Rail transport: quality assurance of administrative data, Apr 2017

Quality assurance report investigating the administrative data sources used in the production of short-term economic output indicators.

Contact:
Glyn Rice-Mundy
stoi.development@ons.gov.uk

Release date:
4 May 2017

Next release:
To be announced

Table of contents

1 . Introduction

1.1 Background

The National Accounts and Economic Statistics (NAES) group within the Office for National Statistics (ONS) downloads data from the Office of Rail and Road (ORR) on the rail transport industry. These data form one source in the calculation of short-term economic output indicators, namely gross domestic product (GDP (O)) and Index of Services (IoS) for the UK.

This report outlines the process data take from initial collection through to the output of the release. It identifies potential risks in data quality and accuracy as well as details of how those risks are mitigated.

This report forms the latest in a series of quality assurance of administrative data (QAAD) reports produced by NAES to investigate the administrative data sources we use in the production of short-term economic output indicators as set out by the UK Statistics Authority. As such, this report specifically focuses on our administrative data use for the rail transport industry SIC 49.1 and 49.2 only. Separate industries where we utilise administrative data will be considered in other QAAD reports in the series.

Further information relating to quality and methodology for the short-term economic output indicators can be found in our Gross domestic product, preliminary estimate and Index of Services QMI.

1.2 Standard Industrial Classification (SIC) overview

The rail transport industry covers all activities under UK SIC 2007 division 49.1 to 2. Based on the UK Standard Industrial Classification (2007) the industry is classified to two groups:

49.1 – Passenger rail transport, interurban

49.2 – Freight rail transport

According to the Inter-Departmental Business Register (IDBR) [1] there were 90 enterprises classified under division 49.1 or 49.2 in March 2016. This is a decrease of approximately 10 enterprises (negative 10.0%) from the previous year (March 2015).

The majority of enterprises within division 49.1 or 49.2 were allocated to 49.1 – Passenger rail transport, interurban, which equates to 60 enterprises (66.7% of the total industry).

Notes for: Introduction

1. The Inter-Departmental Business Register (IDBR) is a comprehensive list of UK businesses that is used by government for statistical purposes. It provides the main sampling frame for business surveys carried out by both the ONS and other government departments. It is also an important data source for analyses of business activity.

2 . Quality assurance of administrative data (QAAD) assessment

2.1 UK Statistics Authority QAAD toolkit

The assessment of our administrative data sources has been carried out in accordance with the UK Statistics Authority Quality Assurance of Administrative Data (QAAD) Toolkit.

Each administrative data source investigated has been evaluated according to the toolkits risk/profile matrix (Table 1) reflecting the level of risk to data quality and the public interest profile of the statistics (Table 1).

Table 1: UK Statistics Authority Quality Assurance of Administrative Data (QAAD) risk/profile matrix

Level of risk of quality concerns Public interest profile	Public interest profile		
	Lower	Medium	Higher
Low	Statistics of lower quality concern and lower public interest [A1]	Statistics of low quality concern and medium public interest [A1/A2]	Statistics of a low quality concern and higher public interest [A1/A2]
Medium	Statistics of medium quality concern and lower public interest [A1/A2]	Statistics of medium quality concern and medium public interest [A2]	Statistics of medium quality concern and higher public interest [A2/A3]
High	Statistics of higher quality concern and lower public interest [A1/A2/A3]	Statistics of higher quality concern and medium public interest [A3]	Statistics of higher quality concern and higher public interest [A3]

Source: Office for National Statistics

The toolkit outlines four specific areas for assurance and the rest of this report will focus on these areas in turn. These are:

- operational context and administrative data collection
- communication with data supply partners
- quality assurance principles, standards and checks applied by data suppliers
- producer’s quality assurance investigations and documentation

In the assurance of our data source we have chosen to give a separate risk/profile matrix score (Figure 1) for each of the four areas of assurance. This will allow us to focus our investigatory efforts on areas of particular risk or interest to our users (Figure 2).

2.2 Assessment and justification against the QAAD risk/profile matrix

Table 2 – QAAD risk/ profile matrix assessment of administrative data used to measure the Rail Transport industry

	Low A1	Medium A2	High A3
Operational Context and Administrative Data Collection	[A1]		
Communication with Data Supply Partners	[A1]		
Quality Assurance Principles, Standards and Checks by Data Supplier	[A1]		
Producers Quality Assurance Investigations and Documentation	[A1]		

Source: Office for National Statistics

The risk of quality concern and public interest profile has been set as “low” due to the small contribution that the rail transport statistics feed into Index of Services (0.4%) and gross domestic product (GDP) (0.3%). As such, a score of A1 is deemed appropriate for this data source.

All scoring was carried out by National Accounts and Economic Statistics (NAES) based on the level of risk of the data and interest of our users. Results for each area of assurance for rail transport are shown in Figure 2. If you feel that this report does not adequately provide this level of assurance or you have any other feedback, please contact stoi.development@ons.gov.uk with your concerns.

3 . Areas of quality assurance of administrative data (QAAD)

3.1 Operational context and administrative data collection (QAAD matrix score – A1)

This relates to the need for statistical producers to gain an understanding of the environment and processes in which the administrative data are being compiled and the factors that might increase the risks to the quality of the administrative data.

The Office of Rail and Road (ORR) is the independent economic and health and safety regulator for the railway industry in the UK. It is accountable to Parliament and the courts and holds statutory powers. Its role is to regulate Network Rail, which involves setting targets and reporting on their performance. ORR also oversees health and safety compliance and manages competition and consumer rights issues. The Secretary of State for Transport appoints all ORR board members for a fixed term of up to 5 years. ORR is a non-ministerial government department, funded by the rail industry for their rail regulation role and the Department for Transport for their highways function.

The ORR reports (below) that we use have National Statistics status and therefore conform to the statistical standards outlined by the Code of Practice for Official Statistics. As an independent regulatory body, ORR publishes statistics impartially.

3.1.1 Passenger rail transport, interurban (SIC 49.1)

National Accounts and Economic Statistics (NAES) uses passenger rail data provided by ORR as a volume measure within the Index of Services (IoS) and gross domestic product (GDP (O)) for SIC 49.1. The statistical release used for this is passenger kilometres by ticket type. Passenger kilometres are derived by taking the distance in kilometres between two stations, and multiplying it by the quantity of passengers on that service. This is then broken down into passengers using season tickets and passengers using full-price tickets as these services have different costs. ORR collects this data in the Latest Earnings Network Nationally over Night (LENNON) ticketing and revenue database; this is reported quarterly and shown on the National Rail Trends (NRT) data portal. This process is illustrated in Figure 3.

The LENNON database captures the majority of passenger rail data overnight; in 2016, the LENNON database recorded 97.3% of all passenger journeys. The LENNON database captures data from all franchised rail companies but also includes some non-franchised companies. The primary use for LENNON data is to allocate revenue between train operators for tickets sold on potentially shared routes.

ORR also collects non-LENNON data; this data comes from ticket sales, from outlets not linked to the LENNON database. These can include tickets that a train operator sells directly and are only valid on the services they operate. For example, Stagecoach sells tickets through their Megatrain website; tickets sold through this website are only for use on Stagecoach-owned franchises, namely, East Midlands and South West Trains. The train operating companies provide this data on a quarterly basis; 2.3% of passenger rail data for 2016 came from non-LENNON sources.

The suppliers of non-LENNON data are fully coherent in the use of LENNON data and know which data are required to be reported as non-LENNON, which diminishes the potential issue of double counting. LENNON and non-LENNON data are then combined, creating the best estimate of passenger ticket sales.

Data on passenger rail usage is published approximately 60 days after the end of the quarter. Therefore, for the IoS some months' data are estimated based on previous trends to compensate for this time lag.

Limitations of the ORR's data

LENNON data are intended for accounting usage so is not tailored for use in creating statistics on rail performance.

Heathrow express service is not included in the LENNON database and does not report non-LENNON data; this makes up approximately 0.35% of passenger rail data.

Merseyrail report on a six-monthly basis, so the quarters between reporting periods are estimated, and then revised once the data has been received. The estimate is based on the change in rail usage figures in the region and for that quarter based on the same quarter a year previously. The estimates have ranged from underestimating the journey numbers by 4.2% to an overestimate of 10.5%. On a national level, however, this equates to between an underestimate of negative 0.05% to an overestimate of 0.12%.

There are also known problems in the data relating to, amongst other things, travel cards and ticketless travel. Ticketless travel relates to unauthorised passengers using rail transport without a ticket, whereas travel cards include staff passes where staff do not need to purchase a ticket. There is no way of measuring this rail usage; however, this issue does not affect the statistics NAES produces.

Another highlighted issue is that LENNON and non-LENNON data do not account for unused tickets or for passengers alighting early, which may affect passenger kilometres data as the assumption is that the ticket bought is used for the entire journey. It is not possible to measure for this; however, as NAES is concerned with output in the rail transport industry this issue will not influence the statistics they produce.

A full list of these issues can be found in the ORR Passenger Rail Usage Quality Report.

Despite the issues highlighted, the data NAES use from ORR has the most comprehensive coverage of the rail industry available, covering more than 99% of passenger rail movements. Many of these issues outlined by ORR also do not have a large impact on the statistics produced by NAES, as NAES are concerned with the value of the passenger kilometres for the Index of Services and GDP (O), not whether the journey was actually carried out.

3.1.2 Freight rail transport (SIC 49.2)

NAES uses ORR freight-moved data as a volume measure for SIC 49.2, the process for which is shown in Figure 4. Network Rail supplies ORR with “freight moved” data at the end of each railway 28-day period (explained below). This is broken down into seven commodity groups. The commodities outlined in each release are coal, metals, construction, oil and petroleum, international, domestic intermodal and other.

Freight-moved data takes into account the quantity of freight moved on the railway, recording both the weight and distance the freight is carried and is measured in net tonne kilometres. This is published quarterly on the ORR data portal. Freight moved data covers all chargeable freight operations. This data is supplied by Network Rail to ORR for producing National Statistics.

The rail industry reports on a periodic basis rather than typical reporting cycles such as quarterly. There are 13 periods, each of 28 days, per financial year. The reporting periods for a given financial year can start on either 31 March or 1 April. As the reporting periods do not always match up with the quarterly releases, some periods are broken up and apportioned according to the overlap created by the reporting periods.

For example, dates in Period 4 cover both quarter 1 and quarter 2. When quarterly data are calculated for 2016 to 2017, 5/28ths of the data are assigned to quarter 1 (covering 26 June to 30 June) and the remaining 23/28ths of the data are assigned to quarter 2 (covering 1 July to 23 July).

One issue identified with this methodology was the potential for skews created. Taking on the example outlined previously, a reporting period may not have consistent rail usage data, for example, the first half of the month may have significantly less usage than the second half. Simply apportioning 5/28ths of the data to one quarter may not reflect the true proportion of data which should be assigned to the quarter in question. ORR has confirmed however, that this method is the most appropriate for converting periodic data and on the whole is unlikely to cause skews. The method does not affect the annual figure, but may impact some quarterly fluctuations. Full details of the methodology are published in the ORR Freight Rail Usage: Quality and Methodology Report.

ORR publishes quarterly data on freight moved on average 75 days after the quarter ends. Therefore, for the IoS some months’ data are estimated based on previous trends to compensate for this time lag.

3.1.3 ORR revisions policy

ORR publishes a revisions policy in line with the Code of Practice for Official statistics. Their policy allows users to acknowledge any changes to the data through their publicly available revisions log.

Types of revision and process:

- minor revisions – updated in the next release, adjusted data will have a revision flag placed alongside it
- intermediate revisions – shown by an amendment to both the Excel and PDF versions of the data, with any changes highlighted by a revisions flag accompanied by a brief description of the revision
- substantial revisions – same procedure as above; however, to further raise awareness, stakeholders are alerted and a prominent note is placed on the NRT page

To reduce the scope for revisions to occur, ORR waits 1 month before extracting any data from the LENNON database. As the majority of revisions are likely to occur within the first 2 weeks of entry into the database, this gives a good balance between timeliness and quality.

Strengths

- Full coverage of all chargeable freight movements.
- Passenger-rail usage coverage in excess of 99% – only service not covered is the Heathrow express.
- National Statistics badge – therefore meet the high quality criteria laid out in the Code of Practice for Official Statistics.
- Transparent revision policy – users made aware of changes that will help improve their quality.
- Independent regulator – government body, impartial statistics.

Weaknesses

- Apportionment of data between reporting periods may affect quarterly fluctuations for freight rail transport.
- Provisional data and use of estimates for Merseyrail – will inevitably lead to small revisions.
- Heathrow express service not covered by ORR.
- Timeliness – releases published 2 months or more after the end of the quarter and therefore NAES uses estimates to compensate for time lag.

3.2 Communication with Data Supply Partners- (QAAD matrix score A1)

This relates to the need to maintain effective relationships with suppliers (through written agreements such as service level agreements or memoranda of understanding). This includes change management processes and the consideration of statistical needs when changes are being made to relevant administrative systems.

The overall flow of communication from data suppliers to NAES is illustrated in figure 5.

3.2.1 ORR communication with Passenger-Kilometres Data Supply Partners

The Office of Rail and Road (ORR) has expressed that it has good working relationships with its data supply partners. ORR is in contact at least once every 28-day railway period, via email, with the train operating companies (TOCs) who supply them with non-LENNON data. The data collection process has been in place for a number of years and therefore the TOCs are familiar with the expectations and requirements outlined in the memorandum of understanding (MoU) held between ORR and the TOCs, namely:

- the data for passenger journeys and passenger miles sold outside of the Lennon system should be filled in each quarter in the ORR template

- ORR should receive the data 21 days after the end of the quarter

- the data are sent to the generic rail statistics inbox

The MoU held with train operators undergoes a review every 12 months; this is conducted through email. ORR has a quarterly bilateral with the Rail Delivery Group (RDG) who own the LENNON database, and are the representative body of all train operators. The purpose of this meeting, however, is wider than just passenger usage statistics; it covers everything that is going on within the respective organisations. The communication ORR has with RDG on rail usage is generally only if something is amiss with the data, for example, a large unexpected variation in an individual train operator or sector total. ORR does not contact RDG every quarter outside of the quarterly bilateral in the same way as with train operators.

3.2.2 ORR communication with freight-moved data supply partners

ORR has expressed that it has a good working relationship with Network Rail, who supply it with data for freight-moved statistics. There are monthly, face-to-face liaison meetings in place, which provide both Network Rail and ORR with a forum to collectively discuss issues on reporting and monitoring. These meetings can also be used to provide assurance on the data. ORR has confirmed that if there are any issues within the data they are corrected and resupplied by Network Rail.

3.2.3 NAES communication with ORR

National Accounts and Economic Statistics (NAES) does not have a formal service level agreement with ORR. The data used for the IoS is publicly available on the ORR data portal and the logistics of implementing and actively managing formal arrangements are considered both prohibitive and unnecessary for this data, considering the low weight of IoS and GDP (O) it comprises.

NAES also does not have regular formal contact with ORR. However, in the past whenever there has been a query about any of the data there has been a contact email address accompanying each dataset, and ORR has been forthcoming and helpful in providing further information. In addition to this, on an occasion when there was a substantial revision to ORR data, ORR contacted NAES directly and made it aware of the change. As outlined in section 3.1.3, ORR communicates minor and intermediate revisions by flagging them on the Excel and PDF versions of the data. NAES checks for these revisions on a regular basis to ensure it is using the most up-to-date data.

NAES is confident that this working relationship with ORR is sufficient for this data source. There is little need for regular contact as the data are readily available, revisions are communicated, and ORR is easy to contact in the case of queries and willing to provide clarification. ORR has noted, however, that it does not have any knowledge of what data we use nor for what purpose. It would therefore benefit NAES if there were some discussion with ORR in order to inform them of how NAES use these data. This could lead to better communication of relevant information.

Strengths

- ORR is in contact with TOCs at least once every 28-day railway period with a regularly reviewed MoU in place.
- ORR has monthly, face-to-face liaison meetings with Network Rail.
- ORR provides contact information and is helpful in response to any data queries from NAES.
- ORR contacted NAES in the case of a large revision.

Weaknesses

- No formal point of contact.
- ORR not aware of NAES uses of its data.

Next steps

- Set up a formal point of contact and have a discussion with ORR to explain which data releases NAES currently uses, and to explain to ORR the uses of these data.

3.3 Quality assurance principles, standards and checks by data supplier (QAAD matrix score A1)

This relates to the validation checks and procedures undertaken by the data supplier, any process of audit of the operational system and any steps taken to determine the accuracy of the administrative data.

3.3.1 Passenger rail transport, interurban (SIC 49.1)

Figure 6 shows the quality assurance process undertaken by Office of Road and Rail (ORR). ORR first extracts the data from the Latest Earnings Network Nationally over Night (LENNON) database. The structure of the data is then checked automatically to ensure it has been correctly entered into the system, for example, if there are data containing letters where it should be containing numbers, this will be rejected by the data warehouse. Once the data have passed the automated checks, the steps for data validation are as follows:

- a member of the Information & Analysis team validates the data warehouse output against the original data file extracted from LENNON
- draft versions of the data tables produced via Business Intelligence Development Studio
- data tables are then quality assured – this includes comparison of data with previous quarters or years, comparison of change with journey and revenue reports, checks of revenue per passenger-kilometre, revisions and so on
- Head of Profession or Deputy Head of Profession then signs off statistical release and portal tables

In addition to the regular process above, ORR occasionally calculates an average journey length using passenger-kilometres and passenger journey data to check whether the output looks realistic.

An arm of the Rail Delivery Group (RDG, formerly ATOC) owns the LENNON database. ORR sends the statistics it produces to RDG for a pre-release check, as they can provide an insight as to why numbers are moving in a particular direction. RDG also publishes passenger demand statistics on an ad hoc basis, therefore ORR may provide them with numbers ahead of pre-release for the purpose of sense checking the data.

The quality assurance measures for non-LENNON data follow the same process as LENNON data, and similarly if there are concerns about non-LENNON data, ORR can go back to the train operators to confirm the numbers.

3.3.2 Freight rail transport (SIC 49.2)

The process for quality assuring freight data is illustrated in Figure 7. National Rail uploads freight-moved data directly to the ORR SharePoint upload site; loading packages then process the data. The loading packages perform a first set of checks for whether uploaded data have been supplied in line with protocol, for example, the file naming convention and data types in each field required are as expected. If they are not, that particular load fails and is assigned “reject” status; Network Rail then resupplies this data. If the data pass these checks, the load progresses and gets stored in the ORR data warehouse (SQL database). The data are assigned a “draft” status until validated.

ORR then validates the loaded data following the steps outlined in section 3.3.1. If the data pass validation, they are assigned “approved” status and are made available for further analysis. If the data do not pass this step, the load is assigned a “reject” status. Data with a reject status are not made available for any further analysis; this can result in the data being resupplied if required and reprocessed through the loading packages. ORR then repeats the validation process. This is repeated until all of the data pass and can be assigned an “approved” status.

In addition to these checks, Arup independently review and report on this dataset. Both Network Rail and ORR appointed Arup as an independent reporter in 2009. Its role is to review data from Network Rail to provide ORR with assurance of its accuracy and reliability.

Strengths

- Clear and extensive data quality-assurance process both using automated and expert checks.
- Data resupplied if errors are found.
- Independent review – freight data reviewed by Arup.

3.4 Producers quality assurance investigations and documentation (QAAD matrix score A1)

This relates to the quality assurance conducted by the statistical producer, including corroboration against other data sources.

National Accounts and Economic Statistics (NAES) uses publically available Office of Road and Rail (ORR) data as a volume measure for the rail industry. These are aggregated up in the Index of Services to provide a figure that encompasses the whole of division 49 (land transport). However, in the gross domestic product (GDP) release there is a full industry breakdown that separates rail transport from the other modes of land transport in division 49. The quality-assurance process for ORR administrative data undertaken by NAES is shown in Figure 8.

NAES has specific and up-to-date desk instructions for collecting and processing rail data. The desk instructions note that the rail tables work on the financial year calendar whereas NAES uses calendar quarters; the instructions define this to mitigate any risk of confusing which data relate to which period.

The NAES process

NAES downloads both the quarterly data tables on passenger-kilometres and the freight-moved data tables.

NAES then carries out sense checks on the data on an Excel spreadsheet to ensure it looks reasonable and is in line with recent trends.

The data at this stage are uploaded to an internal ONS system as a CSV file where seasonal adjustment is applied.

This system is able to impute missing data, to compensate for the time lag between the data quarter and the publication of ORR data, and create a reliable estimate for the months that are missing. The internal system imputes the data at the lowest level, based on specification files inputted by the time-series analysis branch. These files specify which model is the best to use for imputation for this series. In the case of ORR data, the imputation is carried out using an X13-ARIMA-SEATS model.

In the IoS rail, data are published as part of an aggregate for SIC 49 (land transport).

NAES keeps a full audit trail of any previous versions of the data and all previous CSV files are stored on an internal drive. Therefore, if there are any revisions, NAES would be able to identify where and to what extent the change had taken place.

Due to the low weighting of rail transport on IoS (0.4%) and GDP(O) (0.3%), there has been very little public interest in these rail transport statistics. NAES has had no feedback to date in relation to the rail transport industry, even despite a recent correction to the data where NAES reached out directly to users to notify them of this change.

Strengths

- Sense checks – data checked to ensure it is in line with previous trends.
- Audit trail – NAES keeps copies of historic CSV files.
- Reliable estimation – NAES has the capability to estimate missing data in the internal system.

Weaknesses

- Use of estimated data rather than actual values, offset by the requirement to deliver a timely dataset.

4 . Summary

The process for the overall flow of rail transport statistics from suppliers to National Accounts and Economic Statistics (NAES) is presented in Figure 9.

In investigating the administrative source for rail transport, NAES considers the main strengths of the data for our purpose to be:

- high level of coverage for passenger rail kilometres (greater than 99%) and full coverage of all chargeable freight movements
- high level of quality assurance by Office of Road and Rail (ORR) including independent review
- NAES quality checks and comparison with previous year trends
- compliance with Code of Practice for Official Statistics We believe current limitations of this data source are:
- no direct point of contact with ORR
- no discussion with ORR to inform them of the releases used by NAES or the purpose of NAES usage

In constantly seeking to improve our data sources, we will be taking next steps to address these limitations. These will be communicated to users in the future QAAD report updates for this topic.

However, despite these slight limitations, based on the low risk of quality concerns and small contribution that the rail transport statistics feed into Index of Services (0.4%) and GDP (0.3%), NAES considers this data source to be fit for purpose, and to fulfil the requirements of an A1 assurance rating.