ADVISORY PANEL ON CONSUMER PRICES – STAKEHOLDER

Consumer Prices – Alternative Data Sources Roadmap Update January 2021

Status: Final
Expected Publication: Alongside minutes

Purpose

1. The Alternative Data Sources Roadmap presented to APCP-S in January 2020 detailed our timeframe for implementation of new methods and data sources into headline consumer price statistics. This paper provides an update on progress made and our plans over the next 2 years.

Actions

2. Members of the Stakeholder Panel are invited to:
   a) provide feedback on the progress to date
   b) comment on the timelines and scope of the alternative data sources project
   c) comment on the potential flexibility of the parallel run (see paragraph 33)

Background

3. We are currently working through a comprehensive transformation programme for Consumer Prices Index including owner occupiers' housing costs (CPIH) and Consumer Prices Index (CPI)in order to modernise their measurement and make better use of data and methods that are becoming increasingly available to us.
4. At a high level, this involves obtaining robust sources of alternative data, development of statistical systems to work with these data, and methodological research in order to effectively classify, validate and construct high quality price indices from these new data sources. These new data sources will be used in conjunction with traditionally collected data to improve the accuracy, efficacy and representativity of consumer price inflation statistics.
5. The data sources we are investigating are web-scraped data (automated data collection from retailer websites) and scanner data (point-of-sale expenditure and quantity data provided directly by retailers) More information can be found regarding these data sources in our article Introducing alternative data sources into consumer price statistics. These data sources will be implemented for a number of priority categories, including clothing and groceries. These categories were prioritised with stakeholders in 2019.
6. This transformation will be the largest change to consumer price statistics in a generation, and the scale and importance of this work should not be underestimated. We will be reliant on developments in many areas, including the use of new technology platforms and the willingness of retailers to provide us point-of-sale data.
7. In this paper we provide a summary of our progress made over the last 12 months, and an update on our roadmap between now and 2023, when we plan to first incorporate data from alternative sources into our headline measures of consumer price statistics.
8. During this period, we will be liaising regularly with both the APCPs, our users, and the Office for Statistics Regulation, to ensure that our future plans for consumer price inflation measurement are appropriate for improving the quality of our statistics and meeting our ongoing user requirements.

Progress made in 2020

Obtaining robust sources of alternative data

9. We have been engaging with several data suppliers throughout 2020, making significant progress. The table below provides an update on data acquisition for each of our priority categories. We have also included a RAG status that indicates whether the data sources we are looking to acquire will be sufficient for the given category to proceed with implementation in 2023 or may need to be rolled forward into 2024 and beyond. Further details of the progress made can be found in Annex A.

Summary of data acquisition by priority category

Category	Progress summary	RAG
Groceries	Substantial progress in acquiring point-of-sale (scanner) data from 6 retailers, making up a significant majority of the grocery market share.	GREEN
Clothing	Continued use of web scraped data through a new contract with Mobius. Post-2023 development may look to incorporate scanner data (we have acquired data from 2 non-grocery retailers).	GREEN
Tech goods, chart collection items	Focus continues to be on web scraped data covered by the new Mobius contract. Post-2023 development may look to incorporate scanner data (including from the 2 non-grocery retailers).	GREEN
Used cars	Web scraped data (including historic) from AutoTrader in Q1 2021.	GREEN/AMBER
Rail fares	Transaction level rail fares data (including historic) from Rail Delivery Group (RDG) in Q1 2021.	GREEN/AMBER
Package holidays	Web scraped data from Mobius, supplemented with a direct feed from a travel price comparison site. This sector was significantly affected by restrictions in 2020 and we are reviewing whether these data are adequate to continue with research.	AMBER
Air fares	Original aim was to use web scraped data for this category, using two price comparison sites. A review of these data by the end of 2020 revealed limitations of the data and concluded that a new acquisitions strategy is required for airfares data in 2021. This means that we will not be able to incorporate alternative data sources for this item into our headline statistics in 2023 but will continue to work to include them for 2024 and beyond.	RED

10. We have also been scaling up our capability to scrape in-house (see 'Responding to the pandemic' for how we've used these data in the response to the COVID-19 pandemic). This included feeding into the development of ONS's new web scraping policy which supports more flexibility in setting up the scrapers, and transferring the scrapers to a more stable cloud environment. We have also been developing a more standardised framework for scraping that will allow us to more quickly set up and maintain scrapers in future. Our long-term goal is to bring all web scraping in-house.

11. We have also been working with GS1 UK – the provider of industry-standard product identifiers (barcodes) – whose catalogue of product descriptions may help streamline the data collection and classification process. We are on track to acquire their data during the first half of 2021.

Development of statistical systems

12. In 2020, our goal was to integrate the processing of our locally collected data alongside our alternative data sources (ADS) pipeline, which we had built in 2019. The local collection pipeline uses a similar modular approach to the ADS one. The two data sources are aggregated at the end of the pipeline using market share expenditure weights. For our pilot, we used data from one supermarket retailer covering COICOP 01 and 02. Price quotes collected from this retailer were removed from the locally collected data to avoid double counting.
13. The pipeline has also been developed further to include additional functionality in order to facilitate our research programme (see next section). This included adding capability to test the impact of different methods of classification.
14. We were also able to redeploy a version of this pipeline at pace to calculate the weekly online price changes for food & drink items, in response to the COVID-19 pandemic. This demonstrated the benefits of having the pipeline coded in a modular way, we have been able to adapt specific modules quickly to allow for weekly indices to be produced, but keep other relevant areas of the pipeline the same as had already been developed.
15. In 2021, ONS is moving towards use of new cloud-based technology to support the regular production of national statistics, which in the long-run will be more robust and future-proof than current systems. Note that this is different to the platform expected to be used when the roadmap was originally drawn up[1]. We expect that work which is completed can be transferred to the new system easily (Also see paragraph 33).

Methods research

16. In 2020 we have significantly progressed our understanding and research into the use of these alternative data sources. In summary, we have:
17. Published our first series of articles with details of our progress on some of the methods we are investigating for use with scanner and web-scraped data, which we plan to update bi-annually.
18. Used our index number methods quality framework (developed in 2019) to shortlist the preferred methods for use with scanner and web-scraped data and stress tested these methods against synthetic datasets to understand if there were any data properties that might cause the preferred methods to diverge. This has resulted in a handful of methods that we feel are suitable for these data sources, this work is due to be peer reviewed in 2021. As such, the ONS are seeking an independent and external review of the framework and scoring of each method, facilitated by Economic Statistics Centre of Excellence (ESCoE).

[1] See the Consumer Prices Alternative Data Sources Roadmap, presented to the panel in 2019.

19. Made substantial progress in automated classification methodologies and testing the impact on resulting indices. For our most complex classification task, clothing, we have coded all stages of a machine learning-based approach to classify to well over 100 consumption segments, with high performance levels. The latest publication on our machine learning classification models shows high performance with an F1 of 0.77 (between 0 – 1, where 1 is perfect precision and recall)[2]. For grocery scanner classification, where a full machine learning approach is likely to be less suitable, we have begun implementation of a classification strategy that was proposed to the APCP-T in January 2021.
20. Taken significant strides in being able to produce indices from scanner data including: the identification of unique products in the data using appropriate product codes; the standardisation of unit sizes so we can automatically adjust when product sizes change; the identification of product relaunches (cases where a product undergoes a minor change that does not affect the overall quality, causing its associated product codes to change); and understanding the prevalence and impact of refunds.
21. Acquired an overlap of scanner and web-scraped data with the same time coverage to allow us to test methods and assumptions for approximating expenditure shares for web-scraped data when no quantity/expenditure data are available.
22. Worked with National Institute of Economic and Social Research and Economic Statistics Centre of Excellence as well as internally to test methods of outlier detection with scanner and web-scraped data, with a report due to be published in Q1 2021.
23. Identified a source of weights that can be used to weight retailers by market share in each category instead of the traditional multiple/independent split. This will allow us to aggregate scanner and web-scraped data with traditional data sources.
24. The data acquisition, systems development and research work completed in 2020 has allowed us to produce a multitude of provisional, experimental indices including:
    - Research indices using externally contracted web-scraped clothing and laptops data
    - Retailer-specific scanner data indices based on the existing retailer hierarchy's
    - Retailer scanner data indices loosely mapped to the existing CPIH hierarchy so they can be aggregated with locally collected data

    It is not yet possible to share or publish these due to confidentiality constraints and commercial sensitivity (particularly for retailer scanner data) and further methodological improvements required (e.g. classification, product grouping, quality adjustment).

Responding to the pandemic

25. Our progress over the past year has also allowed us to rapidly respond to the coronavirus pandemic, providing the government and other users with timely data showing the impact of coronavirus on price and expenditure changes. This includes weekly production of our online weekly price changes section of the Coronavirus and the latest indicators for the UK economy and society statistical bulletin, as well as providing senior government officials with details of expenditure change and the extent of stockpiling throughout the pandemic. Without these data and technologies, the government's ability to understand these impacts would have been more limited.

[2] For more information on the performance of machine learning classification models please refer to our article on 'Automated classification of web-scraped clothing data in consumer price statistics'

26. Although this has accelerated our understanding of the use of scanner and web-scraped areas in some ways, the redirection of resource has slowed the pace of systems development and some of the research projects which will be continued into 2021, although we continue to ramp up the application phase in parallel.

Roadmap to 2023 and beyond

27. We have split our delivery programme until 2023 into three phases, each lasting a year. In 2021 we will continue to discuss ongoing research and developments with both the Technical and Stakeholder Panels at regular intervals and will publish bi-annual research papers to update users on our progress. In 2022 we enter an engagement phase where we will aim to share aggregate quarterly experimental estimates incorporating alternative data sources.
28. The first phase (research) ran until the end of 2020 and involved:
    - continued engagement with retailers to secure more regular data feeds as well as historic data
    - integration of traditionally collected data into the processing pipeline and understanding the impact of processing scanner data, web-scraped data and traditional data simultaneously
    - research into improvements that can be made to processing of traditional data including developments of new systems (note: some work is still ongoing in 2021 due to the impact of the COVID-19 pandemic)
    - further developments of the processing pipeline for web-scraped and scanner data to enable research and impact analysis
    - research into the methods needed to produce high quality indices using web-scraped and scanner data (a full research programme is outlined in Annex C)
    - initiation of a new and more focussed web-scraping contract with Mobius
    - a review into policy and system changes needed to fulfil the longer-term strategy to bring web-scraping in-house
    - user engagement, including through bi-annual publications

29. The second phase (application) will run throughout 2021 and involves:
    - continued engagement with retailers to expand the amount of scanner data available to us and ensure continuation of regular data feeds
    - a parallel run of using web-scraped data in a production environment (note: delayed from 2020)
    - application of research to specific item categories within the inflation basket as prioritised with APCP-S in September 2019
    - completion of approved methods built into the processing pipeline, while moving towards cloud-based technology at organisation level (see also paragraph 33)
    - initial impact assessments carried out on aggregate measures
    - scaling-up of in-house web-scraping capabilities
    - a summary of research and final recommendations on methods for different priority item categories made
    - planning priority items for beyond 2023
    - continued user engagement, including through bi-annual publications

30. The third phase (engagement) will run through 2022 and involves:
    - quarterly publication of aggregate experimental indices including web-scraped and scanner data in conjunction with traditionally collected data
    - user engagement to discuss methods and changes
    - research and developments for priority items for beyond 2023

31. A visualisation of our high-level roadmap and detail on what we expect to deliver in 2021 can be found in **Annex B**.

32. Beyond 2023 the rollout of alternative data sources will continue. We plan to a) increase the use of alternative data sources in existing priority categories (through adding more retailers), b) continue a programme of research on methods, in line with ongoing work both in the ONS and internationally, and c) rollout new data sources to new item categories. The new item categories we are considering include:
    - Gas & electricity
    - Clothing accessories/footwear
    - Household goods
    - Mobile phone charges
    - Pharmaceuticals
    - Furniture

    We will prioritise these alongside any delayed 2023 categories in a separate panel paper to be discussed with the stakeholder panel before end of 2021.

Risks and mitigations

33. As the programme enters its penultimate year of implementation, we will continue to closely monitor progress. The project team have identified three main risks: systems change at ONS level, funding and resources, and Covid19. We have carefully considered mitigating actions.
34. **Systems change at ONS level:** ONS is moving towards use of new cloud-based technology to support the regular production of national statistics, which in the long-run will be more robust and future-proof than current systems[3]. This is currently in development and is a key dependency for this project. Alternative data sources cannot be fully progressed on the existing platform, which lacks the required functionality. Development of the final processing pipeline on the new platform must be complete by December 2021 to allow for the parallel run of alternative data sources throughout 2022.

[3] This is different to the platform expected to be used when the roadmap was originally drawn up. See the Consumer Prices Alternative Data Sources Roadmap, presented to the panel in 2019.

- **Mitigation (a)**: We have led a series of workshops with the responsible IT teams to clarify prices requirements of the new platform and timelines for implementation. A team to lead on the prices requirements is expected to be approved within the next month.
- **Mitigation (b)**: We have assumed that the parallel run of CPI/CPIH using alternative data sources must run for a full 12 months in 2022. Another potential mitigation is to be flexible with the duration of the parallel run. Stakeholders are invited to comment on the flexibility we can show with the parallel run, and whether the final systems, methods and data must be used for the full 12-month period.

35. **Funding and resourcing** for price statistics development: The 2020 Spending Review (SR20), recognising the effects of the pandemic, prioritised funding across government, resulting in certain changes to expected departmental spending across government. SR20 set departmental budgets for one year (2021/22), and we expect the process will be repeated in 2021 to secure funding for future years.
    - **Mitigation**: Prices is recognised as a key priority across Economic Statistics Group. In the recent spending round the team secured an additional settlement for financial year 2021/22, reflecting the increasing volume of resources required to deliver the roadmap for the upcoming year. The team has additionally secured a share of HMT's Economic Data Innovation Fund. These additional funding sources have allowed us to secure further resource dedicated to the ADS project.
36. **Covid-19 pandemic**: During 2020, additional workstreams using alternative data sources were identified and prioritised in order to respond to the pandemic, while simultaneously price statistics production became more challenging. This meant that some resources were temporarily diverted away from the ADS project, resulting in some delays to the research programme (see paragraph 24).
    - **Mitigation**: As above, the project team has secured additional funding to deliver the roadmap for 2021. However, we will continue to deliver online weekly price changes and remain responsive to user needs.

List of Annexes

Annex A	Data acquisition for priority categories
Annex B	2020:2023 roadmap, including detail for 2021
Annex C	Research programme for the alternative data sources project

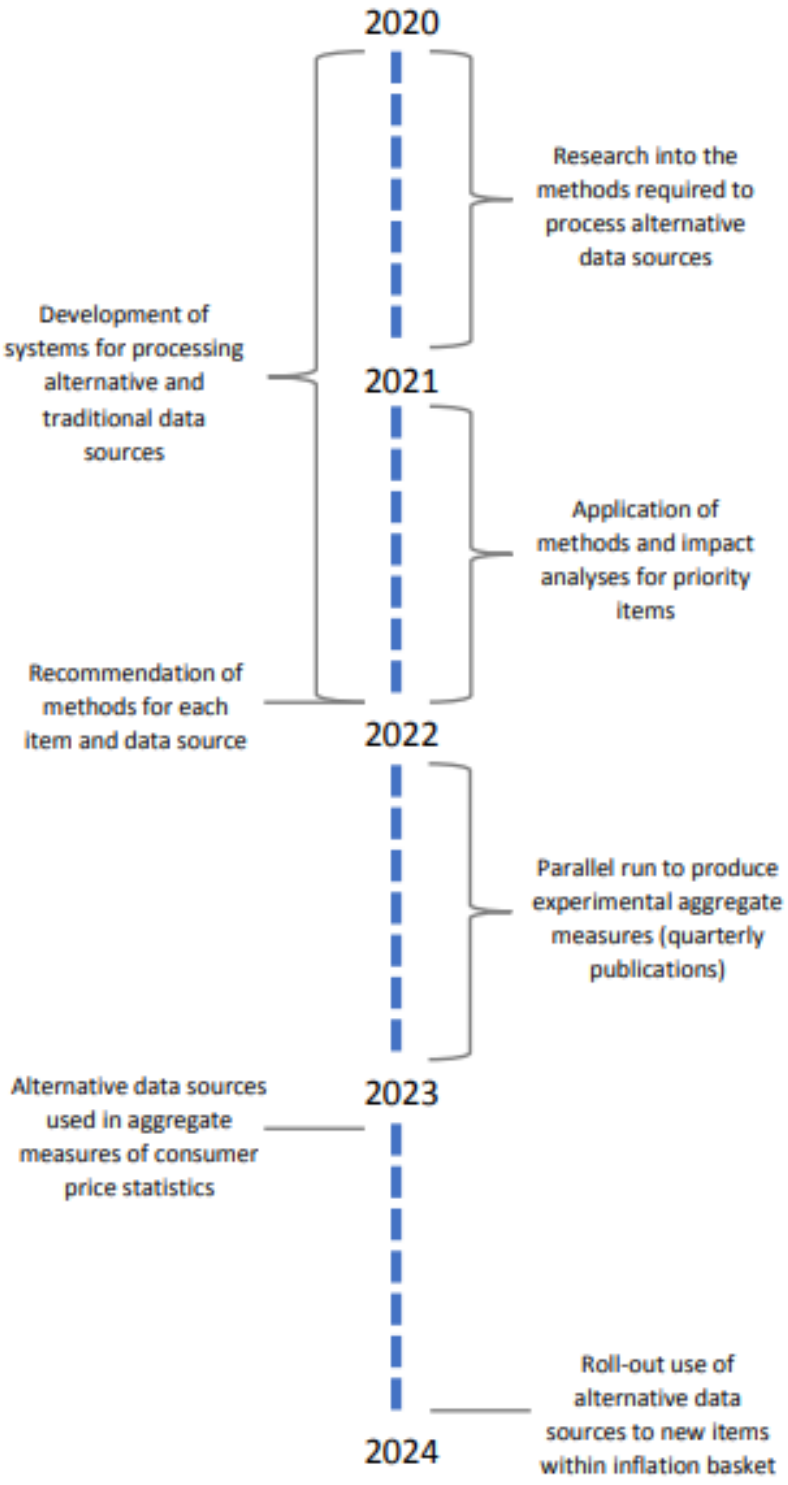
Annex A: Data acquisition for priority categories

a. **Groceries** (COICOP divisions 01 and 02): GREEN. We have made substantial progress in accessing point-of-sale data in 2020. We now have regular data feeds from three of the UKs biggest grocery retailers alongside historical data from one retailer, with work ongoing to acquire historic data from the other two retailers. We have also received test data from a further two grocery retailers and are close to finalising the data specifications with an additional retailer. Together, these six retailers account for a significant majority of the grocery market share. Initial research on using these data was presented to APCP-T on the 15th January, although the paper is restricted due to the disclosive nature of the data presented, as requested by the retailers.
b. **Clothing**: GREEN. Our focus for clothing data has been on the use of web scraped data. The contract with mySupermarket was terminated in May 2020, and a new contract with a new supplier, Mobius, began in June. There are no historical series available with these data so we will need to build up a sufficient time series of high-quality data before a final impact assessment can be completed. As well as the 6 grocery retailers, we now also have regular scanner data feeds from an additional two large UK retailers which include clothing within their offering. For one of these retailers we also have historic data going back to September 2017, and work is ongoing to acquire historic data from the other retailer. A number of the grocery retailers mentioned above also include clothing. While our focus for 2023 will continue to be on using the web scraped data for clothing, we will look to assess the feasibility of using these scanner data in a later iteration of development.
c. **Tech goods, chart collection items**: GREEN. As with clothing, our focus for these items will be on using web scraped data covered by the new Mobius web scraping contract although future development may also look to incorporate some of the scanner data listed above.
d. **Used cars**: GREEN/AMBER. Data for this category is being sourced from Auto Trader. They are the UK's largest online marketplace that specialises in used car sales, including cars sold by private sellers and trade dealers. We are aiming to start receiving these data (including historic) in Q1 2021, which will allow us time to review these data and apply our chosen methods in time for the parallel run in 2022 but this is dependent on the data that we receive.
e. **Rail fares**: GREEN/AMBER. Data for this category are being sourced from the Rail Delivery Group (RDG) which is an organisation that works with all railway companies. The RDG hold a comprehensive dataset of all train journeys taking place in the UK, in essence, transaction data for rail fares. We are aiming to start receiving these data (including historic) in Q1 2021, which will allow us time to review these data and apply our chosen methods in time for the parallel run in 2022 but this is dependent on the data that we receive.
f. **Package holidays**: AMBER. Our focus for package holidays is on using web scraped data from Mobius, but we will also be supplementing this with a direct feed from a travel price comparison site. Due to the turbulence associated with this particular market sector in 2020, we are currently undergoing a review to determine whether these current data sources will be enough to go ahead with methods research and then implementation, but it may be further data is required in which case the 2023 implementation date may not be feasible for this category.
g. **Air fares**: RED. We originally targeted using web scraped data for this category, using two price comparison sites to limit the number of websites that were required to be scraped (there are many individual airline companies). Due to the number of dimensions involved for airfares, we used the same parameters as the current collection (for example, a return flight for a European holiday should be booked 2 weeks after arrival). A review of the data we had collected so far was completed towards the end of 2020 and the following conclusions were drawn:

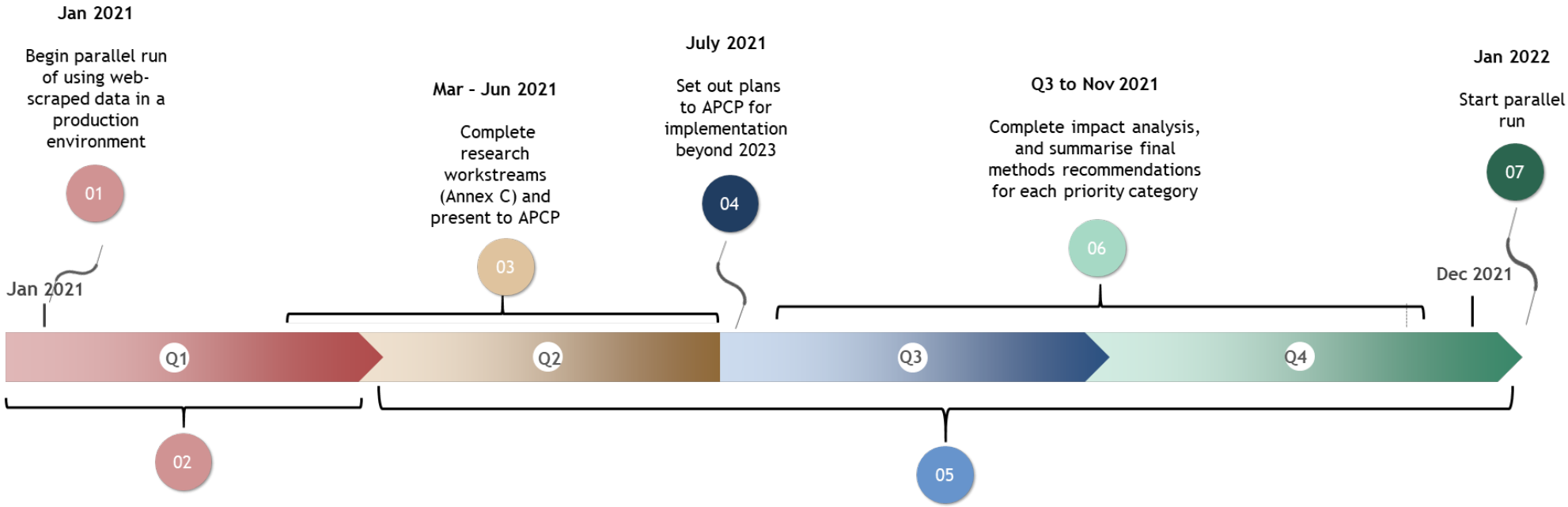
i. Web scraped data allowed us to collect a complete time series (instead of once a month, we were able to scrape daily), and a wider selection of routes/ airline operators than in the current collection. However, we were still limited for some of the dimensions (for example, changing the length of stay) as to scrape every potential parameter value would place significant burden on the website, and risked us being blocked from scraping.
ii. One price comparison web site changed its terms & conditions (T&Cs) during our collection, which meant we were relying on only one other site. We did review whether we could use the individual airline websites, but the majority did not allow scraping in their T&Cs. We concluded that this was too risky for our price collection to be dependent on only one site.
iii. The prices collected from the price comparison sites were different to those collected manually from the individual airline website. Consumers use both price comparison sites and direct websites in their purchase of airfares, so we need to ensure we are representative of both sectors in our data.

These conclusions led us to determine that we would need to review our acquisition strategy for airfares data in 2021, which therefore means we will not be able to incorporate this item into our headline statistics in 2023.

Annex B: 2020:2023 roadmap



Detailed roadmap 2021



1. Receive remaining data feeds for priority categories (RDG and Autotrader)
2. Complete review into package holidays to determine if further data acquisition work is required
3. Continue engagement with retailers to receive regular feeds and historic data from grocery retailers, and the non-grocery retailers

Ongoing - scaling up of in-house web scraping capability
Published updates are provisionally scheduled for March, June and September, pending approval

Annex C: Research programme for the alternative data sources project

1. Research phase (2020, some continued into 2021)

Research into the methods needed to produce high quality indices using web-scraped and scanner data:

- **Classification** techniques for alternative data sources (including machine learning methods): This will enable us to automatically classify large quantities of data into price indices
- **Index number methods framework**: This will allow us to ensure we are choosing the most appropriate index number method dependent on the pricing behaviour of a particular item and characteristics of the dataset
- **Product grouping**: In areas with high product churn, product grouping methods will allow us to follow groups of products over time, rather than individual products
- **Scanner data research**: We need to better understand the scanner data before using it, e.g. how do we account for returns and discounts? Can we apply take-up rates of multibuy discounts to traditionally collected data? How do we identify product relaunches to ensure appropriate quality adjustment takes place?
- **Expenditure proxies**: This research will look at using proxies to weight web-scraped data to ensure that more popular products are given a higher weight in the index
- **Retailer weights**: We need to find data sources that will allow us to weight retailers together with traditionally collected data at the lowest levels of aggregation
- **Outlier detection and imputation**: This will enable us to appropriately identify and remove/validate outliers and identify appropriate imputation methods to handle missing data – whether the data is missing temporarily, permanently, or on a seasonal basis.

2. Application phase (2021)

Applying methods and techniques developed during phase 1 of the research programme to specific item categories. These item categories were agreed and prioritised with our stakeholders in September 2019.

- **Groceries**: With a focus on scanner data as we can cover a large market share with a small number of retailers.
- **Clothing**: With an initial focus on web-scraped data as market widely distributed across retailers. Clothing items have high product churn due to changing fashions and have led to serious problems with price indices in the past (e.g. formula effect).
- **Tech goods**: With an initial focus on web-scraped data as market widely distributed. Tech goods also have a high product churn and current hedonic methods are highly resource intensive.
- **Used cars**: Measurement in used car prices is challenging as need to adjust for age/mileage. Market engagement ongoing to procure a suitable data source that gives sufficient coverage of attributes to allow for quality adjustments.
- **Package holidays**: Current method of producing package holiday indices doesn't align with Eurostat methodology and is incoherent with calculations across the rest of the inflation basket. Initial focus on web-scraped data to provide a larger quantity and higher frequency of data collected.
- **Air fares**: Web-scraped data will provide more frequent collection over a significantly larger sample of routes.
- **Rail fares**: Current methodology uses an imputation based on the cap of regulated rail fares set by the chancellor. Scanner-type data will allow us to use actual transaction data to calculate the index
- **Chart collected items**: DVDs, CDs, Books, Computer games use a methodology that follows chart positions over time rather than individual products. As such the index can be volatile as products change position within the charts.