Improvements in editing methods and processes for use of Value Added Tax data in UK National Accounts
Prepared by Martina Portanti and Robert Breton, Office for National Statistics, United Kingdom

I. Introduction

1. ONS 2013-2023 strategy puts at its core an increased use of administrative data and less reliance upon traditional surveys. Within this backdrop, the office is investing in developing methods and a processing platform to use Value Added Tax data (VAT) data into the UK National Accounts in 2017.

2. ONS carried out an extensive programme of methodological research on use of VAT turnover data as part of the EssNet on the Use of Administrative and Accounts Data for Business Statistics.

3. Since the end of the Essnet in 2013, ONS has continued methodological work to address specific issues related to the use of VAT turnover data to produce short term statistics. At the same time, a SAS system based on the research developed for the Essnet has been used by the VAT analysis team within ONS to process VAT turnover data on a monthly basis.

4. The move from a mainly research-based approach to VAT data to a regular statistical production environment has highlighted a number of limitations with the existing methods and systems. The processing of the very large amount of data required to produce VAT-based estimates of turnover has particularly been problematic. In 2016, ONS took a decision to invest in developing a new platform to process VAT data.

5. The platform redevelopment has provided an opportunity to review the existing methods for VAT and to assess whether they are still fit for purpose. In October 2016, a 8-weeks Discovery was launched to carry out a methodological review of the methods required to use VAT data in National Accounts, including methods for:
   (a) Cleaning
   (b) Linking and apportioning
   (c) Calendarisation
   (d) Estimation

6. This paper focuses on cleaning methods for VAT data, presenting an overview of current methods and areas for improvements. In doing so, we will consider the following aspects related to the data editing process:
   (a) The position of cleaning within the overall VAT data processing sequence
   (b) The validation methods (i.e. the methods to detect errors)
   (c) The cleaning methods (i.e. the methods to treat errors)

7. The methodological review has provided the statistical basis for the system development work. This officially started in January 2017, with a Minimum Viable Product due to be delivered by June
2017. ONS has adopted a new agile approach to the platform development, which provides an opportunity to work collaboratively across methodologists, data scientists and IT developers.

II. Key characteristics of VAT data

8. In order to facilitate comprehension of the rationale behind the design of both the existing VAT cleaning methods and future improvements, some key aspects of the UK VAT data are outlined in this section.

9. Most businesses (89%) report VAT data on a quarterly basis. These businesses will file their VAT returns according to one of 3 patterns (staggers): 1) reporting for January to March; April to June etc; 2) reporting for the period February to April; May to July, etc; 3) reporting for the period March to May; June to August, etc. A small proportion of businesses (around 1%) report on an annual basis instead. Similarly as quarterly businesses, they are allowed any pattern of reporting e.g. year ending in January, year ending in February and so on. Overall, there are therefore 12 annual staggers. Finally, some business report on a monthly basis – these businesses are classified as stagger 0. Monthly businesses represent 10% of total VAT units. ONS receives a monthly feed of VAT data from HMRC. Each monthly file will contain a mix of monthly, quarterly and annual reports.

10. HMRC data is collected via VAT reporters (or VAT units). ONS surveys are instead sampled from Reporting units (RUs). VAT units and RUs can be linked via the enterprise(s) they refer to. The relationships between the three entities (VAT units, Enterprises and RUs) can be quite complex.

(a) One VAT unit is linked to and reports for one and only one Enterprise (One-to-one)
(b) One VAT unit is linked and reports for many Enterprises (One-to-many). HMRC refers to this as a group registration
(c) Many VAT units are linked to one and only one Enterprise (Many-to-one). HMRC refers to this as divisional registration
(d) Many VAT units are linked to many enterprises (Many-to-many). This appears to happen when an Enterprise belonging to a group registration is also linked to other VAT units as in a divisional registrations

11. Reporting units are observational units created within ONS Interdepartmental Business Register (IDBR) which is used as the sampling frame for business data collection. The relationship between RUs and enterprise is hierarchical and can be only one of the following two:

(a) An enterprise has one reporting unit only (one-to-one)
(b) An enterprise has more than one reporting unit (one-to-many).

12. Businesses which consist of a single VAT unit linked to one and only one enterprise which has only one RU are called “simple”; everything else is classified as a “complex” business.

13. In using monthly VAT turnover data in UK national accounts, it is essential to link the VAT unit –level data to RU level using the IDBR, which holds information about a business economic activity (SIC code) and size in term of employment. This information is essential to derive estimates by industry. In order to link VAT and RUs, HMRC turnover needs to be apportioned in a first stage to enterprises; and to Reporting Unit in a second stage. Currently, turnover is apportioned based on each enterprise’s and reporting unit’s share of employment.

III. Existing cleaning methods

14. A high level process map of the overall VAT cleaning process can be found in Appendix I. In summary, validation and cleaning of VAT data takes places in three parts of the overall VAT processing system.

A. Cleaning of VAT records at take-on
15. Basic (e.g. structural) validation and cleaning of VAT records takes place immediately as records are taken on by ONS system. At this point in time the following validation takes place:

   (a) Identification of valid returns
   (b) Identification of duplicate records

16. Businesses may submit more than one VAT return to HMRC for the same reporting period, because of corrections requested by HMRC or identified independently by the businesses. Each monthly VAT file contains returns made by businesses during that month, but they may refer to different reference periods. For example, a business may submit a return for the January-March reference period in the April monthly file but then re-submit a corrected January-March return at a later stage (e.g. in a June monthly file). Multiple returns for the same reference period can therefore be found within the same monthly file (if multiple submissions for the same reference have been made during the same period) or across different HMRC files (if multiple submissions have been made in different periods). Duplicate records are automatically treated by identifying the most recent submission. Previous records are deleted.

B. Cleaning of RU-level turnover at stagger level

17. After VAT unit level turnover is apportioned to RU level, more complex cleaning takes place (content validation). The following validation rules are applied:

Thousand pounds rule

18. When submitting a VAT return, businesses are supposed to report their turnover to HMRC in pounds. It may happen that a business incorrectly submits a return in thousands pounds instead. These errors are identified by first matching all VAT data from the current period to corresponding previously returned turnover values, where available. For all businesses that have both current and previous turnover values present and where turnover is not listed as zero, the current turnover for the business is divided by the previous turnover for the business to calculate the following proportion:

\[ \text{Ratio 1} = \frac{\text{Current VAT Turnover}}{\text{Previous VAT Turnover}} \]

19. If the calculated ratio (Ratio 1) is between 0.00065 and 0.00135 the turnover value for the current month is said to have failed the thousand pounds rule. The current period turnover value for a business that fails the thousands pounds rule is automatically edited by multiplying the original return by 1000.

Quarterly pattern rule

20. This rule enables to understand whether businesses who report on a quarterly basis are reporting true quarterly turnover by identifying suspicious patterns in the data. There are three variations of the rule:

   ● If businesses report exactly the same positive values for any four consecutive quarters, then it can be suggested that the business is actually reporting one annual figure and allocating this equally between the four quarters.
   ● If a RU reports the same positive value in any three consecutive quarters followed by a different value in the fourth quarter it implies that the business is using an approximate estimate of its annual value and allocating it between the four quarters. The fourth quarter is then allocated the residual sum (a balancing item) to sum to the annual value.
   ● If a RU reports a turnover value of zero in any three quarters and then a positive value in the fourth quarter it implies that the business is returning an annual value.

21. In the above examples the annual value is re-assigned to the corresponding quarters by:
Calculating median proportions between the quarters from businesses in the same class with genuine reporting values. A class is defined as a grouping of businesses with the same industry classification and a similar number of employees.

Calculating the annual turnover figure for the business failing any of the quarterly pattern validation rules above

Applying the relevant class quarterly median to the business in error to derive quarterly turnover figures

**Suspicious turnover rule**

22. This rule identifies reporting units that have suspicious turnover for a VAT return and is based on comparing each RU turnover value against the median turnover of businesses in the same class across time. Two scores are calculated:

- Score 1 is calculated by dividing the current VAT turnover of the RU being investigated by the median of the current VAT turnover in the class and stagger that the business belongs to:
  \[ \text{Score1} = \frac{\text{Current VAT Turnover}}{\text{Median of current VAT Turnover in Class and stagger}} \]

- Score 2 is calculated by dividing the previous VAT turnover of the RU being investigated by the median of the previous VAT turnover in the class and stagger that the business belongs to:
  \[ \text{Score2} = \frac{\text{Previous VAT Turnover}}{\text{Median of previous VAT Turnover in Class and stagger}} \]

23. A ratio value (Ratio 2) is then calculated from these scores. If score 2 is larger than score 1, then Ratio 2 is calculated in the following way:

\[ \text{Ratio2} = \frac{\text{Score2}}{\text{Score1}} \]

24. If score 1 is the same as or larger than score 2, then Ratio 2 is calculated as follows:

\[ \text{Ratio2} = \frac{\text{Score1}}{\text{Score2}} \]

25. Finally, Ratio 2 is examined to determine whether the business’s current turnover value can be deemed suspicious:

- For businesses reporting on a monthly basis, any RU with a value of Ratio2 greater than seven is deemed suspicious.
- For businesses reporting on a quarterly basis, any RU with a Ratio2 value of greater than five is marked as failing the suspicious turnover rule.
- For businesses reporting on an annual basis, any business with a value of Ratio2 greater than four is considered to be suspicious and to have failed the cleaning rule.

26. When a business’s turnover value has been marked as suspicious it is replaced by using ratio imputation. First, a ratio (Ratio 3) is calculated as the sum of VAT turnover for the current period divided by the sum of VAT turnover for the previous period (for the same class and stagger as the RU that has reported a suspicious turnover value):

\[ \text{Ratio3} = \frac{\text{Sum of Current period Turnover in Class and stagger}}{\text{Sum of Previous period Turnover in Class and stagger}} \]

---

1 See De Jong (2003)
27. Only businesses with turnover data in both the current and previous periods (for any stagger) and that have not failed any cleaning rules are used in this calculation. The new turnover value for each business that failed the suspicious turnover rule is then imputed as Ratio 3 multiplied by the previous turnover value for the RU:

\[ \text{New Turnover} = \text{Ratio 3} \times \text{Previous Turnover} \]

C. Cleaning of RU-level turnover at monthly level

28. Following the cleaning at RU level by stagger, the data is calendarised to produce monthly series. At this stage of the process, a final round of cleaning takes place and the following rules are re-applied:

- (a) Thousands pound rule
- (b) Suspicious turnover rule

29. Both the validation and cleaning methods mirror those carried out at the stagger level, with the difference being that the previous period data for comparison is the previous month as opposed to the previous month, quarter or year depending on the stagger the business belongs to.

III. Assessment of current cleaning process

30. Based on analysis of the current system, including internal consistency, theoretical fit and performance (including feedback from the VAT analysis team within ONS), improvements to the current cleaning process and methods are advisable. Specific recommendations are outlined below. A summary of the recommended changes is also captured in a flowchart included in Appendix 2

A. Position

31. Only structural validation takes place on VAT unit level records. All the content validation is carried out after VAT unit data is apportioned to RU level. While this is not an issue for simple enterprises (as VAT unit equals enterprise equals Reporting Unit), it is problematic for complex enterprises as it becomes difficult to understand what drives large changes over time for a particular reporting unit. Are there actual large changes in reported turnover from the VAT unit or are there changes in the VAT structures underlying an enterprise/RU that affect the apportionment between reporting units?

32. More in general, a principle of data editing is “earlier the better”, so editing rules should be applied as early in the process as possible, e.g. on VAT unit level data rather than on apportioned RU turnover. More specifically, the following validation rules:

- (a) thousand pounds rule
- (b) quarterly pattern rule

could be applied at the VAT unit level.

33. It should also be noted that the quarterly pattern rule may not identify correctly suspicious quarterly patterns for complex businesses. While a VAT unit may be reporting a suspicious quarterly pattern, this pattern may disappear when the data is apportioned to RU level if there are changes in the apportionment variables during the previous 12 months period (e.g. employment and number of RUs).

34. While detecting quarterly patterns errors could be applied at the VAT unit level, the current cleaning method relies upon using median pattern calculated by class, whereas class is defined as industry by employment sizeband. This cleaning specification will need to continue to be applied at RU level, as it relies upon IDBR information. Alternatively, consideration could be given to calculate median
patterns using HMRC SIC information\(^2\) only, which would be available at VAT unit level.

35. Similarly, the current specification of the suspicious turnover rule relies upon building validation and imputation classes using RU-level IDBR variables (employment and industry) and as such it can only take place after apportionment to RU level. However, suspicious returns at the VAT unit level could be identified using different validation rules, for example looking at large period-on-period percentage changes for each VAT unit.

36. In devising a content validation module at VAT unit level, consideration should be given in whether information on the level of complexity of a business, employment size, SIC etc, which can be derived after linkage to the IDBR, could be fed back to the VAT units. This may help in assessing the suitability of VAT-unit level content validation methods for different type of units and groups of businesses.

37. The latest stage of cleaning, which takes place after calendarisation needs to be revisited. The thousand pound rule appears to be redundant at this stage and could be removed. As for the suspicious turnover rule, it is unclear how much value it adds at this stage. The rule is applied earlier in the process by stagger. There may be a benefit in re-applying the rule after calendarisation to create larger pools for comparisons, as each business is compared to similar businesses across all staggers. This would be particularly beneficial if comparison classes by stagger are too small. However, it does add an additional layer of complexity in the data process and it will be affected by the accuracy of the calendarisation method. The current imputation method for cases that fail the suspicious turnover rule on monthly data also results in the turnover being readjusted using monthly imputation links, hence the calendarised data for these cases is inconsistent\(^3\) with those that did not require further intervention after the RU-level cleaning by stagger.

B. Validation methods

38. The thousands pounds and quarterly pattern rule are easily justifiable to detect very specific known error in the data. The suspicious turnover rule is probably open to more scrutiny. Previous research carried out as part of the Essnet on Administrative data had looked at a number of different methods to detect suspicious turnover. However the research should be revisited to use more recent data and to ensure that large businesses are appropriately accounted for.

39. In assessing different methods to detect suspicious turnover values, consideration needs to be given in separating methods that could be applied at the VAT unit level (e.g. methods using business-level historical information only) and methods that can only be applied at RU level (e.g. methods relying upon information from other businesses in the same industry and employment group).

C. Cleaning methods

40. Surveys normally carry out a mix of automatic cleaning (e.g. for thousands pounds errors) and clerical intervention depending on the nature of the error. Clerical intervention can either consist of an “informed decision” – e.g. validation error is solved based on previous information collected for the survey or on the business – or a “call-back” to the business.

41. With the VAT data, because of the impossibility of recontacting businesses to confirm the results, the strategy has been to carry out automatic error correction only. While this approach is justifiable to correct thousands pounds and quarterly pattern errors, it seems more risky when treating suspicious turnover errors. A suspicious turnover is not necessarily an error, but only an implausible value – automatic imputation of all these validation failures seems an excessively conservative approach.

---

\(^2\) Although quality of HMRC SIC information is deemed worse than ONS SIC information, it may be of sufficient quality to allow provisional cleaning.

\(^3\) A business that failed the monthly-level suspicious return rule will have a different turnover value for each month of the quarter, while a business in the same stagger that did not fail this rule will have the same value for each month of the quarter.
42. Consideration should be given to introduce an “informed decision” stage with a consistent clerical review of a manageable number of returns that fail the suspicious return validation rule. Following review, one of the following actions could be taken:

   (a) accept original value
   (b) use imputed value
   (c) carry out manual construction

43. Because of the large number of VAT returns, a selective editing strategy would need to be adopted, whereas only returns with large impact on estimates are manually investigated.

44. A score for each business can be calculated as:

\[
\text{Score} = \frac{\text{|Previous Year Value} - \text{Current Year Value}| \times 100}{\text{Current Year Value}}
\]

45. The expected value can be the previous period return; or the previous period return multiplied by a growth factor or imputation link. The domain estimate would be the previous period estimate for each industry of interest.

46. Only cases with scores above a certain threshold get normally validated, where the threshold is set to either control bias; or based on resources available to carry out validation. In the case of VAT data, a threshold could be set so that an agreed manageable number (e.g. 300 cases) with the highest scores would be manually checked. The remaining could then be either automatically imputed or left untreated, depending on the overall impact on the estimates. This selective editing strategy will need to be built within the RU-level content validation, as final estimates from VAT data are based on the IDBR SIC classification. However, additional manual editing based on a selective editing approach could be also built earlier in the process within the VAT unit-level content validation by using HMRC SIC information to build provisional groups of businesses.

47. This in essence mirrors the approach currently adopted by the VAT analysis team within ONS, who use macro-editing techniques to validate VAT data. Their macro editing work consists in identifying large movements in the aggregate series and drill down to the businesses which have caused the movements. In most cases, errors with the data are identified and corrected on ad-hoc basis.

48. The advantages of introducing a manual clerical review based on a selective editing approach are:

   (a) identify errors earlier in the process
   (b) apply a more consistent editing approach across all the data
   (c) decrease the time and effort required for macro-editing
   (d) improve knowledge of VAT data
   (e) identify genuine business changes (e.g. restructures) and treat them accordingly
   (f) inform future cleaning methods improvements

49. It would be ideal to carry out an analytical review to assess the performance of a range of imputation methods to treat suspicious turnover returns, including:

   - Ratio imputation
   - Class mean/Median imputation
   - Donor imputation

50. Another option to deal with cases failing the suspicious return validation rule would be to treat them as non-respondents and carry out non-response weighting. In practice, if imputation classes correspond to non-response weighting classes, non-response weighting equates to a class mean
imputation approach. However, the method currently employed within the VAT processing (ratio of means imputation) is a standard imputation method used at ONS and known to perform satisfactorily on business data. For this reason, it has been agreed to maintain the current method specification for a VAT MVP and carry out an imputation methods review at a later stage.

IV. Ongoing development

51. The following editing improvements were agreed as part of the Discovery on VAT methods

(a) Building a content validation module at VAT unit level. This should include, as a minimum, the thousands pound rule and the quarterly pattern rule. Consideration should also be given to include a rule to identify suspiciously large movements in the data, although their treatment (e.g. imputation) may need to take place at the RU level to make use of IDBR information.

(b) Considering whether linkage back of IDBR variables to the VAT-unit level data may enhance VAT content validation.

(c) Assessing the added value of the final cleaning stage on monthly data and consider whether this could be removed.

(d) Retaining the thousands pounds rule and quarterly pattern rule as detection rules within the content validation of VAT data. The suspicious turnover rule based on period-on-period ratios should be reassessed alongside alternative methods once a test system is available.

(e) Introducing a systematic clerical review of a manageable number of suspicious VAT returns based on a selective editing approach. This could be initially built at the RU-level data, but consideration could be given to introduce it also at the earlier VAT-level data stage.

(f) Reassessing the imputation method used to automatically edit cases that fail the suspicious return validation rule

52. In December 2016, a HMRC VAT content validation module was created following the recommendations of the methods discovery. The module included suspicious turnover detection, quarterly patterns detection and thousand pound detection and correction. This module was completed in a four week sprint (following an agile approach).

53. The module was developed on distributing computing technology (Spark being the main component). This platform allowed rapid processing times – 5 minutes to process 4 years of HMRC data, as opposed to a day, using conventional computing. This confers obvious advantages for testing and statistical production systems.

54. The four week sprint also made extensive use of data visualisation (specifically Jupyter notebooks). This allowed weekly demonstrations of progress to methodology and the business area. This enabled continuous feedback and improvement, in the development progress.

55. The other modules are being developed on the platform (to be completed by June), covering:

- Further cleaning
- Apportionment
- Estimation
- Aggregation (pre step to integration into National Accounts)
Appendix I – Cleaning process summary map – AS IS

AS IS cleaning process

- VAT unit level data
  - Structural validation
    - Valid entries and duplicates
  - Cleaning
    - Automatic cleaning
  - Apportionment to RU level
    - RU level data at stagger level
      - Content validation
        - Thousand pounds rule
        - Quarterly pattern
        - Suspicious turnover
      - Cleaning
        - Automatic cleaning
      - Edited RU level data at stagger level
        - Calendarisation
          - Content validation
            - Thousand pounds rule
            - Suspicious turnover
        - Cleaning
          - Automatic cleaning
        - Edited RU level monthly data
          - Validation
            - Cleaning
          - Other methods
Appendix II – Cleaning process summary map – recommended changes

References

De Jong, A. (2003), Impact: Recent developments in harmonized processing and selective editing.
Proceedings of UNECE Work Session on Statistical Data Editing, Madrid, October 2003.


ONS, *Further information about IDBR sources, structure and updating for publications*, ONS website