# Task 2: Dataset Description

**Assessment of Data Provenance, Quality and Licensing**

The Retail Sales Index dataset, collected by the UK Office of National Statistics (ONS), and the weather datasets provided by the Met Office, can be accessed using the following links:

* Retail Sales Index (data accessed 16th December 2023, 18.50 GMT): <https://www.ons.gov.uk/businessindustryandtrade/retailindustry/datasets/retailsalesindexreferencetables/current>
* Historic Weather Data provided by Met Office: [MetOffice: UK and regional series](https://www.metoffice.gov.uk/research/climate/maps-and-data/uk-and-regional-series) (data accessed 12th December 2023, 16.00 GMT)

Both datasets contain public sector information licensed under the Open Government Licence 3.0 (link to the license: <https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/>). This permits users to “copy, publish, distribute and transmit the information” provided a link to this license is provided, as well as attribution to the source.

The *MetOffice* provides thorough documentation on how its *HadUK-Grid* service collects and aggregates meteorological metrics on its page here: <https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/haduk-grid> The data collected at various weather stations throughout the UK is then used to *interpolate* values for a 1km x 1km grid of values for the entire land surface of the country using multiple-regression with inverse-distance weighted interpolation (<https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/methods>). Consequently, aggregate values such as the rainfall and or the mean temperature for the country are calculated as an average of these 1x1*km* grid cell values. Accuracy is monitored and controlled through the calculation of root-mean square errors (RMSE), as a way of comparing real and predicted values (<https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/faq#faq1>). The MetOffice lists the advantages of using gridded data as improving accuracy and **consistency** of data, and improved completeness without missing values, which weather stations can be prone to.

The ONS compiles the Retail Sales dataset for every month using the Monthly Business Survey – Retail Sales Index (MBS-RSI). Thorough and detailed documentation about the methodology and quality checks used to collect this data can be found here on the ONS’s website: <https://www.ons.gov.uk/businessindustryandtrade/retailindustry/methodologies/retailsalesindexrsiqmi>

As many official institutions such as the Bank of England depend on the validity of these observations, the ONS invests a lot of resources into making this data as accurate and reliable as possible. The RSI collects sales data from a sample of about 5000 UK retailers every month. Although this sample size seems small as it only represents about 2.5% of all retailers, it actually represents “approximately 75.6% of all known turnover in the retail industry”. The data is collected from businesses using an electronic survey. Overall, the ONS enhances data reliability using various mechanisms such as “response chasing” by trained staff to ensure a high monthly response rate. A system of automated checks is implemented to “flag” figures which are above a certain threshold, which are then manually assessed to check for errors.

As such, one should bear in mind that while these datasets undergo rigorous error-checking, the inherent limitations of estimation and interpolation may inevitably introduce some error and inaccuracy.

**Description of Dataset Shape, Size and Data Types**

**Retail Sales Volume Data**

The datasets provided by the ONS on Retail Sales come in the form of multiple spreadsheets. The spreadsheets we will use for this analysis are the *KPSA1*, which represents the seasonally-adjusted percentage change in retail sales volume compared to the same month a year earlier, and *KPSA3*, which does the same but comparing instead the percentage change from the previous to the current month. Each percentage-change is represented as a signed floating-point number.

In contrast to the spreadsheets for retail sales *values* (including prices), which use an index (using the base year 2019 as an index of 100), this facilitates the employment of a linear regression model by overcoming the difficulties which arise when using a relative index. As such, we will compare the effect of the weather on two sets of “target” vectors: the percentage change in sales volume compared to the previous month and compared to the same month a year ago, in order to get an overview of the impact of the weather on both the long-term patterns in consumer spending and on short-term fluctuations. The choice to use *seasonally-adjusted* volumes is to assist in isolating the effect of weather on consumer spending due to accounting for events such as Christmas or bank holidays.

Each table consists of 421 rows indexed from January 1988 to October 2023, with 10 columns for different retail sectors. To start with this analysis, we will focus on the set of values in the column called *All retailing excluding automotive fuel* as an attempt to exclude the impact of the volatility of fuel prices.

The weather data consists of 7 tables each one of which represents a different “feature” or weather parameter. Each *row* in these tables is indexed by a year, while each column represents either a month, season, or annual total. Each cell stores the data value representing the thing being measured. The data values are represented as signed floating-point numbers. The units used are degrees Celsius for the temperature tables, number of days for measuring air frost and days with more than 1mm of precipitation, hours of bright sunshine for that month for the sunshine table, and precipitation is measured in mm. Each table has a different number of rows (years recorded), ranging from starting in 1836 to 1960 up until 2023, but as we only have data for retail sales volume from January 1988 to October 2023, 421 rows will be used here as well once the data is restructured to store a Month-Year per row instead of only the year.

# Data Sources and Provenance

* The weather datasets used for this project are collected by the UK *MetOffice*’s *HadUK-Grid* service. Meteorological data is collected from a network of stations all over the UK (<https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/haduk-grid>). The MetOffice provides detailed documentation on how this weather is collected and how the monthly or yearly averages are calculated for different regions and for the whole country.
* The data collected at the specific stations is then “interpolated to a [high-resolution] 1km to 1km grid”, meaning that data values at those specific stations are used to estimate or predict a “grid” of values for every square kilometre across the UK.
* **How the datasets have been produced:** multiple regression with inverse-distance weighted interpolation are used to generate values on a regular grid from an irregular station network, accounting for factors such as altitude and distance from the coast. <https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/methods>
* Each regional average or aggregate value is an average of this 1x1 km grid cell value that fall in the specified geographical region (in this case, the whole of the UK) <https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/methods>
* More details about this process can be found here: <https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/weather/learn-about/uk-past-events/papers/monthly_gridded_datasets_uk.pdf>
* <https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/weather/learn-about/uk-past-events/papers/cm24_generation_of_daily_gridded_datasets.pdf>
* The accuracy of these grids can be found here using the root-mean-square errors (RMSE) at verification stations: <https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/faq#faq1>
* RMSE is a way of quantifying the difference between predicted and observed values. As seen in this table, the RMSE value for the monthly grid for mean temperature measured over 12 test months was 0.36 degrees Celsius.
* Rigorous quality control measures are used at the stations to correct or remove erroneous data (<https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/faq#faq1>)
* The collection of this data is partly funded by multiple UK governmental departments and research councils (<https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/overview>)
* **Licensing:** the data is provided under the *Open Government License* (<https://www.metoffice.gov.uk/research/climate/maps-and-data/data/haduk-grid/overview>)
* <https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/> The Open Government License allows users to “copy, publish, distribute and transmit the information”, change the information, and exploit the information for commercial and non-commercial purposes by combining it with other information, as long as you provide a link to this license and an attribution to the source of the data, as well as stating that this project *Contains public sector information licensed under the Open Government Licence v3.0.*
* The datasets can be downloaded here: <https://www.metoffice.gov.uk/research/climate/maps-and-data/uk-and-regional-series>
* The Retail Sales Index (RSI) which provides data about the volume and value of UK retail sales over time can be accessed here [Office for National Statistics: Retail Sales Index (RSI) QMI](https://www.ons.gov.uk/businessindustryandtrade/retailindustry/methodologies/retailsalesindexrsiqmi). It is also licensed by the Open Government License v3.0, like the MetOffice data shown above.
* The ONS provides a detail report and documentation here about the methodology used to collect the data <https://www.ons.gov.uk/businessindustryandtrade/retailindustry/methodologies/retailsalesindexrsiqmi>
* Reliability and accuracy of the data: as the RSI is another official government dataset used by institutions such as the Bank of England and the Treasury for data-driven decision making, a lot of care has been taken to ensure that this data is as accurate and valid as possible. The sales data is collected with a monthly frequency and the sample covers 5000 retailers. It is compiled using the Monthly Business Survey – Retail Sales Index (MBS-RSI).
* The sample of 5000 retailers includes all large retailers and “a representative sample of smaller businesses”. The known retail industry is approximately 200,000 businesses and while the sample represents 2.5% of the total population, which may seem very small, it represents approximately 75.6% of *all known turnover in the retail industry*.
* Retail turnover provided by the retailers includes internet and telephone sales.
* “The primary purpose of the Retail Sales Index (RSI) is to produce a short-term measure of the changes in the volume and value of sales of goods by retail businesses in Great Britain, providing a timely indicator of economic performance and strength of consumer spending.”
* “The value estimates reflect the total turnover that businesses have collected over a standard period, while the volume estimates are calculated by taking the value estimates and adjusting to remove the impact of price changes.  ”
* The ONS aspires towards achieving a high standard of data quality and reliability by reducing non-sampling/reporting errors as follows: “very effort is made to minimise reporting error, for example, all Monthly Business Survey – Retail Sales Index (MBS-RSI) respondents who do not return data by the stated deadline are response chased to reduce the non-response rate. Non-response error is minimised through the use of imputation. The careful design of questionnaires, intensive training and supervision of editing and validation staff, and efficient data-processing procedures are also used to minimise reporting errors.”

Estimates for the most recent month are provisional and subject to revision because of:

* late responses to the MBS-RSI
* revisions to seasonal adjustment factors, which are re-estimated every month and reviewed annually
* changes from the annual seasonal adjustment review
* annual updating of the business register that forms the basis for the sample for the RSI (usually occurring in January)
* other methodological improvements

ABS: alternative statistics on retail sales 🡪 Cross-validation?

Retail Sales Monitor – another alternative.

he monthly [Retail sales statistical bulletin](https://www.ons.gov.uk/businessindustryandtrade/retailindustry/bulletins/retailsales/previousReleases) is released in a timely manner, approximately 20 days after the reference period, and is an important early economic indicator. Data are collected for standard trading periods that follow a four-week, four-week, five-week cycle. For the reference period 3 April 2022 to the 30 April 2022, the RSI was released on 20 May 2022.

To ensure a good response rate, approximately 60% each month, response chasing is conducted by trained staff and allows respondents to return data over the phone.

This derived, single score for the business’s return is then tested against a methodologically set threshold. If the score is higher than the defined threshold, then the questionnaire will fail and be flagged for manual editing. Returns that fail selective editing are passed to the editing and validation team for manual checking.

Thresholds have been derived and set to ensure minimum bias is introduced from not editing values that may have been edited under the previous edit all returns system. Therefore, selective editing does result in an adverse impact on output quality. To ensure that all errors are captured, before questionnaires pass through selective editing they are subjected to automatic editing and then a number of user-defined checks. For example, these will include checks to detect changes in reporting periods and implausible returns such as turnover being less than zero. Additionally, as of September 2023 businesses returning zero turnover are automatically flagged and queried.

Source: Office for National Statistics licensed under the Open Government Licence v.3.0

Met Office: Unless we advise you in writing that a different attribution statement should be used you must use the following attribution statement when using the Data in your applications, as required by the Open Government Licence:

"Contains public sector information licensed under the Open Government Licence"

* The report shows that the median year-on-year standard error is estimated at 0.9% and the month-on-month standard error is estimated at 0.5%.