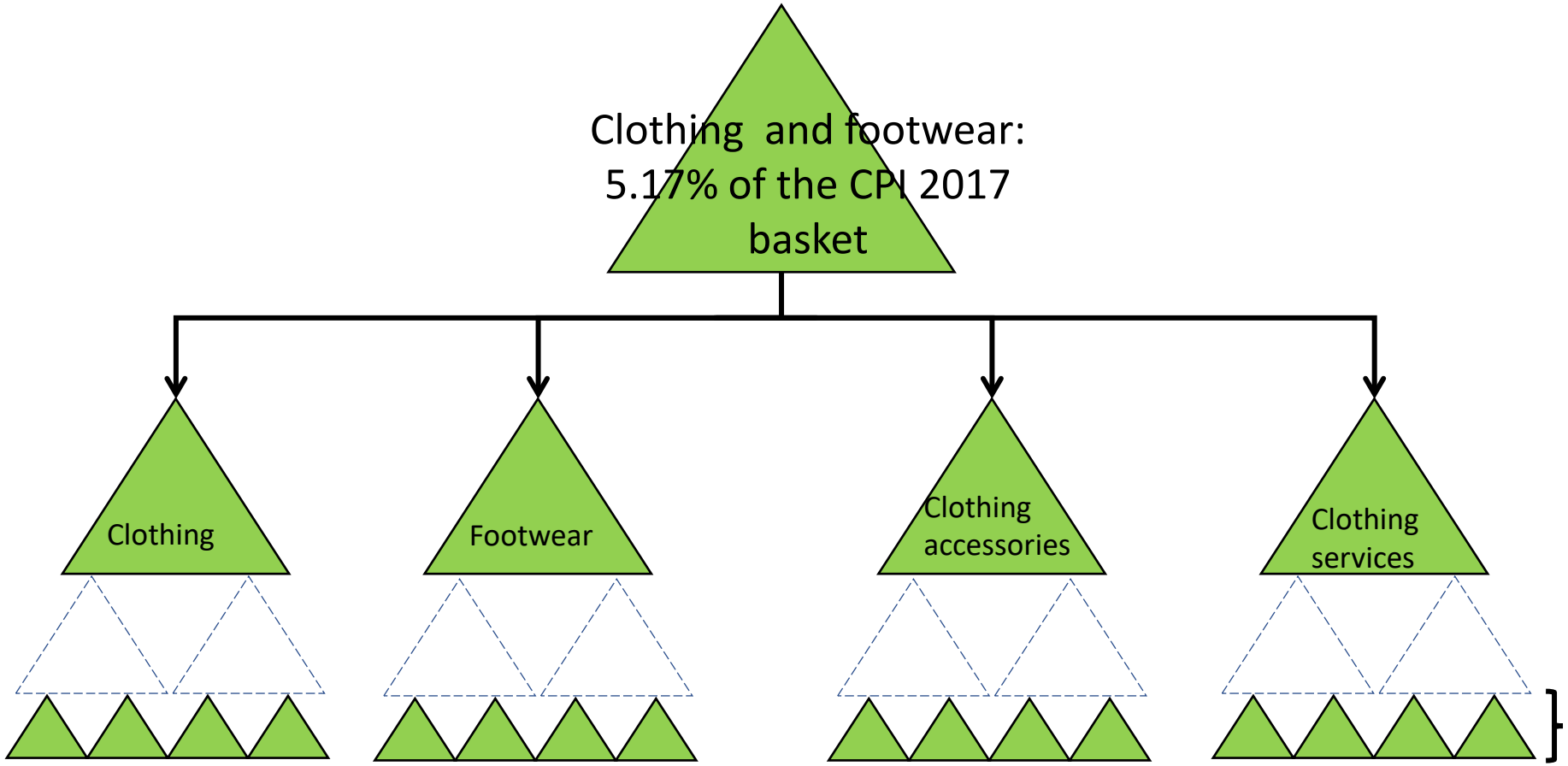
- Product name;
- Category;
- Description;
- Features;
- Stock keeping unit (SKU) identifier;
- Etc.

Main rules used for quality control:

- Size in (GB) of the file received;
- Integrity of variables in the file;
- Number of records in the file;
- Validity of each record.

Classification

Clothing and footwear sub-components



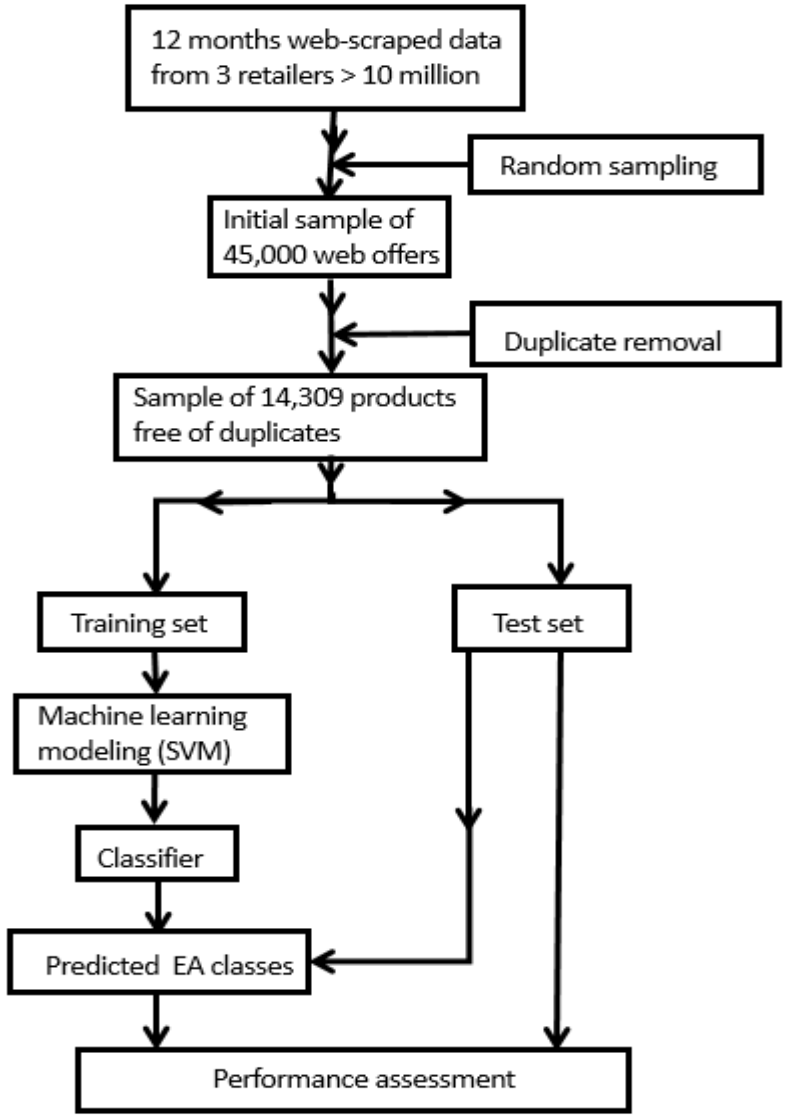
Classification

Development of the Support Vector Machine (SVM) classifier

- **F_1 score** = Average of Precision and Recall.
  - SVM: 92.2%
  - Random Forest: 87.8%

- **Precision** = percentage of relevant classification among the set of products that are assigned to the EA by the classifier

- **Recall** = percentage of relevant classification among the set of products that effectively should belong to the EA



Labelling

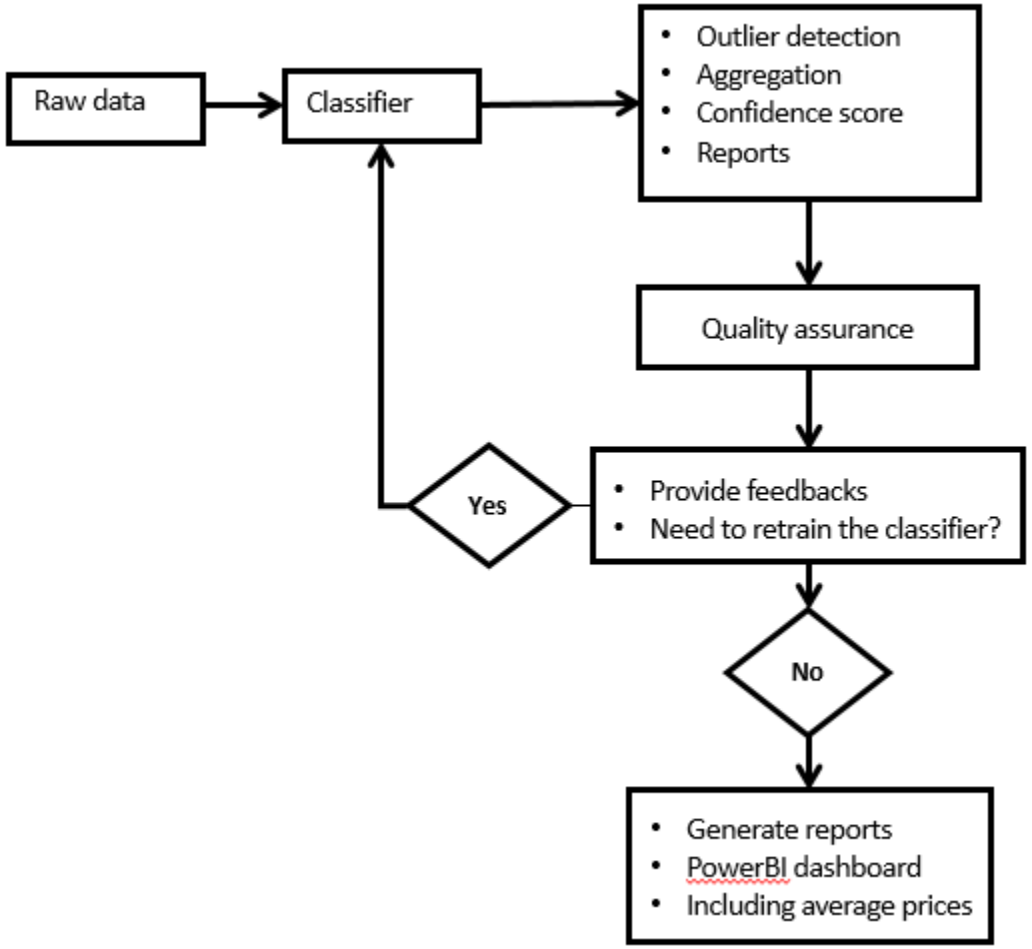
Modelling

Performance assessment using F_1 score

Classification

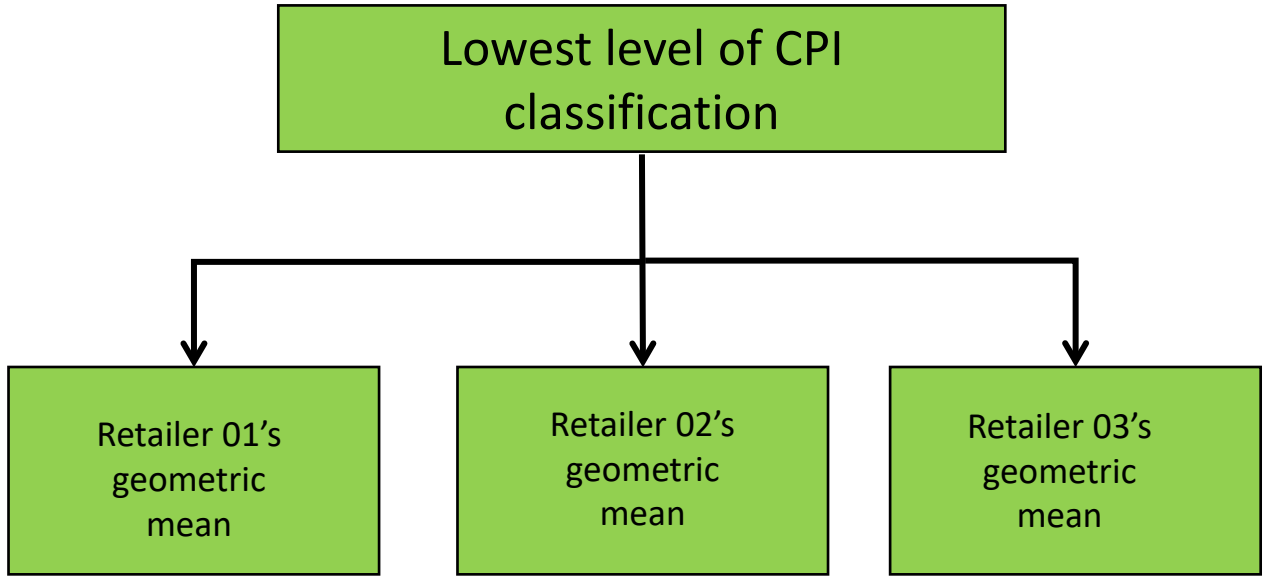
Implementation of the SVM classifier

- **Monthly run of the classifier**



- **Quality assurance (Summary from Jan 2020 to Feb 2021)**
  - Average number of manually verified unique products per month is 3,877;
  - Average precision or average percentage of well-classified products is 95.0%, which is considered a satisfactory performance

Index number formula

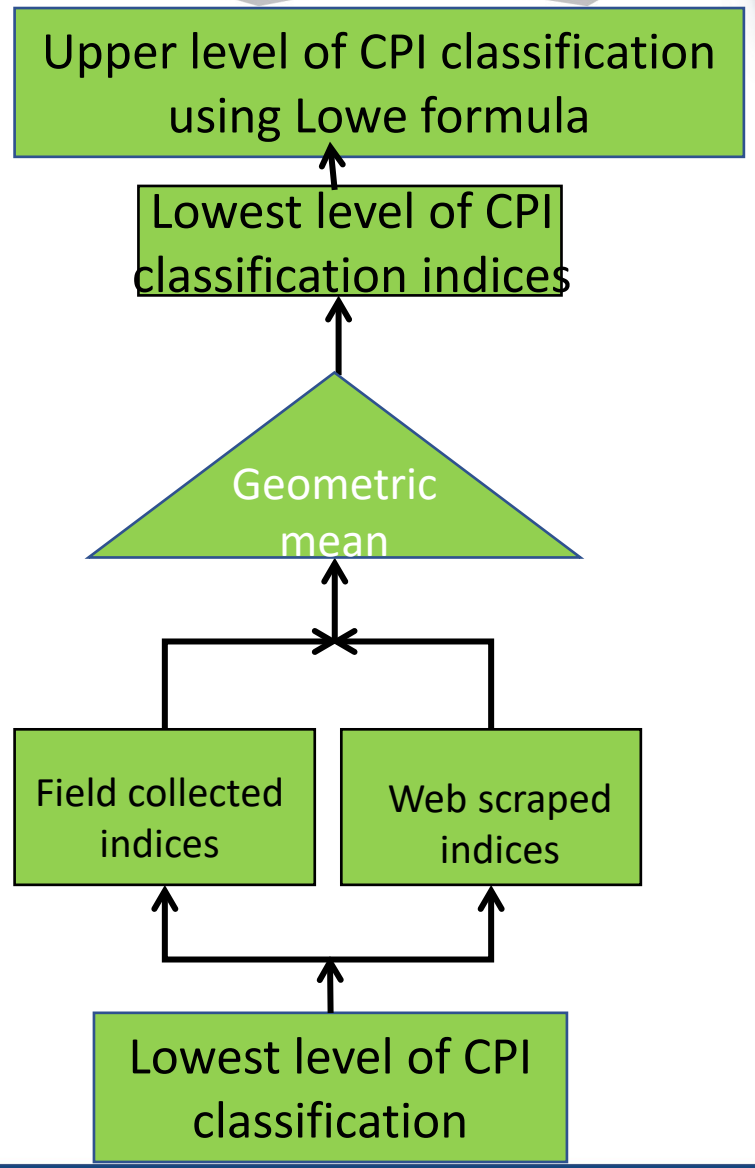


- Use all the product offers

- Consistent with the Jevons formula used to aggregate all field collection price quotes at the lowest level of the CPI classification

Integration of web scraped data

- Assume retailer has a national pricing strategy
- Calculated prices are used in all geographies where the retailer operates through either brick and mortar stores or a warehouse for local product shipping
- Before implementing the web scraping methodology for the three retailers in production, we did parallel runs to assess its effect on the CPI

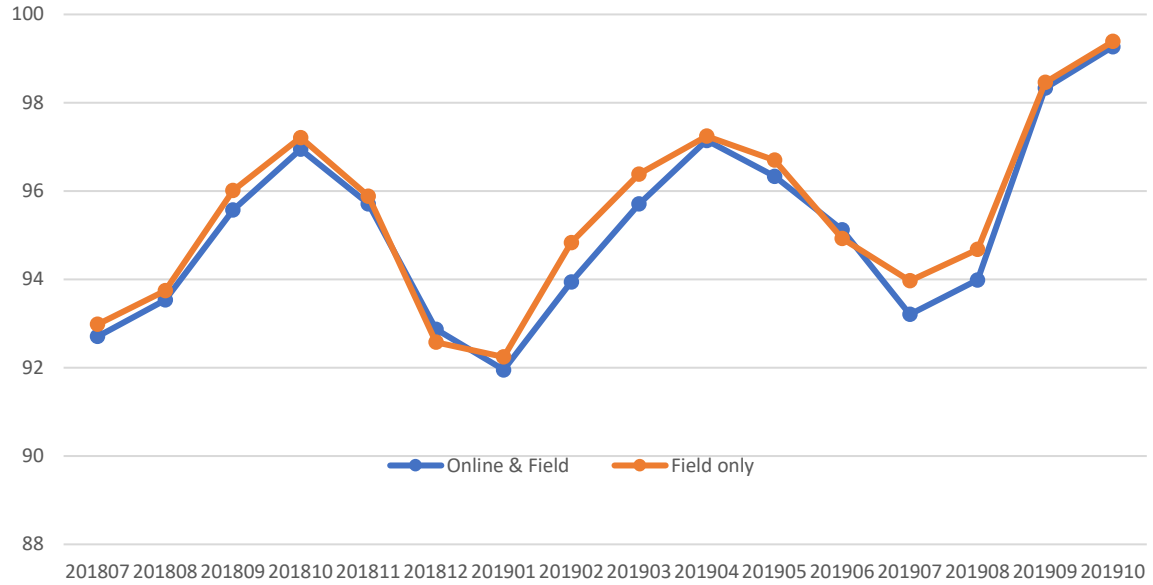


Integration of web-scraped data

Parallel runs

Negligible impact of web-scraped methodology on the all-items CPI

Web-scraped closely tracks the trend in the field collected clothing and footwear CPI



Future work and research development to improve the use of web scraped data in the CPI

- **Development of a user interface application for web scraped data processes**
  - Would efficiently support the web scraped data processes and minimise the risk of errors
- **Active learning**
  - Cost efficient in terms of sample selection for the labelling of future retailers to include in the web scraped methodology;
- **Clustering**
  - Mitigate the impact of product churn;
- **Outlier detection**
  - Efficiently target the candidates for quality assurance (QA)
- **Treatment of outliers**
  - Develop an outlier robust geometric mean M-estimator;

Conclusion

- Statistics Canada is leading the way of unique mode field collected data toward alternative data source to improve the coverage and the quality of the CPI

- Research and development are conducted to smoothly include additional retailers into the web scraped collection and processing method

Thank you!

For more information, please contact:

valery.dongmo-jiongo@Canada.ca