Office for National Statistics

{79}

Survey Methodology Bulletin

May 2019

Contents

Office for National Statistics
Swyddfa Ystadegau Gwladol

The Survey Methodology Bulletin is primarily produced to inform staff in the Office for National Statistics (ONS) and the wider Government Statistical Service (GSS) about ONS survey methodology work. It is produced by ONS, and ONS staff are encouraged to write short articles about methodological projects or issues of general interest. Articles in the bulletin are not professionally refereed, as this would considerably increase the time and effort to produce the bulletin; they are working papers and should be viewed as such.

The bulletin is published twice a year and is available as a download only from the ONS website.

The mission of ONS is to improve understanding of life in the United Kingdom and enable informed decisions through trusted, relevant, and independent statistics and analysis. On 1 April 2008, under the legislative requirements of the 2007 Statistics and Registration Service Act, ONS became the executive office of the UK Statistics Authority. The Authority's objective is to promote and safeguard the production and publication of official statistics that serve the public good and, in doing so, will promote and safeguard (1) the quality of official statistics, (2) good practice in relation to official statistics, and (3) the comprehensiveness of official statistics. The National Statistician is the principal advisor on these matters.

www.ons.gov.uk

Edited by: Philip Lowthian

methodology@ons.gov.uk

The methodological challenges of protecting outputs from a Flexible Dissemination System

Stephanie Blanchard [1]

1. Introduction

There is increasing demand on National Statistical Institutions from users for more data to be made publicly available and released sooner after collection. There is also a greater desire by the Office for National Statistics (ONS) to make better use of the data that is held. A solution to satisfy both sides is to look at new and innovative ways to disseminate data. A project currently underway at ONS is the development of a Flexible Dissemination System (FDS) for the 2021 Census which has produced a 'proof of concept' prototype; the result of collaboration between the Statistical Disclosure Control (SDC) team in Methodology, the 2021 Census Outputs team, Digital Publishing and an external company. Further work will be to investigate the potential of using an FDS beyond the release of Census data.

The core aims of the project are to address user feedback for more flexible, accessible and timely outputs but these user requirements have to be balanced against the requirements for the ONS to protect the data. This has identified a number of methodological challenges to protecting data accessed through an FDS which are described, along with potential solutions, in this paper.

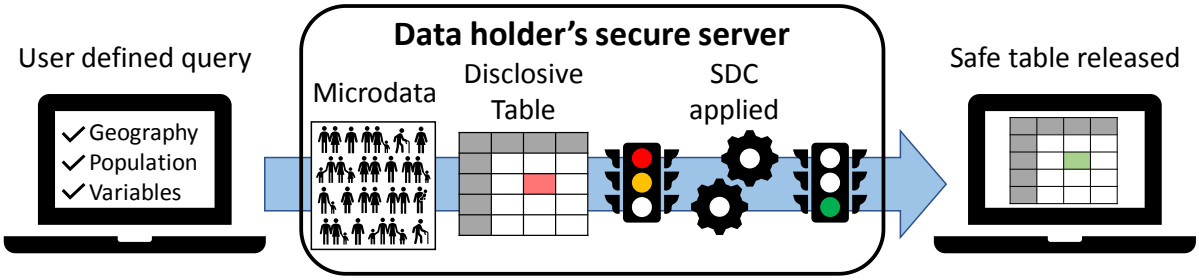
2. A Flexible Dissemination System for outputs

Traditionally, a data output comprises of a series of static tables which once released, cannot be changed. The aim is to design the tables to meet the majority of user needs while maintaining data confidentiality, a vital requirement for all data outputs as described in section 3. If users require an additional table to those published, in some cases a commissioned table service is provided by the data holder, usually with a charge to the customer, which can involve a lengthy negotiation process to agree a table that is acceptable to the user while meeting confidentiality requirements.

Rather than provides a series of tables, an FDS provides an online interface where users can define their own tables by building a query from a list of selections provided by the data holder which could include the level of geography, the table population and the variables. The table is built in real time from the unit record level microdata which is held securely and not accessible by the user. Confidentiality is maintained through the disclosure control methods applied to the microdata and algorithms applied to the table once built, before it is released to the user.

[1] Office for National Statistics stephanie.blanchard@ons.gov.uk

Figure 1: Flow of processes through a Flexible Dissemination System



2.1. Benefits of a Flexible Dissemination System

There are two main benefits of an FDS; flexibility and timeliness.

In an FDS, users can define their own tables based on their individual requirements and priorities. The level of flexibility will depend on the options made available through the FDS but it is likely to lead to more data being made available to the user, either through more detailed versions of tables that had previously been released or through combinations of variables that weren't previously made available. Offering more flexibility to users could reduce the demand on a commission table service as users will be able to get more through the FDS, but the service may still be required for tables of non-standard construction as technical capability may limit the functionality of the FDS.

Outputs will be available sooner through an FDS as time is saved by not having to design the tables, build and disclosure check them. The shorter lead time for producing outputs is achieved by applying automated SDC methods although there is an initial time and resource investment by the data holder to design the content of the FDS and define the parameters of the automated SDC methods. For data releases that comprise a large number of tables released in stages, time gains can also be made by applying automated SDC methods since all the data can be released together.

2.2. Application of a Flexible Dissemination System

The biggest benefits in flexibility and timeliness will be where there is a large number of tables generated from a single data source which has a wide user base with different priorities and requirements. An example of this is the UK Census. Following the 2011 Census, ONS published over 5,000 standard release tables for England and Wales based on 650 table templates covering a variety of geographies. There have also been a further 900 commission tables released and that number is still growing even seven years after census day. Census data are accessed by a large variety of users including Local Authorities, academics, charities, businesses and enquiring citizens.

Following Census day, it took around 16 months for ONS to publish the first outputs and a further 2 years and 8 months until the last outputs were published. The reason for this long time lapse was down to the table building process which included designing the tables based on users' requirements, manual checking for disclosures by the SDC team, table re-design when disclosures were identified and finally building the table for publication. User feedback after the 2011 Census was that they generally liked the SDC

methods used, since it allowed small counts to be produced unlike the 2001 Census methods, but they were disappointed that no significant gains had been made regarding flexibility and timeliness. Based on this feedback, work started in 2015 on the development of an FDS. This will provide the 2021 Census outputs with both flexibility and timeliness; the aim being to publish the first outputs within 12 months of census day and for all outputs to be released within 24 months.

National Records of Scotland (NRS) and the Northern Ireland Statistics and Research Agency (NISRA) are also developing FDSs for disseminating their 2021 Census outputs. ONS, NRS and NISRA are aiming for harmonisation in the dissemination approaches for the 2021 Census wherever possible.

3. What Statistical Disclosure Control is and why we need it

SDC is the application of methods to protect respondents in statistical outputs. A respondent could be an individual, a household, a business or any other statistical unit. A statistical output can take many forms including microdata datasets, frequency tables, magnitude tables, graphs and visualisations. The principle behind many SDC methods is to introduce sufficient uncertainty into the outputs such that a respondent cannot be identified or their characteristics revealed to an intruder, an intruder being someone who either purposefully tries to identify a respondent or someone who inadvertently stumbles upon a respondent through using the data or output. There are many examples in the literature on disclosure risk and methods, including Hundepool et al (2012).

Determining the most appropriate SDC method(s) for an output is/are based on a range of factors with the ultimate aim of maintaining a satisfactory relationship between disclosure risk and data utility. In order to make an output safe for release, the disclosive data must be changed in some way to protect the outputs. This will damage the utility of the data so it essential to select the disclosure control methods that will reduce the disclosure risk to an acceptable level while maximising the data utility in line with user requirements.

Application of disclosure control is required because ONS has legal obligations under the Statistics and Registration Service Act (2007)[2] and the General Data Protection Regulation (2018)[3]. If a breach occurred it could result in action being taken against the organisation as well as the individuals involved including fines and criminal proceedings. ONS also has ethical obligations through the UK Statistics Authority Code of Practice for Official Statistics (2018)[4] along with the pledge made to respondents for all data collections that confidentiality will be maintained. A breach of our ethical obligations could result in reputational damage for ONS and a loss of trust from the public leading to lower response rates across all our surveys which will adversely affect data quality.

[2] https://www.legislation.gov.uk/ukpga/2007/18/contents
[3] https://www.gov.uk/government/publications/guide-to-the-general-data-protection-regulation
[4] https://www.statisticsauthority.gov.uk/code-of-practice/

4. Methodological challenges of a Flexible Dissemination System

During the early development work for the FDS, it was identified that the SDC methods of targeted record swapping and table re-design used for the 2011 Census would not be sufficient on their own to protect data released through an FDS. This is because as well as presenting new opportunities, the FDS also presents new disclosure risks and challenges for which the 2011 methods would not appropriately address. A summary of these challenges is:

- **Univariate uniques –** cell counts of 1 in a marginal total that is apparent in every table produced for that variable revealing large amounts of information on individuals
- **Differencing –** similar tables produced alongside each other, when compared reveal unpublished information
- **Ensuring disclosure risk has been reduced to an acceptable level –** the vast increase in variable combinations available to users through an FDS makes it infeasible to evaluate risk in tables manually
- **How a Flexible Dissemination System affects microdata releases –** ensuring that results derived from microdata products are consistent with those protected through an FDS

While some of these risks are not limited to tables released through an FDS and may occur within a release of static tables, the increased volume of tables made available through an FDS and the greater flexibility offered to users make these risks more prevalent with an FDS. The following sections describe these methodological challenges in more detail and some potential solutions.

4.1. Challenge: univariate uniques

A univariate, or marginal, unique is where there is a unique record in the marginal total of a table. For example, if there was only one person in a particular age group in a table, that record will be a univariate unique in every table that includes that particular age grouping, making it easier to identify that individual and potentially revealing a large amount of information about that person. In figure 2, a single observation in the univariate table for A.3 can be tabulated against other variables to reveal that they also belong to categories B.2 and C.1. Repeating this exercise for all variable combinations will lead to a significant amount of information available for a single observation.

Figure 2: Example of the disclosure risk from univariate uniques

Variable A

Category	Count
A.1	173
A.2	25
A.3	1
A.4	7

Variable A by Variable B

	B.1	B.2	B.3	B.4
A.1	93	36	26	18
A.2	11	9	2	3
A.3	0	1	0	0
A.4	4	2	0	1

Variable A by Variable C

	C.1	C.2	C.3
A.1	106	32	35
A.2	8	14	3
A.3	1	0	0
A.4	5	2	0

Univariate uniques are a risk in any variable, not just sensitive variables, because it is not necessary for an intruder to know who the univariate unique is. Combined with information from the other variables that the unique record can be tabulated against, allows a greater chance for an intruder to make an identification. While it may be possible to find this information across a set of static tables, the data provider has much more control over how much information is available to users. By contrast, an FDS makes it easier for an intruder to generate multiple queries using the same variables creating a greater risk to unique records through an FDS.

Single observations in the marginal totals pose the most significant risk but the risk can be extended to other small cells. A two in a marginal total could lead to one of those two records identifying the other respondent through eliminating their own information. Some characteristics are common amongst household members, for example Ethnic Group or Main Language, so a small count in a marginal total could lead to the identification of a household unit, although there is more uncertainty in this example so may not require as much protection as unique records.

One solution could be to apply a threshold within the FDS that prevents any table with a univariate unique from being published. The disadvantage of this is where there is likely to be a significant number of univariate uniques which would result in the majority of tables being suppressed. Another solution could be to apply targeted record swapping to the microdata that specifically targets univariate uniques.

4.1.1. Targeted record swapping

The method used to protect the 2011 Census outputs was targeted record swapping whereby potentially identifiable households were swapped with similar households from a nearby area (ONS, 2012). This method worked well for 2011 and will continue to be the main source of protection for the 2021 Census outputs with the method being adapted to incorporate specific targeting of univariate uniques.

The basic method of targeted record swapping involves scoring every household in the dataset based on a number of characteristics that are targeted for rarity. The rarer the characteristics, the higher the risk score. These characteristics could involve highly visible variables to target noticeably identifiable households and/or sensitive variables to protect particularly vulnerable characteristics. A sample of households are selected for swapping based on their risk score such that the riskier households have a higher chance of being swapped than the non-riskier households, although every household has a non-zero chance of being selected. For each household in the sample, a matching household is identified from a nearby geographic area that is similar to the risky household on a number of characteristics, such as household size or age group of residents.

The size of the sample for swapping, or swap rate, is determined by how much protection is necessary to apply, taking into account other aspects that affect the quality of the dataset. For example, for a highly sensitive variable, it may be desirable to swap all risky households identified rather than just a sample. Alternatively, if the quality of the data

for a geographic area is particularly poor, for example if a high level of imputation has been required, then applying a lower swap rate would help to maintain data utility.

Adapting record swapping to incorporate targeting of univariate uniques requires that all variables to be included in the FDS are considered for targeting since a univariate unique in any variable could lead to an identification. It is also necessary to swap all univariate uniques and not just a sample since the amount of information available to an intruder is likely to outweigh the doubt introduced from just swapping a sample. Further swapping based on rare characteristics as before, can be carried out on top of the swapping due to adapted targeting.

In practice however, this approach could render the swap rate unacceptably high if the FDS is to include a large number of variables. An option would be to prioritise targeting variables that are more likely to contain univariate uniques and protect the remaining variables through other means, such as disclosure checks outlined in section 4.3. Alternatively, all variables could be targeted using a coarser level of aggregation while the more detailed version of variables could be protected using disclosure checks.

The benefits of targeted record swapping are that the majority of records are swapped within a low level of geography which maintains counts at higher levels of geography. For example, swapping records between Output Areas (the lowest level of geography used for Census outputs) within a Local Authority means that when aggregated, the Local Authority counts are unaffected. Also, by targeting the risky households, it reduces the damage to the data by focusing protection on the records that really need it.

The disadvantages of record swapping are that outputs at a low level of geography will be more affected because that is where the majority of risky records will be identified. It also means small counts are more affected as these are the cells that will be targeted as risky.

4.2. Challenge: differencing

When designing a collection of tables from a single data source, it is important to consider how the tables could be linked together. For example, if two similar tables are produced with slightly different classifications, they can be differenced to reveal previously unpublished information. An example is in Figure 3.

For a series of static tables, it is possible to consider every table within the context of the whole release as each table being produced is known. However, tables produced through an FDS are more susceptible to differencing as the flexibility provided to users increases. For example, if users have the opportunity to specify their own variable classifications, it would be possible for them to create overlapping categories in different tables creating the conditions for differencing.

Figure 3: Example of differencing

Published non-disclosive tables

Population: All usual residents

Marital Status	Frequency
Single	60
Married	78
Widowed	3
Divorced	19
Separated	5

Population: All usual residents in households

Marital Status	Frequency
Single	60
Married	78
Widowed	3
Divorced	19
Separated	4

Unpublished disclosive table

Population: All usual residents in communal establishments

Marital Status	Frequency
Single	0
Married	0
Widowed	0
Divorced	0
Separated	1

One solution would be to design the content of the FDS such that the opportunities for differencing are minimised by only offering standard variable classification, geographies and populations for selection. While this would minimise the chance of differencing, the opportunity may still arise from standard classifications. For example, in the 2011 Census outputs, standard population bases included 'All usual residents' and 'All usual residents in households'. If offered in an FDS, two identical tables can be produced using these populations to reveal information on communal persons at low levels of geography. Removing one of these population bases from the FDS would have a detrimental effect on data utility for a number of users who requires these populations and would also prevent comparability with outputs from previous censuses.

While it may be possible to design the FDS in a way that eliminated differencing, the flexibility of the FDS and in turn the level of utility to the user will be greatly affected as a result. An alternative solution which would allow a greater level of flexibility is the application of cell perturbation to the table prior to release. This provides protection against differencing because every cell has the potential to be perturbed and therefore the difference between two potentially perturbed cells will also be affected by perturbation.

Basic perturbation methods involve adding or subtracting a random value to the true cell values; the random value usually generated from a normal distribution with a mean of zero to avoid adding bias to the data. However, because noise is added randomly, it can create different values for the same cell.

The cell key perturbation methodology, originally developed by the Australian Bureau of Statistics (ABS) to protect their census tables generated through an FDS (Fraser and Wooton, 2009), applies perturbations consistently so that the same cell will always have the same value, even when appearing in a different table.

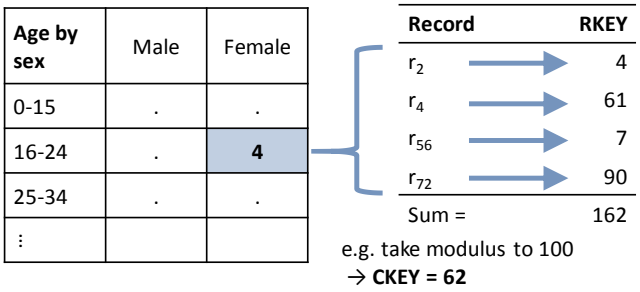
4.2.1. The cell key perturbation method

An outline of the method follows:

Step 1: Assign a random number to every record on the microdata dataset. This random number is called the record key (RKEY) and is from a uniform distribution on a finite range, say 1 to 100. Once RKEY is assigned, it remains unchanged for that record.

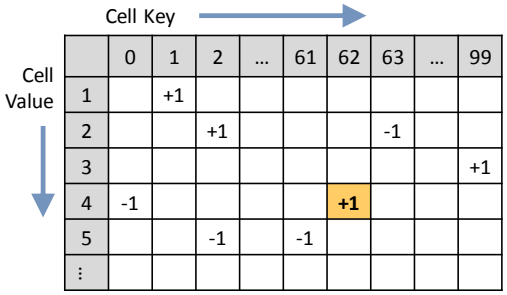
Step 2: For every cell in a table, sum RKEY for the records that contribute to that cell. Apply a function to the RKEY sum, for example taking the modulus to the range maximum, i.e. 100, to get a uniformly distributed value called the cell key (CKEY).

Figure 4: The generation of cell key from the contributing record keys



Step 3: Use the cell value and CKEY in a look up table (PTABLE) of perturbation values (PVALUES).

Figure 5: Example perturbation table



The PTABLE is designed such that the sum of PVALUES across each row equals zero so that the expected change of values in a table after perturbation is zero to reduce bias. The PTABLE can also be tailored to the level of protection required by adjusting the number of non-zero perturbation values across each row and the range of PVALUES. Different cell values can have different level of perturbation, for example small cell values can have a higher perturbation rate.

While CKEY will be within a finite range, cell value may not have a known limit, or the maximum value may be extremely high. To avoid having to create a row for every cell value, the rows within the table can be recycled for larger cell values. For example, if the PTABLE only contained 100 rows, then for any cell value over 100, the corresponding row of the PTABLE is used when cell value is taken to the modulus of 100. The exception is where cell value is divisible by 100, in which case row 100 is used. If a higher perturbation rate is used for small cells, then these should be excluded from being recycled.

Figure 6: PTABLE rows for corresponding cell values

PTABLE row	Corresponding cell values for each PTABLE row			
1	1	101	201	…
2	2	102	202	…
⋮	⋮	⋮	⋮	⋮
99	99	199	299	…
100	100	200	300	…

The design of the PTABLE prevents the cell key perturbation method from creating negative values but any differenced values may contain negative values depending on the perturbation that has been applied. This increases the protection from cell perturbation as it increases doubt that an intruder has generated true differenced values.

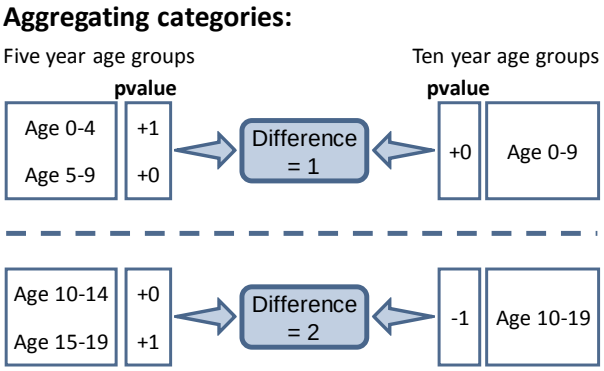
Step 4: Apply the PVALUE to the cell.

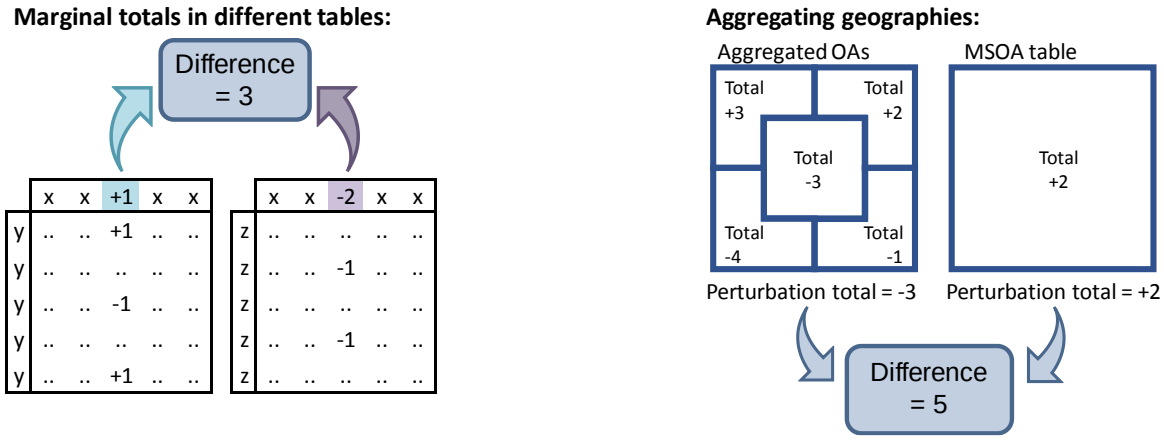
Along with the RKEYs, the PTABLE is created once and remains unchanged. Steps 2 to 4 occur in real time within the FDS once the table has been generated from the microdata.

The benefit of using the cell key perturbation method is that cells are perturbed consistently every time a table is generated or when the same cell appears in different tables. When a set of records are aggregated, the same record keys will lead to the same cell key being generated and the same perturbation value being applied.

While the same set of records will always produce the same set of results, this does not apply when those records are aggregated in different ways. For example, the sum of two perturbed cells will not be the same as a comparable cell in a different table as shown in figure 7. These inconsistencies can also occur when identical marginal totals are compared between tables that have been created by summing perturbed cells and when perturbed lower geographic areas are aggregated and compared to a higher geographic table.

Figure 7: Ways in which inconsistencies can occur from cell key perturbation





The higher the perturbation rate, the greater the inconsistencies in the tables. To limit the effect of inconsistencies for the 2021 Census, ONS are planning on applying a 'light touch' cell key perturbation method which takes into account targeted record swapping as the main source of protection. Targeted record swapping protects the identifiably risky households along with the univariate uniques while cell key perturbation provides a layer of protection against differencing.

In addition to applying cell key perturbation as a 'light touch', the ONS have adapted the method to allow the perturbation of zero value cells. The ABS do not allow cell counts of 1 or 2 to appear in their output tables as small cells are particularly vulnerable in an FDS; these cells are perturbed to either zero or three. For the 2021 Census, ONS are applying targeted record swapping which will provide some protection to these small cells but, in order to be able to apply additional protection through cell key perturbation, ONS have developed a method for consistently perturbing zero value cells. The zeros perturbation method uses similar principles to the cell key method but determines which cells should be perturbed based on contributing categories rather than contributing records. This ensures consistent perturbations as much as possible but as with the cell key method, inconsistent perturbations can occur when tables are constructed in different ways.

4.3. Challenge: ensuring disclosure risk has been reduced to an acceptable level

For publicly available outputs, the level of disclosure risk should be negligible. In theory, this could be achieved by removing all unique records from a dataset but in most cases this would require the removal of a significant number of records which would not maintain data utility. Thus, SDC methods work on the principle of adding a sufficient level of uncertainty to outputs so that disclosure risk is reduced to an acceptable level.

The main types of risk that can occur in frequency tables, illustrated in figure 8, are:

- **Identity disclosure:** where there is only one respondent with a set of characteristics, i.e. a cell count of 1.

- **Attribute disclosure:** where an intruder can use existing information on a respondent to learn something new, i.e. only one cell in a row or column is populated. A group attribute disclosure is an attribute disclosure where the cell value is greater than one, however attribute disclosures with large values are not usually considered as much of a risk as attribute disclosures with small numbers since a 'large' attribute disclosure is unlikely to be a rare characteristic. However, the variables involved are key to determining the level of risk of an attribute disclosure, no matter how large the cell value is.
- **Sparsity:** when a table contains a large proportion of zero value cells or cells containing small counts.

Figure 8: Types of disclosure that can occur in frequency tables using the example table of type of pet by weekly household expenditure

	£0 - £499	£500 - £999	£1,000 - £1,499	£1,500+	
Dog	0	52	0	0	
Cat	0	36	0	0	Sparsity
Fish	0	1	0	0	
Horse	0	0	0	1	Attribute Disclosure
Other	0	1	0	0	
Mixed	0	0	0	0	
None	1	203	0	0	Identity Disclosure

Methods such as targeted record swapping and cell key perturbation aim to protect these risks by adding sufficient uncertainty to the data so that an intruder cannot be sure if a disclosure is real or not. To evaluate the level of uncertainty in tables produced from the protected data, a further risk assessment is usually carried out. If the level of risk is above what is acceptable then either further record swapping or perturbation can be applied to reduce the risks, or tables could be designed in such a way as to reduce the exposure of risky cells.

For a static set of tables, it is usually possible to check all tables to be released for the levels of risk and uncertainty prior to publication and redesign them when necessary. For tables released through an FDS where there are few restrictions on what the user can select, the number of possible tables they can create can be substantial. For an FDS that includes 30 variables allowing a user to select up to 4 variables for a query, there are around 28,000 possible tables. With additional population and geography combinations, this number could become substantially higher. To manually check every table for sufficient uncertainty becomes an infeasible task within reasonable timeframes and resources.

A solution is to apply automatic checks to the table before it is released to the user. If the table fails the automatic checks then the user is informed their table cannot be produced and are able to amend their query to find a table that is acceptable to publish. Thus, the user is in control of the table redesign process than the data holder allowing the user to amend the table based on what their specific requirements are, selecting to retain detail in the variables they are primarily interested in.

Possible rules to prevent the release of tables that contain a high level of risk are:

- **Limit the number of variables a user can select** to prevent the user building a query with detail similar to a microdata record. It also helps to prevent sparse tables.
- **Limit the number of cells in a table** which helps to prevent sparse tables by limiting the level of detail across the dimensions selected.
- **Limit the number of cell counts of 1** to reduces the risk of identity disclosure.
- **Limit the number of cell counts of zero** to prevent sparse tables.
- **A marginal minimum threshold** which protects detailed information being revealed on small populations and can protect univariate uniques.
- **A marginal maximum threshold** which highlights dominance within variables where the majority of records fall into one row or column. This can indicate a lack of diversity in the remainder of the table.
- **An attribute disclosure threshold** to limit the number of attribute disclosures in a table. It could also be applied as a threshold on the size of an attribute disclosure to prevent small attribute disclosures from occurring in tables, or could be applied as a combination of both.

4.4.1. Applying disclosure rules in a Flexible Dissemination System context

The rules required in an FDS, and their parameters, will depend on the source microdata, the content that is made available for selection and the protection already applied to the data from other SDC methods.

The dataset: this includes aspects such as coverage of the data, for example if the dataset is a census or full coverage of a sub-population then a cell count of 1 will pose a bigger risk than for a sample survey where a sample unique may not be a population unique. The quality of the dataset is also a consideration; if the quality of the underlying dataset is poor then the rules may not need to be as strict.

The content: this includes geography, populations and variables. The lower the geographic level of the output, the more chance of risky cells occurring in the dataset, such as cell counts of 1 and attribute disclosures. If different levels of geography are available then the rule parameters may need to vary based on the geography selected, for example, the higher the geography selected, the more variables a user may be allowed to include in the table. Different populations have different levels of risk associated with them, for example a smaller population may be more prone to cell counts of 1. As for variables, the inclusion of sensitive or visible variables within the FDS, may require a stricter set of rules than a dataset that includes very few sensitive variables. It may also be desirable to apply different rules to different variables depending on their sensitivity.

The protection: when determining the rules to apply it is important to consider the protection already applied from any other methods so that the data are not over protected. If the record swapping method from section 4.1.1 has been used to swap all

univariate uniques, then applying a marginal minimum threshold not allowing a marginal value of 1 is unnecessary given the univariate uniques have already been protected.

The benefit of applying automatic rules is that aside from the initial investment to determine what the rules should be, the need for manual intervention in the table building process is reduced, decreasing the time before the data can be made available.

Another benefit is that the rules can be applied in a way the specifically targets the riskier areas of the country. Traditionally, tables are created that are safe to publish for every geographic area, essentially the maximum level of detail that is not disclosive for every area. This means that a table for a highly diverse urban area with no disclosive cells might not be able to be published because it was not safe to publish the same level of detail for some other areas, such as a less diverse rural area that does contain disclosive cells. While this does allow for comparability across the country, for users who are only interested in the urban area, data utility is greatly reduced. An alternative approach would be to evaluate each geographic area individually and publish the maximum level of detail that is safe to do so for that area but, with manual checking, it is not always practical to evaluate each geographic area independently. Automatic rules can be applied such that each geographic area is evaluated independently and if an area passes the rules then it will be released but if the area fails then it will not be released. If the table fails then the user will be able to amend their query to find an acceptable version. Comparability across the country will still be possible as less detailed queries are likely to be available for a wider coverage of areas but it will allow more data to be made available, where previously it was not, for the areas of the country where it is safe to publish.

A possible alternative to disclosure checks is access to the FDS through licencing which will involve users agreeing to a set of terms, usually including the promise not to try to identify respondents or claim to have identified respondents. Following agreement of these terms, users can then set up a user account allowing the data holder to monitor how the FDS is being used. It is still important to ensure the level of risk within output tables is appropriate for the level of access so there may still be a need for disclosure checks but they may not need to be as restrictive as with a public access FDS.

4.4. Challenge: how a Flexible Dissemination System affects microdata releases

In addition to releasing tables, many surveys release microdata, usually under licencing or through secure environments. If the microdata data set is the same as the underlying data of the FDS then there is a risk that the cell key perturbation method can be unpicked and the true values deduced. Tables created directly from the microdata would not have perturbation applied and, if compared to the tables generated through the FDS, then the user can deduce information about the perturbation method applied, lessening its protection to the data. Any pre-tabular methods applied to the microdata will still provide some protection.

If the microdata are being accessed through a secure environment then a condition of exporting the outputs from the environment could be to apply the same cell perturbation algorithm used in the FDS. This would be done by the data holder so the user does not have access to sensitive information about the method. If the microdata are being accessed through licencing then the user can download the data giving them access to both the perturbed and unperturbed tables at the same time greatly increasing the opportunity to unpick the cell perturbation method.

Solutions include designing both the microdata and FDS so that the same tables cannot be produced by both. An extreme way of doing this would be to have no overlapping variables although this is not likely to be practical as key variables, such as age and sex, are vital to users of both tables and microdata. For variables that do appear in both datasets, different classifications could be used so that no cells can be duplicated, for example if the FDS includes the age group 10-19 then the microdata could include the age group 15-24. This would only be made possible because the protection from differencing that the cell key perturbation method provides but again, this is not always practical as some classifications are standard, for example five-year age groups and Ethnic Groups.

A further solution would be to release a microdata dataset that is a sample from the underlying dataset of the FDS. Intruders could not be sure whether differences between tables created from both sources were due to the cell key perturbation method or because the microdata dataset was a sample. The drawback to this approach could be in situations where the underlying data of the FDS is a sample already. Taking a further sample to produce the microdata could result in reduced data utility, but the full dataset could still be made available through secure environments. This solution may only be suitable for instances where the underlying dataset is large, such as a census or large administrative dataset.

5. Parameterisation of a Flexible Dissemination System

The final stage of designing the SDC methods for an FDS is the process of parameter setting, for example, how much record swapping or cell perturbation is required, what disclosure checks need to be applied and what should their thresholds be. There is no set way for how to set the parameters because it will depend on the source data, the SDC protection package as a whole and the output requirements.

Every dataset is different and any SDC methods applied should be tailored to each dataset which will require its own risk assessment before being used in an FDS. This is because the risks in one dataset will be different to the risks in another. For example, a cell count of 1 poses a bigger risk in a census than a sample, and therefore record swapping may not be required for a sample as another method for protecting extreme values may be more suitable, such as design of the variable classifications in the FDS.

The protection required for each dataset should be considered as a whole rather than the application of SDC methods independently. For example, if a pre-tabular SDC method is used alongside cell key perturbation then a lower perturbation rate may be sufficient

whereas if the cell key method is the only source of protection, then a higher perturbation rate may be required.

Finally, the output requirements will influence the SDC methods and parameters required. The more flexibility built into the FDS through similar variable classifications, the higher the chance of differencing which will require a higher perturbation rate. Alternatively, if the lowest level of geography available through the FDS is relatively high, for example Local Authority, then a method such as record swapping may not be the most appropriate method for protecting univariate uniques.

6. Summary

The introduction of the FDS brings with it opportunities to make more data publicly available, granting greater accessibility and flexibility to users, while outputting results in a timelier manner than previously possible. As dissemination tools adapt to give the user more options, the disclosure control methods required to protect the data must also evolve.

The work outlined in this paper is based on the challenges encountered when developing SDC methods required to protect the 2021 Census data in an FDS. It is likely that similar challenges will be common across many ONS outputs along with those from other government departments. It is possible in the future that an FDS could be applied more widely to enable tables to be generated from other high profile datasets. As always, the methods will vary from one dataset to another and should be selected to ensure that disclosure risk is minimised and data utility is maximised.

References

[1] Hundepool, A. Domingo-Ferrer, J., Franconi, L. Giessing, S., Schulte Nordholt, E., Spicer, K. and de Wolf, P.P. (2012) *Statistical Disclosure Control,* Wiley Series in Survey Methodology.

[2] Office for National Statistics (2012) *Statistical disclosure control for 2011 Census*, ONS. Available at: https://webarchive.nationalarchives.gov.uk/20160129234204/http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/index.html (Archived 29 January 2016).

[3] Fraser, B. and Wooton, J. (2005) A proposed method for confidentialising tabular output to protect against differencing. *Joint UNECE/Eurostat work session on statistical data confidentiality* (Geneva, Switzerland, 9-11 November 2005).

An investigation into using data science techniques for the processing of the Living Costs and Foods survey.

Gareth L Jones[1]

Abstract

Working towards the strategic aim, digital by default, this paper explored ways to improve the timeliness of delivery for the Living Costs and Foods Survey (LCF). This included the development of deep learning algorithms to perform optical character recognition on the content from purchase receipts; and thus, enabling automatic classification of products through machine learning classification mechanisms. By developing an Optical Character Recognition(OCR) application, we extracted textual information from the receipts provided as part of the LCF diary process. This extracted text was passed to machine learning algorithms to classify the receipt item text into the LCF item description and associated Classification of individual consumption by purpose (COICOP) code.

It was found that scanning of receipts as a concept had potential to improve the timeliness of LCF diary processing. In practice, there are processing issues as outlined in this paper which impede the performance of such a solution. Quality of receipts had a strong influence on the performance of the application and the quality of the extracted text. Whilst receipts received were acceptable for the manual coding process they proved to be problematic for an OCR scanning solution.

1. Introduction

The Living Costs and Food Survey (LCF) is a household survey whose primary purpose is to collect information about expenditure on goods and services by UK households. The data collection of the LCF is split into two components; a face-to-face questionnaire followed by a self-completion expenditure diary. In 2016, the LCF underwent a National Statistics Quality Review (NSQR)[2]. This review recommended that we explore the possibility of semi-automated coding of purchase information from scanned supermarket receipts.

This paper shows the work carried out to date in relation to this recommendation posed by the NSQR. The paper first looks at the attempts made by other National Statistics Institutions (NSI) towards implementing an automated process. Then we look at the

[1] Gareth L Jones – Data Analytics Apprentice – gareth.l.jones@ons.gov.uk

[2] https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/methodologies/nsqrseries2reportnumber3livingcostsandfoodsurvey

application, how it was developed- including the challenges and results. Finally, the paper looks at the automatic classification of item description to COICOP code.

1.1. Background of the Living Costs and Food Survey

1.1.1 History

A household expenditure survey has been conducted each year in the UK since 1957. From 1957 to March 2001, the Family Expenditure and National Food Surveys (FES and NFS) provided information on household expenditure patterns and food consumption for government and the wider community. In April 2001, these surveys were combined to form the Expenditure and Food Survey (EFS) which was later renamed the Living Costs and Food Survey (LCF) in 2008.

1.1.2 Uses

LCF data are widely used within and outside government. The data are used to provide information on spending patterns for the Retail Prices Index (RPI) as well as provide the weights for the Consumer Price Index (CPI). Other users of the LCF expenditure data include the Statistical Office of the European Communities (EUROSTAT[3]) and other government departments such as Her Majesty's Revenue and Customers (HMRC)[4].

1.1.3 Diary Processing

Each individual age 16 or over in the household is asked to keep a detailed record of daily expenditure for two weeks. Children aged between 7 and 15 years are asked to keep a simplified diary of daily expenditure.

To reduce the burden of completing the diary the respondent(s) are given the option to provide the receipts for their purchases rather than writing the individual purchase(s) in the diary. These transactions are then manually inputted onto the Blaise[5] system by the processing team in Titchfield which currently takes on average 3 hours per diary to complete.

2. Work carried out by international National Statistic Institutions

We considered the work carried out by other National Statistical Institutes (NSI) who have considered the application of semi-automatic coding in there on household budget surveys. Our findings from three NSI's are as follows.

[3] http://ec.europa.eu/eurostat/about/overview

[4] https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/methodologies/livingcostsandfoodsurvey/livingcostsfoodtechnicalreport2015.pdf

[5] https://www.blaise.com/products/general-information

2.1. Sweden

Sweden has been scanning both diaries and recipes for their household budget survey since 2012 using the commercial software EFLOW[6]. Sweden did not report finding any problems with scanning the diaries as the format of the Swedish diary was designed for a scanning method. When it came to extract contextual information from purchase receipts they found this process much more complex. Like the UK, receipts in Sweden are not standardised. The information required can appear anywhere on the receipt and condition was also a problem. Sweden opted for a commercial solution using a customised version of the EFLOW invoice scanning software.

2.2. Finland

Finland have been scanning receipts since 2016 using the KOFAX capture software[7], a commercial solution which encompasses an all in one software solution of scanning and coding items. Finland also employ a restriction of only scanning receipts with 3 items or more. To put this process in place, Finland developed a separate application to interface with the KOFAX software to allow manual editing of data. Using this process 80% of data was extracted from the receipts, with 20% requiring manual entry (short receipts and written data).

2.3. Netherlands

The Netherlands explored a similar project in 2013. Their aim was for respondents to scan receipts as part of a digital diary and then carry out automatic COICOP classification. They achieved a correct receipt processing rate of 50% of scanned receipts, with their application extracting the correct prices amounts for 75% of the receipts. The Netherlands classifier correctly coded 85% of products to COICOP classification, however they found that OCR is a critical factor and poor image quality impacted on performance and thus increasing respondent burden as non-recognised receipts would need manual input. The Netherlands concluded that scanning was not a sustainable option.

3. Receipt Scanning Application

The receipt scanning application for this project was built using the Shiny[8] package from the R programming language[9]. Options in Python were also explored but R was chosen for its flexibility using the Shiny package. Shiny is a package in R that allows you to build browser based applications which can be deployed to the user to work locally or run on a server.

[6] https://www.topimagesystems.com/solutions/content-process-automation/forms-processing/
[7] http://www.kofax.com/document-capture-software/
[8] https://shiny.rstudio.com/
[9] https://www.r-project.org/about.html

The application includes three elements:

- Optical Character recognition
- Image processing
- Item description Classification

The premise of the application is that the extracted text would pass through a classification algorithm to convert the receipt description to an LCF description which would then feed into a COICOP classifier to allocate the COICOP code. To achieve maximum time efficiencies, we need the least amount of human interaction as possible. Thus, achieving a high level of accuracy in the initial receipt item classification was paramount in the application succeeding. If we could not classify the receipt item description to the LCF item description then the coder will still need to convert this intuitively, thus not achieving any efficiency by scanning the receipt.

3.1. Optical Character Recognition

The core part of the application is the Optical Character Recognition (OCR) engine. The OCR engine extracts the textual data from the scanned purchase receipt images and outputs this as a raw text file. For this application, we used Tesseract OCR[10] provided by Google Open Source. This is a pre-trained OCR engine which can recognise up to 100 different languages with capability for further training. This technology was chosen as it is open source, plus both R and Python have ready developed packages to interface with this technology.

Optical character recognition (OCR) is the process of converting images of printed, handwritten or typed text into machine readable/encoded text. At its core OCR is an algorithm which produces a ranked list of candidate characters based on one of two methods.

- **Pattern Matching** using matrix matching, which compares the image of the character to be read against a stored image on a pixel-by-pixel basis. This method relies on the input image to be correctly isolated from the rest of the image and of a similar font and scale to the stored image.
- **Feature extraction** separates the image into "features" such as lines, closed loops, intersections etc. This process reduces the dimensionality of the image allowing the recognition process to perform in a more computational efficient way.

OCR software such as Google Tesseract employ a two-pass approach. This is where the OCR engine processes the image twice, first pass picking out the high confidence letter recognitions and then using these results in the second pass to achieve a better recognition of the remaining letters.

[10] https://opensource.google.com/projects/tesseract

3.2. Image Processing

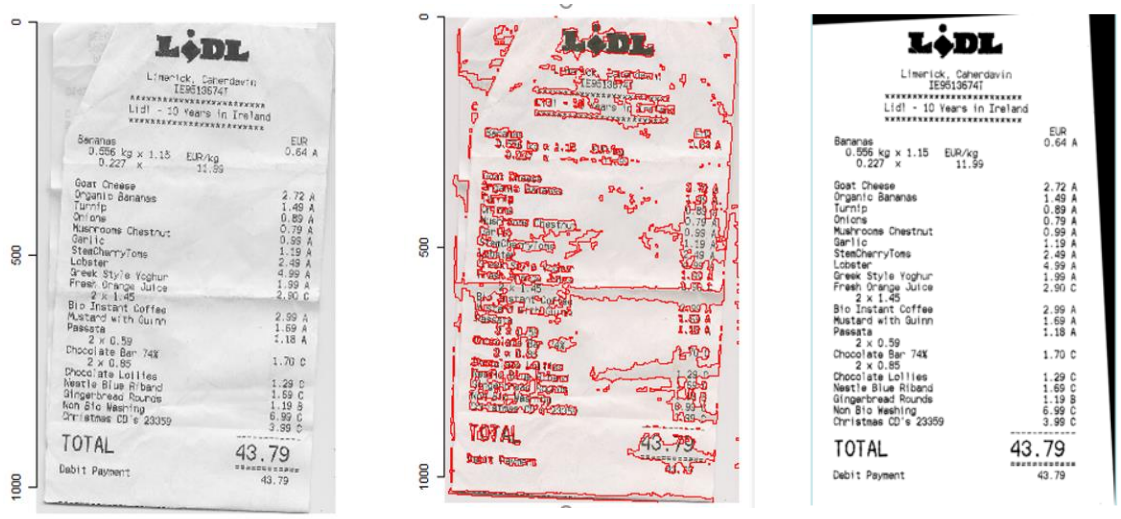
Before the image can be passed to the OCR engine, we needed to clean the image. Purchase receipts as part of the survey process were received in various levels of quality and formats as shown in Figure 1 below. We found that the image cleaning treatment of one receipt would not be applicable for the other. Thus, there was a number of challenges in developing a generic cleaning process.

Figure 1 - Example variations in receipt qualities and formats

The most common quality issues found were images which were damaged, marked, rotated and faded. Using a combination of the magick[11] and imager[12] packages for R, the images were converted to black and white, passed through thresholding algorithms to remove marks, de-skewed to correct any misalignments, cropped and sharpened to counteract fading. The resultant image was then at a state where the OCR can read the text. Figure 2 shows how the image is processed as it passed through the pre-processing steps.

[11] https://cran.r-project.org/web/packages/magick/magick.pdf
[12] https://cran.r-project.org/package=imager

Figure 2 - Receipt image passing through pre-processing steps



Receipt image after first pass of pre-processing converting the image to greyscale.

Receipt image during the edge detection phase to identify and remove folds and marks.

Final image correctly aligned with folds and marks removed and converted to Black and white.

Annotations proved to be more challenging for the application to deal with. As part of the diary process the respondents are asked to annotate their purchase receipts to include weights and measures information to meet LCF stakeholder needs. This confused the application as the OCR engine struggled with handwritten notes over printed text. Another common annotation is marked off items on online grocery receipts, e.g. when a respondent mark off the items they have received. These small marks over the beginning of the item text led to miss reads in the OCR process. As per current practice the receipts are annotated with the following colour coding: blue for the respondent; green for the interviewer; and red for the coder. Using these colours, we explored removing the annotations by means of colour filtering, however we found that we could not remove the annotation in full without degrading the overall quality of the image in the process. A possible solution to this could be the use of specialised pens (such as the Non-repo pen[13]) where the ink is not scannable. These pens work by using a light blue ink which is not picked up by the scanners. Pens such as this could be handed out to the respondents for annotating the receipts. Testing would be required to see how this impacts on OCR results but may provide a compromise with need for annotations and OCR performance.

3.3. Item description Classification

In the UK purchase receipts come in all kinds of formats and the information contained can be abbreviated in various ways. The challenge here was to design an application which could handle variable formats whilst still extracting the required information. Given the number of retailers in the UK each having a variation in receipt format, we decided to focus on receipts from the big five supermarkets (Tesco, Asda, Morison's, Aldi and Lidl). Our initial approach was to create rules based on retailers to direct the application

[13] https://www.jetpens.com/Non-Repro-Blue/ct/539

where to find the relevant information. This became unsustainable with changing receipt formats (i.e. locations of weights and quantities), also image quality caused receipt type recognition problems. For example, Tesco online shows quantity of items and weights whilst the standard Tesco receipt does not. The risk with this approach is that if the application could not distinguish that this was a Tesco receipt and which kind of Tesco receipt, then it could not apply the receipt specific rules and extract the text.

We switched to a more generic rule based system where the application looked for matching patterns to indicate item and price and extract this. This limited the application to data variables which are common across all the tested receipts such as item and price. As weights and measures are not present in all receipts and when present not in a common format or location this was not extracted as part of the process. What was found with weights and measures is that the same shop and receipt type had different locations for this data. The exception for this was Tesco online where the weights quantities were in a consistent location however the weights were not.

Once the raw text was extracted it was passed through a series of regular expression string matching algorithms which manipulated the data into a format that could be uploaded to the Blaise survey system. Figure 3 is an extract from the application showing how the extracted text has been manipulated via the string matching algorithms.

Figure 3 - Raw extracted text separated into Item description and Price

	OCR Text	Recipt item	Price (in pence)
1	Goat Cheese 2.72 A	Goat Cheese	272
2	Organic Bananas 1.49 A	Organic Bananas	149
3	Turnip 0.89 A	Turnip	89
4	Onions 0.79 A	Onions	79
5	Mushrooms Chestnut 0.99 A	Mushrooms Chestnut	99
6	Garlic 1.19 A	Garlic	119
7	StemCherryToms 2.49 A	StemCherryToms	249
8	Lobster 4.99 A	Lobster	499
9	Greek Style Yoghur 1.99 A	Greek Style Yoghur	199
10	Fresh Orange Juice 7 ‘ 2.90 C ’	Fresh Orange Juice ‘ ’	290

When we compared the extracted and processed text to the LCF data it was found that the item descriptions on the receipts were different to the item descriptions inputted during the manual coding process, which can be seen in Figure 4 below.

Figure 4 - Differences between the receipt item text and the LCF item description

Each retailer describes items on receipts in a different way and thus a process of standardising the item description is carried out by the coder as part of the coding process, following the coders guidelines. To replicate this process a machine learning classifier was developed which could take the receipt item text and classify it to a standardised LCF description.

4. Receipt Text Classifier

In order to train a classifier, we needed a training dataset which included the receipt item text along with the item descript which would be entered into the LCF survey.To obtain this an additional field was added to the Blaise coding system so that the receipt text was captured as part of the manual coding process. This was carried out in the last quarter of 2017 and the classifier was trained using data from November 2017.

The classifier used was a Support Vector Machine (SVM) which is a machine learning algorithm that can be used for both classification and regression. For this application the classification functionality was used to create a text classifier. We trained the classifier using the following split of training, testing and validating data.

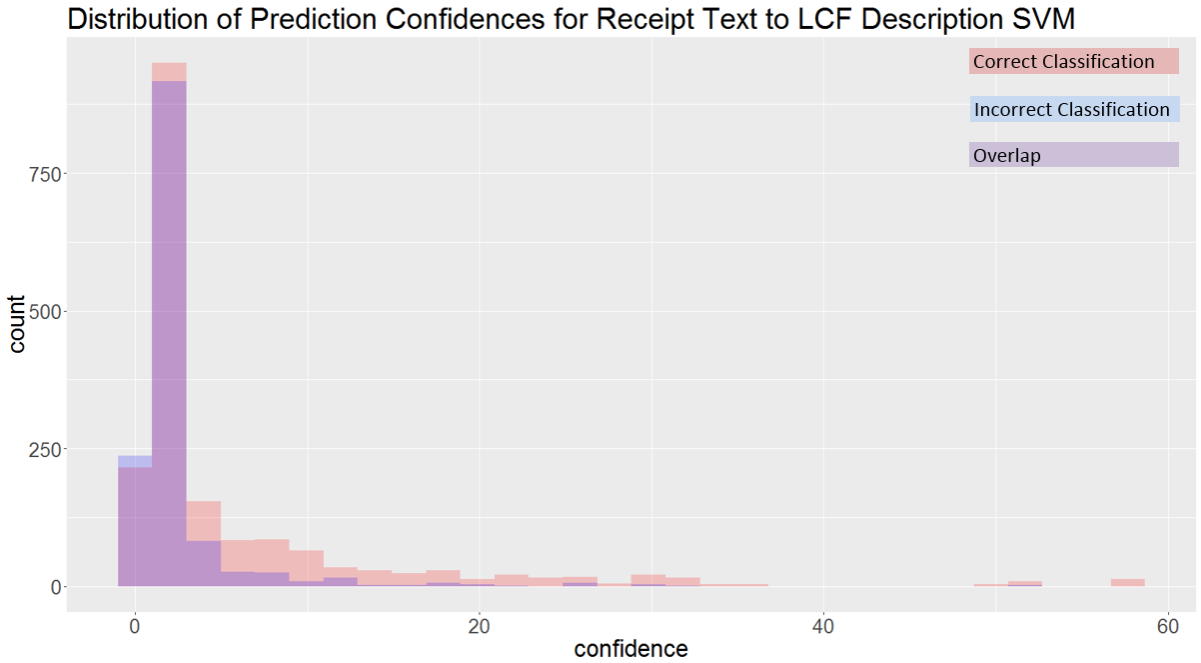
Table 1 - Breakdown of number of items used to train, test and validate the classifier

Training dataset	80% of data	25,184 Items
Testing dataset	10% of data	3,148 Items
Validating dataset	10% of data	3,148 items

The classifier yielded a 30% classification accuracy in receipt item text to LCF item description. We define accuracy in this case as the percentage of total receipt item texts correctly classified to LCF item description. One cause of low accuracy was due to the way items are represented on receipts. What was found is that there can be many different receipt item descriptions relating to the same LCF item description. In a similar fashion, there can be different LCF descriptions relating to the same or close to the same receipt item description (despite the standardised instructions set for coders). This made it difficult for the SVM to give a correct and confident classification. The results were either a highly confident incorrect classification or a "lucky guess" where the SVM classified the correct description with a 10% confidence in that classification. Figure 5 below shows the counts of correct matches in red with the incorrect matches in blue.

The purple area shows the overlap of confidences in both correct and incorrect classifications.

Figure 5 - Distribution of Prediction Confidences for Receipt Text to LCF Description SVM



To productionise this process, we needed to ascertain a confidence level where we could tell the application to accept the classification or not. Any matches not accepted would need to be manually amended in the application. It is clear from the confidence distribution that the classifier was equally confident in both the incorrect and correct matches little to no confidence in theses classification. To create a process which could be used in "Business as usual process" we needed to see a distinction between the confidences of correct and incorrect predictions. This would have given us a confidence cut off point where we can tell the application to reject the classification and indicate that the item would need manual coding. Currently as there are very little correct matches at a confidence level where there are no incorrect matches the classifier would reject most classifications.

To achieve a higher level of accuracy in the receipt item text classifier there needed to be a smaller hierarchy of classifications than is currently present in the LCF item description. This would mean a change in the way item descriptions are stored in the LCF and investigation on how this will impact on current LCF outputs would be required.

Linking the data with supermarket scanner data could also provide a better hierarchy for receipt text classification using the standard Global Trade Item Number (GTIN) description of item.

5. COICOP Classification

The Classification of individual consumption by purpose (COICOP), is a classification developed by the United Nations Statistics Division to classify and analyse individual

consumption expenditures incurred by households[14]. COICOP has 5 levels of detail which is currently allocated by the coder in Titchfield. Our aim here was to see if we can build a classifier which can be used in a production environment to automatically classify items down to the 5th level of COICOP.

Initial work carried out within ONS assessed three different types of machine learning algorithms: Naïve Bayes; Random Forest; and SVM. Out of the three, SVM yielded the better accuracies with predictions up to 97% accuracy. The code for this model was adapted so that confidences in classification could be extracted and further investigation could be carried out to see if the method is production viable. The modified model resulted in a reduced accuracy due to the changes needed to extract confidences.

To train the COICOP classifier the 2016 LCF dataset was used and split into a training (133,984 item dataset) and testing (38,598 item dataset). The accuracy rates for the classifier was as follows.

Table 2 - Breakdown of Classification accuracy by COICOP level

COICOP Level	Prediction Accuracy
First	96%
Second	95%
Third	92%
Fourth	89%
Fifth (Full COICOP)	87%

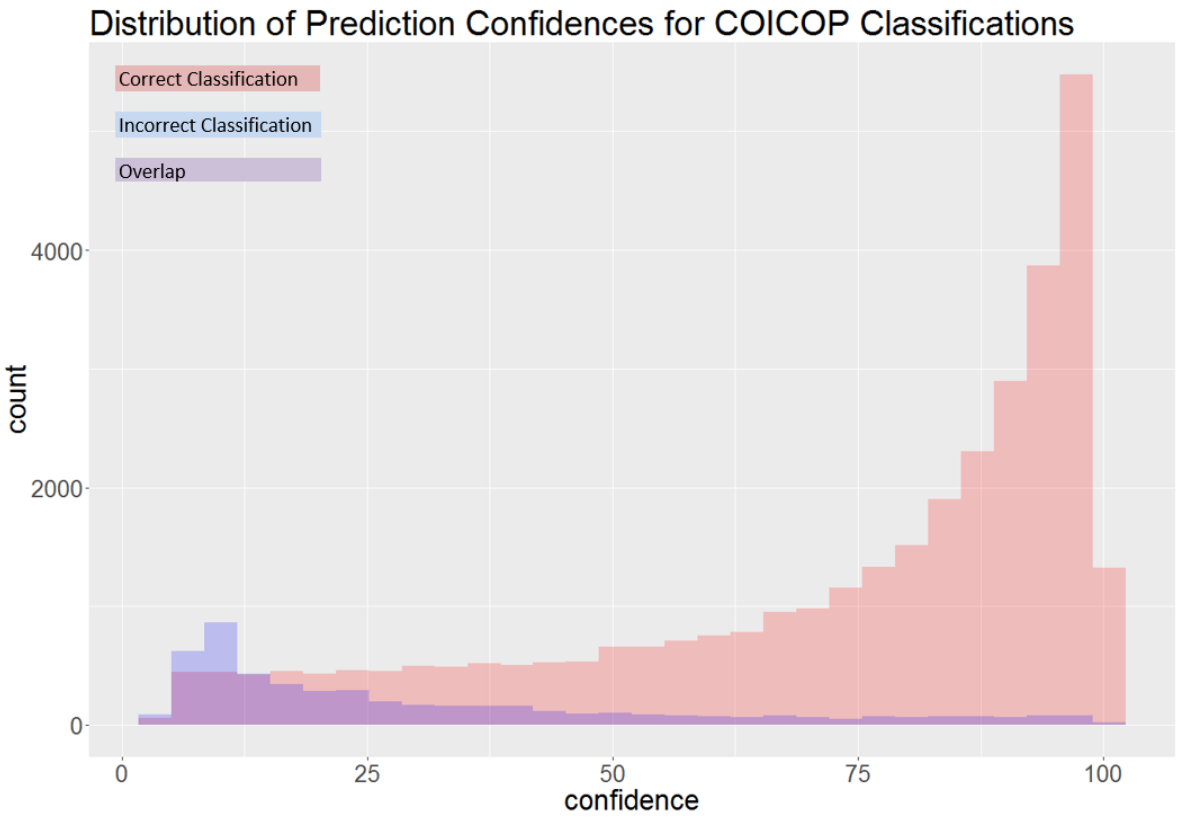
Out of the testing dataset the classifier correctly classified 87% (33,194) of items passed through at full COICOP level.

Prediction confidences were needed because to proceed into production we needed a way to extract the predictions that were at low confidence so that manual amendments could be done. We knew the overall accuracy of the model was 87% at full COICOP, but we didn't know how confident the model was in these predictions. The amendments to the model allowed us to extract the individual confidences for each prediction. Using this we were able to see if there was a point where we could separate correct and incorrect predictions by confidence level.

The extracted confidences ranged from low confidence correct predictions to highly confident incorrect predictions as shown in figure 6 below.

[14] http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Classification_of_individual_consumption_by_purpose_(COICOP)

Figure 6 - Distribution of SVM classifier confidences in COCIOP code predictions



The above distribution shows the correct classifications in red with the incorrect in blue. Unlike the receipt description classifier, the confidences in correct and incorrect classifications are more distinctive. We found there were a lot higher confidence in correct predictions where the incorrect predictions had lower confidences as expected. However, there is concern that there are incorrect classifications with high confidence. Ideally, we did not want the confidence distributions to overlap so that we could separate the correct from incorrect classifications. With regards to the overlapping confidences we don't know whether the cause of this is with the process used for training the classifier or within the training data itself. Further analysis is needed into the incorrect classifications to see why the classifier is so confident in the wrong prediction whilst having such low confidence in a correct classification.

6. Conclusion and areas for further study

6.1. Conclusion

Our research has identified some key findings in the process of semi – automated data collection. Scanning of receipts as a concept does have potential to improve the timeliness of LCF diary processing. In practice, there are processing issues which impede the performance of such a solution. Quality of receipts had a strong influence on the performance of OCR and the quality of the extracted text. Whilst receipts received were

acceptable for the manual coding process they proved to be problematic for an OCR scanning solution.

Item descriptions in LCF as recorded by coders are substantially different to how items are represented on receipts. Automatic classification of the receipt item text to the current LCF item description was too complex to achieve an accuracy needed for production value. The derived classifier correctly classified 30% of receipt item texts to LCF description classification using the current LCF item description hierarchy. Reducing the hierarchy in the LCF item descriptions showed improved classification accuracies, however we will need to ensure that this level of detail will satisfy LCF stakeholder needs.

Automatic classification of COICOP code showed promise in improving timeliness of LCF processing. Results from the classifier showed it was possible to train a classifier to produce highly confident prediction of COICOP code at first level (96% of records were correctly classified to COICOP) however accuracy reduced as COICOP level increased. Classification of full COICOP code was correct in 87% of cases.

6.2. Areas for further study

Our research identified that the level of variation in LCF description impacted on performance of the receipt text classifier. The volume of abbreviations used to add detail to the item description restricts the accuracy of automation as there is a many to many relationship between receipt and LCF data. Reducing the variation in item descriptions could potentially enable a higher accuracy in classification however further study is needed to see the impact this has on current statistical outputs.

Linking the data with supermarket scanner data could provide a better hierarchy for receipt text classification using the standard GTIN[15] description of item. Acquisition of commercial data such as store scanner and GTIN data is being pursued by ONS. Subject to acquisition of data, further research will be carried out to explore the feasibility of integrating commercial data into LCF processing. In addition to enhancing LCF data such sources could provide a smaller and standardised hierarchy of item descriptions which would support further development of the OCR application whilst building a framework for additional innovations such as barcode scanning and mobile phone applications.

Automating COICOP classification has potential production value. For such a solution to be put in place further study is needed into why incorrect COICOP codes have been predicted with high confidence whilst some correct codes have been predicted with low confidence. Further investigation into why this has happened which should naturally lead to investigating into how automatic classification could be implement in to the BAU process. This element of the initial research project is being taken forward for further study within the LCF team.

[15] https://www.gtin.info/

Integration of survey and Value Added Tax (VAT) data at ONS for short term output indicators: methodological challenges and approaches

Jennifer Davies[1]
(Dominic Brown, Gary Brown, Katie Davies, Claire Dobbins, Duncan Elliott, Rhonda Hypolite, Megan Pope)[2]

1. Introduction

Since 2017, ONS has undertaken several projects investigating whether Value Added Tax (VAT) turnover data from HM Revenue and Customs (HMRC) can be used to enhance or replace survey data. This is part of a wider strategy to "develop our capability to integrate administrative and commercial data sources, supported by appropriate methods and standards" outlined in the UK Statistics Authority Strategy for Statistics 2015 to 2020. Administrative data have many advantages, however as they are typically not collected for statistical purposes, pose several challenges if to be used in the production of official statistics.

This paper discusses the challenges faced, and proposed methodological solutions, for using VAT turnover data in the production of short-term output indicators for the Distributive Trades industries (UK SIC 2007[3] divisions 45, 46 and 47). The results provided are indicative based on the research methods used as part of this work.

The paper is structured as follows. Section 2 provides background to the coverage of the Distributive Trades (DTrades) industries. Section 3 provides an introduction to features of VAT turnover data. Section 4 discusses the challenges presented by VAT turnover data, research to date, and current recommendations for methods to address these. Section 5 outlines the proposed statistical design approach for combining VAT and survey data for DTrades. It also lists the acceptance criteria for deciding where VAT turnover data are appropriate to use. Section 6 then presents results of applying the methods recommended in section 5 to produce possible outputs combining both survey and VAT turnover data for DTrades. It provides examples where the resulting estimates do not meet the acceptance criteria and a discussion of the remaining issues. The paper ends with conclusions and suggestions for further research in section 7.

[1] Office for National Statistics, on secondment to Australian Bureau of Statistics Jennie.Davies@abs.gov.au

[2] Office for National Statistics. Contact name Claire.Dobbins@ons.gov.uk

[3] https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007

2. Distributive Trades

ONS produces several short-term output indicators for industry sectors of the United Kingdom (UK) or Great Britain (GB) economy. These are the Index of Production (IoP), Index of Services (IoS), Construction Output and Retail Sales Index (RSI). Short-term turnover statistics are used as inputs in calculating these indices. These use data collected by three surveys under the umbrella of the Monthly Business Survey (MBS). There are separate surveys for Monthly Business Survey - Production and Services, Monthly Business Survey – Construction and Allied Trades, and Monthly Business Survey - Retail Sales Inquiry.

As part of ONS' ongoing commitment to transform economic statistics, data sources, data collection, methodology and technology used to produce short-term output statistics are being reviewed. One change is to introduce a new survey: the Monthly Turnover Survey (MTS). This survey will initially collect data from businesses in UK SIC 2007 divisions 45 (motor trade), 46 (wholesale) and 47 (retail). The intention is that the coverage will then be extended to cover the remaining industries measured in the MBS.

In addition to a new questionnaire, there is the opportunity to use both survey and VAT turnover data as inputs to the short-term turnover statistics with the aim of transforming data collection activity and reducing ONS' reliance on large surveys.

3. VAT Turnover Data

This section provides a brief overview of some of the key features of VAT turnover data which lead to some of the challenges of its use for the short-term turnover statistics. The requirement is to produce monthly estimates of total turnover. The properties of the VAT dataset are comprehensively covered in an ONS published document titled 'Quality assurance of administrative data (QAAD) report for Value Added Tax turnover data'[4].

Businesses are required to register with HMRC for VAT if their annual VAT-taxable turnover exceeds £85,000 (correct as of April 2017). On registration, businesses are allocated to one of three quarterly reporting periods, (quarters ranging Jan-Mar, Feb-Apr or Mar-May). The different reporting periods are referred to as staggers. Businesses can request to change their reporting period to another quarterly stagger, or to report on a monthly or annual basis. In total, there are 16 staggers; one for monthly reporters, three for quarterly and 12 for annual (year ending Jan, Feb, …, Dec).

The deadline for submitting VAT returns to HMRC is 1 month and 7 days after the end of the reporting period for monthly and quarterly returns, 2 months for annual returns. For

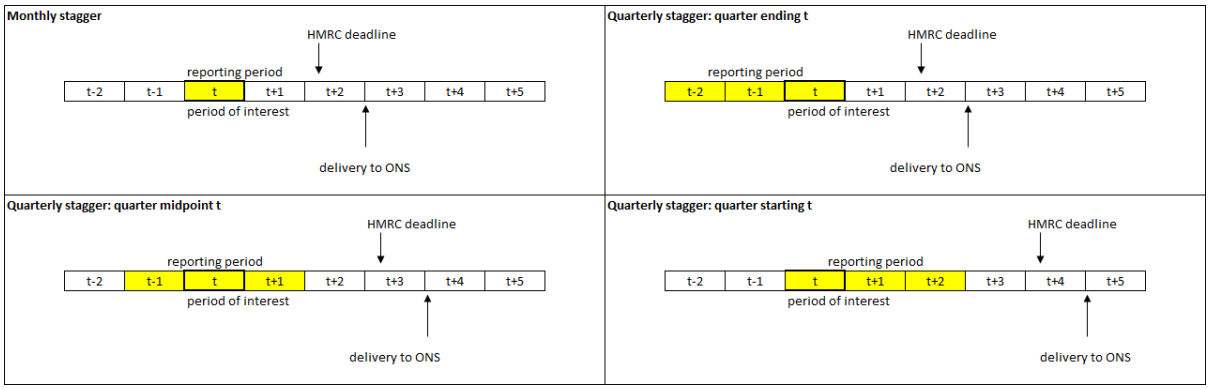
[4]https://www.ons.gov.uk/economy/economicoutputandproductivity/output/methodologies/qualityassuranceofadministrativedataqaadreportforvalueaddedtaxturnoverdata

example, for a VAT return covering the quarter January to March 2018, the deadline for submission was 7 May 2018.

ONS receives a file of VAT data from HMRC on the first working day of the month. It comprises returns received in the previous month, which could cover multiple different reporting periods. The effect of the different staggers, the deadline and the delivery date to ONS mean that data for any given month can be received in different data deliveries.

For a given month of interest t, consider the monthly and quarterly staggers. If the business reports monthly, then it should submit a return for that month. If it reports quarterly it should provide a return where the month of interest t is either the start, middle or end month of the quarter. Assume that a business submits its VAT return on the day of the HMRC deadline. If the business reports monthly, then this deadline will fall at the start of month $t+2$. For the three quarterly staggers this will fall either at the start of $t+2$ (t is the end month of the quarter), $t+3$ (t is the middle month of the quarter) or $t+4$ (t is the start month of the quarter). The data will then be delivered to ONS at the start of the next calendar month. This will either be at the start of $t+3$ (monthly and quarter ending t), $t+4$ (quarter with t as the middle month), or $t+5$ (quarter starting t). These timelines are illustrated in Figure 1.

Figure 1: Timeline for ONS receiving VAT data for a given month t delivered on the HMRC deadline day from the monthly and quarterly staggers.



Data from early VAT returns may be received ahead of these timescales, and likewise, late returns could be received after these timescales. For a reporting period ending in a particular month, say January 2019, 1.2% of the returns are, on average, available after 1 month (February 2019), 60.6% after 2 months (March 2019) and 95.9% after 3 months (April 2019).

This is not saying that 95.9% of data covering January 2019 is available in April 2019. In fact, it would be expected that by April 2019, for monthly and quarterly reporting periods covering January 2019:

- 95.9% of data for the monthly reporting period January 2019 is expected to be available in April 2019
- 95.9% of data for the quarterly reporting period ending January 2019 (Nov 2018 - Jan 2019) is expected to be available in April 2019
- 60.6% of data for the quarterly reporting period ending February 2019 (Dec 2018 - Feb 2019) is expected to be available in April 2019
- 1.2% of data for the quarterly reporting period ending March 2019 (Jan – Mar 2019) is expected to be available in April 2019

Table 1: Percentage of overall VAT returns received by ONS between 1 and 5 months after the end of the reporting period

Number of months from the end of reporting period	Average percentage of VAT returns received by ONS
1	0.4%
2	20.6%
3	52.6%
4	84.8%
5	97.4%

4. Challenges

4.1 Error Detection and Correction

Both survey and VAT data can contain errors. MBS and RSI use a selective-editing approach to error detection. Selective editing calculates a score for each business based on the impact that a value, if incorrect and remains unchanged, would have on the final output. The score is then used to prioritise which businesses require further validation of their responses. In such cases, businesses are re-contacted either to verify that the data are correct or amended if incorrect. The approach was implemented in 2010 and an overview is provided in Skentelbery (2011).

The size of the VAT dataset requires an efficient cleaning strategy. On business surveys, when potential errors are identified, re-contacting businesses is the preferred option for error correction. For VAT data, re-contact is not a viable option. Options for error correction that are available include rules-based editing, imputation and estimation methods, or comparison with other data sources.

Previous research has identified two common systematic errors, the thousand-pound error and the quarterly-pattern error. The thousand-pound error is where a business has provided turnover in thousands of pounds sterling rather than in pounds sterling (i.e. 1000 times smaller). The quarterly pattern error is where a business which has provided quarterly VAT data has provided the same value for four consecutive quarters, or the same value for three consecutive quarters and a different value for the final quarter. These patterns are not thought to be true representations of quarterly turnover, rather annual figures spread over a year, or estimates balanced to a final annual figure.

The quarterly-pattern errors are automatically treated. It is assumed that an annual total has been provided. This annual total is redistributed using the median proportions for each quarter based on businesses in a homogenous class. This method was recommended in ESSnet (2011).

The thousand-pound errors are detected by calculating a ratio of the current unedited value with the previous edited value. If this value falls between a set range then it is flagged as a thousand-pound error and multiplied by 1000 (ESSnet, 2011).

However, these automatic rules do not capture all potential errors and further cleaning is required. A similar philosophy to selective editing on survey data has been tested and proposed for error detection on VAT data. Analysis found that a modification was required to the score to more effectively identify influential errors. For error correction, it was found that automatic error correction by imputation was over-treating the data and smoothing-out genuine movements. It is therefore proposed that where potential errors are identified, they are investigated by subject-matter experts. If they are deemed to be errors, then this decision is recorded and the value is replaced with an imputed value. The proposed method and results of testing are presented in Davies (2018).

Davies (2018) found that after selective editing, some suspicious values remained in the aggregated time series. It is therefore recommended that macro-editing is undertaken to compliment micro-editing. Time-series modelling with automatic outlier identification was found to perform well as an outlier detection method; however, using estimated values from the model to correct errors was again found to over-clean the data. Therefore, it is recommended to use this approach to identify potential errors in aggregated data for further manual inspection at the micro-level. This approach could be adopted more widely to improve the efficiency of macro-editing, which is largely a manual process on survey data.

This cleaning research was conducted in parallel with the analysis to identify where survey data could be replaced by VAT turnover data presented in section 6; therefore, the data presented in section 6 have not been cleaned under the proposed methods and contain some errors.

4.2 Missing data and timeliness

There are two reasons why VAT data may be missing: timeliness and under-coverage.

On timeliness, RSI is currently published 18 or 19 days after the end of the reference period. IoP and IoS, which MBS data feed into, are published around 40 days after the end of the reference period. As described in section 3, the deadline for submitting VAT returns to HMRC is generally 1 month plus 7 days after the end of the reporting period. ONS receives data from HMRC on the first calendar day of the month. Therefore, for a given month, considering the monthly and quarterly reporters only, it can in theory take up to 5 months for data from quarterly VAT returns to be received by ONS. This doesn't account for late returns.

Lack of timeliness is not an issue unique of VAT data, it similarly occurs on survey data. At the time of publication, RSI has a response rate of around 61%, accounting for 87% of sampled register turnover (ONS, 2017)[5]. In this case, ratio imputation is used to

[5]https://www.ons.gov.uk/businessindustryandtrade/retailindustry/methodologies/retailsalesindexrsiqmi

impute values for non-responders to minimise any potential non-response bias. Re-weighting is an alternative approach to dealing with non-response. This calculates weights by assuming that the achieved sample is the complete real sample.

The target populations for most ONS business surveys, including RSI and MBS, are based on the Inter-Departmental Business Register (IDBR). There are differences between the IDBR and VAT populations including:

- Businesses registered for PAYE but not VAT feature on the IDBR
- Businesses that have previously registered for VAT but have fallen below the de-registration threshold
- Births and deaths. The IDBR is fed from several sources, including VAT unit births and deaths, and there may be a delay in taking these changes on to the IDBR population
- Differences in unit definitions (see section 4.3)

Calibration estimation, in particular ratio estimation, is a technique used widely on sample surveys in ONS and internationally in other National Statistics Institutes (NSIs). The method calculates weights for each business in the sample, so that a known population total is reproduced. It requires an auxiliary variable that is well correlated with the target variable and is available for the entire population. For more information, see Särndal et al (2003).

MBS and RSI both use ratio estimation with the auxiliary variable of register turnover; an annual turnover value maintained on IDBR updated by several sources including VAT data and survey returns. Ratio estimation was one method tested as part of the ESSnet on Administrative data and was found to perform well on estimating variables from administrative data in comparison to survey estimates from the Annual Business Survey (Lewis, de Waal, 2011).

Some assumptions of ratio estimation under stratified simple random sampling include: the sample is a random sample, all businesses have a non-zero probability of selection, and the probabilities of selection for businesses within the same strata are equal. These assumptions are not necessarily true of VAT data.

1. Assumption 1: the sample is a random sample. The realised VAT sample comprises only those VAT units that have returned to date, which may not be random. Also, as the sample evolves between $t+1, t+2$ etc the main difference is which staggers become available. Staggers are initially allocated by HMRC, however businesses can change so there is an element of self-selection. If self-selection is related to the target variable and the calibration does not account for the difference between businesses in different staggers, then this can lead to bias in estimates if it is assumed that the sample is representative of the population at any point in time.
2. Assumption 2: all businesses have a non-zero probability of selection. This is not true for all businesses, including where there are differences between the populations. For

example, a business that is not registered for VAT but is in the population has a zero probability of submitting a VAT return.

3. Assumption 3: the probabilities of selection for businesses within the same strata are equal. This may not be true of VAT data, especially with stagger patterns, and if some businesses are more likely to response earlier or later than others.

To understand any potential impact of these assumptions, analysis was conducted of the evolution of estimates produced using ratio estimation at $t+1, t+2, \ldots, t+6$. The estimates were found to stabilise at $t+3$, however were generally determined unsuitable for use before.

The recommendation is to use ARIMA models to forecast the VAT series for $t+1$ and $t+2$ then use estimate from ratio estimation for $t+3$ onwards. The performance of the ARIMA forecasts and evolution of estimate from ratio estimation was investigated in the partition analysis presented in section 6. If they were not satisfactory then VAT data was not considered suitable for use.

4.3 Definitional differences

Both ONS and HMRC collect turnover data from businesses, however the definitions of turnover can differ. A list of items to include in and exclude from turnover reported to HMRC can be found at HMRC (2019). Exclusions and inclusions vary by survey and industry for MBS and RSI and are not reported here.

The analysis in section 6.2 gives the example of dispensing chemists where a notable definitional difference exists between the definition of VAT turnover and RSI turnover. VAT turnover includes prescriptions while RSI asks respondents to exclude prescriptions, leading to a level difference in the two series.

In addition to differences in definition of variables collected, there are differences in the units that data are collected for.

VAT returns are provided for a VAT unit while the statistical unit of interest is the reporting unit.

VAT units are linked to reporting units via enterprises. Over 90% of reporting units have a simple relationship, where one VAT unit is linked to one enterprise which is linked to one reporting unit. However, other relationships can arise. The relationship between VAT unit and enterprise can be one of: one-to-one, one-to-many, many-to-one or many-to-many. The relationship between enterprise and reporting unit can be either one-to-one or one-to-many. These structures are referred to as complex in all but the simplest case of a one-to-one-to-one VAT unit-enterprise-RU relationship.

While most reporting units are in simple structures, these businesses are typically small in terms of employment and turnover. It is the large businesses, which have a disproportionally large contribution to the total turnover of many industries, that are typically complex.

For producing estimates of turnover by industry, it is a requirement that turnover data are available at a reporting unit level. The process of converting data at VAT unit level to reporting unit level is called apportionment. Apportionment can mean either splitting out (in the case of one-to-many) or combining (many-to-one) VAT data together, or a combination of both (many-to-many), to provide estimated values for reporting units.

In the many-to-one and many-to-many cases the combining of VAT data presents an additional challenge when combined with timeliness when not all VAT data linked to a reporting unit for a period of interest are available.

VAT returns are apportioned first to enterprises, then to reporting units using proportions based on the employment headcount numbers.

The effect of apportionment is limited firstly by maintaining a survey for the largest businesses, which are more likely to be complex, and secondly by taking into consideration the number of complex businesses as part of the acceptance criteria in the partitioning analysis in section 6.

4.4 Periodicity

This section discusses challenges arising from differences in the periodicity of available data and the required output. The process of converting data available on one periodicity to another is hereafter referred to as calendarisation.

Monthly outputs are required to be given on a calendar month. Not all survey or VAT data will be available for a calendar month. Therefore, methods are required to use data available for different periods to produce calendar-month estimates.

On business surveys, if a business cannot provide data for the required reporting period then it is permitted to provide for a period of its choosing. If these data are used without accounting for the alternative reporting period then this can introduce a source of bias into the resulting output.

These returns are automatically edited. Because of large differences in activity on different days of the week, industry-specific trading day weights are assigned to each day of the week. The total of the weights in the standard reporting period is divided by the total of the weights in the period that the business has provided. This ratio is applied to the returned data to produce an adjusted value.

With VAT, businesses are permitted to provide HMRC with either monthly, quarterly or annual turnover data. A method is therefore required to produce monthly estimates using data on many different periodicities.

Previous research by ONS (Parkin, 2010) compared different combinations of methods for interpolation and extrapolation. Performance was measured by revisions to monthly estimates of levels and growth rates. For growth rates, the statistics of interest for the short-term output indicators, no method performed consistently best. In addition, they all demonstrated extreme poor performance in some instances on test data.

Research into the comparison between benchmarking and other more simplified methods have been considered, including a proportional allocation using seasonal factors and state space modelling. The research did not identify a consistently best performing method. The main conclusion was that a good indicator of the monthly path, such as the existing survey, is essential to any of these methods. The issue of calendarisation remains an open one and there is further research continuing by ESCoE (Labonne, 2018).

The method used on survey data, would provide the benefit of accounting for the different trading days in each month, but again, would not address seasonality. However, a modification to this approach, which accounts for the difference in seasonality between months can be proposed.

1. Using monthly data that are available, for example survey data from businesses in the same industry remaining in the survey, perform a time series decomposition to estimate the combined seasonal and trading day component.
2. Use this series to provide monthly weights, which are then used to calendarise quarterly and annual data.

If the seasonal pattern in the sampled population differs to that in the VAT population then this method will impact the estimates. The analysis in section 6 uses VAT data calendarised using the seasonal pattern from large businesses in the same industry assumed to remain in the sample. This analysis would identify any industries where this calendarisation method did not produce overall estimates of an acceptable quality.

5. Statistical Design

The proposed statistical design for DTrades is to split the population into mutually exclusive parts. One part will continue to be estimated for using a survey. The other will be estimated for using VAT data. The splitting of the population into two will be referred to as the partition.

Initially, the boundaries for partitioning will be based on the current sampling strata for the MBS and RSI surveys. Later it is proposed to refine these, for example to partition on complexity. The strata are based on industry classification and employment size-band. For RSI, the employment size-bands are consistent across industries. On MBS, the definitions of the size-bands vary. To provide an idea of the size-bands, the current RSI size-bands are provided in table 2.

Table 2: RSI existing survey size-bands

Size-band	Definition
1	0-4 employment
2	5-9 employment
3	10-99 employment and IDBR turnover <=£60m

4	100+ employment
5	10-99 employment and IDBR turnover > £60m

Both size-bands 4 and 5 are fully enumerated, meaning that all businesses in those size-bands are sampled, that is, taking a census. MBS also has 5 size-bands with size-bands 4 and 5 being full enumerated.

It has been assumed that all size-band 4 and 5 businesses will remain in the survey population. VAT data have only been considered for replacing size-bands 1 to 3 (i.e. the smaller sized businesses).

An extensive piece of partitioning analysis was undertaken to determine where aggregate VAT data was a suitable replacement for survey data estimates.

The decision on whether VAT data are a suitable replacement for bands 1 to 3 was based on:

1. The difference in the level estimates. VAT estimates were assessed against survey estimates and their 95% confidence intervals.
2. The growth in the estimates.
3. Proportion of complex VAT turnover in each size-band.
4. The size-band contribution to the industry turnover.
5. Size of revisions from forecasting VAT data. Any revision over 1% was investigated further.

Where VAT data were not suitable to replace survey data, cut-off sampling was considered as an alternative option to reduce the sample size and hence cost and burden. This method sets a threshold, below which business are not sampled. Larger businesses are then used to estimate for the cut-off population. This works well if the relationship between the variable of interest and auxiliary information available for the whole population, including the cut-off is the same above and below the threshold. Cut-off sampling was only considered for size-bands 1 and 2.

If neither VAT nor cut-off sampling was considered acceptable then the whole industry would continue to be sampled.

It is recommended to revisit these decisions periodically to ensure that any assumptions about relationships continue to hold, and when any changes are made to the data sources. Where the survey data are replaced it will not be possible to replicate this initial

work, however using data from another survey, or conducting a one-off quality assurance survey are options to explore.

6. Partitioning analysis

Pope and Brown (2018) present results of the partitioning analysis for divisions 45, 46 and 47. This section provides several examples where VAT data have and have not been currently recommended for use. The examples where it has not been recommended have been categorised by which of the challenges in section 4 have not been addressed. The partition will continually be placed under review and reassessed on a regular basis.

6.1 VAT data recommended for use

Figure 2 plots estimates of turnover for the motor-trade industry sale of motor vehicles. It provides estimate based on MBS with 95% confidence intervals. It then provides an estimate based on cut-off sampling where size-bands 1 and 2 are not sampled. The third estimate uses VAT data to estimate for size-bands 1 and 2 and survey for bands 3-5.

Generally, all three estimates are very similar. The confidence intervals are very narrow as this industry is dominated by the large businesses, which are fully enumerated.

Figure 2: Example monthly turnover for sale of motor vehicles based on current MBS estimate, cut-off sample for size-bands 1-2, and a VAT hybrid series with VAT data used for size-bands 1-2 and MBS for size-bands 3-5

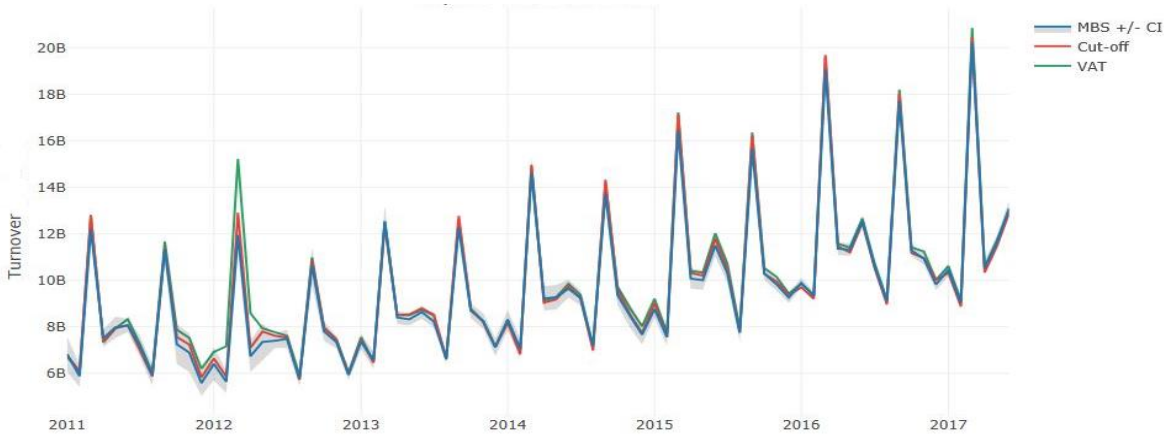
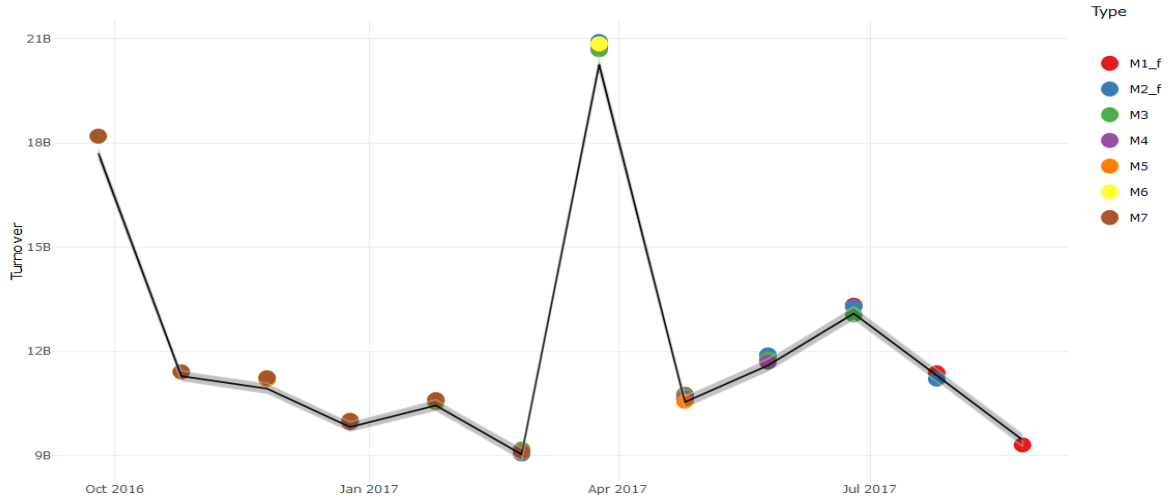


Figure 2 is based on mature VAT data. It was important to assess the scale of revisions to the estimates as the VAT estimation method changed from forecasting to ratio estimation, and as more VAT data became available.

Figure 3 shows how the estimates for the VAT hybrid series evolve for the period October 2016 to September 2017. The black line indicates the MBS series. The dots are the estimates at different points in time. There is very little variation between the estimates as more VAT data become available.

Figure 3: Example revisions to hybrid VAT estimate for sale of motor vehicles.



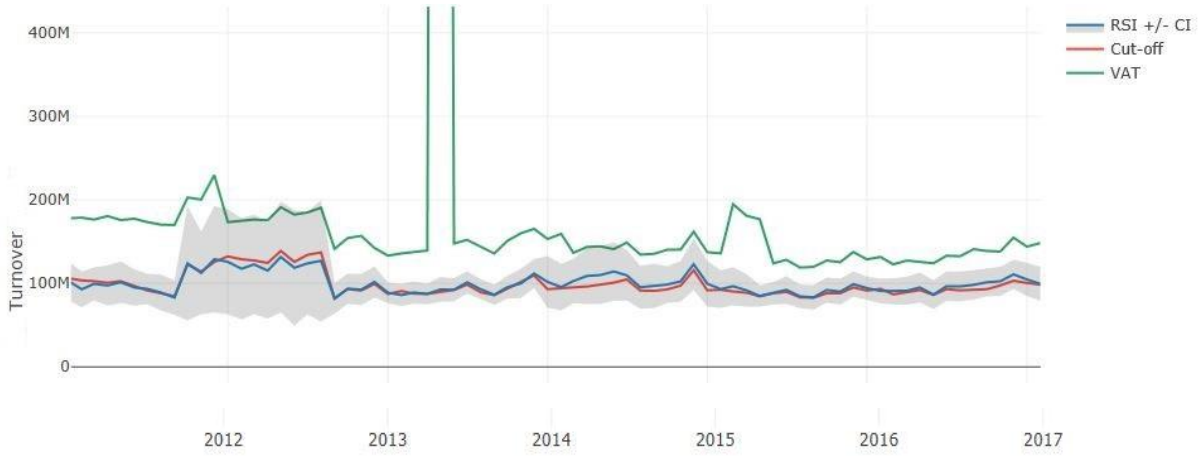
For this industry, it would be appropriate to replace size-bands 1 and 2 with VAT data. Further analysis was conducted to assess the possibility of using VAT data for size-band 3, however at this stage it was not deemed acceptable quality.

6.2 Definitional Difference

Figure 4 presents estimates of total retail turnover for dispensing chemists. The figure compares the current RSI estimate with 95% confidence intervals, an estimate using cut-off sampling for size-bands 1 and 2, and an estimate using VAT data for size-bands 1 and 2 and survey for bands 3-5.

As noted in section 4.1, the VAT data used in this analysis had not been fully cleaned, and there are several periods in 2013 in the VAT data affected by an error. Excluding this error, the VAT series is consistently higher than both the survey and cut-off estimates. This is an example of a definitional difference where use of survey data may be more appropriate. As detailed in section 4.3, one difference between RSI and VAT data for chemists is the treatment of prescriptions; they are included in VAT data but excluded from RSI.

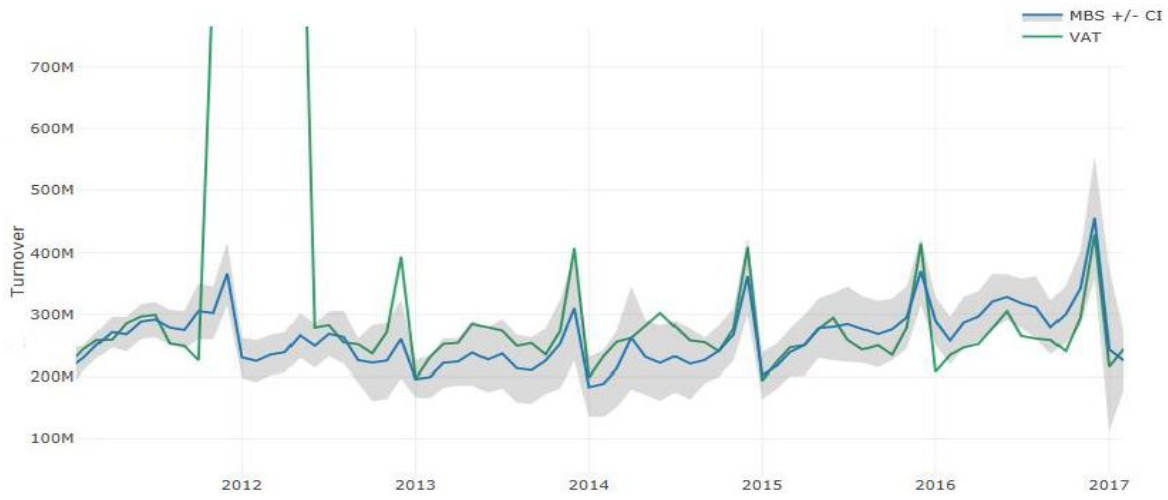
Figure 4: Example monthly retail turnover for dispensing chemists based on current MBS estimate, cut-off sample for size-bands 1-2, and a VAT hybrid series with VAT data used for size-bands 1-2 and MBS for size-bands 3-5



6.3 Seasonality

An example where calendarisation is having an impact is for retail sale of alcoholic drinks, beverages and tobacco. Figure 5 compares two estimates of retail turnover of alcoholic drinks, beverages and tobacco from RSI data, the second with RSI data for size-bands 4 and 5, and VAT data for size-bands 1-3 replaced with VAT data. Again, there is an error in the VAT series affecting data at the end of 2011 and beginning of 2012. However, there are differences in the seasonality. The VAT series has a consistent peak in December, while the December peak has been increasing in the survey data. The VAT series also has a different pattern in the rest of the year, with a more pronounced peak in June. The different seasonal patterns will be captured as part of any update to the seasonal adjustment parameters and settings.

Figure 5: Example retail sales of alcoholic drink, beverages, and tobacco. Turnover estimates from MBS and using VAT for size-bands 1-3 and MBS for size-band 4-5.

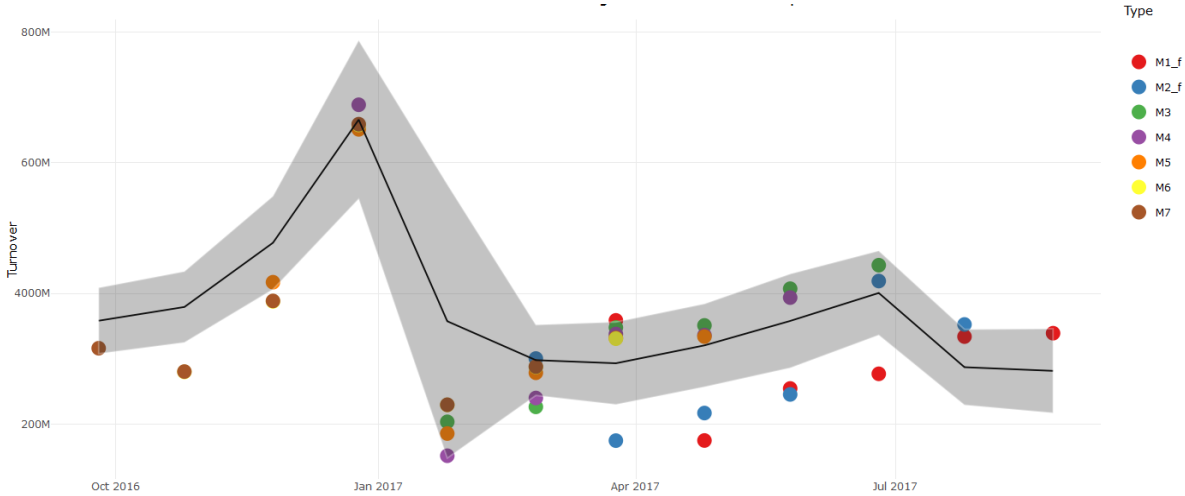


6.4 Timeliness and missing data

Additionally, retail sales of alcoholic drinks, beverages and tobacco proved challenging for estimation. Figure 6 shows how the estimates of turnover for October 2016 to September 2017 evolve as more data become available. The estimates in the first two months are produced using ARIMA models to forecast the VAT data; denoted as M1_f and M2_f. These are the red and blue dots. All remaining months are denoted as M3 to M7. The M3 to M7 points are estimates produced using VAT data and ratio estimation. These evolve as more data are received between 3 and 7 months after the reference period.

The M1 and M2 forecasts consistently underestimate the later estimates from ratio estimation. They also regularly fall out of the confidence intervals around the original RSI estimate (black line). There are also some large revisions to the ratio estimates as more data become available, for example in March 2017 the M3 estimate is outside the confidence interval for the survey data but by M7 falls closer to the RSI estimate.

Figure 6: Example revisions to hybrid VAT estimate for retail sale of alcoholic drink, beverages and tobacco.



7. Conclusion

This paper has provided a discussion of the challenges of using VAT data for short-term turnover statistics. The challenges discussed were definitional differences, error detection and correction, missing data and timeliness, and periodicity. Methods to address these problems have been proposed and used on VAT data to produce test data. These test data were combined with survey data to produce outputs as-if the population were partitioned into a part estimated for using a survey and the other using VAT data to inform areas for potential use of VAT turnover data.

To increase the use of VAT data in the short-term output indicators there are several challenges and pieces of further research that should be undertaken. These are:

- Consider alternative methods for apportionment, or consider a partition based on using VAT data for simple units and survey for complex.
- Investigate the use of modelling where there are definitional differences between the survey and VAT data.
- Machine learning techniques for cleaning could be investigated as alternatives.
- Continue research into calendarisation methods.
- Test changing the order of estimation and calendarisation.

References

[1] Davies J (2017a) *RSI Transformation Discovery: Exploring the use of benchmarking for calendarisation of VAT data*, ONS internal paper

[2] Davies J (2017b) *State space modelling for calendarisation of VAT data*, ONS internal paper

[3] Davies K (2018) *Investigating methods of efficient detection of errors in VAT data*, UNECE Conference of European Statisticians https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T3_UK_DAVIES_Paper.pdf

[4] ESSnet (2011) *Use of Administrative and Accounts Data in Business Statistics WP2 Reference document Section 3: "Checking for Errors and Cleaning the Incoming data"*, https://ec.europa.eu/eurostat/cros/system/files/SGA%202011_Deliverable_2.3_1.pdf

[5] HMRC (2019) *Guidance: How to fill in and submit your VAT Return (VAT Notice 700/12)* https://www.gov.uk/guidance/how-to-fill-in-and-submit-your-vat-return-vat-notice-70012

[6] Labonne P and Weale M (2018) *Methods for temporal disaggregation of rolling quarterly VAT-based turnover*, https://www.escoe.ac.uk/wp-content/uploads/2018/06/EM2018-Labonne-and-Weale.pdf

[7] Lewis D, de Waal T (2011) *Deliverable 3.4: Guide to estimation methods,* ESSnet: Use of administrative and accounts data in business statistics

[8] ONS (2017a) *Retail Sales Index (RSI) QMI* https://www.ons.gov.uk/businessindustryandtrade/retailindustry/methodologies/retailsalesindexrsiqmi

[9] ONS (2017b) *Transforming short-term turnover statistics: October 2017* https://www.ons.gov.uk/businessindustryandtrade/retailindustry/articles/transformingshorttermturnoverstatisticsoctober2017/2017-10-19

[10] Parkin N (2010) *Interpolation and Extrapolation from Value Added Tax Returns*, ONS internal paper

[11] Pope M, Brown D (2018) *Distributive Trade – Partitioning*, ONS internal paper

[12] Särndal CE, Swensson B, Wretman, J (2003) *Model assisted survey sampling*, Springer Science & Business Media.

[13] Skentelbery R, Finselbach H, Dobbins C (2011) *Improving the efficiency of editing for ONS business surveys*, UNECE Conference of European Statisticians

[14] UKSA (2015) *Better Statistics, Better Decisions* https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/Better-Statistics-Better-Decisions.pdf

Forthcoming Courses

GSS Statistical Training Programme

A series of government specific short courses (between 0.5 and 2 days in length) delivered by methodological experts in the field. These courses are delivered at ONS sites in London, Newport and Titchfield.

For further information on learning and development offered by the GSS see the link below

https://gss.civilservice.gov.uk/learning-development/

or contact gss.capability@ons.gov.uk

Details of specific courses can be found here

https://gss.civilservice.gov.uk/training-courses/

Details of additional opportunities for learning can also be found in in the training events page. In summary these are:

MSc in Data Analytics for Government

This is available at the following universities: University College London, Oxford Brookes University and Southampton University. More details can be found via this link.

https://gss.civilservice.gov.uk/learning-development/the-msc-in-data-analytics-for-government/

Details on the modules offered by each MDataGov provider can be accessed from the same link.

European Statistical Training Programme 2019

The purpose of the European Statistical Training Programme (ESTP) is to provide statisticians the opportunity to participate in international training courses, workshops and seminars at postgraduate level. It comprises courses in Official Statistics, IT applications, Research and Development and Statistical Management. More information on the core program for 2019 can be found on the Eurostat website

https://ec.europa.eu/eurostat/documents/747709/6103606/2019-ESTP-catalogue-final.pdf

Methodology Advisory Service (MAS)

The Methodology Advisory Service is a service of the Office for National Statistics (ONS); it aims to spread best practice and improve quality across official statistics through methodological work and training activity. The ONS has about one hundred methodologists - highly qualified statisticians and researchers; their primary role is to provide expert support, advice and methodological leadership to the ONS in producing and analysing National Statistics.

Methodology staff are arranged into Centres of Expertise, each comprising a team of specialists who keep abreast of research and developments in their area of expertise through contacts with academia, other national statistical institutes and the wider research community. Many of these Centres have international reputations and present research and applied work at conferences and at other meetings of experts in their fields. Examples of these centres are Sample Design and Estimation and Time Series Analysis.

The Methodology Advisory Service has a remit to extend the services of ONS methodologists beyond ONS into other public sector organisations. Every year, MAS carries out projects with customers addressing a wide range of statistical requirements. As well as calling on methodology staff, MAS can also draw on the wider expertise of statisticians, researchers and subject area specialists across the ONS. Further expertise is available though links with Universities.

Contact ***MAS@ons.gov.uk***

GSS Methodology Series

Latest reports in the GSS Methodology Series:

38. *100 Years of the Census of Production in the UK, Paul Smith*
39. *Quality of the 2010 Electoral Register in England & Wales, Neil Hopper*
40. *Modelling sample data from smart-type electricity meters to assess potential within Official Statistics, Susan Williams and Karen Gask*
41. *Using geolocated Twitter traces to infer residence and mobility, Nigel Swier, Bence Komarniczky and Ben Clapperton*
42. *Assessing the Generalised Structure Preserving Estimator (GSPREE) for Local Authority Population Estimates by Ethnic Group in England, Solange Correa-Onel, Alison Whitworth and Kirsten Piller*

Reports are available from:

https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/currentmethodologyarticles

Enquiries

We aim to publish the Survey Methodology Bulletin twice a year, in Spring and Autumn. Copies of many previous editions are available electronically at:

http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/survey-methodology-bulletin/index.html

If you would like to be added to the distribution list please email ONS Methodology at:

methodology@ons.gsi.gov.uk

Or write to us at:

Philip Lowthian
Survey Methodology Bulletin
2nd Floor
Office for National Statistics
Drummond Gate
London
SW1V 2QQ

ons.gov.uk
visual.ons.gov.uk