On the Unit Problem in Business Statistics

Statement Paper Prepared for Participants' Discussion at EESW17

1. Introduction and Aim

Statistical methodology has thus far been lagging in providing guidance for considering and treating business units in the production and use of business and economic statistics. The Steering Committee of the European Network for Better Establishment Statistics (ENBES) has in late 2016 formed a working group[1] to prepare a short statement paper on the subject, for discussion at the 2017 European Establishment Statistics Workshop (EESW17).

The aim of this paper[2] 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 methodology for evaluating and treating the unit error from a total survey error perspective. Comments, thoughts, and suggestions of EESW17 participants, as well as other interested parties, are invited and warmly welcome.

2. 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 would find in every country (their definitions vary over countries, however). Another example is tax units, which exist in some countries, that are created for taxation purposes 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) (Figure 1). Or, as another example, the Institutional Unit, which is closely related to the ENT, may be subdivided

1 The working group consisted of Arnout van Delden, Sanjiv Mahajan, Norbert Rainer, Peter Struijs, Li-Chun Zhang, and Boris Lorenc (who acted as its moderator).

2 The paper was written by Arnout van Delden, Boris Lorenc, Peter Struijs, and Li-Chun Zhang, drawing on input from the working group's members.

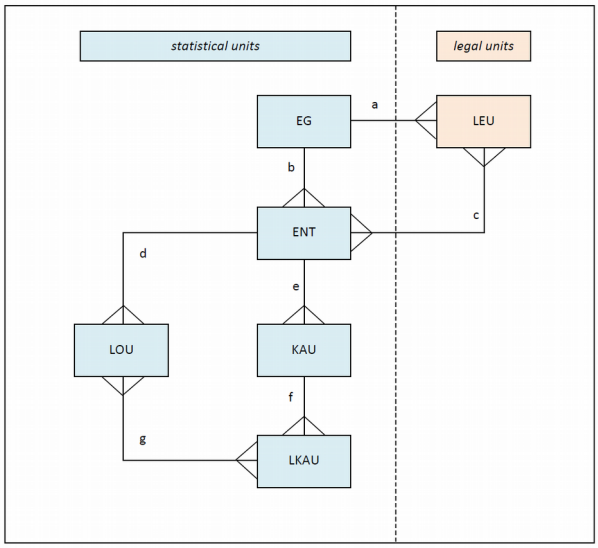

Figure 1: A system of statistical units (Eurostat, 2014)

into units of homogeneous production (UHP) for the purpose of national accounts, where the UHPs are not the same unit type as the KAUs.

3. Needs and Challenges Regarding 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 they may deliver data that refer to some proxy unit (e.g. LeU) instead.

Survey methodology sets up a framework which distinguishes between the study unit (of the target population), the sampling unit, the reporting unit (i.e. the entity within a business that is expected to actually deliver the required data), etc. Systems of statistical units for business statistics and national accounts are 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.

4. Unit Error and Unit Problem

Choice of the business unit type is a cornerstone decision in the design and operation management for a statistical product. Recognising its importance, we are here highlighting the

concept of the “unit error”. Unit error refers to errors in the statistical output that are caused by discrepancies in 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 error and our efforts to deal with it. The unit problem may be related to the practices of the business statistical products, the design of relevant statistical processes, as well as the conceptualisation of the system of statistical units.

5. Generic Situations for Unit Errors

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

1. There is **observation error** in the obtained data, 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 may result in erroneous statistical units, which tend to have a large impact on the output[3].
   • Consolidation (or apportioning) error may be unavoidable when the obtained data needs to be transformed (or disaggregated), because the required data is missing or simply unavailable[4]. For example, turnover of the VAT units may need to be transformed to that of ENTs, where the two units have many-to-many relationships. Or, the quarterly data may need to be disaggregated to monthly data.

2. **Implementation error** may be the case with respect to the relevant regulation or statistical unit model (e.g. the one in Figure 1). The discrepancy is between the results from correct and incorrect implementation. For instance, the regulation on business registers 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 Figure 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

3 Profiling is the activity to delineate the statistical units associated with large or complex businesses, including relationships.

4 Consolidation concerns excluding internal flows from values reported by units that are underlying a targeted composite unit type. De-consolidation is the opposite situation.

error in statistical products. More discussions regarding the conceptualisation difficulties can be found in the Village Bakery Example (Appendix 2).

6. Dealing with the Unit Problem Under the Total Survey Error Framework

We believe that unit error should be included and recognised in the Total Survey Error (TSE) framework, in the same spirit alongside the other types of error, such as sampling error, nonresponse 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 the 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.

7. Evaluation of the Unit Error Through User Value Criteria

Regardless of one's approach to the unit problem, the effects of the unit error that remain in the statistical output 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, Ph., T. Deroyon, & E. Gros (2014). “A first assessment of profiling in France”. Presentation given at the ENBES workshop “The Unit Problem in Business Statistics” (UNECE, Geneva, 10 November 2014). Available at: https://statswiki.unece.org/display/ENBES/ENBES+Workshop%2C+2014+%28Geneva%29+-+The+Unit+Problem+in+Business+Statistics?preview=/126353571/128024595/Unit%20Problem%20Workshop%202014.Brion-Deroyon-Gros.pdf.

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

Eurostat (2014). The Statistical Units Model, version presented to the Business Statistics Directors Group Meeting 24 June 2014. 2014. BSDG/24 June 2014/Doc. BSDG 201406 02b. Available at: https://circabc.europa.eu/d/a/workspace/SpacesStore/ac15c6fe-94ef-4e8d-ab4a-186da203aaf7/Background%20Doc%20201406%2002b%20The%20Statistical%20Units%20Model.pdf

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.

Appendix 1: Errors in Actual Data

With the actual obtained data we mean data obtained by applying the operational rules to actual units found in data sets (for instance an enumeration of legal units). We also consider actually obtained measures of the properties of interest of those units.

The Actual Units Themselves

Usually one needs two types of data to delineate statistical units. Firstly, one requires an enumeration of a set of base units from which a population of statistical units is derived. Second, one needs information on the relation between those base units, for instance information about ownership relations. Within this data set on relations also identification variables concerning the units are present. The base units and their relations are not necessarily present within the same set, it may well be that they need to be linked. Using the combined data, statistical units may be derived from base units using (operational) derivation rules. In addition to that: for the large and complex units, for instance units that operate globally, the statistical units and their relations may be derived manually, through profiling.

The following errors may occur in the actual obtained data (see van Delden, 2017), and lead to the unit error in produced statistics:

- Identification issues

  There are errors in the actual identification values of the base units. For instance the units may be identified by variables like “name” , “postal code”, and “phone number” and errors in their values (misspellings) might have occurred.

- Characterisation issues

  For classification variables of a statistical unit, that we need to derive by aggregation of classification variables for legal units, we usually select the main category (e.g. NACE code). Errors in the derivation of this main category might occur.

- Linkage issues

  There may be linkage errors between the actual (base) units and a data set with relations between the units that is needed to delineate the units. Those errors in turn might occur when linkage variables are not unique (such as “name” and “postal code”).

- Delineation issues

  Relations that exist between units are erroneously missing.

  > There are errors in the relations between the base units from which the statistical units are derived, for instance the level of ownership (unit A has 70% of the shares in unit B, whereas it should be 50%).

  Errors in the profiling of statistical units.

- Time delays

  Data available for the operations above are generally not reflecting the current situation, but the situation at a time t that has already passed. (Sometimes, it is not known which time t the data reflect.) If treatment of time is not done consistently within the production of statistics, so that it is not known which time point or period the data represent, this may lead to the errors in units. For instance, the information on the relation between base units is delayed, or the values of identification variable are outdated (old id-values are reported).

Measurement Issues

We are interested to obtain measures at statistical unit level but because this unit is “created”, and/or because the issues above have been at work, errors might occur when deriving these measures. This might occur when the data supplier (respondent) makes this derivation, but also when the statistical office makes this derivation for instance when we have administrative source data. We subdivide this into two situations:

The source data needs to be aggregated for many-to-one relations between source and statistical units. For continuous variables, one might need to correct for internal flows (consolidation). When the statistical office performs the aggregation, they might not have the information to correct for the internal flows.

The source data need to be disaggregated, for one-to-many relations between source and statistical units. This is an even more difficult situation than the first one. We need some statistical method to split up the data and in case of continuous data: to deconsolidate them. One might use auxiliary information to split the data.

Note that these two situations might occur within the same data set, for example when deriving turnover values available for fiscal units towards turnover for statistical units.

Appendix 2: Village Bakery Example

A bakery in a village, a family business, is organised as two LEUs, one for the operational business (producing and selling baked products) and one for the real estate (building). The reason for this arrangement is to reduce risk and have something set aside for inheritance (not subject to business liabilities etc. and good for tax reasons). The statistics authority is viewing this as a single ENT (def.: autonomy in decision making); which is also the idea behind but the concept: the firm = an economic actor. This bakery in fact is also a small supermarket.

The baker and his wife work daily in the bakery and shop. The wife is officially employed; the baker is the owner and manager. His sister does bookkeeping and administration for him (as well as for her husband, the baker's brother-in-law), which she is paid for, but not officially. Occasionally (mainly weekends and evenings), their son helps in the shop or for running errands (transport to or from the shop, etc). He is not formally employed, as he is a secondary school pupil. In deciding what products to bake, the baker takes into consideration the production of his brother-in-law, who is a farmer; in turn, the brother-in-law takes the needs of the baker into account when deciding on next season's crops. In short, they coordinate their business activities.

The work is done by four persons, though not all of them full-time; the location is the immovable (a section of the family home); the capital comes from running of the business, plus some years ago the baker borrowed from a bank to modernise his production line, which he now is repaying; the raw material input (flour) comes mainly from the brother-in-law; the retail products come mainly from other local producers plus some selected distributors; energy in part from the local wind farm, and in part from the main electricity network; the bakery does not produce any significant waste.

Some of the "unit errors" that can arise in this situation:

- Identification of units
  - How many units?

    In the legal framework, two. In the statistical framework, one.

    Comment: In producing business demography, this will give different result.
- Delineation of the ENT
  - Is there a sufficient level of autonomy in the baker's decision-making?

    If not, joint decision-making of the baker and his brother-in-law would need to be known to the statistical institute in order to make a correct delineation of the enterprise?
- Characterisation of units
  - What economic activity? Assuming one activity per unit:

    In LeU framework:

    bakery: baking

    building: renting business space

    In ENT framework

    baking

    Comment: Due to 'one activity per unit', in the ENT framework use of business space activity is not visible; in both frameworks the retail activity is not visible. Alternatively, allow the retail part to be visible somewhere, as it is adding value.

- How many employees/FTEs?
    - 1, 2, 3, 4 depending on whether only wife, or the baker, sister and son are also included.

A wider comment: a business generically has these properties: location (building), equipment, human workforce, material input, energy input, capital input, output (product, service), waste. Variations that result from some of these missing in a particular business arrangement introduce irregularities in produced statistics.

- factoryless production (no location)
- owing versus not owing (i.e. renting) the location
- outsourcing of work
- etc