Topic (1): Editing Administrative Data and Combined Data Sources

Discussants: Natalie Shlomo (Southampton University, UK) and Heather Wagstaff (ONS, UK)

INTRODUCTION

There are 8 papers in Topic (1):

WP2 - Use of Administrative Data in Statistics Canada's Annual Survey of Manufactures, Canada

WP3 - Conceptual Modeling of Administrative Register Information and XML - Taxation Metadata as an Example, Finland

WP4 - Use and Editing of Administrative Data in the Business Indicators Unit, New Zealand

WP5 - Detecting Outliers in Price Quotes for the Canadian Consumer Price Index, Canada

WP6 - Imputation of External Trade Data in Denmark, Denmark

WP7 - Evaluation of Editing and Imputation Supported by Administrative Records, Israel

WP8 - Editing and Imputation for the Creation of a Linked Micro File from Base Registers and Other Administrative Data, Norway

WP9 - The Use of Administrative Data in the Annual Survey of Retail, Wholesale and Services, United States

Most of the papers deal with aspects of edit and imputation when incorporating administrative data into the statistical processes for economic and business data (WP2, WP4, WP5, WP6, and WP9). The papers emphasize the need for pre-processing of administrative data to obtain common classifications, definitions and time frames before combining with survey data. This is carried out through the prior assessment of completeness, coherence and quality of the administrative data and edit and imputation processes. Editing at source point by data suppliers is another method for ensuring high quality administrative data. The development of generic modules for editing and imputation for administrative data is particularly challenging since data collection methods and formats of administrative data vary greatly depending on the source.

In WP2, a pseudo census is developed from combined administrative data with survey data, and high quality statistical data is obtained. The quality indicators include variance estimates of the totals which take into account the imputation model. Other quality indicators are used for assessing the administrative data. In WP9, administrative data replaces survey data for the smaller statistical units, and the impact on the quality of the data is also measured through estimates and their variances. Papers WP4 and WP5 deal with selective editing techniques. WP5 concentrates on outlier detection methods as a selective editing technique taking into account the skewness of economic data. WP4 focuses on selective editing through thresholds where the goal is to obtain "fit for use" data based on user requirements as opposed to "perfect" data obtained by manually editing all units. WP4 and WP6 specifically give examples of edit and imputation processes for trade statistics based on administrative data where WP6 presents a macro-imputation approach based on estimating a total for the non-respondents and then implementing an apportionment of the total.

Two papers deal with edit and imputation for social surveys, i.e. the statistical unit is an individual (WP7 and WP8). WP8 focuses on linking multiple data sources to obtain employment characteristics

of an individual taking into account that the more sources linked the higher the possibility for errors. Prior knowledge of the administrative sources is fundamental to obtaining high quality data. WP8 gives an excellent review of the impact on using multi-source administrative data on the quality of statistical data, and in particular for the edit and imputation processes: error localization, outlier detection, minimizing the need for edit and imputation, and selective editing techniques. Other important statistical processes are supported by the administrative data, such as: record linkage, coding and imputation of new variables.

WP3 deals with the development of a systematic methodology for constructing metadata for administrative data. This process is very important in order to combine administrative metadata with statistical metadata obtained throughout the various data processing stages. A final, complete and comprehensive metadata structure is then available for users of statistical outputs based on integrated data sources.

The underlying theme in all of the papers is how the use of administrative data in the survey processing and efficient edit and imputation procedures based on error localization techniques, sophisticated modeling and selective editing methods, increases the quality of the statistical data, and reduces respondent burden and survey costs.

POINTS FOR DISCUSSION

- Can we automatically assume that administrative data is of higher quality than survey data? How can we optimize and integrate lower quality administrative data? How should thresholds be set below which administrative data should not be used at all? How can we reconcile the differences in definitions of the collection units, variable classifications and definitions without introducing new bias in the statistical data? Some papers show discrepancies between administrative and reported data and yet they are used because it reduces cost and response burden and the impact on the final estimates is minimal. More research needs to go into quality indicators and statistical measures (based on the trade-off between the bias and variance) which assess the impact of using administrative data in the survey processes, in particular when the administrative data is of low quality.
- Metadata from administrative registers is not in a user friendly format. How can we best integrate content information about administrative registers into the metadata describing the overall statistical processing operation? The use of metadata throughout the data processing stages provide key explanations to end users on changes observed in the series, and thus it is important to develop good systems for preserving metadata.
- Is there a mechanism by which we can influence the methods of data collection by suppliers of administrative data in terms of content and format, to ensure more generic pre-processing and edit and imputation processes?
- Some papers describe edit rules for administrative data based on historical data and past experiences. Administrative data, however, can be constantly changing and edit rules need to reflect and capture these changes. Edits that are obsolete and irrelevant need to be deleted and new edits need to be added when the content of the administrative data changes. How can edit rules be updated? Are there automatic methods for checking and updating edit rules?
- In some papers, thresholds for selective editing are set by budget constraints and not necessarily on an evaluation of the impact on the quality of the data and the final estimates. How can thresholds

be set for selective editing? How can historical data obtained from administrative data and external knowledge feed into the selective editing thresholds?

- A very important problem when using selective editing techniques on administrative data is that they generally target the most influential and larger units for editing. However, as seen in many papers, administrative data is often used to replace survey data, in particular for small businesses in business surveys. These small businesses however do not undergo any editing in the administrative data and this means extended editing processes for the business surveys that use the administrative data. Is there a way to efficiently edit smaller businesses through selective editing technique in the administrative data?
- Outlier detection methods are good selective editing techniques, but they don't necessarily target the most influential units which have greater impact on the estimates. How can the selection of influential units be incorporated into an outlier detection methodology and how would the thresholds be defined?
- How can better imputation models for administrative data be implemented? Are there specific model choice techniques when using administrative data as opposed to survey data, i.e. more use of historical data or other data sources? For example, an imputation model for enterprises reporting quarterly or half-yearly use as covariates the enterprises that report monthly. However, these enterprises are usually very different in their size and other factors.
- The availability of several data sources at the data collection phase provides unique challenges for the editing process and their integration can result in the introduction of new errors. For item non-response, for example, it is not clear which source should be used for imputation. How should the quality of the variables be assessed in a multi-source data collection?

SUMMARY OF PAPERS

WP 2: USE OF ADMINISTRATIVE DATA IN STATISTICS CANADA'S ANNUAL SURVEY OF MANUFACTURES

Steve Mathews, Canada Post and Wesley Yung, Statistics Canada

1. SUMMARY

This paper reports on activities at Statistics Canada to increase the use of administrative data in its economic survey programs with the aim to improve data quality, reduce respondent burden and to reduce the cost of conducting surveys. In particular, the paper focuses on one survey, the Annual Survey of Manufactures (ASM) which is an annual survey covering all manufacturing establishments. The survey collects information on financial variables (revenues, expenses and salaries) and on commodity variables (manufacturing inputs and outputs).

The main focus of the paper is how a pseudo – census can be developed by combining administrative tax data with the survey data, in particular when not all of the establishments have tax data. The administrative tax data is split for the incorporated and non-incorporated establishments. For the non-incorporated establishments, only estimates of totals for specific variables at an upper geographical region and industry are obtained as explained below. The incorporated establishments have much more information, but only the section totals and net income/loss are mandatory and are complete. This data also undergoes edit and imputation to assess the quality of the administrative data, to clean records and to ensure that the financial totals balance.

A detailed description of the ASM is given:

- Target population - all businesses in the manufacturing industry as recorded in the business register and includes the description of the hierarchy of the businesses, companies and establishments. No data is collected for the smaller units which represent less than 10% of the economic activity.
- Sample design - stratified simple random sample with take all and take some strata.
- Questionnaire design - collects both commodity and financial variables, though the focus of the paper is only on financial variables. The mode of investigation is mail out / mail back with telephone follow up which is managed by a score function to prioritize units.
- Edit and imputation - based on the administrative tax data as described below.
- Estimation - aggregation of the pseudo-census microdata, except for the take none strata where only total revenues are calculated as follows:
  - incorporated businesses - data obtained from the tax data;
  - non-incorporated businesses - estimates obtained by drawing a simple random sample from the list and calibrating sampling weights benchmarked to the total revenues of the non-incorporated businesses as obtained from the Business Register (BR).

There is a different definition for the collection unit based on the administrative data (legal entity level) and the sample data (statistical entity level), thus all complex establishments are in take all strata and therefore tax data is only used for simple businesses.

Only about half of the variables on the administrative tax data may be equivalent in concept to the survey variables. To build the pseudo - census, an analysis was carried out to examine which variables could be directly replaced by the tax data based on:

- Correlations - check whether a linear relationship exists between the two data sources,

- Consistency - same direction and level of reported values and the number of reported zeros,
- Distribution of ratios – ratios calculated between survey to tax data for each unit and for each variable,
- Comparison - weighted population estimates of totals compared.

Results showed that for general concepts, i.e. total expenses and total revenues, there was consistency between the data sources and therefore replacing survey data with tax data would not greatly effect the quality of the data. For more detailed items, however, there were wide discrepancies. Thus, for 7 general variables representing the totals in the different sections, tax data is replaced directly as input into the pseudo – census, while for the other financial variables, the tax data is used as auxiliary data for imputation as follows:

- Historical imputation - for each section, the distributions of items from past data (if available) are prorated to the current totals obtained from the tax data,
- Ratio imputation - for each section, the ratio between the detailed items and the total within imputation groups (defined by geography and industry classifications) are calculated and these ratios are applied to the totals of the business from the tax data,
- Nearest neighbor donor imputation - a donor most similar to the recipient business is found and the distributions are donated which are then applied to the totals of the business from the tax data.

Note that the imputation is used for both non-responding and non-sampled units.

The quality of the method for developing the pseudo-census was assessed by comparing estimates and their variances to the weighted survey and to a pseudo -census assuming no tax data is available and using the BR with revenues as auxiliary data for imputation. The variance takes into account both the imputation model and the sample design through an innovative approach based on the assumption that the variance is not conditional on the response mechanism. The analysis was carried out on 2 variables: total expenses (directly replaced tax data) and total energy expenses (imputed from tax data), and excluded imputation classes having less than 5 responding units. Also, the analysis assumes that all complex units responded.

The tables show the differences in the estimates and their CV's between the different methods. The authors report little differences between the methods, however, for the estimates we note, for example, that for total expenses for all manufacturing, the relative difference from the Horvitz-Thompson (unbiased) estimator is 1.8% for the pseudo - census without tax data but only 0.5% for the pseudo - census with tax data. This seems to be a large difference though the base for the percentages is not given. The CV's are considerably lower for the pseudo - census using tax data for total expenses but for the imputation based total energy expenses, there is almost no difference between the methods, and in some cases the variance is higher than the Horvitz-Thompson estimators. Since there is the underlying need for a pseudo-census, the best option is based on the pseudo-census which uses the tax data since it overall gives smaller CV's and less bias.

2. MAIN ISSUES

- The paper demonstrates how administrative data can be used to build censuses out of survey data while maintaining high quality data, reduce bias, and obtain minimal CV's for the target estimates.
- Before incorporating administrative data into the statistical processes, it needs to be carefully assessed for its completeness, coherence and quality, and must undergo extensive edit and imputation processes on its own in order to get the administrative variables in line with the survey target variables.

- Several methods are used for assessing the quality of the data: differences in definitions of variables between the sources as measured by statistical criteria and differences in point estimates through the bias and variance. The authors gave a good interpretation of the results of the comparisons.
- The paper concludes with some ideas for improvement: better edit and imputation systems for administrative data to increase its quality prior to its use in the statistical processes; better modeling techniques between survey data and administrative data using more covariates and an intercept term; and the development of quality indicators specifically focused on the integration of administrative data into statistical systems.

3. PROBLEMS

- Can we automatically assume that administrative data, in particular tax data, is of higher quality than the survey data?
- How can we optimize and integrate lower quality administrative data? The paper shows the need to develop edit and imputation procedures for the administrative data to clean the records prior to its use. Results in the paper show, for example, large discrepancies between the survey and administrative data and yet it is still used because of user requirements.
- How far can we go with administrative data of low quality? How to set thresholds based on quality indicators below which administrative data can't be used? This paper shows a good analysis for assessing the quality of the end product. In this paper, some of the use of the administrative data are through estimates that are subject to sampling errors. There is no mention if these sampling errors are taken into account in the variance of the final estimates.
- How to handle differences in collection units and differences in variable definitions between the survey and administrative data?

4. RELATED ASPECTS

- More research needs to go into quality indicators and statistical measures which assess the impact of using administrative data, in particular when the administrative data is of low quality. The paper shows quality indicators based on measuring the impact on bias and variance of the estimates. There may be other quality indicators measuring the impact of the massive imputation that is implemented in this case.
- Metadata is particularly important to users on the definitions, strengths and limitations of the administrative data, sampling errors and measurement errors that are directly obtained from the administrative data, etc.
- Imputation methods can be expanded to include more sophisticated models.
- Quality indicators that focus on combining administrative and survey data.

WP3: CONCEPTUAL MODELING OF ADMINISTRATIVE REGISTER INFORMATION AND XML - TAXATION METADATA AS AN EXAMPLE

Heikki Rouhuvirta

1. SUMMARY

The paper first considers how best to understand the information content of administrative registers. To find information about the content we would normally refer to paper or electronic handbooks produced by the data owners. When producing statistics we are unable to integrate the owners

handbooks into statistical systems and tend to use them manually. This leads us to consider the best way for producers of statistics to communicate the contents of the registers to data users. It is important to ensure that users gain a good understanding of the content and make accurate interpretations.

The unambiguous meaning, or interpretation, of the administrative data should feed through every stage of the production process to ensure that it is well understood at all stages. This is especially important for edit and imputation. Data models should be revised so as to facilitate the presentation of statistical information and the interpretation of register data through a uniform reference frame and in a standard format. There are no ready made solutions, either for producers of statistics or for the producers of administrative data. Hence, as a starting point, the author considers solutions for the conceptualisation of statistical metadata already applied at Statistics Finland. This is a reasonable approach since semantic data can be considered as a question of metadata. The semantic description of administrative information should be fully integrated with the statistical metadata derived from a processing operation.

The paper briefly describes the Common Structure of Statistical Information (CoSSI) method developed by Statistics Finland. The aim is to produce a fully defined and logical information structure which is characteristic of the current administrative information, within the frame of which the semantic description of administrative data can be produced. The method is applied to personal taxation data and to the administrative information describing it.

The concept of taxation is described as an information system. Here we are not concerned with the payment of taxes but rather with what type of information the taxation process is based on. The sole purpose is to produce descriptions of the data contained in the taxation material that will explain the meaning of the data collected on the register which results from the payment of taxes. By not considering tax itself, we can then consider the concept of income that is formed from income sources and deductions. This is aptly described in the paper by consideration of two figures showing the flows of types and sources of income followed by income tax deductions.

The main source of information about the content of data from tax registers is contained in instructions and handbooks developed by the Tax Administration who also provide technical descriptions of their information systems. A large source of contextual information describing in detail personal taxation is available annually as a printed document. An electronic copy, which is normally used for printing is available.

The paper discusses the practical issues concerned with translating information about the data into structured metadata. It goes on to discuss how a structure for the taxation information can be created which utilises the existing electronic data, which defines the content of the information, and transfers this to users without losing any of the information. A logical concept model for the taxation information was defined and shown in the paper as a tree structure. When implemented the information structure of taxation is hierarchical and the defined structure also allows the presentation of the information from the taxation handbook.

In order to test the models for register information, an electronic version of the taxation handbook was used. This provided structured register metadata which was then linked to records selected from the relational database. The metadata for each tax record can be searched by using prescribed codes. An example screen is displayed in the document. The use of structured register metadata as part of statistical metadata is illustrated by an example from income distribution statistics.

The paper concludes by considering moving towards a metadata driven statistical production. This becomes a reality once the metadata from registers can be fully integrated with the statistical metadata derived from the production of statistics. In a metadata driven production process rich metadata is present and available at all stages of the process. This has particular importance and relevance for editing and imputation.

2. MAIN ISSUES

- The paper introduces the very important issue that information about administrative registers is held by the owners of the registers but often not in an electronic format suitable for use in the statistical production process.
- Unambiguous interpretation of the meaning of the content of the administrative data should feed through every stage of the production process to ensure that it is well understood at all stages. This is especially important for edit and imputation.
- Where register data has been used as an integral part of the survey process, producers of statistics need to ensure that users gain a good understanding of the content and make accurate interpretations.

3. PROBLEMS

- Metadata from administrative registers is not in a user friendly format. How can we best integrate content information about administrative registers into the statistical processing operation?

4. RELATED ASPECTS

- Is there a mechanism by which we can influence the methods of data collection used by the owners of administrative registers, in terms of content and format, to ensure that the data can be used for the statistical processing?

WP4: USE AND EDITING OF ADMINISTRATIVE DATA IN THE BUSINESS INDICATORS UNIT, STATISTICS NEW ZEALAND

Vera Costa and Blair Cardno, Statistics New Zealand

1. SUMMARY

The Business Indicators Unit (BIU) of Statistics New Zealand is responsible for measuring and reporting on indicators of economic activity and trade flows for different market sectors. Administrative data is used as stand alone data for output and to supplement existing surveys, thus reducing response burden and collection costs.

A description of the sources of the administrative data is given in the paper:

- The BAI series is developed from several administrative sources, including revenue and tax data. This data is used to supplement several surveys, such as the Wholesale Trade, Retail Trade and Manufacturing Surveys, by removing the need to survey small to medium businesses (which comprise about 15% of the total value).
- Trade data is obtained from the Customs Service which supplies information on export and import transactions. The Overseas Trade Indices are developed from this data.

- Local government agencies provide building consent forms for building statistics which are also used as frames for drawing samples for the quarterly Building Activity Survey.

Suppliers of administrative data generally have an agreed electronic format for the transfer of data. The exception is the Building Consents forms which need to be pre-processed. Some administrative data need to have their industry classifications added to the data, either automatically or manually depending on the file.

The paper differentiates between editing strategies for:

- Perfect Data – editing of every record and complete follow up with the aim of obtaining accurate data. No consideration is taken of the user needs or the final output. The trade data and building consent forms use this strategy where every record is edited at the micro level.
- Fit for Use Data- editing what is possible under the given resources. This requires an understanding of what the data is used for and then specifying the data quality level. The BAI uses this strategy.

More attempts are being carried out to obtain other administrative data sources, in particular from private companies. This involves defining the data and ensuring data supply.

The paper discusses the use of administrative data in the BIU and presents two examples: Overseas Merchandise Trade and the Business Activity Indicator (BAI) Series. In the conclusions of the paper, the challenges for developing a generic selective editing strategy which would take into account both examples is underlined.

Overseas Merchandise Trade provides statistical information on import and export of merchandise goods. The data is obtained from the customs service and much of the pre- editing is carried out by the data supplier for both its own needs as well as the specific needs of Statistics New Zealand. These include reality checks on values and correcting missing entry identifier information. Statistics New Zealand also has an employee working in the Customs audit team one day a week to deal with queries.

A text file is provided with new import and export entries, amendments deletions, client details, excise data, etc. At Statistics New Zealand the data undergoes initial editing for correct formats and duplications. Errors are repaired before loading the file into the editing data base.

A system of automatic edits are carried out for simple errors which are corrected deterministically. These involve the correct classification of codes. The impact of the automatic edits on the quality of the data has not been quantified, but it has reduced the number of manual edits and has saved resources and time. Manual edits made up about 31% of all edits in 2004 compared with 100% prior to the introduction of the automatic edits.

After the data is loaded into the editing data base, the data is checked against edit constraints. These have been developed over time and are based on historical experiences. The flags define:

- Errors - an impossible combination or definition that needs to be fixed,
- Warnings - suspicious situations that require further investigation.

The values are checked against the “unit value range” which are the expected prices per unit. These are updated on a quarterly basis to ensure that the unit value range moves with the market. The validation of the ranges is carried out manually at a very detailed level by checking the largest and smallest differences between the old and updated ranges. This allows insight on the impact of having a small number of entries for the item as well as on the detection of outliers.

Every entry with an error/warning is checked manually. Historical and other important information is supplied to the editors for their checking processes. Final checks are carried out on the data before moving to the output environment, including macro checks. These include checking for influential transactions on the totals as well as the ranges and percentages of the totals out of the overall total. Discrepancies are checked with other information such as industry knowledge. The final stage validates the output data. This is a very important step since it provides metadata and future key explanations to users for movements in the Trade Series. The entire process takes about 4 months. The data is then moved to the output data base, where gross weight is apportioned across different items.

Editing accounts for about 41 percent of the budget for Overseas Trade Statistics. Threshold editing will be introduced in order to reduce the amount of editing. The proposal is that small valued entries with non-missing dollar values will not be brought up for manual editing for the warning edits. In an initial analysis, almost no change at upper aggregated levels were detected as a results of changes made to the lower raw values of entries. The threshold below which no editing will be carried out still needs to be determined and will be based on the amount that is saved from the editing process and historical data in order to control the impact on the quality of the data. This initial project on the Overseas Trade Statistics will be part of a large scale project at Statistics New Zealand for moving to a more efficient business model, including generalized and parameterized modules to support various outputs.

Business Activity Indicators uses GST data matched to the Business Frame (BF) to provide monthly indicators of business activity by industry in New Zealand 2 months after the reference month. The processing includes: checking for invalid numbers; data extraction; data matching to the BF; data manipulation which include apportioning GST values among enterprises within a group and apportioning GST values monthly through modeling for those businesses submitting 2 monthly or 6 monthly; and data editing which is carried out manually.

Several statistical outputs are generated by the GST data. It is also used in the Quarterly Manufacturing Survey and other surveys to model quarterly data for small businesses in order to reduce response burden as well as for size variables for drawing samples. To move from the experimental to production stage, the system was redeveloped and revised and a new editing methodology was introduced.

Missing monthly GST returns result from businesses who report 6 monthly or 2 monthly returns and these businesses need their returns apportioned either back over the reference period (temporal apportionment) or forecasted for the next reference period (forward estimation). The temporal apportionment uses seasonal factors in the 2 monthly units and simple division for the 6 monthly units. Forward estimates are imputed using historical or mean imputation. In addition, non-respondents are imputed within imputation cells (industry*filing frequency) where mean imputation is used for the births and historical imputation is used for existing businesses or if there is no historical data than mean imputation is used. The ratios of the previous and current month are manually checked and corrected if necessary as a result of large capital movements within an industry. An important note is that the selective editing methodology typically targets large and influential enterprises, however in most surveys it is the small enterprises that have their data replaced by administrative data which haven't gone through any editing processes.

Focus is placed on developing a better processing system to minimize delays which will include developing an automated editing system. In particular, the new process will first implement data editing and then carry out the data manipulation. The data editing in the new system will include outlier detection for editing and these outliers are prioritized based on their contribution to the totals. The values with a contribution over a fixed threshold are manually edited and the ones below automatically

edited. The new editing system takes into account that the data is numerical, is available at the unit record level and a long time series exists for most units.

Outliers are detected based on ranges developed from time series of historical data. The mean value up to the current month is a direct function of both the mean value up to the previous month and the total number of months in the series. An outlier is outside the range defined by the confidence interval around the mean value. To manually or automatically correct for an outlier depends on whether the contribution of the new GST value to the group's total is above the fixed threshold. The automatic editing is based on editing rules. For example, the main editing rule is to use the previous year's mean value for the unit and not the new mean value. The groups for editing and imputation are defined by industry code. Thresholds are set to take into account available resources for editing.

1. MAIN ISSUES

- The paper introduces a very important concept of perfect data vs. fit for use data, and this requires an understanding of the needs of the users and what the data will be used for.
- Editing at source point, including an employee of Statistics New Zealand to work in the data supplier's auditing team. This greatly enhances the quality of the administrative data.
- Selective editing techniques and setting thresholds for manual or automatic editing is based both on historical data and the resources available for editing.
- The modernization of the BIU includes much more use of available administrative data, better editing processes to save time and resources, and developing unified and generic modules for editing and imputation, in particular when data sources are diverse and may include both numerical and categorical data.

2. PROBLEMS

- Edits are defined by historical data and past experiences. Administrative data can be constantly changing and the edit constraints need to reflect and capture these changes. In addition, some edits become obsolete and irrelevant and need to be deleted. Are there better ways to update edit constraints?
- How can thresholds be set for selective editing? They can be based on budget constraints as mentioned in the paper, but they should also be based on evaluating the impact on the quality of the data. The use of historical data and external knowledge should also feed into the selective editing thresholds.
- A very important problem when using selective editing techniques which generally target the most influential units for editing is when administrative data are also used to replace survey data for the small businesses in business surveys. These businesses however do not undergo any editing. This means extended editing processes for the particular business surveys that use the administrative data.

3. RELATED ASPECTS

- A times series approach to the analysis of historical data is necessary to take into account the seasonality and trends of the import and export data for setting thresholds for selective editing.
- The use of metadata obtained from the detailed micro and macro editing stages of the trade data provide key explanations to end users on movements in the trade series.

WP5: DETECTION OF OUTLIERS IN THE CANADIAN CONSUMER PRICE INDEX

Abdelnasser Saidi and Susana Rubin Bleuer, Statistics Canada

1. SUMMARY

For each item in a fixed basket of commodities, a price relative which is equal to the ratio of the price of an item in the current period to the price in the previous period, is calculated. The consumer price index (CPI) is a weighted sum of means of price relatives. There are 8 major components in the CPI with a total of 169 basic classes. The items in the basic class are surveyed for their prices at sampled outlets. Micro indices for an item in the class is a geometric mean of price changes (price relatives) obtained from sampled outlets. These micro indices are then weighted using an arithmetic average for a class level price index for a certain geographical area. The weights are obtained from the Survey of Household Spending and reflect the relative importance of an item in the class.

Data editing for the consumer price index focuses on the detection of outliers and the identification of influential observations from among the price relatives. Outliers arise from mistakes in coding, non-recorded specialized prices (specials) or other unusual economic events.

The old outlier detection method was based on a combination of manual detection and fixed stationary bounds. A large amount of resources was used to correct a small number of errors. At the point of data collection, prices are standardized, special prices are coded, and all records with special prices or price relatives that have increased or decreased by 15% are selected and evaluated manually. Resources are wasted on manually checking price relatives that are 1 but both the previous and current prices are specials. In addition, price relatives that have regular and special prices in one of the months are not checked if the price relative has not increased (or decreased) by 15%. In other words, the old method used a problematic selective editing threshold where all failures are edited manually and all others are deemed reasonable and not checked.

The new methods for detecting outliers and influential observations take into account the skewness of the distribution, small sample sizes and variations in prices as a result of specials . Price relatives with drastic movements as a result of special prices are placed into separate editing classes when they can be identified. The aim is to detect outliers for the current month assuming that the previous month are already verified. Thus, there is an automatic historical edit built into the process.

5 different methods are compared in an analysis that was carried out between 2003 and 2004 on price relatives that reflect different behaviors over periods of time resulting from fluctuations in the market and other economic upheavals, i.e. price of meat during the mad cow scare. Because of the problem with specials, detection methods are favored with higher breakdown points. In addition, the skewed distribution of the price relatives must be dealt with either through transformations, i.e. log-transformation, which make them more symmetric or through asymmetric bounds (fences). Editing groups are defined based on an item (or a set of items in a basic class)*geography. Each group must have at least 15 regular observations.

The methods are described in depth in the paper and include: Tukey's algorithm and variants, Quartile method or Asymmetric Resistant Fence method (QM), Resistant Fence Method (RFM), MAD, and the Hidiroglou-Berthelot variant (HB) which is the QM method on a different transformation of the data.

The paper describes the analysis on the item hand towels. The log transformation was carried out to make the data more symmetrical based on statistical tests. Results for the detection of outliers are very

similar for the above methods except for the Tukey's algorithm which automatically selects values that are not extreme and also selects much fewer extreme values than the other methods. The recommendation is to use the QM with the log transformation since it is simple to implement and yields similar good results. Using this method, the analysis showed that all the corrected points had been detected as outliers.

For the items in the basic class "gas", 14 outliers were detected among 1662 observations. None of the outliers detected by the QM method in this case was corrected by the old manual method by evaluators. The contribution of each outlier to the basic class index was evaluated. More work needs to go into developing thresholds for determining when an outlier is influential or not.

2. MAIN ISSUES

- Outlier detection and influential observations are essentially a data editing problem that needs to be managed. The edit rule is basically a single rule based on historical data. A selective editing approach is adopted.
- Sound statistical methods for detecting outliers driven by the skewness of the distribution and the number of units in the editing class.

3. PROBLEMS

- Can a price relative be influential without being an outlier? More work needs to go into this area. The selective editing described based on outlier detection methods determines the outliers but does not determine the most influential price relatives. The influential price relatives should also undergo selective editing. How to set thresholds for influential price relatives and can it be incorporated into the outlier detection methodology?
- Only one item is evaluated in the example in the paper. How do you reach conclusions and set thresholds based on many items over different periods of time? Does the QM method always produce the best results? Need to elaborate on the difference between the manual editing and the outlier detection algorithms.
- No discussion on how outliers are treated – are they removed from the indices or are imputations carried out?

4. RELATED ASPECTS

- Outlier detection as a data editing problem.

WP6: IMPUTATION OF EXTERNAL TRADE DATA IN DENMARK

Carsten Zornig, Marius Ejby Poulsen, Peter Ottosen

1. SUMMARY

With the emergence of the EU in 1993, registration of trade between national borders of member countries ceased. The data collection changed from a full coverage of commodities to a system based on direct reporting to the statistical agencies from selected companies according to EU regulations. For trade with countries outside the EU, administrative sources are still available. Trade statistics at Statistics Denmark are based on:

- Intrastat-monthly declarations from a panel of companies (about 10,000) selected based on the size of their annual trade. Administrative VAT data and previous Intrastat data are used for determining the companies.
- Extrastat-administrative data from Customs. Data consists of flow, countries, commodities, quantity and value. All transactions except small ones below a threshold are sent to Statistics Denmark.

Some imputation is carried out for Extrastat for transactions below the threshold. Imputation is carried out on a macro level. For example, instead of imputing a single transaction, imputation is carried out for an aggregated value, such as for a specific company in a specific month by partner country and commodities, i.e. a cell in the commodity/partner country matrix.

Imputation is mostly carried out for Intrastat because of non-response (total and partial) and trade below thresholds. Non-response is problematic. For 2003, 10.3 arrivals were not reported and 3% were below the threshold, so 13% needs to be estimated. For the dispatches, 5.2% were non response and 1.8% were below thresholds so 7% needs to be estimated. Trade estimates need to be constantly revised. Data can be reported by individual transactions or aggregated transactions so there is no definitive number of units. Therefore, the statistical unit is not the reporting firm or company, rather the cells in a partner country/commodity matrix.

Statistics Denmark has a central business register and administrative VAT data which contains important variables: Total EU buys and Total EU sales. VAT data is used for setting thresholds and for imputation. Imputation is carried out as follows:

- Flash Model (current months) - Based on a linear regression model for obtaining initial estimates for about 60% of the reported trade across different kinds of enterprises. To be included in the estimation group, 3 criteria have to be met: timely reports for the last 12 months, correctness of reports and validity of reports. Historical data are generated, and an OLS regression model is carried out with the dependent variable the arrivals or dispatches of trade, and the independent variables the estimation group trade, seasonality dummies, and other dummies for outliers. The estimation of missing trade is the difference between the total estimate and the reported trade. Missing trade is spread over partner countries/commodities matrix.
- Master Model (revised months) - Imputation is carried out for all enterprises based on VAT data. The trade data collected under Intrastat are divided by transaction. Large discrepancies with the VAT data are treated separately. VAT data needs to undergo imputation since enterprises may report either monthly, quarterly, or half-yearly. After obtaining a complete data set for the VAT data, this provides monthly information on EU trade for all enterprises. The imputation model is as follows: Based on enterprises reporting monthly for the last 24 months, a trend for both import and export figures are calculated based on an average least squares model and parameters are estimated. These parameters are used for imputation of monthly VAT which then undergoes seasonal adjustment.

Checks for errors are carried out:

- large values are checked manually;
- detailed checking of VAT data for outlier reports based on deviations from a range defined by 2*standard deviation;
- manual investigation of all enterprises having major discrepancies between Instrastat and VAT.

After error checks, Intrastat and VAT data are merged. Imputation is carried out within groupings according to the type of non-response or whether below thresholds. If Intrastat exceeds VAT value, no

imputation is carried out. If partial non response (VAT exceeds Intrastat) a percentage of the VAT data is used taking into account that some data is available. For total non-response and below thresholds, the VAT replaces the reporting data. Distribution of the non-reported trade among the country/ commodity matrix is based on the trade pattern from previous reports, and if there are no previous reports, then a trade pattern is calculated from similar enterprises within the same industry. The imputed non-reported trade is aggregated into one record and benchmarked.

Future research is presented in the paper based on using other data sources. In addition the timeliness issues are discussed since only a very short time is available for error checking before the first release. Other methods need to be reviewed. One particular method is to incorporate the flash model in the imputation of the VAT and also to develop a new model for the imputation of the /commodity matrix.

2. MAIN ISSUES

- Imputing trade data, both reported and administrative. Data is imputed at a macro level and not for a basic statistical unit (firm or company). The data are stored in matrices: partner country*commodity and each cell contains the value of the trade.
- Imputation models (mostly regression models) relies heavily on administrative data which is used as a direct source for trade data.
- Error checking carried out manually and not enough time is allowed for this process.

3. PROBLEMS

- Because trade can be reported by individual transactions or aggregated transactions, the number of units in the population is unknown.
- Direct replacement of administrative (or imputed administrative) data. Some checks are carried out on whether administrative data can directly replace reported data. The imputation method should be assessed for its quality.
- A model is used for obtaining first initial trade estimates on about 60% of the data (different than the imputation model for non-reported trade), resulting in large differences between subsequent estimates. It may be preferable to readjust the timetable and produce estimates of higher quality.
- Can the enterprises reporting monthly be used for defining the parameters of the regression imputation for enterprises that do not report monthly? Usually the frequency of reporting is based on the size of the enterprise, and therefore, enterprises reporting monthly do not represent smaller enterprises reporting quarterly or half-yearly.

4. RELATED ASPECTS

- Imputation on a macro level as opposed to a micro (per-unit) level. This method involves estimating the total value of non-response and distributing this mass to cells of the table according to some kind of apportionment.

WP7: EVALUATION OF EDITING & IMPUTATION SUPPORTED BY ADMINISTRATIVE FILES

Olivia Blum, Israel Central Bureau of Statistics

1. SUMMARY

The paper emphasises that the error detection process requires either the implementation of logical rules or a reference-file against which to make comparisons. Similarly, the error correction process requires either a well-specified model for prediction, together with sufficient data to implement the model, or the use of an external true value for the imputation. While error detection can be improved by using administrative data sets, error correction often relies on them, if they are available and are of good enough quality.

There are three main mechanisms and sets of merits by which administrative data can support the edit and imputation process:

The first mechanism is by supplementing the collected information. At the editing stage, administrative records help with the error localisation process by identifying unexplained differences in the values of variables involved in the failed edit-checks. Administrative records also allow better model specification for imputation in two ways: by adding variables to the post-edited data file, not already in the data file, which can refine the selection of imputation classes and improve the pool of potential donors; or by adding records to improve the representation of smaller groups. The second mechanism is use as a reference file. Administrative records can support the editing process in error localisation, in confirming the values of variables failing the edit checks and by improving the identification of erroneous variables. Also, comparing the edited file and the reference file can minimise false-positive errors. If the administrative data are of a good enough quality then they can be used directly for imputation. For example, the population register holds date and place of birth and can be used to identify missing records in the edited data file and used for unit level imputation. The third mechanism is the use of administrative records as a continuous quality assurance process. This is achieved by a comparison of the processed data and the administrative data in order to identify missing values, errors in the collected data and errors generated by the processing operation. The merits include a significant reduction in errors and in the need for imputation.

The author describes some of the wider aspects of the editing process benefit from the use of administrative data: record linkage, coding and imputation of new variables. One use of record linkage is to confirm the identity of an individual in the data file, and once the identity is confirmed then the values of critical variables can be confirmed. A potential use of administrative data in complex coding is described, that is, to provide additional information in order to manually assign an appropriate code. The paper describes two aspects of imputing new variables: direct imputation of variables not on the data set and the replacement of the data source by administrative data. The first is described in terms of an index of well being while the author acknowledges that the second is more complex.

The paper discusses how administrative data can be used to enhance the dimensions of quality. The dimensions of quality related to the edit and imputation process are coherence within a dataset, consistency between datasets, comparability, completeness and accuracy. If a single file is being considered, then error detection is by a set of edit checks which are based on logical rules about relationships between the variables within the dataset. Resolution is normally based on the Felligi-Holt principle of minimum change. Hot deck and model based imputation methods are dependent on the richness of the attributes in the dataset and the correlations between them. Normally editing is at a

micro-level, macro-level errors are not easy to detect and, since no external data is available, correction is often not possible. Quality can be evaluated by internal cohesiveness and by completeness related to the attributes within the file, whether they are item non-response or erased errors, and completeness related to unit non-response within the sample.

Independent data sources, other surveys or administrative files, widen the scope of the information available to the edit and imputation process. They provide macro editing with the supporting information needed for detecting and correcting error. When more sources become available cold deck imputation can be expanded to more variables and has more candidate values to be imputed. Furthermore, the introduction of additional data, coming from independent sources enables a better-specified imputation model. The quality evaluation is more reliable, especially the accuracy dimension, since this wealthy information facilitates the development of an accepted reference-file.

The uniqueness of registers, with regard to other administrative sources, is the completeness of the frame and continuous updated information. Registers are reference-files and as such, they influence directly and indirectly processes and results. Hot-deck imputation is an example of an indirectly influenced process. After exhausting the cold-deck imputation with the support of the register, hot-deck imputation is engaged for missing values of variables not included in the registers. The register information improves the stratification of the population in the edited file, needed for hot-deck imputation to be performed within homogeneous strata. Other sources of information may do it as well, but registers have the advantage of containing the whole population, as censuses do in infrequent points of time. This attribute contributes to the completeness dimension when evaluating coverage of a relevant population.

The author discusses the unique challenges arising from the availability of several data sources at the data collection phase and questions what this means for the editing and other statistical processes. The discussion highlights the fact that new types of errors are introduced. The quality considerations in a multi-source data collection are considered in three stages: selection and integration of data sources to build a raw data file; edit and imputation of the raw data file; and evaluation of the end-file.

The main conclusion of this paper is that the boundary between editing and imputation is constantly moving due to the use of multiple sources of data. Administrative records support the detection and correction of errors and yet they are evaluated as potential source data by applying the same logic and rules that are used for editing. When developing a reference 'truth' file, the process and results of the selection of data that have passed pre-defined quality threshold is actually limiting the options that were opened for imputation.

2. MAIN ISSUES

- The paper discusses three mechanisms and sets of associated merits by which administrative data can support the editing and imputation process and further, can support record linkage, coding and imputation of new variables.
- Independent data sources widen the scope of information available to the editing and imputation process, the dimensions of quality relevant to different imputation procedures are stipulated by the different data sources.
- When editing is applied in a multi-source data collection, the considerations of quality vary dependent on the processing stage.
- The paper concludes by observing that the boundary between editing and imputation is constantly moving due to the use of multiple sources of data and some thoughts about future developments.

3. PROBLEMS

- How do we decide when administrative data is good enough to be used in the production of statistics? Is it possible to develop a set of universally accepted criteria?
- The availability of several sources at the data collection phase provides unique challenges for the editing process and integration can result in the introduction of new errors.
- Administrative records support the detection and correction of errors and yet they are evaluated as potential source data by applying the same logic and rules that are used for editing.

4. RELATED ASPECTS

- The clear need for standardisation/harmonisation of definitions and concepts to facilitate the use of register and other administrative data within the survey process.

WP8: METHODS ON EDITING AND IMPUTATION IN CREATING A LINKED MICRO FILE OF BASE REGISTERS AND OTHER ADMINISTRATIVE DATA

Svein Gåsemyr

1. SUMMARY

The paper starts by describing the infrastructure that has been developed in Norway to support electronic collection and processing of data. This has resulted in reduced respondent burden for both enterprises and households. The infrastructure and data collection strategy include: a computerised central administrative base register for persons, business and dwellings together with the assignment of a unique reference number for each of the three type of units. Also under development is the operation of a common portal for EDI reporting, methods for designing electronic questionnaires and common meta information systems.

The use of a common portal and electronic questionnaire directly effect methods for editing and processing. The use of electronic questionnaires means that respondents are actively involved in a large part of the editing process. Editing is performed interactively in real time which lead to respondents becoming users of the central NSI system. They expect an immediate reply to advise them whether their responses have been excepted or not. Electronic data collection has a significant impact on the internal workflow and responsibilities in respect of the data collection process. The author makes the point that these impacts occur instantaneously when electronic reporting is provided as an alternative to paper surveys.

Mixed mode data collection emphasises the need for a central repository for all incoming responses. The editing which is developed for electronic questionnaires has the potential to be reused both on file extracts and during the electronic capture and verification of paper responses.

Government policy in Norway is that an enterprise should only report a variable to government agencies once. Statistics Norway uses administrative data as a source for statistics where ever possible. Currently about 50% of official statistics collected are based administrative data. The author comments that the volume of data (records*variables) is much larger for administrative sources than for social surveys.

The main focus of the paper is on the editing and imputation issues associated with the creation of a system of linked files. These are highlighted by description of the problems and methods involved in

creating the job file. This was initially developed for the Norwegian register-based Population Census 2001 and formed from the integration of a number of administrative and statistical units. The integrated job file has a key role in developing integrated and coherent social statistics and also in the integration of economic and social statistics. The author warns that the job file might be the most difficult component to develop within a system of register-based statistics.

The paper provides a detailed description of the methods of data linking employed to identify the unit of employee job across the differing sources. This is followed by a description of the methods to ensure that there is consistency in the dates of starting and terminating each job when a person changes jobs. The analysis identified that there is a time lag between the employer's reporting of start and termination of a job to Social Security system and that more delays exist for termination than for start. The result of this practice is that, according to the Social Security system, a person might be registered with two active full-time jobs for the same day.

The analysis leads to the identification of differing patterns of employment. Some people in employment perform a secondary job in parallel to their main job. Some people perform both employee jobs and self-employed jobs in parallel or move from one type of job to another during the calendar year. The methodology to link employee jobs and self-employed jobs is described.

The paper briefly describes the editing and imputation process applied to those variables not used to identify the unit of job. The imputation of wage rates is discussed, and is based on the formulation of a set of groups of jobs that are expected to be homogeneous with respect to other variables. To date the average wage rate for a group of jobs has been used to calculate the hours paid for variable. The paper states that future imputation of wage rates, for example, will be based on statistical models and would include other data sources and include occupation. Currently, occupation is only recorded on the Social Security register and is missing for three groups in the job file: some 6.4% of those registered on the Social Security and Tax Agency have no occupation; those working for an employer based on the Tax Agency system only; and those self employed based on the Tax system only. The imputation of the item non-response from the Social Security register is based on complete records from within that source whilst the remaining records are imputed using data from the LFS. The paper makes recommendations about how the imputation of occupation could be improved.

2. MAIN ISSUES

- Electronic data collection has a large impact on the internal work flow and responsibilities concerning the data collection process.
- When respondents report electronically they need immediate feed back as they deliver their information. This introduces the need for common procedures to verify that their information has been received and they have completed their reporting duty.
- Record linkage integrates units of employee jobs and creates a linked file on the units of job; this plays a key role in developing integrated and coherent statistical systems but how can we be sure we have uniquely identified a unit across the sources.

3. PROBLEMS

- When item non-response is present within a data set formed from linked sources it is not clear which source should be used for imputation.

- The paper discusses challenges in the process of editing and imputation: how to ensure that the dating of events is consistent across sources, and methods for calculation and imputation variables of the linked file.
- When there is a delay in updating information on administrative registers there is a knock on effect for the production of statistics.

4. RELATED ASPECTS

- To support the introduction of electronic questionnaires respondents need immediate feed back when they supply information. This introduces the need for common verification procedures to ensure that information has been received and that reporting obligations have been completed.
- Given that the volume of administrative data is much larger for administrative data than for social surveys are there any implications for storage? Should NSI's hold redundant information or keep the minimum set for their purposes?

WP 9: THE USE OF ADMINISTRATIVE DATA IN THE ANNUAL SURVEYS OF RETAIL, WHOLESALE AND SERVICES

Carol S. King, US Census Bureau

1. SUMMARY

Administrative Records at the US Census Bureau are used for constructing frames for censuses and surveys, improvements in measures of size for sampling, coverage issues for births and deaths of businesses and firms, and imputation for non-response to a census or survey form. The business register (BR) is constructed from various sources: Internal Revenue Service (IRS) , Social Security Administration (SSA) and the Bureau of Labour Statistics (BLS). The goal is to use administrative record data to improve statistical programs and reduce response burden. 3 surveys are examined: Annual Retail Trade Survey (ARTS), Annual Trade Survey (ATS) and the Service Annual Survey (SAS) which is subdivided into 6 surveys. The above surveys are composed of units selected from a probability based stratified random sample for employer businesses and supplemented with administrative data for nonemployer businesses. Employer Businesses are selected in strata with certainty (probability of 1) and noncertainty strata.

A study in 1998 was carried out to evaluate the use of administrative receipts in place of survey responses for the noncertainty businesses in the above 1996 surveys. The study showed that the survey data could be replaced for businesses below specified payroll cutoffs. This reduces both collection cost and response burden. Only slight changes were found in the overall estimates from the surveys.

For the 1999 SAS survey, the Census Bureau used administrative receipts for small businesses when few data items other than revenue were collected. For some industries, both SAS forms and administrative receipts data were available, including Travel Agencies and Tour Operators. Nonignorable changes in total revenue estimates were obtained when using the administrative data which could be a result of the inclusion of commissions on the administrative receipts which are excluded from the revenue data. All businesses with an annual payroll less than a cut off appropriate to the industry were not mailed SAS forms. If administrative receipts were not available for a non-mailed form, revenue was imputed by an estimated revenue to payroll regression coefficient from businesses in the same industry which was computed from the 1997 Economic Census Data. Besides revenue,

other values are imputed using data from reporting units that responded to the 1999 SAS. For SAS-transport, 4% of the units had a payroll below the cutoff making up 0.71% of the total receipts but only 0.45% had administrative receipts. For SAS-general, 19% of the units had a payroll below the cutoff making up 9.3% of the total receipts but only 5.7% had administrative receipts.

For the 2000 ARTS, small businesses classified in Accommodation and Food Services were excluded from the mailing based on the annual payroll cutoffs from the 1997 Economic Census. The nonmailed businesses contributed about 15% to the total sales of the industry and the corresponding administrative data percents were about 7%. Imputation of sales for this survey uses the data from the Monthly Retail Trade Survey if available, otherwise other imputation methods are used including the substitution of available administrative receipts.

A comparison was carried out between administrative inventory from the IRS to reported inventory from the 1999 ARTS. For about half of the businesses, the administrative data was roughly 100 times that of reported data as a result of cents keyed as dollars. This was automatically corrected. Edit rules for acceptable ranges and an inventory to sales ratio based on the ARTS survey was used to edit the administrative data and an analysis of replacing administrative inventory data for non respondents as well as for small businesses was carried out. The estimates of total inventory had minimal differences when incorporating the administrative inventory, though the CV's were slightly increased.

Expense data was also examined to replace total expenses collected in the SAS. The analysis showed that after editing expense data, it was feasible to use administrative expenses as an imputation method for SAS. Substituting administrative expenses for survey expenses had small effects on the total estimates and standard errors.

Future research will improve the editing methodology of administrative sources, study the agreement between individual sampling units on reported and administrative data (and not just the effects on the totals), study and understand the administrative data and why some businesses with reported sales do not have administrative receipts in the business register, and finally the use of administrative data for enterprises with multiple businesses.

2. MAIN ISSUES

- Thorough description of the use of administrative sources in many of the Census Bureau's business surveys, thus reducing costs and respondent burden.
- The impact on the quality of the data after replacing with administrative data is examined through totals and their variances.

3. PROBLEMS

- There is no mention in the paper on whether administrative data has to be preprocessed and edited before use. Is it always reliable and no adjustments are needed?
- Must understand the administrative source, what factors are included in the administrative records that are not included in the reported survey data. For example, it was shown large discrepancies between administrative and reported data for tour operators and travel agencies. How was this resolved?
- Is administrative data always used as described even if the quality may be low?
- How is the variance (CV's) of the totals calculated when using imputed administrative data? This is important to understand since it is a measure of the quality of the end data.

4. RELATED ASPECTS

- Variance estimation for totals obtained from both survey data and imputed data based on administrative sources.
- How to process and edit administrative sources to make it fit for use.