APCP-T(22)02

ADVISORY PANEL ON CONSUMER PRICES – TECHNICAL

Transformation of UK consumer price indices: rail fares

Status: Work in progress
Expected publication: For publication alongside minutes

Purpose

1. In 2021 ONS obtained transaction level data for rail fares in Great Britain sourced from the rail industry's Latest Earnings Nationally Networked Over Night (LENNON) ticketing and revenue system. In this paper we look at how we can use these new data to produce detailed, informative and accurate statistics regarding price changes for rail fares, and the expected impact of including new price indices for rail fares in UK consumer price statistics

Actions

2. Members of the Panel are invited to:
   a) consider the suitability of the proposed methods for calculating a GB rail fares index for use in UK consumer price statistics
   b) consider the appropriateness of suggested methods to assign regions to rail tickets in order to produce regional indices for rail travel, or whether a national index is more suitable conceptually
   c) discuss whether it is appropriate for compositional effects (such as rail cards and child fares) to affect UK consumer price statistics

Background

3. Rail fares in the UK are complex, with around 40 per cent of rail fares being 'regulated'. Regulated fares are standard class fares including saver returns, standard returns, off-peak fares between major cities, and season tickets for most journeys. Unregulated fares include first class, advance purchase, and saver tickets. Train operators are free to determine these latter fares, although they can be capped in certain circumstances.

4. Price changes for regulated fares in Great Britain (GB) are all capped by the government based on the annual change in the RPI in July of each year. This annual uplift to fares (reported by the Rail Delivery Group each year) is what is currently used to calculate the consumer price index for GB rail fares, which is then aggregated with a similar annual figure provided by contacts in the Northern Ireland Transport Holding Company, Translink. The weights for this aggregation of GB and NI are based on the total franchised passenger revenue published by the Office of Rail and Road (ORR) vs the total passenger receipts as published by the Department of Infrastructure in Northern Ireland.

5. While the current method is simple to implement and aligns with information already in the public domain regarding price changes for rail fares, new data will improve our coverage of rail fares, allowing us to better understand price changes for unregulated fares, seasonal fluctuations in price, and geographical variations.

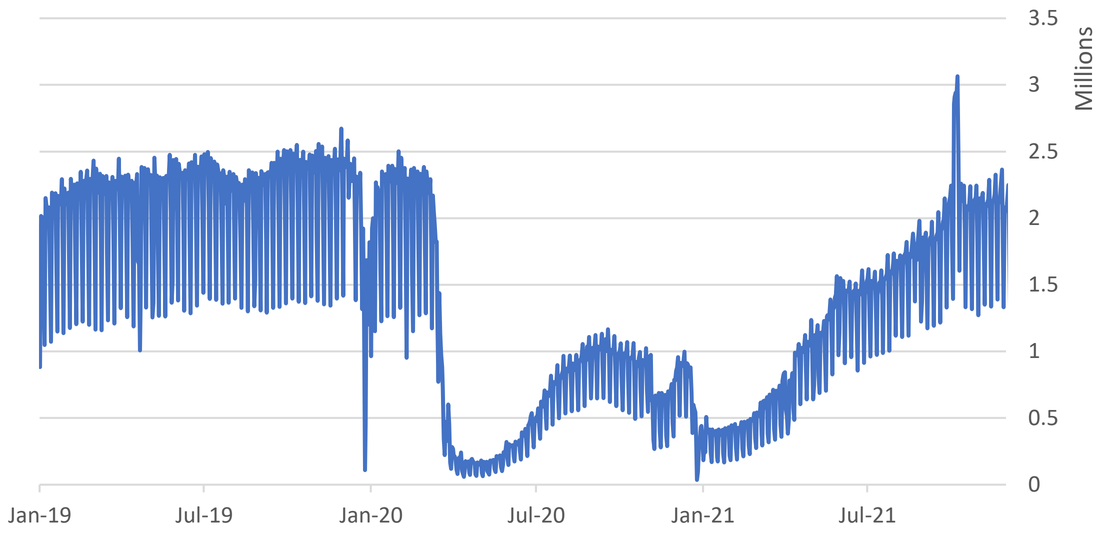
6. In 2021, rail fares had a weight of 3.55 parts per thousand (0.36%) in CPIH, and 4.97 parts per thousand in CPI (0.5%).

New data

7. In 2021, ONS obtained access to transaction level data for rail fares in GB sourced from the rail industry's LENNON ticketing and revenue system, dating back to January 2019. As these are transaction level data, explicit information is available on quantities of each product purchased and the data are comparable to those that we refer to as scanner data. These data are expected to cover a near-census of transactions for rail fares in GB.

8. These transaction level data are delivered daily. Although some are delivered at a lag, we have found that 85% of data are delivered within 1 day of purchase and 97% of data are delivered within 1 week, making these data extremely timely. We receive approximately 2 million transactions per day, equating to approximately 60 million per month (in non-COVID times), though the number of daily transactions shows seasonality (Figure 1).

Figure 1: Daily transactions by issuing date for rail fares (millions)



9. The data are highly informative, including variables such as: sale and processing dates, sales values and quantities, origin and destination stations including product names and route types, fare product groups (e.g. advance, peak, off-peak), journey factor (which tells how many journeys on a single ticket – i.e. a return ticket would be a journey factor of 2), ticket class (e.g. first or standard), station postcode, number of adults and children. A data dictionary describing the full set of variables is provided in **Annex A.**

10. The data include additional transactions for car parking, rail card purchases, business fares and other non-rail fare related expenses. After filtering and cleaning the data, we can use approximately 40 million transactions per month in calculating consumer price indices. Further details regarding the data cleaning carried out prior to producing the analysis in this paper are provided in **Annex B**. Additional, more advanced, outlier detection techniques are being considered for use on our range of alternative data sources, but these methods will be brought to the Panel for consideration at a later date.

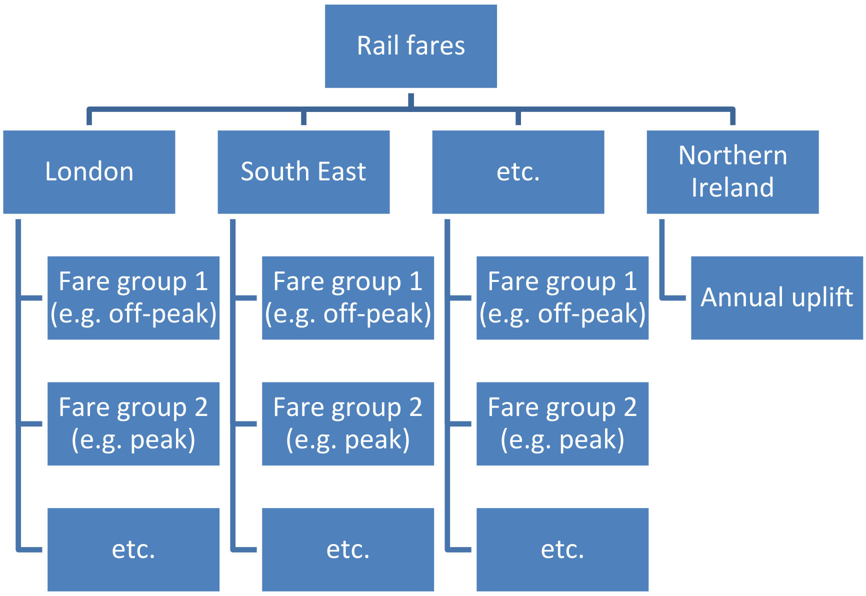
11. The data are inclusive of underground and metro fares that are in a separate subclass in the COICOP hierarchy. We have ongoing work to identify and remove these stations from the data

to avoid double counting price movements for these fares, but this is a complex task as there is no easy way to differentiate underground and national rail travel from stations who provide both services. They remain in the data for the current analysis.

Index methods

12. Our previous work and corresponding international guidance has pointed towards multilateral methods being most appropriate for producing price indices using large, dynamic datasets. Our work in choosing the most appropriate index number method is ongoing; for the purpose of this analysis, we focus on a GEKS-Törnqvist index because of its reduced run-time comparatively to other methods.
13. We stratify our indices to a fare product group (e.g. advance, peak, off-peak) within each region (Figure 2). We have also considered stratifying indices by ticket class (standard or first class) to provide an additional layer of information when interpreting the indices but have found the coverage of these fares to be low.

Figure 2: Future hierarchy for UK rail fares index



14. A GEKS-Törnqvist using a mean splice on the published series with a 25-month window is used for calculation of these low-level stratum indices. Consistent with our traditional practices of CPI construction, above the stratum level we use a Lowe formula to aggregate to higher levels.
15. Stratum level indices for GB rail fares in this analysis are aggregated based on the previous year (y-1) expenditure shares. For the first year of the analysis, where no historic data are available, they are based on the first year (y). Regional indices are aggregated with the existing index for Northern Ireland, using existing weights, to produce a UK rail fares index.

Defining a unique product for rail fares

16. An initial challenge with using these data for the calculation of rail fares indices is in determining what variables should be used to define unique products that we can follow the price change of over time. There are a number of variables that can be used; using all variables would result in a lower product match rate over time and using a limited number of variables would allow compositional effects (e.g. inclusion of more saver returns) to impact the resulting indices.

17. For the analysis presented in this paper we have used the following variables to define a unique product (for more details of these variables refer to the data dictionary in **Annex A**):

    - origin region (implicit through stratification, e.g. Wales)
    - fare product group (implicit through stratification, e.g. Anytime / Peak)
    - origin station (e.g. Cardiff Central)
    - destination station (e.g. London Paddington)
    - product name (e.g. STANDARD DY RTN 2BAF)
    - route (e.g. via London)
    - ticket class (e.g. Standard)

Regional indices for rail fares

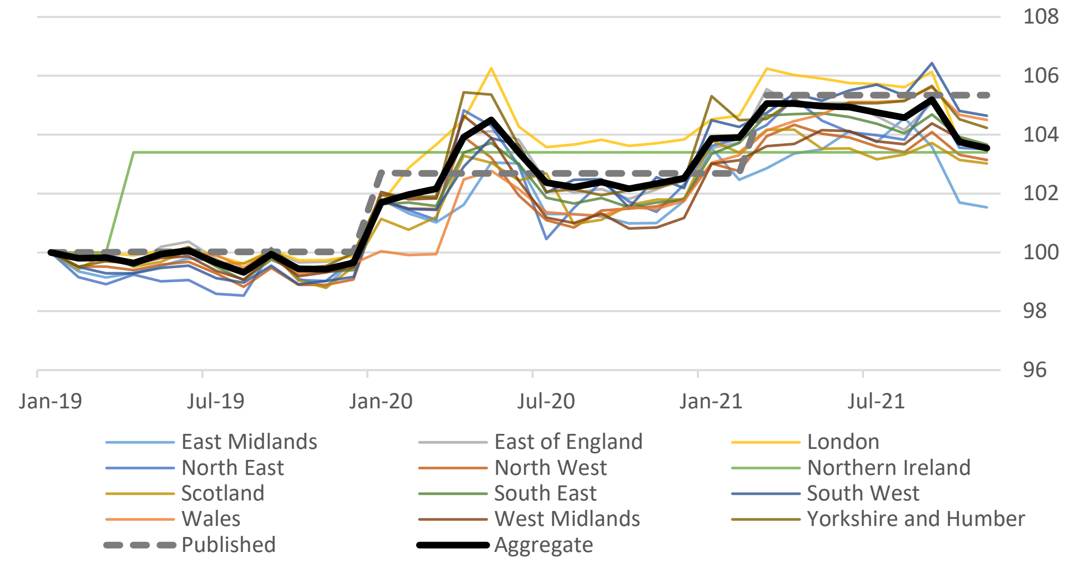
18. One of the objectives of using new data sources in the UK is that they provide better geographical variation that in future will allow us to produce price indices on a regional basis more readily. However, there is something of a conceptual challenge in defining regions for rail fares. The region could be based on the origin station, destination station or a single national index could be produced.

19. The question depends on the concept for which one might want to measure regional price indices. If regional indices are intended to reflect price changes faced by households who live within a said region, then the origin station would be preferred. Even then, sometimes travellers purchase separate tickets for both the outward and return journey (such as two advance single tickets, or buying two separate singles on the day), or they may purchase multiple tickets covering a single journey (known as ticket splitting). We do not know how common these practices are.

20. Another challenge is in allocating stations to a region, as region is not a variable contained within these data. Of the 4347 unique stations in the data, there is a postcode provided for 2700 stations - which we can then use to map the station to a region. These 2700 stations make up approximately 70% of the expenditure, but by manually assigning a region to 50 of the unmapped stations we have improved this coverage to 97% expenditure.

21. Figure 3 shows regional indices for rail fares in Great Britain, as determined based on the origin station, compared to the aggregate index and the existing (published) index. While there is some regional variation in rail fares, broadly speaking the regional indices follow a similar trend. The aggregate index does experience an annual uplift broadly in line with our published index, though is somewhat more staggered. There is also more temporal variation in the index produced using new data and methods, which is further explored in the following sections.

Figure 3: Regional indices for rail fares compared to the new aggregate rail fares index and the currently published rail fares index, Jan 2019 = 100

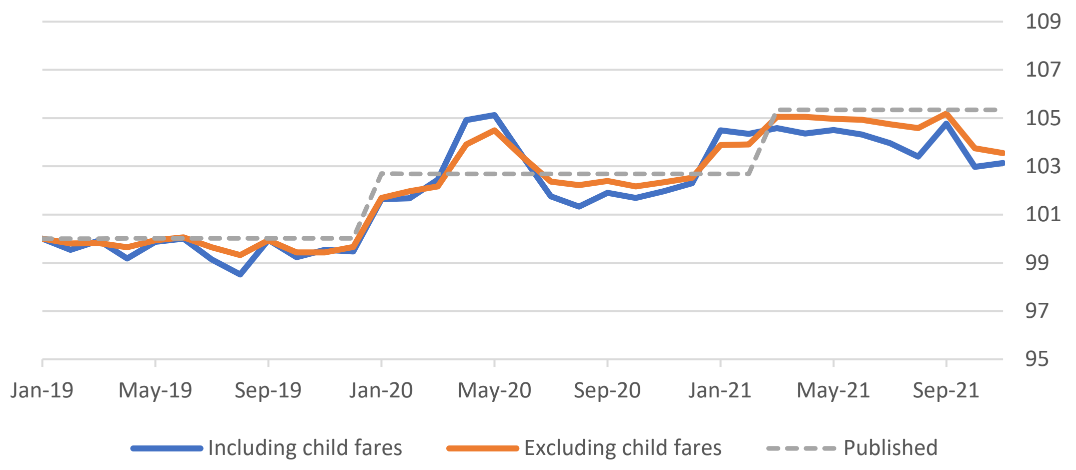


Accounting for children fares

22. As the data are transaction level data, there can be both adult and child tickets bought within a single transaction, and the price for each cannot be easily differentiated. If we were to not appropriately adjust indices for the inclusion of children, the price movements may be impacted by composition of travellers rather than genuine price changes. For example, if more children travel in August due to the school holidays than in September when they have returned, our price index might decrease in August and increase in September, despite train companies not imparting any genuine price changes.

23. As child fares make up a small proportion of the data, at approximately 2% of the total expenditure, we have removed any transactions that have at least one child fare included. This reduces the volatility of the index (Figure 3), suggesting that the composition of adult and children fares within our transactional data were contributing to price change. For our further analysis we have therefore removed any transactions related to child fares.

Figure 4: Rail fares index including and excluding child fares, Jan 2019 = 100



The rail card conundrum

24. As shown in Figure 4, there is still seasonal variation in the price of rail fares, as well as a notable spike in fare prices during the initial months of the COVID pandemic, even after accounting for child fare effects. While this may reflect genuine price change, it may also partially reflect seasonal (or pandemic-driven) trends in rail card usage, something we are not able to account for in the current data.

25. Rail cards offer groups of the population discount for travelling, such as young, senior, and disabled travellers, and those who travel with friends and family. There are even some rail cards available that apply to specific geographies.

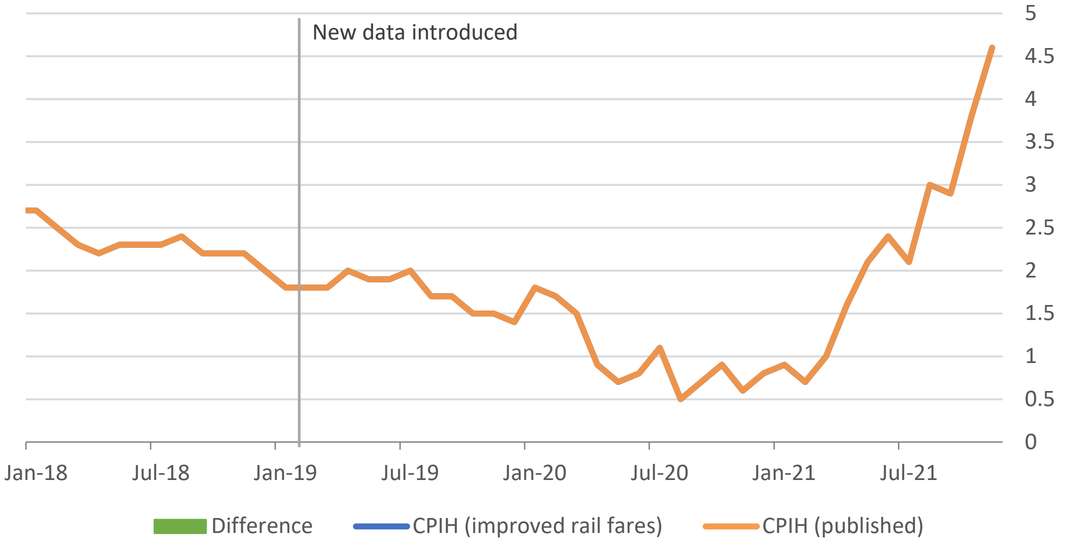
26. The question of whether it is appropriate for these rail card (or other compositional) effects to be influencing our consumer price statistics depends on the underlying purpose of the CPI. As the transactions within the index are weighted, rail card prices are only included in proportion to the take-up rate that the rail card discounts have. This gives us a good assessment of the typical price being paid in each month from the perspective of the average consumer.

27. However, if we are purely interested in price changes for fixed services (assuming no changes in composition), it may be hard to establish whether this is truly the case or whether there is something of a railcard effect. This is not something we are appropriately able to assess using the current data.

Impact of new data and methods for rail fares on headline consumer price statistics

28. The aggregate index for CPIH (Figure 5) has been produced between January 2018 and November 2021, including the new rail fares index from February 2019 onwards so that growth rates in the year of introduction can be seen as well as annual growth in the years following introduction. The new index is aggregated together with published series using the existing annual weights and chain-linking methodology.

Figure 5: Impact of new data and methods for rail fares on CPIH annual growth rate (%)



29. While our index for rail fares was more responsive to the pandemic as well as seasonal effects, the impact on CPIH as a result of this change is negligible (see Figure 5 that CPIH (published) and CPIH (improved rail fares) overlap), with an average absolute difference of 0.0053 percentage points across the entire period. The maximum difference is 0.02 percentage points lower for improved rail fares in May 2021, and this is also the maximum impact on CPI.

30. Note that since March 2020 there have been a number of unavailable items that have been imputed in some periods based on price movements of the headline index. For this impact analysis we haven't recalculated these imputations due to the complexity of their calculations, but we would expect the impact of recalculating imputations to be negligible based on the minimal impact of these new data and methods on the headline indices.

31. Impacts of the introduction of these data are discussed further in **APCP-T(22)03 Transformation of UK consumer price indices: Impact Analysis, 2022**.

Future work

32. There is still some remaining work to refine our indices for this category as we ready for our first publication of experimental statistics using these data in May 2022. We do not expect any of the remaining work to significantly change the impacts presented in this paper.

33. In particular, future work will consider:

    a. considering whether product "relaunches" may be a problem for rail fares, for example if a product code changes or a station changes name

    b. trying to find the best way to separate fares for underground travel, to avoid any potential double counting in consumer price indices (a price index for underground travel is calculated in another COICOP5 heading)

    c. investigating whether using less than a full month of data (to ensure the timeliness of our price indices) would have a substantial impact on the resulting price index.

Joe Barker & Helen Sands
Prices Division, Office for National Statistics
January 2022

List of Annexes

Annex A	Data dictionary
Annex B	Data cleaning and filtering

Annex A – Data dictionary

Field	Description	Our use	Example
processing_date	The date the Lennon system processed the settlement. This may be different to the date the ticket was sold or issued e.g if there is a delay in receiving data from the retailer.		2019-05-28 00:00:00
sale_date	The date the ticket was sold.		2019-05-18 00:00:00
issuing_datetime	Date and time when the ticket was issued. With Ticket on Purchase (ToP), ticket issue and sale are simultaneous, so the Date of Issue is the same as Date of Sale. With Ticket on Departure (ToD), the sale and issue occur at separate locations and/or on separate days, so the Date of Issue is usually after the Date of Sale, reflecting the date when the ticket was collected.	Collection date	2019-05-27 11:10:00
origin_code	4-character National Location Code.		1947
origin_desc	Origin description. This is the fare location which might not be a physical location e.g it could be a logical fare group such as “LONDON BR” (London Terminals)	Used in defining a unique product and joining postcode / region information	LEICESTER
destination_code	Similar to origin		1072
destination_desc	Similar to origin	Used in defining a unique product	LONDON BR
route_code	Indicates what restrictions, if any, apply on a journey from A to B. For example, a route can denote that travel is restricted by Train Operator, or via a specific station, or is valid on any permitted route.		01000
route_desc	Route description	Used in defining a unique product	ANY PERMITTED
product_code	Lennon Type of Ticket. This is a 4-character code. The first character represents the class (1 for First, 2 for Standard and 9 for other) and the other characters are the product code.		1BAF
product_desc	The product on the transaction, in the case of a travel ticket this indicates the type of ticket the passenger has bought, for example 1AAA, FIRST SINGLE	Used in defining a unique product	FIRST DY RTN 1BAF
pro_fpg_description	Fares Product Group: Advance Anytime / Peak Off-Peak Other tickets Seasons Super Off-Peak	Filtering out non rail tickets and stratification	Anytime / Peak
product_ticket_class	Standard, First or No Class		First Class

channel_type	Channel through which the sale took place (Station booking office, TVM, Web TIS,…)		001 – Station Booking Office
method_of_fulfilment	001 – Smartcard Direct 002 – Smartcard Indirect 004 – Self Print 005 – Oyster Top Up 006 – M-ticket 007 – e-Ticket 008 – Paper Roll Ticket 009 – ISRN Unknown 010 – Plastic Railcard 011 – Digital Railcard No Value Unrecorded – this usually relates to magstripe (orange paper tickets) or TfL bulk inputs Null		007 – e-Ticket
cross_london_desc	Indicates whether the ticket permits cross-London travel. This is shown on tickets with the Maltese cross		Not Via London
passenger_journeys	The number of passenger journeys represented by the transaction. Calculated by multiplying the number of people by the journey factor, or for season tickets of 1 month +, the Season Ticket Journey Weightings.		2
number_of_tickets	The number of tickets issued e.g. if two single adult tickets were purchased this value would be 2. If two return adult tickets were purchased this value would still be 2.	Quantity	1
jof_journey_factor	The number of journeys reported for each issue of the product. For example, a product for a return ticket will have a Journey Factor of 2, a product for a single will have 1. Journeys for season ticket products use Season Ticket Journey Weightings not Journey Factors. Journey Factors for season tickets are reported as 999.99 by default.		2
selling_retail_channel	This field is based on a mapping table which is manually maintained by rail delivery group using information provided by third party retailers and TOCs.	Filtering out non consumer tickets	Ticket Office
selling_high_level_mapping	This field is based on a mapping table which is manually maintained by RDG using information provided by third party retailers and TOCs.	Filtering out non consumer tickets	B2C
number_of_adults	The number of adults that the ticket is valid for		1
number_of_children	The number of children that the ticket is valid for	Filtering out child fares	0
ticket_miles	The number of miles represented by the transaction. Calculated by multiplying the		75.45

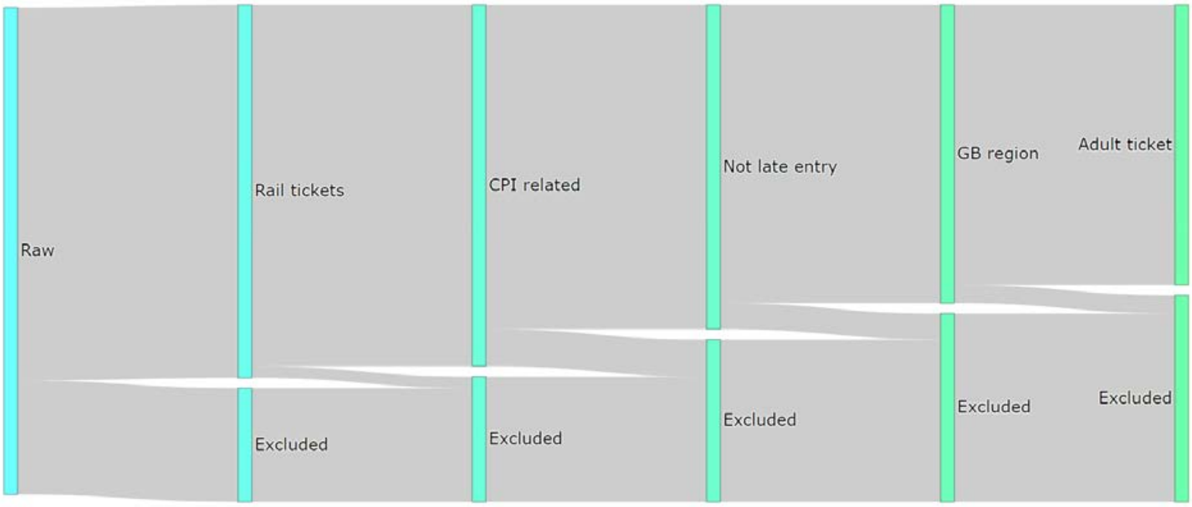
	journeys by the Ticket Miles for the flow. Also known as passenger miles.		
gross_receipt_sterling	Face value of the ticket/transaction; it is inclusive of VAT and any private settlement element.	Sales value of the ticket (including discounts)	95.5

Annex B – Data cleaning

1. The data include additional transactions that are not relevant to the typical consumer purchasing a rail ticket. This includes car parking tickets, business prices and transactions that we cannot assign to a region. In order to produce meaningful indices for rail fares, we need to remove these data.
2. The data include non-rail tickets, such as car parking tickets, ferry tickets and seat reservations. We can filter out these tickets based on the fare product group of a transaction. This field usually takes values such as Peak / Off Peak but we exclude the values where the fare product group is either “N/A” or “Other Tickets”. The “N/A” group contains tickets such as car parking and seat reservations, whereas “Other Tickets” contains more obscure products that we might not want to track, such as train+bus tickets or Liverpool underground tickets. This removes 23% of the data, which accounts for 2% of expenditure.
3. Since we are looking to produce consumer price indices with these data, we need to make sure that we are only capturing price changes that are available to the consumer. We can exclude transactions that are business to business (B2B) and business travel services (BTS) as well as other corporate tickets based on the values in the selling_high_level_mapping and selling_retail_channel columns. This is in line with suggestions from the data supplier. As a result, we exclude 3% of remaining data, which accounts for 10% of expenditure.
4. The majority of data are very timely, as highlighted in Figure X, but in the case that the lag on receipt may cause the processing date of a transaction to be in a month following the collection date, it is necessary to remove these transactions to ensure that previous month's indices are not revised once they have been published. Excluding these transactions accounts for a further 10% of the remaining data and 6% of expenditure.
5. The dataset does not include Northern Ireland train operating companies. Despite this, there are still a few Northern Ireland stations that are present when one of the origin / destination stations is in GB; these are primarily ferry ports where the transaction relates to a ferry + train ticket. Along with these NI stations, we also exclude the transactions that do not have a region assigned to them, as outlined in Paragraph 21 so that we are able to produce regional indices that can be further aggregated to a UK value. This removes a further 8% of the data and 2% of expenditure in the data cleaning process.
6. As referenced in Paragraph 24, we have removed transactions that contain child fares to reduce any compositional effects on the price change. Fortunately, the number_of_children field provides exactly that information, so we can exclude any transactions that have at least one child with relative ease. This accounts for 6% of data and 2% of expenditure, though also removes some adult fares (if they are travelling with children).

7. Figure 6 provides a Sankey diagram detailing the flow of data through the data cleaning process, highlighting the amount of data that is removed with each subsequent step. After all data cleaning has taken place, we are left with 58% of the original data set.

Figure 6: The flow of data through the data cleaning process


Raw
Rail tickets
Excluded
CPI related
Excluded
Not late entry
Excluded
GB region
Excluded
Adult ticket
Excluded