I. Introduction

1. Recently there has been a push for National Statistics Institutes to make greater use of administrative data in the production of business statistics. This should reduce burden on businesses and could lead to increased efficiency. In many cases, it is not possible to completely replace surveys with administrative data, so a mixed-source approach is required instead. The Office for National Statistics (ONS) is developing a method for producing mixed-source estimates for the Monthly Business Survey (MBS), based on a combination of Value Added Tax (VAT) Turnover data and a reduced survey.

2. When the VAT data reach ONS they have generally not been subject to much cleaning. Any cleaning that does happen is likely to focus on the needs of the tax office rather than statistical uses. Therefore, to ensure the quality of the MBS estimates, VAT Turnover data should be cleaned to deal with suspected errors.

3. This paper discusses methods for detecting and correcting errors in VAT data that are to be used in the production of mixed-source short term Turnover estimates for the MBS. Section II outlines some pre-established principles of cleaning VAT data. Section III describes several methods for detecting suspicious VAT Turnover values, as well as summarising previous work on detecting and correcting suspicious patterns and unit errors in VAT Turnover data. Section IV contains the evaluation of each method for use in the MBS. Section V provides a summary of recommended methods to correct for suspicious VAT Turnover values. In section VI some practical implementation issues are considered, and the process of fine tuning of parameters for the recommended detection method for the MBS is illustrated.

II. Principles of cleaning VAT data

4. Seyb et al (2009) list seven key principles for editing administrative data. The first six of these provide useful advice for developing methods and implementing systems for detecting (and correcting) errors in VAT Turnover. Lewis and Lewis (2011) summarised the first six principles in the context of VAT Turnover data, and added a seventh for completeness. In the report, they recommend that these principles, shown below, are adhered to when considering methods for detecting errors in VAT Turnover.

(a) Maintain the original data supplied by the tax office whenever possible.
(b) Choose methods that take account of all statistical uses of the data – macro and micro level if appropriate.

(c) Make good use of historical and auxiliary data.

(d) Automate the process of detecting and correcting errors in VAT Turnover data as much as possible, to ensure best use of resources and consistent results.

(e) Keep an audit trail for any changes to VAT Turnover data – this should include keeping the original (unaltered) data and producing diagnostics for the detection and correction process.

(f) The implementation of automated detection and correction methods should be flexible for future improvements.

(g) Make every effort to understand how the VAT Turnover data are returned and processed as this is a useful indicator of potential sources of error. Good dialogue with the tax office is essential for this.

III. Detecting suspicious VAT Turnover

5. When using VAT Turnover for statistical purposes it is important to ensure that the reported figures returned look reasonable. This includes identifying suspiciously small or large Turnover values as well as suspicious patterns and unit errors in the VAT data. Methods for detecting suspicious Turnover patterns and methods for correcting errors have been investigated in a previous study; findings are summarised later in this section.

A. Methods for detecting suspicious VAT Turnover Values

6. Seven methods are described below for identifying suspicious VAT Turnover values. The results of testing these methods with VAT Turnover data can be found in section V. In some cases, the methods concentrate on unusually large Turnover values, since these are the ones that have the potential to cause the biggest errors in statistical outputs. Other methods detect both unusually large and unusually small Turnover values.

(a) Method 1: Quartile distances in industry Turnover

7. Method 1 is based on a method described in Hoogland and Van Haren (2007) to identify unusually large or small Turnover by locating extreme values in the distribution of VAT Turnover within a particular industry and size class. The method identifies suspicious businesses as those with Turnover greater than a specified multiple of inter-quartile distances from the third quartile, Q3 (for large Turnover) or less than the same multiple of inter-quartile distances from the first quartile, Q1 (for small Turnover). More formally, suspicious Turnover is identified as follows.

If \( \text{VAT Turnover} > Q3 + [C \times (Q3 – \text{Median})] \)

or \( \text{VAT Turnover} < Q1 – [C \times (\text{Median} – Q1)] \)

then VAT Turnover is suspicious.

8. The quartiles and median are derived from the VAT Turnover data within the industry and size class for the period under consideration, and the parameter C is derived from analysing past data to find a threshold that successfully identifies extreme values. C may be given different values for different industry and size classes, dependent on the distribution of data in those classes.
(b) Method 2: Period on period ratios

9. Method 2 comes from De Jong (2003) and involves calculating period on period ratios for each business based on the contribution that business’s Turnover makes to its class. As in method 1, the class would commonly be defined as a cross-classification of industry and size. The period on period ratios are called test ratios. Any business with a test ratio above a pre-defined threshold is judged as suspicious. It is then recommended to check the data to identify which of the periods’ data looks most likely to be in error. The method is described below:

\begin{enumerate}
\item Calculate \( \text{Score} = \frac{\text{VAT Turnover}}{\text{Median VAT Turnover in class}} \) for each business.
\item Then calculate \( \text{TestRatio} = \begin{cases} 
\frac{\text{Score}_t}{\text{Score}_{t-1}} & \text{if } \text{Score}_t > \text{Score}_{t-1} \\
\frac{\text{Score}_{t-1}}{\text{Score}_t} & \text{otherwise.}
\end{cases} \)
\end{enumerate}

Where \( \text{Score}_t \) is the value of Score in period \( t \) and \( \text{Score}_{t-1} \) is the value of Score in period \( t-1 \).

10. Any business with a value of TestRatio greater than a pre-defined threshold is deemed to be suspicious. The threshold can be set by analysing past data to see at what point the values of TestRatio appear to indicate suspicious returns. De Jong (2003) describes using a threshold of 40 when using VAT data for monthly statistics. Note that the method identifies Turnover values that are both suspiciously large and suspiciously small, since the TestRatio can be large for either of these reasons.

(c) Methods 3 and 4: Quartile distances combined with a measure of influence

11. Method 3 is a refinement to method 1 and is inspired by Hoogland et al (2009). The idea is to combine the detection of suspicious values using quartile differences with some measure of the influence or importance of the business. In this method we calculate the influence as the proportion of VAT Turnover the business contributes to the total VAT Turnover in the industry and size class.

12. The method is as follows:

\begin{enumerate}
\item Identify unusual Turnover values using the quartile distances measure described in method 1.
\item Calculate \( \text{Influence} = \frac{\text{VAT Turnover}}{\text{Total VAT Turnover in class}} \) for each business.
\item Treat any business which has both unusual Turnover values from the quartile distance method and whose Influence is greater than a pre-defined threshold as suspicious.
\end{enumerate}

13. The threshold for influential businesses should be set by analysing past VAT Turnover data. Note that this method effectively subsets businesses failing the quartile distance method, so that only the most influential are viewed as being suspicious. One practical way to implement method 3 is to first set reasonable thresholds for method 1, ensuring that the businesses being identified as suspicious appear to be genuine extremes in the distribution. Then, for method 3, slightly reduce the value of parameter C in the quartile distance method so that more businesses are viewed as being extreme. The threshold for influential businesses can then be set to ensure that a similar number of businesses fail method 3 as failed method 1.

14. The suspicious businesses in method 3 are determined as a combination of their relationship to the distribution of VAT Turnover in the class and their influence in that class. Method 1 does not take account of the influence.

15. Method 4 is identical to method 3, but uses a higher value of C to check the sensitivity of the method.
Method 5: Hidiroglou-Berthelot method

16. Method 5 is the Hidiroglou-Berthelot method, which is based on the simple ratio between the VAT Turnover value for a business in the current period and the VAT Turnover value for the same business in the previous period. The method adds some refinements to overcome various drawbacks of using a simple ratio and also to allow for the size of the business to be taken into account. The method comes from Hidiroglou and Berthelot (1986).

17. The method is defined as follows.

(1) Calculate the ratio \( r = \frac{\text{current VAT Turnover}}{\text{previous VAT Turnover}} \) for each business.

(2) Transform the ratio so that the data are symmetric and centred around zero (this gives an equal chance of identifying large and small errors):

Calculate the median of the ratios, \( r \).

If \( r < \text{median} \) then calculate \( t = \frac{r - \text{median}}{r} \)

Otherwise calculate \( t = \frac{r - \text{median}}{\text{median}} \)

(3) Define \( E = t \times \max(\text{current VAT Turnover, previous VAT Turnover})^V \)

This step gives the option of giving greater importance to businesses with larger VAT Turnover. The parameter \( V \) can take any value between 0 and 1. A value of 0 results in every business having the same importance, whereas a value of 1 gives greatest importance to businesses with higher Turnover. When setting a value for \( V \), it is recommended to consider whether errors in businesses with larger Turnover are more important. The final choice of \( V \) should be chosen after analysing its effect on past data.

(4) Calculate the first and third quartiles (Q1 and Q3) and the median (Q2) of the E values calculated in step 3. Then calculate

\[
\begin{align*}
    d_{q1} &= \max \left( Q2 - Q1, |A \times Q2| \right) \\
    d_{q3} &= \max \left( |Q3 - Q2|, A \times Q2 \right)
\end{align*}
\]

In most cases, these distances are just the difference between the median and the first and third quartile of the E values respectively. However, in some unusual distributions it is possible for the data to group around the median leading to businesses being erroneously identified as suspicious. To safeguard against this happening, a small multiple (A) of the median of the E values is used instead. The parameter A is commonly given the value 0.05 for this purpose.

(5) Suspicious businesses are then identified as follows:

If \( E < Q2 - C \times d_{q1} \) Or \( E > Q2 + C \times d_{q3} \) then treat as suspicious.

18. The parameter \( C \) effectively controls the number of businesses being viewed as suspicious, therefore past data should be analysed to determine the value of \( C \) that appears to accurately identify suspicious Turnover.
(e) Methods 6 and 7: Period on period ratios with a measure of influence

19. Unusually large or small businesses are identified using method 2 (period on period ratios). The most influential of those businesses are deemed suspicious.

1. For both current and previous periods calculate the score, and then calculate the TestRatio, as described in steps (1) and (2) of method 2.

2. Then, for each business calculate \( \text{Influence} = \frac{\text{VAT Turnover}}{\text{Total VAT Turnover in class}} \).

20. Any business with a TestRatio greater than a pre-specified threshold and Influence > \( K \) is deemed suspicious.

21. Method 7 is identical to method 6, but uses a higher threshold for TestRatio to check the sensitivity of the method.

B. Detecting suspicious patterns and unit errors in VAT Turnover

22. It is often possible to identify errors in VAT Turnover data by considering the pattern of reported Turnover over a period. Hoogland (2011) describes a variety of suspicious patterns identified in quarterly VAT Turnover. For example, when businesses return positive Turnover for some quarters, a zero response suggests that the true data may have been withheld. Similarly, when the same Turnover is reported for four quarters, it suggests that the business may have simply divided their annual figure into four. The same logic may be applied to monthly VAT Turnover.

23. Hoogland (2011) notes that the presence of these patterns does not always imply that data are suspicious, and for some seasonal businesses it may be reasonable to have zero Turnover in some months but not others. However, Lewis and Lewis (2011) recommend that these patterns are marked as suspicious before implementing any other error detection method.

24. It may be possible to deal effectively and relatively easily with some errors in VAT Turnover, if they are due to a systematic cause. One common systematic error in the reporting of financial data is a unit error. For example, Tax offices often require VAT Turnover in pounds (GBP), but businesses used to reporting in thousands of pounds for other purposes sometimes report in thousands of pounds for VAT Turnover. This leads to a large error for the business, but it is relatively straightforward to identify and correct that error as long as something is known about the expected Turnover for that business. Lewis and Lewis (2011) recommend methods for identifying and correcting such unit errors.

IV. Evaluation

A. Evaluation of detection methods for the MBS

25. A key difference between survey and administrative data is that with administrative data it is often not possible to re-contact the business and ask them to confirm any suspicious values. One implication of this is that the only option for adjusting suspicious values is to remove the original value and impute a figure which is more in line with expectations. Because of this, it is generally not possible to discover the truth about suspicious VAT Turnover. This means that the evaluation of detection methods is not straightforward and cannot usually be definitive.

26. If a business is detected as an error and has VAT Turnover similar to the survey Turnover, then this can be viewed as being a likely ‘false hit’. It is possible to get a clearer indication of false hits for
MBS by comparing quarterly VAT Turnover (as around 90% of businesses report VAT quarterly) with quarterly Turnover figures derived by summing MBS returns from the appropriate months.

27. The following section describes the results of testing these methods with VAT Turnover data available in the ONS. Each method was set to fail the same proportion of businesses, 4%, to allow for comparison between methods. It is generally not possible to discover the truth about any suspicious VAT Turnover returns and therefore evaluation is difficult. To give an indication of the behaviour of the methods, the average VAT Turnover and average business register employment for suspicious and non-suspicious businesses are calculated. Following this, a range of estimated false hit measure are calculated, based on comparison with quarterly MBS Turnover in the services industries.

B. Results of the evaluation for the MBS

28. Table 1 shows the average VAT Turnover for businesses identified as suspicious, and those not identified as suspicious by each method tested.

Table 1. Mean VAT Turnover (£000) for suspicious and non-suspicious businesses

<table>
<thead>
<tr>
<th>Method</th>
<th>Suspicious</th>
<th>Non-suspicious</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>3,216</td>
<td>162</td>
</tr>
<tr>
<td>Method 2</td>
<td>495</td>
<td>275</td>
</tr>
<tr>
<td>Method 3</td>
<td>3,557</td>
<td>148</td>
</tr>
<tr>
<td>Method 4</td>
<td>3,434</td>
<td>153</td>
</tr>
<tr>
<td>Method 5</td>
<td>2,070</td>
<td>210</td>
</tr>
<tr>
<td>Method 6</td>
<td>635</td>
<td>269</td>
</tr>
<tr>
<td>Method 7</td>
<td>568</td>
<td>272</td>
</tr>
</tbody>
</table>

29. Note that mean VAT Turnover (£000) in the whole raw data set is 284. The results in Table 1 show that the methods based on using quartile distances to identify extremes in the distribution of current period Turnover (methods 1, 3 and 4) tend to identify the most extreme businesses. This is a natural consequence of the method, and consistent with results in Lewis and Lewis (2011). The methods using period on period ratios identify much less extreme businesses on average.

30. Table 2 shows the average register employment for businesses identified as suspicious, and those not identified as suspicious by each method tested.

Table 2. Mean Business Register Employment for suspicious and non-suspicious businesses

<table>
<thead>
<tr>
<th>Method</th>
<th>Suspicious</th>
<th>Non-suspicious</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>121</td>
<td>88</td>
</tr>
<tr>
<td>Method 2</td>
<td>282</td>
<td>81</td>
</tr>
<tr>
<td>Method 3</td>
<td>180</td>
<td>86</td>
</tr>
<tr>
<td>Method 4</td>
<td>176</td>
<td>86</td>
</tr>
<tr>
<td>Method 5</td>
<td>261</td>
<td>82</td>
</tr>
<tr>
<td>Method 6</td>
<td>308</td>
<td>80</td>
</tr>
<tr>
<td>Method 7</td>
<td>302</td>
<td>81</td>
</tr>
</tbody>
</table>

31. Note that mean Business Register Employment in the whole raw data set is 89. The analysis shows that the methods based on period on period ratios identify businesses with larger than average register employment than the other methods. This is the opposite of what was seen with VAT Turnover; however the differences between the methods are not as large.

32. The estimated false hits in Table 3, below, are based on a reduced data set of 19,517 businesses, for which both survey and VAT Turnover data are available. The table shows the percentage of failures in that data set that are within 10% or 20% of the survey Turnover. These are described as ‘False hit 10%’ and ‘False hit 20%’ respectively. The percentage of failures in this reduced data set is also shown.
Table 3. Estimated false hits based on closeness to survey returns and proportion of failures in matched survey and VAT data set (all expressed as percentages)

<table>
<thead>
<tr>
<th>Method</th>
<th>False hit 10%</th>
<th>False hit 20%</th>
<th>Failures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>30</td>
<td>45</td>
<td>6.7</td>
</tr>
<tr>
<td>Method 2</td>
<td>17</td>
<td>24</td>
<td>3.2</td>
</tr>
<tr>
<td>Method 3</td>
<td>32</td>
<td>47</td>
<td>8.9</td>
</tr>
<tr>
<td>Method 4</td>
<td>31</td>
<td>46</td>
<td>8.0</td>
</tr>
<tr>
<td>Method 5</td>
<td>23</td>
<td>33</td>
<td>5.2</td>
</tr>
<tr>
<td>Method 6</td>
<td>20</td>
<td>29</td>
<td>4.1</td>
</tr>
<tr>
<td>Method 7</td>
<td>19</td>
<td>28</td>
<td>3.5</td>
</tr>
</tbody>
</table>

33. The results in Table 3 show that there are less false hits for the period on period ratio methods (2, 6 and 7) than for any other method. In particular, the period on period ratio methods perform much better than the quartile distance methods (1, 3 and 4) in this respect. This suggests that some of the businesses with extreme VAT Turnover identified as errors by the quartile distance methods are merely large businesses reporting their correct Turnover. Comparing with the previous period allows us to identify more accurately whether the business has reported in a different way to what would be expected.

34. To put the impact of the estimated false hits into context, Table 4 shows how many there are as a proportion of the whole matched data set.

Table 4. False hits as a proportion of whole (survey-matched) data set (%)

<table>
<thead>
<tr>
<th>Method</th>
<th>False hit 10%</th>
<th>False hit 20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1</td>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Method 2</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Method 3</td>
<td>2.8</td>
<td>4.1</td>
</tr>
<tr>
<td>Method 4</td>
<td>2.5</td>
<td>3.7</td>
</tr>
<tr>
<td>Method 5</td>
<td>1.2</td>
<td>1.7</td>
</tr>
<tr>
<td>Method 6</td>
<td>0.8</td>
<td>1.2</td>
</tr>
<tr>
<td>Method 7</td>
<td>0.7</td>
<td>1.0</td>
</tr>
</tbody>
</table>

35. The estimated false hits for the period on period ratio method make up 0.5% of the matched data set. Imputing an alternative value for these false hits will introduce some error. However, this should be far outweighed by the benefit of correcting actual errors. The imputed values will all be similar to the non-suspicious values and the overall result will be to produce a smoother data set.

36. The results in Tables 1 to 4 suggest that the period on period ratio method (method 2) is the most suitable for identifying suspiciously small or large VAT Turnover values. The Hidiroglou-Berthelot method (method 5) also performs well, but slightly higher false hits and the fact that it is more complicated to implement mean that the period on period ratio method should be preferred. A suitable threshold needs to be derived for this method, and this is discussed for our application to the MBS in section VI.

V. Methods for correcting suspicious values in VAT Turnover

37. Once an error or suspicious value has been identified in VAT Turnover data, there are a number of options for dealing with it. This includes marking suspicious businesses, removing them from the data set or replacing their value with an imputed value. For ONS, like many National Statistics Institutes, legal restrictions prohibit the ability to contact suspicious businesses to confirm their values; however, there are some statistics offices where this is a possibility.

38. The choice of method will depend on the use of the data, the resource available as well as consideration of legal restrictions and burden on business. However, where it is required to produce a
value for a suspicious business it is anticipated that the most common treatment will be automatic imputation.

39. Lewis and Lewis (2011) propose and test a range of imputation methods for two scenarios: a suspicious VAT Turnover value in a single period where it is necessary to impute a single value based on available information; and a suspicious pattern of quarterly VAT Turnover, indicating that the business was only able to provide reliable annual Turnover, where it is necessary to apportion the annual Turnover into quarters.

40. Full details of the methods tested are given in Lewis and Lewis (2011), but we summarise the recommendations here. When imputing a VAT Turnover value in a single period, ratio imputation was found to be the most accurate method, in particular the ratio of means method using data from the previous period. The method involves using a previous value for the business and multiplying it by a ratio based on the growth between the previous and current period for similar businesses. To apportion annual Turnover into quarterly Turnover, the most accurate method was taking the median quarterly proportions from similar businesses and multiplying them by the annual figure.

VI. Implementation in ONS

A. Fine tuning of threshold for period on period ratios method

41. Now that we have chosen the period on period ratio method (method 2), we need to decide a threshold based on detecting suspicious VAT Turnover values. Recall from section III, this method involves calculating period on period ratios for each business, based on the contribution that business’s VAT Turnover makes to its class. The graphs below show the distribution of these ratios for a quarterly return of VAT data, across three consecutive periods (to check the robustness of the outcome). In all of the graphs, a small number of extreme observations have been removed in order to focus on the main body of the distributions.

*Figure 1. Distribution of ratios for quarterly stagger 1 in June 2008*
42. The distribution is fairly similar across multiple quarters. A threshold of 5 for the ratio looks appropriate and gives 18,013 failures, which is 5.1% of the data set. There are an estimated 0.8% of businesses flagged as errors using this threshold that are within 10% of matched survey returns. This compares with 0.6% of businesses when failing 4% of the data set.

43. A similar exercise was completed for monthly and annual VAT returns, and a threshold of 7 and 4 are recommended respectively.

B. Other considerations for implementation

44. The next stage for implementation is to specify the recommended methods for the ONS IT system, and define the edit rules for system developers. It has been agreed that the cleaned VAT data will be held as a new variable on the ONS Inter-Departmental Business Register, as well as the original value provided by the tax office. Additional markers will also be created, to allow identification of suspicious values in VAT Turnover. Markers will also be used to flag suspicious patterns in the VAT Turnover returned, and to indicate where a unit error has been automatically corrected.

45. Further work is needed to fully understand the current process of updating the business register with VAT Turnover data, and how that may be impacted by introducing the editing method. VAT Turnover data has to be apportioned on arrival at ONS, to enable a match to the statistical units held on the Inter-Departmental Business Register, in order for it to be more usable by ONS. The current apportionment method is being reviewed to consider the use of VAT Turnover for mixed-source estimates. VAT Turnover returns of zero are also used to identify deaths of businesses. The affect of editing suspicious values and patterns in the VAT Turnover on these zero returns, and consequently on how the ONS determine businesses that have died, needs more research.
46. When used in practice, we will need to review the editing strategy for the reduced sample of MBS. Selective editing is currently used for the MBS; see Hooper and Lewis (2010) for detailed information on the methodology. When the MBS changes to a mixed-source approach, the method will need to be re-evaluated, and thresholds updated to suit the new estimates.

VII. Conclusion

47. In common with many statistical offices, the ONS has been investigating methods for making greater use of administrative data in the production of business statistics. This paper has described and evaluated the methods for detecting and correcting errors in VAT Turnover data to be used in the production of mixed-source short term Turnover estimates. The practicalities of implementing the recommended method have also been considered.

48. This paper has also described suspicious patterns and unit errors that are commonly observed in VAT Turnover data. These are easily identified and generally imply that the business has made a reporting error. It is recommended that VAT data are checked for these potential errors before implementing any other error detection method. Methods for detecting suspicious Turnover patterns and correcting unit errors are not likely to be affected by the specific output; detection and correction methods recommended in Lewis and Lewis (2011) should be applicable to other surveys, and other National Statistics Institutes.

49. Based on the evaluation presented, it is recommended to use the period on period ratio method to detect suspiciously small and large values in VAT Turnover, for the use in mixed-source estimates of the MBS. Any business with a period on period ratio (test ratio) above a pre-defined threshold is judged as suspicious. The thresholds should be reviewed within a year of implementation to check they are working appropriately. The effective values of these thresholds are likely to differ between data sources, and it is recommended to test the performance of methods outlined in section III on any data source itself before implementation. The choice in methods to detect and correct errors in administrative data, and the possible evaluation measures available, will ultimately depend on the data used to produce mixed-source statistics.

VIII. References


