Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Neuchâtel, Switzerland, 18-20 September 2018)

Investigating methods of efficient detection of errors in VAT data

Prepared by Katie Davies, Office for National Statistics, UK

I. Introduction

1. As part of Business Statistics Transformation, the Office for National Statistics (ONS) are investigating including administrative data sources such as VAT data to supplement or replace survey data to reduce burden. The Distributive Trade Transformation project is one such instance where this is being explored.
2. The Distributive Trade output will produce estimates of the level of monthly turnover in NACE Rev. 2 Section G (Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles). It is proposed that VAT turnover data will be used to replace some survey data. HMRC collect this information but it is not validated, as this is additional information that they collect, meaning that there could be more errors in this data. Investigation is required to identify methods to detect and remove errors from the data.
3. Error detection methods need to be efficient, because of the volume of data, minimising resources required. Two main approaches were considered: Selective Editing to identify and treat suspicious turnover at a micro level, and Outlier Detection to identify further errors followed by treatment at the macro level.
4. The results presented in this paper are based on analysis of Division 47, (Retail trade, except of motor vehicles and motorcycles).

A. Automatic Cleaning

5. The following automatic edit rules are already applied to VAT data (ons.gov.uk, 2016):
   (a) Thousand Pound Rule: Accounting for businesses that have returned values 1000 times too small, as for surveys turnover is often asked for in thousands of pounds.
   (b) Quarterly Pattern Rule: Applied only to calendar quarter stagger; identifying cases where businesses estimate turnover during the first three quarters then use the fourth quarter to balance turnover for the year. These are then adjusted using ‘genuine’ quarterly pattern information defined by businesses with similar characteristics.
6. The cleaning methods tested in this paper were applied after these automatic edits.

B. Staggers

7. Businesses can provide data in different ‘staggers’ as portrayed in Table 1. Most businesses return on a monthly or quarterly stagger, with 1% providing annual returns (ons.gov.uk,2016). For efficiency, cleaning methods were not applied to annual returns.

Stagger	Returns
00	Monthly
01	Quarterly; calendar quarters (Jan-Mar etc)
02	Quarterly; Feb-Apr, May-Jul, Aug-Oct, Nov-Jan
03	Quarterly; Mar-May, Jun-Aug, Sep-Nov, Dec-Feb
04-15	Annual starting with different months 05 starting in January 06 starting in February … 15 starting in December

Table 1: Structure of returns for each stagger

8. An issue encountered whilst cleaning the VAT data was that there were some businesses reporting in periods that they were not required to according to their staggers. There was some uncertainty as to whether these returns were referring to a period of the same length or the time between last return and this extra return. It was deemed that the stagger variable was more reliable than the period in which they returned so all returns made outside of the stagger reported were removed from the data before micro level treatment has been applied.

C. Calendarisation

9. As Distributive Trade estimates will be reported monthly, calendarisation is required to produce monthly values from the quarterly and annual values. Calendarisation reduced the effect of some of the errors seen in the quarterly data as they were spread across three months. Therefore, detection and treatment of errors was required prior to calendarisation. To allow for this each stagger has been treated separately. For the cell[1] level plots provided, the monthly and quarterly data have been combined using a simpler calendarisation method to provide an indication of how the methods tested work. As the simpler method involves dividing the quarterly values by 3 to allocate to each month in the quarter the plots provided will be smoother than what will be seen when it is run through the system fully. The cell level plots have also been provided without further estimation through weighting.

D. Estimation

10. The deadline for VAT returns to be submitted is 1 month and 7 days after the end of the reference period. For a business reporting for the quarter January to March, the deadline is May 7th. ONS receives the data from HMRC on the first working day of the month. This means that VAT returns could be received as late as the first working day of June. If the month of interest were January then calendarised data from this stagger for January could take 5 months to become available.

11. Estimation is used to account for VAT data which has not yet been received, because of the timeliness issue. Additionally, there is a small proportion of RUs which, for various reasons, never receive VAT data.

12. The estimation method used is ratio estimation with register turnover[2] as the auxiliary variable. Data used in this analysis were taken once all VAT staggers had been received so was close to complete. In practice cleaning will be applied in real time therefore estimation will need to be accounted for.

II. Selective Editing

13. Selective Editing is a method for prioritising possible errors in data. An edit score is calculated for all VAT returns, with higher scores indicating more suspicious values. VAT returns with the highest

[1] Reporting Units are assigned to cells based on their industry and employment size band. Cells are used for estimation purposes.

[2] Register turnover used here was the turnover from the Inter-Departmental Business Register. This will be from the Statistical Business Register in the future.

selective editing scores were manually validated, and where returns were determined to be errors, treated using Ratio of Means imputation (ec.europa.eu, 2008). Each stagger was treated separately.

A. Score function

14. A variety of score functions were tested as listed in the Annex. The score was calculated for each Reporting Unit (RU) and any alterations are transferred back to the VAT unit level in cases where there is a one-to-one match.

15. The leading function is detailed in Equation 1. Introducing the standard deviation to the denominator proved more effective as there were some RUs that were much more volatile than others. Without the introduction of the standard deviation, the score was prioritising returns from RUs that were known to be more volatile, but given their history, these values were typically plausible and were not contributing to errors at the aggregate level.

16. This score was tested from the start of 2012 to the first quarter of 2017 to assess its performance over time. Analysis assumed a weight of 1 for each RU as the majority of, if not all of, the population of interest would be available for the periods of interest. In live production as the VAT data are cleaned in real time, this would not necessarily be the case as there are multiple deliveries of the data. To account for this, weights may not equal 1 and should be calculated using register turnover.

$$Score = \frac{w|x - \mu|}{\sigma \hat{t}_{d,t-1}}$$

Equation 1: Selective Editing score function

Where

$x = Turnover\ for\ current\ period$
$\mu = Mean\ turnover\ over\ the\ last\ 3\ years$
$w = \frac{Register\ turnover\ for\ population\ in\ current\ period}{Register\ turnover\ for\ sample\ in\ current\ period}$
$\sigma = Standard\ deviation\ of\ turnover\ over\ the\ last\ 3\ years$
$\hat{t}_{d,t-1} = Total\ turnover\ for\ domain\ d\ in\ previous\ period\ t-1$

B. Treatment

17. Initially the top 5 cases were treated for each stagger using Ratio of Means Imputation for each period consecutively. As the businesses are spread across different staggers, imputation classes were created based on the last time the business returned and only used RUs from NACE Rev. 2 Division 47. Further breakdown by industry, for example, would make the classes too small and therefore not necessarily providing a suitable growth factor. The last time the business returned is not necessarily the same as the stagger as a business may have changed stagger between the last return and return to be considered. Testing found the errors with the largest impact were in the quarterly staggers. From this point the monthly stagger has not been treated using Selective Editing.

18. Selective Editing was set to take one point of interest at a time, cleaning initial time points before considering more recent time points. This is reflected in Figures 1 and 2 where there is only one point of interest for the treatment and a further time series has been included for reference. The data past the point of interest will not have been considered for cleaning at that time.

19. In Figure 1 there is a large peak in the original data, for the quarter ending April 2012, that is highly likely to be an error. This RU was identified in the top 5 scores for the given period. The treated line, using Ratio of Means imputation, produces a value closer to what could reasonably be expected given the history and future values. But there are also cases where treatment would not be appropriate. Examples are given in Figure 2.

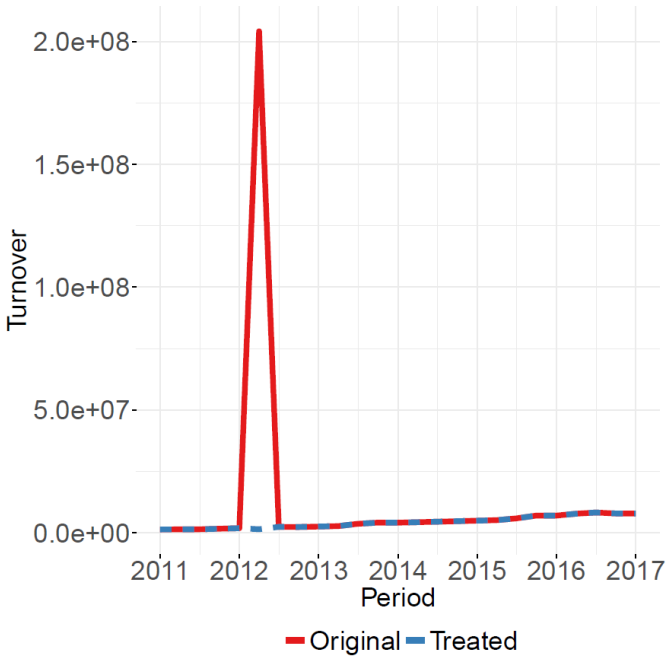


Figure 1: Time series of VAT turnover for an RU with a suspicious value. Treated value is imputed value under ratio imputation

20. For the quarter ending June 2012 in Example 1 of Figure 2 the imputed value is more extreme than the original value that was identified as suspicious. This is because other businesses are generally experiencing a growth of a larger magnitude in the same direction. Therefore, the original value no longer appears to be suspicious and should be kept. Identification can be made at the time of treatment with no future knowledge required so would not result in revisions being required.

21. For the quarter ending April 2012 in Example 2 of Figure 2 the imputed value is working in the opposite direction to the original value. The future returns suggest that this is because other businesses generally experiencing growth in a different direction. The future returns help provide evidence in agreement of this so revisions would be required.

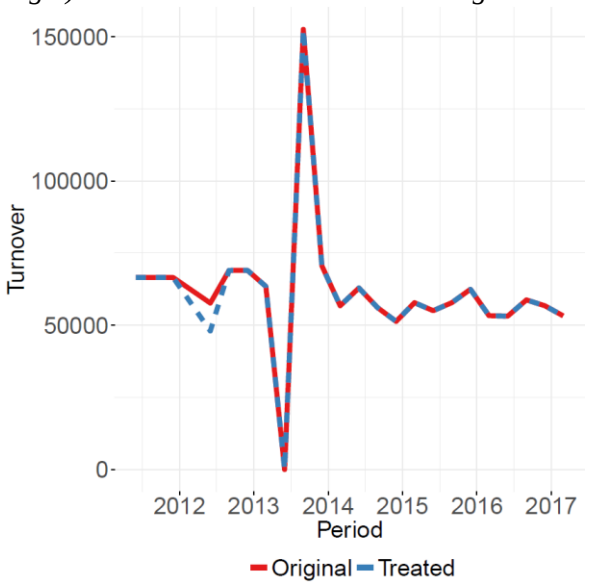
22. The quarter ending June 2013 in Example 3 of Figure 2 is the start of a level shift. Cases like this initially look like errors until future returns are supplied. In this case, the point would be treated initially but when the next period of data has been returned this would need to be revised as otherwise the whole time series after the level shift is likely to be cleaned out and this could have an impact on the trends seen at the cell level.

23. The quarter ending December 2014 in Example 4 of Figure 2 illustrates a large peak before reducing to zero or not returning. This could be where a business is selling off all its goods before it dies. Such values were not treated, and this did not lead to an impact on the cell level estimates. These cases would look like errors when they were initially returned and therefore would likely be treated. With future knowledge these cases would be identifiable and therefore revisions would be required for the next period or two.

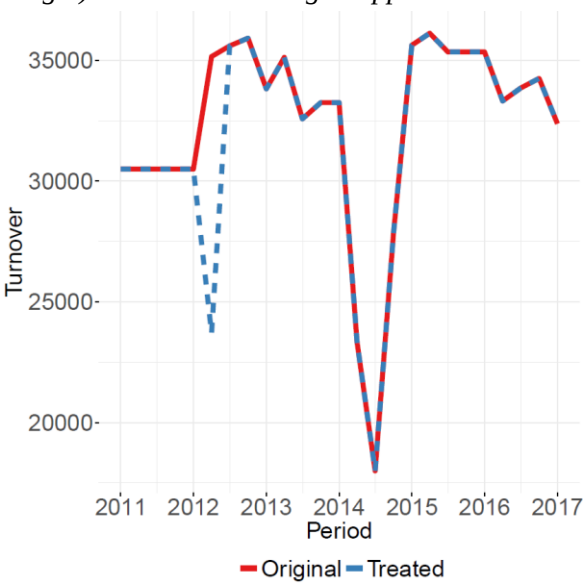
24. Example 5 of Figure 2 illustrates a case where having future knowledge suggests the original value, given in the quarter ending March 2014, may be true for a reason other than those outlined in the other examples. These cases would originally look like errors and be treated but then should be revised after future returns.

25. These examples demonstrate that it is not appropriate to automatically edit RUs identified as suspicious by selective editing. There is a requirement for manual validation of these cases. Additionally, as a result of cases such as those given by Examples 3-5 in Figure 2, initial decisions for treatment could potentially be revised as future data arrive.

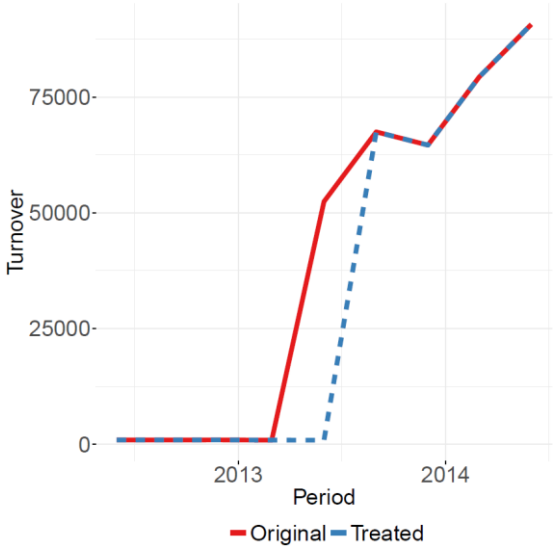
E.g.1) Treatment more extreme than original



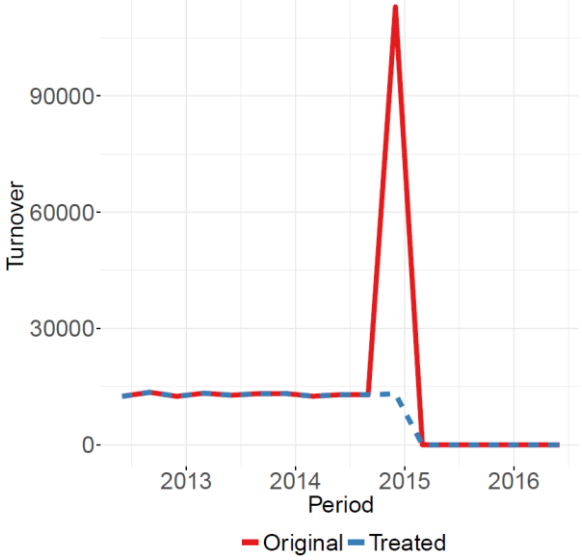
E.g.2) Treatment working in opposite direction



E.g.3) Level shift



E.g.4) Large peak before reducing to zero



E.g.5) Future suggesting original value may be true

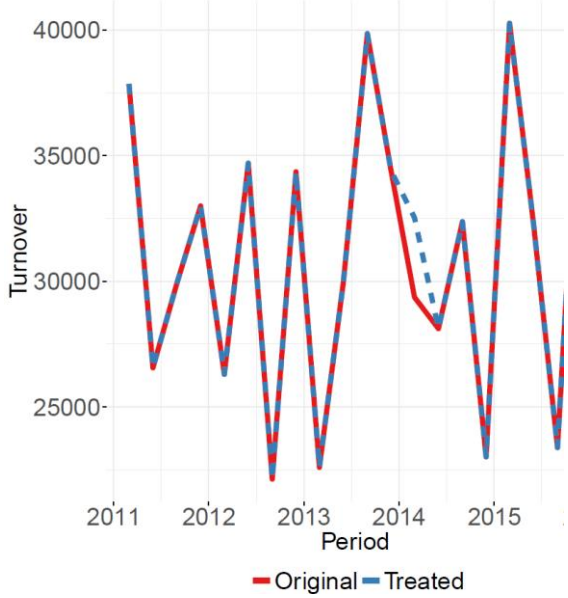


Figure 2: Examples top 5 RUs that require further validation before decision on treatment

III. Outlier Detection

26. The seas function in R has been used to detect outliers in cell level time series. The seas function identifies a seasonal ARIMA model and uses regressors to identify significant outliers (Sax, C., 2017). The regression model can then be used to estimate the size of the outlier. This error detection method is a macro editing method and will identify both outliers that are caused by errors in the data and genuine outliers.

27. The seas function was applied to staggers 00-03 (monthly and quarterly staggers) separately. This data had not been calendarised and did not include estimation. This was because calendarisation was found to cause outliers in the quarterly data to be spread across the months in the quarter. These were then picked up as level shifts instead of outliers, or not detected at all. This presented a problem initially as RUs changing among staggers were detected as outliers. To reduce the impact of no estimation, the following approach was used:

(a) For each stagger, separately:
- For each time point, calculate a ratio of total turnover and register turnover for the units that had responded in the VAT stagger
- Use automatic Outlier Detection on the time series of ratios. Where outliers are detected, replace the ratio with its expected, modelled value. This becomes the treated ratio series.
- Multiply each point in treated ratio time series by the register turnover total for that period.
- Convert the quarterly series to monthly by dividing by 3

(b) Sum treated, calendarised series across staggers to allow for comparison plots to be created to assess the impact of the method.

28. Initially, to give an idea of where the Outlier Detection was picking up errors that the Selective Editing method was not, automatic corrections using the coefficients provided by the seas function were applied. This removes all outliers whether they are caused by errors or are genuine movements. In practice, the purpose of cleaning is to remove errors in the micro data. As such, any errors detected at the macro level would be treated at the micro level so as not to remove genuine outliers.

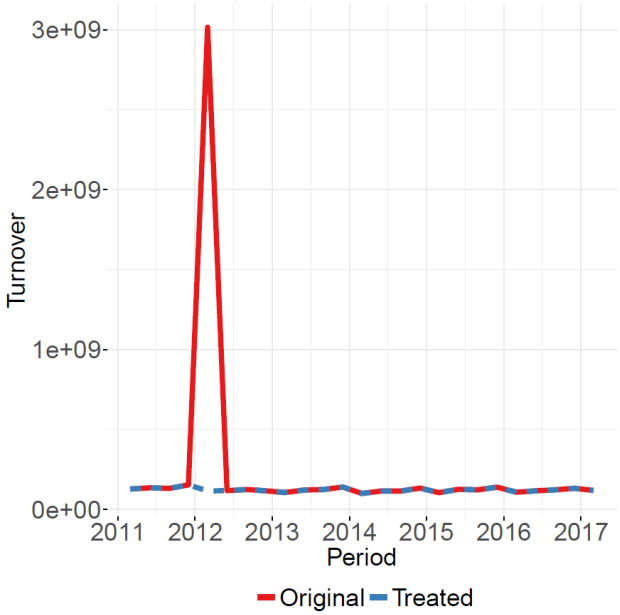


Figure 3: Cell level estimate of turnover and treated value under automatic outlier correction

29. There is a large peak in the quarter ending in March 2012 of Figure 3. This has been identified as an outlier which is highly likely to be an error. Automatic treatment using coefficients provided by the results of the seas function reduce the value to be more consistent with the rest of the time series.

IV. Comparison of Selective Editing and Outlier Detection Methods

30. The performance of the Selective Editing and Outlier Detection methods was assessed by visually comparing cell level estimates produced using the two cleaning methods. It was found that there were some cases where one or the other method worked better to detect the particular error.

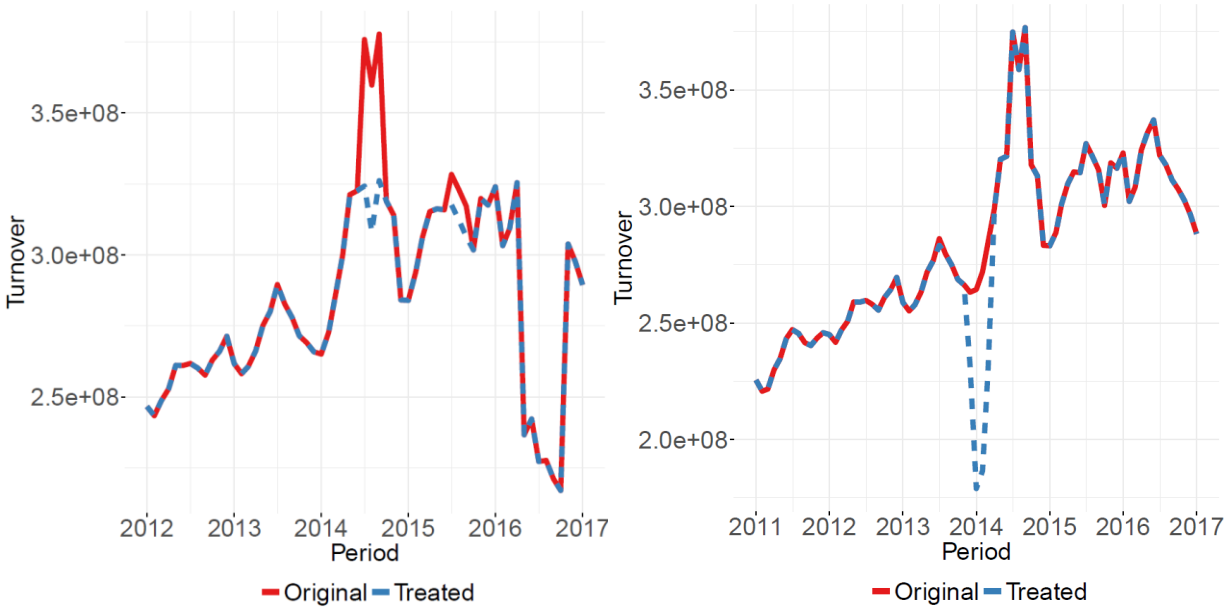


Figure 4: Selective Editing (left) picking up error not picked up by Outlier Detection (right) for a cell

31. In Figure 4 Selective Editing (left) has changed the double peak in 2014 while Outlier Detection with automatic treatment (right) has not detected. What is left after Outlier Detection still looks suspicious. In this case Selective Editing has cleaned more errors than Outlier Detection. This is not always the case.

32. In the Outlier Detection plot (right) from Figure 4 a trough has been produced by the automatic treatment method. This looks like it is creating an outlier but could have been produced from an outlying ratio. The automatic treatment method was only used for initial comparisons while a final treatment is discussed in Section V.

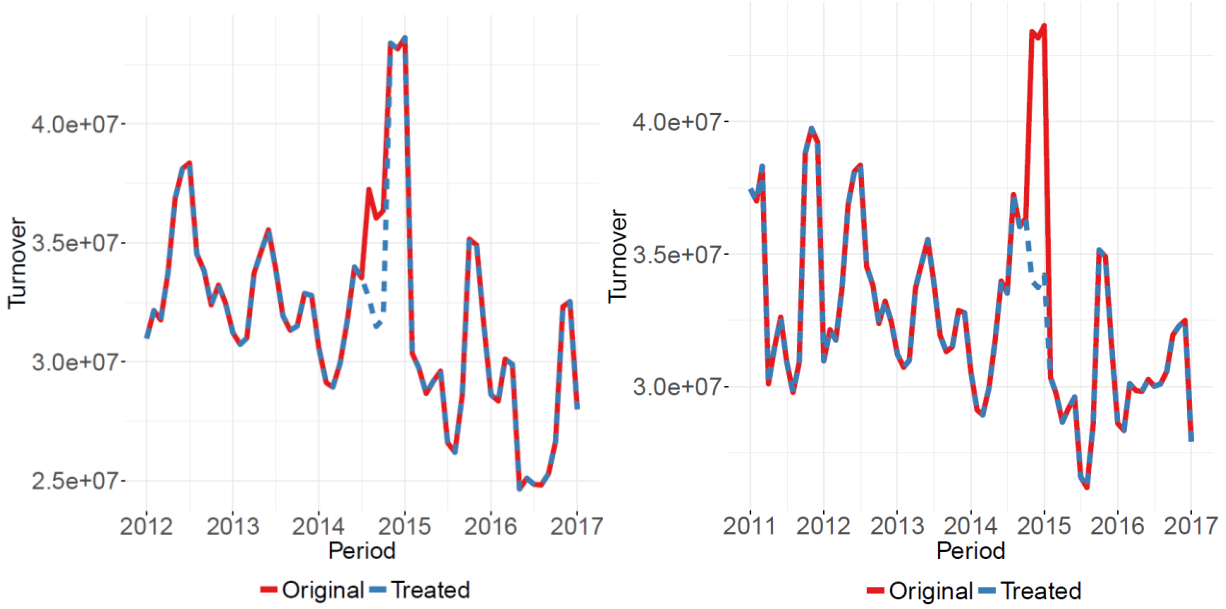


Figure 5: Outlier Detection (right) picking up error not picked up by Selective Editing (left) for a cell

33. In 2015 of Figure 5 there is a prolonged peak that is identified by Outlier Detection with automatic treatment (right) while this has not been picked up through Selective Editing (left). In this case the Outlier Detection method has cleaned more errors than Selective Editing.

34. As both methods pick up some different errors further analysis was conducted into how Outlier Detection and Selective Editing could be applied together.

V. Outlier Detection of Selective Edited Data

35. As seen in Section VI neither method discussed so far picks up all the errors in the data. Therefore, these methods have been tested in combination to determine if all errors can then be picked up.

36. The data, which were treated following Selective Editing, were aggregated to cell level for each stagger and run through the Outlier Detection methodology. The following process was then used to determine the cases to treat:

(a) For each cell, if an outlier was detected it was flagged indicating the period and stagger from which it originated.

(b) Referring to RU level, all the RUs within the cell are flagged accordingly.

(c) A Selective Editing score is recalculated, using score function given in Subsection II.A, accounting for the data having been treated after the previous round of Selective Editing for the period and stagger of interest.

(d) The case with the highest score is then treated using Ratio of Means imputation within the cell, period and stagger of interest.

(e) Plots were then produced to determine the performance of the treatment method.

37. The peaks in the original series in Figure 6 are clear outliers. The automatic treatment method does not always produce sensible cleaned values as seen towards the end of the series but returning to the RUs and treating the case with the highest Selective Editing score produces sensible cleaned values with the relating line following the trend for the cell.

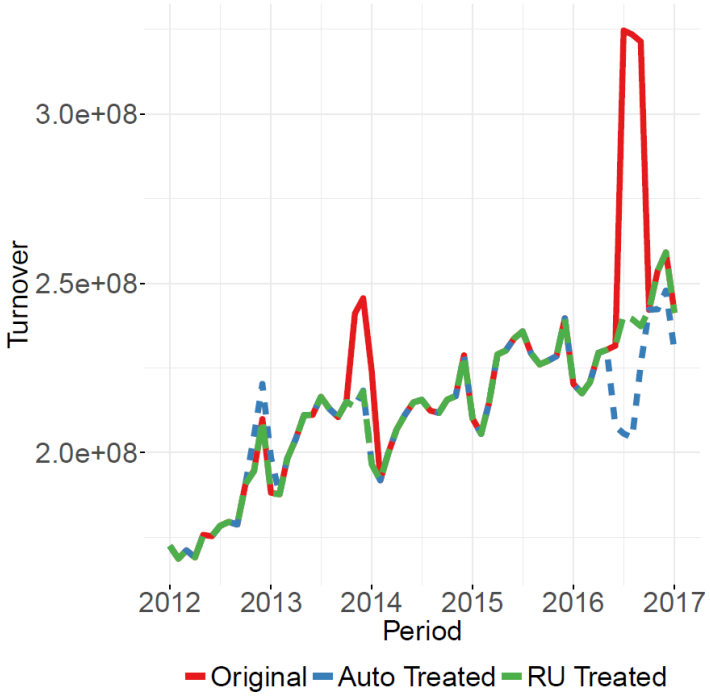


Figure 6: Cell level example of further outliers being detected on Selective Edited data

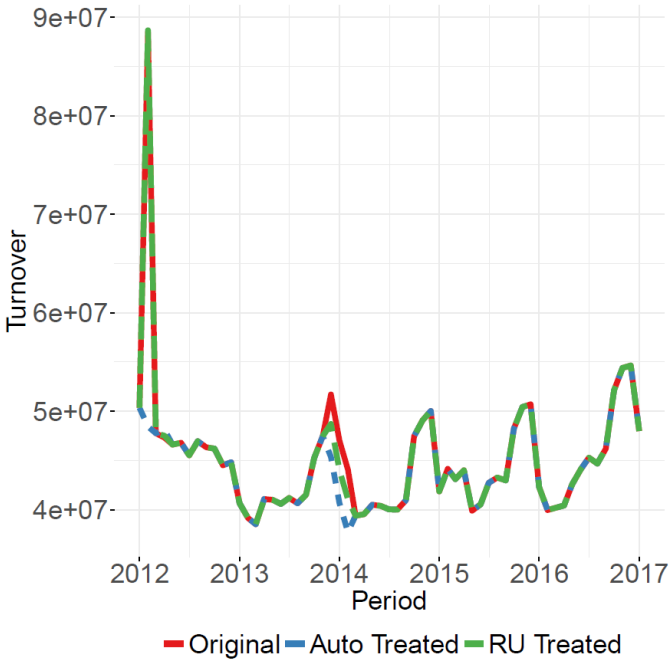


Figure 7: Cell level example where further investigation is required

38. Figure 7 identifies an outlier at the start of 2012 that was picked up by the Outlier Detection method. The RU causing this outlier was not identified as having the top-ranking score and therefore not treated. Further investigation is needed for cases like this as these could be genuine outliers. For example, all the businesses within the cell experiencing a sudden increase in turnover followed by a sudden decrease in turnover, or a business that has quite volatile which has dominated the change in the cell. In this example, it was found that the RU contributing to this outlier was an antiques dealer with very volatile turnover. This was not identified through selective editing as there was not enough previous data for the RU to calculate the score required.

39. Estimates for a period are required before the VAT data for that period has been returned. As a result, the VAT data will be forecast initially to produce these estimates. To allow for this there is a greater importance on the trend of the data so cases such as those given in Figure 7 would need to be treated.

40. For manual changes, the logic behind them should be recorded so that they may be reviewed to determine if there are rules that can be applied to account for them. For example, some of the outliers detected as described as given in Figure 7 can be attributed to one business. The business can be identified quite clearly by calculating the ratio of the current turnover of all RUs in the cell and the cell total for the previous year. But the ratios will need to be considered manually and treatment is likely to be manual until enough cases are acquired to allow a set threshold to be determined.

VI. Further Cleaning

41. Some manual input is required to consider the cases that have been detected using the Outlier Detection method that are not subsequently accounted for by further Selective Editing as described in Section V.

42. The peak in 2012 of Figure 8 provides an example where an error was not picked up through Selective Editing or Outlier Detection. This suggests that there may be a need for further manual intervention to look at the plots to determine if there are any other cases that should be treated.

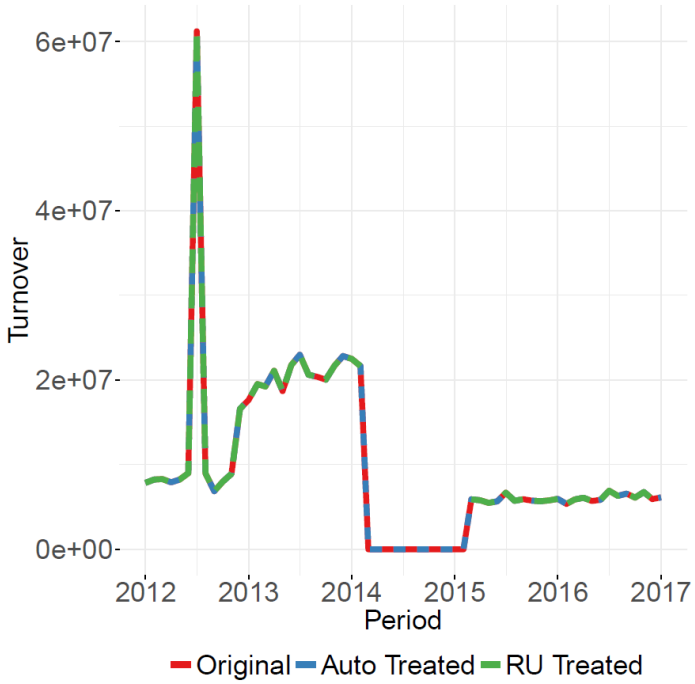


Figure 8: Example of cell where error not picked up by any method so far

43. As stated in Section V, manual changes and logic behind them should be recorded so that they may be reviewed to determine if there is a rule that can be applied to account for them.

VII. Revisions

44. Tests on the treatment so far have been done where future knowledge is available. This will not be the case in the production system with newly returned data. Returns provided to HMRC can be altered or returned up to three years after the reference period. As a result, the data in a rolling three-year window changes. Therefore, with each new period the data for the three years before and the period itself will be cleaned again. This may cause revisions in what has been treated as new data may result in different rankings for each RU. It will also allow the cases described in Figure 2 of Subsection II.B to be identified and returned to original values accordingly. Again, these decisions will need to be made manually.

45. Recording the manual changes and logic behind them may help derive rules that could reduce the need for manual intervention in the future.

VIII. Conclusions and Recommendations

46. In conclusion the following process is recommended:

(a) Calculate a Selective Editing scores as specified in Subsection II.A on the pre-calendarised RU level data for each quarterly stagger separately.

(b) For the 5 highest scores within each of the quarterly staggers, treat the RU using Ratio of Means imputation and produce plots to see how the method has worked.

(c) From the plots determine if the treatment is required accounting for the changes that may be genuine such as through seasonality. Use this to produce final ‘Selective Edited’ data.

(d) Aggregate the ‘Selective Edited’ data to cell level and run through the Outlier Detection method to identify outliers for the monthly and quarterly staggers.

(e) Identify RUs that in the cells identified to have outliers.

(f) Recalculate Selective Editing score for each stagger including the monthly stagger to account for the data that has been cleaned in the previous round of Selective Editing.

(g) Use Ratio of Means imputation to treat the RU with the highest score in the stagger and period required for cells in which an outlier has been detected.

(h) Use cell level plots to determine where outliers were detected and treatment has not had a notable impact. Investigate these cases to determine if there is a particular business causing the outlier that could be treated manually. Record logic/decisions.
(i) Check final cell level plots to check if there are any other obvious errors that may occur and not be treated because of a lack of data/seasonal pattern. Investigate these manually recording logic/decisions.
(j) Trace changes at RU level back to VAT unit level.

47. This method will be run for a three-year rolling window at noted in Section VII causing some revisions to previous data and allowing cases such as those in Figure 2 of Subsection II.B to be identified. The data can then be run through the next steps of the VAT pipeline (calendarisation, estimation etc) before being incorporated with the survey data to produce estimates required for Distributive Trade. When more periods have been cleaned it may be possible to derive further rules from the manual changes that have been recorded. The logic and decisions should be reviewed regularly to reduce the need for manual intervention.

IX. Next Steps

(a) The methods will be presented to Methodological Assurance for Statistical Transformation (MAST) before final sign off.
(b) The work so far has only been applied to Division 47. This will need to be replicated for Divisions 45 and 46 to check that the method works for them in addition.
(c) Determine if further rules can be adopted to reduce time and resource requirements.

Annex – Score functions tested

In Section II it was noted that a variety of score functions were tested. They include:

$$score\ 1 = \frac{|x-\mu|}{\hat{t}_{d,t-1}}$$

$$score\ 2 = \frac{w|x-\mu|}{\sigma\hat{t}_{d,t-1}}$$

$$score\ 3 = \begin{cases} \frac{w|x+2\sigma-\mu|}{\hat{t}_{d,t-1}} & when\ \mu > x \\ \frac{w|x-2\sigma-\mu|}{\hat{t}_{d,t-1}} & when\ \mu < x \end{cases}$$

$$score\ 4 = \frac{w|x-\mu|}{\sigma}$$

Where

$$x = Turnover\ for\ current\ period$$
$$\mu = Mean\ turnover\ over\ the\ last\ 3\ years$$
$$w = \frac{Register\ turnover\ for\ population\ in\ current\ period}{Register\ turnover\ for\ sample\ in\ current\ period}$$
$$\sigma = Standard\ deviation\ of\ turnover\ over\ the\ last\ 3\ years$$
$$\hat{t}_{d,t-1} = Total\ turnover\ for\ domain\ d\ in\ previous\ period\ t-1$$

References

ons.gov.uk, (2016) 'VAT turnover, initial research analysis, UK: Jan 2014 to Mar 2016' [online] Available at: www.ons.gov.uk/economy/grossdomesticproductgdp/articles/vatturnoverinitialresearchanalysisuk/jan2014tomar2016

ec.europa.eu, (2008) ‘Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys’ [online] Available at: http://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf Section C.4.2

Sax, C (2017) ‘Package ‘seasonal’’ [online] Available at: https://cran.r-project.org/web/packages/seasonal/seasonal.pdf