ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

# Proposed pipeline for processing alternative data sources

Status: final
Expected publication: alongside minutes

## Purpose

1. The Prices Alternative Data Sources project has completed exploratory analysis on suitable methods that can be used to process alternative data sources from raw data to final item level indices.
2. The aim of this preliminary research was to sketch out a proposed end to end pipeline, comprised of individual modules required to process the data, for example 'classification'.
3. This research work has also provided an opportunity to identify some of the main discussion points that will need to be reviewed in the next phase of the project.

## Actions

4. Members of the panel are invited to:
    a) comment on the analysis in Annex A, focusing on the points raised in each of the discussion points sections
    b) advise on the direction of further research

**Tanya Flower**
**Prices, ONS**
**September, 2018**

## List of Annexes

| | |
|---|---|
| **Annex A** | Proposed pipeline for processing alternative data sources |

**Annex A - Proposed pipeline for processing alternative data sources**

**Introduction**

1. Alternative data sources such as web scraped and point of sale scanner price datasets are becoming more commonly available, providing large sources of price data from which measures of consumer inflation could potentially be calculated. The ONS has been carrying out research into these data sources since 2014. The first stage of the project was completed in 2017, with the work summarised in Research indices using web scraped price data: August 2017 update. The recommendations from this work were fed into the development of a Consumer Prices Data Collection strategy, taken to the Advisory Panel on Consumer Prices-Stakeholder (APCP-S) in January 2018 and published alongside the minutes of this meeting.
2. The work summarised in this paper represents the next stage of the project. Over the last few months, the Prices Alternative Data Sources project has completed exploratory analysis on suitable methods that can be used to process alternative data sources from raw data to final item level indices. This research included a literature review of international experience. There is no agreed best practice for processing these data sources yet (although a chapter on scanner data will be included in the next update of the ILO consumer prices manual). However, a number of research papers from different countries have been very helpful in completing this analysis.
3. The aim of this preliminary research was to sketch out a proposed end to end pipeline, comprised of individual modules required to process the data, for example 'classification'. For each module, we looked at the different methods that could be used, and how they differed for the different data sources. This research work has also provided an opportunity to identify some of the discussion points and key questions that will need to be reviewed in the next phase of the project.
4. This paper summarises the work that has been completed on the pipeline so far.

**Data sources**

5. We have 3 possible data sources that could feed into the future basket for consumer prices:

    a. Web-scraped data (i.e. prices collected automatically from online websites, this includes the Robot Tool[1] which checks prices online but does not scrape them directly)
    b. Scanner data (i.e. point of sale data from retailers)
    c. Data as currently collected (including local collections, admin data, and some central collections that can't be replaced by web scraping e.g. phone calls to local services)

6. The status of these datasets are as follows –
    a. Web scraping – we are now receiving web scraped data from mySupermarket. These data are for 30 categories in the basket covering items such as clothing, electronic items and package holidays. There is no back series with these data so we will need to build up a time series before a final impact assessment can be completed.

[1] The robot tool is essentially a price monitoring tool that notifies collectors when there has been a change in price of an item advertised online. It is suited most to items that do not incur a price change on a frequent basis (for instance, 'golf club membership fees'). It is not worth setting up a web scraper for these items because of these infrequent price changes, but the robot tool can monitor these items and notify collectors the day that the price changes.

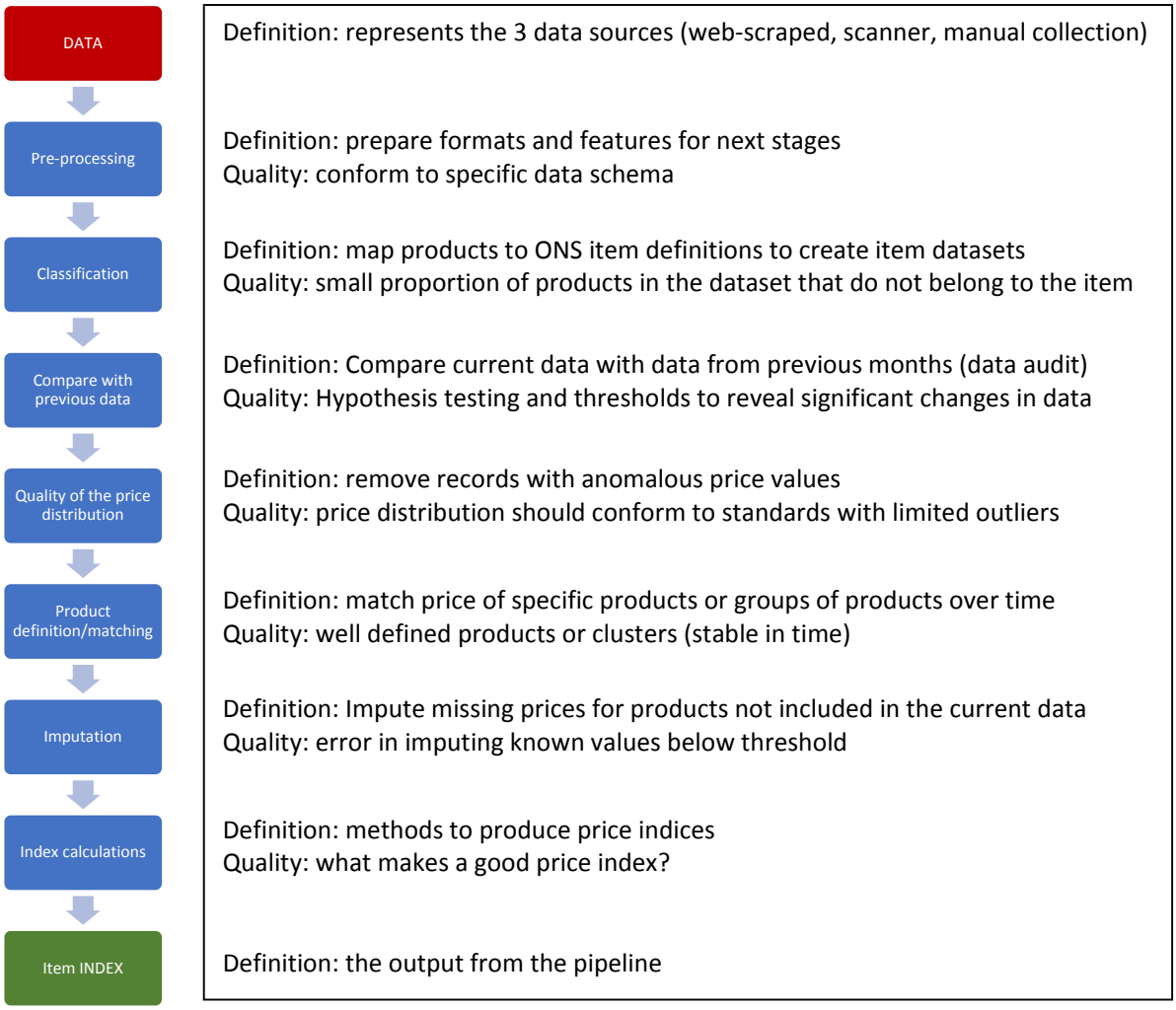
b. Scanner data – we are continuing to engage with retailers on receiving some transaction data, targeting some of the largest retailers that we currently collect prices from. This will cover areas of the basket such as groceries and department stores. We may be able to receive a historical back series with scanner data, so less time may be required for the impact analysis.
c. Manual collection data – we will always need some element of the current manual collection for those items where alternative data is not available or feasible to collect (for example, some local services which are currently collected by phone call each month).

7. An article on the composition of the future basket will be taken to APCP-Technical (APCP-T) in a later meeting.

**Pipeline and methods**

8. Once these datasets are ingested, they then need to be processed and aggregated to a format that can be used by the final production platform. In practice, this means that we need a pipeline that takes the raw input data, processes it, and outputs item level indices which are required as inputs into this final platform.
9. Figure 1 includes all the modules of the pipeline, a short definition/description and how we may ensure quality for each. It should be noted that the final version of the pipeline will be dependent on some of the decisions we need to make in the next phase of the work. For example, certain index methodologies don't require any imputation. The order of these modules may also change depending on the item and methods chosen.

**Figure 1: Flow diagram of proposed pipeline for processing prices data**




10. We have prepared documents for each of these stages which summarises the methods that we could use, what the input/output data may look like at each stage, and suggested quality checks for the output. We have also flagged up potential interdependencies between modules (for example, some price index methods require the price distribution to be checked, others may require the price relative distribution), and any differences that may occur if the module was used for scanner data instead of web scraped data.
11. The next section goes through some of these modules in more detail, indicating some of the key questions and discussion points we need to review in the next stage of the work.

**Module 1: Pre-processing**

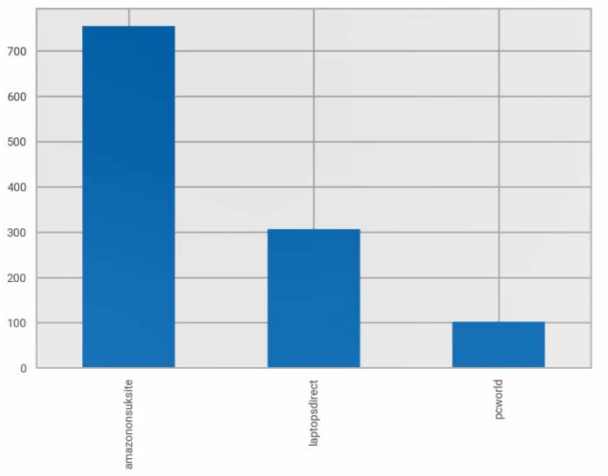
12. Pre-processing is the first stage of the pipeline and is needed to prepare the raw data for the following stages. Without it, the following stages are not going to work properly and will probably crash. It includes the following steps:
    a. Cleaning
    b. Feature engineering
    c. Excessive missing rates and imbalanced variables

13. Each item and data source is going to have its own functions for pre-processing depending on its own unique characteristics. For example, a function that corrects the format for the attribute column “RAM size” is applicable only for laptops. However, some functions may be generic enough to be applicable across all items. For example, a function that flags any column that has more than 50% of its values missing is applicable to any item dataset.
14. So far, we have been using the mySupermarket data to carry out exploratory analysis and assess the data quality of the web scraped data. This includes univariate descriptive statistics and plots. For example, for categorical variables, we check their unique values and frequency distributions using bar charts. We have been exploring if all the retailers we expect are available in the data. We also check if they are equally represented in the data by looking at their frequency distributions:

**Figure 2: Frequency plot for Desktops, by retailer (June 2018)**

```
Number of unique values:  3

amazononsuksite    755
laptopsdirect      306
pcworld            102
```



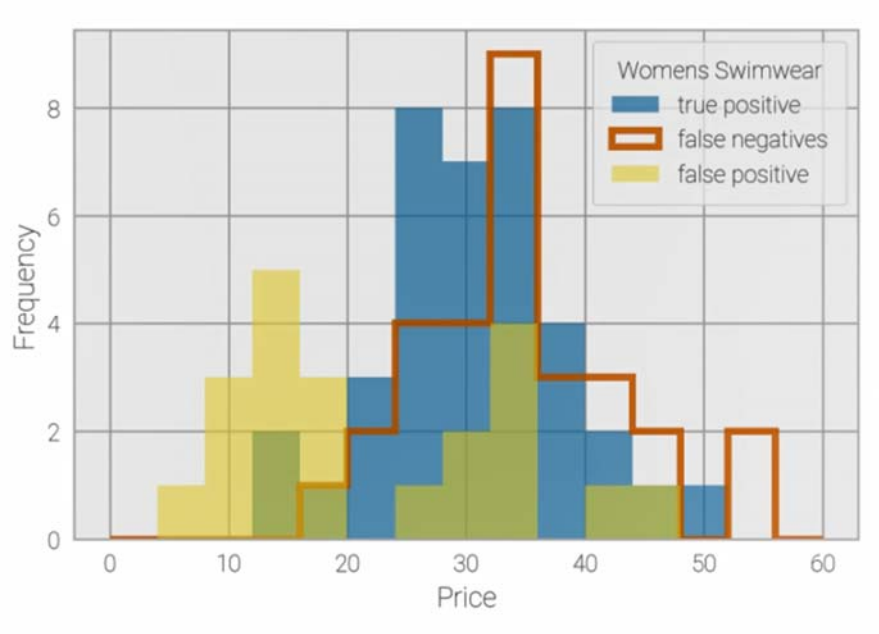

**Module 2: Classification**

15. Classification is about ensuring that we have the right products (rows) in each individual dataset to produce an index for a specific item of the basket i.e. the aim of this stage is to create a specific dataset for each defined ONS item. To achieve that we need to make a decision about whether each product/row in the initial dataset should be included or excluded from the individual item dataset. These items are those that are defined in the concurrent consumer basket. At the moment, we are not considering items that do not fall within this manual collection criteria (see classification questions).
16. The classifier makes a decision depending on whether a product belongs to the item of interest or not and whether the classifier eventually allocated the particular product into the item of interest or not. There are 4 possibilities:

| Belongs to item? / Allocated to item? | Yes | No |
|---|---|---|
| Yes | True positive | False positive |
| No | False negative | True negative |

17. The higher the proportion of true positives, the better the performance of the classifier. That may differ between items. For example, prices for separate women's swimwear bikini bottoms or tops are not collected in the current collection according to the current item definition but the classifier was not able to distinguish them and as a result they were incorrectly included in the item dataset (false positives).
18. The price distribution can also be distorted by false positives and false negatives. This can be illustrated by the histogram below for women's swimwear from the mySupermarket data. The false positives (largely the individual bottoms/tops) have a different price distribution compared with the true positives (the complete sets).

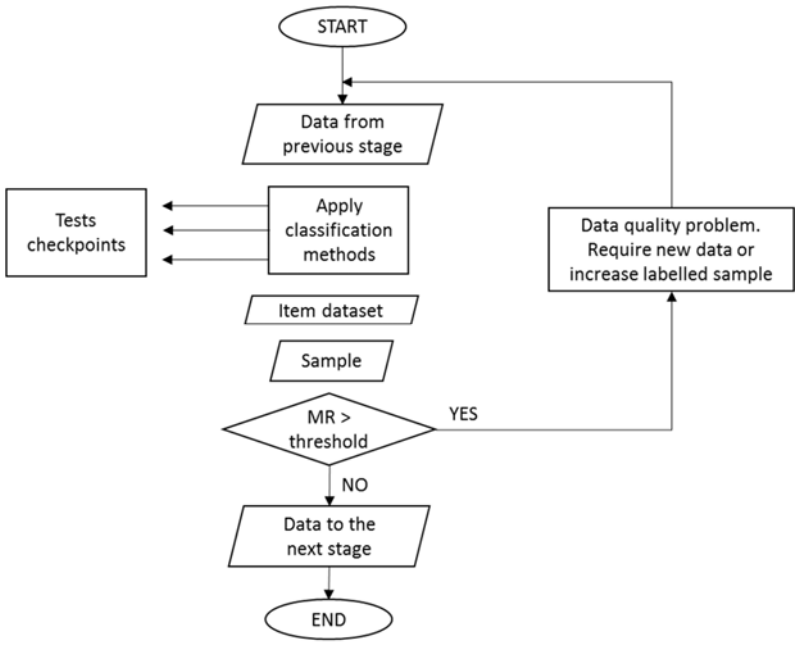
**Figure 3: Price distribution of products labelled as true positives, false negatives and false positives; women's swimwear**




Methods

19. There are a number of methods that can be used to classify the data, ranging from the simplest (mapping the existing retailers' classification to an ONS item definition) to the most complicated (using supervised machine learning techniques such as a support vector machine). The process that could be carried out within the classification module is defined in Figure 4.

**Figure 4: flowchart for the classification process (includes test for false positives where the misclassification rate (MR) is greater than a given threshold, but not test for false negatives)**




Note: The sample here refers to drawing a sample of products to test the misclassification rate. Subject to the misclassification rate from this sample falling below a specified threshold, the full item dataset will pass to the next stage.

Classification discussion points

20. Each of these methods will need to be tested with different parameters and on different items and data sources. The final process will be chosen based on assessing the methods against the given quality checks (for example, testing the false positives rate above).
21. In this stage of the work, we have prioritised the problem of classifying products to the item definitions that are defined by the current manual collection criteria and classification hierarchy (the item level underneath COICOP). However, there are potentially a lot of relevant products which will be dropped using this method. In a future phase, we will need to consider extending the classification structure. Paper APCP-T(18)09 - Extending the classification structure using web scraped data presented at the May 2018 APCP-T describes this in more detail, and includes a list of considerations that will be need to be taken into account before making a final decision about the classification structure.

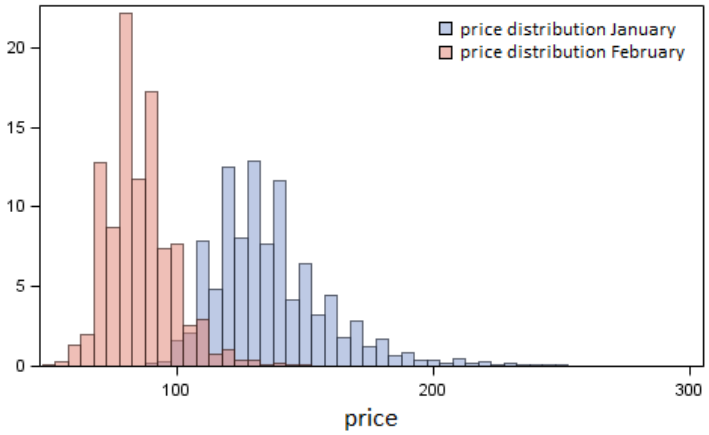
**Module 3: Compare current data with data from previous months (data audit)**

22. Once an item dataset has been created, it is possible to compare it with previous data. This can be useful because if there are any significant changes in the data they may affect the index. We will also be able to explain index movements based on the findings from this stage.
23. This stage of the pipeline is about checking the following information and comparing it with previous data:
    a. Variables distributions:
        i. Numerical
        ii. Class
        iii. Text

b. Correlations between variables
c. Product clusters

24. For example, we can look at the distribution of numerical variables. Price is the most critical variable. If the price distribution in the most recent data is significantly different from the price distribution in the data from the previous month, we may expect this to have an impact on the index, which we can then explain given these tests.
25. Hypothesis testing can be used to compare means or other metrics related to the shape between the two distributions. Comparing the range can also be important although anomalies/outliers in the price distribution will be handled at another stage of the pipeline. This process refers to price but can be applied to any other numerical variable in the data.
26. Figure 5 shows an example. Prices for this item for February have shifted to the left compared to January. That may reflect a drop in the prices for this item on average or it may be an issue with the products that have been scraped for February (e.g. a problem with the scrapers).

**Figure 5: example distributions for an item, comparing January to February data**

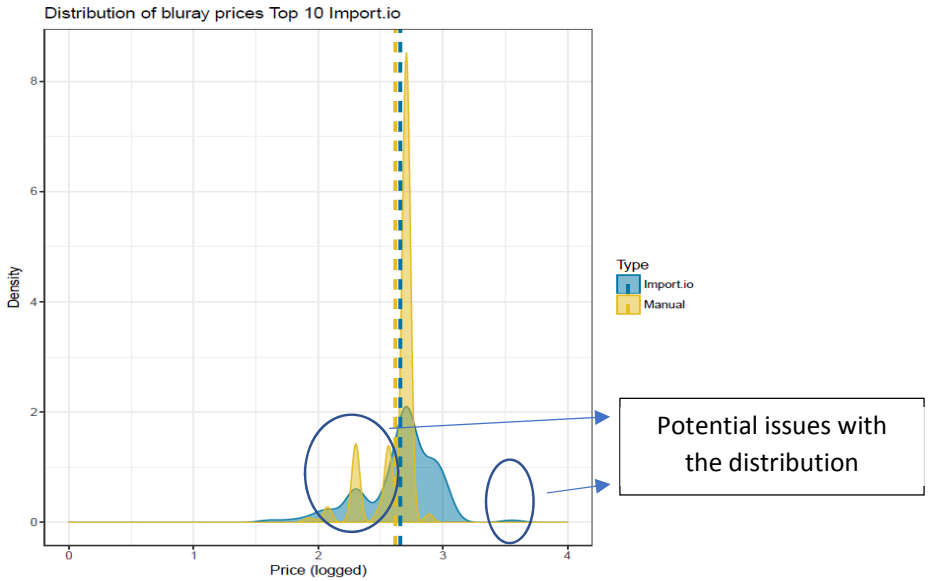



27.

**Module 4: Quality of the price distribution**

28. The quality of the price distributed can be affected in two ways. Firstly, the distribution may be of a different shape to what is assumed by the price index methodology (for example, the dataset could be multimodal (e.g. more than one peak in the data), where we may only want a log-normal dataset). Secondly, the distribution may contain outliers (for example, a pint of milk priced at £50 in the dataset while its real price is £0.50). These outliers may be due to errors in collection. An example is presented in Figure 6.

**Figure 6: Distribution of blu ray prices (web scraped and manual online collection)**




29. Anomaly detection is a technique used to identify unusual patterns that do not follow the expected behaviour (i.e. outliers). The Tukey Algorithm is used in the traditional collection to detect outliers. It works by identifying price movements which differ significantly from the norm for a particular item (Section 3.3.3.4 of the Consumer Prices Technical Manual).
30. The price distribution in Figure 6 is for blu ray discs collected manually and through web scraping. The price distribution consists of multiple peaks and is asymmetric. We assume that the multimodal distribution is not an issue for our index methodology. However, there might be outliers that lie within the multiple peaks and in the tails of the price distribution. The removal of these anomalies can be tested using a sample of products to look at the rate of true and false positives.

Quality of the price distribution discussion points

31. In terms of the prices data, it is not clear that a particular distribution shape (e.g. unimodal or multimodal; normal or log normal etc) is required for a particular chosen index methodology but this assumption needs to be checked further. It is more likely that outliers will affect the index methodology. In the milk example above, the mean of the price relative (price of milk in previous month compared to price of milk in the current month) may be skewed. Therefore, this product should be detected as an anomaly and removed. However, we will need to define what an outlier looks like in the data (e.g. if we use the Tukey algorithm, what parameters should we use?).
32. Some index methods use prices as input (unmatched models), whereas others use price relatives (matched models). Therefore, there is also further work that is required on whether we need to do these checks on prices, on price relatives, or both. Price relatives are also defined differently for certain index methods. For example, a price relative is essentially a ratio of two prices. This ratio could be the ratio of current price over the base price (fixed base Jevons), or current price over previous price from the month before (chained Jevons).

**Module 5: Product matching/definition**

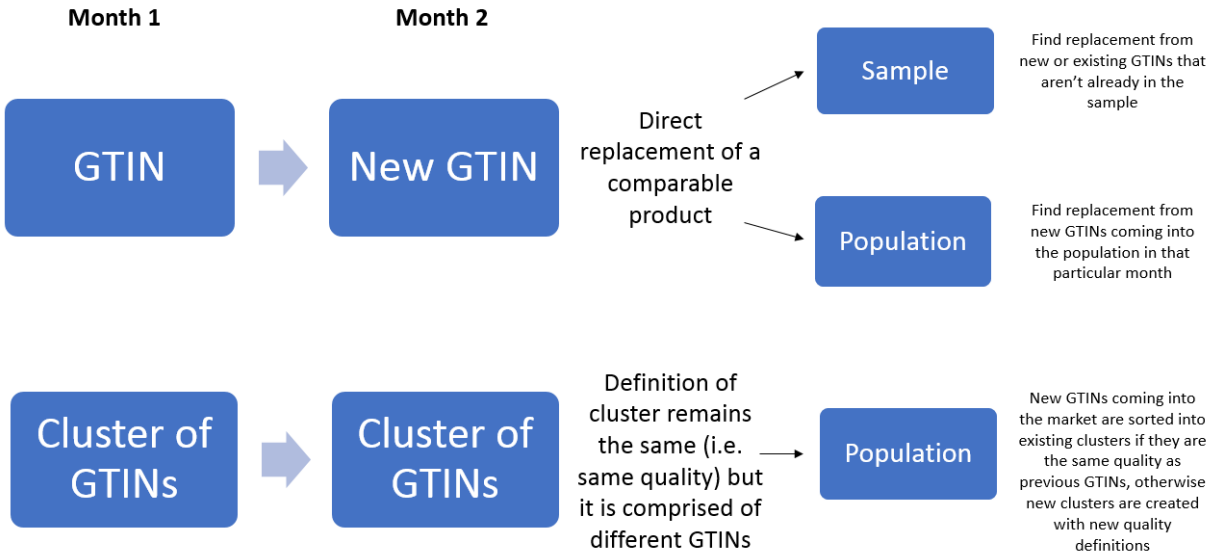
33. In the traditional methodology, an individual product is chosen in the base month by the price collector and then the price of that same product is collected over time. If the product is no longer available, the price collector identifies a comparable product with the same quality characteristics (if possible). If a comparable replacement is not available, a non-comparable replacement is chosen with a suitable quality adjustment applied to the price in the base period.
34. With alternative data sources, it may not be possible to identify a suitable comparable replacement if a product goes out of stock due to the sheer volume of data collected. Generally, a product is defined and linked at the GTIN level (i.e. the individual barcode). However, defining products at this level can cause problems where there are lots of "relaunches". This is where existing products are re-introduced into the stores with a new GTIN, but there are no changes in the quality of this product.
35. For example, such GTIN changes are usually the result of changes to the packaging of items, which may be reshaped in order to fit a new product line, while quality characteristics such as brand, package volume and composition/ ingredients often stay the same. If GTINs are treated as unique products, the price differences between old and new items will then be fully ascribed to quality changes. However, if these new GTINs are of comparable quality to old GTINs, we are then missing this price change. The same problem occurs if new models are introduced which are of comparable quality to existing models in the dataset.
36. Defining products at GTIN level can therefore affect the trend of price indices calculated using matched methods (i.e. use price relatives as input, for example GEKS or traditional bilateral methods). Research has shown that these indices show a strong downward trend when applied to consumer goods with high rates of churn and with prices that decrease after their introduction to the market.

Methods

37. For market segments that experience high rates of product churn, a different approach may be to define a broader concept of product that combines different GTINs of homogenous quality into the same unit. Index methods could then be applied to the average unit value and expenditure of these "clusters", rather than the individual GTIN level. These clusters could be defined using a number of different methods including filtering on product attributes (identified either manually by expert price collectors or automatically by an optimisation algorithm – similar to what is currently being developed by Stats Netherlands), or by using unsupervised machine learning techniques to cluster the data following certain rules (similar to the approach used in the CLIP index developed by ONS).
38. Figure 7 presents three different scenarios for how products could be matched over time. The first two scenarios are based on using the individual GTIN level to define products. In month 2, the GTIN from the previous period is not available in the data. We would like to directly replace this GTIN with a comparable product. There are two possibilities – firstly, if we are using a sample of GTINs (i.e. to replicate the existing methodology), we could draw a comparable replacement from new or existing GTINs that aren't already in the sample (analogous to the methodology used for rental equivalence, detailed in the CPIH compendium). If we are using the population of GTINs, the only option to draw a replacement is from the new GTINs that are coming into the sample in month 2. New GTINs not matched to existing products will be deemed new products into the index.

39. The third scenario in Figure 7 is when we use clusters to define a product. From month 1 to month 2, the definition of the cluster remains the same i.e. it is of comparable quality. However, it may be comprised of different GTINs from month to month depending on the churn in the data. New GTINs coming into the market are sorted into existing clusters if they are of the same quality as previous GTINs, otherwise new clusters are created with new quality definitions. Depending on the index methodology chosen, base prices could then be imputed for these new clusters.

**Figure 7: Different scenarios for matching products over time**




Product definition/matching discussion points

40. Methods to define suitable clusters of homogeneous products are an open question in the international research at the moment. There is a high resource burden attached to manually selecting the filters required to generate clusters. Stats Netherlands have produced some exploratory analysis of an optimisation method that balances the homogeneity of the cluster against the product churn in the data (MARS), but there is still some further research to do on the robustness of this method and application to different data sources. One question is when to define these clusters. Should they be stable across time (e.g. defined in the base period as used in the CLIP index) or can they vary depending on the index methodology chosen.

**Module 6: Imputation**

41. Imputation is a procedure for entering a value for a specific data item where the response is missing. Rather than leaving missing items blank, imputation allows us to reduce non-response bias; manage systematic bias and preserve variance; and preserve relationships between variables.
42. Missing prices can be a problem particularly for price indices constructed from alternative data sources, which may have higher rate of product churn than in the current collection. Most price indices use price relatives as an input into their series, so if a price is missing from one period because it is out of stock on the website, the imputation module allows the product to remain in the population for a period in case it comes back in stock.

43. Imputing prices is a good way to deal with missing prices so that a more consistent sample size is kept over all the period of interest, but sometimes a product may go out of stock for a significant period and either get reintroduced or removed from the market all together. Therefore, it may be unwise to continually impute the prices in either of these situations, as the index may not be representative of actual price movements.

Imputation methods

44. While there are many methods for imputation, the three methods below have been tested previously in Imputing web scraped prices. This work found that carrying forward the previous price minimised the relative imputation bias.

    a. Carry forward the previous price
    b. Impute the mean by store or item type, using:
        i. Arithmetic mean
        ii. Geometric mean
        iii. Harmonic mean

       Note: It should be noted though that the use of average price should be used as a last resort as it artificially reduces the estimate of variance.
    c. Impute the average growths of the rest of the items then multiply this by the previous price, using:
        i. Arithmetic mean
        ii. Geometric mean
        iii. Harmonic mean

Imputation discussion points

45. The choice of whether to impute or not is highly interdependent with the index methodology chosen (i.e. matched or unmatched). Some indices use price levels in which case imputation is not necessarily required. However, indices which use a fixed basket approach and use the ratio of prices from the current month over the base month will require imputation. For example, the multilateral indices may not require imputation as they take multiple periods into account and do not follow a fixed basket approach (e.g. GEKS). Additionally, some multilateral indices (for example, the ITRYGEKS, FEWS) impute automatically while implicitly or explicitly quality adjusting prices.

**Module 7: Index calculations**

46. The index number methods module takes the cleaned and classified data from previous modules in the pipeline. It uses price and expenditure data (where available) to calculate a price index at the elementary aggregate level. The different methods that can be used to create price indices from alternative data sources have been well documented in the international literature over the last few years.

Definitions

47. Direct bilateral indices: compare prices from the current period relative to an earlier base period (e.g. period 0 to 1, period 0 to 2); problems include product churn decreasing the number of matched products over time.

48. Indirect (chained) bilateral indices: compare prices from consecutive time periods (e.g. period 0 to 1, period 1 to 2) which can be chained together to form a continuous series; overcome the issue of product churn but these typically suffer from chain drift
49. Multilateral indices: compare prices across 3 or more time periods (e.g. for a price change between periods 0 and 2, you could include price changes from period 0 to 2, period 0 to 1 and period 1 to 2); overcome chain drift issue and sample attrition but depending on time period could experience loss of characteristicity.

Extension methods

50. When a multilateral method is used to produce a temporal index, each bilateral price comparison depends on prices observed in other periods of the multilateral comparison window. As a result, incorporating a new period into the multilateral comparison window may alter the price comparisons of earlier periods.
51. Therefore, to overcome the issues of revisions and characteristicity, we also need an extension method for multilateral indices. This allows for new data to be incorporated without the need to revise, by chaining the new price change onto the existing time series. The choice of "link" period determines which extension method is used.
52. The choice of length for the estimation and splicing windows has generally defaulted to 13 months, but further research by ABS and Stats New Zealand have suggested different window lengths may be more suitable for different items (for example, ABS use 25 months for their groceries data, and Stats New Zealand use an 8-year window for rental prices).
53. An example of an extension method is the movement splice, which uses the previous period index as the pivot and applies the price movement estimated from the new multilateral window. Other methods include the direct extension, mean splice and window splice. For more information, please see the multilateral extension methods section in this paper by ABS.

Index methods

54. The tables below contain a summary of index methods that we might use in practice, split into bilateral and multilateral methods. There are variants of these methods which we have not covered here, but are not used elsewhere and we would be unlikely to use in practice.
55. The tables also contain information on examples of where they have been used internationally, and whether the methods need expenditure data. Those that require expenditure data may still be able to be used for web scraped data, subject to an appropriate proxy expenditure weight being found.

*Bilateral indices*

| **Bilateral method** | **Need expenditure data?** | **Application** |
|---|---|---|
| Jevons | N | Elementary aggregates in current CPI methodology. The chained bilateral Jevons (i.e. the indirect version) is included as "dynamic method" in Eurostat guidance (2017) on processing scanner data and is also used by most countries in Europe for supermarket scanner data. |
| Dutot | N | Elementary aggregates in current CPI methodology. The chained bilateral Dutot Is used by Statistics Netherlands for web scraped clothing and footwear data (Griffioen and ten Bosch, 2016). |
| Lowe ("Laspeyres type") | Y | Item-level and above aggregation in current CPI methodology. |

| | | |
|---|---|---|
| Fisher | Y | No CPI applications as bilateral method are known, but it is usually combined with the GEKS method. The Törnqvist index is often preferred because it is easier to work with in the GEKS from an analytical point of view. |
| Törnqvist | Y | It is usually combined with the GEKS method due to data availability. |
| Unit value index | N | No CPI applications are known, but it is part of the CLIP index developed by ONS. |

Note: (1) indices that use expenditure data (for example, Laspeyres) are not transitive and therefore can suffer from chain drift when they are chained monthly.

*Multilateral indices*

56. In theory, each extension method can be applied to each multilateral index although this may not always be appropriate. Where the two coincide to produce a recognised index, this is also noted e.g. the window splice on the time product dummy is equivalent to the FEWS index.

| **Multilateral method** | **Expenditure data needed?** | **Application** | **Extension method**[1] | | | |
|---|---|---|---|---|---|---|
| | | | Monthly expanding | Movement Splice | Window Splice | Mean Splice |
| Time product dummy (TPD) | Y/N: Can be weighted or unweighted. | Rental prices, Stats New Zealand (Bentley, 2018) | | | FEWS | |
| Time dummy hedonic (TD) | Y/N: Can be weighted or unweighted; also needs detailed attribute info. | Is usually considered for consumer electronics. | | | TDWS | |
| Geary-Khamis (QU)(2) | Y | CBS use this method now for almost all transaction data, also supermarkets (Chessa et al., 2018) | FBME | | | |
| CCDI | Y | Grocery data, ABS (ABS, 2016) | | | | Used by ABS |
| GEKS-Törnqvist (imputed) | Y | Electronics, Stats New Zealand (2014) | | ITRYGEKS | | |
| GEKS-Jevons(3) | N | | | RYGEKS-J | | |

Notes: (1) Other extension methods could potentially be applied, with different linking periods, but these have not been considered in the literature and we're unlikely to use them in practice. An exception is the half splice method (de Haan, 2015).

(2) The Geary-Khamis method is a special case of the family of Quality Adjusted Unit Value (QU or QAUV) index methods. Different methods of standardising quantities can be thought of, but these are not used internationally and have shown to give marginally different results compared to GK (Chessa, 2016a).

(3) Although the GEKS-Jevons is not used in the CPI, it is used in a number of comparative studies (de Haan and van der Grient, 2011; Van Loon and Roels, 2018).

Implementation

57. There are a number of stages that could be implemented in tandem or separately for this module. Should we decide to take a phased approach to the implementation of these methods, then we could likely move across in up to three stages. These stages could also be

staggered across different COICOP groups so that we move to subsequent stages one group at a time. The possible stages are:

d. Try to replicate the existing methodology: sample items on or around index day using probability proportional to expenditure sampling. If possible, the sample should be geographically stratified and we should aim to sample the same number of items as are collected in the local collection. Use a Jevons, Dutot or Carli methodology with current to base period (fixed base) comparisons. This will require a methodology for identifying comparable replacements; however, such a method should be applicable across items given sufficient metadata.
e. Adopt a similar methodology: for example, a Jevons, Dutot, or Carli using current to previous period comparisons and chaining monthly. This could be based on the whole dataset or it could be based on a sample, whether drawn on a similar basis to point 1 or otherwise. This approach should only be adopted where product churn is demonstrably low, so that any chain drift from changing samples is minimised.
f. Introduce expenditure weights: this could be on a basis consistent with how higher-level aggregates are calculated; for example, a Lowe or Laspeyres weighting scheme. This should be on a fixed-base basis where a method for identifying comparable replacements can be implemented, as frequently chaining a weighted index could lead to significant chain drift.
g. Introduce expenditure weights on a basis consistent with the target multilateral method; for example, average expenditure shares if we intend to implement a GEKS-Törnqvist method in the future. As with point 3 this should also be on a fixed base basis.
h. Implement multilateral methods:
    i. Geary-Khamis(QU), ITRYGEKS, and TDWS methods all require attribute data and an analyst's time to build models. Therefore, these approaches should be deployed strategically. ITRYGEKS and TDWS are most applicable where product churn is high; for example, electronic goods. Geary-Khamis(QU) is also an approach that controls for quality changes associated with incoming products into the sample, but is potentially less resource intensive and so could be applied to a wider selection of items than hedonic methods. The quality adjustment review currently being undertaken by the Prices Development team should provide some guidance on the best approach to applying quality adjustment methodologies.
    ii. FEWS is not resource intensive and does not require additional attribute data; however, in cases where differences in quality are negligible and product churn is slow, this approach may not add much value.
    iii. GEKS and RYGEKS make no imputation for the changing quality of the sample. Therefore, we would need to be relatively confident that quality change is negligible in any given category, or that the impact of any quality changes on higher levels is minimal.

58. The aim of using a staged approach to implementing alternative data sources is to minimise the impact for users. There is an implementation trade-off between the quality of the indices and the impact. The longer the staging process the more minimal the impacts; however, we are not making full use of new data sources, and we could easily fall behind other NSIs.
59. The stages should all be run in parallel from the outset to allow Prices Division to assess the impact that each stage shift will have. The staging can then be designed in such a way that any impact on headline indices is effectively smoothed across the transition period.

60. An alternative approach could be to stage according to a particular group's importance (i.e. weight); however, given that the aim of staging is to minimise impact this would not be the optimal approach as we might expect impacts to be larger for higher weight items.
61. Staging might also involve switching from web scraped to scanner data sources (for example, we are working on web scraped clothing data at the moment but we may get scanner data from clothing retailers in future). In this case the staging approach outlined above can still be used, but an additional stage could be introduced where we switch from a web-scraped Jevons (for example) to a scanner data Jevons using the same methodology. This is conceptually equivalent to switching from ticket prices to sale prices (the price actually paid rather than the advertised price).

Index calculations discussion points

62. This suggested implementation plan raises some important questions, including how long should the implementation phase be and whether we need to use all of the proposed stages above. There is also discussion required around what COICOP level should we apply stages to, e.g. item level or higher.
63. As well as the process used to implement these alternative data sources in production, there are other questions we need to consider. For example, chain linking may interact with the extension methods proposed above (in particular, how is the double chain link affected by splicing?).
64. Another area of interest is how these alternative data sources interact with one another. For example, laptops is a single item that is currently collected by one mode of collection – online. We can therefore reproduce an item level index by just using web scraped data. However, there are some items (for example, groceries) where local collection is supplemented by manual online collection. Scanner data may be able to replace all these data, or it might just be able to replace some of the quotes and we may still need data for these items from the local collection (for example, independent grocers). There is a question of how these data sources are aggregated up together, possibly using different index methods.
65. A number of index methods use regression models as input (for example, the TDWS). Do these regression models need to be rebuilt every period to reflect the new window, or can they be updated less frequently, e.g. in line with current practice? These models generally need manual analyst resource to determine, so constructing these models may need to happen outside of the main processing pipeline. It is also unclear how higher level product definitions (e.g. clusters of homogeneous products) could be fed through hedonic models.
66. Introducing expenditure weights from scanner data at the elementary aggregate level will be a big change for the index and would generally infer that expenditure information should be updated month to month, but how did this impact on the weights at a higher level of aggregation?