From field collection to alternative price data at Stats NZ

Paper presented at the UNECE Expert meeting on Statistical Data Collection, Geneva
12-14 June 2023

Mark Colville, Frances Krsinich, Prices, Stats NZ
P O Box 2922
Wellington, New Zealand
info@stats.govt.nz
www.stats.govt.nz

Disclaimer
Conference papers represent the views of the authors, and do not imply commitment by Stats NZ to adopt any findings, methodologies, or recommendations. Any data analysis was carried out under the security and confidentiality provisions of the Data and Statistics Act 2022.

Liability statement
Stats NZ gives no warranty that the information or data supplied in this paper is error free. All care and diligence has been used, however, in processing, analysing, and extracting information. Statistics New Zealand will not be liable for any loss or damage suffered by customers consequent upon the use directly, or indirectly, of the information in this paper.

Citation
Colville, M and F Krsinich (2023, June). *From field collection to alternative price data at Stats NZ.* Paper presented at the UNECE Expert meeting on Statistical Data Collection, Geneva

Executive summary

Stats NZ has increasingly been using alternative data for inflation measurement over the last 20 years. In particular, scanner data was introduced into the NZ CPI for consumer electronics products in 2014, replacing both field collection and significant in-office resources spent on largely subjective quality adjustment.

Efforts to get scanner data for supermarket products were given a kickstart during the COVID pandemic and average prices for products in the CPI basket of goods have been sourced from scanner data rather than field collection since then. With the recent agreement from supermarket retailers to provide expenditure data in addition to prices for all products, we are now able to improve the index quality by using new methodology to incorporate data for all supermarket products.

Stats NZ is currently developing a generalised production process for these new methodologies, called MAP (the ‘multilateral application pipeline’) – which uses a common process after the initial data wrangling stage, with methods and parameters set specific to the data source. The production of price indexes from supermarket scanner data will join that of consumer electronics products (scanner data), used cars (survey and vehicle registration data), rents (tenancy bond data) and overseas trade indexes (customs data) in production using MAP. We will present and discuss this development in terms of its efficiency gains and futureproofing of our ability to use alternative data sources for inflation measurement.

Introduction

In this paper we give a history of Stats NZ’s[1] use of multilateral methods for alternative prices data and explain why we are now developing a generalised research and production system in R called the Multilateral Application Pipeline (MAP).

In addition to index estimation, other processes are required in production, and these need to be automated and standardised across different price indexes and data sources to aid transparency, robustness, and efficiency. These include:

1. **input diagnostics** to explore and validate source data
2. **output diagnostics** to validate the results of index estimation against those of previous periods
3. **analytical measures** such as decompositions, or ‘points effects’, to aid insights into the aggregate-level price indexes
4. processes to identify and deal with **changes in the coding of characteristics**, if those characteristics are used for explicit hedonic modelling[2], or if they are required for the creation of unique product identifiers[3]

Because Stats NZ has adopted a range of multilateral methods gradually for a number of data sources over the last 20 years, over time a range of production processes across SAS, Excel, and R, were introduced, with different levels of automation and robustness. The development of MAP

[1] Statistics NZ is now called ‘Stats NZ’.

[2] For example, in the time dummy hedonic (TDH) or Imputation Törnqvist Rolling Year GEKS (ITRYGEKS) indexes

[3] Such as required for consumer electronics scanner data where model name is masked for those products sold predominantly by one retailer, to protect the confidentiality of that retailer.

enables us to simultaneously improve our current production processes and pre-build the development and production system for future adoption of new alternative price data sources.

Since Stats NZ has started using these methods, the theory of multilateral price indexes has developed, and we are now in a position to develop a system that generalises all the production processes once the source data has been transformed into a consistent format. The appropriate index estimation can then be specified with parameters for each choice of a multilateral index method, a splicing method, and an estimation window length, with flexibility to easily change these settings in response to future theoretical findings.

Multilateral price indexes

Traditional index methods do not work well with alternative prices data[4] such as scanner data, administrative data, and web-scraped online data for two main reasons:

1. Chained superlative indexes[5] tend to exhibit 'chain drift' when frequent sales result in asymmetric volatility in prices and quantities.
2. Matched-model methods omit the implicit price movements associated with the introduction of new products.

Over the last 20 years, there has been a significant amount of research and development in this area, resulting in the adoption of multilateral index methods, such as:

- the Time Dummy Hedonic (TDH)
- the Rolling Year GEKS (RYGEKS) (Ivancic, Diewert and Fox, 2011)
- the Time Product Dummy (TPD) (ibid.) or FEWS (Krsinich, 2016)[6]
- the Imputation-Törnqvist RYGEKS (ITRYGEKS) (de Haan and Krsinich, 2014)

Evolution since 2001 at Stats NZ

Stats NZ has used alternative data and multilateral methods in the New Zealand Consumers Price Index (NZ CPI) for **used cars** from 2001; **consumer electronics** from 2014 and **housing rentals** from 2019. In the NZ Overseas Trade Index (OTI), a multilateral method was used for **mobile phones and televisions** from 2013 before being fully adopted **for all price indexes from customs data in the OTI** in 2020.

Krsinich (2014) explains the adoption of multilateral methods at Stats NZ in the wider context of the history of quality adjustment in the New Zealand Consumers Price Index (NZ CPI).

Used cars

Stats NZ first used a multilateral index in production in 2001, when a time-dummy hedonic (TDH) index was adopted as a more efficient and accurate way of estimating price change from a large-

[4] Also known as 'non-traditional data' or 'big data' in the context of price measurement, though many argue that these data sources are not strictly 'big data'. A more accurate term might be 'bigger data'.

[5] The seemingly appropriate way to estimate representative price indexes in the context of rapidly changing product universes and full-coverage data.

[6] The FEWS index explicitly combined window-splicing with a TPD index to address the systematic bias that would result from using a TPD in production for a non-revisable index such as the CPI. Now that splicing (of more than just the latest period) is recognised as an important element in the specification of multilateral methods, the distinction between TPD and FEWS is no longer required and so we will now tend to use the original term 'TPD' to refer to this method.

scale survey of all used cars sold by a sample of used-car dealers. In 2011 the hedonic formulation was improved and in 2017 administrative data on used cars' characteristics from the New Zealand Transport Authority was incorporated to reduce respondent burden.

Rental prices

In 2009 a time-product dummy (TPD) was used to benchmark the performance of the then matched-model rental index based on a longitudinal survey of landlords. Exploring the properties of this approach then motivated further research by Stats NZ into the potential of using fixed-effects (or time-product dummy) indexes with splicing more generally, for any longitudinal price data with insufficient data on product characteristics to exploit explicit hedonic methods such as the TDH.

In 2019 Stats NZ then redeveloped the rental index in production as a TPD[7] index based on tenancy bond data (Stats NZ, 2019a; Bentley, 2022).

Overseas trade indexes

In 2013 the TPD was used to estimate price indexes for mobile phones and televisions from import data in the overseas trade index then, in 2020, Stats NZ fully adopted the TPD for estimation of all price indexes from customs data for the NZ OTI (Stansfield, 2019; Stats NZ, 2019b).

Consumer electronics

In 2014 the Imputation Törnqvist Rolling Year GEKS (ITRYGEKS) (de Haan and Krsinich, 2014) index was adopted to estimate price indexes from scanner data for consumer electronics products in the NZ CPI (Stats NZ, 2014).

Stats NZ's strategy for future use of alternative price data

Bentley and Krsinich (2017) gave an overview of the potential for alternative data in the NZ CPI. Following this, in 2021 an internal review by Stats NZ recommended a strategy for the future of using alternative data in the NZ CPI. The internal report's key recommendation was that Stats NZ should pursue the development of a generalised processing system to consolidate the existing production processes and provide a solid basis for the future incorporation of alternative data sources. The paper by Stansfield and Krsinich (2021) presents some of the conclusions and empirical testing undertaken during that review.

Production processes are non-trivial

At price index conferences and in the literature, most of the focus on the use of alternative data has centred around index methodology. In particular, on the still-evolving concepts, limitations and empirical results relating to multilateral index methods.

However, the estimation of indexes is just one element of what must be dealt with when using alternative data in production. It is also crucial in the production of price indexes to understand 1. what drives aggregate price movements and 2. the impact on the most recent index movement of

[7] With a geomean splice.

the splicing procedure used[8]. Issues also arise when dealing with incomplete or inconsistent-across-time source data which, in Stats NZ's experience, is the rule rather than the exception with this data.

Many of the processes required for these insights and mitigations become non-trivial to automate at scale. The iterative development of the MAP system is therefore incorporating an automation of processes which, in the past, have involved relatively time-consuming analytical work, often at-least partially using Excel.

Stansfield and Krsinich (2022) show in more detail the implications of inconsistent coding over time of scanner data for consumer electronics products, and the need for production processes to deal with this.

Empirical testing at scale

The ability to automate and scale up both the index estimation and many of the associated production processes is also important when determining the appropriate methods for new data sources. Decisions are required about which underlying multilateral index methods to use (e.g., TDH, GEKS-T, GEKS-IT[9], TPD) and what their appropriate settings should be in terms of splicing method and estimation window length. While some methods will be better than others based on theoretical considerations, we acknowledge that the theory is still evolving. This heightens the importance of empirical testing – across methods and their parameters, and against historical series (where they exist) to help justify those decisions.

The Multilateral Application Pipeline

As already mentioned, until recently the processing of alternative data sources with multilateral methods at Stats NZ has used bespoke systems across a variety of different languages and operational systems, namely Excel, SAS, and R, with varying degrees of manual intervention required by analysts.

The earliest implemented processes, such as those for used cars and consumer electronics, are inefficient in various ways by today's standards. For example, the splicing[10] of the most recently estimated quarter's movement for used cars is done in Excel in a relatively manual way, rather than coded into the production of the index. For consumer electronics, the identification and treatment of changed coding for characteristics has been done in excel and is labour-intensive and relatively opaque without documentation of decisions and treatments incorporated into the system itself.

By centralising the process in Stats NZ's new Multilateral Application Pipeline (MAP) system, the integration of alternative data sources and multilateral methods can be consolidated and

[8] The use of splicing (where the splicing period is greater than just the latest period) trades off the quality of the most recent movement in favour of the longer term index. While this is generally a desirable property, focus is often on the most recent movement (either annual or monthly) so the NSO should understand the impact of the implicit revisioning implied by the splicing.

[9] The multilateral package refers to ITRYGEKS as GEKS-IT (GEKS Imputation Törnqvist) as a more standardised naming convention. Similarly, GEKS-T (or the CCDI) is the Rolling Year GEKS based on a Törnqvist index and GEKS-J is the GEKS based on a Jevons index.

[10] The CPI is non-revisable. Multilateral methods, however, re-estimate a back series with each successive period. This means that new results must be 'spliced' onto the published index such that they preserve the integrity of the published series (by incorporating a 'revision' factor). See de Haan (2019) for a discussion of different splicing approaches.

streamlined. Creating a centralised system also brings transparency to these complex processes and a platform upon which team members can learn, with links to documentation and instructions.

While making production processes for the existing use of alternative data more robust and transparent, this generalised system also lays much of the groundwork for future implementations of new data sources.

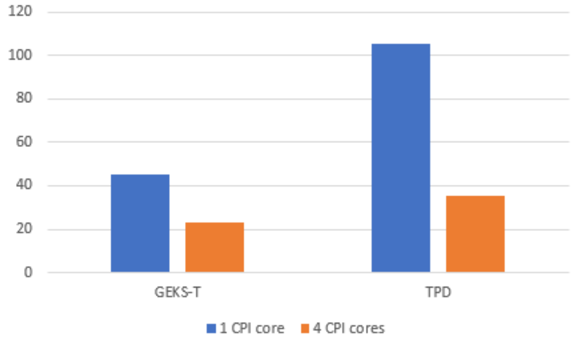
Multilateral R package for index estimation

Over the past few years, we have developed an R package for estimating all the multilateral indexes in production at Stats NZ, the *multilateral* package, which is now available at CRAN[11]. Some of the underlying functions are an implementation of the *IndexNumR* package[12] by Graham White. We have also added multilateral methods that use hedonic regression modelling, such as the time dummy hedonic (TDH) and the Imputation Törnqvist Rolling Year GEKS (ITRYGEKS[13]).

Stats NZ built our own package internally to ensure full transparency, particularly for our validation against existing SAS-based implementations, and with consideration of speed and the flexibility to change between methods and parameters easily. For speed of processing, the package allows parallel processing and optimized functions like sparse matrices and memory efficient operations. The extra hedonic regression functionality is computationally intensive and requires this optimization.

Figure 1 shows the relative processing times within the Stats NZ environment using parallel processing (with four CPU cores) compared to standard runs (one CPU core) on two years of data of approximately 50 million observations. Both the GEKS-T and TPD methods use geomean splicing and an estimation window length of 13 months.

Figure 1: The effect of parallel processing on run-time (in minutes) of the GEKS and TPD indexes



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

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

[11] Comprehensive R Archive Network https://cran.r-project.org/web/packages/multilateral/index.html

[12] https://mirrors.pku.edu.cn/cran/web/packages/IndexNumR/index.html

[13] Referred to as GEKS-IT in the *multilateral* R package.

Overview of the MAP system

The goal of the Multilateral Application Pipeline (MAP) is to be a generic system capable of consuming raw data, processing it, producing statistics, and presenting diagnostic information to the end user (internal analysts). The main use-case of MAP at Stats NZ is to calculate multilateral price indexes on a range of product categories using the "multilateral" R package (Stansfield, 2022), to validate those outputs using a variety of diagnostic measures, and then to be collated with other indexes for dissemination. The system is designed to be very user-friendly, requiring no prior coding experience, and during typical usage of MAP manual intervention should not be required. The system is operated using a simple interface of button prompts and text entry fields, resembling a stand-alone application.

Architecture

The MAP system is written in R and R Markdown built into an R package alongside a secure file storage location for data steady-states and metadata. The system uses a single high-level function to run end-to-end, and a Shiny application is included as the primary intended method for non-developers to use MAP which streamlines their interaction with the system.

Modularity

The system is designed to be flexible with functionality separated into discrete steps, including Initialization, Storage Setup, Data Ingestion, Editing and Imputation, Index Calculation, Data Export, Diagnostics and Cleanup. Individual steps can be run in isolation or repeated, such as when an error occurs, removing the need for running redundant steps.

To streamline using MAP, all indexes produced from a distinct alternative data source are bundled together into an "output". These outputs are typically aligned to a specific statistical output, such as the Rent Price Index (RPI), or a homogenous group of products such as supermarket products. Each output has a corresponding metadata file that describes calculation parameters, in addition to discrete data ingestion processes, diagnostics and output data structures.

Steady-states and version control

To maintain strict reproducibility of our statistics, data steady-states are produced during the processing of data sources. These states vary depending on the origin and specifications of the data source, but typically include the data in its raw unadjusted format, processed states before and after editing and imputation, followed by the production statistics. Each steady-state is date-time-stamped, allowing traceability of any statistical output from the system.

MAP is version-controlled using Gitlab, making use of branches to allow development of the system to occur alongside production outputs. Designated releases additionally make it easier to trace any specific statistical output, to provide documentation on changes and to simplify troubleshooting.

Documentation and Diagnostics

Due to the ability to version-control MAP, it was beneficial to incorporate documentation directly into the system. User guides and process documentation are written in R Markdown and are built directly into the Shiny application used by internal analysts running MAP.

Diagnostic reports are written in R Markdown, with default diagnostics for input data and output indexes, with the ability to create tailored diagnostics for product groups, or specific products. Desirable diagnostics include analysis of input data such as column/row count, expected variables and simple averages of key variables. More complex diagnostics such as interactive graphs of multilateral index splicing are also available, allowing visualisation of complex analysis.

Example: migrating the consumer electronics scanner data system

The original production system for consumer electronics scanner data was implemented in 2014 in SAS, with diagnostic and analysis processes largely executed in Excel. The system required manual intervention from analysts to produce and respond to diagnostic processes. At the time it was introduced, the system was well understood but with staff turnover the process has slowly turned into a 'black box' with gradual loss of understanding of the purpose underlying key steps. This has made the system quite fragile and overly dependent on a few senior technical staff to deal with ad-hoc issues.

The system usually took about four days for an analyst to run, as issues often arose requiring bespoke problem-solving. If no issues at all arose the fastest possible run (which incorporated some quite laborious semi-automated work in Excel) took approximately 4 hours.

With its migration to MAP this system can now run in less than 5 minutes end-to-end, with all reports automatically produced. The new process has required little manual intervention and is significantly simpler to maintain and interpret.

Future plans to migrate into MAP

To date, used cars, consumer electronics and rents have been migrated to MAP. The next system to migrate is overseas trade indexes (which use customs data). Although this already has its own relatively robust R-based systems, it will be rebuilt in the generalized MAP system to enable full consolidation and streamlining. Despite the overseas trade index migration not yet being complete, we are already observing a ~60% decrease in processing time based on the consolidation into MAP.

Likely future data sources to be developed in the MAP system:

- **Supermarket scanner data** is in the exploration stage for use of multilateral methods, with the testing of methods and parameters, and investigation of the raw data.
- A **prototype official house price index** able to be disaggregated into land and structure indexes, using local councils' valuation and sales data (see Krsinich, 2019)

We are also now exploring the potential to use MAP inside the NZ GS1[14] environment to produce indexes securely with release to Stats NZ of the aggregate-level indexes. This is looking very promising.

Conclusion

In addition to the methodological challenges of using alternative data for price index estimation, there are non-trivial issues associated with production at scale. Our development of the R-based Multilateral Application Pipeline (MAP) helps to automate, consolidate, and generalise these production processes.

The development of MAP has been iterative, starting with the migration of existing production systems for used cars and consumer electronics products, from SAS and Excel. More recently, the Rent Price Indexes (based on tenancy bond administrative data) were consolidated from existing R

[14] GS1 hold price and quantity information corresponding to their barcode information from a market research company, meaning that sufficient information for (non-hedonic) multilateral index (such as TPD or GEKS) methods is available within their secure environment, though not able to be released at that level of disaggregation.

systems, and we are currently migrating the R-based systems for the Overseas Trade Index (based on customs data).

We plan to develop supermarket scanner data and a prototype house price index using the MAP system, and we are currently exploring the use of MAP inside NZ GS1's secure environment to enable the safe use of confidential price and quantity data linked to barcode information.

For Stats NZ, there are multiple benefits of the MAP system:

- A reduction in manual, error-prone processes – everything that can be automated will be automated.
- More transparency, with the underlying code open for review and reuse by others.
- Diagnostics, monitoring, and analysis are incorporated alongside index estimation.
- Index estimation is done with our in-house developed *multilateral* R package, which enables the full range of multilateral methods already in production at Stats NZ, and performs well at scale through optimised functionality and parallel processing.
- Consistent interfaces and processes across product types, data sources and methods.
- The potential for incorporation of links to training and documentation in the user interface.

In addition to the multilateral package (Stansfield, 2022) the rest of MAP's R packages will be made open source and available from CRAN - we hope that other agencies and researchers will also make use of them in their research and development.

References

Bentley, A and F Krsinich (2017) *Towards a big data CPI for New Zealand* Paper presented at the 2017 Ottawa Group, Eltville, Germany

Bentley, A (2022) *Rentals for Housing: A Property Fixed-Effects Estimator of Inflation from Administrative Data* Journal of Official Statistics, 38(1)

Bentley, A and Krsinich, F (2022) *Timely Rental Price Indices for thin markets: Revisiting a chained property fixed-effects estimator* Paper presented at the 2022 Ottawa Group conference, Rome, Italy

de Haan, J and Krsinich, F (2014) *Scanner data and the treatment of quality change in nonrevisable price indexes* Journal of Business and Economic Statistics, 32(3)

de Haan, J (2019) *Rolling Year Time Dummy Indexes and the Choice of Splicing Method* Paper presented at the 2019 Ottawa Group conference, Rio de Janeiro, Brazil

Ivancic, L, W E Diewert and K J Fox (2011) *Scanner Data, Time Aggregation and the Construction of Price Indexes* Journal of Econometrics 161, 24-35

Krsinich, F (2014) *Quality Adjustment in the New Zealand Consumers Price Index* Chapter from *The New Zealand CPI at 100. History and Interpretation* Publisher: Victoria University Press. Editors: Sharleen Forbes, Antong Victorio

Krsinich, F (2016) *The FEWS index: Fixed effects with a window splice* Journal of Official Statistics 32(2)

Krsinich, F (2019) *Land prices: UNCOVERED! Extricating land price indexes from improved property price indexes for New Zealand* Paper presented at the 2019 New Zealand Association of Economists conference, Wellington, New Zealand

Stansfield, M (2019) *Import and export price indexes using fixed-effects window-splicing* Paper presented at the 2019 New Zealand Association of Economists conference, Wellington, New Zealand

Stansfield, M and F Krsinich (2021) *Bigger, better, faster: further progress in using non-traditional data to measure price inflation* Paper presented at the 2021 New Zealand Association of Economists conference, Wellington, New Zealand

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*