The use of administrative and accounts data for economic statistics - Introduction

Daniel Lewis
Office for National Statistics
daniel.lewis@ons.gsi.gov.uk
Introduction

• Growing topic in official statistics
• Use of admin data not new – important source of auxiliary information
• Recent objective to make greater use for economic statistics
• Largely driven by reduction in budgets and desire to reduce burden on businesses
• Statistical theory still being developed
• European Statistical System network project (ESSnet Admin Data) carried out large body of work
  • http://essnet.admindata.eu
Schedule

2. Admin data sources for economic statistics
3. Uses of admin data for economic statistics
4. Integration of data from administrative sources
5. Detection and treatment of suspicious values in admin data
6. Sample designs for statistics based on admin data
7. Estimation of Structural Business Statistics using admin data
8. Estimation of Short Term business Statistics using admin data
9. Quality measurement for statistics using admin data
10. Summary
Administrative data sources for economic statistics

Daniel Lewis
Office for National Statistics
daniel.lewis@ons.gsi.gov.uk
Introduction

• Before looking at uses and methods, we examine main sources of admin data for economic statistics

• Availability and legal access differs between countries

• Discuss three key sources
  • Value Added Tax
  • Social Security
  • Company Accounts

• Focus on admin data that can be useful for Structural Business Statistics (SBS) and Short Term business Statistics (STS)
Value Added Tax (VAT)

- Widely available to statistics offices
- Focus of significant recent research
- Key variable is Turnover
- Other variables may be available, e.g. Expenditure
- Majority of businesses provide quarterly, although some monthly (in most countries...)
- VAT Turnover commonly used to update business register and for survey activities such as calibration and imputation
Social Security data

• Key source of Employment data for statistics offices
• Available either monthly or quarterly
• Also used to update business register and for related survey activities
• Potential further uses of Employment for economic statistics mirror those of Turnover
Company accounts data

- Businesses are required to complete annual company accounts
- Often contain a wide range of financial variables – depending on size of business
- Big potential for use in economic statistics – already happening in some countries
- Most useful information is often only available in the notes – not easy to process electronically
- Accounting definitions do not always match well with statistical definitions...
Sources used directly for SBS (across ESS)

- Annexes I-IV:
  - Social security (employment variables)
  - VAT (turnover variables)
  - Published business accounts
  - Corporate tax (wide range of variables)
  - Personal income tax
  - Chambers of commerce / professional associations
Sources used directly for SBS (across ESS)

- Annex V (insurance):
  - Regulatory authorities
  - Central bank
  - Published business accounts
  - Social security (subset of variables)
  - Corporate tax
Sources used directly for SBS (across ESS)

- Annex VI (credit institutions):
  - Central bank (main admin source for most countries)
  - Regulatory authorities
  - Corporate tax
Sources used directly for SBS (across ESS)

- Annex VII (pension funds):
  - Regulatory authorities

- Annex VIII (business services):
  - Social security (employment)
  - VAT (turnover)
  - Published business accounts (turnover)
  - Corporate tax (for turnover variables, estimation for rest)
  - Chamber of commerce / professional associations
Sources used directly for SBS (across ESS)

• Annex IX (business demography):
  • Personal income tax (turnover variables)
  • Social security (employment variables)
  • VAT (turnover variables)
  • Corporate tax (wide range of variables)
Sources used directly for STS (across ESS)

• Turnover variables:
  • VAT
  • Regulatory authorities
  • Customs and excise
Sources used directly for STS (across ESS)

- Employment:
  - Social security
  - Personal income tax

- Production prices / costs:
  - Customs and excise
  - Local authorities
  - Ministries and other governmental bodies
  - Regulatory authorities
Administrative data sources for economic statistics

Daniel Lewis
Office for National Statistics
daniel.lewis@ons.gsi.gov.uk
Introduction

• Before looking at uses and methods, we examine main sources of admin data for economic statistics
• Availability and legal access differs between countries
• Discuss three key sources
  • Value Added Tax
  • Social Security
  • Company Accounts
• Focus on admin data that can be useful for Structural Business Statistics (SBS) and Short Term business Statistics (STS)
Value Added Tax (VAT)

- Widely available to statistics offices
- Focus of significant recent research
- Key variable is Turnover
- Other variables may be available, e.g. Expenditure
- Majority of businesses provide quarterly, although some monthly (in most countries...)
- VAT Turnover commonly used to update business register and for survey activities such as calibration and imputation
Social Security data

• Key source of Employment data for statistics offices
• Available either monthly or quarterly
• Also used to update business register and for related survey activities
• Potential further uses of Employment for economic statistics mirror those of Turnover
Company accounts data

- Businesses are required to complete annual company accounts
- Often contain a wide range of financial variables – depending on size of business
- Big potential for use in economic statistics – already happening in some countries
- Most useful information is often only available in the notes – not easy to process electronically
- Accounting definitions do not always match well with statistical definitions...
Sources used directly for SBS (across ESS)

• Annexes I-IV:
  • Social security (employment variables)
  • VAT (turnover variables)
  • Published business accounts
  • Corporate tax (wide range of variables)
  • Personal income tax
  • Chambers of commerce / professional associations
Sources used directly for SBS (across ESS)

- Annex V (insurance):
  - Regulatory authorities
  - Central bank
  - Published business accounts
  - Social security (subset of variables)
  - Corporate tax
Sources used directly for SBS (across ESS)

- Annex VI (credit institutions):
  - Central bank (main admin source for most countries)
  - Regulatory authorities
  - Corporate tax
Sources used directly for SBS (across ESS)

• Annex VII (pension funds):
  • Regulatory authorities

• Annex VIII (business services):
  • Social security (employment)
  • VAT (turnover)
  • Published business accounts (turnover)
  • Corporate tax (for turnover variables, estimation for rest)
  • Chamber of commerce / professional associations
Sources used directly for SBS (across ESS)

- Annex IX (business demography):
  - Personal income tax (turnover variables)
  - Social security (employment variables)
  - VAT (turnover variables)
  - Corporate tax (wide range of variables)
Sources used directly for STS (across ESS)

- Turnover variables:
  - VAT
  - Regulatory authorities
  - Customs and excise
Sources used directly for STS (across ESS)

- **Employment:**
  - Social security
  - Personal income tax

- **Production prices / costs:**
  - Customs and excise
  - Local authorities
  - Ministries and other governmental bodies
  - Regulatory authorities
Integration of data from administrative sources

Daniel Lewis
Office for National Statistics
daniel.lewis@ons.gsi.gov.uk
Introduction

• To construct statistics from admin data, usually need to integrate them with statistical data in some way
• Even when admin data cover the required population completely, statistical data may be needed to ensure the data can be aggregated to the correct output domains
• Matching can often be done via the business register
• Integration is not always straightforward – we discuss issues with linking, coverage and definitions
Linking admin data to statistical data

• Due to historical use of admin data for the business register, it is common to find a unique identifier between admin and survey sources

• However, differences in reporting structures often mean that there is not a one-to-one match between sources

• Four types of links may be found...
Links between admin data and statistical data

1. One admin data unit links to one statistical unit
2. One admin data unit links to multiple statistical units
3. Multiple admin data units link to one statistical unit
4. Multiple admin data units link to multiple statistical units

- When the link is to one statistical unit, it is straightforward to identify appropriate values in the admin data set
- It is more problematic if the link is to multiple statistical units...
Linking to multiple statistical units

Without contacting the business for further information, the only option is to approximate the split between units.

UK tested methods for doing this with VAT data:

- Equal split amongst units
- Proportional to Employment
- Using ratio of Turnover / Employment

Turnover and Employment values came from business register.

Evaluated by comparing to SBS survey data.

Best method was Proportional to Employment.
Linking to the correct period

• When matching admin data, also important to ensure data correspond to required periods
• May need to manipulate admin data in some way to achieve this
• Relatively straightforward to sum sub-annual data to create an annual total
• Splitting data from reporting periods which span across those required for outputs is more difficult...
• One real life example is using quarterly VAT data to produce monthly outputs!
• Possible solutions are discussed later in the course
Coverage of admin data

- Admin data populations are not generally the same as statistical populations
- There are some units in the admin data that are not required for statistics (overcoverage)
- This can be dealt with if it is possible to identify them
- Undercoverage can be more problematic
- If a full dataset is required, imputation is necessary
- For aggregates, estimation techniques can be used
Coverage of admin data

Survey population

MATCHED PART

Administrative data
Differences in definitions

- Definitions used for admin data variables are often different from those required for statistics.
- For example, the treatment of tax in Turnover definitions can vary between sources.
- This problem often needs to be overcome to use admin data for economic statistics.
- First step is to understand the admin data better by talking to the admin data holder.
- In some cases the differences may be small enough to be acceptable.
- If bigger differences are understood fully, it may be possible to correct them either directly or through modelling.
- Modelling techniques are discussed later in the course...
Detection and treatment of suspicious values in admin data

Daniel Lewis
Office for National Statistics
daniel.lewis@ons.gsi.gov.uk
Introduction

- Admin data often assumed to be free of errors as they are collected for important official reasons.
- In practice errors exist and need to be identified and treated in some way.
- Admin data holder may have done some cleaning.
- Traditional survey data editing methods are not always appropriate for admin data.
  - Can’t re-contact businesses.
  - Effect on outputs less clear.
- We discuss methods for detecting and correcting suspicious values in admin data.
- Examples for VAT and Company accounts data.
Key principles for editing admin data

1. Maintain original data whenever possible
2. Methods should take account of all uses of data
3. Make good use of historical and auxiliary data
4. Automate detection and correction process where possible
5. Keep audit trail for any changes to admin data
6. Methods should be flexible for future improvements
7. Important to understand how admin data are returned and processed
Types of error found in admin data

• **Systematic errors**
  - Generated by well known causes, e.g. “unit errors” and “scanning errors”
  - Usually possible to identify and treat efficiently

• **Random errors**
  - No systematic cause of errors, e.g. may be due to respondents making mistake
  - Require more sophisticated detection and treatment methods
General methods for identifying errors

- Check for invalid and inconsistent data
  - Requires identification of valid combinations
  - Inconsistencies within source and by comparison to others (e.g. business register)
- Duplicates and missing data
- Negative values
- Balancing errors

- Plausibility checking of aggregate data is always useful
  - Graphically
  - Summary statistics
General methods for identifying errors

- Check for any suspicious patterns in data
- Identification of unusual values – “outliers”
- Check with historic / time series data
- Comparison with other sources
- The admin data holder may be aware of systematic problems with the dataset, which can help in identifying errors
- Some standard editing techniques may be useful – testing is important. Examples later...
Treatment of suspicious values in admin data

• Choice in how to deal with suspicious values
  • Ignore
  • Remove from data set
  • Flag (potentially allow differential treatment for different uses)
  • Query with admin data holder / business
  • Manual imputation
  • Automatic imputation

• Choice depends on time, resource, quality requirements and legal provision for querying data
Case study 1: Methods for cleaning VAT data

- Case study using UK data to develop and test methods
- Identification of different types of error
- How to deal with each of those errors once identified
- Evaluation of different options to choose preferred method
- Implementation
UK VAT data

• Available from tax office – monthly updates
• 90% of businesses respond quarterly, most of the rest monthly
• Two main variables
  • Turnover
  • Expenditure
• Tested using two calendar years of VAT data
• Methods tested drawn from mainstream and admin data specific literature
Unit errors

• Common in survey data where businesses are often asked to report in thousands of Euros
• Also present in VAT data
• In UK, businesses should report VAT in pounds
• Some businesses report in thousands, by habit
• Easy to correct, if they can be identified
• Most effective identification by comparison to previous data (need to be careful if only have two periods...)
• Also possible to compare to other source – e.g. business register
Unit errors

- Identify unit errors using formula of the type

\[ A < \frac{\text{current VAT data}}{\text{previous VAT data}} < B \]

- In UK, following values chosen based on experience with surveys

\[ 0.00065 < \frac{\text{current VAT data}}{\text{previous VAT data}} < 0.00135 \]

- Values in this range can then be automatically corrected by multiplying by 1000
Scanning errors

- May occur during data entry
- Often can be identified as highly implausible value
- System may use special characters to denote scanning error
  - For UK VAT data, they are set to 99999999999
  - Doesn’t necessary capture every scanning error...
- Simple to identify pre-set code
- May need additional edits for other scanning errors
- Clearly not appropriate to use those values in production of statistics...
Suspicious quarterly data patterns

- By analysing four consecutive quarters, the following patterns are sometimes observed:
  1. Zero value in one quarter, positive in other three
  2. Zero value in three quarters, positive in other one
  3. Exactly the same value in all four quarters
  4. Same value for three quarters, different for fourth
  5. Negative VAT data in any of the quarters
Suspicious quarterly data patterns

• All suggest the possibility of erroneous data
• Zero value in one quarter suggests data may have been withheld for one period
• Patterns 2, 3, 4 suggest an annual figure has been provided rather than quarterly
• Negative VAT values are impossible...
• Fairly simple to identify patterns, how to deal with them?
Suspicious quarterly data patterns - correction

• Where one period looks incorrect, impute using preferred method for a single erroneous or missing value (see later)

• Negative values should not be treated by simply reversing the sign! Instead impute new value

• Where it is suspected an annual figure has been provided, re-apportion the annual figure amongst the quarters
Suspicious quarterly data patterns - correction

- Quarterly proportions:
  - (Trimmed) Mean proportion from similar businesses
  - Median proportion from similar businesses
  - Proportions for same business from previous year
  - Proportions for same business from alternative source
Suspicious quarterly data patterns - correction

- Tested methods by randomly creating suspicious patterns in clean data and comparing imputations to true values
- Median proportion from similar businesses was found to be most accurate
Methods for detecting random errors

1. Quartile distances
2. Period on period ratios
3. Comparison with business reporting history
4. Quartile distance with measure of influence
5. Hidiroglou-Berthelot method
6. Period on period ratios with measure of influence
Methods for detecting random errors

1. Quartile distances

If VAT return $> Q3 + [C \times (Q3 – Median)]$

or VAT return $< Q1 – [C \times (Median – Q1)]$

then treat as suspicious

- Quartiles and median calculated within industry and size class
- $C$ derived from analysing past data to find threshold that identifies extreme values
- $C$ may vary by class
Methods for detecting random errors

2. Period on period ratios (within class)
   • For each business, calculate
     
     \[
     \text{Score} = \frac{\text{VAT return}}{\text{Median VAT return in class}}
     \]

     \[
     \text{TestRatio} = \begin{cases} 
     \frac{\text{Score}_t}{\text{Score}_{t-1}} & \text{if } \text{Score}_t > \text{Score}_{t-1} \\
     \frac{\text{Score}_{t-1}}{\text{Score}_t} & \text{otherwise.}
     \end{cases}
     \]

   • For periods \( t, t-1 \)
   • Any business with test ratio above a pre-defined threshold is treated as suspicious
Methods for detecting random errors

3. Comparison with business reporting history

If VAT return > € 100 million
And VAT return > 10 \times \text{mean VAT return for the business in the past 24 months},
then treat as suspicious.

- Can enhance method by varying thresholds by class
Methods for detecting random errors

4. Quartile distance with measure of influence
   • Refinement to method 1
   • Identify suspicious businesses using quartile distances
   • Then calculate influence for each business

\[
\text{Influence} = \frac{\text{VAT value}}{\text{Total VAT in class}}
\]

• Treat businesses with high influence and outside quartile range as suspicious
Methods for detecting random errors

5. Hidiroglou-Berthelot method

\[ r = \frac{\text{current VAT value}}{\text{previous VAT value}} \]

- Calculate median of ratios, \( r \)
- If \( r < \text{median} \) then \( t = \frac{r - \text{median}}{r} \)
- Otherwise \( t = \frac{r - \text{median}}{\text{median}} \)
- Define

\[ E = t \times \max (\text{current VAT Turnover}, \text{previous VAT Turnover})^y \]
5. Hidiroglou-Berthelot method

- Calculate Q1, Q3 and median of E values
- Then
  
  \[
  \text{If } E < Q2 - C \times (Q2 - Q1) \\
  \text{Or } E > Q2 + C \times (Q3 - Q2)
  \]

  treat as suspicious.

- C is set to identify appropriate number of suspicious businesses
6. Period on period ratios with measure of influence

- Refinement to method 2, using the measure of influence

\[
\text{Influence} = \frac{\text{VAT value}}{\text{Total VAT in class}}
\]

- Treat businesses with high influence and test ratio over threshold as suspicious
Evaluation of detection methods

- Usually not possible to re-contact business to confirm suspicious values
- Generally not possible to find the truth – this makes evaluation of methods difficult
- Diagnostics may be helpful
  - Proportion of suspicious businesses identified
  - Average size of businesses identified as suspicious
  - Average VAT values of suspicious businesses
- May be possible to compare with independent source
  - Careful of differences in definition
Evaluation of detection methods

• Based on the criteria above, study found that best performing methods were
  2. Period on period ratios
  4. Hidiroglou-Berthelot
  6. Period on period ratios with influence measure

• Note that all of these methods rely on comparison with previous period data
Correcting random errors in VAT data

• If a complete data set is desired, imputation is often the only option for “correcting” random errors
• A range of imputation methods were tested

1. Mean imputation
2. Trimmed mean imputation
3. Median imputation
4. Ratio imputation
5. Nearest neighbour donor imputation
Imputation methods for VAT data

1. Mean imputation

   • Replace suspicious value with mean of VAT data values from all non-suspicious businesses
   • Accuracy may be improved by calculating means within homogeneous classes – e.g. industry by employment size
2. Trimmed mean imputation

- Means are influenced by outliers, so often makes sense to trim extreme values
- e.g. Remove highest and lowest 10% of values before calculating the mean
3. Median imputation

- Replace suspicious value with median of VAT data values from non-suspicious businesses (within class)
- Median is robust to outliers
Imputation methods for VAT data

4. Ratio imputation

\[
\text{Previous value for business } \times \frac{\sum_{i \in \text{class}} \text{current period value for business } i}{\sum_{i \in \text{class}} \text{previous period value for business } i}
\]

- Only include businesses with (non-suspicious) data for both periods
- Ratio of mean values, reduces to ratio of sums
5. Nearest neighbour donor imputation

- Use VAT value from donor business
- Donor is business (in same class) that is most similar to suspicious business
- Identify similar businesses by calculating distances between suspicious business and potential donors
- Distances calculated for specified auxiliary variables – e.g. Turnover, Employment, industry
- Donor chosen has smallest combined distance
Evaluation of imputation methods

• Use “clean” data set and artificially contaminate with suspicious values
• Replace suspicious values with imputations, using each of methods 1 to 5
• Compare resulting data sets to “true” values to evaluate performance of methods
• Iterate process many times over (e.g. 200), randomly creating different suspicious values each time
• Estimate bias from simulation
Evaluation of imputation methods

- Relative Imputation Bias (RIB)
  \[
  RIB = \frac{\sum_{\text{Population}} \text{(Imputed VAT value - Original VAT return)}}{\sum_{\text{Population}} \text{Original VAT return}}
  \]

- Relative Imputation Error (RIE)
  \[
  RIE = \frac{\sum_{\text{Population}} \left| \text{Imputed VAT value - Original VAT return} \right|}{\sum_{\text{Population}} \text{Original VAT return}}
  \]
Evaluation of imputation methods

• In the study, ratio imputation was found to be the most effective imputation method for VAT data
Case study 2: Cleaning Company accounts data

- Company accounts variables sometimes have similar terminology as statistical variables but do not necessarily share the same concepts and may use different valuation methods.

- Errors may be committed
  1. At the time of preparing the accounts
  2. During the conceptual adjustment process
Errors caused when preparing accounts

• Errors of omission
  • Information left out of accounting statement by mistake or recorded in the wrong place

• Errors of commission
  • Data are recorded or calculated inaccurately – e.g. adding variables instead of subtracting
  • Use of software has reduced such errors

• Errors of principle
  • Basic accounting principles applied improperly, resulting in errors within a form or across forms – e.g. variables may not agree between Balance sheet and Profit & Loss accounts
Errors caused when preparing accounts

- **Involuntary data entry errors**
  - Entering the incorrect number or sign, typographic errors
- **Voluntary data entry errors**
  - Respondents may not have access to required data, so they invent a “typical” value
  - Can be very difficult to identify
- **Unit errors**
  - e.g. thousand Euros error
- **Zero values**
  - Variables may not be required to be included in accounts, but may have positive value in reality
Errors introduced during conceptual adjustment

• Adjustment may be required to be able to use accounts data for statistical purposes
• This adjustment can introduce additional errors
• Matching complex businesses across accounting and statistical units leads to inaccuracies
• Accounting period may differ from that required for statistics
  • Many businesses report accounts at different dates to end of calendar year
  • May need to combine data from multiple accounting years to deal with this
  • Problematic for volatile variables (e.g. Inventories)
Detecting and correcting errors in company accounts

- First step should be to focus on identifying and dealing with obvious and very large errors.
- Following that, consistency edits are useful in identifying errors.
- For example:
  - Total Assets = Total Liabilities + Total Shareholders' Equity
  - Total Revenue – Total Expense = Net Income after Adjustments
  - Sum of components within section = Section Total
  - Gross Profit = Turnover – Costs of Sales
  - Gross Profit – Expenses = Net Profit/Loss
Detecting and correcting errors in company accounts

- **Turnover - Expenses = Net Profit/Loss**
- **Operating Profit/Loss = Gross Profit/Loss – Admin Expenses**
- **Opening stock + purchases + Other direct costs = Costs of Sales + Closing stock**
- **Fixed assets = Tangible fixed assets + Intangible fixed assets**
- **Current assets = Stocks + debtors + Cash in Bank and in Hand**

- In practice hundreds of logic checks of this type may be used
Sample designs for statistics based on administrative data

Daniel Lewis
Office for National Statistics
daniel.lewis@ons.gsi.gov.uk
Introduction

- Use of admin data for economic statistics gives rise to specific challenges in sample design
- Aim to maximise use of admin data whilst keeping quality of resulting outputs at required level

1. Use solely admin data to cover whole population
2. Use admin data for whole population, with survey-based adjustments to deal with imperfections
3. Use combination of admin and survey data to cover population
4. Use survey to cover whole population, but utilise admin data to improve efficiency
1. Solely admin data for whole population

- This is the ideal scenario
- Need admin data to match very well with statistical requirements
- Admin data needs to cover whole population, contain all required variables
- Then simply sum admin data to produce aggregates
- Unfortunately, rare situation in practice...
2. Admin data with adjustments

- Even where admin data covers all (or most) of the statistical population, likely to need some adjustment to ensure quality

- e.g. to deal with:
  - Differences in definition
  - Full data set not being available on time
  - Missingness
  - Undercoverage
  - Imperfections in the data
2. Admin data with adjustments

- In some cases may be possible to adjust using only current or past admin data
- Often, survey data is required to make adjustments
  - May utilise existing survey
  - Might need to commission a special survey
  - Where admin data are being used to replace an existing survey, one option is to reduce sample size and use remaining survey data to adjust
- Could be simple aggregate level adjustment or more complex modelling techniques
- More on estimation methods in following chapters...
2. Admin data with adjustments

- Choice of sample design for survey element driven by amount of sample required in each domain to accurately adjust admin data
- Sample size dependent on how far admin data are from statistical requirements
- Sample should be set by analysis of historic data
3. Combination of admin and survey data

- Most common use of admin data for economic statistics
- Small survey usually needed to meet quality requirements
- Wide range of options for combining admin and survey data
- Usually essential to maintain (fully enumerated) survey of largest businesses
  - Importance in statistics
  - Complex nature makes admin data impossible to use
3. Combination of admin and survey data

- Option (i): Survey for largest, admin data for rest
3. Combination of admin and survey data

• Option (i): Survey for largest, admin data for rest
• This design works well if admin data are of sufficient quality
• Note that quality requirements of admin data are now less strict since the most important businesses are surveyed
• If this option does not work, may need to sample some of smaller businesses to adjust admin data, either at overall industry level or by size class
3. Combination of admin and survey data

- Option (ii): Survey adjusts admin data, industry level
3. Combination of admin and survey data

- Option (iii): Survey adjusts admin data by size class
3. Combination of admin and survey data

- Finally, admin data may be used to replace all of the very smallest businesses
- Combination of survey and admin data used for medium businesses
- Sometimes known as “cut-off sampling” – see later...
3. Combination of admin and survey data

- Option (iv): Cut-off sampling
4. Survey whole population, utilise admin data

- Not always possible to directly replace survey, even for part of population
  - Lack of data
  - Quality concerns
- Can still use admin data to reduce survey costs and respondent burden
- Improve efficiency by using admin data variables not traditionally used with survey data
  - To define more effective strata
  - As auxiliary variables in GREG estimation
  - As comparison to survey data to assist with editing
  - To improve imputation
Estimation of Structural Business Statistics using administrative data

Daniel Lewis
Office for National Statistics
daniel.lewis@ons.gsi.gov.uk
Introduction

• Some Structural Business Statistics (SBS) variables can be taken directly from admin sources
  • e.g. VAT data or Company accounts
• However, there are many variables that are not directly available
• Here we concentrate on estimation methods to deal with the range of variables required for SBS
Simple use of available admin variables

- Some variables may be available fairly directly from admin data – either on their own or by combining available variables
- It may still be necessary to use modelling techniques to deal with coverage or definitional problems
- e.g. Lithuania uses the following formula to estimate Gross investments in tangible goods:

\[
\text{Gross investment in tangible goods} = [\text{tangible assets at the end of the year}] - [\text{tangible assets at the beginning of the year}] +/ - [\text{revaluation}]
\]
Simple use of available admin variables

• Germany has investigated the following method to estimate Full Time Equivalent employees (FTE) using data from a range of sources:

\[ \text{FTE} = \text{Full time employees} + \alpha \cdot \text{Part time employees} \]

Where \[ \alpha = \frac{\text{Hours of part time employees}}{\text{Hours of full time employees}} \]
Cut-off sampling

- It is often not possible to use admin data to estimate SBS variables for the whole population
- However, cut-off sampling may be used to reduce the size of the survey
- A variable $x_1$ is used to partition the population
  \[ i \in U_c : x_{1i} \leq c \]
  \[ i \in U_m : c < x_{1i} \leq m \]
  \[ i \in U_e : x_{1i} > m \]
- Where $c$ is the cut-off, $m$ defines medium businesses used to help estimate in the cut-off, $e$ defines fully enumerated (large) businesses
Cut-off sampling

- The population total $Y$ can then be written
  
  $$ Y = Y_c + Y_m + Y_e $$

  where for $\phi$ equal to $c$, $m$ or $e$

  $$ Y_\phi = \sum_{i \in U_\phi} y_i $$

- Cut-off sampling can be an effective way of reducing the cost and burden of a survey, whilst maintaining the accuracy of outputs

- Some estimation methods are designed specifically to be used with cut-off sampling
Simple ratio adjustment

- Used with cut-off sampling to provide an estimate for non-surveyed part of population
- Uses single admin variable, correlated to variable of interest
- Simple method with good results
- Cut-off usually businesses with <10 employees
Regression modelling

- Simple ratio is based on implied ratio model
- Sometimes possible to make better use of admin data using linear regression models

\[ Y_i = \alpha + \beta_i X_i + \varepsilon_i \]

- Option to use multiple admin variables
- Can be used with cut-off sampling or to model whole population
  - Need strategy to update the model
Regression modelling

• Logarithmic transformation may be needed to deal with skewed nature of variables
  • Need back transformation
• For variables with many zeros, use two stage approach, using logistic regression to model zeros
• Some variables require robust linear regression methods to deal with outliers
  • e.g. using Least Trimmed Squares

• Regression gives good results, but often no significant improvement on simple ratio adjustment
Generalised calibration estimation

• Method that aims to reduce bias introduced by cut-off sampling (Haziza et al, 2010)
• Method builds on calibration estimation framework
• Requires two vectors of auxiliary variables
  i. $X_i$ - auxiliaries related to variable of interest
  ii. $Z_i$ - auxiliaries to explain whether unit falls above or below cut-off
• Method should work well if no residual relationship between probability of being above cut-off and $X_i$ variables
Generalised calibration estimation

- Estimator for whole population

\[ \hat{Y} = \sum_{i \in s} \tilde{w}_i y_i \]

\[ \tilde{w}_i = a_i \left\{ 1 + \frac{X - \hat{X}}{\hat{T}} z_i \right\} \]

\( a_i \) is design weight for unit \( i \)

\[ X = \sum_{U} x_i \quad ; \quad \hat{X} = \sum_{i \in s} a_i x_i \quad ; \quad \hat{T} = \sum_{i \in s} a_i z_i x_i \]
Generalised calibration estimation

- UK experience in the ESSnet Admin Data found problems implementing the method successfully
- Difficult to identify appropriate, sufficiently well correlated auxiliaries
  - Mainly tested for difficult variables – Purchases and Investments
- Different variables work well in different periods
- Results can be acceptable for a single period, but not consistent over time
- No improvement on simple ratio adjustment
Inflation of survey weights

• Ratio estimation is commonly used for SBS variables

\[
\hat{Y} = \sum_{s} a_i g_i y_i
\]

• Calibration weight is calculated as

\[
g_i = \frac{\sum_{U} x_i}{\frac{N}{n} \sum_{S} x_i}
\]
Inflation of survey weights

- When the population is partitioned into $U_c$, $U_m$ and $U_e$ the calibration weight can be expressed

$$g_i = \frac{\sum_{U_c} x_i + \sum_{U_m} x_i + \sum_{U_e} x_i}{\frac{N_c}{n_c} \sum_{s_c} x_i + \frac{N_m}{n_m} \sum_{s_m} x_i + \frac{N_e}{n_e} \sum_{s_e} x_i}$$
Inflation of survey weights

• If enterprises in cut-off band are no longer sampled, can use inflated calibration weight to estimate for population

\[ g_i = \frac{\sum_{U_c} x_i + \sum_{U_m} x_i + \sum_{U_e} x_i}{\frac{N_m}{n_m} \sum_{s_m} x_i + \frac{N_e}{n_e} \sum_{s_e} x_i} \]

• Simple method which works well in some instances
Estimating components

- Specific methods can be used to estimate components
  i. If the corresponding total is available
  ii. If neither components or total are available
Estimating components

i. If the corresponding total is available
   • Apply average proportions from similar businesses to the known total
     • When using cut-off sampling, similar businesses may be larger businesses in the same industry
   • Can use mean or median proportions
Estimating components

ii. If neither components or total are available
   • It can be effective to use nearest neighbour donor imputation (see section 5)
   • Using same donor to impute all variables ensures that relationships are maintained
Choosing an estimation method

- Important to analyse potential methods to test accuracy
- One evaluation method is comparison with survey estimates – but care is needed here...
- Need consistency over multiple periods
- Bootstrap study useful to verify robustness of results – should be little variation in estimates
- For regression methods, important to check validity of chosen model
- May find methods work well for some industries, but not all
Estimation of Short Term Statistics using administrative data

Daniel Lewis
Office for National Statistics
daniel.lewis@ons.gsi.gov.uk
Introduction

• Estimating Short Term Statistics (STS) using admin data has specific challenges
• Turnover and Employment are commonly available
  • VAT and Social Security sources
• Admin data can fall short of timeliness and punctuality requirements
• We discuss three cases, depending on amount of VAT data available at right time and punctuality

1. All VAT data available
2. Most VAT data available
3. Limited VAT data available
Case 1: All VAT data available

- This is the ideal situation
- After receiving and processing the VAT data, it just remains to aggregate to produce Turnover statistics
- Unfortunately for most statistical offices, this possibility is far from reality....
Case 2: Most VAT data available

• With around 80% or more data available, simple imputation methods may be used to fill in missing data

• In most ESS countries it is rare to have sufficient data available for monthly STS outputs

• But the imputation methods described here can be useful for quarterly STS outputs

• In each case, methods work better when applied within homogeneous classes
  • Important that classes are not too small, minimum size around 10-20
Case 2: Most VAT data available

i. Ratio imputation based on growth from previous period

\[
\hat{y}_{i,t} = \left( \frac{\sum_{j \in \text{class}} y_{j,t}}{\sum_{j \in \text{class}} y_{j,t-1}} \right) \times y_{i,t-1}
\]

• Works well if growth rate in class is fairly stable
Case 2: Most VAT data available

ii. Ratio imputation based on growth from same period in previous year

\[ \hat{y}_{i,t} = \left( \frac{\sum_{j \in \text{class}} y_{j,t}}{\sum_{j \in \text{class}} y_{j,t-12}} \right) \times y_{i,t-12} \]

- Works well when there is strong seasonality in the series
- Susceptible to big economic changes – 12 month lag
iii. Imputation when previous period is not available

- May be possible to use growth from an alternative time point
  - e.g. 2 months previously
- For new businesses, simplest to use mean or median value from homogeneous class
- Sensible to trim extreme values when using mean (e.g. top and bottom 10%)
Case 2: Most VAT data available

- Experience shows that any of the ratio imputation methods works well in practice
- However, important to be able to accurately identify businesses in target population
- Lag in information on births, deaths, etc. can lead to under- or over-estimation
- Need process to deal with this in practice
Case 3: Limited VAT data available

- With limited VAT data, imputation is unlikely to give good results
- Estimation, modelling or forecasting methods are required instead
- Choice partly governed by limitations of VAT data
- We examine:
  i. Use of quarterly data for monthly estimates
  ii. Benchmarking
  iii. Forecasting previous data
  iv. Estimation techniques
Case 3: Limited VAT data available

i. Use of quarterly data for monthly estimates
   • Common for a large proportion of businesses to return VAT data on a quarterly basis
   • Problematic for monthly outputs...
   • One solution is the cubic spline
     • Fit polynomial through quarterly data and read off monthly values
     • Can provide reasonable results, but note nothing in the method to prevent negative values
   • In testing, spline surprisingly worked no better than simply dividing quarterly values by three!
Case 3: Limited VAT data available

ii. Benchmarking

• Method for combining series of high frequency data with series of less frequent data

• Aims to preserve movements in high frequency series, whilst keeping consistent with less frequent data

• May be used to improve quality of monthly Turnover estimates derived from incomplete VAT, using quarterly estimates from almost complete VAT
Case 3: Limited VAT data available

ii. Benchmarking

- To use this for monthly estimates before corresponding quarterly estimates are available, need benchmark forecast

\[ Y_{\text{benchmark forecast}} = Y_{\text{monthly}} \times C_1 \]

- \( C_1 \) summarises adjustments from previous periods
  - e.g. mean adjustment from set number of previous periods

- Benchmarking works well if quarterly series is of good quality and not affected by level shifts
Case 3: Limited VAT data available

iii. Forecasting previous data

- When VAT data not available on time, may be possible to forecast from previous period

- Following methods tested in UK:
  
  i. Apply growth between last two periods to last level
  
  ii. Holt-Winters model with seasonal factors and either linear or quadratic trend
  
  iii. Univariate ARIMA model
  
  iv. Singular Spectrum Analysis – based on estimating structure of process that generated time series

- For level estimates, Holt-Winters worked best

- No clear favourite for growth estimates
iv. Estimation techniques

- May be possible to estimate STS Turnover from VAT data using simple estimation techniques
- General approach:

\[ \hat{Y}_t = SL_t + R \times VS_{t-a} \]

- \( SL_t \) is (fully enumerated) survey estimate for large businesses
- \( R \) is a ratio adjustment
- \( VS_{t-a} \) is VAT data for small and medium businesses
Case 3: Limited VAT data available

iv. Estimation techniques

\[ \hat{Y}_t = SL_t + R \times VS_{t-a} \]

• Note that \( R \) can be put equal to 1 if VAT data are not to be adjusted
• \( a \) can be 0 for forecast data
Case 3: Limited VAT data available

iv. Estimation techniques

\[ \hat{Y}_t = SL_t + R \times VS_{t-\alpha} \]

• Three types of method tested
  
  i. Add (forecast or previous) VAT data for small and medium businesses to sampled large businesses (\( R=1 \))

  ii. Calculate \( R \) based on (aggregate level) ratio of survey to VAT data for large businesses

  iii. Maintain small survey for small and medium businesses and calculate \( R \) based on ratio of survey to VAT data from sample small and medium businesses
iv. Estimation techniques

\[ \hat{Y}_t = SL_t + R \times VS_{t-a} \]

- Choice of ratio based on assumed relationship between businesses used to fit model and businesses being estimated
- Likely to get better results calculating ratio separately for different industries
- In practice, methods often only give good results in certain industries
Quality measurement for statistics based on administrative sources

Daniel Lewis
Office for National Statistics
daniel.lewis@ons.gsi.gov.uk
Introduction

• As with any output, important for any statistics using admin data to be accompanied by quality statements
• Well established procedure for reporting on quality of survey statistics, much less in place for admin data

• We discuss the following quality reporting methods
  i. Basic quality indicators for statistics using admin data
  ii. Composite quality indicators for quality themes
  iii. Estimating error for mixed source statistics
Basic quality indicators

- Survey quality indicators do not always apply to admin data based outputs
- Potential quality issues are different
- Lack of well developed framework for inference
- ESSnet Admin Data produced list of 23 quantitative indicators covering 8 quality themes
  - Relevance
  - Accuracy
  - Timeliness and punctuality
  - Accessibility and clarity
  - Comparability
  - Coherence
  - Cost and efficiency
  - Use of administrative data
Basic quality indicators

• Each indicator gives information on output quality
• However, some take account more directly of inputs and processes
• Not the intention that all 23 indicators should be used for each output – focus on most relevant
• In published guide, each indicator is accompanied by an example to explain use
Basic quality indicators

- Accuracy
- Item non-response (% of units with missing values)
- Misclassification rate
- Undercoverage
- Overcoverage
- % of units for which reference period differs from that required
- Size of revisions from different versions of admin data
- % of units which fail checks
- % of units for which data have been adjusted
- % of imputed values in the admin data
Basic quality indicators

• Timeliness and punctuality

• Periodicity (frequency of arrival of admin data)

• Delay to accessing / receiving data from admin sources
Basic quality indicators

- Comparability

- Discontinuity in estimate when moving from a survey-based output to an admin data-based output
Basic quality indicators

• Coherence
  • % of common units across two or more admin sources
  • % of common units when combining admin and survey data
  • % of consistent items for common variables in more than one source
  • % of relevant units in admin data which have to be adjusted to create statistical units
Basic quality indicators

- **Cost and efficiency**
- % of items obtained from admin source and also collected by survey
- % reduction of sample size when moving from survey to admin data
- Cost of converting admin data to statistical data
- Efficiency gain in using admin data
Basic quality indicators

- Use of administrative data
- Number of admin sources used
- % of items obtained exclusively from admin data
- % of required variables derived from admin data that are used as a proxy
Basic quality indicators

- Quantitative indicators are accompanied by qualitative indicators
- Descriptive information to help users understand more about quality
  - Descriptions of sources and their characteristics
  - Methods used to process data and compile outputs
  - Reasons for discontinuities
  - Differences in concepts, classifications and definitions
  - Assessment of sources and likely impact of errors

- Example of full information for quality indicator ‘Undercoverage’ on next slide...
Basic quality indicators - example

Quality Indicator: Undercoverage

\[
\text{Undercoverage} = \left( \frac{\text{No. of relevant units in reference population but NOT in admin data}}{\text{No. of relevant units in reference population}} \right) \times 100\%
\]

* e.g. 16% of units in reference population are not in tax data, making up 4% of Turnover (weighted indicator)

Qualitative indicators / Descriptive information:

- Describe the coverage of the administrative data and any known problems
- Describe methods used to deal with coverage issues
- Assess likely impact of coverage error on key estimates
Composite quality indicators

• Collating some of the basic quality indicators into themes
• Aim to provide a more holistic, overall view of the quality for a particular theme of the statistical output
• Considered each theme, but decided composite indicators only useful for:
  • Accuracy
  • Timeliness and punctuality
  • Coherence
Methods for calculating composite indicators

- **Step 1:** Identify which quality theme and which constituent basic indicators will be used
Methods for calculating composite indicators

- e.g. choose whether to use some or all of indicators for the Accuracy theme
- Item non-response
- Misclassification rate
- Unit missingness
- Overcoverage
- % of units with different reference period
- Size of revisions from different versions of data
- % of units which fail checks
- % of units for which data have been adjusted
- % of imputed values
Methods for calculating composite indicators

- **Step 1:** Identify which quality theme and which constituent basic indicators will be used
- **Step 2:** Standardise the basic indicators so that their values have a similar meaning
Standardisation

• Need to standardise indicators to be comparable:

\[
\text{Indicator value} - \text{Reference value} \over \text{Max value} - \text{Min value}
\]

• Reference value denotes the point where quality becomes unacceptable
• Can then combine indicators using a mean
• Negative values denote acceptable quality
Reference values – min and max
Methods for calculating composite indicators

- **Step 1:** Identify which quality theme and which constituent basic indicators will be used.
- **Step 2:** Standardise the basic indicators so that their values have a similar meaning.
- **Step 3:** Combine and weight the basic indicators to form the composite indicator using a simple mean.
Weighted indicators
Methods for calculating composite indicators

• **Step 1:** Identify which quality theme and which constituent basic indicators will be used

• **Step 2:** Standardise the basic indicators so that their values have a similar meaning

• **Step 3:** Combine and weight the basic indicators to form the composite indicator using a simple mean

• **Step 4:** Is the quality ‘acceptable’ or ‘unacceptable’?
Estimating error for mixed source outputs

- UK and Netherlands investigated a number of case studies for estimating variance and bias for mixed source outputs
- Combination of tax and survey data for short term statistics, based on cut-off sampling
- Variance and bias estimated for following estimators
  i. Inflation of calibration weights
  ii. Simple ratio adjustment
  iii. Regression modelling
Estimating error for mixed source outputs

i. Inflation of calibration weights

- Estimator calculated as

\[
\hat{Y} = \sum_{i \in s} a_i g_i y_i
\]

\[
a_i = \frac{N_h}{n_h} \quad \text{for } i \in h
\]

\[
g_i = \frac{\sum_{i \in U_{g(h)}} x_i}{\sum_{h: g(h) = g} \frac{N_h}{n_h} \sum_{i \in s_h} x_i} \quad \text{for } i \in g
\]
Estimating error for mixed source outputs

i. Inflation of calibration weights
   • Analytic expression for total variance of estimator

\[
\hat{V} = \sum_{g=1}^{G} \left( \frac{\sum_{i \in U_g} x_i}{\sum_{i \in U_{g\setminus c}} x_i} \right)^2 \sum_{h: g(h) = g} \frac{N_h^2 (1 - f_h)}{n_h (1 - n_h)} \sum_{i=1}^{n_h} \left( y_i - \hat{R}_{g(h)} x_i \right)^2
\]
Estimating error for mixed source outputs

i. Inflation of calibration weights
   • Use bootstrap approach to estimate bias
   • Re-sample many times over with replacement, sample size n-1
   • Re-calculate estimates each time and estimate bias as average difference between bootstrap and full sample survey estimates

\[
\hat{B} = \frac{1}{K} \sum_{k=1}^{K} (\hat{Y}_{\text{cut-off}, k} - \hat{Y}_{\text{full}})
\]

• \( k \) denotes individual bootstrap replications
Estimating error for mixed source outputs

ii. Simple ratio adjustment

iii. Regression modelling

• No analytic expression for variance with these two methods, so use bootstrap method for both variance and bias

• Variance for cut-off part of population is

\[
\hat{V}_c = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{Y}_k - \overline{Y}_{est})^2
\]

\(\overline{Y}_{est}\) is mean of estimates from bootstrap replications
Estimating error for mixed source outputs

- Netherlands extended the approach to investigate the likely impact of measurement error in the survey variables and misclassification error on the frame
- Incorporated as parameters based on probabilities of these errors for different sources and in different industries
- Analysis found source specific misclassification error is important enough to require including in bias estimation
The use of administrative and accounts data for economic statistics - Summary

Daniel Lewis
Office for National Statistics
daniel.lewis@ons.gsi.gov.uk
Summary

- Examined range of methods for making greater use of admin and accounts data in official statistics

- Not straightforward...

- Need to integrate admin and statistical data
  - Issues with linking, definition and coverage

- Data should be cleaned
  - Various methods
  - Difficulty of being unable to contact businesses to query data
Summary

• Estimation methods should aim to maintain quality
• In many cases will need to retain small sample survey

• Investigated methods for dealing with:
  • Lack of timeliness in admin data
  • Admin data not of required periodicity
  • Not all variables available in admin data

• All statistics produced should be accompanied by quality statements
  • Range of options for doing this
Summary

• Due to lack of well established statistical theory for use of admin data...

• Important to carry out thorough analysis of methods and their quality implications before implementing them

• For more information: http://essnet.admindata.eu