ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

Assessing the feasibility of web scraped data within the current collection methods

Status: final
Expected publication: alongside minutes

Purpose

1. The aim of this paper is to discuss the feasibility of web scraped data to supplement the current collection process for the centrally collected items of the CPIH basket of goods and services.

Actions

2. Members of the Panel are invited to comment on:

   a) the overall methodology ONS is undertaking to carry out this feasibility study and the approach towards using web scraped data;
   b) the items that are being used and tested as part of the feasibility study;
   c) the nature of web scraped data and the bespoke cleaning techniques used to validate the data;
   d) the frequency of collection

Background

3. Currently, price quotes are collected via two methods: 'local' and 'central' collections. Local/field collections are those that require a price collector to manually obtain prices for items by physically visiting outlets. Central collections are performed manually by ONS price collectors (emails, phone, CDs, brochures and online collection). There are about 730 items in total in the CPIH basket, out of which 74% of the items are collected locally, and the remaining 26% are collected centrally by ONS price collectors.

4. Around 40% of price quotes collected though the central collections are able to be web scraped. At present, the price collectors would manually have to visit websites or search CDs or brochures to obtain prices for individual items in the basket. The introduction of web scraping (automated collection of data via point and click software) would automate the process of price collection, which could potentially introduce substantial time and cost efficiencies.

5. Additionally, at present due to time and resource constraints, the price collectors are capped at the number of price quotes they can obtain each month for each item. However, automating the process implies an increase in the number of observations that can be obtained each month, resulting in a greater coverage. This can be in terms of more products observed or more time periods covered (for example, a daily collection).

6. Presently the items that have been selected for this pilot study are:

   - chart based collections (cds, blu-ray discs and dvds) – since Jan 2016, COICOP weight= 1/1000
   - PC peripherals (printers and routers) – since Jan 2016, COICOP weight= 1.6/1000

- laptops – since Feb 2017, COICOP weight = 1.25/1000
- package holidays –since November 2015, COICOP weight= 25.48/1000

7. The collections are primarily made through a web scraping tool called import.io. Additionally, the project is also exploring other web scraping tools such as dexi.io, to determine which product might be best for the purpose of collection (for example, ease of use for collectors and the flexibility of the scraping tool are considerations). Future data collection methods will require flexibility and adaption to new and emerging technology.
8. These items were chosen because they provided a good range of different test cases (for example, package holidays are quite a volatile category so investigating a more frequent data source could provide some useful insight; PC peripherals have quite stable prices but require collection from a number of different websites, therefore the intention was to explore if the current collection methods can be replicated with less resource requirements).
9. The intention is to further expand the collection to include items such as 'airfares' that are challenging and time consuming to obtain manually, and therefore larger sample sizes on a more frequent basis could improve the quality of the index.
10. The aim of the project is not to replicate the current collection process, but to test if web scraped data can be used in ways to improve the statistics (for example, to supplement it with increased data coverage on a more frequent basis). While the idea is to map the process as closely as possible to the current manual collections, due to the nature of web scraped and large datasets in general, replicating the manual process may not be the optimum solution. For example, mapping the manual collection process has not been entirely feasible primarily due to compliance issues with websites 'terms and conditions'. Many websites strictly prohibit the use of automated robots or web scrapers, which limits the websites available to scrape prices from. Current methods of validation and scrutiny will also not be able to be matched given the increase in the amount of data to check. Therefore, in this pilot we have taken a slightly different approach to the manual collections.

11. For more information about the central collections project, please see Section 6 of our latest research update.

Methodology

12. The process for the items being collected above is as follows:

    a) Chart Collections (cds, blu-ray discs & dvds) – as part of the manual collections, top 10 chart positions are obtained from an online charts provider, prices for which are obtained online from retailers. It is assumed that the top 10 chart positions are fully representative of consumer purchases, and are capturing the majority of expenditure weights. Due to a lack of expenditure data it is hard to comment on the

accuracy of this assumption. As part of the pilot study, top 100 positions are being web scraped from an online charts provider, prices for which are then obtained from fewer retailers. The reduction in the number of retailers being included is due to compliance with terms & conditions.

b) PC peripherals- prices for 'printers' and 'routers' are collected online from five retailers. This is similar to the manual collections, with the only difference being the sample sizes, which are significantly greater for the web scraped collection. As part of the manual collections, 'Printers' and 'Routers' are collected separately, but clustered together to form one price index for 'PC peripherals'. This accumulates a total weight of 1.6/1000 in the CPIH basket, and items are not weighted on an individual level. This is simply because weights are only available on an item level that is, 'PC peripherals', and not on a product level. Prices for both items are manually scrutinised by price collectors, before index construction. Similarly, in the case of web scraped data, prices for both the items are being collected together and the data for both the different items are clustered together to represent one price index that is, 'PC peripherals'. However, given the differences in sample sizes and the cleaning methods for both the manual and web scraped data it has been recommended to separate the collection of these items, and to construct separate price indices. This would be more representative of the price variations experienced by each item, given the greater sample sizes.

c) Laptops- collections are primarily made online from one retailer. This is different to the manual laptop collections, which are collected online over a range of five to six retailers (again, compliance with website terms & conditions has meant we have been unable to cover the full set of retailers). The web scraped data includes the associated attributes, which are matched to attributes that are selected when manually collecting prices for laptops. These attributes are used in the manual collection for the hedonic regression models for laptops. It is hoped that the web scraped data will be used to form these regression models instead as it is much more efficient, but currently we are not able to cover the full range of retailers.

d) Package holidays – a range of retailers are covered following a similar process to the manual collection (that is, holidays feed into the index month for when they actually are, rather for when they were purchased). However, the current manual collection is largely done using brochures and therefore can't be compared explicitly with the web scraped data.

Cleaning & validation

13. Initially the cleaning and validation for all web scraped items (excluding package holidays) was completed using the outlier detection method 'The Tukey Algorithm'. This was chosen because it was more suited to the non-parametric nature of the datasets, which followed multimodal distributions (Annex A).

14. Further work has shown that using 'The Tukey Algorithm' on these items may not be the best solution. For example, with the chart collections, the top (1st) position could be the

highest price in the dataset, while the bottom (100th) position will be the lowest price in the dataset. There is a risk that these prices would be flagged up by 'The Tukey Algorithm' as outliers. Therefore, using 'The Tukey Algorithm' in this instance introduces the risk of excluding actual prices from the dataset. To combat this risk, a substantial amount of manual scrutiny is required on the flagged prices. However given the greater sample size for the web scraped data, manually scrutinising the prices will be time consuming and will take away from the benefits of web scraping.

15. As a result, the intention is now to develop bespoke validation techniques for each of the different items being collected, which better suit the needs of each individual dataset without introducing a potential bias. This is because the needs of each individual item dataset are so different from one another, due to differences in sample sizes and attributes that they each require tailored validation techniques. One option is to use the error detection method used on the web scraped grocery data (detailed in section 5 of our research paper).

Limitations & propositions

16. The limitations to the current approach for each item is as follows:

    a) Chart Collections- the increase in the sample size to top 100 may not actually be representative of consumer purchases. It may be the case that the top 10 are capturing consumer purchasing trends, and the increase in sample size does not add much further value. Similarly, the increase in sample size may also include products that are not comparable with the other products. For instance, the collection for Blu-ray and DVDs may also be including 'Box-sets', which are set at a higher price than individual items. A possible solution is to identify the box-sets and exclude them from the dataset, to avoid any potential bias. Additionally, obtaining expenditure data would be a huge advantage as it will enable us to analyse whether the top 10 chart positions are capturing the maximum expenditure weight, in which case the increased sample size may not be needed.

    b) PC peripherals- similar to the above, the increase in sample size also implies the inclusion of some unwanted items that may not necessarily represent consumer purchases. In this instance it is the inclusion of 'office printers' (which are priced significantly higher than the day-to-day printers and wouldn't be deemed as consumption goods) that may introduce unwanted inflation within the web scraped price index. It has been recommended that the collection is separated by the two different items being collected; 'Routers' and 'Printers'. Similar to the chart collections, these datasets could possibly be cleaned and validated through a clustering algorithm, which forms data clusters based on the various item attributes, for instance 'item description'. This will enable identification of unwanted items such as 'office printers', which can be removed from the dataset once identified.

    c) Laptops- a recommendation has been put forward to expand the retailers we are currently web scraping from, to more closely map the process to the manual collections. However, at present numerous challenges remain around the websites

terms and conditions, where some websites specifically prohibit the use of 'automated web scrapers' to store and capture their data. This majorly limits us in regards to the number of online retailers that the prices can be web scraped from.

d) At present work on 'Package Holidays' is ongoing and no robust analysis has been undertaken yet. However, we intend to expand our analysis by testing bespoke validation techniques as well as constructing a price index that better suits the needs of the web scraped data.

Frequency of collection

17. At present the frequency of collection for the web scraped data collection matches that of the manual collections (that is, prices are web scraped on the second or third Tuesday of each month, index day)

    a) Should the frequency of collection remain to be monthly? Or should this expand to be a weekly or daily collection?

    b) Should the frequency of collection be specific to the item? For instance, the more volatile items should be collected more frequently, whereas the less volatile could remain to be collected monthly?

Himanshi Bhardwaj
Prices, ONS
September 2017

List of Annexes

Annex A	Distribution charts for the web scraped and manually collected items The distributions below are drawn from raw web scraped and manual data for all items in the pilot (excluding package holidays). These distributions are based on logged prices which helped in reducing a positive skew and asymmetries in the data, particularly for the web scraped data. It is apparent from the distributions that both the datasets experience multiple peaks and follow a multimodal distribution. Multiple peaks are more evident in the web scraped data; this is primarily due to the existence of anomalous prices as these distributions are based on raw data. In contrast, the manual data experiences narrow peaks, reflecting its smaller sample sizes.

Annex A – Distribution charts for the web scraped and manually collected items

Figure 1: Histogram for web scraped and manually-collected prices of 'Blu-ray discs'

January 2016 to May 2017

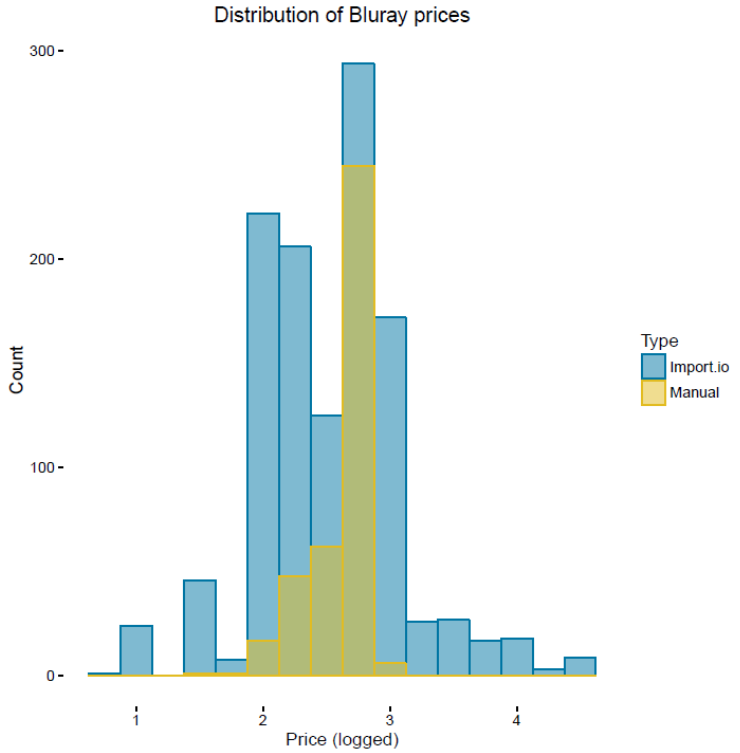

Figure 2: Kernel density function for web scraped and manually-collected prices of 'Blu-ray discs'

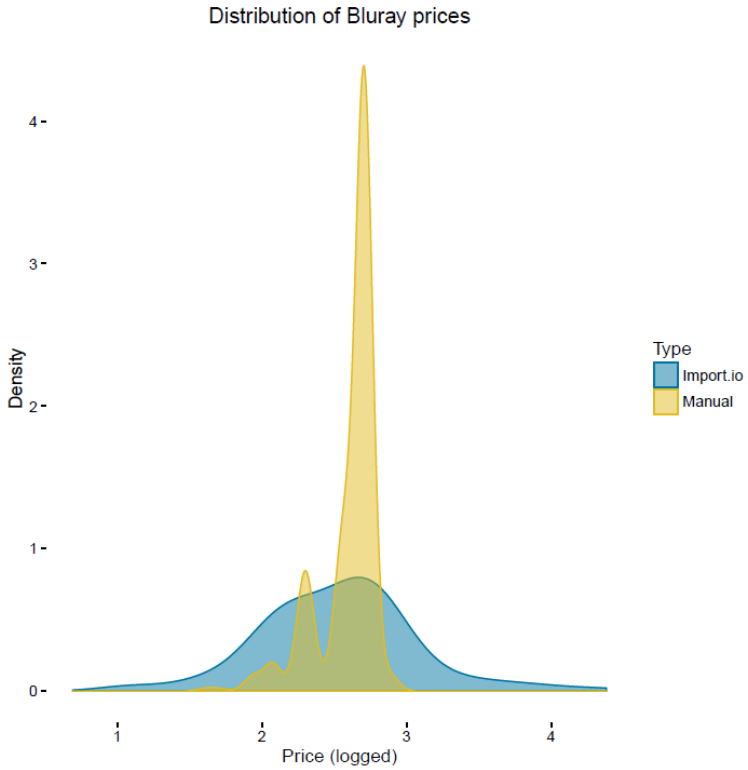

Figure 3: Histogram for web scraped and manually-collected prices of ‘compact discs’

January 2016 to May 2017

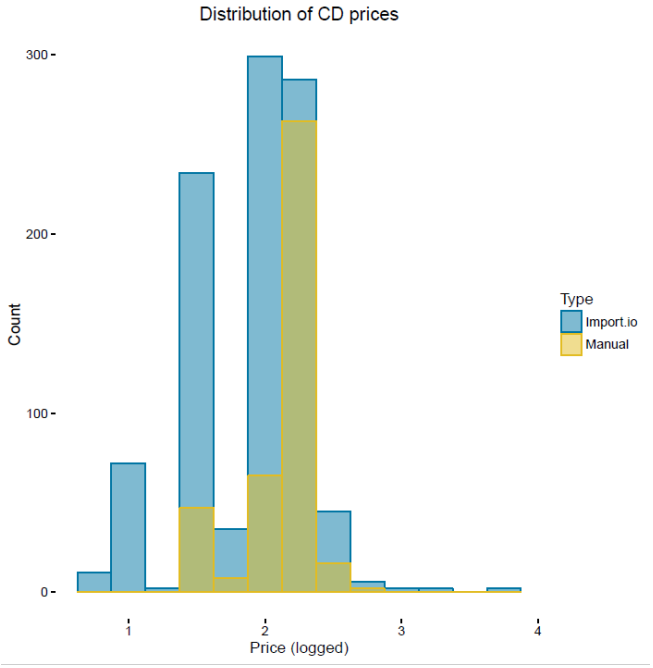

Figure 4: Kernel density distribution for web scraped and manually-collected prices of ‘compact discs’

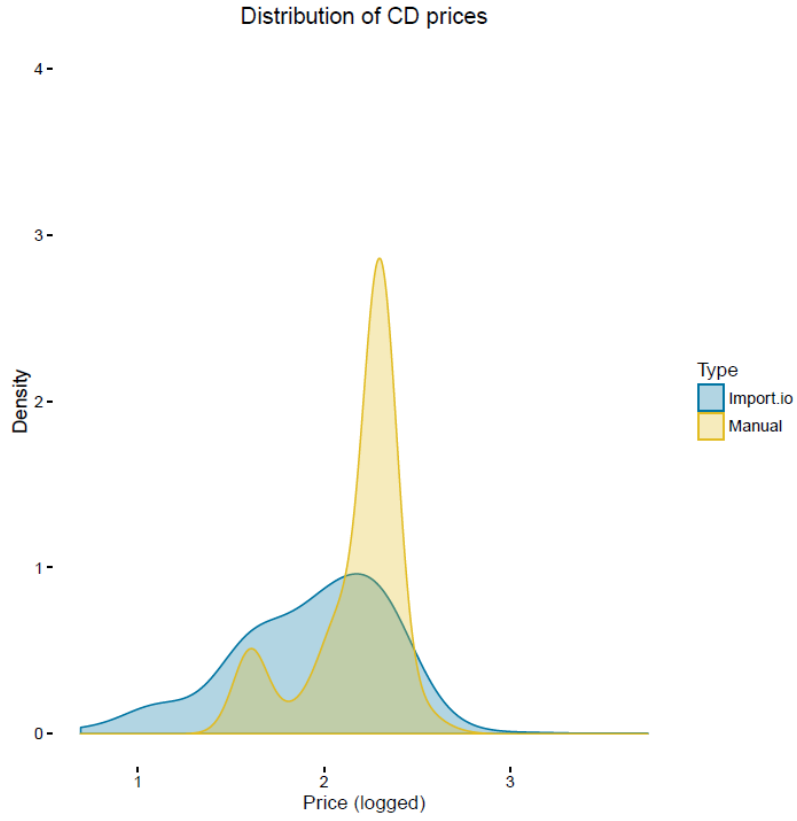

Figure 5: Histogram for web scraped and manually-collected prices of 'DVDs'

January 2016 to May 2017

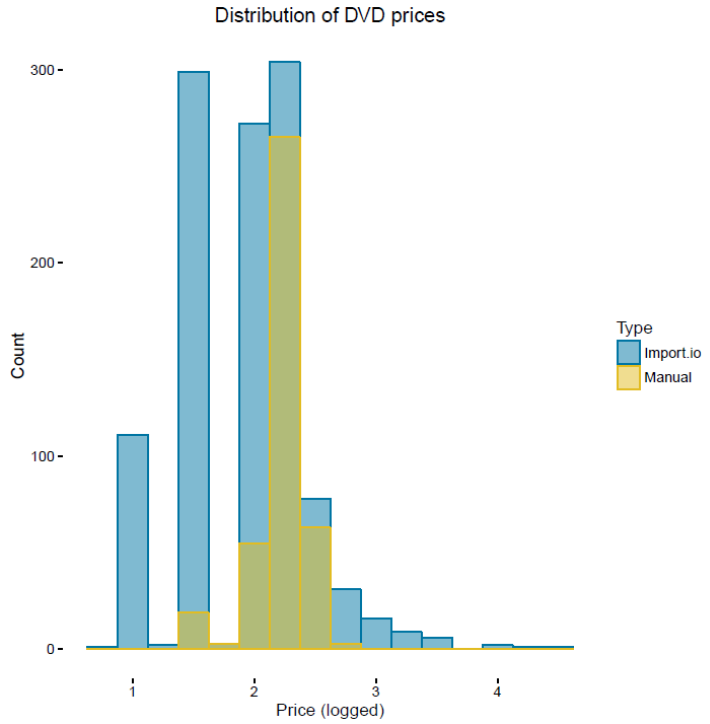

Figure 6: Kernel density distribution for web scraped and manually-collected prices of 'DVDs'

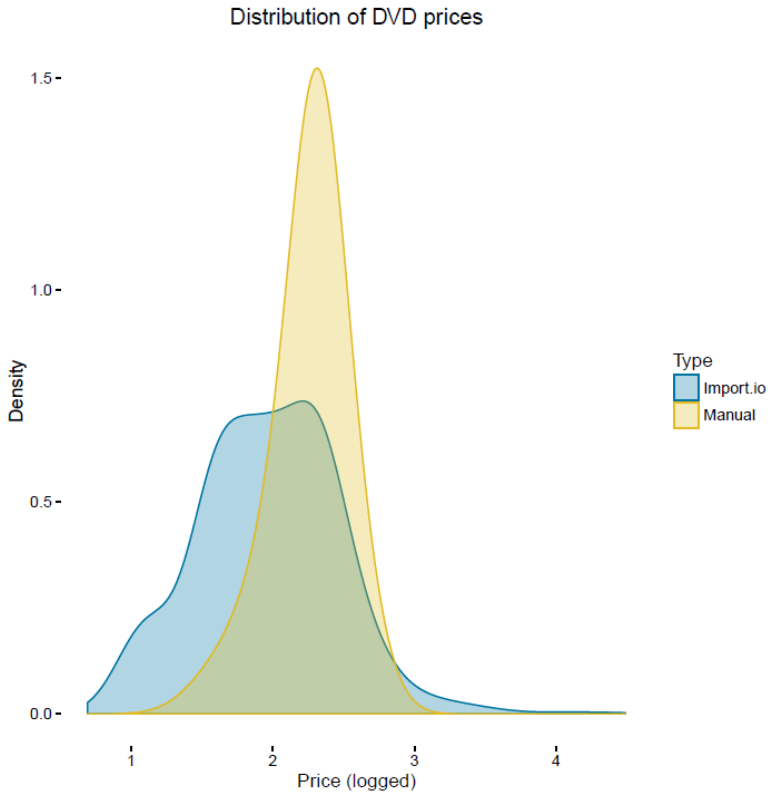

Figure 7: Histogram for web scraped and manually-collected prices of 'PC peripherals'

January 2016 to May 2017

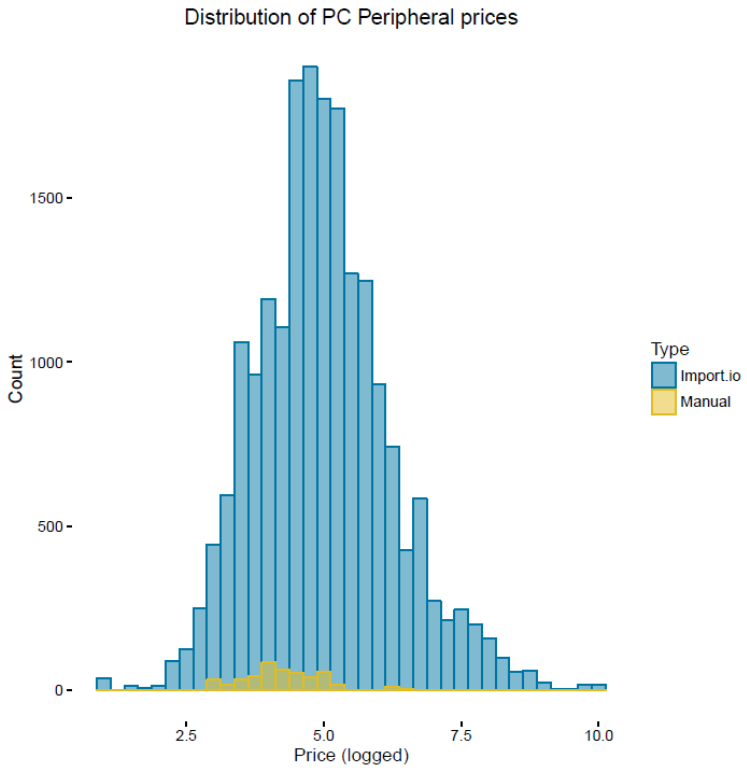

Figure 8: Kernel density distribution for web scraped and manually-collected of 'PC peripherals'

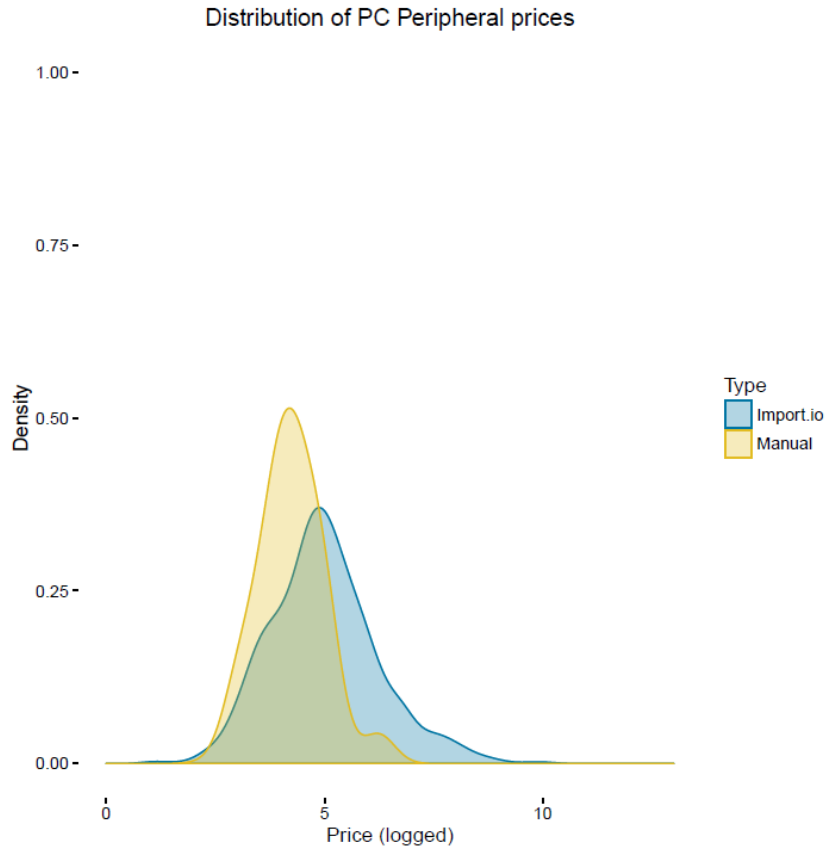

Figure 9: Histogram for web scraped and manually-collected prices of 'laptops'

February 2017 to May 2017

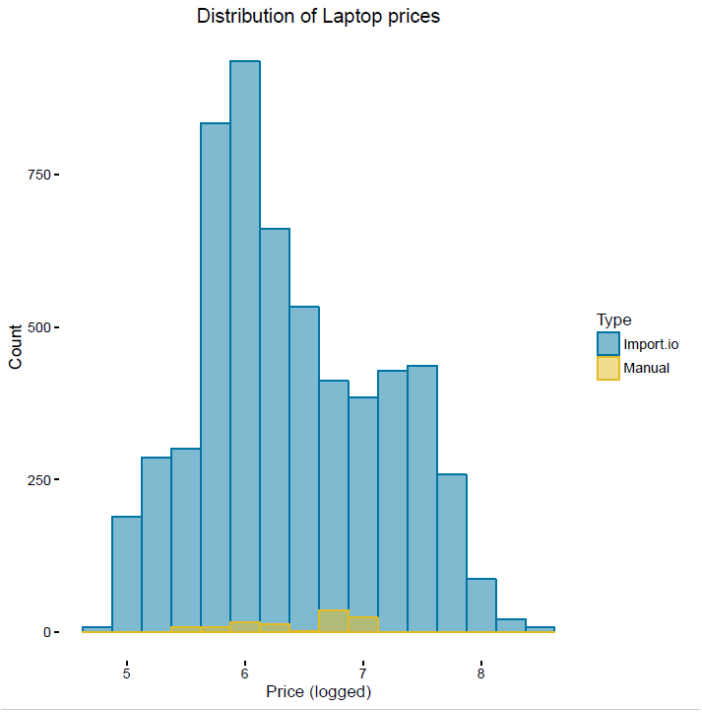

Figure 10: Kernel density distribution for web scraped and manually-collected prices of 'laptops'