The Unit Problem and Other Current Topics in Business Survey Methodology

The Unit Problem and Other Current Topics in Business Survey Methodology

Edited Proceedings of the European Establishment Statistics Workshop 2017

Edited by

Boris Lorenc,
Paul A. Smith,
Mojca Bavdaž,
Gustav Haraldsen,
Desislava Nedyalkova,
Li-Chun Zhang
and Thomas Zimmermann

Cambridge
Scholars
Publishing

The Unit Problem and Other Current Topics in Business Survey Methodology

Edited by Boris Lorenc, Paul A. Smith, Mojca Bavdaž, Gustav Haraldsen, Desislava Nedyalkova, Li-Chun Zhang and Thomas Zimmermann

This book first published 2018

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-1661-X
ISBN (13): 978-1-5275-1661-8

CONTENTS

Preface vii

The European Network for Better Establishment Statistics..... ix

1. Introduction 1
   Paul A. Smith

2. The unit problem: an overview..... 7
   Paul A. Smith, Boris Lorenc and Arnout van Delden

3. The unit problem from a statistical business register perspective..... 19
   Roland Sturm

4. How to improve the quality of the statistics by combining different statistical units?..... 31
   Olivier Haag

5. Issues when integrating data sets with different unit types..... 47
   Arnout van Delden

6. Producing business indicators using multiple territorial domains..... 65
   Daniela Ichim

7. Improving the efficiency of enterprise profiling..... 79
   Johan Lammers

8. The impact of profiling on sampling: how to optimise sample design when statistical units differ from data collection units..... 91
   Emmanuel Gros and Ronan Le Gleut

9. Coping with new requirements for the sampling design in the survey of the service sector..... 107
   Thomas Zimmermann, Sven Schmiedel and Kai Lorentz

10. Sampling coordination of business surveys at Statistics Netherlands..... 127
    Marc Smeets and Harm Jan Boonstra

11. Sample coordination and response burden for business surveys: methodology and practice of the procedure implemented at INSEE 139
Emmanuel Gros and Ronan Le Gleut

12. Response processes and response quality in business surveys 155
Gustav Haraldsen

13. Paradata as an aide to questionnaire design 177
Jordan Stewart, Ian Sidney and Emma Timm

14. Studying the impact of embedded validation on response burden, data quality and costs 199
Boris Lorenc, Anders Norberg and Magnus Ohlsson

15. Adaptations of winsorization caused by profiling 213
Arnaud Fizzala

16. Big data price index 229
Li-Chun Zhang

17. Analysis of scanner data for the consumer price index at Statistics Canada 237
Catherine Deshaies-Moreault, Brett Harper and Wesley Yung

18. Surveys on prices at the Statistical Office of the Republic of Slovenia 253
Mojca Noč Razinger

19. An overview of data visualisation 267
José Vila, José L. Cervera-Ferri, Jorge Camões, Irena Bolko and Mojca Bavdaž

Index 285

PREFACE

During preparations for the 2017 European Establishment Statistics Workshop (EESW17), an opportunity was offered to the EESW17 Scientific Committee to produce a Proceedings volume. As books dedicated to methodology for producing business statistics are few and tend to appear rarely, the committee decided to explore this possibility further by consulting the potential authors. We received an overwhelmingly positive response, which settled our course through an intensive period of work in late 2017 and the first half of 2018. The result is this edited Proceedings of selected, considerably revised papers from EESW17, which we – the chapters' authors and volume's co-editors – are now proud to present to a wider audience of business statistics methodologists and practitioners.

The first chapter of the volume, Introduction, is a conspectus of the book's content: an overview of the current topics in business statistics methodology treated in the chapters. We trust these topics will be of relevance to other colleagues too.

By placing the chapters side by side, some trends emerge. Among these, the unit problem is a major new one. In a belief that it is a technical problem only, solvable by proper regulation, the problem – both its formulation and development of methodologies for addressing its components – have been put aside for decades. However, the five chapters that address the unit problem here explicitly signal that it is time to acknowledge the unit problem as one of the constituent error sources in the Total Survey Error framework and start addressing it properly. Several further papers report changes in methods to deal with changes in units models.

Other trends include the greater use of alternative data sources, seen particularly in the papers on price indices, the development of coordinated sampling practices, and (in one summary paper) the use of data visualisation to assist with the dissemination of statistical outputs.

In their work towards publishing the volume, the co-editors – who were also members of the Scientific Committee of EESW17 – have had the great pleasure to work with a group of talented, able and patient authors who have worked hard on improving the chapters' content through repeated revisions. We are grateful to all the authors for their endurance and accomplishment.

We look forward to meeting our readers at the next EESWs and hope that the chapters in this volume provide a timely input for further developments of business statistics methodology and its practices.

Boris Lorenc
Paul Smith
Mojca Bavdaž
Gustav Haraldsen
Desislava Nedyalkova
Li-Chun Zhang
Thomas Zimmermann

June 2018

THE EUROPEAN NETWORK FOR BETTER ESTABLISHMENT STATISTICS

The European Network for Better Establishment Statistics (ENBES) is the network which organises European Establishment Statistics Workshops. ENBES was launched in 2009 at a meeting during the New Techniques and Technologies for Statistics Conference in Brussels. ENBES is dedicated to improving cooperation and sharing knowledge on theory, methodology and practices within European establishment statistics. "Establishment" is used with a similar definition to that in the International Conference on Establishment Surveys (ICES) conference series to refer to businesses, farms, hospitals, schools and other similar institutions, for which there was no pre-existing generic term in English (Cox and Chinnappa, 1995).

ENBES encourages cooperation on and development of the methods and practices for enterprise statistics, primarily through a biennial series of multi-day workshops: 2009 in Stockholm, 2011 in Neuchâtel, 2013 in Nuremberg, 2015 in Poznań, and 2017 in Southampton. It also holds occasional one-day workshops, which are usually devoted to a specific theme.

Since 2013 ENBES has been affiliated to the Royal Statistical Society in the UK, whose continuing organisational support has allowed it to operate without the need for a formal organisational registration.

ENBES's biennial workshops provide an opportunity for discussion of the methodologies and practices of business statistics, covering a variety of different topics. They are designed to allow more space for development of ideas and discussions than a traditional conference, to encourage progress in business statistics through the synergy of the participants. The chapters in this volume derive from the ENBES workshop which took place from 30 August to 1 September 2017 in Southampton, UK, and have allowed the authors an opportunity to further develop the ideas which were presented there.

Reference

Cox, B.G. and Chinnappa, B.N. (1995). Unique features of business surveys. In Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. and Kott, P.S. (eds.), *Business survey methods*, pp1-17. New York: Wiley.

CHAPTER 1

INTRODUCTION

PAUL A. SMITH[1]

Abstract

This chapter introduces the topics of the different chapters and sets them in the context of current views of business survey methodology and practice. The chapter also highlights how topics are related to each other and how they demonstrate the features which are peculiar to business surveys.

The 2017 instance of the European Establishment Statistics Workshop (EESW17) consisted of 35 contributed papers and three posters. In comparison to a large conference these numbers are modest, however the contributions covered a broad range of themes within business statistics. The classical themes stretched from the design and sampling for business – and more broadly, establishment – surveys, through data collection, and data editing to estimation. Additional themes covered data provider-oriented and user-oriented matter like response burden management and communication with data providers and with the users of statistics. Further themes involved aspects of modernisation of systems for business statistics production, the impact of new data sources on production of indices (such as the consumer price index), and some considerations regarding nonresponse in business surveys. And, most relevantly, as is reflected in the title of this volume, the workshop covered the emerging theme of the unit problem in business and establishment statistics.

Unlike a large conference, the EESW series is tailored to enable considerable feedback from colleagues after presenting a paper, and thus enables greater exchange among peers. The papers included here have

[1] S3RI, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Email: p.a.smith@soton.ac.uk.

benefitted from this feedback and discussion. In the remainder of this chapter, we introduce the topics of the different chapters and sets them in the context of current views of business survey methodology and practice. Taken together, the chapters reflect the current advances in establishment statistics methodology as pursued by European business survey methodologists.

Chapter 2 gives an overview of the unit problem: the issues that emerge when the type of units used to produce statistics differs from the type of unit that the statistical concept is intended to represent. This has several aspects, including when the unit type to calculate the statistics is different from the unit type on which the data are collected, when unit types are in a hierarchical relation, and when different data sources need to be merged to improve data content and to produce more detailed statistics. Chapters 3 to 7 all describe approaches to deal with aspects of the unit problem, particularly the delineation of businesses. This activity is fundamental to the construction of business statistics, where we need to know what units we are trying to construct statistics about. The availability of a business register is a key feature that sets business surveys apart from social surveys (where generally a corresponding register doesn't exist, except in a few countries with a longstanding tradition of population registers). Such a register is based on administrative data usually derived from tax and/or employment registration processes. The business register lists all the businesses that are known to be present in a country, and usually contains a limited amount of information on their characteristics, particularly their size. Such a list supports detailed sampling designs, which are important since it is necessary to include all the largest businesses in a survey in order to obtain accurate results.

Many, particularly larger, businesses have complex structures, so the delineation of businesses is needed to support this detailed sampling by distinguishing separate businesses. This process is dependent on a *units model* which describes how the different units which make a business are treated in the register. This could lead to different statistical outputs from different units models, and this is the essence of one element of the Unit Problem, discussed in more detail in Chapter 2. Additional aspects of the unit problem include understanding the intended unit structure at data collection, and communicating statistics to users, where the conceptual interpretation may be different to the interpretation 'needed' by the user).

The European Union had attempted to update the definition of enterprise within the units regulation, but there was no agreement to change and therefore the original definition is now being implemented. In Chapter 3, Sturm discusses this implementation in Germany and how this relates to

the structures and profiling activities within the German business register. In Chapter 4 Haag discusses a similar revision of the French business register as part of a change to a system with statistics based on the enterprise, but where the legal unit is retained as the observation unit. (This leads later in Chapters 8 and 15 to further developments of the sampling design and outlier processing, changes which are needed to support this approach.)

Chapter 5 also discusses units, but goes further to consider the whole system of statistical data integration when there are inputs derived from different units. This is a particularly important issue when there are multiple data sources with different underlying units models, as it underpins the ability to put all these sources together to develop new outputs and realise efficiencies in statistical production. Chapter 6 also deals with the challenge of making estimates using different unit definitions, more specifically how to make regional estimates according to several different definitions of "region", derived from different regional classifications of unit structures. Chapter 7 also addresses profiling, but examines the efficiency of the profiling activities, and how they can achieve the most value for the least cost.

Following the order of processing of a business survey, we then consider the process of sample design, which continues to be an important one for business surveys. Two chapters deal with sample design challenges: chapter 8 continues the story of the implementation of the business units regulation in France by designing a cluster-based business survey where enterprises (clusters of legal units) are selected at the first stage, and then all the legal units within selected enterprises are surveyed as the observation units. Cluster designs in business surveys are unusual, although there are other similar cases of 1-stage designs (two-stage cluster designs are much rarer), and this is therefore an interesting example. In Chapter 9 Zimmermann *et al*. investigate changes to a sample design brought about by a court decision on equal treatment which is at odds with efficient sample design. This mismatch between design considerations and the interpretation of the legal framework is a lesson in itself of the unintended consequences that can arise in setting up a national statistical system. The immediate need is for a design with minimum (and justified) use of take-all strata, but with as small an effect on accuracy as possible; a change to the estimator helps to retrieve some of the lost accuracy. Longer term it would be sensible to seek an appropriate update to the legal framework.

Sample coordination is a generic term for ensuring that surveys (or different occasions of the same survey) either have some units in common (positive coordination) or that the same units are not included (negative

coordination). The former gives more accurate estimates of changes, and the latter a fairer distribution of the burden of completing questionnaires, and there is some trade-off between these extremes in any system. There is often some negative coordination in social surveys, but there the large size of the population means that its impact is usually negligible. Business surveys, however, often cover the same population and need to include high sampling fractions. And they may have rotating designs without fixed waves. Therefore sample coordination has been an important topic in business surveys, often implemented through a permanent random number based system (Ohlsson 1995). Here Chapters 10 and 11 describe the methods and implementations of coordinated sampling in the Netherlands and France respectively. These are interesting case studies, since coordination has many varieties and implementations (see also Ernst *et al*. 2000, Lindblom 2003, Nedyalkova *et al*. 2009) and there is no recent review of the aims and approaches in this topic area.

A series of chapters deals with the process of obtaining information from businesses. Chapter 12 categorises respondents to a business survey questionnaire by their information retrieval process, and uses this to investigate the effect of business complexity and size on the response process, specifically the quality of the responses and the burden of responding. Chapter 13 discusses the challenges of converting a paper questionnaire to a web questionnaire, and in particular deals with ways in which the paradata collected in the process of administering the web version can feed back to improvements in the questionnaire design. The authors use a classification of the ways in which questionnaires are completed to help in this analysis. Chapter 14 examines experimentally the impact of increasing the number of embedded edit checks in an on-line questionnaire (which already contains some such edits). It shows minimal impact on data quality and efficiency for a small increase in embedded edits, but further testing over a wider range of numbers of embedded edits is needed. The main message from this work is the need for cognitive survey methodologists to focus much more on aspects of embedded editing.

Chapter 15 is related to the change to the enterprise as the basic unit for business statistics in France, and examines how to implement winsorisation in this new sampling context.

There are also challenging data collection problems in surveys of prices as the basis of price indices, and complex clustered sample designs are sometimes needed to make such collections practical. Many National Statistical Institutes are investigating the possibility of using alternative data sources to supplement or replace these, and three chapters deal with this topic. Chapter 16 provides a strategic overview of the challenges of

using alternative data sources for prices. Chapter 17 describes an exploratory study using a variety of methods of index calculation with scanner data from a restricted range of products in Canada. The extra possibilities from the availability of weighting information at lower levels and higher frequencies are interesting, but do not always have the desired effect on the indices, and may just add noise. Chapter 18 provides an overview of price collections in Slovenia and how they are moving to more cost efficient procedures, particularly how agreements with retailers for the provision of scanner data are leading to these being introduced to the main consumer price index calculation.

Finally Chapter 19 summarises several presentations from the ENBES workshop that dealt with the topic of data visualisation. This has clear applications in business statistics, but is a general and much wider subject. The way in which users interact with visualisations, and how they can be used to convey important information and, particularly, stories in the data, is the subject of some empirical investigation. These experiments are also related to existing research on user interactions with graphics, and some commentary on the opportunities and challenges of modern devices.

In summary, this volume presents a range of recent developments in business statistics, many related to aspects of the Units Problem and providing additional information to develop the assessment of quality in relation to the choice of units. A number of current research areas are represented, and some future avenues are suggested. We hope that these contributions will also act as a spur to further research on the methodology for establishment statistics and its application to solve real problems and challenges. If we inspire some activity then ENBES will be continuing to achieve its objectives for promoting knowledge sharing and cooperation in this important area. So if you find these topics interesting we encourage you to participate in ENBES and its events and activities at www.enbes.org.

References

Ernst, L.R., Valliant, R. and Casady, R.J. (2000). Permanent and collocated random number sampling and the coverage of births and deaths. *Journal of Official Statistics* 16, 211-228.

Lindblom, A. (2003). *SAMU – The System for Coordination of Frame Populations and Samples from the Business Register at Statistics Sweden*. Background Facts on Economic Statistics 2003:3, Statistics Sweden. Available at: http://www.scb.se/statistik/OV/AA9999/2003M00/X100ST0303.pdf (accessed 6 April 2018).

Nedyalkova, D., Qualité, L. and Tillé, Y. (2009). General framework for the rotation of units in repeated survey sampling. *Statistica Neerlandica* 6, 269-293.

Ohlsson, E. (1995). Coordination of samples using permanent random numbers. In Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. and Kott, P.S. (eds.), *Business Survey Methods*, pp. 153-170. New York: Wiley.

CHAPTER 2

THE UNIT PROBLEM: AN OVERVIEW

PAUL A. SMITH[1], BORIS LORENC[2] AND ARNOUT VAN DELDEN[3]

Abstract

This chapter introduces unit error and the unit problem and gives some historical background. It then sets the unit error and unit problem within the contexts of two familiar survey methodology frameworks: the European Statistical System (ESS) quality dimensions and the Generic Statistical Business Process Model (GSBPM). To exemplify different aspects of the unit problem, connections to the chapters in this collection that treat the unit problem are made.

1. Introduction

This chapter introduces unit error and the unit problem and gives some historical background, as well as referring to some of the latest developments in measuring and understanding its effects on survey quality. The unit problem, work on which has been encouraged by ENBES over several years, is a statistical quality issue more pronounced in business statistics due to businesses' complex organisational structures. After defining the unit error in Section 2 and giving some historical background in Section 3, the chapter proceeds to review the unit error and unit problem

[1] S3RI, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Email: p.a.smith@soton.ac.uk
[2] Bright Lynx LLC, Vaarika 1-1, 10614 Tallinn, Estonia. Email: boris.lorenc@blresearch.ee
[3] Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague, The Netherlands. Email: a.vandelden@cbs.nl

within the contexts of two by now familiar survey methodology frameworks: the European Statistical System (ESS) quality dimensions in Section 4 and the Generic Statistical Business Process Model (GSBPM) in Section 5. To exemplify different aspects of the unit problem, we make connections to the chapters in this collection that treat the unit problem.

2. The unit problem

The *unit problem* is based on the *unit error* (Zhang, 2011; van Delden *et al*., 2018), which refers to errors in a statistical output that are caused by deviations from an ideal case in identification, characterisation and delineation of the units and in establishing relationships between the units relevant for producing a desired statistical output. With that starting point, the unit problem refers to the challenges and obstacles to understanding of the unit error and to efforts to deal with it. The unit problem therefore indicates a paucity of investigation of the unit errors. A number of chapters in this volume, the statement paper "On the Unit Problem in Business Statistics" prepared for EESW17 (Lorenc *et al*., 2017), and the van Delden *et al*. (2018) letter to the editor of the Journal of Official Statistics, are contributions towards initiating a change to this paucity by focusing the methodological attention of survey statisticians on the unit error.

Awareness of unit errors and the unit problem is not new. In the opening chapter of a compilation containing a selection of edited papers from the first major international conference on business surveys held in 1993, Cox and Chinnappa (1995) reviewed a number of issues related to units that are treated in subsequent chapters of that volume. The issues include: that unit types in business surveys are not natural and often defined for the purposes of the statistics produced; that the hierarchy of units in business surveys can be complex, including criteria such as location and legal and administrative structures; that actual business structures are difficult to relate to units for sampling and reporting, indicating possible issues with availability of data and the correctness of the collected data; and that business populations are extraordinarily dynamic (Cox and Chinnappa, 1995). As pointed out by Pietsch (1995) in the same volume, even in the ideal case of no conceptual and operational mismatch in unit type definition and application, the fragmentary approach taken by surveys in gauging businesses' operations (different areas of business operations targeted, different unit types, different reference periods, different data providers, etc), contributes to the variability in the statistical output produced, due in part to issues now referred to as the unit error.

Meeting annually since 1986, what is now the Wiesbaden Group on Business Registers had already become well established as the body treating issues relating to units by the early 1990s. Struijs and Willeboordse (1992) reported on a survey of the state of affairs regarding units at that time. From the distant perspective of 2018, it can perhaps be said that the low visibility of survey methodology and sampling theory in the management of units in business registers, or more generally in treating the impact of what we now call the unit error on produced statistics, has characterised the whole period up until the 2010s.

3. Broader approaches

Only more recently have more encompassing approaches emerged. Based on the survey lifecycle model (Groves *et al*., 2004, Fig. 2.5) reflecting the context of directly collected data from individuals and households, Bakker (2013) and Zhang (2012) have developed two-phase life cycle models of integrated statistical microdata for the production of statistics using administrative and survey data, thus spanning several "lifecycles" of data. Around the same time Zhang (2011) introduced the unit error in the context of household surveys. Business registers are not created for the purpose of conducting a single survey, and can therefore be seen as an earlier stage within a specific survey's lifecycle. Creation of units when they enter the business register can lead to unit errors, so the two-phase perspective applies even here. Van Delden (2018, Chapter 5 in this volume) highlights this approach.

Lorenc *et al*. (2017) and van Delden *et al*. (2018) express the need to address the unit error and unit problem in a methodological way, for which they argue that the Total Survey Error (TSE) approach is the most appropriate. The unit error then gets acknowledged as one among the other – more well-known – types of errors associated with the production of statistics, such as sampling error, nonresponse error, measurement error, and so on. Within that framework, part of the challenge is to obtain estimates of the effect of the units model on the statistical outputs. Ichim (2018, Chapter 6 in this volume) implicitly provides such estimates, as the difference between using different unit types in regional estimation, and some micro-level comparison of the effect of switching units in employment surveys in the UK is presented in Smith *et al*. (2003). This is helpful, but how to integrate these estimates into a survey error framework is an open problem. Should every statistic in principle include an estimate of the "unit model error" which characterises how much the outputs could vary under a range of different unit models? That would be a complex set

of calculations, which could only be done over a small number of case studies – both a small number of statistics and a small number of unit models – because of the effort required to define units and recast collected data to those units. Even to use a single ideal measure would be resource-intensive, and it is far from clear in many situations what the ideal unit structure is.

However, when done, such studies can indicate the contribution of the unit error to the uncertainty of the produced statistics. For instance, the study by INSEE referred to in Haag (2018, Chapter 4 in this volume) shows that when legal units are chosen to present the large enterprises' share of exports from France that share is 22%. But, when the enterprise is the unit on which the statistics are calculated, that share is 52%. It also shows that important breakdowns of variables like turnover and exports can vary substantially – there are differences of up to 5 percentage points in the breakdown of turnover between NACE sectors and 30 percentage points in the breakdown of exports between small and large units (which, although a big difference, is perhaps less surprising because of the way large businesses are typically built up of multiple smaller units). This uncertainty due to unit choice is much larger than the usual sampling error in business surveys and thus deserves much more research than it has received thus far. Therefore, in spite of being an additional effort, studies to estimate the impact of the choice of unit type – a component of the unit error – are much needed.

The European approach to the unit problem has been to harmonise (initially with limited take-up, but now more stringently) on one particular units model (Eurostat, 2014) so that statistics are comparable. But this does not give any indication of how important the model is in defining the outputs. The model in itself allows inconsistencies to arise, as discussed among others by Sturm (2015) and van Delden *et al*. (2018), contributing to the variability in statistical outputs that use this model. Further, as Sturm (2018, Chapter 3 in this volume) discusses, the stringent harmonisation goes down to a certain level, but still leaves open a wide variety of choices below that level in implementing the particular units model, leading to further deviations and variability.

Measurement errors can contribute to the unit error in two ways. At one level, the unit error could be considered to be a measurement problem, since it is necessary to define the object of interest before attempting to collect data from it, and any mismatch between the objective and what is actually collected will be a type of measurement error. Nonetheless, the unit problem itself is wider, because it also encompasses how the object of interest is defined. The second contribution of measurement error is in the

variables from which the units are profiled (profiling is how the units model is implemented). Any errors in the variables on which profiling is based may cause errors in the unit structure, even if the model is conceptually perfect.

Studies of data collection in business statistics have in recent years tried to understand how the data collection process within companies works, and whether one can make adjustments to this process to possibly reduce the impact of unit errors or at least to trace their effects. Haraldsen (2018, Chapter 12 in this volume) is an example of such a contribution.

Haraldsen (2013) gives an overview of practical problems affecting the quality of data while conducting business surveys, where many of the examples arise as a result of unit error.

In the remainder of this overview, we consider the unit problem in relation to the European Statistical System's standard quality dimensions (Eurostat 2015) (and these thoughts can be straightforwardly adapted to other quality classifications), and also to the Generic Statistical Business Process Model (GSBPM) (UNECE 2013).

4. ESS quality dimensions perspective

The ESS uses five dimensions of output quality, and we consider each in turn:

Relevance, assessment of user needs and perceptions. This is a challenging component because enterprise statistics have a wide range of uses and users. To begin with, the National Accounts (NA) by tradition have been a very important user, and their whole-economy coverage has promoted the use of a consistent units model.

However, it has been noted that in practice different countries use kind-of-activity units, enterprises or even enterprise groups as the basic statistical unit underlying their supply and use tables. In the institutional sector accounts, some countries use legal units as the best approximation of institutional units (the unit type used by NA), while other countries apply enterprises or enterprise groups as being equivalent to institutional units. This is said to have a clear impact on the national accounts aggregates (OECD, 2016).

The growing importance of the input-output framework to underpin NA leads to a wish for more detailed outputs by product. This level of disaggregation makes the units model less critical, since the more detailed components, once estimated, can be added up in any way as necessary. However, a prerequisite for such detailed outputs is that the required data are available. A rich source of such data can be found in the actual

transactions in business administrations. The problem is however, that not all businesses classify their transactions in the same way. The use of a central, reference code set to which different business classifying systems are mapped, might be a solution to this problem (Buiten *et al*., 2016). There is also the issue that a readiness to make these transactions available to external entities (NSIs included) is not always to be counted on, as well as that providing data is a burden on businesses (Haraldsen, Chapter 8).

Other users require different breakdowns, such as regional or by country of ownership, or by any of myriad other variables. Where such variables are clearly defined by the unit model *and* the relevant data are available from a survey or other source, the outputs should be directly estimable and relevant, but where they are not we may expect a mismatch between the actual estimates and some conceptual true value, diminishing the relevance of the statistical outputs.

There is also a component of relevance related to user perceptions and the choice of the model itself. The model must be close enough to perceptions of reality that its use to describe business structures (and the inevitable approximations) is accepted by most users as a satisfactory representation for statistical purposes. There is an aspect of communication of statistics involved in this, at two stages in the production process: (i) at the data collection stage, whether the intended concepts related to units and their characterising variables have been understood correctly by the data provider, and (ii) similarly at the dissemination stage, whether the declared concepts related to units have been understood correctly by the users (including whether the user understands the uncertainty that may emerge at the data collection stage). This comes down to whether the statistics producer can vouch that the intended concepts regarding units have been understood and applied correctly during data collection, and whether any remaining uncertainty has been communicated clearly to the user.

Accuracy and reliability. There are several components of the application of the units model which contribute to accuracy. One is in the choice of the model itself, which leads to a "units model assumption error". A second relates to the data requirement for implementation of the model (and its timeliness, see below), and therefore whether the profiling of a business using that model is done correctly or with error. There are therefore two components of accuracy, one related to the range of values achievable from different unit models, which is a kind of variance, and one related to the difference between what is measurable and the target concept, a kind of bias and closely related to relevance.

Regarding NA, when the unit type that is chosen for compiling NA differs from the unit type that is chosen for the data that are used to compile

NA, which is currently the case, this not only affects coherence but also the accuracy (variance). The input data that enter NA need modifications to make them suitable for the NA unit type. The NA unit type is more 'homogeneous in economic activity' than its input, to enable the production of a detailed input-output table. So, in compiling the NA, data of the largest (inhomogeneous) companies are adjusted to fit the NA unit type, which likely affects accuracy. Such adjustments are usually done when a company has shifts in their composition, when it splits off part of its legal units for instance.

Timeliness and punctuality. Business populations are dynamic, and the processes of collecting information and using it to profile a business in line with a model are at best contemporaneous and at worst occur with a long lag. Therefore, the units model could be correctly implemented with current evidence, but the evidence itself may be out of date. There is some published information on lags in updating registers (Hedlin *et al*. 2001, Smith and Perry 2001) but more work in this area and how it relates to applications of units models and profiling would be beneficial.

Accessibility and clarity. The profiling and units models are not in general publicly available, so they do not score highly on this quality dimension. There may be a case for more openness about such structures, in part as a way to obtain feedback from users and researchers on how well the models work in different situations.

Coherence and comparability. A consistent units model across the whole economy promotes coherence within the national accounts, and similarly consistent models in different countries promote comparability of models, methods and outputs across the international statistical system. How to measure this comparability is however challenging – one possibility is to relate it to the accuracy and relevance dimensions by considering how great the mismatch in statistics would be under alternative models. However, as discussed earlier, this is likely to be an expensive type of case study because of the need to recode business structures onto an alternative model. Some elements can be treated in the same way as classification updates, since a new units model effectively applies a different classification of units; see Smith and James (2017) for a review of reclassification approaches in official statistics.

Lorenc *et al*. (2017) also consider cost and response burden, which are important considerations, although not quality components in the ESS framework. There is a cost to the statistical system (e.g. of profiling), which is directly affected by the operational features (and Lammers (2018, Chapter 5 in this volume) considers the efficiency of this process in the Netherlands), but can ultimately be attributed to the conceptualisation. The

response burden is also strongly affected by the conceptualisation in that it drives whether the data that exist in businesses' accounting systems or administrative sources can be used.

Finally, although there are many complications in the unit model, van de Ven (2018) makes an appeal that the approach to units should be as simple and practical as possible – for example, so that data linkage can be undertaken straightforwardly.

5. GSBPM perspective

The GSBPM (UNECE 2013) codifies the steps in the design and processing of official statistics, and it is interesting to consider in which stages the units model has the greatest impact. The level 1 processes in the GSBPM are shown in Fig. 2-1 (there is a more detailed subprocess level, not shown).

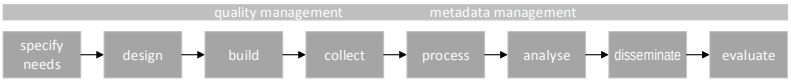

Figure 2-1: Level 1 (high level) processes in the GSBPM (redrawn from UNECE 2013).

The unit model has most effect in the early stages of the GSBPM, since it provides a foundation for establishment statistics. ***Specify needs*** and ***Design*** are closely related to the relevance dimension of quality, where the chosen unit model needs to be capable of supporting a range of different user requirements, in as consistent and comparable a way as possible. These elements therefore imply appropriate decisions over the choice of unit model. ***Build*** means that the methods and software to support the implementation of the units model need to be constructed, and these are populated with information derived from the ***Collect*** processes. The data are then ***Process***ed to implement the units model, which may include some additional stages such as imputing for missing data. The subsequent stages rely on these underpinnings, but do not directly relate to the unit model. The two overarching processes, for ***Quality*** and ***Metadata Management*** are however important – the assessment of the quality impacts of the choice of the unit model is the essence of the unit problem (van Delden *et al*., 2018), and it is vital to document the sources of evidence and decisions which are taken to apply the unit model structures for a particular business, so that these can be revisited when necessary.

There is a feedback loop in the GSBPM where the quality of the outputs is ***Evaluate***d, and used to improve the processing steps in the next cycle. To some extent it is this feedback loop which has not been clearly implemented in business statistics, leading to the situation where the unit problem is clearly an area of weakness. But with the extra attention being brought to bear on this issue, we can hope that progress on improving the units model and the quality measures that describe its effects can be made.

References

Bakker, B.F.M. (2013). Micro-integration: State of the art. In *Report on WP1 of ESSnet on Data Integration*. Available at: https://www.istat.it/it/files/2013/12/FinalReport_WP1.pdf (accessed 2017-07-04).

Buiten, G., Boom, R. van den, Roos, M. and Snijkers, G. (2016). Issues in automated financial data collection in the Netherlands. *Proceedings of the Fifth International Conference of Establishment Surveys*, June 20-23, 2016, Geneva, Switzerland. Virginia: American Statistical Association.

Cox, B.G. and Chinnappa, B.N. (1995). Unique features of business surveys. In Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. and Kott, P.S. (eds.), *Business survey methods*, pp1-17. New York: Wiley.

van Delden, A. (2018). Issues when integrating data sets with different unit types. In Lorenc, B., Smith, P.A., Bavdaž, M., Haraldsen, G., Nedyalkova, D., Zhang, L.-C. and Zimmermann, T. (2018) (eds.). *The unit problem and other current topics in business survey methodology*, pp 47-63. Newcastle upon Tyne: Cambridge Scholars.

van Delden, A., Lorenc, B., Struijs, P. and Zhang, L.-C. (2018). Letter to the Editor: On statistical unit errors in business statistics. *Journal of Official Statistics* 34, 573-580.

ENBES (2014) ENBES Workshop "The Unit Problem in Business Statistics Methodology" held in Geneva, Switzerland, on November 10th 2014. Available at https://statswiki.unece.org/download/attachments/126353571/ENBES%20Workshop%20Unit%20Problem%20Summary_20150220_logo.pdf?version=1&modificationDate=1477298058445&api=v2 (accessed 2018-04-03).

Eurostat (2014). *The statistical units model* (Version: 15 May 2014), version presented to the Business Statistics Directors Group Meeting 24 June 2014.

Eurostat (2015) *ESS handbook for quality reports, 2014 edition*. Luxembourg: Publications Office of the European Union.

Groves, R.M., Fowler Jr., F.J., Couper, M., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). *Survey methodology*. New York: Wiley.

Haag, O. (2018). How to improve the quality of the statistics by combining different statistical units? In Lorenc, B., Smith, P.A., Bavdaž, M., Haraldsen, G., Nedyalkova, D., Zhang, L.-C. and Zimmermann, T. (2018) (eds.). *The unit problem and other current topics in business survey methodology*, pp 31-46. Newcastle upon Tyne: Cambridge Scholars.

Haraldsen, G. (2013). Quality issues in business surveys. In Snijkers, G., G. Haraldsen, J. Jones, and D.K. Willimack, *Designing and conducting business surveys*, pp 83-125. Hoboken, New Jersey: Wiley.

Haraldsen, G. (2018). Response processes and response quality in business surveys. In Lorenc, B., Smith, P.A., Bavdaž, M., Haraldsen, G., Nedyalkova, D., Zhang, L.-C. and Zimmermann, T. (2018) (eds.). *The unit problem and other current topics in business survey methodology*, pp 155-176. Newcastle upon Tyne: Cambridge Scholars.

Hedlin, D., Pont, M.E. and Fenton, T.S. (2001) Estimating the effects of birth and death lags on a business register. In *ICES-II: Proceedings of the Second International Conference on Establishment Surveys*. Contributed Papers (CD), pp 1099-1104. Virginia: American Statistical Association.

Ichim, D. (2018). Producing business indicators using multiple territorial domains. In Lorenc, B., Smith, P.A., Bavdaž, M., Haraldsen, G., Nedyalkova, D., Zhang, L.-C. and Zimmermann, T. (2018) (eds.). *The unit problem and other current topics in business survey methodology*, pp 65-78. Newcastle upon Tyne: Cambridge Scholars.

Lorenc, B., van Delden, A., Struijs, P. and Zhang, L.-C. (2017). *Statement on the unit problem in business statistics*. Paper written for the European Establishment Statistics Workshop 2017, 30 August – 1 September, 2017, Southampton, UK. Available at: https://statswiki.unece.org/download/attachments/122325493/ENBES%20Unit%20Problem%20Statement.pdf?version=1&modificationDate=1500370501222&api=v2 (accessed 2018-04).

OECD (2016). Reassessment of the role of the statistical unit in the System of National Accounts. Prepared for the Fifteenth session of the Group of Experts on National Accounts, Geneva, 17-20 May 2016. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.20/2016/ECE.CES.GE.20.20_OECD.pdf (accessed 2018-06-13).

Pietsch, L. (1995). Profiling large businesses to derive frame units. In Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. and Kott, P.S. (eds.), *Business survey methods*, pp. 101-114. New York: Wiley.

Smith, P.A. and James, G.G. (2017). Changing industrial classification to SIC (2007) at the UK Office for National Statistics. *Journal of Official Statistics* 33, 223-247.

Smith, P. and Perry, J. (2001). Surveys of business register quality in central European countries. In *ICES-II: Proceedings of the Second International Conference on Establishment Surveys*. Contributed Papers (CD) pp 1105-1110. Virginia: American Statistical Association.

Smith, P., Pont, M. and Jones, T. (2003). Developments in business survey methodology in the Office for National Statistics, 1994-2000 (with discussion). *Journal of the Royal Statistical Society, Series D* 52, 257-295.

Struijs, P. and Willeboordse, A. (1992). Terminology, definitions and use of statistical units. 7th Round Table on Business Registers, Copenhagen 12-16 October 1992. Available at https://circabc.europa.eu/sd/a/f975e6c9-4a3c-44c4-b750-eae48cb4efb6/Terminology%252c%20Definitions%20and%20Use%20of%20Statistical%20Units.pdf (accessed 18 Jun 2018).

Sturm, R. (2015). Revised definitions for statistical units–methodology, application and user needs. The main conceptual issues of the "units discussion" of the years 2009–2014. *Statistika* 95, 55-63.

Sturm, R. (2018). The unit problem from a statistical business register perspective. In Lorenc, B., Smith, P.A., Bavdaž, M., Haraldsen, G., Nedyalkova, D., Zhang, L.-C. and Zimmermann, T. (2018) (eds.). *The unit problem and other current topics in business survey methodology*, pp 19-30. Newcastle upon Tyne: Cambridge Scholars.

UNECE (2013) *Generic Statistical Business Process Model GSBPM (Version 5.0)*. Geneva: UNECE. Available at: http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0 (accessed 16 April 2018).

van de Ven, P. (2018). Economic statistics: how to become lean and mean? *Journal of Official Statistics* 34, 309–321.

Zhang, L.-C. (2011). A unit-error theory for register-based household statistics. *Journal of Official Statistics* 27, 415-432.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica* 66, 41-63.

CHAPTER 3

THE UNIT PROBLEM FROM A STATISTICAL BUSINESS REGISTER PERSPECTIVE

ROLAND STURM[1]

Abstract

The chapter builds on work in progress in Germany on implementing profiling as a component in regularly updating the statistical business register. I am addressing three questions. (1) What are we looking for and what do we get? I highlight borderline issues between enterprise and kind-of-activity unit in practical statistics production. (2) Aiming for harmony or aiming for harmonization? Since many ways to do profiling work are in use in the member states of the European Union, statisticians may provide a wide range of practical outcomes regarding units for everyone to follow, but the definition of enterprise is not harmonised in the same way. (3) Is the enterprise definition settled for good? I present some conceptual and empirical findings on the enterprise concept, focusing on two issues: splitting of legal units and a whole enterprise group classified as one enterprise. These findings suggest that the proposals for a revision of the wording of the enterprise definition which were formulated by the Statistical Units Task Force in 2014 are still relevant.

1. Introduction

Continued attention is given to the unit problem as a major obstacle to producing and communicating good business statistics. Regarding statistical units, a distinction can be postulated between three stages of abstraction from reality (Sturm, 2014):

[1] Destatis, Gustav-Stresemann-Ring 11, 65189 Wiesbaden, Germany. Email: roland.sturm@destatis.de.

- The definitions should capture decisive characteristics of units which are important from conceptual and analytical points of view. As concepts and analyses serve practical purposes these definitions are of course driven by issues of real life, they are not purely academic.
- The operational rules describe in more detail how the definitions should be understood or how they could be handled in reality, that is, in application. Therefore, the operational rules build a bridge between definitions – which should be concise but also as short as possible in wording – and application. When drafting definitions and operational rules, sometimes it has to be worked out what belongs to the pure definition and what is already practical and therefore belongs to the operational rule. Still, the operational rules should be general and not particular for only one specific context or situation.
- The application of the definitions starts by deciding which unit to choose for which statistical purpose. Operational rules often have to be elaborated in further detail regarding a certain context and it has to be worked out how to handle the manifold practical aspects, e.g. how to collect data from respondents about observation units and how to transform this data to produce figures about the statistical units.

One of the crucial components in dealing with statistical units is the statistical business register. Register experts have been involved intensively in Eurostat's efforts to improve the consistent application of the enterprise concept in recent years (Sturm and Redecker, 2016). Since the 2014 ENBES workshop (ENBES 2014), which dealt with the unit issue, major developments have taken place and have been dealt with in the working groups for Structural Business Statistics (SBS), for Short Term Statistics (STS) and for Business Registers and Statistical Units (BR&SU) at Eurostat.

Firstly, in 2015, Eurostat abandoned the attempt to change the enterprise definition of Regulation 696/93 (the "Statistical Units Regulation", European Union 1993), or at least postponed any major discussion about changes of the unit definitions for the coming years (Eurostat, 2015). Secondly, all Member States which do not apply the enterprise definition appropriately have provided Eurostat with action plans to implement the enterprise definition in Structural Business Statistics (SBS). Thirdly, profiling as a method to identify enterprises is being established in many statistical offices – commonly in the register