From field collection to alternative prices data at Stats NZ

Mark Colville
Frances Krsinich
14 June 2023
Introduction

- Stats NZ has been moving from prices survey data to ‘alternative data’ and the methods associated with those since 2001 for inflation measurement:
  - used cars - introduced a ‘multilateral method’ (hedonics) from 2001 (on large-scale survey data) then incorporated admin data in 2017
  - Rent price index from tenancy bonds data (2019)
  - Overseas trade index (import data for TVs and phones: 2013, customs data all imports: 2020)
- Time saved and quality improved **but** risks from bespoke systems becoming ‘black boxes’ over time
- So, after 20 years, Stats NZ is building MAP (Multilateral Application Pipeline) to generalise our production processes
Multilateral price indexes

• Traditional methods don’t work well with alternative data
  - chain drift (asymmetrical price/quantities due to sales)
  - implicit price movements associated with new products

• Over the last 20 years, significant research on multilateral methods
  - TDH, GEKS, TPD, GK, ITRYGEKS

• Stats NZ has adopted multilateral methods in production since 2001
  - used cars (2001, TDH), consumer electronics (2014, ITRYGEKS), rents (2019, TPD),
    overseas trade index (2013, 2020, TPD)

• 2019 internal review recommended consolidation of processes for both
  production and R&D
Production processes are needed in addition to the index estimation itself:

- **input diagnostics** to explore and validate source data
- **output diagnostics** to validate indexes, and compare them to previous production runs, effect of splicing on most recent movement
- **analytical measures** such as decomposition (i.e. what drives change)
- processes to identify and **deal with changes** – e.g. to coding of characteristics
Version control

Version 1.5 release

- Assets
  - Active branches
    - main [default, protected]
      - c32cafbc
    - OTI_migration
      - cec26646
    - scanner-data-testing
      - 784ceed4
    - epicTesting
      - 9196db1

- Version 1.4 release
  - c32cafbc v1.5.0 Created 2 hours ago by
Data Storage

- MAP
  - prd
    - RPIQ
      - 2022.09
      - 2022.12
      - 2023.03
        - Diagnostics
        - Files
        - Inputs
          - 01 Raw Data
          - 02 Cleaned Data
          - 03 Edited Data
        - Outputs
          - Contribution
          - Indexes
          - LoadTable
  - uat
    - RPIQ
      - 2022.09

- Table:
  - Name: `celc_results 2023-05-02.rds`
    - Date modified: 2/05/2023 4:26 pm
    - Type: RDS File
    - Size: 12 KB
  - Name: `celc_results 2023-05-03.rds`
    - Date modified: 3/05/2023 11:03 am
    - Type: RDS File
    - Size: 12 KB
  - Name: `celc_results 2023-05-04.rds`
    - Date modified: 4/05/2023 6:12 pm
    - Type: RDS File
    - Size: 12 KB

- Table:
  - Name: `countPriceNA_diagnostic 2023-05-02.csv`
    - Date modified: 2/05/2023 4:22 pm
    - Type: Microsoft Excel CSV
    - Size: 1 KB
  - Name: `countPriceNA_diagnostic 2023-05-03.csv`
    - Date modified: 3/05/2023 10:50 am
    - Type: Microsoft Excel CSV
    - Size: 1 KB
  - Name: `countPriceNA_diagnostic 2023-05-04.csv`
    - Date modified: 4/05/2023 6:08 pm
    - Type: Microsoft Excel CSV
    - Size: 1 KB
  - Name: `countPriceNull_diagnostic 2023-05-02.csv`
    - Date modified: 2/05/2023 4:22 pm
    - Type: Microsoft Excel CSV
    - Size: 1 KB
  - Name: `countPriceNull_diagnostic 2023-05-03.csv`
    - Date modified: 3/05/2023 10:59 am
    - Type: Microsoft Excel CSV
    - Size: 1 KB
  - Name: `countPriceNull_diagnostic 2023-05-04.csv`
    - Date modified: 4/05/2023 6:08 pm
    - Type: Microsoft Excel CSV
    - Size: 1 KB
Interface

System level logging
• Timestamps for each production run incl. Topic, Period, User and System Version

mcolvill
v.1.4.0 – Complete

2023-05-05 08:36:18] PRD_RPIM_2023.04
mcolvill
v.1.4.0 – Complete

“Run” level logging
• Timestamps each step of the production run

2023-04-04 11:32:21] 00 Log File Initialisation
2023-04-04 11:32:21] 01 Folder structure created
2023-04-04 11:32:21] 02 Running data ingest script

…

Thank you!

…and we welcome any questions or feedback:

mark.colville@stats.govt.nz
frances.krsinich@stats.govt.nz
Performance time

The *multilateral* R package is the index-estimating R package that sits within the wider Multilateral Application Pipeline (MAP) R-based system.

Relative processing times (in minutes) using *multilateral* within the Stats NZ environment using parallel processing (with four CPU cores) compared to standard runs (one CPU core) on two years of supermarket scanner data - approximately 50 million observations.

(Note – in this example both the GEKS-Tornqvist and TPD (time-product dummy) methods use *geomean splicing* and an *estimation window* length of 13 months).

GEKS-T 45 min (1 core), 23 min (4 cores) TPD 105 min (1 core), 36 min (4 cores)
References


Stansfield, M and F Krsinich (2022, June). *A MAP for the future of price indexes at Stats NZ* Paper presented at the 17th Ottawa Group 2022, Rome, Italy

Stansfield, M (2022) *Multilateral R package* available on the Comprehensive R Archive Network (CRAN)

Stats NZ (2014) *Measuring price change for consumer electronics using scanner data*

Stats NZ (2019a) *New methodology for rental prices in the CPI*

Stats NZ (2019b) *Overseas trade price indexes through a multilateral method*