Estimating computers and peripherals price indices using web-scraped data

Lance Taylor & Roobina Keshishbanoosy, Statistics Canada

ESCoE Conference on Economic Measurement 2021

May 12th 2021

Delivering insight through data for a better Canada

Outline

- Background
- Data
- Quality adjustment
- Aggregation
- Concluding remarks

Background (1/2)

- Computers and peripherals are used to price the Computer equipment, software and supplies index in the Canadian CPI
  - About .42% of total Canadian basket weight

- January 2021 CPI commenced new methodology with corresponding introduction of web scraped data source

Background (2/2)

- Products covered
  - Laptops
  - Desktops
  - Monitors
  - Printers
- Challenges with the previous approach
  - Data cost
  - Timeliness – two months lag
  - Quality adjustment

Data (1/4)

- Prices collected and delivered weekly from retailer websites
  - Eliminates a month of lag from the previous data source
  - Price changes more likely to be recorded in the period they occur
- Multiple outlets per retailer
  - Currently 3 retailers
- Average price taken for each item (SKU) within a retailer across outlets and weeks
- Weights used in modelling and aggregating are sourced from industry reports and Statistics Canada's Merchant Retail Trade Survey

Data (2/4)

- Most of the cleaning and prep is automated while allowing subject-matter intervention at various stages
  - Previously human cleaning added a month lag, automation eliminates this
  - Current production now takes one employee one full day
- Product characteristics contained in semi-structured text (descriptions, names, crumb trails, etc.)
  - Processing to identify characteristics, harmonize units for continuous variables, standardize discrete variables etc.
- Currently no classifier used
  - Rules based product and characteristic identification
  - Items are by default in scope if they have the characteristics of the product in question (after cleaning)

Data (3/4)

- **desktop 1 text:**

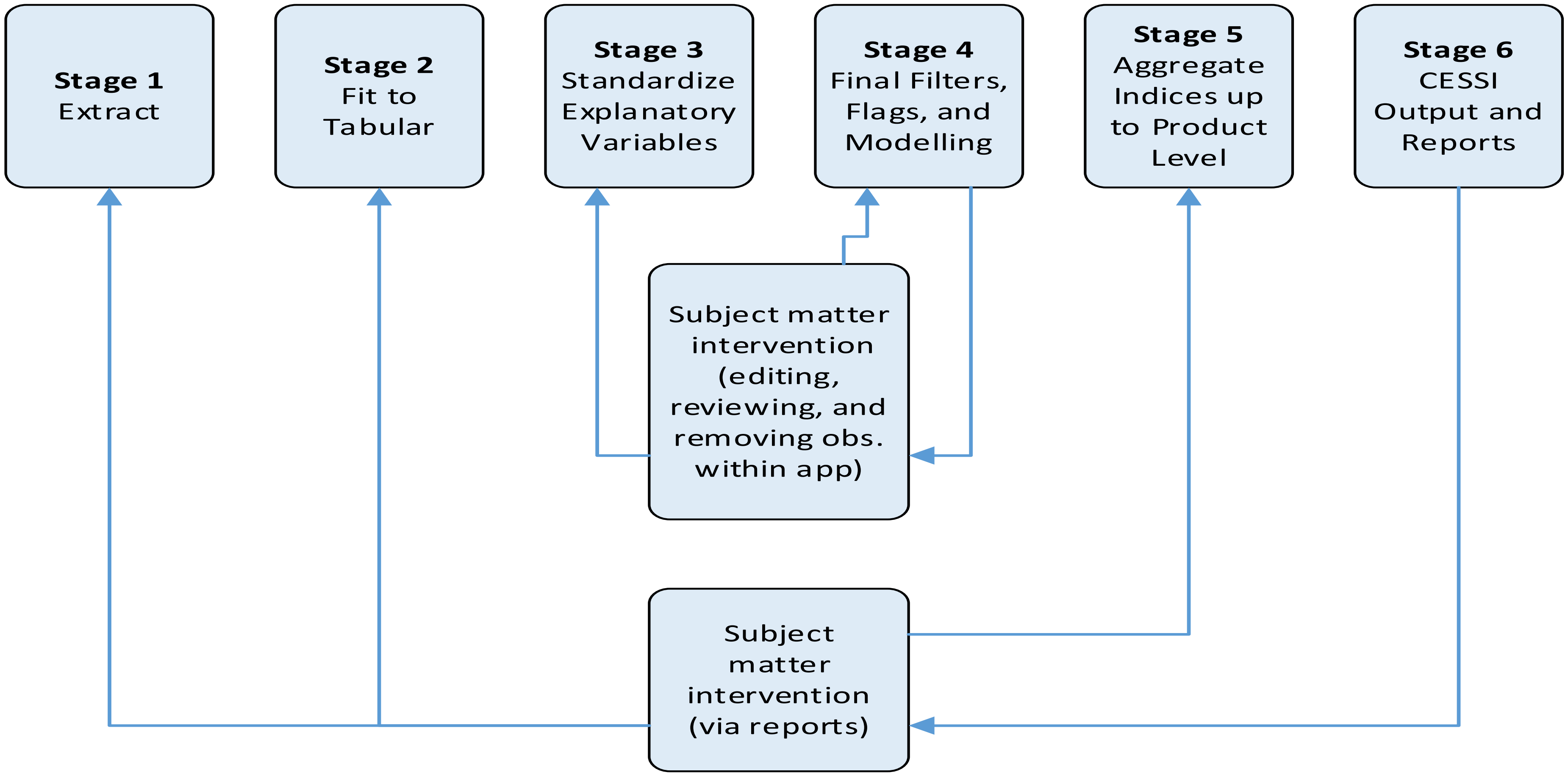
Audio Output: 1 x Microphone/Headphone Combo. Brand Description: ASUS. Colour: Silver. Dimensions (WxDxH): 24.8" x 18.2" x 2-7.4“. Display: 27” LED-backlit Touchscreen 1920 x 1080. Hard Drive: 1TB 5400RPM HDD. Input Device: USB Keyboard / Mouse. Model: V272UA-DS501T. Networking: Gigabit Ethernet, IEEE 802.11 AC (1*1), Bluetooth 4.1. Operating System: Windows 10. Power: 90W AC Adapter. Processor: Intel Core i5-8250U 1.6GHz Quad-Core. **Ram: *8GB DDR4***. Webcam: 1.0M 720p. Weight: 8.48 kg.

- **desktop 2 text:**

Assembled Depth (in.): 25 in, Assembled Length (in.): 21.5 in, Assembled Weight (lbs.): 35 lb, Assembled Width (in.): 12.5 in, Compatible Memory Cards: Not Applicable, Connectivity: 3.5mm Audio3.5mm JackHDMIPS/2USB 2.0USB 3.1Wi-FiWLANDisplay, Graphics card: NVIDIA GeForce RTX 2070, Hard Drive Storage: 1000 GB, Hard Drive Type: Solid State Drive (SSD), Memory (RAM) Size (GB): 16 GB, **Memory (RAM) Size (GB): *16 GB***, **Memory Type: *DDR4 SDRAM***, Operating System Version: Microsoft Windows 10 Home, Optical drive: None, Processor Speed (GHz): 3.6 GHz, Processor type: Intel Core i7-9900K, Product Type: Gaming, Software Included: Not Applicable, Sound Card: Integrated, Brand: CyberpowerPC, Walmart Item #: 31648578, Model #: SLC10200CPG, SKU: 6000199242199, UPC: 81184205770

- Text is made up of feature (underlined) description (italicized) pairs
- Need to assign a feature for each variable used in quality adjustment
- E.g. will want the final data to have a format so each desktop has a column describing the size of its RAM and a column describing the type of RAM
- Variables must then be standardized

Data (4/4)

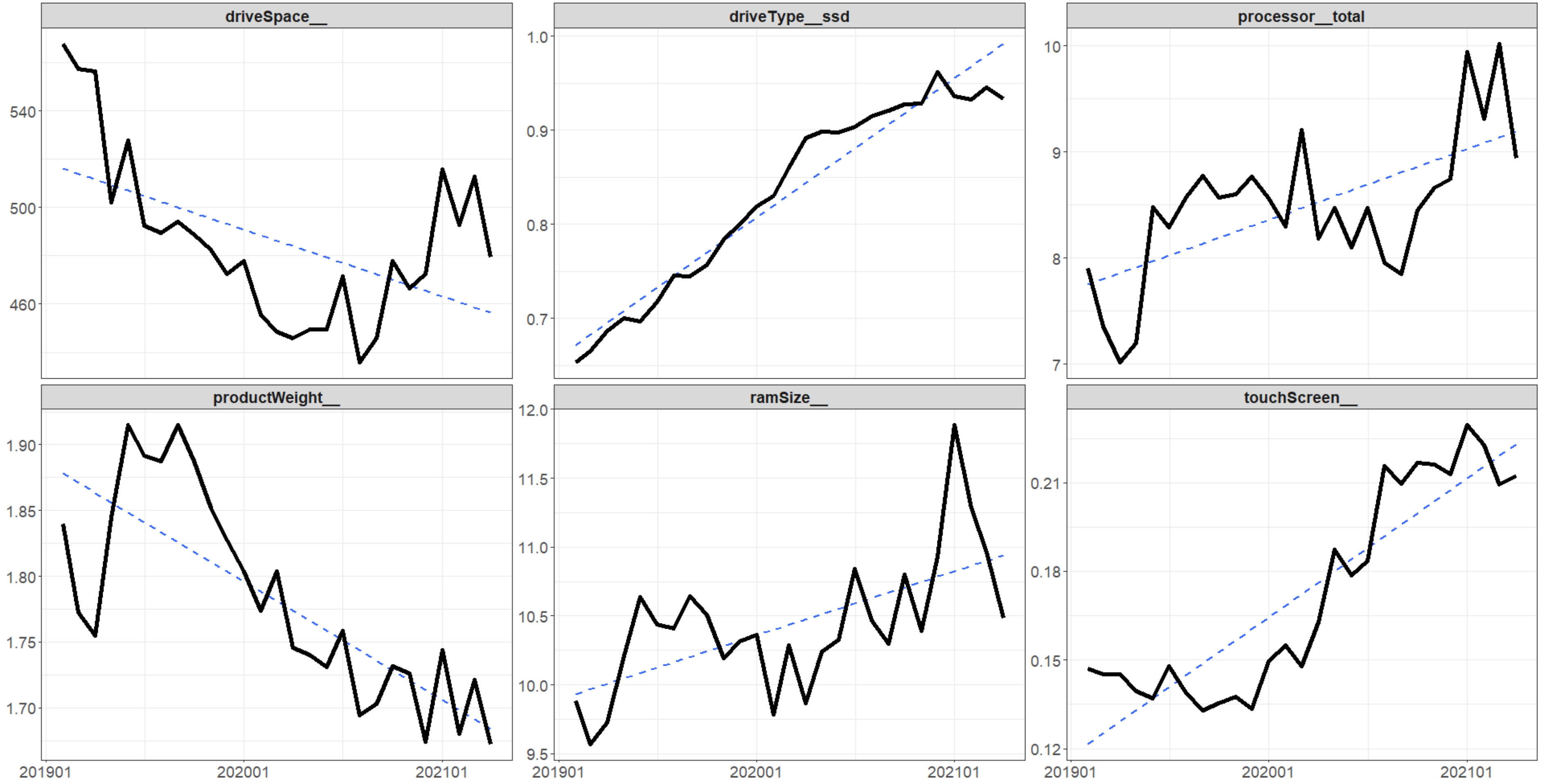

Stage 1
Extract
Stage 2
Fit to Tabular
Stage 3
Standardize Explanatory Variables
Stage 4
Final Filters, Flags, and Modelling
Stage 5
Aggregate Indices up to Product Level
Stage 6
CESSI Output and Reports
Subject matter intervention (editing, reviewing, and removing obs. within app)
Subject matter intervention (via reports)

Quality Adjustments (1/7)

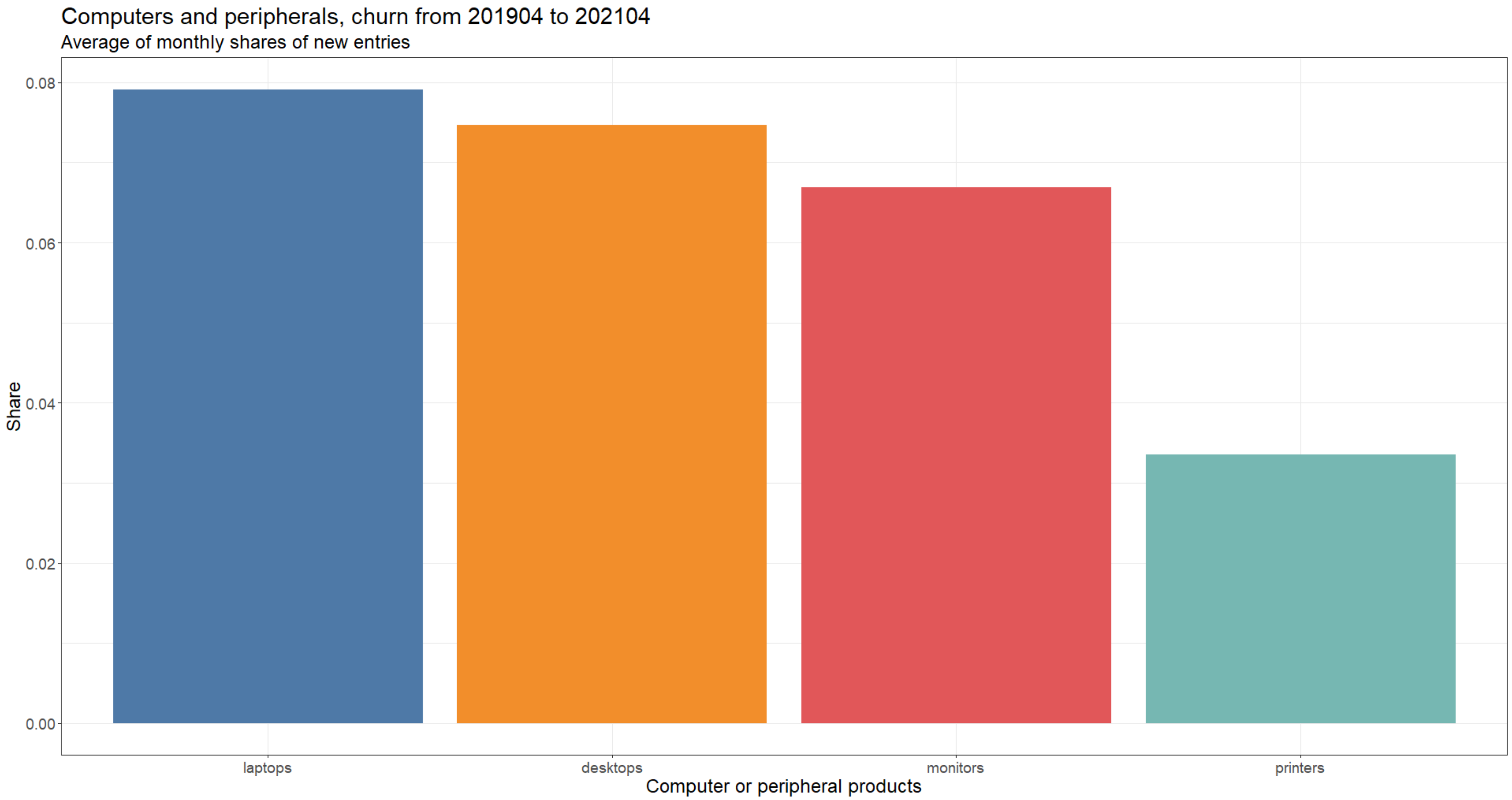
- Computer and peripheral products vary in their rate of technological change and churn
  - Laptops and desktops have a much greater churn rate, and greater pace of technological change than monitors or printers
  - Missing entering and exiting items could cause bias in an index based on only the continuities between periods
- Hence, laptops and desktops use a hedonic approach to account for missing prices of entries and exits

Quality adjustment (2/7)

Laptops, select characteristics over time



Quality adjustment (3/7)



Quality Adjustments (4/7)

- Entries and exits have missing prices imputed via a model from the corresponding period
  - i.e., an item entering in month T will have its T-1 price given by the product's model from T-1
- A random forest is used to model prices for laptops and desktops:
  - Handled outliers much better than OLS models did during testing
  - Expect less decline in performance improvement as data increases
  - In testing, the random forest had much better out of sample fits
  - In testing, resulted in a less volatile index than the corresponding regression model, with a similar trend

Quality Adjustments (5/7)

- Random forest:
  - Decision trees continually split data into subsets
  - Splits are based on whichever variable can minimize intra-class variance in the outcome of interest (log price)
  - Overfitting is countered by the use of random subsets of variables to choose from when performing each split, as well as the use of bootstrap sampling for each tree
  - A given tree's prediction for an item with a set of characteristics is the average outcome of observations in its training set that satisfied the conditions of each of the nodes above it (e.g. RAM size > 2, processor cores < 6, brand = *B*, etc.)
  - Random forest's prediction is the average of each tree's prediction i.e. $\widehat{f_{rf}(X)} = \frac{1}{M}\sum_{m}^{M}\widehat{f_m(X)}$
  - Can easily capture complexities and non-linearities without specifying the functional form
    - E.g. repeatedly splitting on different values of RAM size allows the nonlinear relation between RAM size and price to be captured
    - E.g. branches that involve splits on storage space and storage type can easily capture the interactions between the two

Quality Adjustments (6/7)

- The model fitted is $p_{it} = f_{t,product}(X_i) + \varepsilon_{it}$
  - p_{it} is log price of item i at time t
  - $f_{t,product}$ is the model for the given product (laptops or desktops) at time t

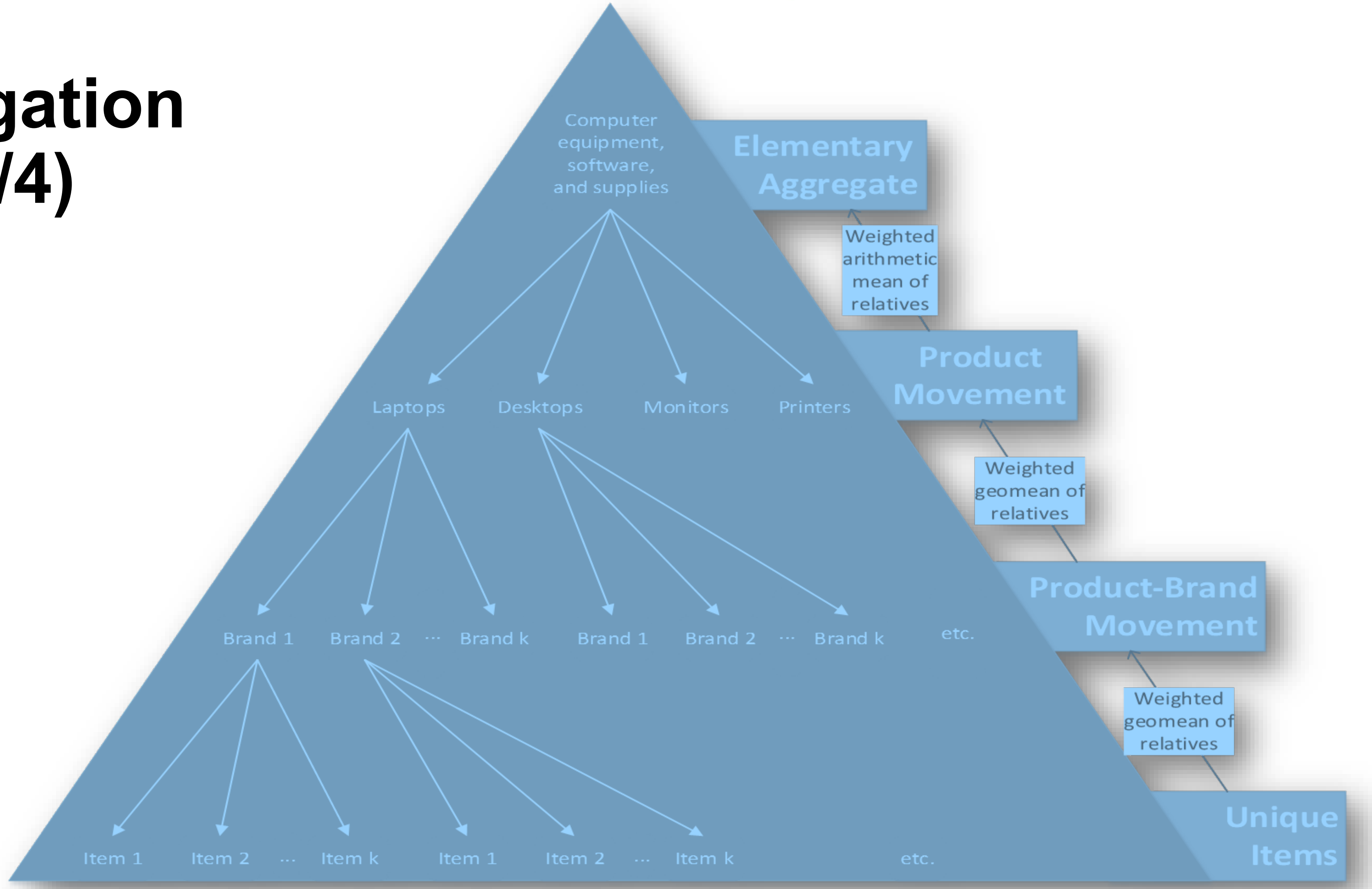
- X_i are time invariant characteristics used in modelling log price
  - Storage (space, type)
  - Memory (space, type)
  - CPU (speed, cores, brand)
  - GPU brand
  - Display (size, touch*, resolution*)
  - Weight (kg)
  - Manufacturer
  - Retailer

*Laptops only

Quality Adjustments (7/7)

- Monitors and printers use a matched model instead
  - These two products have much lower aggregation weights than laptops or desktops
  - Had lower churn rates and observed technological change
  - We found much less pronounced difference between the matched-model and hedonic imputation indices for these two products

Aggregation (1/4)



Aggregation (2/4)

$$I_{t,brand,product} = \exp\left(\sum_{i}^{n_{t,brand,product}} \Delta\,\tilde{p}_{t,i,brand,product} * w_{t,product,i,retailer}\right)$$

$$\Delta\tilde{p}_{t,i,brand,product} = \begin{cases} p_{t,i,brand,product} - p_{t-1,i,brand,product} \text{ for continuities (all products)} \\ \hat{p}_{t,i,brand,product} - p_{t-1,i,brand,product} \text{ for entering (laptops + desktops)} \\ p_{t,i,brand,product} - \hat{p}_{t-1,i,brand,product} \text{ for exiting (laptops + desktops)} \end{cases}$$

- $w_{t,product,i,retailer}$ is designed to prevent sample composition affects on the price movements by keeping retailer weight constant
- $w_{t,product,i,retailer} = \frac{s_{y-1,retailer}}{n_{t,retailer,product} * \sum_{retailer} s_{y-1,retailer}}$

Aggregation (3/4)

$$I_{t,product} = \prod_{brand} I_{t,brand,product}^{w_{t,brand,product}}$$

$$w_{t,brand,product} = \frac{s_{t-1,brand,product}}{\sum_{brand,product} s_{t-1,brand,product}}$$

$$s_{t,brand,product} = s_{t-1,brand,product} * I_{t,brand,product}$$

Aggregation (4/4)

$$I_{t,71010301} = \sum_{product} I_{t,product} * w_{t,product}$$

$$w_{t,product} = \frac{s_{t-1,product}}{\sum_{product} s_{t-1,product}}$$

$$s_{t,product} = \sum_{brand} s_{t,brand,product}$$

Note: applied nationally

Final Remarks

- New methodology tackled the previous issues of timeliness and cost, while allowing for important improvements:
  - Modelling
    - Monthly model allows better measurement of price change in a market with evolving conditions and technology
    - Nonparametric approach allows complexities of product pricing to be better captured
    - Updated inputs
  - Coverage
    - Sample size
    - Brands
    - Multiple prices per month for items to better capture price change in the period it occurs
- New retailers are in the process of being added
- Data and methods can be applied to more consumer electronics
  - Currently investigating

End

Thank you for your attention.

Questions?

Answers:

Lance.Taylor@canada.ca
Roobina.Keshishbanoosy@canada.ca