Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Statistical Office FSO

Book of abstracts

2011 European Establishment Statistics Workshop

12 – 14 September 2011, Neuchâtel, Switzerland

Abstracts

Small Area Estimations in the Industrial Survey

Iosune Azula, María Victoria García Olea, Patxi Garrido, Haritz Olaeta, EUSTAT, Spain

Aware of the increasing demand of small area estimations, Eustat started publishing estimates of the main variables of the Industrial Survey for comarcas (i.e., administrative clusters of municipalities) in 2005. There are 20 such clusters within the three provinces of the Basque Country, being some of them extremely small in terms of industrial activity.

The Industrial Survey is designed to provide estimates at province level, so that the use of small area estimation models is necessary in order to obtain reliable estimates. The estimation process combines both a Fixed Effects Model and a Linear Mixed Model in order to borrow strength from wider areas where the available sample information is adequate.

In this work, the small area estimation process is explained in detail together with the specially relevant issue of coherence with the estimates that provides the Industrial Survey at province level and for the whole of the Basque Country. The last estimates of the 2009 Industrial Survey will be used to analyse the main issues and to address the difficulties that arise in the necessary integration of different estimation processes at local and regional level.

Keywords: Fixed Effects Models, Linear Mixed Models, Local Estimates, Coherence

Combined Firm Data for Germany

Stefan Bender, Anja Gruhl and Tanja Hethey-Maier, Institute for Employment Research (IAB)

In many countries the everyday practice of National Statistical Institutes (NSIs) are to establish surveys for individuals, households or firms, which are not organized in an integrated system. In some countries – like Germany – the collection of data are done by different data producers with different data generating processes (like process generated or survey data) without any possibility of linking the data. Because of data protection reasons it is often the case that combined datasets out of this processes are forbidden by law. This leads to an increase of response burden by the units of interest and to an unknown data quality, which is – maybe – not as good as expected.

To overcome this limitations the KombiFiD project (Combined Firm data for Germany) tried to merge company data collected by the Federal Statistical Office, the Deutsche Bundesbank and the Federal Employment Agency. One important criterion of the combined datasets is the quality of the links. First we will give an overview about the combined data, because not every company will be part of every single data source. In a second step we will identify companies with only one work place to be sure, to have in every single data set the same company. In the third step we will compare variables that originally appeared in more than one of the linked datasets with identical content. At the end of the day we will have measures for problematic variables in one or more of the data sets and for hard to measure companies. In the final step we will discuss – in the light of our results – whether the daily practice of a separated collection of data is adequate or the NSIs should push forward in an integrated system for collecting their data.

Using survey data collection as a tool for improving the survey process

Silvia Biffignandi (University of Bergamo), Giulio Perani (Istat) Antonio Laureti (Istat)

The paper focuses on how paradata information can improve survey methodology and quality.

Collecting data via Web allows for server-side paradata (i.e. log files describing access time, number of accesses and so on) and client-side paradata (namely the answering process within the questionnaire, i.e., insight into the sequencing and completeness of responses. Respondent behaviour is traced on each Web page as they answer the survey). If we consider the business surveys, business register data are available, too. This data may be linked to survey data. Therefore, an integrated set of data becomes available and may be used not only for descriptive purposes of substantive information, but for improving the data collection process at different steps of the survey.

This paper is analyzing data collected using a web questionnaire in Italy (Innovation survey on Businesses, carried out from Istat) and discusses how survey response, client-side paradata, survey-side parade and auxiliary variables of the business registrer can be allocated in the framework of the survey process.

Abstracts

Cognitive aspects of business surveys – presentation of thesis proposal

Irena Bolko, Faulty of Economics, University of Ljubljana

My thesis will focus on the interdisciplinary field of survey methodology and cognitive psychology. The findings of cognitive psychology have already proved useful in survey methodology but to my current knowledge, they have not been applied to those steps of statistics production that heavily rely on expert knowledge.

Paper will summarise the first part of my thesis, namely theoretical background with focus on cognitive psychological findings and their potential extension on business surveys, and propose the design of the empirical study that would be conducted afterwards.

Two phases of statistical production in business surveys will be studied, namely filling in a questionnaire and editing collected data. Given that survey respondents as well as data editors are considered experts that use specific expert knowledge, I will prepare review of research findings on experts and expert knowledge with special attention given to two key elements of cognitive system – short-term (working) memory that is involved in decision making processes (e.g. formulating the response, editing decision) and long-term memory where knowledge is stored and later in the process of decision making retrieved in working memory. Along these lines I will be interested in several questions, e.g. how is information structured in expert knowledge, what is the role of implicit knowledge and use of heuristics in experts decision making.

By applying the theoretical findings into empirical study I would like to provide a starting point for improvement of questionnaires (where cognitive processes of experts –respondents would be taken into consideration) aiming for better data quality, as well as optimising the work of experts –editors.

Development of an integrated business statistics program

Jean-François Carbonneau, Statistics Canada

In 2010, Statistics Canada introduced the Corporate Business Architecture Initiative (CBA) to conduct a comprehensive review and revision of Statistics Canada's business architecture in order to harvest efficiency on ongoing operating costs, enhance quality assurance through implementation of more robust systems and processes and improved responsiveness in delivery of new statistical programs. Combined with the challenge of improving relevance while maintaining quality, this requirement will force Statistics Canada to achieve a much greater efficiency than at any time in the past. Initiated in 2010, the IBSP is to be based on the existing model the Unified Enterprise Statistics program (UES) which had been conceptualized in 1996 and implemented through different phases up until 2008.

Currently the Unified Enterprise Surveys Program (UES) is applied to 59 annual business surveys and covers the sample and questionnaire design, edits and imputation, allocation and estimation processes. By 2016 close to 130 annual and sub-annual surveys in ten different programs will be integrated into the IBSP framework. The surveys under the umbrella of the IBSP will use the Statistics Canada's Business Register as a common frame. They will adopt electronic data collection as the principal mode of collection, focus on a common sampling approach, use tax data universe for the estimation of financial information, apply a common editing strategy for automated and manual editing, establish an earlier collection cut-off and expand on the use of data warehouses. All processing methodologies will be driven by a common metadata framework and they will share common tools to analyse, edit and correct data.

A parallel run will be conducted on the 2011 production data. More specifically we will assess the newly introduced functionalities such as 1)Monitoring collection and analysis progress and prioritizing collection and analysis efforts using Quality Indicators, 2) Replacing financial Survey data with Tax data for simple units, 3) Impact of an iterative process on timeliness and data quality, 4) the use of administrative data in replacement of survey information related to the allocation process and 5) providing follow-up instruction to collection services.

Since this parallel run will begin in June 2011 a summary of our findings will be shared with the work session participants.

Abstracts

Optimal Allocation in the Multi-way Stratification Design for Business Survey

Demetrio Falorsi, Paolo Righi, Piero, Italian National Institute (Istat)

Commonly, the business surveys produce estimates for a huge number of domains that define two or more partitions of the target population. When domain indicator variables are known at population level then a multi-way (or incomplete) stratification design can be used, guaranteeing a sample with planned size in each domain. The multi-way approach has some advantages with respect to the standard approach (using a one-way stratified design where the strata are obtained combining the domains of the partitions) such as: the sample allocation is more efficient (smaller sample size with same sampling errors); the response burden is reduced both in a given survey occasion and considering several survey occasions (for the combining strata with small population sizes the one-way design selects with high probability or sometime with certainty some business units in each survey occasion producing a great statistical burden). The paper shows a procedure defining an optimal sample allocation for the multi-way design according to the definition proposed by Bethel (1989). The procedure is suitable in the multivariate-multidomain case. An algorithm implementing the procedure is also explained. Furthermore, it is suggested of using the Cube algorithm (Deville and Tillé, 2004) to achieve the multi-way random sample selection. Then, the paper gives the tools to really apply the proposed approach on business surveys.

Survey Data and Administrative Data Integration for the Estimation of Structural Business Statistics Preliminary Results

Salvatore Filiberti, Italian National Institute (Istat)

According to the Structural Business Statistics (SBS) Regulation (EC) No 295/2008 Member States have to provide Eurostat with information about a subset of economic variables at 3-digit NACE Rev. 2 breakdown 10 months after the year of reference. Such as timeliness compels Italian National Statistics Institute (Istat) to use only statistical and administrative data sources that effectively are available at the time of data transmission. Survey data are collected by the sample survey on the enterprises with 1–99 persons employed and by the census survey on the enterprises with 100 and more persons employed, but those data are partially available for the year of interest at the time of preliminary data transmission. Administrative data are represented by corporate enterprises (their financial statements are provided by Chambers of commerce) and by enterprises with employees (social security data) so, coverage of the population of interest is not complete. Also, Business Register is not available for the year of interest and, consequently, the traditional estimation methodology cannot be applied.

The choice of the approach adopted for obtaining preliminary results is then constrained to data availability and to their characteristics. Aim of the paper is to suggest improvements concerning some crucial phases of the SBS preliminary results estimation procedure using auxiliary information. In particular, an application regarding the administrative data treatment and a revised estimation methodology are suggested.

Construction of Full Time Equivalent for the Swiss Business Frame

Monique Graf and Jann Potterat, Statistical Methods Unit, Swiss Federal Statistical Office

Business statistics in Switzerland face a paradigm shift. The business census was held for the last time in 2008. It will now be replaced by the use of registers and complementary surveys. The two main sources are the business register (BR) that provides information at the enterprise and local unit levels, and the social security register (SR) that provides information at the enterprise level only. The record linkage (on the basis of the enterprise name and address) between BR and SR is ongoing and should be finished by the end of 2011. The business register will record the new businesses and update the economic activity. The social security register will provide information about gender, employment and wages at the level of the enterprise. A survey called «profiling» will (among others) allocate employment to local units for enterprises with more than one local unit.

The employment data recorded in SR are the months worked per employee. From this information, it is possible to deduce the number of employees per month and gender, but not the corresponding full-time equivalents (FTE). Full-time equivalents per gender (FTE) will be reconstructed using a model based on the combined register and results of the Quarterly Survey of Employment (JobStat). This survey gives the total employment and FTE by gender for approximately 36'000 enterprises. The model will be first applied for production in 2013 on the 2011 data.

The purpose of the presentation is to describe the foreseen model, and methods in place for its validation.

Abstracts

Key elements of quality frameworks, to be applied to statistical processes at NSI's

Robert Griffioen, Arnout van Delden and Peter-Paul de Wolf, Statistics Netherlands

As part of the BLUE ETS programme, Statistics Netherlands started a research project in 2010 on how to improve and control quality in a chain of interlinked statistical processes. At National Statistical Institutes (NSI's), the outcomes of statistical processes may be reused in other statistics, for example as input or for weighting purposes. For example, in economic statistics at Statistics Netherlands, outcomes of turnover statistics are input to statistics on final consumption, on the production index and on structural business statistics, which in turn are input to National Accounts.

The core question of the project is how we can translate quality needs of users of statistical outcomes, to quality requirements that can be measured and controlled during the production of statistical processes, including the requirements to the inputs. The emphasis of this work is on improvement of the complete chain, not only on a single process or single data storage point. To find ideas, we studied literature on existing quality systems in different domains: NSI's, food industry, logistics and quality control systems. In the current paper we present the results.

We explain the existing general frameworks: Total Quality Management (TQM), Six Sigma, LEAN and business process management (BPM). From these frameworks we extract key elements for a quality framework. Next, we relate these elements to data processing chains, where we focus on three objects: a process, a data storage point and the statistical chain itself. For each of these three objects we discuss what is meant by quality and how quality can be improved. Finally, we synthesize the findings into first ideas for a quality framework for a statistical chain. In a next paper, we wish apply the framework and work out some specific examples of statistical chains at NSI's.

A mixed mode survey on book prices among booksellers

Beat Hulliger, University of Applied Sciences Northwestern Switzerland

To establish whether the actual book prices in Switzerland are different from list prices in 2008 a survey among booksellers on actual prices of a basket of book titles was carried out. The main survey was carried out online and obtained a response rate of 37%. An additional survey was administered for quality control reasons where enumerators visited book shops and recorded prices of the titles of the basket. While the online survey was exhaustive, for the quality control survey in book shops a regionally stratified sample was selected. The response rate for the quality survey was 79%. Due to the low response in the online survey the quality survey results were joined with the online survey, taking into account the response and sampling probabilities in both surveys. Since the basket of book titles was considered too large to be answered by a bookshop the basket was split into 6 sub-baskets and each bookshop was attributed a sub-basket list, thereby reducing the response burden of a bookshop. Exceptions were made for the largest bookshops, which were asked for prices of the complete basket. Thus from the point of view of a book title also the online survey was a sample survey. Data preparation had to cope with missing values and outliers and with the reconciliation of the answers from a few shops which were in both surveys.

Abstracts

Designing Linkage between Patents and Business Registers: the Italian Experience

Daniela Ichim, Giulio Perani, Giovanni Seri, Italian National Statistical Institute (Istat)

The paper describes the record linkage strategy followed at the Italian national statistical institute to match micro-data on patent application from the international database PATSTAT with the data available from the Italian Official Business Register (ASIA).

The target data in PATSTAT are the applicants based in Italy registering patent/s in the period 1985–2010. Patents applicants can be ‹individuals› or ‹establishments›. In this last category we aim at identifying business enterprises who were active (as recorded in ASIA) in the period 1989–2008. The wishing output of the linkage process is, for each patenting enterprise, a pair composed by the ‹applicant identification code in PATSTAT' and the ‹enterprise identification number in ASIA›. This last allows for accessing the repositories of the official statistical data and, therefore, linking economic data to patenting enterprises. Statistical analysis such as: identifying the premises of patenting propensity; evaluate the impact of patenting on the enterprise profitability; etc. can be then performed.

On the methodological side, linkage of patent data has to rely on the ‹applicants' names›. Consequently, a great effort has been put in the pre-processing phase of the process to standardise the applicant/enterprise names and extract the ‹legal form› from the name string. In order to face computational problems (ASIA 2008 counts around 4500k records) an approach based on the assumption that at least one word in a name is registered in the same (a similar) manner in both archives has been followed: a ‹neighbourhood› has been defined as the set of ASIA enterprises having at least one word in common with (or similar to a word of) the patent applicant name. Reducing the search space to a neighbourhood by blocking for patent applicant produces ‹small› problems that can be easily thought to be solved as a 1:1 linkage.

Encouraging results has been obtained investigating most of the ASIA archives. In the next future, different approaches (probabilistic record linkage) will be tested and further archives (such as the List of companies' owners and partners) will be used to identify ‹individuals› (names without legal form).

What makes business statistics different?

Wim Kloek, Eurostat

In many statistical institutes the world of business statistics is distinct from statistics on persons and households in organisation, culture and language. The question is what are the real differences from a methodological point of view, and what do these differences suggest on the organisation of official statistics. The distinction is close but not identical to the distinction between economic statistics and social statistics.

Main differences to be discussed:

- The statistical units: business statistics has a system of statistical units, units have an internal organisation and the boundaries of the unit are not always self evident;
- Skewed distributions for most of the target variables (e.g. value added);
- Business statistics usually collects ‹hard› accounting data;
- The pressure on timeliness of the data.

There are also secondary effects to be taken into account. For instance, the skewed distributions make a design with equal inclusion probabilities inefficient. For small units the option of modelling becomes prominent; for the largest units a take all strategy will automatically produce longitudinal data, which offers several new options in the statistical production process.

The differences should be considered in the design throughout the statistical production process; this will be illustrated with some examples. On this basis we will reflect on the organisation of official statistics from the integration perspective of reuse of methods and tools and the industrialisation of production processes, both at the level of the statistical institute and at the level of methodological support within the institute.

Abstracts

Possibilities of exploiting administrative data in short term statistics in Poland

Jacek Kowalewski, Statistical Office in Poznań

A dynamic development of new techniques, an ever closer integration within the European Union as well as a necessity to reduce respondent burden all call for a modification of the data collection system in the field of economic statistics, which includes most surveys conducted by Central Statistical Office. This need also means a wider use of administrative registers. A team of employees of Statistical Office in Poznan, aided by employees of Statistical Office in Katowice and statisticians from University of Economics in Poznan conducted a research program aimed at exploring the possibilities of using existing administrative registers in short term statistics. The scope of work comprised an assessment of potential usefulness of administrative sources, an analysis of ways to decrease response burden for some companies and improve data completeness and quality. The last objective involved mainly assessing the possibilities of exploiting register data for purposes of calibration in the case of incomplete data.

Seasonal Adjustments: Causes of Revisions

Øyvind Langsrud, Statistics Norway

X-12-ARIMA is a frequently used tool for seasonal adjustment. To find the best decomposition into trend, seasonal and irregular components several modeling decisions have to be taken. When a new data point is available we also have new information about the seasonal pattern and the decomposition model can be updated. The consequence is that seasonally adjusted data are revised. Choosing seasonal adjustment methodology can be viewed as a question of balancing the requirement of optimal seasonal adjustment at each time point against the requirement of minimal revisions.

In this paper, history analyses of 52 Norwegian economic time series has been conducted. Seasonal adjustment revisions are mainly caused by revisions of seasonal factors. Revisions of prior adjustments (calendar effects) are less important. This paper demonstrates how several modeling choices (ARIMA model, trading day, holyday treatment) affect revisions. The treatment of outliers (extreme observations and level shifts) is related to both prior adjustments and seasonal factors. When a completely automatic procedure for detecting outliers is applied, re-identification of outliers leads to big revisions. This paper demonstrates how revisions and out-of-sample forecasts (quality of model) are affected by the outlier detection limit. The analyses were made by running X-12-ARIMA via the R programming language.

Estimating structural business statistics based on administrative data: the case of the Italian small and medium enterprises

Orietta Luzi, V. De Giorgi, U. Guarnera M. Rinaldi, G. Seri, Italian National Statistical Institute (Istat)

The Italian National Statistical Institute has recently started a re-design project in the area of structural business statistics (SBS) with the main objective of reducing statistical burden on enterprises and statistical production costs. Key elements of the project are widening the use of administrative information on businesses and moving from a traditional stove pipe model for the statistical production to an integrated model where administrative data represent the information core, and statistical surveys aim at estimating sub-populations/variables which are not available in external archives. The project also aims at consistently reduce non response rates and eliminating survey questionnaires redundancies.

In the paper, some of the developments and results for the Small and Medium-sized Enterprise survey (SME) are illustrated. The SME sample survey collects information mainly on profit-and-loss accounts of enterprises with less than 100 employees. The SME sample consists of about 105,000 enterprises. Available administrative sources on target population are Financial statements, Tax Authority sources, and the Italian Business Register.

Incorporating administrative data in the estimation process implies the re-design of the organizational and methodological survey framework. Besides the analysis of the usability of administrative data in terms of information contents and population coverage, in the paper an experimental evaluation of the potential biasing effects due to integrating administrative and survey data for estimating SME parameters is illustrated. The main focus is on methods adopted for measurement error detection (in particular, identification of influential errors) and for non responses imputation. The study considers both key survey variables directly available in external sources (e.g. Turnover and Number of employees) and variables which need a modelling effort based on existing administrative auxiliary information (e.g. Change in stocks of goods and services). Parts of the results have been obtained in the context of the ESS-net on the use of Administrative data for business statistics, which has been funded by Eurostat in 2009 in the framework of the MEETS Program.

Abstracts

The Pros and Cons of Automatic Data Extraction

Ken Moore and Steve MacFeely, Central Statistics Office,Ireland

In 2005 the Central Statistics Office in Ireland introduced, with the cooperation of all the major payroll (software) systems providers, a facility whereby statistical returns for the Earnings, Hours and Employment Costs Survey (EHECS) and the National Employment Survey (NES) could be returned automatically via XML.

In the 2010 the ambitions and design of these surveys were reassessed following a review to understand why the original objectives were not being achieved. From this review a number of important lessons were learned.

This paper presents a frank summary of the strengths and weaknesses (pros and cons) of this approach to data collection.

The EuroGroup Register as a coordinated frame for the European Statistical System

Enrica Morganti, Harrie Van der Ven, Italian National Statistical Institute (Istat)

The Eurogroup Register (abbreviated EGR) is the statistical register of multinational enterprise groups created by the cooperation between Eurostat (the statistical agency of the European Commission) and the statistical authorities of European Member States and EFTA countries. The EGR contains structural economic information on enterprises that are part of multinational groups with an interest in Europe.

The EuroGroup Register is the first shared statistical register created at European level integrating information coming from administrative sources and statistical sources from 27 Member States and 4 EFTA Countries at micro level.

It is foreseen to become the common coordination frame for all European statistical authorities, National Statistical Institutes and National Central Banks, for surveying and the production of consistent statistics on globalization, since it will offer statistical compilers access to integrated and up-to-date register data on those enterprise groups which have relevant transnational operations (financial and non-financial) in at least one of the European countries.

The paper addresses the need for a coordinated frame for producing Inward and Outward FATS statistics by European countries, describing the scope and the structure of the two, and discusses how the EGR could improve the overall quality and the statistics produced.

Keywords: Enterprise groups, globalisation, foreign affiliate statistics, coordinated statistical frame.

Collaborative networks in the European Statistical System

Jean-Marc Museux, Eurostat

ESSnets are projects on the basis of collaborative networks in the European Statistical System (ESS), co-financed by Eurostat. The results of the projects should come to the benefit of the ESS, not only to the benefit of the participating countries. For this reason the work is usually of a methodological nature. The purpose of this presentation is to offer an overview of past and ongoing projects relevant to the domain of business statistics. Most of these projects are part of the framework Modernisation of European Enterprise and Trade Statistics (MEETS), but some of the projects are of a more general methodological nature, other projects have a more research character and are in the BLUE-ETS (Enterprise and Trade Statistics) framework. This is also reflected in the participants list of this workshop; most participants are somehow linked to one or more of such projects. The aim of the overview is to produce more synergies between the projects. We will also present some ideas on future project with the intention to profit from the feedback in the workshop.

Abstracts

Using the Commercial Register to Reduce Response Burden in Economic Structural Statistics

Haritz Olaeta, María Victoria García Olea and Patxi Garrido, EUSTAT, Spain

The increasing burden that respondents are undertaking increases the difficulties that might arise from high non-response rates such as non-response bias estimation and the analysis of effects of different imputation techniques.

Eustat, the Statistical Office of the Basque Country, is really concerned with this phenomenon and is actually working on a long-term project that deals with how to increase the quality of our Economic Structural Statistics without increasing the response burden of our respondents. Actually, the ultimate goal is to increase the quality (not only in terms of errors but also in terms of timeliness, non-wasteful duplications, etc.) by means of the use of data from different sources such as administrative registers.

The Commercial or Mercantile Register due to its nature in Spain, constitutes and important data repository for Economic Structural Statistics. It provides exhaustive information of many of the variables of interest for an important part of the population under study in Economic Structural Statistics. It does not include, for instance, self-employed people, that constitute an important part of the reference population in most of the Economic Structural Statistics.

In this work, an estimation procedure based on a composite estimator, a linear combination of a direct ratio estimator and an indirect synthetic estimator is used to estimate variables for the population not covered by the Commercial Register. This estimated enhanced Commercial Register could derive in a substantial reduction of the size of the samples for many Economic Structural Statistics and, at the same time, assure a quality and efficiency improvement. The shortcomings of the data contained in the Commercial Register are described such as sectors and companies not included in the database, localisation of the economic activity of companies, etc. Attention is focused on the yearly Industrial Statistics.

Keywords: Commercial Register, Reduction of Response Burden, Composite Estimation

Annual Bookkeeping Report as the primary administrative source for the production of structural business statistics: current experiences and future plans

Tiina Pärson, Statistics Estonia

By the Estonian Business Act all the commercial undertakings, non-profit associations and foundations are obliged to provide their annual bookkeeping reports only in electronic format since 01.01. 2010.

The main objective is to simplify the current data collection system, which is burdensome for data providers as well as for the state. There has been the possibility of submitting the annual bookkeeping reports electronically, albeit in format (PDF, RTF), which requires time-consuming data entering at the Centre of Registers and Information Systems. At the same time, there was double data collection as the other state agencies (statistics, agricultural registers and information board) collected the same information by questionnaires and declarations.

Since January 1 2010 all state agencies (statistics, tax office etc) are obliged to use the information provided in electronic format to the Centre of Registers and Information Systems and the double data collection should be finished. The XBRL was chosen as data transmitting standard for annual bookkeeping reports. The new data collection format enables the immediate data processing without additional data entering etc. The national XBRL taxonomy was created and it follows Estonian accounting principles.

On the process of taking use the administrative data current tasks have been done:

- created the correspondence tables to transmit data from administrative data files to statistical data files as the XBRL uses the specific coding system of variables,
- created rules, formulas and tables to calculate statistical variables using information from annual bookkeeping reports as bookkeeping variables are not always directly usable,
- harmonized statistical surveys using information from annual bookkeeping reports to facilitate the data transmission from administrative data sources.

Using the created correspondence tables, the information from electronically transmitted annual bookkeeping reports was used for imputation of item and unit non-response of reference year 2009. A project to complement the eSTAT (Statistics Estonia electronic data submission environment) is in a way, which enables to preload the information received from administrative sources. The information from annual reports will be preloaded and data provider has to fill in only the gaps, and the response burden of enterprises will be diminished

Abstracts

considerably — information not available form annual report and the response burden of enterprises will be diminished considerably. eSTAT will be complemented since the reference year 2011 data collection. For enterprises in sample the information is displayed (prefilled) in electronic statistical questionnaires. Thus enterprises have to provide only statistical information not available from annual bookkeeping reports, the double data collection will be avoided and the response burden of enterprises will be diminished.

Return of experience on the Swiss survey coordination system

Lionel Qualité, Swiss Federal Statistical Office

In 2009, the Swiss Federal Statistical Office (SFSO) has introduced a coordinated sampling system to select its business survey samples. This system allows for the selection of samples with maximal positive coordination or maximal negative coordination in a dynamic population. Through its use, panels and rotating panels can be selected, as well as one-time surveys, and the response burden is spread as evenly as possible on the whole population.

While the system gives optimal coordination, its impact in term of burden distribution is still modest, as it should be on a population that is split between large units that are exhaustively selected and where no coordination is possible, and little units with very small inclusion probabilities. However, this effect will increase with the number of surveys conducted.

The fact that this system can be used with a dynamic population allowed us to update our sampling frame in 2010, and use recorded demographic events in this process. Our rotating panel for the production of the value-added statistic has then been updated effortlessly in spite of some businesses having changed of ‹strata›, and of a modified sampling design.

Finally, we will see how we intend to modify our planification procedures in order to cope for the increased size variability of our samples, due to the fact that our coordinated selection system produces Poisson random size transversal samples.

Estimation of GDP by municipalities

Marta Salvador, María Victoria García Olea, Alaitz Gallastegi, EUSTAT, Spain

EUSTAT started publishing estimates of the GDP by municipalities in 1996. Thereafter, estimates for 2000 and 2005 have been published. For 2010 estimates a full revision of the whole estimation process has been proposed that is described in this work.

In the Basque Country there are 251 municipalities grouped into 20 clusters in the three provinces. Most statistics in Eustat are designed to offer accurate estimates at province level. Therefore, it is by all means necessary to make use of a battery of other types of available economic information and indicators in order to estimate the GDP at municipal level by economic main branches (Primary Sector, Industry, Construction and Services).

Due to the high demand of statistics information at local level, Eustat has decided to offer local estimates of the GDP at a higher frequency, every two years. In addition, with the adoption of Nace Rev. 2, estimates for more sectors will be published.

The estimation process is highly based on composite indicators for each of the 251 municipalities that compile information from different sources to obtain indirect estimates for each of the 87 sectors (Eustat´s own classification). A wide range of sources are used for each of the sectors, both register and sample based. Coherence with small area estimations that Eustat provides for some sectors at municipality clusters and with the estimates at province level from the Economic Accounts is achieved by benchmarking.

In this work, the attention focuses in the methodology used for constructing composite indicators and benchmarking techniques in some of the 87 sectors. Several difficulties that have arisen in the estimation process will be addressed.

Keywords: Composite indicators, Indirect estimation, Local Estimates, Coherence, Benchmarking, Multi source data (administrative and sample).

Abstracts

Methods for estimating Structural Business Statistics variables not available from administrative sources

Ria Sanderson on behalf of WP3 of the Admin data ESSnet. Office for National Statistics

We present results from Work Package 3 (WP3) of the ESSnet on the use of administrative and accounts data for business statistics. The aim of WP3 is to recommend estimation methods for variables not available from administrative data sources. This work is in part motivated by the common circumstance of statistical offices desiring to, or needing to, replace survey data with administrative sources. Suitable administrative data are not always available for all variables of interest, so the focus of WP3 is on recommending estimation methods in the case where administrative data sources cannot directly replace survey data.

We report the results of the investigations carried out within WP3 on a number of Structural Business Statistics (SBS), and on the Short Term Statistic (STS) «New orders». We describe the evaluation criteria applied to the methods, and identify the key methods of use to statistical offices, including the administrative data requirements of the proposed methods. We present our recommendations of estimation methods for a first wave of variables, which includes the STS variable «New orders», and the SBS variables «Payments of goods for agency workers», «Purchases of goods for resale in the same condition as received», «Number of employees in Full time equivalent» and «Changes in stocks of goods».

Sampling coordination of business surveys conducted by INSEE

Olivier Sautory, INSEE

The issue of coordination of samples in the field of business statistics has led to many theoretical works. Coordination is often seen as a method to reduce the statistical burden (negative coordination), but it is also used to obtain an overlap between samples, in panel updating for example (positive coordination).

The method presently used at INSEE for negative coordination is based on the use of random numbers, drawn according to a uniform distribution in [0, 1]. In a given stratum, the n units with the n smallest numbers are selected. After the draw, a rotation is carried out on the random numbers, which gives the units that have just been selected a lower probability of being selected in the next drawing, while maintaining good properties to classical statistical estimators. Other schemes are used to achieve positive coordination.

Ch. Hesse (INSEE, 2001) has proposed a method which generalizes the current technique: the random number is replaced by a «coordination function», which is a function which transforms the random numbers, and has the characteristic of preserving uniform probability. This function changes with each selection, depending of the desired type of coordination. This leads to an automatic process for selecting separate samples or updating panels, which takes into account the cumulative response burden over several samples. The paper will present the first experiments of this method, which can be used with Poisson sampling and stratified simple random sampling.

Spatial robust small area estimation applied on business data

Timo Schmid, Ralf Münnich and Thomas Zimmermann, University of Trier

Economic and political decision processes are increasingly based on specific indicators and other statistical information. Nowadays the necessity of developing regional indicator values or disaggregated values is evident in order to allow for regional or group-specific comparisons. Surveys which shall deliver the necessary information for these indicators, however, are generally constructed for larger areas, e.g. countries or NUTS2 domains. Hence, sample information on lower levels such as NUTS3 is rarely available so that classical estimates lead to high variances of the estimates.

Applying small area estimation methods may lead to highly improved accuracy of the estimates of interest. Especially in business statistics outliers in connection with small sample sizes lead to severe problems while applying standard models which are based on normal assumptions due to the high sensitivity of the model estimates towards these influential units. One way to overcome these peculiarities is the application of robust methods.

Two such robust small area methods are the robust EBLUP estimator and the robust M-quantile approach. But in business data, spatial dependencies often occur, so there is a need to enhance these models, which is already done for the M-quantile approach. In this talk we present an overview of the recently used robust small area methods and present a spatial extension of the robust EBLUP estimator and its MSE estimation.

Abstracts

To start with, the spatial and non-spatial robust estimators are compared by means of model-based and design-based simulations. In model-based simulations we test the performance of the methods when the outliers occur or not and investigate the performance when a spatial dependency exists within the data or not. Finally, the methodology is tested in a design-based simulation study using BLUE-ETS data.

The research is conducted within the BLUE-ETS project which is financially supported by the European Commission within the 7th Framework Programme.

Sampling error estimation – SORS practice

Rudi Seljak, Petra Blažič, Statistical Office of the Republic of Slovenia

Statistical results, derived from the data of the sampling surveys, are by definition «contaminated» with the sampling errors. Although precision (determined by the sampling error) is in the modern understanding of quality assessment just one of the quality dimensions, it still represents a strong indicator of the reliability of the provided information. It is hence important that even in the tight schedule of the official statistics production these errors are estimated and presented on the regular basis.

It is well known that the sampling error estimation can be a complex tax, especially in the case of non-linear statistics or complex sampling designs used. Therefore many times in the case of the regular statistical production, the estimation is done by using pragmatic approach with certain degree of simplification.

In the paper we present the general application for sampling error estimation, developed at the Statistical Office of the Republic of Slovenia (SORS). The application is build on the bases of so called metadata driven principle, meaning that there is one general program code which is then for the particular survey parameterized through the (process) metadata tables. The core part of the application uses the Taylor linearization method, built in the SAS SURVEYMEANS and SURVEYFREQ procedures. In the first part of the paper the theoretical background along with some pragmatic implementation approaches will be described. In the second part we focus on the description of the practical usage of the application. The theoretical descriptions will be supplemented by the concrete examples from the ICT survey.

Keywords: sampling error, general application, ICT survey

The Missing Link: From Concepts to Questions in Economic Surveys

Ger Snijkers Statistics Netherlands Diane Willimack, US Census Bureau

Data collected in business surveys typically rely on technical definitions and well-specified measurements. Moreover, the desired data are expected to be available in business records. However, underlying concepts may not be as clear-cut or as measurable as expected. Even a concept as seemingly straightforward as «employment» has different dimensions. E.g., a straight «head count» or number of Full-Time-Equivalents? Should part-time workers be included? What about «temps,» leased employees, or contractors? Should employees on paid or unpaid leave be counted? And so on…

Questions in business surveys are intended to provide valid measurements of underlying economic concepts (e.g., construct validity) that often have many attributes, some measurable, some not, some know and some unknown. In survey practice, the variety of attributes may lead to mismatches with a respondent's interpretation or with available data, resulting in measurement error, as collected data fail to meet the intent of the survey question, the underlying concept, or the needs of data users. Survey designers are often unable to identify this ambiguity in the questions or concepts until cognitive pretesting or even data collection is complete, demonstrating that the missing link from concepts to questions and data is often overlooked in business surveys.

The aim of this paper is to put a spotlight on this missing link, not only for survey questions but also for evaluating the validity of register data. We will discuss research methods that can be used to investigate concepts, ascertain attributes, identify measurements, and specify questions, to achieve construct validity at the design stage. We distinguish between top-down theory-driven and bottom-up data-driven approaches. These methods are, among others: dimension/indicator analysis, feasibility or exploratory studies, early stage scoping, record keeping studies, focus groups, concept mapping, and factor analysis. A number of these methods will be described and illustrated with examples.

Abstracts

Approach to the Ukrainian Current Labour Statistics survey designing

Ganna Tereshchenko, National Academy of Sciences of Ukraine

The Current Labour Statistics (CLS) survey in Ukraine is designed to provide monthly estimates of levels and month-to-month trends of payroll, employment, paid hours and earnings. The data are compiled at detailed industrial levels for Ukraine and regions. Some indicators should be estimated for subregional level. The target population is composed of all enterprises of Ukraine with more than 10 employees, except those primarily involved in private household services, services provided by extra-territorial organizations.

The CLS is organized using combined approach: as census for enterprises with more than 50 employees (big enterprises), as sample survey for enterprises with number of employees from 10 to 49 (small enterprises).

Monthly survey data consists of data from approximately 50,000 enterprises (40,000 big enterprises and 9,000 small enterprises). The sample covers approximately one-tenth of small enterprises population. The sample frame is obtained from actualized data of Business Register. The realized rotation scheme provides replacement of 50% of sample each year.

Paper presents approaches to build an effective CLS survey design and to overcome of problems in sample frame construction and indicator estimation.

Automatic editing of numerical data with R.

Mark van der Loo, Edwin de Jonge and Sander Scholtus, Statistics Netherlands

R is a free1 interactive software environment and statistical computing language which hasevolved to the lingua franca of statistics over the last years. R originated as a reimplementation of the statistical language S (Becker et al. 1988) by Ihaka and Gentleman (1996) and is now developed and maintained by a group of about 20 (mainly) academics who maintain and develop the R core. R has a modular architecture, allowing users to add functionality through so-called packages. If a package meets certain technical and documentation standards, it can be offered to other users by uploading it to the Comprehensive R Archive Network (CRAN).

Since 2010, Statistics Netherlands (SN) has adopted R as a scientific programming tool to build (parts of) statistical production systems. Using R facilitates a fast roll-out of new methodology to the production floor because of its shallow learning curve compared to traditional programming languages and its massive statistical functionality.

Recently, Sander Scholtus of SN developed a number of algorithms which are able to use the information in erroneous records to repair records under linear (in)equality restrictions. Specifically, they can solve simple typing errors, rounding errors and sign errors or value interchanges (Scholtus 2008 and 2009). These algorithms deviate from Fellegi and Holt's famous paradigm2, in that they do not necessarily change as few fields as possible, but they do leave as much data as possible intact. For example by swapping values, flipping signs or swapping digits in a number. Generalisations of Scholtus' algorithms have now been implemented and published as an R-package (Van der Loo, de Jonge and Scholtus, 2011). A second package was developed to facilitate the definition, checking and manipulation of edit rules with R (De Jonge and Van der Loo, 2011). In this paper and talk we will discuss some details of the implementation, the algorithms used, and demonstrate the effect of them on realistic Structural Business Survey data.

1 Under General Public License version 2. You may download, use, alter and redistribute the most of the program under mild conditions. See http://www.gnu.org/licenses/gpl-2.0.html for details.

2 Under the assumption that errors occur randomly, and no information on their cause is available, change as few fields as possible to make a record consistent.

References

Becker, R.A., Chambers, J.M. and Wilks, A.R (1988). The new S language: A programming Environment for Data Analysis and Graphics. Wadsworth&Brooks/Cole

De Jonge, E. and Van der Loo, M. (2011) editrules: Convert readable linear (in)equalities into matrix form. R package version 0.4. http://cran.r-project.org/package=editrules

Ihaka, R. and Gentleman, R (1996). R: A language for data analysis and graphics. J. Comp. Graph. Stat. 5 299-314.

Scholtus, S. (2009). Automatic correction of simple typing errors in numerical data with balance edits. Discussion paper 09046, Statistics Netherlands, The Hague/Heerlen. Accepted for publication in the Journal of Official Statistics.

Abstracts

Scholtus, S. (2008). Algorithms for correcting some obvious inconsistencies and rounding errors in business survey data. Discussion paper 08015, Statistics Netherlands, The Hague/Heerlen. Accepted for publication in the Journal of Official Statistics.

Van der Loo, M. De Jonge, E. and Scholtus, S. (2011). Deducorrect: Deductive correction of simple rounding, typing and sign errors. R package version 0.9-1. http://cran.rproject. org/package=deducorrect

Checking the Usefulness and Initial Quality of Administrative Data

Frank Verschaeren, Statistics Belgium

Work Package 2 (WP2) of the ESSnet on the use of administrative and accounts data for business statistics aims to provide guidelines for NSIs examining the quality of administrative data as input for the statistical production process.

The work package is set up to meet two objectives:

- To help Member States examining the usefulness of available administrative data for business statistics
- To help Member States for checking initial quality of administrative data before introducing it into statistical data base

A checklist is being developed that helps NSIs consider all the relevant issues in evaluating the usefulness of administrative data before a new source is acquired or if an existing one is to be changed.

Once the data become available, efforts are needed to ensure the quality of the data, because very often the administrative data are not fully compliant with statistical needs. This WP looks closer into methods of detecting and resolving quality issues at the initial stage of receiving the data.

The paper will give an overview of the work planned until June 2013 and present the first results of this WP in elaborating a checklist and on methods for investigating initial data quality issues.

Industrialization of editing

Li-Chun Zhang, Statistics Norway

Many National Statistical Institutes (NSIs) are currently facing similar challenges on three fronts: (a) budget constraints, (b) increasing user demands, both in terms of the scope of the statistical outputs and a greater emphasis on the various quality dimensions, and (c) rapid changes in the information-related technologies and data/metadata infrastructure of the global society to-day.

Statistical data editing (SDE) is an area of focus because experiences show that typically SDE-related processes command 30%–40% of the total budget in business statistics. Nevertheless, such processes are indispensable for ensuring the quality of the statistical products.

There is thus a need for greater production efficiency, and many NSIs have started the transition from numerous stovepipe-like chains of production solutions to corporate-wise standardized systems. Such modernization programs, however, have raised two pressing issues. Firstly, there is the conflict between the resources for continuous production and those for modernization. Secondly, despite the emergence of the likes of GSBPM and SDMX, there is a lack of internationally accepted common references and standards when it comes to the design and implementation of SDE-processes.

Stronger international collaboration was recognized as an important measure in dealing with both issues, at a meeting in June, 2010 in Paris by the initiation and leadership of the Australian Bureau of Statistics, with the additional participants from Canada, New Zealand, Norway, Sweden and UK. An overall vision of industrialization of statistical production was forged. Editing was chosen as one of the areas for advancing, and Statistics Norway was to lead the effort.

In this presentation we would like to inform on the spirit of industrialization, as well as to report on the progress in the field of SDE, with particular emphasis on the Generic Statistical Data Editing Processes (GSDEP), the Common Statistical Data Reference (CSDR), the overall design of generic SDE functionalities, the repository of standard statistical methods, IT tools and platforms.

Abstracts

Restricted neighbor imputation with adjustments

Li-Chun Zhang, Statistics Norway

We start with the task of reconciling conflicting information in statistical micro data that may arise due to partial imputation. The missing values are imputed either by the corresponding values of a suitable complete record (i.e. donor) or by statistical estimation. The imputed record then consists of two parts with data from different sources. One part contains the observed values from the original record (i.e. receiver) and the other the imputed values. Edit rules and statistical relationships that involve the variables from both parts will often be violated.

For instance in business statistics we may have that Turnover must be equal to the sum of Profit and Cost, where Cost is again the sum of costs for material, personnel, etc. and all variables except Profit must be non-negative. If some of the variables are missing, the imputed values taken from a suitable donor may not satisfy the various restrictions, together with the observed values of the receiver. Moreover, suppose a variable such as the number of employees is used for donor selection, but it does not match exactly between the chosen donor and the receiver, then statistical adjustments may be desirable on this account, regardless of whether the additional edit rules are satisfied or not.

A general methodology for adjusting the initial (donor) values has recently been developed. The edit rules are specified as linear equality-/inequality-constraints on the variables. The adjustments are then obtained by minimizing a suitable distance metric subjected to these constraints.

We consider two extensions. Firstly, both the construction of the linear constraints and the choice of the distance metric need to be modified, in situations which involve data that are categorical or semi-continuous. We propose to decompose the overall distance metric into a part for additive adjustments (which can be used to handle categorical variables) and another one for multiplicative adjustments (which often seem more natural for continuous variables). Secondly, the partial imputation approach can be applied for the purpose of constructing statistical registers, which are subjected to benchmark constraints at various aggregated levels in order to improve the efficiency. For a unified approach we generalize the setting to include unit imputation as a special case. We shall illustrate the approach of Restricted Neighbor Imputation with Adjustments (ReNIA) using data from the Norwegian agriculture census 2010.