ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

Review of web-scraped price indices

Purpose

1. This paper sets out the initial plans for a methodological review of the price index formulae that are available for use to calculate indices from high frequency and high volume datasets.

Actions

2. Members of the Panel are invited to consider:

   a) the approach ONS are taking to carry out this review;
   b) the criteria used to assess the different methodologies;
   c) the relative merits and weaknesses of the methodologies presented in Annex A, with a view to prioritising some methods for assessment

Background

3. The Office for National Statistics (ONS) is exploring alternative and innovative ways of collecting data. A part of this program has been to investigate the use of web-scraped price data.

4. One aspect of this work, referred to as the 'bulk collection', is the use of scrapers built in Python by the ONS to web-scrape prices for 33 Consumer Prices Index (CPI) items from 3 retailers that have an online presence. These have been scraped since June 2014. This data has been used to compile price indices using different formulae and at different frequencies. Further details regarding this work can be found in "Research indices using web scraped price data" (ONS, 2016).

5. Another aspect of this work involves using web-scrapers to collect other items in the CPI basket using point and click software, such as Import.io and Dexi.io, to replicate the manual collection of centrally collected items in the CPI price collection.

6. Recently, ONS has also acquired web-scraped data from third party sources, such as a set of clothing data from the fashion forecaster WGSN (research on this data will be released as part of the GSS Methodology Series in the near future).

7. There have been several methods developed to handle these larger and high frequency datasets, as traditional methods often prove ineffective. These methods may be more or less appropriate, dependent on the product that is being measured.

8. The web-scraping project is supported by a Eurostat grant which expires in May 2017. One of the grant objectives is to explore and assess different methods for compiling price indices using high frequency data, with a view to finalising a price index methodology for use of web scraped data in the UK CPI. This review aims to meet this objective.

Review of web-scraped price indices

9. In order to decide the most appropriate methodology to use for web scraped data, a number of points need to be considered:

    a. A number of methods have been developed that are more suited to high frequency and high volume data. These are as follows (See Annex A for full descriptions, and Annex B for advantages and disadvantages):

        i. Chained bilateral indices between contiguous periods (for example, the daily chained Jevons)

        ii. A Unit value Index

        iii. A fixed base Jevons

        iv. The GEKS family of methods

        v. The Fixed Effects Window Splice (FEWS)

        vi. Clustering Large datasets Into Price indices (CLIP)

    b. Some index number formulae might be more suitable for some products than others. For example, from preliminary research, the IntGEKSJ and FEWS indices are not suited for clothing, whereas the CLIP index may be more appropriate. Statistics New Zealand suggest that the FEWS index is more suitable for technological goods due to their rapid change in quality, *Krsinich (2015)*. Some examples of where different methodologies have been applied to items from the bulk collection can be seen in Annex C.

    c. Further considerations are required regarding frequency:

        i. with what frequency should these indices be calculated - monthly, weekly, or even daily?

        ii. which growth rates should be published for a daily or weekly index - movement since the same time period a year ago, a month ago, from the previous period, or from the previous CPI release?

        iii. should average monthly or weekly unit prices be used in the calculation of the monthly or weekly price indices?

    d. The CPI is also the UK's Harmonised Index of Consumer Prices[1]HICP. If web-scraped data is to be incorporated in the CPI, there are two aspects of methodology that may not be compliant with current Eurostat regulations (*Eurostat*):

        i. "elementary aggregate indices are computed as the ratio of geometric average prices or the ratio of arithmetic average prices"

        ii. "prices should then be observed for the selected products over time",

[1] A measure of inflation designed by Eurostat to be comparable across member states

Some of the aforementioned index number formulae are not compliant with these points. There may be potential to separate the CPIH from the CPI to avoid these issues in the future, although this may be dependent on the future development of production systems.

Matthew Mayhew
Methodology, ONS
December, 2016

References

Breton, R et al, (2016), Research indices using web scraped price data: May 2016 update

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

EUROSTAT, HICP Methodology

Diewert, W.E., Fox K,J., and Ivancic, L. (2009) Scanner data, time aggregation and the construction of price indexes, Journal of Econometrics 161 (1) pp 24-35.

Krsinich, F (2014), The FEWS index: fixed-effects with a window-splice

De Haan, J and van der Grient, H. (2009), Eliminating chain drift in price indexes based on scanner data, Journal of Econometrics 161 (1) pp 24-35

De Haan, j and Krsinich, F. (2012), The treatment of unmatched items in rolling year GEKS prices indexes: evidence from New Zealand Scanner data.

Metcalfe, E et al. (2016), Research indices using web scraped price data: clustering large datasets into price indices (CLIP)

ILO, Consumer price index manual: Theory and Practice

Howard, A, Dunford, Using transactions data to enhance the Australian CPI

List of Annexes

Annex A	Index Formulae
Annex B	Advantages and Disadvantages to the index formulae
Annex C	Indices for different products.

Annex A – Index formulae

1. Fixed based Jevons:

   The fixed based Jevons fixes the base period to the first period in the dataset, and matches the products common to all periods. It compares the current period price back to the base periods. The formula is defined as follows:

$$P_{FBJ}^{0,t} = \prod_{j \in S^*} \left(\frac{p_j^t}{p_j^0} \right)^{\frac{1}{n^*}}$$

   where p_j^t is the price of product *j* in period *t*, S^* is the set of products common to all periods, and n^* is the number of products in S^*.

2. Chained Bilateral Jevons Indices:

   The chained bilateral index involves constructing bilateral Jevons indices between period *t* and *t-1* and then chaining them together. The formula is defined as follows:

$$P_{CJ}^{0,t} = \prod_{i=1}^{t} P_J^{i-1,i} = \prod_{i=1}^{t} \left(\prod_{j \in S^{i-1,i}} \frac{p_j^i}{p_j^{i-1}} \right)^{\frac{1}{n^{i-1,i}}}$$

   where $P_J^{i-1,i}$ is the Jevons index between the current period and the previous period, $p_j^i$is the price of product *j* at time *i,* $S^{i-1,i}$ is the set of products observed in both period *i* and *i-1*, and $n^{i-1,i}$ is the number of products in $S^{i-1,i}$.

3. Unit value Index:

   The Unit value index[2] is defined to be the ratio of averages between two unmatched sets of products between period 0 and period *t.* The formula is defined as follows:

$$P_{UV}^{0,t} = \frac{\left(\prod_{j \in S^t} p_j^t \right)^{\frac{1}{n^t}}}{\left(\prod_{j \in S^0} p_j^0 \right)^{\frac{1}{n^0}}}$$

   where S^0 is the set of products in period 0, and n^0 is the number of products in S^0, S^t is the set of products in period *t*, and n^t is the number of products in S^t. The geometric average has been used so that it is consistent with the other indices presented in this paper.

4. The GEKS Family of Indices is a set of indices that is based on a formula devised by Gini, Eltetö, Köves and Szulc:

   a. The GEKS-J Index:

      The GEKS-J index is a multilateral index, as it is calculated using all routes between two time periods. It was originally developed for Purchasing Power Parities but adapted for the time domain in *Diewert, W.E., Fox K,J., and Ivancic, L. (2009)* The

[2] Unit value is usually defined as the value of a product divided by the quantity bought, but since the data isn't available, an average price is taken for the web scraped data.

GEKS-J price index for period t with period 0 as the base period is the geometric mean of the chained Jevons price indices between period 0 and period t with every intermediate point (i = 1,...,t-1) as a link period. The formula is defined as follows:

$$P_{GEKSJ}^{0,t} = \prod_{i=0}^{t} \left(P_J^{0,i} P_J^{i,t}\right)^{\frac{1}{t+1}}$$

A product is included in the index if it is in the period *i* and either period 0 or period *t*.

b. RYGEKS-J:

RYGEKS-J or Rolling Year GEKS-J extends the GEKS-J to allow for a moving base period and allows for a longer series to be calculated without the need to revise the back series constantly. The formula is defined as follows:

$$P_{RYGEKS-J}^{0,t} = \begin{cases} \prod_{i=0}^{t} \left(P_J^{0,i} P_J^{i,t}\right)^{\frac{1}{t+1}} & t < d \\ \prod_{i=0}^{d-1} \left(P_J^{0,i} P_J^{i,d-1}\right)^{\frac{1}{d}} \prod_{k=d}^{t} \left(\prod_{i=k-d+1}^{k} \left(P_J^{k-1,i} P_J^{i,k}\right)^{\frac{1}{d}} \right) \end{cases}$$

where *d* is the window length, for a monthly series d=13. A formal definition of RYGEKS is in *De Haan and van der Grient (2009).*

c. ITRYGEKS:

As new products are introduced on the market and old products disappear an implicit quality change may occur - this often happens in technological goods. Hence, there is an implicit price movement which isn't captured in the RYGEKS method. In a market where consumers increase their purchase of higher quality goods these implicit movements need to be captured. De Haan and Krsinich (2012) propose using an imputed Törnqvist as the base of the RYGEKS. An imputed Törnqvist is a hedonically adjusted Törnqvist index, where the prices of new or disappeared products are imputed using a hedonic regression in the current or base period respectively. A hedonic regression assumes that the price of a product is uniquely defined by a set of *K* characteristics. The Imputed Törnqvist index is defined as follows:

$$P_{IT}^{0,t} = \prod_{j \in S^{0,t}} \left(\frac{p_j^t}{p_j^0}\right)^{\frac{w_j^0 + w_j^t}{2}} \prod_{j \in S_{N(0)}^t} \left(\frac{p_j^t}{\widehat{p_j^0}}\right)^{\frac{w_j^0}{2}} \prod_{j \in S_{D(t)}^0} \left(\frac{\widehat{p_j^t}}{p_j^0}\right)^{\frac{w_j^t}{2}}$$

where w_j^0 is the expenditure share of item *j* at time 0, w_j^t is the expenditure share for item *j* at time *t*, $\widehat{p_j^t}$ is the estimated price for a missing product at time *t*, $S^{0,t}$ is the set of products observed in both periods, $S_{N(0)}^t$ is the set of new products at time *t* but weren't available at time 0, and $S_{D(t)}^0$ is the set of products at time 0 that have disappeared from the market at time *t*. De Haan and Krsinich suggest three different imputation methods, these are:

i. The linear characteristics model:

Estimate the characteristic parameters using a separate regression model for each period. The imputed price is calculated as follows:

$$\widehat{p_j^t} = \exp\left(\widehat{\alpha^t} + \sum_{k=1}^{K} \widehat{\beta_k^t}\, z_{jk}\right)$$

where $\widehat{\alpha^t}$ is the estimate of the intercept, $\widehat{\beta_k^t}$ is the estimate of the effect characteristic *k* has on the price, and z_{ik} is the value of characteristic *k* for product *j.*

ii. The weighted time dummy hedonic method:

This method assumes parameter estimates for characteristics don't change over time, and includes a dummy variable, D_j^t, for which period the product was collected. In this method the imputed price is calculated by:

$$\widehat{p_j^t} = \exp\left(\hat{\alpha} + \widehat{\delta^t} D_j^t + \sum_{k=1}^{K} \widehat{\beta_k}\, z_{jk}\right)$$

where $\widehat{\delta^t}$ is the time specific parameter estimate.

iii. The weighted time-product dummy method:

This method can be used when detailed characteristic information is not available, and a dummy variable, D_j, for the product is created. The missing price is then estimated using:

$$\widehat{p_j^t} = \exp\left(\hat{\alpha} + \widehat{\delta^t} D_j^t + \sum_{j=1}^{N-1} \widehat{\gamma_J}\, D_j\right)$$

where $\widehat{\gamma_J}$is the estimate of the product specific dummy, the N^{th} product is taken as the reference product. This method assumes that the quality of each distinct product is different to the quality of other products to a consumer. It is a reasonable assumption as the number of potential characteristics is large and not all of them are observable.

For each of these methods, a weighted least squares regression is used, with the expenditure shares as the weights.

d. The Intersection-GEKS-J or IntGEKS-J:

The IntGEKS was devised by *Krsinich and Lamboray (2015),* to deal with an apparent flattening of RYGEKS under longer window lengths, though this was found to be an error in applying the weights. It removes the asymmetry in the match sets between periods 0 and *i* and between periods *i* and *t*, by including products the matched sets only if they appear in all three periods, the set $S^{0,i,t}$. The formula is defined as follows:

$$P_{IntGEKSJ}^{0,t} = \prod_{i=0}^{t} \left(P_{J,j\in S^{0,i,t}}^{0,i} P_{J,j\in S^{0,i,t}}^{i,t} \right)^{\frac{1}{t+1}}$$

If there is no product churn (products coming in and out of stock) then the IntGEKS-J reduces to the standard GEKS-J. The IntGEKS-J has more chance of "failing" than a standard GEKS-J as the products need to appear in more periods.

5. FEWS:

The Fixed Effects Window Splice produces a non-revisable and fully quality-adjusted price index where there is longitudinal price and quantity information at a detailed product specification level. It is based around the Fixed Effects index which is defined as follows:

$$P_{FE}^{0,t} = \frac{\prod_{j\in S^t} \left(p_j^t\right)^{\frac{1}{n^t}}}{\prod_{j\in S^0} \left(p_j^0\right)^{\frac{1}{n^0}}} \exp\left(\overline{\hat{\gamma}^0} - \overline{\hat{\gamma}^t}\right)$$

where $\overline{\hat{\gamma}^0}$ is the average of the estimated fixed effects regression coefficient at time 0. Using a fixed effects regression overcomes some of the disadvantages of using the time dummy ITRYGEKS, whilst being equivalent to it. Like the RYGEKS, after the initial estimation window, the new series is spliced onto the current series for subsequent periods; this is called a window splice. The window splice essentially uses the price movement over the duration of the estimation window, rather than the price movement in the latest period. This approach has the advantage of incorporating implicit price movements of new products at a lag. There is a trade-off, then, between the quality of the index in the current period and in the long term. Over the long term, the FEWS method will remove any systematic bias due to not adjusting for the implicit price movements of new and disappearing items. A full description of the method can be found in *Krsinich (2014).*

6. CLIP:

Clustering Large datasets Into Price indices is a recently developed price index from ONS. The CLIP groups products into clusters and tracks those clusters over time. In the base period the products are clustered according to their characteristics, for example if the product was on offer, as it assumes consumers would buy within a certain set of products on offer. Clusters are formed using the same rules over time, but the products that form the cluster can change over time, allowing for product churn. The geometric mean of the clusters in two periods are compared, creating a unit value index for each cluster, which are then aggregated using the size of the cluster in the base period. Mathematically, the formula is defined as follows:

$$P_{CLIP}^{0,t} = \frac{\sum_k |C_{k,0}| \frac{\left(\prod_{j\in k,t} p_j^t\right)^{\frac{1}{|C_{k,t}|}}}{\left(\prod_{j\in k,0} p_j^t\right)^{\frac{1}{|C_{k,0}|}}}}{\sum_k |C_{k,0}|}$$

where $C_{k,0}$ is cluster *k* in period *0*, $C_{k,t}$ is cluster *k* in period *t*, and $|C_{k,0}|$ is the size of a cluster. For a full description, please read *Metcalfe et al. (2016).*

Annex B – Advantages and Disadvantages to the index formulae

Table 1 – Advantages and Disadvantages

Index Formulae	Advantages	Disadvantages
Fixed Based Jevons	• Direct comparison from the base period to current period. • Tracks the same products over time. • Relatively straight forward, and easy to explain.	• The matched set may be small as products have to be common to all periods, and will possibly get smaller the further away the current period is from the base period, as products can't be followed through time in the same way as the CPI collection.
Chained Bilateral Jevons	• Uses more data • Products need only exist in contiguous periods to be included in the index. • Useful in a production environment as we only need the current and previous period. • Computationally straight forward, and easy to explain.	• High chance of chain drift, especially if there is high product churn. •
Unit Value	• Uses all the data in both time periods. • Relatively straight forward, and easy to explain	• Does not track products over time. • Not strictly a price index (See *ILO* paragraph 9.70)
GEKS-J	• Uses more data, as product needs to be observed in either the base and intermediate period, or intermediate period and current period. • Eliminates chain drift, a problem of the Chained Bilateral Jevons	• Revised when adding a new period. • Needs the whole time series of prices to create the index, so not suitable for a production environment. • In later periods, there is a loss of characteristicity. • Does not account for quality change. • The method is complex and computationally intensive.
RYGEKS-J	• Uses more data • Avoids the revision problem of standard GEKS. • Only the window length of data is needed to calculate the price index, so do not need the full time series.	• Loss of transitivity (although impact is thought to be negligible). • Does not account for quality changes of new and disappearing items. • This method is complex

	• Avoids the loss of characteristicity in later time periods.	and computationally intensive.
ITRYGEKS	• Allows for implicit price movements due to quality changes of new products, • Shares the other advantages of the RYGEKS method	• Depends on the hedonic regression model chosen and the availability of the characteristics of the products to impute. • De Haan and Krsinich recommended the use of the time dummy model, which is not possible with our data. • Resource intensive to develop the models needed. • Possibly more suited to be an analytical tool rather than for using in the production of consumer price indices. • This method is complex and computationally intensive.
IntGEKS	• Shares the advantages of other GEKS methods	• Uses less data than other GEKS methods • As the input indices cancel, the IntGEKS-J is equivalent to the average of t Jevons indices based on slightly different samples
FEWS	• Does not need a full set of characteristics for the quality adjustment to be done. • Doesn't revise as the GEKS does. • Does not require an analyst to develop the model. The Window Splice eliminates long term bias.	• Has some undesirable properties due to the quality adjustment and the window length, as investigated in *Howard et al (2015)*
CLIP	• Tracks groups of products over time to overcome products churn. • Uses all available characteristics that are available to cluster the data, so does not need as much information as an ITRYGEKS.	• Needs a certain amount of products in order to cluster, and a minimum amount of clusters to perform the weighting, else it reverts to being a Unit value. • This is a very experimental index so little is known

		about it

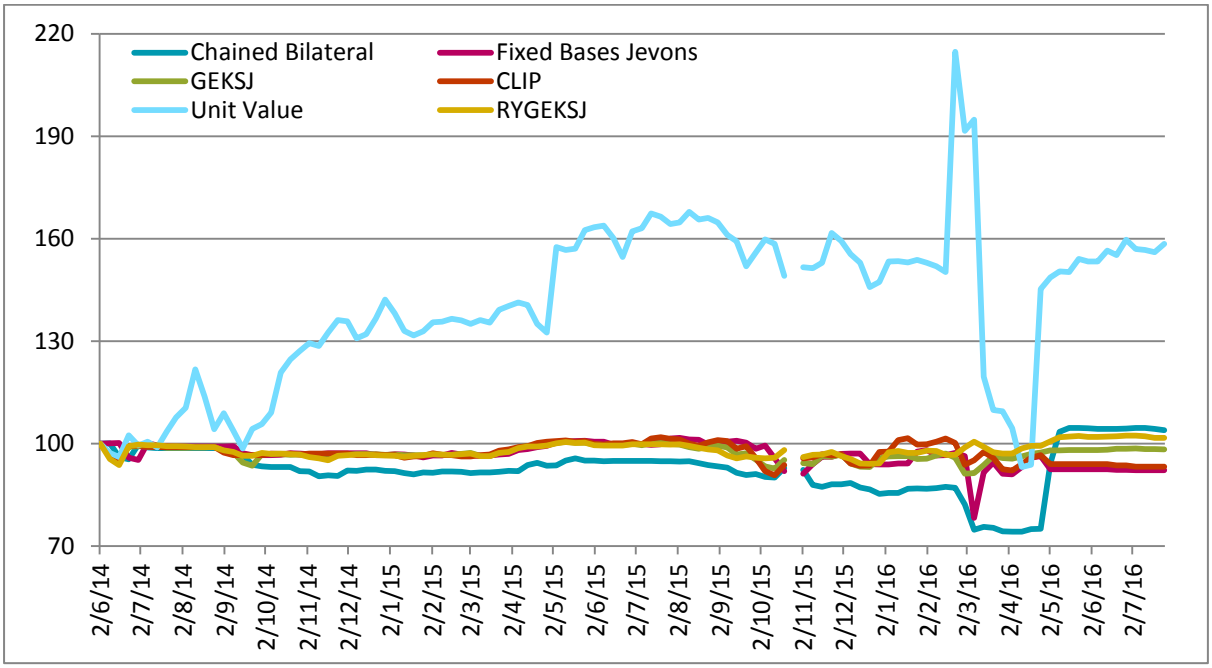
Annex C – Examples of the index series from each index.

The following graphs show the following index number formulae:

1. Fixed Based Jevons Index
2. Chained Bilateral Jevons Index
3. Unit value Index
4. GEKS-J Index
5. RYGEKS-J Index
6. CLIP Index

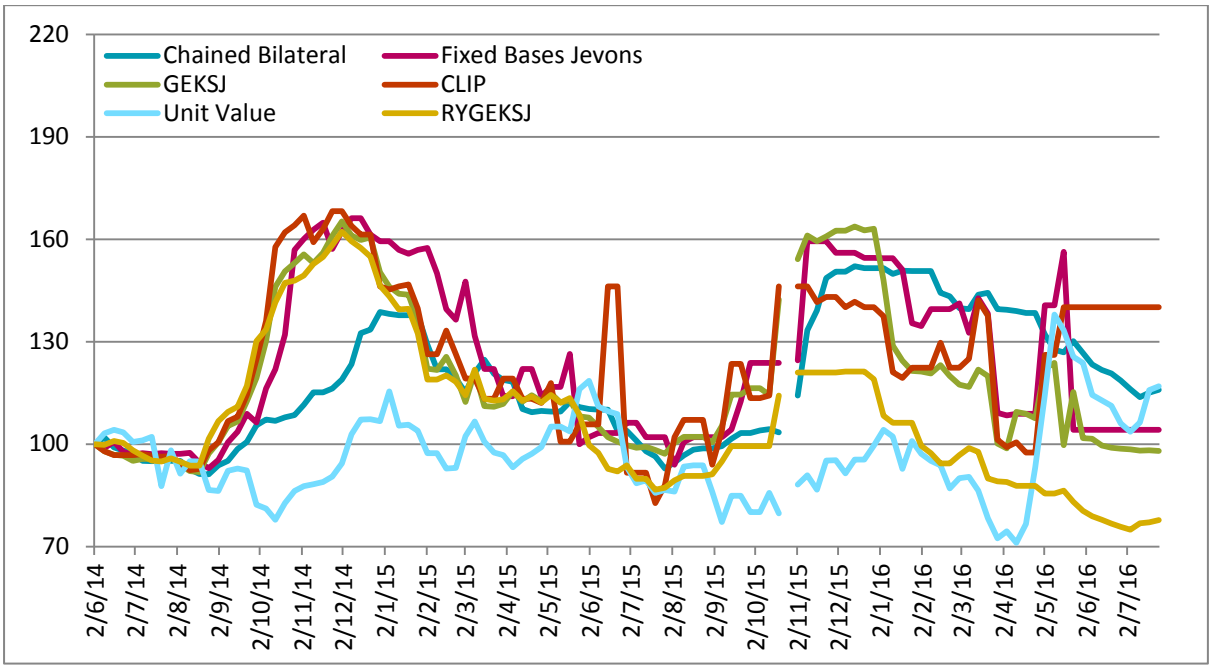
ITRYGEKS has not be calculated due to lack of characteristics needed for the hedonic regressions, FEWS has not calculated as the code to produce the indices is still in development, and IntGEKS has not been calculated for similar reasons.

Figure 1 – Apples (dessert)



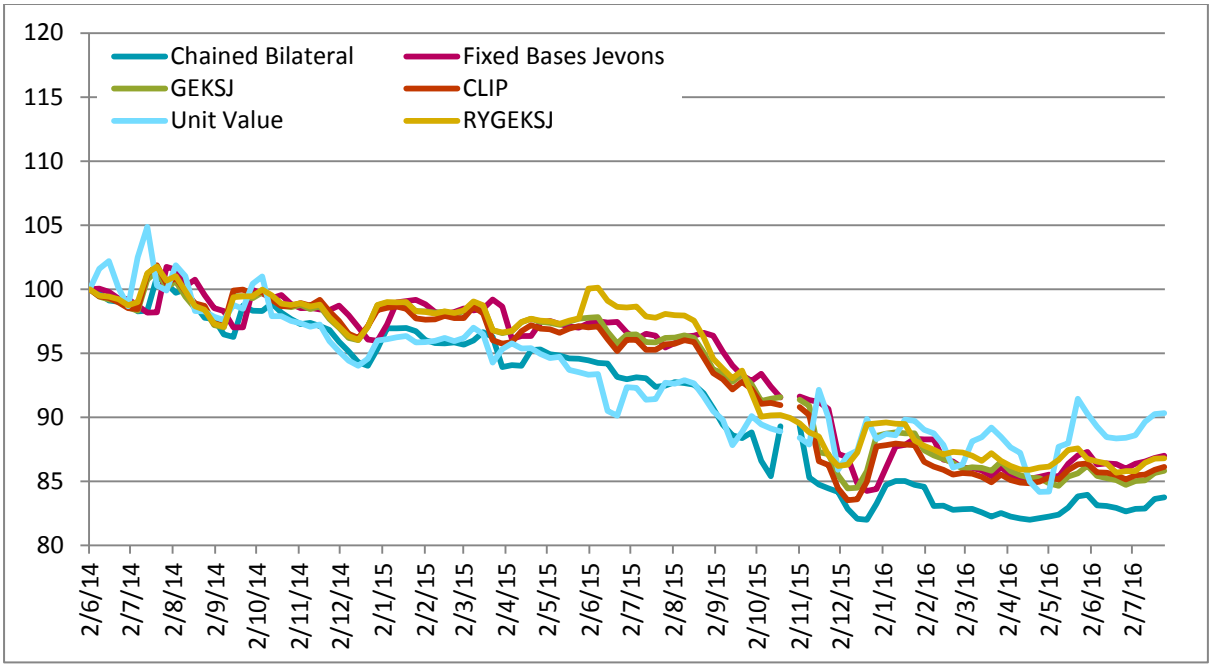
For Apples (dessert), Figure 1, the majority of the indices follow the same pattern, apart from the Unit Value Index, this might be down to more expensive products being introduced later. The RYGEKSJ and GEKSJ have a similar pattern except for towards the end of 2015 about the time the Unit Value index has its sharp jumps.

Figure 2 – Strawberries



The indices seem to be well behaved when calculated for Strawberries, a seasonal item in the web scraped dataset, though the Unit Value and the Chained Bilateral lag behind the other indices when a seasonal peak happens (Figure 2). Again, towards the end of the 2015 the RYGEKS and GEKS are moving further apart.

Figure 3 – Red Wine



For Red Wine, Figure 3, the indices follow similar patterns, though for this one the fixed based Jevons lags behind the rest of the indices by a couple of weeks. The RYGEKSJ changes its behaviour from the GEKSJ soon after the window period, perhaps due to the loss of transitivity.