LESSEON AND EXPERIENCES FROM AUSTRALIA, JAPAN, AND NEW ZEALAND

INCORPORATING NON-TRADITIONAL DATA SOURCES INTO OFFICIAL STATISTICS

THE CASE OF CONSUMER PRICE INDEXES

UNITED NATIONS ESCAP

Economics and Social Commission for Asia and the Pacific
Table of contents

1. Introduction........................................................................................................................................... 2
   1.1 Background........................................................................................................................................ 2
   1.2 Purpose of this report........................................................................................................................ 3
   1.3 Traditional and non-traditional data sources for price measurement................................. 3
   1.4 Why use non-traditional data for price measurement?............................................................... 5
   1.5 Potential issues with adopting non-traditional data sources...................................................... 7
2. Case Studies............................................................................................................................................. 11
   2.1 Introduction to the case studies........................................................................................................ 11
   2.2 Australian Bureau of Statistics (ABS)............................................................................................ 12
   2.3 Statistics Bureau of Japan.................................................................................................................. 14
   3.4 Statistics New Zealand (Stats NZ).................................................................................................. 16
3. Priorities in pursuing non-traditional data sources........................................................................... 18
4. Summary lessons learnt from experiences so far............................................................................. 19
5. Conclusion.............................................................................................................................................. 24
6. References and resources...................................................................................................................... 24

Acknowledgements

This report was drafted by Peter Campion. The author would like to thank and acknowledge the contributions from the Australian Bureau of Statistics, the Statistics Bureau of Japan, and Stats NZ. A special thank you to Leigh Merrington, and Andrew Tomadini (ABS), Naruhiko Yamagata (SBJ), Fiona Smillie, and Paul Pascoe (Stats NZ), as well as Gemma Van Halderen and Tanja Sejersen (UNESCAP), who all made valuable contributions.

Photo credit for cover page

- Shutterstock ID 177686234 by Foxaon 1987
- Shutterstock ID 157804964 by Foxaon 1987
- Shutterstock ID 667543681 by Tong Patong
1. Introduction
1.1 Background

Consumer Price Indexes (CPIs) are one of the highest profile official statistics produced by any National Statistical Office (NSO). CPIs measure the changing price of the goods and services that households buy. While internationally there some differences in the concepts and methods used to compile CPIs, they are generally used for similar purposes in most countries. These purposes include: being a tool for monetary policy, an economic indicator, and being used for the indexation of contracts and government transfer payments.

The possibilities offered by scanner data have captured the imagination of price statisticians and academics for decades. Price statisticians have dreamed of what they might be able to achieve if they had access to timely and comprehensive scanner data. However, until quite recently very few NSOs have successfully implemented scanner data into their CPIs.

Statistics Netherlands has been a leader in the use of scanner data for price measurement and was one of the first to implement scanner data into their CPI. In more recent years, the Statistics Bureau of Japan, Stats NZ, and the Australian Bureau of Statistics have implemented scanner and other forms of non-traditional data sources into their CPIs.

Traditionally, the high cost of obtaining prices data meant that NSOs enjoyed a monopoly in the production of regular, nontrivial, price indexes. However, the launch of MIT’s Billion Prices Project in 2008 proved that relatively comprehensive price indexes could be produced by other organisations. The Billion Prices Project makes use of millions of prices automatically collected from retailer’s websites through a process known as web-scraping. The Billion Prices Project grew to a point where MIT are collecting millions of prices every day from websites around the world and producing daily price indexes. The success of the project resulted in the initiative being expanded into a commercial venture called PriceStats. Furthermore, in 2016, the software company Adobe launched their Digital Price Index as part of their Digital Economy project. Adobe’s indexes initially covered a smaller range of products, but grew over time to a more comprehensive range including, but not limited to; consumer electronics, food, airfares, furniture, and non-prescription medicines. The Adobe Digital Price Index made use of transactions data obtained from a selection of Adobe’s clients using their digital marketing tool.

1 Billion Prices Project
2 Adobe’s Digital Economy Project
1.2 Purpose of this report

The Asian and Pacific statistics community have a shared vision that by 2030 national statistical systems, including NSOs, are enabled and empowered to lead the development of innovative, trusted and timely statistical products and services, for the evolving statistical requirements of the 2030 Agenda for Sustainable Development. This report documents and draws on the experience of three NSOs in the Asia-Pacific region who have successfully implemented non-traditional data sources, specifically scanner data, into their CPIs. Lessons can be learnt from the experience of the Australian Bureau of Statistics, the Statistics Bureau of Japan, and Stats NZ that will help other NSOs in their efforts to incorporate similar approaches within their CPIs or other price indexes.

A great deal of work has been put into researching and developing methods that can be used to work with non-traditional data sources. This work has been well documented in many papers which have been published (e.g. on the Global Working Group on Big Data for Official Statistics website) or presented at international conferences (e.g. the UN Ottawa Group on Price Statistics), some of which have been referenced at the end of this report.

This report focuses more on the practical side of implementing non-traditional data sources into CPIs rather than on the technical side. The goal of this report is to help NSOs with decisions on whether and how non-traditional data should be implemented into their CPIs, and to help NSOs do this more efficiently. Additionally, this report might help manage the expectations of how far along with this work NSOs are and should help raise awareness of how much work is required to implement non-traditional data sources into CPIs.

1.3 Traditional and non-traditional data sources for price measurement

Steering clear of the definition of ‘big data’, this report focuses on the use of non-traditional data sources in CPIs.

Traditional data for CPIs is usually manually collected directly from retailers via personal visits, phone calls, or manually from websites. The traditional data collection methods usually result in relatively small, but well targeted, samples of products and services. Because these samples are small and relatively expensive to collect, it is often hard for these samples to reflect the impact of promotional sales and new varieties of goods and services entering the market. Traditional price data collection can also be burdensome on retailers. This is because price collection staff will be regularly visiting stores capturing the prices of goods and services. Furthermore, in situations where the product or service being sampled is no longer available,
the price collector will often need to talk to staff in the retail store to understand why that product or service is no longer available and what a suitable replacement should be.

For this report, non-traditional data sources for price measurement include scanner data, online data, and administrative data.

Scanner data is one of the best-known forms of transaction data used in CPIs. Scanner data are electronic records of transactions, usually where a barcode has been scanned. This data generally contains a weekly, or monthly, summary of all transactions for a retail chain, or location, at the most detailed (individual) product level. The data usually contains total value of sales as well as the total quantity of sales for the reference period. Using this information, unit prices can be calculated, which account for the different quantities of products sold at different prices over the reference period. As a result, CPIs using scanner data have the opportunity to better reflect the true impact of promotional sales than they would using traditional, manually collected data.

Scanner data can be obtained either directly from retailers, or in some cases it can be purchased from market research companies. However, it is good to be aware that, in some cases, data from market research companies might not be as timely as data sourced directly from retailers. Having said that the delay in getting data from market research companies could be due to them being busy adding value to the data, like aggregating some of the data which makes the volume of data received more manageable.

Online data is data that is electronically collected from retailer’s websites. For this report, online data includes price and product/service data that has been automatically collected from websites by software through a process known as web-scraping, or data collected from website’s Application Programming Interfaces (APIs). Accessing a website’s APIs allows NSOs to more efficiently access the price and product/service data that underlie the website. Web-scraping is the same process used by the MIT Billion Prices Project to collect the online data they are using in their price indexes. Accessing data from websites via APIs, when they are available and access is granted to them, can be seen as preferable to web-scraping because a web-scraping tool doesn’t need to navigate the retailer’s website hence the process for collection is more efficient and creates less traffic on the retailer’s website.

Administrative data is another important non-traditional data source that can be used for CPI. Administrative data is data that has been compiled by organisations for administrative purposes, for example, price lists or fee schedules for prescription drugs. Administrative data
has been included within many country’s CPIs for some time. In recent years, some NSOs have looked to make more use of administrative data sources to limit burden on respondents, improve the accuracy of their CPIs, or reduce the costs of data collection. As NSOs have worked to develop new methods that are more suited to large datasets, this has expanded the potential use of administrative data sets, as these methods can be applied to the administrative datasets as well.

1.4 Why use non-traditional data for price measurement?

User expectations are being influenced and changed in a number of different ways. Since the Global Financial Crisis, official statistics, especially economic and financial statistics, have faced greater scrutiny from the user community. Stakeholders are expecting more from NSOs at a time when these organisations are often being asked to operate with less funding. Additionally, because inflation has been so flat in many countries for an extended period, some users are starting to ask more detailed questions about how CPIs are constructed in an effort to understand why inflation has been so low.

These influences on user expectations have meant some users are looking for more granularity in official statistics, including CPIs. Non-traditional data sources can potentially help NSOs deliver more granular CPI series by significantly increasing sample sizes and the geographical coverage of indexes. However, when dealing with non-traditional data sources, more granular data might not always be available, especially when it comes to the geographical granularity. This is true for online data, when a retail chain might offer national pricing online, but allows individual stores to set their own prices. Also, when NSOs are negotiating for the supply of scanner data they might need to make some trade-offs with the level of granularity available within the data. For example, some retailers might only agree to supply scanner data that has been aggregated to a broad regional level rather than supplying data for specific locations. This might be done to maintain the confidentiality of individual stores or franchise partners.

User and stakeholder expectations are also being influenced by what has been achieved by initiatives like the MIT Billion Prices Project and Adobe’s Digital Price Indexes. NSOs no longer enjoy the monopoly on producing the type of information they once did. Additionally, because these initiatives are already making use of non-traditional data sources, they are constructed using a very large number of data points which in many cases dwarf that of the official CPIs. This can lead users to question the quality of the CPI and ask why similar methods are not being used for the official CPI.
Non-traditional data sources allow NSOs to increase the coverage of their CPIs by including nearly all the products sold by sampled retailers rather than just sampling a small number of products. This will significantly increase the same size, capture new products as they enter the market, and in the case of scanner data, will allow weights to be used more dynamically or updated more frequently.

While online shopping has been around for many years, it is still growing at a fast rate. Most major retail chains have a growing online presence, and more is being invested in sophisticated methods where these retailers monitor each other’s websites and adjust their prices accordingly. In an environment where prices are more dynamic and change more frequently, traditional manual data collection methods could fail to capture this and risk being unrepresentative. This means that sticking with the status quo could be both a quality and a reputational risk that NSOs need to be aware of.

CPIs tend to be some of the timeliest statistics produced by NSOs. Some have suggested that non-traditional data sources will help CPI’s become even more timely, pointing to the speed and frequency with which the Billion Prices Project produce their indexes. However, comprehensive measures of inflation cannot be published before the least timely price component of the index is captured, checked, and processed. This means that CPIs cannot really be any timelier than the data for the least timely good or service. This means that most non-traditional data sources will not help make CPIs any timelier than they are now.

Additionally, official statistics must be of fit-for-purpose quality. Given the importance of the CPI the significant policy uses of them (e.g. large sums of transfer payments indexed by them) and the non-revisable nature of most CPIs, there should always be a strong focus on the accuracy of the index. Generally, important stakeholders like central banks prioritise improvements to the accuracy and granularity of the CPI ahead of improved timeliness.

While NSOs are facing greater expectations from their users and stakeholders, technological advancements are providing them with greater opportunities to meet those demands. As retailers capture more of their transaction records electronically and are making greater use of that information for their own marketing purposes, more data sets are available now than they once were. Technological advances are also making it easier for NSOs to process and store very large datasets. Additionally, new university graduates are entering the job market with greater coding, and other data science skills, than they were in the past.
For many NSOs, limiting or reducing the burden placed on households and businesses has been an organisational priority for a long time. Some NSOs have even introduced policies where administrative and other non-traditional data sources need to be investigated and ruled out before they are allowed to develop any new traditional data collections. In many cases, the introduction of non-traditional data sources will mean NSOs are able to limit the burden on individual retailers. In addition to this, at a time when society is becoming more concerned with the environmental impacts of our actions, the use of non-traditional data sources could reduce the motor vehicle use and therefore carbon footprint involved in producing CPIs.

Furthermore, there has been a lot of research done into price index methods that can be used for different types of non-traditional data. The existence of this research, and knowledge, should significantly speed up the development and implementation of price series based on non-traditional data sources than would be the case if NSOs needed to start from scratch. This research and knowledge can be accessed through the websites of the UN Ottawa Group on Price Statistics, UN Economic Commission for Europe, University of New South Wales’s Economic Measurement Group, the UN’s Global Working Group on Big Data, and NSOs.

1.5 Potential issues with adopting non-traditional data sources

Fundamental principles of official statistics

In recognition of the important role that official statistics play in the functioning of democracy, the economy, and society, the Conference of European Statisticians developed and adopted the Fundamental Principles of Official Statistics in 1991. Since then, in the acknowledgement of the importance and relevance of these principles, they have been endorsed and adopted by the United Nations Statistical Commission and many producers of official statistics.

The ten fundamental principles are:

<table>
<thead>
<tr>
<th>Relevance, impartiality, and equal access</th>
<th>Professional standards and ethics</th>
<th>Accountability and transparency</th>
<th>Prevention of misuse</th>
<th>Sources of official statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confidentiality</td>
<td>Legislation</td>
<td>National coordination</td>
<td>Use of internal standards</td>
<td>International cooperation</td>
</tr>
</tbody>
</table>

All these principles are relevant to the introduction and use of non-traditional data in CPIs, with some being especially relevant.
Relevance, impartiality, and equal access

The use of non-traditional data sources can help improve the quality and relevance of CPIs. Quality improvement has been the motivation behind the introduction of these data sources for most NSOs. Given that non-traditional data can be challenging to obtain there is potentially the need for NSOs to partner with private businesses to obtain the desired data. In doing so NSOs need to take due diligence in selecting businesses to work with as well as being transparent in their decisions. NSOs need to avoid any potential pressure from businesses who might want to put their interests ahead of the public interest. It has been suggested that non-traditional data could be crosschecked to help ensure that it remains free from manipulation.

Professional standards and ethics

When implementing non-traditional data into the CPI, NSOs should ensure that they are using an appropriate price index method. There are a large number of papers written on developing methods to use with non-traditional data sources which can be easily accessed. Any NSO looking to implement this type of data for the first time would be advised to talk with other NSOs who have done so successfully.

Accountability and transparency

Given the high-profile nature of CPIs NSOs will always want to be transparent about the sources and methods used to compile the index. As some of the methods developed for use with non-traditional data are less straightforward than traditional index methods, NSOs would be well served to provide users with both technical explanations and explanations in relatively simple language. Additionally, as the NSOs examined as part of this report have done, publishing empirical studies that demonstrate the impact that the new data sources and methods would have had to the official CPI prior to implementation is recommended.

Sources of official statistics

Fundamental Principle five says that NSOs should choose their data sources with regard to quality, timeliness, costs and the burden on respondents. These are some of the reasons why NSOs are choosing to use non-traditional data and will continue to make their selection of different sources across the CPI basket based on these criteria.

Confidentiality

As discussed in more than one section of this report, confidentiality needs to be a key consideration in the use of non-traditional data sources. When data is obtained on the scale
that non-traditional data sources bring, especially scanner data, it can give greater insights to the operation of retailers. This means that retailers will consider the data to be far more commercially sensitive than traditional CPI data. Data providers need to be assured that their data will be kept confidential and will only be used for statistical purposes. Additionally, if data providers are providing NSOs with data where specific customers can be identified, then the data provider should inform the NSO whether their customers are aware that data about them is being delivered to the NSO.

**Legislation**

Legislation plays an important role in the operation of nearly every NSO. Legislation helps the NSO operate while also protecting respondents in terms of the need to keep data confidential and secure. In this respect, legislation can help NSOs obtain non-traditional data. However, in many countries, the statistical legislation has not been fully reviewed for some time and might be a little out of date when it comes to the new types of data that are available. This means that in some cases a review of the legislation might be required. Examples if this might include explicit sections that cover web-scraping or the use of cloud computing. Most countries statistical legislation was written at a time before web-scraping or cloud computing were invented.

**International cooperation**

International cooperation has been a key reason why countries like Australia and New Zealand have been able to implement changes to their CPIs to include non-traditional data sources. This report goes some way into sharing the experiences of three NSOs, but there is far more to be learnt from other countries and more cooperation and collaboration should be encouraged.

**Building user engagement and social license**

When making changes to the way that official statistics are produced, it is important to be transparent about the changes. Given the importance of the CPI, it is especially important that the user community trust the index and the changes that are made to it. Introducing non-traditional data sources into the index, especially if combined with changes in methodology (e.g. the introduction of multilateral methods), can represent a significant change that could potentially concern some stakeholders. Introducing non-traditional data into the CPI therefore will require significant effort to engage stakeholders to ensure that they understand the changes that are being proposed, the reasons for the changes, and the timing of the changes and so that any concerns that they might have are addressed.
The central bank will have an important role to play here as the credibility of the CPI is a key element to the functioning of monetary policy in most countries.

There are a number of ways that NSOs can build user engagement and social license. The first is being transparent with the sources and methods that are used to compile the index. Papers and notes to users can be written to discuss the NSOs intentions and empirically demonstrate the impact that the introduction of these measures might have. Additionally, as the ABS did, NSOs can seek endorsement from external experts, which can be reassuring to the user community.

**Continuity of data supply**

As one of the most important statistics produced by any NSO, the timely and accurate publication of the CPI is vitally important. This means that any NSO looking to incorporate non-traditional sources of data into their CPI needs to be very confident that there are not going to be any major disruptions or delays to the supply of the data.

As using non-traditional data sources inevitably will mean the NSO will have less control over data collection process than they would have with traditional data, they need to be prepared for the risks this could bring. Contingency plans should be in place to ensure that the CPI can still be published to an acceptable level of quality should data supply be interrupted. What the acceptable level of quality means in practice will be situation specific and should be considered as part of the NSO’s overall business continuity planning.

In situations where NSOs have contracted with market research companies for the supply of scanner data, this form of supply could be seen as relatively safe. However, the NSO will want to ensure that their contracts extend well into the future because renegotiating contracts on an annual basis can bring additional risk. When NSOs are obtaining scanner data directly from retailers, some form of agreement (e.g. a memorandum of understanding (MoU)) is likely to be in place between the NSO and the retailer. However, that agreement is unlikely to be a legal (contractual) agreement, unless there is some form of consideration from the NSO (potentially in the form of some tailored statistical product). This means that the retailer could choose to stop supplying data to the NSO at any point. In this situation it will be important to ensure that there is a good relationship between the NSO and the retailer.

When it comes to the use of online data, the supply, or availability, of online data can be more inconsistent and therefore riskier. A lot of websites have tools attached to them that block web-
scrapers. This is done for security reasons and so the performance of the website does not suffer because of the unwanted web-scraping traffic. Even if an NSO has explicit permission to scrape the retailer’s website the performance of the web-scraping tools needs to be constantly monitored as retailers regularly install additional security measures to their websites.

**Data quality**

One concern with the introduction of non-traditional data into CPIs is how to ensure the quality of the data being used in the index. Because non-traditional data sets are usually much larger than traditional data sets, the idea of quality assurance can be a little daunting to NSOs. Some of the quality assurance methods used with traditionally sourced data will still be relevant, for example, analysing higher-level movements as well as looking at the contributors to, and distributions of, change. However, other methods need to be employed with non-traditional datasets. Some of these might include checking the characteristics of the datasets (e.g. number of observations, proportion of products changing price) to ensure that they are similar to the previous period. Additionally, derived timeseries can be analysed (e.g. total volumes of products sold each period). The specific methods available to ensure that the data is of acceptable will be situation specific and will require some critical thinking.

**New methods are needed**

Traditional index calculation formula/methods are based on fixed basket concepts and don’t lend themselves well to making the most of non-traditional data sources. If NSOs are looking to make full use of non-traditional price data for their CPI, it is more than likely they will need to use a multilateral price index formula in a similar way to how the Australian Bureau of Statistics and Stats NZ have.

2. **Case Studies**

2.1 **Introduction to the case studies**

Three case studies have been included below, where experiences from the Australian Bureau of Statistics, the Statistics Bureau of Japan, and Stats NZ have been summarized. These three organisations were selected as each of them has made good progress incorporating non-traditional data sources within their CPIs. While all three NSOs have shared their experiences through forums like the UN Ottawa City Group on Price Statistics meetings, each country has had a different experience. Information was gathered from each organisation through a combination of email exchanges, telephone interviews, and in person meetings.
2.2 Australian Bureau of Statistics (ABS)

Timeline

2011-12: First obtained scanner data
2014: Implemented some scanner data using replacement approach
2015-17: Explored the use of multilateral price index methods
2016: Started web-scrapping
2017: Implemented multilateral scanner indexes and some webscraped data
2018: Started experimenting with multilateral methods for online clothing data

Background

The ABS started exploring the use of non-traditional data sources, both scanner and online data, in earnest around 2011 with the receipt of their first set of scanner data from an Australian retailer. Before using the scanner data in the production of Australia’s official CPI (from 2014), the ABS undertook extensive testing and experimentation. During this process the ABS collaborated with other NSOs (e.g. Stats NZ, Statistics Netherlands), academics, and other price index experts on how to maximise the use of the scanner data, particularly what methods and price index formula should be used to make full use of the scanner data.

Following extensive testing and experimentation, in 2014 scanner data was introduced into Australia’s official CPI. The ABS took a phased approach to the introduction of scanner data into their CPI. Initially, a selection of prices for products already tracked by traditional data sources were replaced with unit prices extracted from the scanner data. The unit prices were calculated by dividing the revenue generated by the product by the quantity sold for the reference period. This allowed the ABS to introduce scanner data into the CPI in a relatively risk-free way, while improving the quality of the CPI at the same time.

Following the initial introduction of scanner data into the CPI in 2014, the ABS explored the use of multilateral price index calculation methods to make fuller use of the scanner data. This work culminated in the implementation of CPI sub-indexes for some food, tobacco, and cleaning products using scanner data and the RYGEKS-Törnqvist index formula with mean splicing\(^3\) in 2017.

\(^3\) Australian Bureau of Statistics
In parallel with the work using scanner data, the ABS also explored obtaining and using online data using web-scraping techniques. In 2017, CPI sub-indexes for clothing, hardware, alcoholic beverages, and small appliances making use of web-scraped data (but with no change to index formula) were implemented in the CPI. In addition, sub-indexes for some health services, childcare, and insurance services that make use of administrative data were also implemented.

The ABS continues to develop multilateral methods and processes to better measure price change for clothing using web-scraped data but has not introduced these new methods into the clothing sub-index yet.

Currently, about 40 percent of the of the ABS’s CPI basket (by weight) is represented using price movements calculated from non-traditional data sources. Of that 40 percent, about half uses scanner data.

**Obtaining the data**

Obtaining scanner data from retailers required significant time and effort. This required time from senior staff and assurances that the data would be kept confidential and secure. A good relationship with the retailers needed to be built and nurtured given the commercially sensitive nature of the data. This relationship included signing deeds of confidentiality which require ABS not to divulge the name of the data supplier. Additionally, in one case the ABS needed to negotiate a memorandum of understanding (MoU) with a retailer, which limited the ABS’s use of their data to three specific statistical programs (CPI, Retail Trade, and National Accounts).

For the online data, the ABS obtain price and product information from a number of retailer’s websites through web-scraping. The ABS has chosen to undertake the web-scraping themselves, rather than outsourcing it. This decision was driven by both the relative cost of outsourcing and because the ABS would retain more control over the process. The ABS places significant effort into monitoring and maintaining the web-scraping tool that they developed; however, they are looking at moving over to a privately sourced tool.

**Communication and collaboration**

Throughout the development and implementation process, the ABS undertook significant consultation on proposed changes to the Australian CPI. They worked closely with international experts, academics, other NSOs, and the UN Ottawa City Group meeting of price
index experts. The ABS also held workshops with key stakeholders, including the central bank, to discuss the proposed use of non-traditional data sources and the implementation of multilateral methods in the CPI.

The ABS published a range of information papers as they have progressed their work with non-traditional data sources. An information paper was published in 2013 ahead of their initial implementation of scanner data in 2014. A roadmap to enhancing the CPI, and an information paper on making greater use of transaction data, were published in 2016. In mid-2017, the ABS published an information paper letting users know of their intention to implement methods that would maximize the use of non-traditional data in the CPI from the December 2017 quarter CPI onwards. The paper explained the multilateral index formula that would be used for a selection of CPI sub-indexes using non-traditional data sources. The paper also presented three years’ worth of data to empirically demonstrate the differences between the existing (old) and proposed methods for selected products. Also included within this paper was a review and endorsement of two well-known price index experts: Professor Jan de Hann of Statistics Netherlands and the Delft University of Technology, Netherlands, and Professor Kevin Fox of the University of New South Wales, Australia.

2.3 Statistics Bureau of Japan

Timeline

- **2000**: Implemented scanner data for digital cameras using a hedonic model
- **2003**: Implemented scanner data for computers using a hedonic model
- **IN 2020**: Scheduled to implement online (web-scrapped) data for airfares, overseas package tours, and hotel charges in 2020
- **IN 2020**: Scheduled to implement scanner data for TVs, computer printers, and video recorders in 2020
Summary

The Statistics Bureau of Japan (SBJ) was one of the first NSOs to implement scanner data into their CPI.

In 2000, the SBJ implemented a hedonic model that makes use of scanner data for their measure of price change for personal computers. Subsequently, in 2003, the SBJ implemented a hedonic model for digital cameras. For both of these measures, the SBJ purchase their scanner data from a market research company.

The SBJ plan to implement CPI sub-indexes that make use of scanner data for televisions, computer printers, and video recorders in 2020. They plan to use a hedonic model to calculate price change for televisions. The SBJ are also examining whether scanner data will be suitable for the calculation of price change for mobile phones.

SBJ source their scanner data from a market research company.

The SBJ has also been trialling web-scraping as a way to obtain online price and service data for airfares, overseas package tours, and hotel charges. The SBJ has chosen to outsource the web-scraping to a private company. There are a large number of private companies offering web-scraping services in Japan and the SBJ chose to take this option as they felt that it would be more efficient than undertaking the web-scraping themselves. The SBJ plan to implement new measures based on the online data in 2020. They don’t, however, plan to make use of multilateral methods at that point. They will be closely monitoring the experience and research from other NSOs before making a decision regarding multilateral methods.

Communication and collaboration

Prior to the implementation of scanner data in the CPI, the SBJ met with external price index experts to discuss ideas and opinions on the use of scanner data in CPIs. They also engaged with international experts through regular UN Ottawa City Group meetings and UNECE meetings of experts on consumer price indices.

The SBJ publish a methodology guide on their website that explains the methods used to calculate price change for the CPI, including a specific section that explains how price change is measured for product categories that are using scanner data.
3.4 **Statistics New Zealand (Stats NZ)**

**Timeline**

- **1999**: Implemented a hedonic model for used cars
- **2011**: Implemented scanner data for consumer electronics
- **2014**: Improved the used car hedonic model
- **2017**: Implemented online data for accommodation and transport services
- **2019**: Implemented administrative data (multilateral index formula) for rent

**Background**

Stats NZ has been making use of administrative transaction data to measure the price change for used cars since 1999. This transaction data is obtained from a range of used car dealers from across New Zealand. The data includes a number of important details on the characteristics of every car that is sold from the sampled dealers. These characteristic details/variables along with the transaction price feed into a hedonic model which estimates price change for the used car category of the CPI.

From 2011, Stats NZ worked with academics, other NSOs, and price index experts to develop methods for measuring price change for goods and services using scanner data. One of the first categories Stats NZ focused on was consumer electronics products. These products were targeted as they experience significant technological changes over relatively short periods of time, making them challenging to measure price change for using traditional data and methods. These products were also selected because consolidated scanner data was available from a market research company.

Three years later, in 2014, Stats NZ implemented scanner data for a range of consumer electronics products into the official New Zealand CPI and at the same time chose to make use of the Imputation Törnqvist Rolling Year GEKS (ITRYGEKS) index formula method, which is a sophisticated hedonic measure that was developed in collaboration with Statistics
Netherlands⁴. This method was selected because it accounts for quality change in the index as well as the impact that products with high churn and volatile prices have on the index. The scanner data that Stats NZ uses includes the total monthly values and quantities of products sold for a comprehensive selection of electronics retailers. The data also includes a range of important characteristics for each product which feed into the ITRYGEKS model.

In 2017, Stats NZ implemented web-scraped online data for some accommodation and transport services into their CPI. Stats NZ developed their own web-scraping capability and carry out the collection themselves. They chose to do this because they would have more control over the process and would be able to easily make adjustments when needed. However, the web-scraping process does require monitoring and for frequent adjustment to ensure the successful collection of the desired price data. Stats NZ consider this workload to be acceptable and are looking to expand their web-scraping capabilities.

In 2019, Stats NZ implemented CPI sub-indexes for rental accommodation based on administrative (tenancy bond) data, collected by another government department. The method used with this data is the Fixed Effects Window Splice (FEWS) formula, which was developed to be used with both scanner and online data. The implementation of the administrative data and this new method helped Stats NZ improve the measure for rental accommodation while reducing respondent burden by ending a specific rent survey.

Obtaining the data

Stats NZ purchases the data on consumer electronics from a private market research company. While Stats NZ is purchasing data from the market research company, some negotiation was required to ensure the right data was delivered on a timely basis and at an acceptable price.

Stats NZ is planning to make more use of web-scraped online prices. These online prices will be used to replace other traditionally collected prices. Before starting to web-scrape from a given website, they notify the website owner (the retailer) of their intention to do so. Website owners are reassured that they will be efficient in their actions and invites the owner to get in contact if they have any concerns.

⁴ Krsinich (2015)
Communication and collaboration

Stats NZ undertook extensive collaboration with Statistics Netherlands, international experts, academics, and other NSOs when developing the ITRYGEKS method. Prior to implementing these measures, Stats NZ presented papers at the UN Ottawa City Group, UNECE Meeting of the Group of Experts on Consumer Price Indices, the Economic Measurement Group (hosted by the University of New South Wales Business School), and the New Zealand Association of Economists Conference. A similar process was undertaken prior to the implementation of the new data source and method for rental accommodation.

In 2013 Stats NZ convened a CPI advisory committee. The committee was comprised of a range of representatives from the CPI user community and also included academics and an international expert on CPIs. One issue the committee was asked to consider was the introduction of scanner data measures in the CPI. The committee were very supportive of the initiative and endorsed the approach that Stats NZ was taking.

Stats NZ publishes information papers prior to the implementation of new methods. Additionally, Stats NZ works hard to maintain a good relationship with, and the trust of, key stakeholders. They do this by meeting with stakeholders, including the central bank, on a regular basis to discuss methods, and developments. They also make information about how the index is compiled available to stakeholders though methodological papers and reports. This transparency helps Stats NZ maintain the trust of their stakeholders. This trust, in turn, makes it much easier for them to implement changes into the CPI.

3. Priorities in pursuing non-traditional data sources

While administrators and price statisticians have been very keen on including non-traditional data sources for many years, until recently, only a few NSOs have successfully implemented this type of data in a meaningful way. This lack of success has not come without trying. The reality is that getting hold of the right data and making a substantial change to a key official statistic, the CPI, can be challenging and requires patience, relationship building, and negotiation.

Every potential non-traditional data source will have its own unique characteristics and potential issues to overcome. A great deal of work has been done to develop price index formula and methods which can be applied with non-traditional data sources in different scenarios, so that NSOs can deal with trade-offs when required.
When NSOs look where they can make use of non-traditional data sources it might be tempting to prioritise food or other categories that have relatively consolidated markets and large manual data collection costs. This prioritisation makes sense as there are potentially the biggest savings to be made by perusing non-traditional data in these areas. However, these categories can also be some of the most challenging to obtain data for. NSOs should also look to be opportunistic when they start to pursue non-traditional data sources, balancing categories that strategically make sense against those that are more achievable.

Another factor that needs to be considered is the type of index methods that the NSO is going to be comfortable implementing. A lot of work has been done developing methods for using non-traditional data for products like food and consumer electronics. Methods that maximise the use of non-traditional data sources for clothing, however, are still very much in development.

4. Summary lessons learnt from experiences so far

4.1 Obtaining data

Obtaining scanner data directly from retailers (or retail chains) can be challenging. Retailers are not necessarily going to be keen to handover their scanner data to NSOs. This is because they often consider the data to be commercially sensitive and providing it to the NSO presents them with a risk that they might consider to be unnecessary. NSOs will often need to demonstrate that they have secure systems and processes to ensure the confidentiality and security of data.

One option that is starting to be explored more and more is the idea of offering potential suppliers of scanner data something in return for their data. This might be in the form of a customized data product or a training session on how the respondent might be able to better use statistics produced by the NSO in their business.

Obtaining scanner data from market research companies can also prove to be challenging. Market research companies often deal with private companies looking for information to help them maximize profits. This is a very different situation to an NSO looking for data for the CPI. Because of this, NSOs can sometimes find that data from market research companies can be prohibitively expensive. Sometimes market research companies might offer NSOs a discount provided that their company is sighted as the source of the data. Additionally, with scanner data, trade-offs might need to be made to ensure that the NSO receives the data they can use with enough time to quality assure and process it. Trade-offs might include receiving data that
is more aggregated than would be ideal or receiving data that does not capture the full reference period.

Obtaining online data using web-scraping tools can bring its own challenges as web-scraping can be disrupted for a number of reasons. For instance, websites will often have their own security measures to help protect the website’s performance and ensure the security of online transactions. These security measures might block visitors who appear to be web-scraping, or they could limit the speed that the website can be navigated which will slow the web-scraping process. Even in situations where the retailer (website owner) has explicitly granted the NSO permission to scrape their website there is no guarantee that the NSO will not be automatically blocked. NSOs will need to make decisions on how frequently they scrape from websites. More frequent scraping gives the CPI a better chance at capturing the prices that households are paying, but more frequent scraping can put additional burden on the website and will increase data storage and processing requirements. Whatever the NSO decides, it would make sense to build some redundancy into the process to mitigate the risk of interruptions.

NSOs should always notify website owners letting them know of their intention to scrape their website. If they choose not to, the NSO could face some reputational risks because the website owner will be able to see that the NSO’s IP address has been accessing their website regularly.

None of the three NSOs featured in this report required changes to be made to their statistical legislation to use non-traditional data in their official CPI. However, other countries might find the strengthening of statistical legislation is required to obtain some forms of non-traditional data. Additionally, most statistical acts were written quite some time ago and might need updating so they are relevant in the current environment.

4.2 Classification of products and services

As NSOs experiment with non-traditional data sources, one of the major issues they are facing is the classification of the goods and services contained in the data. To be useful to the NSO, the goods and services included within non-traditional data need to be classified to categories within their respective CPI classification systems. Scanner, online, and administrative datasets can include the retailer’s own categories. However, these do not necessarily align with the classification systems used by NSOs.
NSOs have experimented with a number of approaches for aligning a retailer’s classification system with statistical classification systems using some form of text recognition. However, the most promising option appears to be the development of supervised machine learning tools, like the ones used by the Billion Prices Project.

4.3 Hybrid approaches are needed

When NSOs pursue non-traditional data sources they need to be aware that there is no one solution that will work for all the selected products across the CPI basket. It is highly likely that different types of data will be available for different products or services, for example, scanner data might be available for some products, where online API data could be available for some services and there will be other products and services where non-traditional data sources will not be available at all.

Given that each potential data provider (retailer, market research company, website, or government agency) could provide different types of data or data containing different variables, different approaches (data sources and methods) will be required for each of the different subcomponents of the CPI.

4.4 New methods and skills are required

For NSOs to realize the full benefits from using non-traditional data sources, they will need to employ new price index methods. This is because the traditional price index methods do not work with such large datasets, especially when lower-level expenditure weights are allowed to dynamically adjust.

A great deal of work has been put into developing new multilateral price index methods that work well with large non-traditional data sources. Collaboration between NSOs, academics, and price index experts was a key element to the successful development of these new methods. As NSOs and others look to develop or implement new methods, there is still a lot to be gained from collaboration and learning from others’ experiences. The potential for learning should extend well beyond what can be taken from the largely technical papers explaining how these methods work.

The new multilateral methods that have been developed for non-traditional data sources are more complicated than the traditional price index methods. This means that NSOs will require a new set of skills to be able to develop, implement, run, and maintain these new index methods. Data science skills are becoming much more important to CPI development and
compilation teams. However, it is not just data science skills that are in demand, NSOs are saying that they need people who have a broad understanding of economic and price index concepts as well as data science. This means additional training of new and existing staff is nearly always required.

Requiring more skilled and specialized staff also increases the risks associated with staff turnover. Specialized staff are harder and more expensive to replace. This means that the risks associated with staff turnover are greater than they once were. One way that NSOs are trying to mitigate this risk is by trying to standardize the new tools and methods as much as possible.

4.5 The need to build user engagement and trust

As with all official statistics, significant data and methodological changes need to be implemented in a way where the confidence of users, and other stakeholders, is maintained. The circumstances that any NSO faces when doing this will be unique and specific to that NSO.

When a country’s statistical office is trusted, implementing changes to data sources and methodologies is an easier task. NSOs will have often built this trust through consistency and transparency, accordingly the NSO should endeavour to be transparent about their intentions to make changes to the data sources and methodology to the CPI and invite questions and feedback before the implementation of these changes. If trust is lost it can be hard to regain.

However, in some countries, the user communities are more sceptical of official statistics and changes being made to them. This can be especially true when it comes to CPIs, given the large values of social transfer payments that are indexed by the CPI and given that the public, who are usually more sensitive to price increases than decreases, often perceive the CPI is understating price change. In this situation, the NSO will need to be more proactive in engaging users and explaining any proposed changes. Either way, the NSO will want to explain the changes that are taking place to the index in both a technical and non-technical way and present some empirical results. However, in doing this, the NSO wants to be careful not to discredit the existing measure but explain that they are doing the best they can with the existing data.
4.6 Costs

One reason that NSOs look to incorporate non-traditional data into their CPIs is to save on manual and person-intensive data collection costs. Any replacement of manual price collection should result in lower manual collection costs. Probably the greatest opportunity for savings would come from replacing manually collected food prices with non-traditional data sources. This comes about because food prices tend to be more volatile than prices for most other products, often resulting in NSOs collecting food prices more frequently. Additionally, as there are so many different varieties of food, there tend to be more prices collected for food products relative to their expenditure weight than other products in the CPI basket.

It should be noted that the potential savings from replacing manual price collection, with non-traditional data, for non-food products is likely to result in less savings than are possible with food data.

A lot of work has been done in developing price index methods for calculating price change using non-traditional data. The development costs, however, are not likely to be trivial, especially for major categories within the CPI. The development will take the time and effort of skilled analysts. There is also the need to engage in communications activities to ensure that key users are well informed of the proposed changes before they are made. Furthermore, while incorporating non-traditional data sources in the CPI should cut data collection costs, these savings are likely to be partly offset by other increased costs. Increased costs can come from; additional analysis that might be required, additional data storage needs, the need for the additional training of staff, the potential need for a secure data transfer system, and the potential purchase of the data.

4.7 Data governance

The introduction of non-traditional data into the CPI can bring data governance issues that NSOs will need to take into account. One of these issues comes about because the non-traditional datasets can be extremely large. They can be so large that they put a strain on the NSOs existing data storage infrastructure. Investment decisions will then need to be made and options like cloud computing could become attractive.

While cloud computing offers some appealing benefits like easier access to greater storage capacity and potential cost savings, there can be downsides to cloud computing like reduced data security. Additionally, cloud computing might not fit with the current statistical legislation, or government policies, so it might not be an option for some.
There could also be issues over who can access the data provided to the NSO. While data confidentiality and security are taken very seriously by any NSO, due to the nature of some non-traditional data sources, additional pressure might be placed on the organisation. There have been instances where companies have supplied NSOs with scanner data but have requested that the data is only used for the CPI, or other instances where potential data suppliers have been scared off because the NSO offers researchers access to anonymised unit record data.

5. Conclusion

A number of NSOs in the Asia Pacific region have successfully implemented non-traditional data and methods into their CPIs. NSOs from Australia, Japan, and New Zealand have successfully invested in developing processes and methods to obtain and use this new type of data, while other NSOs are at different points in their development journeys. Collaboration and knowledge sharing have been key elements to the success of the three NSOs featured in this report.

While every country and potential data source will have their own challenges, this report has discussed some of the issues that need to be considered and/or addressed before these new data sources, and in most cases methods, are implemented in official CPIs.

6. References and resources

- The UN Ottawa Group - International Working Group on Price Indices
- UNECE/ILO - Meeting of the Group of Experts on Consumer Price Indices
- UN Global Working Group on Big Data
- UNSW Centre for Applied Economic Research – Economic Measurement Group
- ILO manual on price statistics
- PriceStats
- Billion Prices Project


Australian Bureau of Statistics (2016) *Making greater use of transactions data to compile the consumer price index*, Information Paper, Canberra

Australian Bureau of Statistics (2017) *An implementation plan to maximise the use of transactions data in the CPI*, Information paper, Canberra


Diewert, E & K Fox (2017) *Substitution bias in multilateral methods for CPI construction using scanner data*, Ottawa Group 2017 (and presentation)

Fast company (2016) *E-commerce gets its own consumer price index* (Abobe price index)


International Monetary Fund (2019) Update of the ILO consumer price index manual. Scanner data chapter 10

The Japan Times (2019) Bank of Japan considers using online data for price statistics

van Kints, M (2017) Enhancing the Australian CPI: making greater use of transactions data to compile the Australian CPI, Ottawa Group 2017 (and presentation)


van Kints, M (2017) Measuring price change for new dwellings (poster), Ottawa Group 2017

Krsinich, F (2015) Price indexes from online data using the fixed-effects window-splice (FEWS) index, Ottawa Group 2015


Lamboray, C & Krsinich, F (2015) A modification of the GEKS index when product turnover is high, Ottawa Group 2015

Merrington, L (2018) Continuing to unlock the potential of new and existing data sources, IAOS/OECD Conference 2018


Pascoe, P & Smillie, F (2019) Creating a digital Food Price Index from Web-Scraped Data, Ottawa Group 2019

Shinozaki, K (2017) Webscraped prices comparison method adopted for the PPI in Japan (poster), Ottawa Group 2017


Tomadini, A & Ratnayake, R (2019) Opportunities and challenges: Leveraging alternate data sources to compile the CPI Rents Index (poster), Ottawa Group 2019

Tomadini, A (2019) Experimental clothing indexes using Australian web scraped data Ottawa Group 2019


UNESCAP (2017) Innovative Big Data Approaches for Capturing and Analyzing Data to Monitor and Achieve the SDGs.

UN GWG Project Inventory (2015). Using web-scraping price data for price index of e-commerce

UN GWG Project Inventory (2015). Web scraping and scanner data for price indexes (Japan)

Watanabe, T (2016) Nowcasting the BOJ's Core-Core Inflation Rate Based on the Nikkei CPINow-S.
