On the Unit Problem in Business Statistics

On the Unit Problem in Business Statistics

Statement Paper Prepared for Participants' Discussion at EESW17

Introduction/Aim

Statistical methodology has thus far been lagging in providing guidance for consideration and treatment of business units in the production and use of business and economic statistics. The Scientific Committee of EESW17 has formed a working group (footnote: Members: Boris Lorenc (chair), Arnout van Delden, Peter Struijs and Li-Chun Zhang) to prepare a short statement paper on the subject. Its aim is three-fold: (i) to summarise the core aspects of the unit error and the associated unit problem, (ii) to stimulate the discussions to clarify and improve our understanding of the system of statistical units, which is needed for the production of National Account and various relevant national and international business/economic statistics, (iii) to provide the background for an integrated approach to the unit problem in business statistics, including the development of necessary statistical methods for the evaluation and treatment of unit error from a total survey error perspective. Comments, thoughts, and suggestions of EESW17 participants are invited and warmly welcome.

Administrative vs. Statistical Business Units

There is a distinction between administrative and statistical business units. Administrative units are created for administrative purposes outside the statistical system. For instance, legal units (LeU) are a type of administrative units that one expects to find in every country, even though their definition varies over countries. Another example is tax units that exist in some countries, which are created for taxation and do not coincide with the legal units. Administrative business units are generally maintained by external owners and imported to the statistical system more or less frequently. They are also the starting points for creation and maintenance of statistical business units.

Statistical business units are created within the statistical system for the purpose of producing statistics. Typically, intrinsic relationships between statistical units are inferred and articulated in terms of a classification, or a model of units. For example, the current Eurostat model of statistical units consists of the unit types enterprise group (EG), enterprise (ENT), local unit (LoU), kind of activity unit (KAU) and local kind of activity unit (LKAU) (fig.). Or, the Institutional Unit, which is closely related to the ENT, may be subdivided into units of homogeneous production (UHP) for the purpose of National Accounts, where the UHPs are not the same unit type as the KAUs.

Needs and Challenges of Statistical Units

Creation of statistical units is necessary because administrative units do not exist for statistical purposes nor are they seen as able to fully meet the needs of the users of statistics. For instance, many economic theories are based on the assumption that the business units possess a level of autonomy in decision making. In contrast, the administrative unit LeU entails legal (or fiscal) accountability, the structure of which does not necessarily coincide with that of business decision making.

Ideally speaking the system of statistical units should mirror business data availability as well as possible so as to improve data collection. For instance, the EG and ENT are created to capture better than the LeU the business management, operation and accounting structure. But in practice this is not always achieved. Especially, some of the lower-level statistical units types may present challenges to businesses' own understanding of reality. For example, the business may find it difficult to extract the required data (e.g. sales, purchases, profit) for the LKAU, in which case the delivered data may refer to some proxy unit (e.g. LeU) instead.

In survey methodology, there is a distinction between the study unit (of the target population), the sampling unit, the reporting unit (i.e. the entity within a business that is actually able to deliver the required data), etc. The system of statistical units is created having the user needs in mind. The two approaches do not fully align with each other, in the sense that there does not always exist a many-to-one mapping from one set (of units) to the other. Moreover, the administrative units are relevant with respect to survey compliance, or the reduction of response burden and survey cost by the increasing uptake of administrative data.

Unit Error and Unit Problem

Choice of the business unit type is an important decision in the design and operation management for a statistical product. By the term “unit error” we refer to the errors in statistical outputs, which are caused by the identification, characterisation and delineation of the relevant statistical units and the relationships between them. In addition, by the term “unit problem” we refer to the challenges and obstacles to our understanding of the unit errors and our efforts to deal with them. The unit problem may be related to the practice of business surveys, the design of relevant statistical processes, as well as the conceptualisation of the system of statistical units.

Generic Situations for Unit Errors

Unit errors can be appreciated in terms of a discrepancy between ‘what one aims to obtain’ and ‘what one obtains’. Below are some generic situations where the discrepancy arises.

1. There is **observation error** in the data that is available, such as when a value is missing or misreported. The discrepancy is between the results based on the true data and the erroneous data. For instance, the administrative record shows that an LeU is active in the economic sector “12345”, whereas it is in fact active in the sector “21345”. This can potentially affect the characterisation (e.g. NACE) and identification (e.g. inclusion in the frame) of the statistical units related to this LeU. (Some other examples of observation errors in data are given in Appendix 1.)

Van Delden (2017) provides a breakdown of observational errors in different stages of data processing for statistical purposes. We would like to mention specifically two types of observation error here.

- Profiling error of large and complex business units (footnote: Profiling is the activity to delineate the statistical units associated with large or complex businesses, including relationships) may result in erroneous statistical units, which tend to have a large impact on the output.
- Consolidation (or apportion) error may be unavoidable when the available data needs to be transformed (or disaggregated), because the required data is missing or simply unavailable. For example, turnover of the VAT units may need to be transformed to that of ENTs, where the two units have many-many relationships. Or, the quarterly data may need to be disaggregated to monthly data. (footnote: Consolidation concerns excluding internal flows from values reported by units that are underlying a targeted composite unit type. Deconsolidation is the opposite situation)

2. **Implementation error** may be the case with respect to the relevant regulation or statistical unit model (e.g. Fig. 1). The discrepancy is between the results from correct and incorrect implementation. For instance, the regulation of Business Register (BR) may be misinterpreted, or it may not cover the extra complications in a given country (e.g. existence of tax units in addition to LeUs), etc.
3. There may exist inconsistencies and shortcomings in the statistical unit model or relevant regulation, the **definition error**. The discrepancy is between the results with and without such inconsistencies. For example, the unit model depicted in Fig. 1 does not include the means to guarantee that one obtains the same ENTs directly from the LeUs or indirectly via the EGs. Another example, the definition actually allows an LoU to have activity in different locations (towns).

4. There is ultimately the discrepancy between the ideally delineated units under a consistent unit model and the units that the users need or expect for their purposes. For instance, Brion et al. (2014) have documented such discrepancies between the actual business demography of SMEs in France which is based on the LeUs, and the user expectation of business demography based on autonomous units like the ENTs. One may refer to this as the **conceptualisation error**. As long as there is a (non-negligible) conceptual mismatch, improving the implementation of existing relevant regulation cannot overcome the unit error in statistical products. More discussions regarding the conceptualisation difficulties can be found in the Village Bakery Example (Appendix 2).

Dealing with unit problem under the Total Survey Error (TSE) framework

We believe that unit error should be included and recognised in the TSE framework, in the same spirit alongside the other types of error, such as sampling error, non-response error, measurement error, etc. In other words, while it is important and helpful to try to reduce the unit errors in individual data, it is necessary to approach the unit problem from a more integrated perspective. As indicated by the above analysis of the various generic situations that can lead to unit error, a single-minded focus on the operational aspects of the statistical process will have little effect at all regarding the conceptualisation and definition errors, and only limited and potentially biasing effects on the observation error.

The effects of the actual unit errors in the collected data, their treatment in data processing and adjustment in statistical estimation need to be understood and articulated under the TSE framework. In terms of data collection and integration, the unit error is rooted in the representation side of TSE framework (Groves et al, 2004; Zhang, 2012). The different situations of discrepancy that can cause the unit error are inter-related, so that it is important to keep such ‘interactions’ in mind when dealing with the unit problem. Moreover, the unit error will also affect measurement errors and relevance errors on the measurement side of the TSE framework, whereas the causes of potential errors on the measurement side can as well affect one’s approach to the unit problem.

Evaluation of Unit Error through User Value Criteria

Regardless of one’s approach to the unit problem, the effects of the unit error that remains in the data need to be evaluated with respect to the User Value Criteria below, including the so-called quality dimensions.

a) Relevance (e.g. output that make sense to users) is ultimately rooted in the conceptualisation of the system of statistical units.

b) Coherence (e.g. between annual and short term statistics, national totals based on different units, etc.) seems mostly related to the various types of observation error and to the conceptualisation of the system of statistical units.

c) Accuracy (e.g. avoiding bias of various causes) is amply discussed under the TSE framework.

d) Timeliness does not call for a treatment that is specific to the unit error.

e) Comparability (e.g. if the unit system or classification changes) can be a challenge with respect to all types of unit error over time.

f) Accessibility (e.g. with regard to unduly complex system) seems above all related to the conceptualisation of the system of units.

g) Cost to the statistical system (e.g. profiling) is directly affected by the operational features, but can ultimately be attributed to the conceptualisation.

h) Response burden (e.g. using – or not using – data that exists in business accounting systems or administrative sources) is again rooted in the conceptualisation.

References

Brion et al. (2014). The Geneva workshop. (Will provide ref later. Perhaps they also published somewhere else, later.)

Delden, A. van (2017). Issues when integrating data sets with different unit types. CBS Discussion Paper 2017-05. Available from www.cbs.nl.

Groves, R.M., F.J. Fowler jr., M.P. Couper, J.M. Lepkowski, E. Singer, & R.Tourangeau (2004). *Survey Methodology* (New York: Wiley Interscience).

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