Chapter on scanner data (retail transaction data) from the discussion paper for the New Zealand CPI revision advisory committee 2013.

Frances Krsinich, Prices Unit, Statistics New Zealand
April 2013

Papers and recommendations from the committee available here:

http://www.stats.govt.nz/browse_for_stats/economic_indicators/CPI_inflation/2013-cpi-review-advisory-committee.aspx

The recommendation of the committee re: retail transaction data was:

Recommendation 11:

i. That Statistics NZ endeavours to use retail transaction (scanner) data to measure price change in the CPI.

ii. That the methods used align with international best practice.

iii. That arrangements be in place to ensure the continuity of retail transaction data supply.

iv. That there is a suitable contingency plan in the event that the supply of retail transaction data is unavoidably interrupted.

v. That Statistics NZ periodically audits prices in retail outlets to maintain confidence that the retail transaction data meets quality requirements.

5. Retail transaction data

5.1 Executive summary

This chapter discusses both the potential benefits and challenges of incorporating scanner data into the production of the New Zealand CPI.

New methodologies are required. Statistics New Zealand has contributed to international research in this area. In particular, collaborative work with Statistics Netherlands, developing and testing new methods on a detailed research dataset of New Zealand consumer electronics scanner data, has given important new insights.

5.2 Issues to consider

The issues Statistics NZ would like the committee and CPI user community to consider are:

- Should Statistics NZ continue to pursue the use of scanner data – to measure price change in the CPI and reflect shifts in the relative importance of goods?
- Does the committee agree the tentative conclusions on methodology (ie the use of ITRYGEKS (TD) for consumer electronics and RYTPD for supermarket data) are appropriate?
- Should Statistics NZ look to make full use of the price and quantity information available in scanner data?
- Should Statistics NZ continue with the current treatment of seasonal goods (eg fresh fruit and vegetables) until the introduction of scanner data, which would result in improved price measurement for these products?

5.3 Introduction to using scanner data

Many products, for example those purchased at supermarkets, have their barcodes scanned at the time of purchase. This retail transaction data – or 'scanner’ data – records prices, quantities sold, and associated information for all transactions (not just a sample) across the full reference period.

From the 2006 Consumers Price Index Review onwards, aggregated scanner data for supermarket products and for consumer electronics has been well used to:

- determine the expenditure weights of some goods in the CPI basket
- determine whether expenditure weight adjustments are required to reflect volume changes since the weight reference period (but before implementation of reviews) and, if so, by how much
- select representative products to survey when price collectors visit retail outlets each month or quarter
- ensure the mix of brands in the CPI price samples reflect market shares.

Since 2008, Statistics NZ has been actively researching the further potential of using more-detailed scanner data for directly estimating price change for products sold through supermarkets, and for consumer electronics products. The focus has been on determining appropriate methodologies. Traditional index formulae are problematic when applied to scanner data for two reasons:

- the volatility of prices and quantities, due to discounting and seasonality
- the high degree of ‘churn’ – ie new products entering and old products leaving the market.

5.3.1 Identifying which method to use

Statistics NZ has collaborated with Statistics Netherlands by empirically testing, on New Zealand consumer electronics scanner data, a new benchmark index method called the Imputation Törnqvist Rolling Year GEKS (ITRYGEKS). This method extends the rolling year GEKS (RYGEKS) method proposed by Ivancic, Diewart, and Fox (2011), which had been seen as the benchmark method for scanner data. The research showed RYGEKS can be biased – by not accounting for the implicit price change associated with new and disappearing products. The ITRYGEKS method incorporates hedonic modelling[1] into the RYGEKS approach in such a way that new and disappearing products are dealt with appropriately.

The ITRYGEKS method appears to be feasible and appropriate for the consumer electronics scanner data, which has extensive information on product characteristics available for the hedonic models that the ITRYGEKS method utilises.

However, for supermarket scanner data, the ITRYGEKS method is unlikely to be able to be fully implemented, as there is likely to be insufficient information on product characteristics in the data to implement the hedonic modelling the method requires. Research is still underway into the most-appropriate methodology to use for supermarkets. At this stage, a rolling year 'time product dummy' (RYTPD) hedonic method appears to be an improvement over the RYGEKS method for supermarket products.

5.3.2 Benefits of using scanner data

The potential benefits of using retail transaction data to measure price change include:

- improved accuracy, due to greater coverage of transactions and availability of real-time quantities
- ability to use existing administrative-type data sources
- improved treatment of seasonal commodities
- ability to account for commodity and product substitution between reweights.

5.4. The advantages of scanner data

5.4.1 It’s more accurate

Currently, the CPI relies on sampling prices across several dimensions – commodities, products, outlets, and time. Quantities are based on information acquired during the Household Economic Survey reference period, and are updated only every three years. It is difficult to accurately estimate the sampling error associated with current practice, as it is largely based on informed judgement rather than probabilistic sampling.

[1] Hedonic modelling is the use of regression modelling to control for compositional shifts in the characteristics of goods being sold. Price (or, more usually, the log of price) is modelled against time (eg ‘month’) and price-determining characteristics (eg TV screen size), and the index is derived from the parameters estimated for time. This ensures that the effect, on the average price, of the change in quality composition of goods being sold (in terms of the characteristics observed and included in the hedonic regression model), is removed from the price index.

In contrast, scanner data has the potential to give a more complete picture of both prices and quantities sold at any point in time. Depending on the scanner data's source, there may also be information on the characteristics of each product, which can be utilised for quality adjustment.

For supermarket scanner data it would be possible to disaggregate the data regionally. This would potentially mean having more accurate regional disaggregation for these products. Region is not currently available on the consumer electronics scanner data, but there is likely to be only moderate regional variation for these products.

5.4.2 It re-uses existing data

Given that scanner data is already collated by or for businesses, it would be good practice to re-use it to generate official statistics. This would reduce fieldwork, and the respondent load associated with collection, which involves observing products and prices in stores, and discussing product changes with store staff, particularly for consumer electronics. While there is likely to be an increase in the analytical resource required to introduce and maintain a robust scanner data production system, emerging consensus among national statistical offices is that this is the direction a modern statistical office should be moving in. The net result is likely to be a similar level of resources required to produce more accurate statistics.

New Zealand consumer electronics data

Statistics NZ has been using scanner data for consumer electronics products from market research company GfK for several years, to inform expenditure weighting in the CPI. This data is close to full-coverage of New Zealand's market, and contains sales values and quantities aggregated to quarterly levels for combinations of brand, and up to six characteristics.

Recently, Statistics NZ purchased a more detailed dataset for mid-2008 to mid-2011 for eight products:

- camcorders
- desktop computers
- digital cameras
- DVD players and recorders
- laptop computers
- microwaves[2]
- televisions
- portable media players.

Monthly sales values and quantities are disaggregated by brand, model, and around 40 characteristics. This data was used in the research on methods described in this chapter.

5.4.3 It can handle seasonal variation in quantities

Price change for some products is currently difficult to measure accurately because of the seasonality of quantity shares. The current fixed-basket approach, when applied to seasonal prices and quantities, has the potential to magnify or distort actual price movements. In the extreme case, some products have seasonal periods of unavailability

[2] Microwaves are not strictly a 'consumer electronics' product but, as a product with less rapid technological change, they provide a useful comparison of how different price index methods perform.

– which poses a particular challenge. Index estimation could be improved by modifying the traditional index formula / methods used (eg by using seasonal baskets).

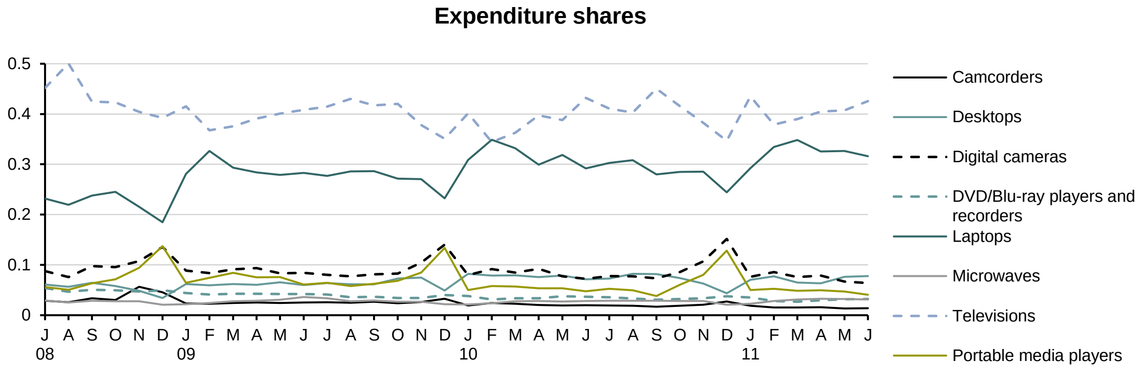
However, the data limitations associated with current practice mean that even the best solution applied to traditional data sources will still be less than optimal. In particular, is the fact that the quantities underlying expenditure shares are updated only every three years (although they vary within these periods).

Scanner data, which has real-time information on both prices and quantities, is more suited to accurately estimating price movements for these seasonal goods. ITRYGEKS (for consumer electronics) or RYTPD (for supermarkets) would accurately reflect the price movements, taking the corresponding actual quantity changes into account appropriately.

Note that some products with seasonal expenditure shares / availability, would still display seasonality in price indexes derived from scanner data – for example, many fresh fruit and vegetables. Whether seasonal adjustment (to try and identify the underlying trend in price movement) is necessary and/or feasible, is a separate issue. It may be more desirable to focus on annual than monthly movements for these products. Fresh fruit and vegetables used to be seasonally adjusted in the CPI, but the high irregular component, along with not being able to revise, meant it was not adding a lot of value, and it was discontinued after the 2006 review. Seasonal adjustment of a revisable analytical series seems a more plausible option for future consideration.

Figure 5.1, from Krsinich (2012), shows the seasonal patterns in expenditure shares of the eight consumer electronics products investigated.

Figure 5.1

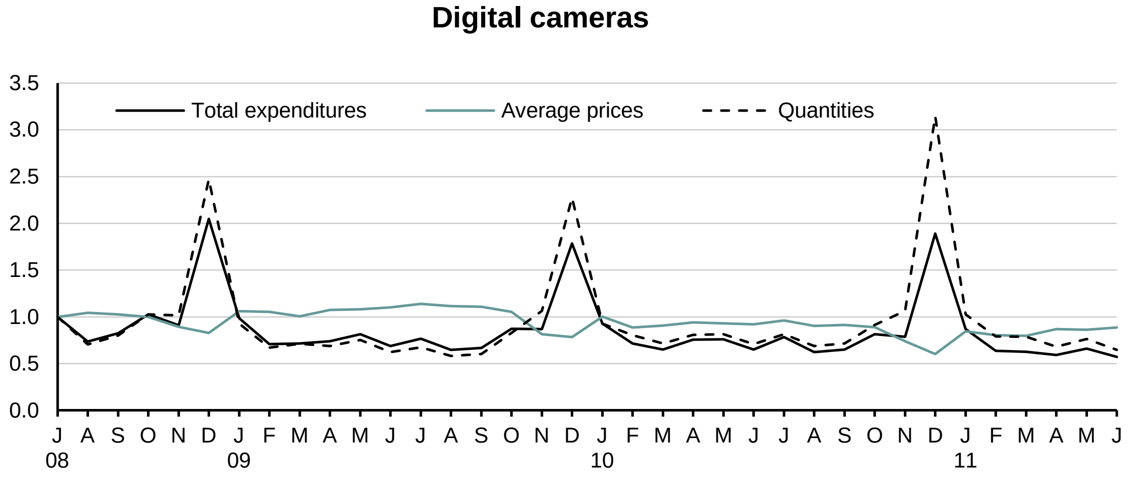


Source: Statistics New Zealand, GfK

In figure 5.2,[3] looking at digital cameras, seasonality is present in total expenditure, quantities, and average price (unadjusted for quality change). Obviously, this would be a challenging product to accurately measure price movements for – using prices sampled across time and quantities averaged across the year.

[3] Note that, for confidentiality reasons, expenditure, quantity and average price values were each rescaled so the July 2008 value is equal to 1. The relative levels are unaffected by this rescaling.

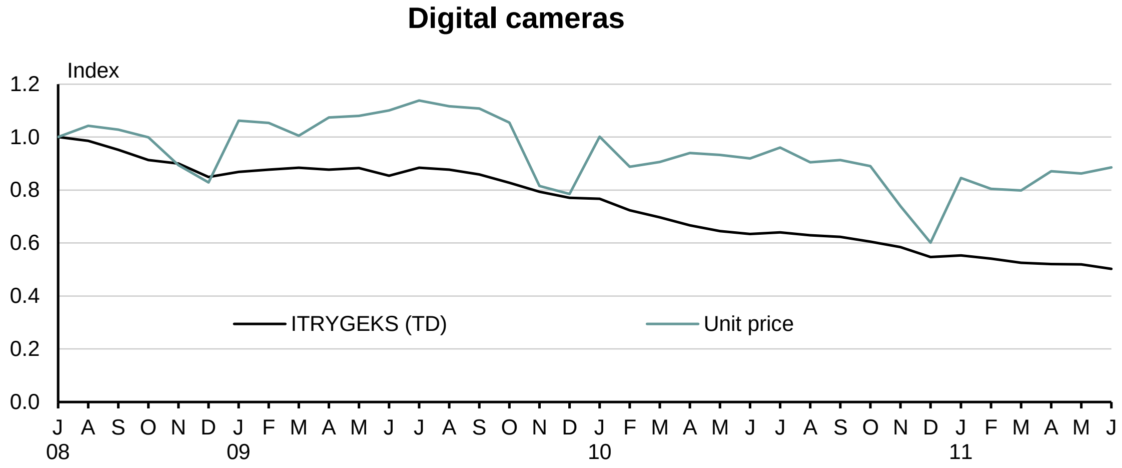
Figure 5.2



Source: Statistics New Zealand, GfK

By using an appropriate index methodology, a quality-adjusted price index that deals with the seasonality of expenditure shares and quality composition appropriately can be produced. In this case it would be ITRYGEKS, which utilises all prices and quantities appropriately and imputes price movements for new and disappearing goods based on hedonic models. Figure 5.3 shows this. ITRYGEKS is shown next to the index based on the unit price, which reflects the seasonality of quality composition within digital cameras – there is no obvious seasonality remaining in the quality-adjusted price movements.

Figure 5.3



Source: Statistics New Zealand, GfK

5.4.4 It accounts for substitution

Because scanner data has the potential to provide prices and quantities for the most-detailed level of product specification (ie barcode level), price indexes estimated from this data by using appropriate methodology will reflect consumers' substitution behaviour across products with different features. Scanner data can also be used to empirically test substitution effects for different product categories, to infer the appropriate level at which to fix expenditure shares.

5.5 The challenges of scanner data

The previous section emphasised the advantages of fully utilising the rich information content of scanner data. However, the behaviour of prices and quantities at the barcode level, as reflected by scanner data, mean that traditional index methodologies do not work well.

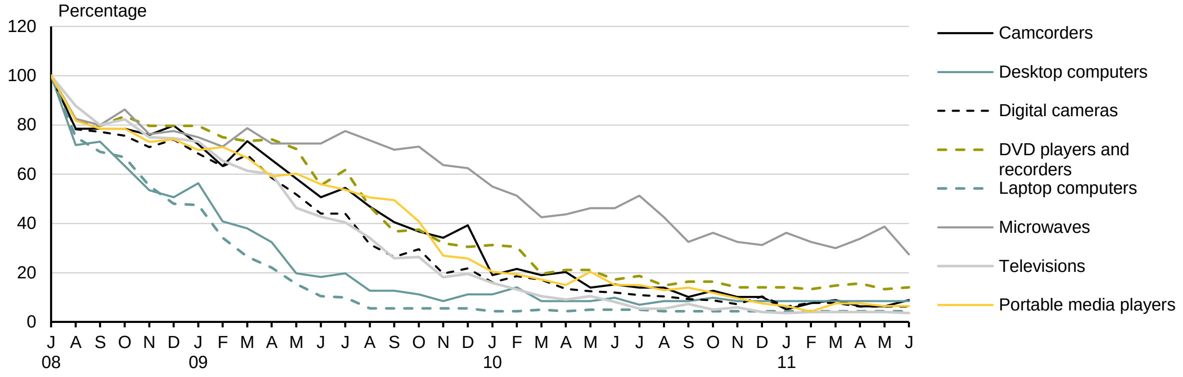
There are two main reasons for this – product churn, and the volatile nature of prices and quantities due to discounting, seasonality, and the life-cycle of products.

'Churn' refers to the turnover in products being sold in the market. For some products, in particular consumer electronics, this turnover can be significant. Figure 5.4 shows the percentage of products sold in July 2008 still on sale each month over the subsequent three years, for the eight consumer electronics products investigated. For laptop computers, only around 10 percent of the July 2008 products were still being sold a year later.

Note that current CPI practice for consumer electronics is to regularly update products being priced during the time between the three-yearly expenditure weight updates. Therefore, the basket will be more representative than is implied by figure 5.4.

Figure 5.4

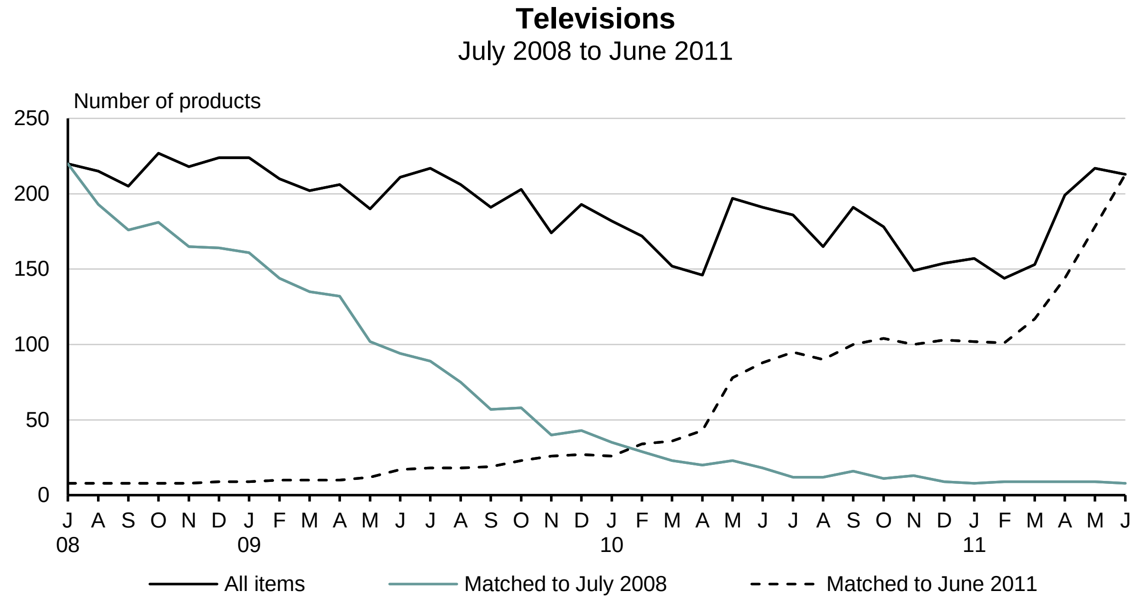
Percentage of July 2008 models available
July 2008 to June 2011



Source: Statistics New Zealand, GfK

Figure 5.5 shows a more detailed picture of the product churn for televisions. It shows the number of July 2008 products still being sold each month over the three-year period, but also the number of June 2011 products being sold in the three years up to that month, and the total number of distinct products sold each month.

Figure 5.5



Source: Statistics New Zealand, GfK

Given this high degree of product churn, the obvious solution is to use a high-frequency (eg monthly) chained superlative index (eg Törnqvist).This would maximise the number of matched products included in the index calculation and reflect substitution across products by incorporating updated quantities each month.

A superlative index is one that utilises both current- and reference-period quantity shares symmetrically, which results in substitution between products being accounted for in the index appropriately.

In a monthly chained index, the basket of products and their associated quantities or expenditure shares are updated every month. The index is calculated between the previous and current month on the basis of these updated products and weights and linked onto the previous index value.

It has been shown by researchers (eg Ivancic, Diewart, & Fox, 2011) that high-frequency chained superlative indexes can result in substantial 'chain drift' when applied to supermarket scanner data, due to a significant amount of price and quantity 'bouncing' because of discounting and seasonality.

Chain drift is the bias that occurs when a chained index diverges, or systematically 'drifts' away, from its direct (ie unchained) counterpart. A chained index in which the return of prices and quantities to previous levels does not correspond to the index also returning to the previous level, is exhibiting chain drift.

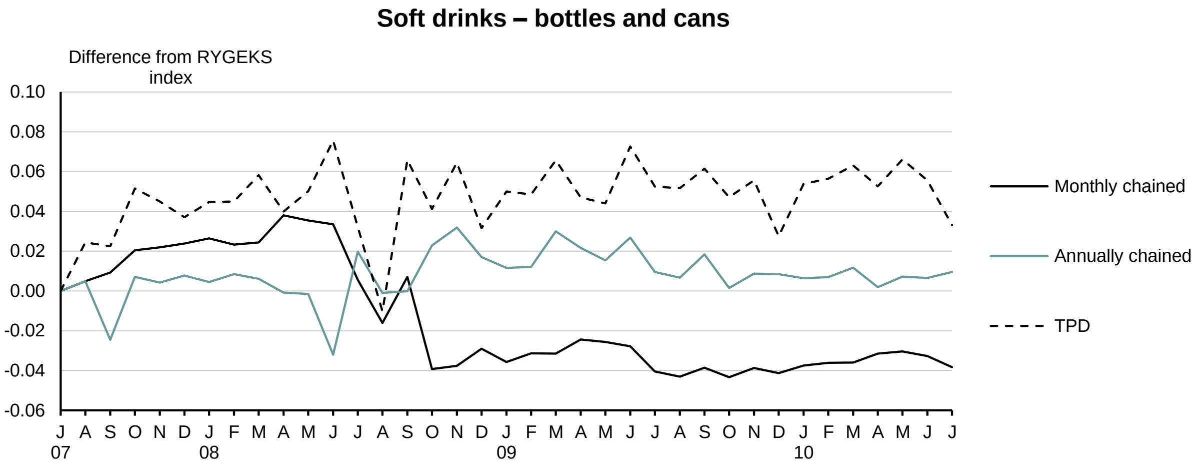
Supermarket products tend to be discounted frequently and, as might be expected, quantities bought increase sharply in response to these discounts.

In 2010, Statistics NZ and Statistics Netherlands were contracted as peer-reviewers of the Australian Bureau of Statistics research into supermarket scanner data. This provided access to the research dataset, which was used as a basis for developing a suite of SAS programs to run several different index methods (see appendix 5a). Krsinich (2011a) shows results from applying these methods.

For confidentiality reasons, price indexes from this data could not be published directly, but the indexes in terms of their difference from the then benchmark method (the rolling year GEKS, or RYGEKS) were published. RYGEKS is explained in the next section.

Figure 5.6 shows how three different methods differ from RYGEKS. In this example, the monthly chained Törnqvist has a downward drift in comparison with the RYGEKS benchmark.

Figure 5.6



Source: Statistics New Zealand, GfK

5.6 The solutions so far

5.6.1 The rolling year GEKS

To address the problems outlined in the previous section, Ivancic, Diewart, and Fox (2011) proposed a method for producing price indexes from scanner data that uses all the prices and quantities in the data, and is free of chain drift. This is called the rolling year GEKS (RYGEKS). RYGEKS is based on the Gini, Eltetö and Köves, and Szulc (GEKS) method, which is used for multilateral spatial price indexes such as purchasing power parities (these compare prices in different countries at a point in time).

Within a window of time (usually one year, plus a period to allow for seasonal unavailability, so five quarters or 13 months) the RYGEKS index between periods t1 and t2 is the geometric mean of all the superlative bilateral indexes (such as the Törnqvist or, in Ivancic et al (2011), the Fisher) between:

1. t1 and all the other periods in the window,
2. t2 and all the other periods in the window.

Formulating the monthly RYGEKS with a 13-month rolling window is as follows:

For the first window, ie t=0 to 12, RYGEKS is equal to GEKS:

$$P_{GEKS}^{0T} = \prod_{t=0}^{T} \left[P^{0t} \times P^{tT} \right]^{1/(T+1)}$$

Where P^{ij} is any superlative index (eg Törnqvist) between periods i and j.

From t=13 onwards, RYGEKS links on the most-recent movement from the GEKS calculated on the next window (ie from t=1 to 13, then from t= 2 to 14, and so on) as follows:

$$P_{RYGEKS}^{0,13} = P_{GEKS}^{0,12} \prod_{t=1}^{13} \left[P^{12,t} \times P^{t,13} \right]^{1/13} = \prod_{t=0}^{12} \left[P^{0t} \times P^{t,12} \right]^{1/13} \prod_{t=1}^{13} \left[P^{12,t} \times P^{t,13} \right]^{1/13}$$

$$P_{RYGEKS}^{0,14} = P_{RYGEKS}^{0,13} \prod_{t=2}^{14} \left[P^{13,t} \times P^{t,14} \right]^{1/13}$$

and so on.

However, a problem with the RYGEKS method is that it relies on the price movements between matched products only. Any implicit price change associated with new or disappearing products is effectively 'linked out'. So, for example, if the initial price of the latest model of a mobile phone is high relative to the features (ie the quality) of the phone then this implicit price increase is not included in RYGEKS. Conversely, if this new model is introduced at a price that is low relative to its features then this implicit price decrease will similarly not be reflected in the RYGEKS index.

5.6.2 The imputation Törnqvist rolling year GEKS

Jan de Haan, of Statistics Netherlands, proposed an extension to the RYGEKS method that addresses this limitation of RYGEKS. This uses hedonic models to impute for new and disappearing products, and is called the imputation Törnqvist rolling year GEKS (ITRYGEKS). The method was empirically tested on the New Zealand consumer electronics scanner data. This work was outlined in de Haan and Krsinich (2012), and presented at these conferences and workshops during 2012, from which useful feedback was gained:

- UNECE/ILO meeting of the group of experts on consumer price indexes, Geneva, May 2012
- Statistics Sweden's workshop on scanner data, Stockholm, June 2012
- New Zealand Association of Economists conference, Palmerston North, June 2012
- Economic Measurement Group workshop, University of New South Wales, Sydney, November 2012.

Unlike RYGEKS, which is based on superlative indexes (eg Fisher or Törnqvist), ITRYGEKS is based on geometric means of hedonic bilateral indexes. The formulation is as above for RYGEKS, with the difference that the P^{ij} are bilateral, weighted, time-dummy hedonic indexes. Three versions of ITRYGEKS are proposed in the paper:

- based on explicit imputation – which requires numeric data and/or categorical data that does not have new or disappearing categories
- based on time product dummy hedonic indexes – which is shown algebraically to be redundant as it is equivalent to doing no imputation for new and disappearing specifications
- based on time dummy hedonic indexes – ITRYGEKS(TD), this was applied to the consumer electronics data.

By taking the mean expenditure shares as weights for the matched items, and half of the expenditure shares for the unmatched products, the paper shows that ITRYGEKS(TD) is algebraically equivalent to a Törnqvist for the matched items. For new and disappearing products, it applies a Törnqvist formula to prices predicted from hedonic models for the period in which there is no price available.

That is, ITRYGEKS(TD) from period 0 to t can be expressed as follows:

$$P_{TD}^{0t} = \exp \hat{\delta}^t = \prod_{i \in U^{0t}} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{s_i^0 + s_i^t}{2}} \prod_{i \in U_{D(t)}^0} \left(\frac{\hat{p}_i^t}{p_i^0} \right)^{\frac{s_i^0}{2}} \prod_{i \in U_{N(0)}^t} \left(\frac{p_i^t}{\hat{p}_i^0} \right)^{\frac{s_i^t}{2}}$$

Where

U^{0t} is the set of matched products with respect to periods 0 and t

$U^{0}_{D(t)}$ is the set of 'disappearing' products with respect to periods 0 and t

$U^{0}_{N(t)}$ is the set of 'new' products with respect to periods 0 and t

Note that this expression can be generalised to ITRYGEKS between any two periods i and j.

ITRYGEKS(TD) was applied to the New Zealand consumer electronics data for eight products:

- camcorders
- desktop computers
- digital cameras
- DVD players and recorders
- laptop computers
- microwaves
- televisions, and
- portable media players.

In addition to applying the ITRYGEKS(TD) method, indexes were produced from the scanner data using other methods, including two different approaches to hedonic indexes based on rolling 13-month windows of data:

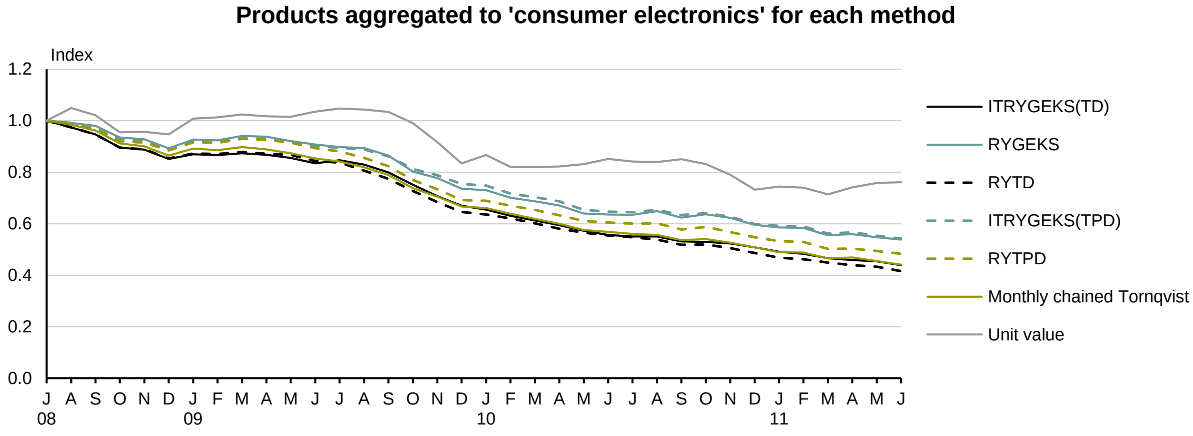
- a rolling year ‘time dummy’ hedonic index (RYTD), which is derived from regression models where the log of price is modelled against time and the characteristics of the products
- a rolling year ‘time product dummy’ index (RYTPD), which includes the product identifiers as the ‘characteristics’ being controlled for by the regression model.

The following conclusions were reached:

1. The monthly chained Törnqvist is not a viable method for consumer electronics, as it appears to have downward chain drift for most products examined, except microwaves and portable media players.
2. RYGEKS shows evidence of bias, due to not accounting for the price movements of new and disappearing products, particularly computers.
3. The easier-to-implement rolling year time dummy (RYTD) hedonic index (which explicitly incorporates characteristics into hedonic models) gives similar results to ITRYGEKS(TD), in particular for computers.
4. In some cases, such as supermarket data, few or no characteristics are likely to be available and so neither ITRYGEKS nor RYTD, which are both based on time dummy hedonic models, will be feasible. Results suggest that in this situation a rolling year time product dummy (RYTPD) – a 'fixed effects' approach, which controls for the product identifiers rather than characteristics – does some adjustment for quality change and is therefore preferable to RYGEKS.

Aggregation of the eight products (see figure 5.7), using their relative expenditure weights from the scanner data, shows that the matched-product RYGEKS is upwards biased compared with ITRYGEKS(TD), which imputes the implicit price movements of new and disappearing products. The RYTD hedonic index and the monthly chained Törnqvist both track the benchmark ITRYGEKS(TD) closely, but for the monthly chained Törnqvist this appears to be a coincidental cancelling out of biases in two opposite directions – chain drift and quality-change bias – for televisions, which have by far the most significant weight of all the eight products.

Figure 5.7



Source: Statistics New Zealand, GfK

5.6.3 Rolling year time product dummy (RYTPD)

ITRYGEKS(TD) is a feasible method only when there are sufficient characteristics in the data to estimate the time dummy hedonic indexes. The research data obtained for consumer electronics has extensive characteristic data, which could be incorporated into a future production system to estimate price indexes for these products in the CPI.

However, for supermarket data, the only explicit characteristics likely to be available are weight, volume, or size. This means ITRYGEKS(TD) may not be feasible for supermarket products.

Because de Haan and Krsinich (2012) showed empirically that the rolling year time product dummy (RYTPD) hedonic approach tends to sit closer to the ITRYGEKS(TD) benchmark than the RYGEKS approach does, at this stage RYTPD looks to be the most promising method to apply for supermarket data.

Theoretical work is underway to determine what the implicit imputations for new and disappearing products are, under the RYTPD method.

Note that the RYTPD method is now the focus of further collaborative research with Statistics Netherlands, and is the topic of the paper to be presented by Statistics NZ at the Ottawa Group meeting of CPI experts in Copenhagen in May 2013 (see appendix 5b).

5.7 Current international practice

A significant body of research exists on scanner data, but actually implementing it in production in official statistical agencies has been limited.

Statistics Netherlands and Statistics Norway are the only countries to comprehensively incorporate supermarket scanner data into their CPI production systems. Some other countries (eg Switzerland) implement scanner data in a partial way by using it as a price source for the selected products in the traditional fixed-basket approach for supermarkets. The disadvantage of this approach is that substitution across products will not be reflected.

Although RYGEKS has tended to be considered a good benchmark method for scanner data, it is also relatively new and untested, and has therefore not been adopted into production. Statistics Netherlands' use of scanner data in production uses unweighted Jevons[4] indexes at the elementary aggregate level, combined with cut-off sampling based on expenditure shares. They run RYGEKS separately as a shadow system, and use it as a benchmark to help inform the thresholds for sampling. (See de Haan & van der Grient, 2011.)

Statistics Norway use a monthly chained Törnqvist approach to estimating price indexes for food and non-alcoholic beverages, but this approach has been shown to be downwardly biased (see Johansen, I & Nygaard, R, 2011).

The official statistical agencies of many other countries are researching the potential of putting scanner data into production. This includes the Australian Bureau of Statistics, which aims to implement scanner data into production for supermarkets in the near future.

5.8 The current research focus

In New Zealand, it appears to be feasible to put consumer electronics scanner data into production, if access to the data at a disaggregated level with detailed characteristics can be gained on a timely basis. The ITRYGEKS(TD) method to estimate unbiased price indexes for these products could be implemented.

The research focus now is on what method will be appropriate for estimating price indexes for supermarket data, which would not have the detailed characteristics necessary to apply ITRYGEKS(TD). At this stage the RYTPD method is looking promising. Its potential will be discussed at the next Ottawa Group meeting. Note that using a version of this method for rental dwellings has been researched – here there are limited characteristics, but there is reason to believe there is implicit price change associated with rental dwellings entering and leaving the rental market (Krsinich, 2011b).

Other options that can be considered for supermarket scanner data are:

- the potential for a cut-down version of the ITRYGEKS(TD) method if the weight, size, or volume characteristics, in combination with categorical variables derived from product identifiers, explain enough of the price variation
- manual intervention – such as identifying, linking, and appropriately quality-adjusting between barcodes, where the only change in product is a change in weight, size, or volume.

See appendix 5b for the abstract of the paper accepted for the next Ottawa Group meeting in May 2013 – *Using the rolling year time product dummy method for quality adjustment in the case of unobserved characteristics.*

5.9 Potential issues associated with putting scanner data into production

The research so far has focused on which methodology will be most appropriate for consumer electronics and for supermarket scanner data. Other more practical issues are likely to arise around putting these methods into production. These will be addressed as data supply arrangements are clarified.

Timing of supply for supermarket data may mean that figures based on the full month may not be available in time to incorporate into production. This would require investigation of the options and implications of partial or lagged data. Note that this kind

[4] The Jevons index is the unweighted geometric mean of price relatives. A price relative is the ratio of the price of an individual product in one period to the price of that same product in some other period.

of data limitation is dealt with in production in existing Statistics NZ outputs such as the Retail Trade Survey, which rates up partial periods of data where appropriate.

Data supply arrangements may not enable full coverage of all products. Even so, the quality of price indexes using scanner data would be an improvement over the current situation, but investigation of how best to incorporate any partial coverage would be required.

The following table shows how limitations on the level of detail and coverage available from the scanner data affect the ability to reflect quality-adjusted price movements accurately. In the table, 'product' means the finest level of specification available – for example, chocolate biscuits, manufacturer A, brand b, type C, weight 200g. 'Item' means the good – for example, 'chocolate biscuits'.

Table 5.1

Implications of different levels of detail from scanner data

Level of detail and information available	**Implications**
1. Full coverage of prices and quantities across all items and products at all points in time, with explicit information about price-determining characteristics at the product level.	This is the ideal situation where our benchmark method – the ITRYGEKS(TD) – can be applied to create price indexes which are both fully quality adjusted and fully reflect consumer substitution at the level of items and products. **Example** This level of detail and information exists in the GfK consumer electronics scanner data, although the data is aggregated across time to the monthly level and across transactions to the model level, and region and outlet information is not available.
2. Full coverage of prices and quantities across all items and products at all points in time, with limited or no information about price-determining characteristics at the product level.	There is less scope for fully quality adjusting price indexes without explicit information on characteristics, but the RYTPD method discussed in the paper is looking promising. Consumer substitution at the level of items and products can be fully reflected. **Example** This level of detail potentially exists in supermarket scanner data and, unlike consumer electronics scanner data, information on region and outlet is potentially available (which would enable disaggregation to region and reflection of consumer substitution across outlets).
3. Full coverage of prices across all products, without corresponding quantities. Limited or no information about price-determining characteristics at the product level.	Without quantities, changes in the expenditure shares of different products is not able to be reflected and consumer substitution across items and products over time cannot be adequately reflected. Some external source of weighting, such as from a household expenditure survey needs to

	be incorporated but this will be at discrete intervals so substitution between these points in time is not reflected. Also, the weighting will be at the item level or higher due to the limitations of the survey data.
4. Sample of items, with full coverage of product prices and quantities across all points in time within those items.	Consumer substitution across products within items is reflected, but consumer substitution across items (eg if consumers move towards buying savoury biscuits rather than chocolate biscuits in response to their lower price movements) is not necessarily able to reflected accurately due to partial coverage of substitutable items.
5. Prices and quantities at all points in time for a sample of products (for example, those in the current CPI basket)	There is improved measurement of price change for the sample of products, as all transactions of the products are included, including those at discounted prices. Consumer substitution across products cannot be reflected accurately due to partial coverage of substitutable products, and there will be subjective judgement required when replacing products that are discontinued or at periodic reviews to maintain representativeness.

Monitoring processes will be very important. Although there are good reasons to set up scanner data production systems to be as automated as possible, for reasons of efficiency for example, there will need to be good procedures for checking outliers, consistency of classifications over time, and stability against past results. These kinds of processes are already in place for the CPI ‘used cars’ system, which uses a hedonic regression approach on a rolling window of data, so aspects of that process can be incorporated.

5.10 Conclusion

For both consumer electronics and supermarket products, implementing scanner data in production for the New Zealand CPI appears to be feasible. By using scanner data we could improve accuracy (through improving coverage of transactions and making use of quantity information) and reduce fieldwork resources by making use of existing privately-held administrative data. We would also have the potential to increase the frequency of price measurement for consumer electronics products from quarterly to monthly.

We have established that the ITRYGEKS(TD) method could be used to estimate indexes for consumer electronics in such a way that the implicit price changes of new and disappearing products are incorporated appropriately.

For supermarket data, where characteristics information is lacking, the RYTPD method is likely to produce better estimates of price change than the RYGEKS method, and research is on-going to establish the theoretical properties of this method.

References

de Haan, J, & van der Grient, HA (2011). Eliminating chain drift in price indexes based on scanner data. *Journal of Econometrics, 161*(1), 36–46.

de Haan, J, & Krsinich, F (2012, November). Scanner data and the treatment of quality change in rolling year GEKS price indexes. Paper presented at the Economic Measurement Group Workshop, Sydney, Australia. Available from www.asb.unsw.edu.au.

Ivancic, L, Diewert, WE, & Fox, KJ (2011). Scanner data, time aggregation and the construction of price indexes. *Journal of Econometrics, 161*(1), 24–35.

Johansen, I, & Nygaard, R (2011, May). Dealing with bias in the Norwegian superlative price index of food and non-alcoholic beverages. Paper presented at the twelfth meeting of the International Working Group on Price Indices, Wellington, New Zealand. Available from www.ottawagroup.org.

Krsinich, F (2011a, May). Price indexes from scanner data: A comparison of different methods. Paper presented at the twelfth meeting of the International Working Group on Price Indices, Wellington, New Zealand. Available from www.ottawagroup.org.

Krsinich, F (2011b, May). Measuring the price movements of used cars and residential rents in the New Zealand consumers price index. Paper presented at the twelfth meeting of the International Working Group on Price Indices, Wellington, New Zealand. Available at www.ottawagroup.org.

Krsinich, F (2012). A fresh look at patterns in gadget sales. *Economic News*, Statistics New Zealand, April 2012. Available from www.stats.govt.nz.

Appendix 5a: SAS code for scanner data methods

Over the past four years a SAS system to run all the different scanner data index methods has been developed. This was used to produce results for the research to date, and will likely form the basis for a future production system for scanner data at Statistics NZ.

The methods that are currently incorporated are:

- a range of bilateral index methods (Laspeyres, Paasche, Fisher, Walsh, and Törnqvist) – both direct and chained
- annually chained indexes, with ‘annually smoothed’ expenditure shares that use the monthly prices in conjunction with the year-to-date’s quantities (or, alternatively, expenditures) based on the above range of bilateral indexes
- RYGEKS – a rolling year GEKS based on any of the (superlative) bilateral indexes above (the window length can be specified by the user)
- RYTD – a rolling year time dummy hedonic index (ie a hedonic regression with characteristics of the products explicitly incorporated). The index is estimated on a rolling window (whose length can be specified) with the most-recent movement linked to the index at the previous period
- RYTPD – a rolling year time product dummy index that uses a fixed-effects 'hedonic' regression method where the product identifiers (ie barcodes) are the 'characteristics' controlled for. The index is estimated on a rolling window (length can be specified) with the most recent movement linked to the index at the previous period
- ITRYGEKS(TD) – imputation Törnqvist RYGEKS, where the inputs to RYGEKS are bilateral indexes from a time dummy hedonic regression.

Note that an updated linking approach is being incorporated, which will better combine the movements of new products in the RYTPD methods.

Versions of this code (with accompanying documentation) were shared with Statistics Netherlands, the Australian Bureau of Statistics, Statistics Israel, and Statistics Belgium.

Appendix 5b: Abstract for 2013 Ottawa Group

Abstract accepted for 13th Ottawa Group meeting in Copenhagen, May 2013.

Using the rolling year time product dummy method for quality adjustment in the case of unobserved characteristics

Frances Krsinich, Statistics New Zealand

This paper will discuss the use of the rolling year time product dummy (RYTPD) method in situations where characteristics are unavailable for explicitly incorporating into hedonic regression models. This builds on results from two earlier pieces of work. Krsinich (2011) used a fixed-effects hedonic model (ie a pooled TPD method) to benchmark the current matched-model approach to estimating the rental index for New Zealand, where we have few observed characteristics in the longitudinal rental data. More recently, de Haan and Krsinich (2012) extended the RYGEKS method to impute price movements for new and disappearing items using hedonic regression. Results from applying this 'imputation Törnqvist' (IT) RYGEKS to New Zealand consumer electronics scanner data were compared with a range of other methods. The RYTPD was the best performing of those methods that do not explicitly incorporate characteristics. We will argue that the RYTPD is a viable method in situations where characteristics are unobserved, such as supermarket scanner data. Unlike matched-model approaches, the RYTPD is doing some imputation for new and disappearing items, though in production a more sophisticated linking approach would be required to deal with the lagged capturing of price movements for new products.