Statistics Belgium
Working Paper

La Direction générale Statistique et Information économique propose des informations statistiques impartiales. Les informations sont diffusées conformément à la loi, notamment pour ce qui concerne leur confidentialité.

Nous classons les statistiques en huit domaines :

- Généralités
- Territoire et environnement
- Population
- Société
- Économie et finances
- Agriculture et activités assimilées
- Industrie
- Services, commerce et transports

Éditeur responsable : A. VERSONNEN
© 2008, DIRECTION GENERALE STATISTIQUE ET INFORMATION ECONOMIQUE | éditeur
B-1000 Bruxelles – 44 rue de Louvain

Statistics Belgium
Working Paper n. 19

The Integrated Business Survey (IBS): a challenge at Statistics Belgium

Maria Caterina Bramati[1]

February 2008

[1]Statistics Belgium, rue de Louvain 44, B-1000 Brussels. Maria-Caterina.Bramati@economie.fgov.be

Table of contents

Contents

The Integrated Business Survey (IBS): a challenge at Statistics Belgium

Maria Caterina BRAMATI

Abstract

In the past few years the modernization of business surveys has been extensively discussed by several NSI's in Europe. The MEETS program (*Modernisation of European Enterprise and Trade Statistics*) has been recently launched for the years 2008-2013 by EUROSTAT for encouraging the reform process at the European level in order to identify new areas for business statistics, to enhance the integration of data collection and treatment and to improve the harmonization of methods and concepts in business statistics.

At Statistics Belgium the debate has been brought especially through a revision of concepts and methods in business surveys with the aim of reducing the survey costs for the Administration and the response burden for the enterprises. Therefore, a project for exploring the feasibility of the integration of business surveys has been launched at the end of 2006, called the Integrated Business Survey (IBS) project.

The aim of this paper is to point out some concerns and to raise questions related to statistical methods in view of the IBS project, suggesting some solution paths.

In particular, this paper focuses on the sampling design methods which are illustrated by means of some numerical examples with data of businesses belonging to the sector of Construction.

The main guidelines on the steps of the statistical production process are given from both a theoretical and practical viewpoint. Feasibility constraints (IT technologies, divergence of fiscal and economic definitions, quality issues in the KBO register) are considered throughout the paper.

Acknowledgments

This work could not have been accomplished without the advice and support of many people.

I am deeply indebted to Denis Luminet for the many insightful conversations, his wisdom, his understanding and helpful comments on the text.

Many colleagues at Statistics Belgium provided a stimulating and fun environment in which to learn and grow professionally. I am grateful to them all.

1 Introduction: why integration?

The issue of survey integration has been object of several discussions in many Statistical Institutes of European and non-European countries.

First of all, by integration it is meant the process of combining the currently existing business surveys into a unique and coherent structure characterized by a whole coordinated statistical production process. Therefore, the horizontal complexity of the stove-pipe production process currently in use, in which each surveying step is led independently from the others is replaced by a vertical increase in complexity.

Many key-points have been addressed as advantages coming with integration, such as a reduction in costs, gain in efficiency, statistical productivity and data quality from the Statistical Institute viewpoint, and reduction of burden response from the businesses viewpoint.

Going towards integration is more than a need for Statistics Belgium, since it provides a solution to many problematic issues raised by the current way of surveying businesses, such as

- no harmonization of the definitions (economic, statistical...): impact on the universe, comparability of results, coherence, difficult use of the output (first need to understand the definitions used, the methods...);
- no harmonization of statistical methods (sampling, calibration, indexes);
- no scale advantages (no 'économies d'échelles') in using surveys-output as input for other surveys;
- no synergies/cooperation are activated between statisticians for improvement of the procedures and of the data treatment;
- ad-hoc IT implementation and intervention for each statistician/survey;
- no monitoring of the survey process;
- too many and different tools for survey management and websurveys;
- lack of documentation;
- no easy use/access of the past survey outputs (historical);
- no systematic approach to data quality.

Though on the one hand, integration boosts the improvement of the statistical practices, allowing for harmonization of concepts, definitions and methods, for synergies in the use of statistical tools, for monitoring and coordination of the statistical production process enhancing data quality, on the other hand it involves some increase in complexity of the process at both theoretical and practical level. This in turns implies some non negligible investment in terms of human capital and of IT tools.

However, the most challenging target, at least at this step of the project is to act at the conceptual level, enlarging the current notion of many single parallel surveys to one coherent variable-oriented structure able to satisfy the whole end-users' demand.

Not only the impact of integration will push changes in the perspectives of survey organization and IT implementation, but also it will require some adjustments and redistribution of the human resources. Therefore, the timing to achieve and complete the process is highly dependent mainly on two binding constraints such as the available infrastructure and the conceptual resistance.

The experience of integration is not new in the current practice of some Statistical Insitutes such as CBS, ONS and Statistics Canada among the others.

2 The methodological framework

The logic of the integration goes further beyond the simple summation of the existing surveys into one. The IBS (Integrated Business Survey) should be considered as a new *variable-oriented* structure which is not directly related to the current surveys structure. In other terms, one should consider the current organizational structure of 24 business surveys as one possible way of grouping (or a partition of) some output variables. Such a partition contains some redundancies (overlapping variables, information which could be derived by other sources, extra information not officially required etc..) which should be eliminated by the integration process.

To be more concrete, the starting point of the process should be indeed the required final outcome of the surveying process. This means that first of all one needs to look at the output side of the problem, which is represented by a series of variables (say *output* variables) to 'produce' for some final users.

Therefore, if the final objective of the surveys consists of a set of variables to deliver under some constraints (time, target population...) imposed by the statistical and federal authorities, the correct approach should be based on such variables. For this reason the variable-oriented structure of the IBS is suggested.

The integration process should be driven mainly at 4 levels: the **architecture** of the statistical production process, the statistical **methods**, the **data flows** (collection, storage, rule-based processing, metadata) and the **revision** of the economic variables involved in the surveys.

The objectives of the methodological framework are *standardization of methods*, *improvement of quality*, *transparency* and *IT-standardization* (to achieve by linking methods to tools).

In the next sections are explained the main guidelines and principles at the basis of the methodological framework.

2.1 The variable-oriented approach

From the viewpoint of the final outcome to deliver in terms of *output-variables* v_{o} contained in the set $\mathrm{V_o}$, the basic assumption on which the following analysis relies is that there exists a one-to-one mapping $f_{\mathrm{o}} : \mathbf{\Omega_o} \rightarrow V_{\mathrm{o}}$, where the domain $\mathbf{\Omega_o} = \mathrm{ID_o} \times \mathrm{SU_o} \times \mathbf{T_o} \times \mathbf{TP_o}$ is the cartesian product of sets $\mathrm{ID_o} = \{\text{identification of the variable}\}$, $\mathrm{SU_o} = \{\text{sampling unit}\}$, $\mathbf{T_o} = \{\text{frequency, delivering time}\}$, $\mathbf{TP_o} = \{\text{target population}\}$[2]. In other words, the output required by the final users of business surveys is a set of variables which are uniquely identified by 4 attributes: the **kind** of variable (added value, number of full-time equivalent employees etc..), the reference **sampling unit** (establishment, statistical unit or group of enterprises), the **time** (which might be an array of

[2]boldface symbols are used to distinguish sets and variables characterized by more complex objects, like arrays

information concerning time, like the survey frequency, the delivering time...), the **target population** (i.e. information concerning the breakdown details of the variable, like NACE class or size ONSS). For instance, an output-variable i is a combination of those characteristics exogenously[3] given, i.e.

$$v_{\mathrm{o}}^{i} = f_{\mathrm{o}}(\mathrm{id}_{\mathrm{o}}^{i}, \mathrm{su}_{\mathrm{o}}^{i}, \mathbf{t_o}^{i}, \mathbf{tp_o}^{i}).$$

Those attributes are defined by the requirements of the statistical authorities and more in general by recipients like EUROSTAT, National Bank of Belgium, etc. (indicated in the diagram by R).

Therefore, a survey j from the output side is a collection of output-variables $\mathrm{V}_{\mathrm{o}}^{j} \subseteq \mathrm{V}_{\mathrm{o}}$ subset of the class of all variables. In this view and under the hypothesis that no redundant variables are present across the current business surveys[4], the current survey structure can be seen as a way of partitioning the output variables in $\mathrm{V_o}$ into clusters (which are currently 24).

One of the advantages of this approach is that the structure of the survey is flexible to changes in the criteria defining the target population. To those changes would correspond new output variables to produce.

Now, ascending in the process to the *input-side* of the variable 'production', i.e. the **collection strategy**, similar assumptions to those of the output-side can be made.

An *input-variable* $\mathrm{v}_{\mathrm{I}}^{i}$ is an intermediate product in the survey process, a piece of information concerning the output-variable which is either collected (part of a questionnaire), or obtained by external sources (administrative records or other sources). In symbols

$$v_{\mathrm{I}}^{i} = f_{\mathrm{I}}(\mathrm{id}_{\mathrm{o}}^{i}, \mathrm{su}_{\mathrm{I}}^{i}, \mathbf{t_I}^{i}, \mathbf{tp_I}^{i}, \mathbf{es}^{i}),$$

where $f_{\mathrm{I}} : \mathbf{\Omega_I} \to V_{\mathrm{I}}$ is a one-to-one mapping with domain given by the cartesian product $\mathbf{\Omega_I} = \mathrm{ID_o} \times \mathrm{SU_I} \times \mathbf{T_I} \times \mathbf{TP_I} \times \mathbf{ES}$, where the subscript I indicates the input-side of the process.

The input-variables are the result of the combination of 5 attributes, the variable *identification*, the reference *sampling unit*, the *time*, the *target population* and the *external sources*. Those attributes in this case are defined by the statistician under some constraints determined by the outcome to deliver. The constraints which are generally respected in the current business surveys are

$$\mathrm{su}_{\mathrm{I}}^{i} = \mathrm{su}_{\mathrm{o}}^{i} \quad \mathbf{t_I}^{i} = \mathbf{t_o}^{i} \quad \mathbf{tp_I}^{i} = \mathbf{tp_o}^{i} \quad \forall \mathrm{v}_{\mathrm{I}}^{i} \in \mathrm{V_I}, \quad \mathrm{v}_{\mathrm{o}}^{i} \in \mathrm{V_o}$$

where it is assumed an input/output correspondence between the input variable $\mathrm{v}_{\mathrm{I}}^{i}$ and the output variable $\mathrm{v}_{\mathrm{o}}^{i}$. With those constraints it is clear that the statistician at present does not dispose of a lot of flexibility in the effort of optimizing and simplifying the business survey structure. Basically the statistician is allowed to act on the external source component only. A more intensive use of the administrative records was indeed the policy adopted by many statistical institutes in the past few years in view of a more efficient surveying process and of a reduction of the statistical burdens for enterprises. However, external sources are not always such a flexible tool for the statistician, since their use is constrained to the delivering time, the sampling units to which they refer (very often the legal unit) and the target population coverage. Therefore, several transformations and approximations are often required. It is clear that without the use of external sources and keeping the constraints enumerated above, to produce K output variables, almost K input variables are

[3] not chosen by the statistician, but defined by the end-user

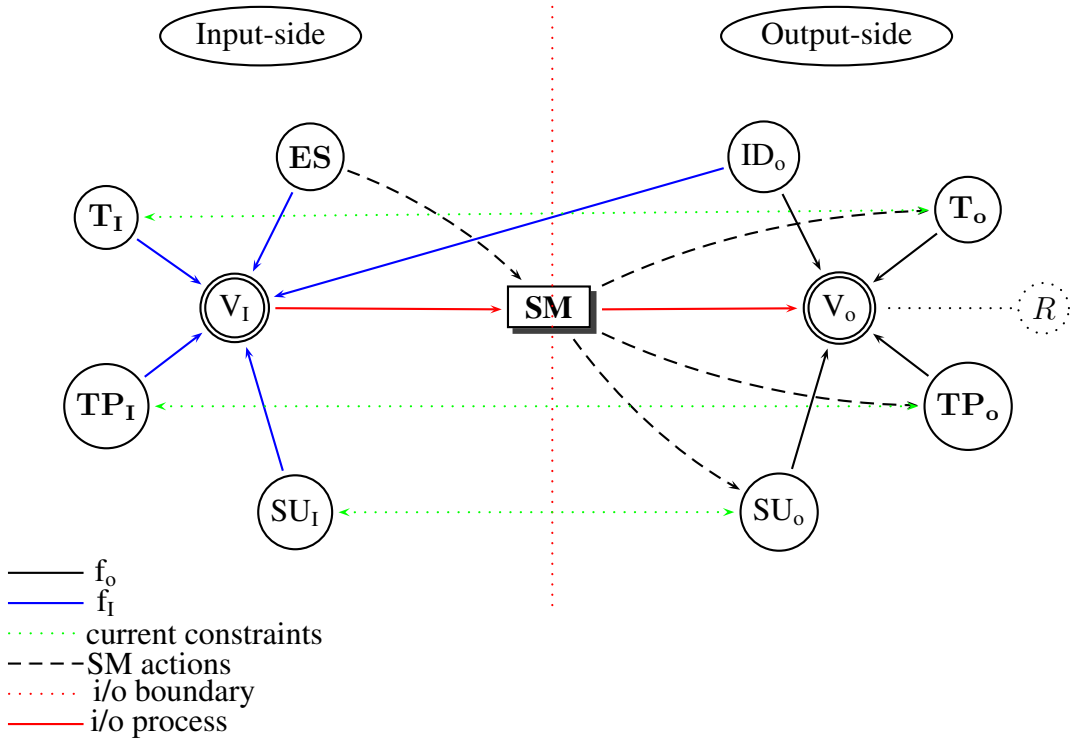
[4] this is to ensure the invertibility of mapping f_{o}

needed. Moreover, to redundancies in the output set V_o might correspond redundancies in the input set V_I.

Now, as in the case of the production process of a good, the crucial step of the variable production process is the transformation of the input variable into the output variable. This transformation is operated by a set of techniques which are known under the name of Statistical Methods. An appropriate and extensive use of those methods may help in relaxing some constraints under which the current business survey structure runs. At this point it is worth mentioning two remarks.

1. Statistical methods are designed to work in a 'simplified reduction' of the reality, in settings which correspond to mathematical models defined by sets of assumptions (sometimes strong and unrealistic, sometimes mild and reasonable). It is the job of the statistician to find the optimal model in the class of those which fit best the reality.

2. From the previous remark it is clear that statistical methods are not the only key-players in the transformation step. The statistician needs some other knowledge linked to the reality that he wants to model. The knowledge of the available economic information on the structure of businesses, of the definition of the economic variables of interest, of whether and how they can be approximated or predicted is crucial in the choice of the statistical techniques.



In synthesis, once that redundancies in V_I are eliminated, the main objective of the IBS project is to act on the input-side of the survey process (collect), relaxing some of the constraints imposed by the output-variables by means of the statistical methods and the economic knowledge. This can entail a reduction in the number of input-variables.

For instance, for some variables the time constraint could be relaxed combining the information coming from external sources and estimation techniques. Also the target population constraint could be partially relaxed finding convenient estimators of the total (with properties like the *stratification-equivariance/invariance*). As for the sampling unit constraint, conversion methods from one definition to another would be helpful.

3 Classification issues

As stated in the previous section, surveys can be seen as clusters of variables. Since variables are characterized by attributes like time, target population, sampling unit etc., variable classification can be done by one single attribute or by a combination of them.

The classification is a requirement for grouping surveys according to their 'similarities'. There is no evidence of the advantage of keeping the current partitioned structure of surveys in the sampling design, collect and treatment of data. Of course, to respond to institutional obligations the output (at the dissemination step) could be converted by an interface into the usual disaggregated survey structure.

Several clustering criteria might be adopted. Basically, 4 main grouping criteria can be considered:

1. **Economic** criterion. It is based on the category of variables of interest in a given survey. For instance, surveys can be clustered by groups of variables which summarize the economic structure of a business. These groups can be the *activity*, the *expenditure*, the *investment*, the *employment* and the *innovation* of the firm. However, this classification of variables into groups might represent a wrong picture of the structure of a business. Therefore, it would be interesting to identify the (hopefully) correct classification of the economic activities across businesses by the *association criterion* which clusters into the same group the activities often *associated* in the businesses. A definition of "proximity" between activities needs to be introduced in order to reflect similiarities between their distributions.

2. **Overlapping population** criterion. Surveys are grouped according to their target population. Survey clusters should be formed on a quantitative basis, i.e. computing and comparing the overlapping population rates for surveys.

3. **Enterprise-based** criterion. This criterion is *business-centric* in the sense of adopting the viewpoint of the business. To each business is attributed a code according to the survey for which it enters the target population. In this way every business can be identified by a series of numbers which represents the list of surveys to which it is concerned. Then, enterprises can be grouped by the survey code obtaining clusters of *similar* businesses in terms of survey subject. A ranking of the clusters can be made (in principle if the surveys are 24, there are 2^{24} possible clusters for the enterprises, but only a subset of them currently existing) and descriptive statistics can be easily calculated (like the mode cluster). Not only this classification might lead to a new way of classifying enterprises (according to the *survey-similarity*) but also it provides crucial insights on the way in which surveys are likely to naturally cluster between them. This topic should be object of further discussion and research.

4. **Geographical** criterion. Geographical dimension is an important component for understanding economic interactions of businesses. For instance, territorial proximity can be used as a measure of *structural similarity* regarding some economic aspects. Innovation might be one of them (some literature available on the geography of innovation). Little attention has been payed until now to the impact of the locational component for businesses to the sectorial economic activity and to the geographical agglomeration (industrial districts). It would be interesting to increase the efforts in this direction.

Of course, some of these criteria can be used as complementary, some others might be non interesting because not feasible or inefficient. Comparisons between criteria should be done on a quantitative basis, using (when possible) cluster analysis techniques.

All of these criteria present both advantages and disadvantages. However, the *enterprise-based* criterion might be privileged (at least as a starting point for research in classification) because it entails a new clustering approach in line with the philosophy of the Integrated Business Survey project.

For instance, the classification criterion of surveys by *overlapping population* is equivalent to a classification of variables by the target population attribute.

The classification according the *enterprise* criterion is also a classification of variables by the target population attribute. Indeed, it can be seen as the dual of the overlapping population criterion.

Also classification by the time attributes is possible. For example, variables can be grouped by the time frequency, i.e. monthly, quarterly, bi-annual, yearly etc.. variables. In the same way, classification by delivering time is possible, or by the sampling unit attribute, or even by kind of variables (*economic* criterion).

Clearly, each of these classification methods has the limit to consider only one of the variable attributes at time. Of course, a classification by a combination of the variable attributes could be a more efficient solution. In what follows, 2 ways of classifying variables (and therefore surveys) by target population are considered.

Some further effort needs to be done in order to construct a classification by the combination of the variable attributes.

3.1 Classification by target population

Target populations for each surveys are constructed using the NACE and the size (turnover and number of employees) requirements. The business surveys considered are 23 (the survey on transports of goods is not taken into account). To make comparisons feasible, the statistical unit used in this exercise is the establishment and the time unit is the year. Cluster Analysis, also called data segmentation, has a variety of goals. All relate to grouping or segmenting a collection of objects (also called observations, individuals, cases, or data rows) into subsets or "clusters", such that those within each cluster are more closely related to one another than objects assigned to different clusters. Central to all of the goals of cluster analysis is the notion of degree of similarity (or dissimilarity) between the individual objects being clustered. There are two major methods of clustering: hierarchical clustering and k-means clustering. The k-means algorithm assigns each point to the cluster whose center (also called *centroid*) is nearest. The center is the average of all points in the cluster, i.e. its coordinates are the arithmetic mean for each dimension separately over

all the points in the cluster. Since this approach is based on averages, it does not fit our case in which categorical variables are considered. In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to n clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the n objects into groups, and divisive methods, which separate n objects successively into finer groupings. Agglomerative methods An agglomerative hierarchical clustering procedure produces a series of partitions of the data, $P_n, P_{n-1}, \ldots, P_1$. The first P_n consists of n single object 'clusters', the last P_1, consists of single group containing all n cases. At each particular stage the method joins together the two clusters which are closest together (most similar). (At the first stage, of course, this amounts to joining together the two objects that are closest together, since at the initial stage each cluster has one object.) Differences between methods arise because of the different ways of defining distance (or similarity) between clusters.

3.2 Ward's hierarchical clustering method

Ward (1963) proposed a clustering procedure seeking to form the partitions $P_n, P_{n-1}, \ldots, P_1$ in a manner that minimizes the loss associated with each grouping, and to quantify that loss in a form that is readily interpretable. At each step in the analysis, the union of every possible cluster pair is considered and the two clusters whose fusion results in minimum increase in *information loss* are combined. Information loss is defined by Ward in terms of an error sum-of-squares criterion (ESS).

3.3 Classification by the overlapping population criterion

The aim of this classification is to form clusters of surveys by means of similarities between the sampling units belonging to each survey universe.

3.3.1 Clusters of surveys

Surveys are grouped by means of the hierarchical cluster approach, using Ward's method with the Squared Euclidean distance and the Lance and Williams measure (based on non-metric multidimensional scaling, see Williams and Lance, 1965).

The squared euclidean distance for a binary variable takes value $d^2_{ijk} = 0$ if cases i and j both have attribute k "present" or both "absent", or 1 if attribute k is "present" in one case and "absent" in the other case. For instance, if we consider a partition of surveys into 5 groups (unfortunately hierarchical methods do not provide with the optimal number of clusters, which is of choice of the end-user), the following groups are formed:

1. SBS, Afval-ind, Dethand
2. Arbeidk, Losa, ICT, CVTS
3. Prod-ind, Lonen
4. Prod-prijs, Toer, Cine, Kredieten, Wijn, Graan, Melk, Slacht, Groente, Fruit

5. Oogsten, Novenq, Afval-land, Landbouw-telling.

whereas if Euclidean distance is used, the following clusters are made

1. SBS, Afval-ind
2. Arbeidk, Losa, ICT, CVTS
3. Dethand
4. Prod-prijs, Prod-ind, Lonen, Toer, Cine, Kredieten, Wijn, Graan, Melk, Slacht, Groente, Fruit
5. Oogsten, Novenq, Afval-land, Landbouwtelling

The two classification are quite close. It is clear that in both cases 'agricultural' surveys are grouped together; the SBS is grouped with the wastes of the industry (in some cases with the Dethand). The surveys concerning the human capital and technology are also clustered. When the survey populations are considered in terms of year-equivalent units, the previous results are partially confirmed. Adding the time constraint in the classification analysis does not produce conclusive results (at least in this approach).

3.4 Classification by the business criterion

Establishments are clustered by combination of surveys.

3.4.1 Clusters of businesses

More than half of the establishments are subjected to be surveyed by SBS and Afval-ind. There are few ways in which surveys are combined.

Two-step cluster is used. This cluster method is designed to handle very large datasets. It has two steps 1) pre-cluster the cases into many small sub-clusters and 2) cluster the sub-clusters resulting from the pre-cluster step into a number of clusters automatically selected. If the partition into 5 groups of units is considered, the following clusters are made.

Clusters of establishments by surveys:

1. Oogsten, novenq, afval-land, landbouwtelling, groenten, fruit
2. SBS (69.2%), afval-ind(67.4%)
3. SBS (11.1%), afval-ind (13.4%), arbeidkost (18.6%), LOSA (21.6%), Dethand (4.1%), Kred (78.4%), prod-prjis (84.8%), prod-ind (3.8%), CVTS (7.3%), Tour (83.4%), ICT (37.4%), Cine (68.6%), wjin (94.9%), graan (93.9%), melk (64.9%), slacht (85.7%)
4. SBS (12.7%), Afval-ind (12.4%), dethand (85.9%)
5. SBS (7%), Afval-ind (6.8%), dethand (10.1%), arbeidkost (81.4%), Losa (78.4%), CVTS (92.7%), lonen (100%), prod-ind (96.2%), Kred (21.6%), prod-prjis (15.6%), Tour (16.6%), ICT (62.6%), Cine (31.4%), wjin (5.1%), graan (6.1%), melk (35.1%), slacht (14.3%).

In parenthesis are indicated percentages of those establishments between the ones entering potentially the survey, which are selected for the given group. For example, group 1 is formed by the 69.2% of establishment potentially surveyed by SBS and by the 67.4% of establishments potentially participating to the survey on industrial wastes. Of course, in this group there might be also establishments which enter potentially both surveys, and therefore contained at the same time in the 69.2% of SBS and 67.4% of Afval-ind.

3.5 Comparisons between classification

The results obtained by classification analysis give some interesting insights on the way in which the survey integration should be conducted. Of course, it is rather difficult to judge which of the proposed classifications perform best in view of the integration of surveys.

However, comparisons can be made by means of multinomial logistic regression models. The idea is to evaluate whether the probability for a given establishment to belong to one specific cluster is well-explained by the NACE and ONSS criteria set.

From the estimation of the logistic model parameters, the following conditional probabilities are obtained for each classification at 5% significance level of regression coefficients

Overlapping population classification	Enterprise-based classification
$\hat{\pi}_1 = [1 + \exp\{.06y_1 + .63y_2 - .79y_3 - 1.34y_4\}]^{-1}$	$\hat{\pi}_2 = [1 + \exp\{-.89y_3 - 1.37y_4 - 1.29y_{10}\}]^{-1}$
$\hat{\pi}_2 = [1 + \exp\{-.27y_1 - 2.71y_2 + 6.27y_3 - .92y_5$ $+6.36y_6 - 4.49y_7 + 1.37y_8\}]^{-1}$	$\hat{\pi}_3 = [1 + \exp\{-35.12 - .23y_1 - 10.59y_2 + 5.82y_3$ $+4.33y_6 - 10.01y_7 + 6.76y_{10} + 5.96y_{11}$ $+7.18y_{12} + 6.58y_{13}\}]^{-1}$
$\hat{\pi}_3 = [1 + \exp\{-.11y_1 - .56y_2 - 1.62y_3 + 6.55y_4$ $+1.79y_5\}]^{-1}$	$\hat{\pi}_4 = [1 + \exp\{-.11y_1 - 1.62y_3 + 6.55y_4 + 1.79y_5$ $-1.92y_6\}]^{-1}$
$\hat{\pi}_4 = [1 + \exp\{.16y_1 + .55y_2 - 1.84y_3 + .82y_5$ $-6.22y_9 + 13.27y_{10}\}]^{-1}$	$\hat{\pi}_5 = [1 + \exp\{.17y_1 + 5.60y_2 + .83y_3$ $+1.72y_6 + 9.36y_7\}]^{-1}$

where the variable y_1^i is the number of employees in the establishment i (ONSS source), whereas y_j, $j = 2, \ldots, 13$ are binary variables defined for establishment i as

$$y_2^i = 1 - I_{\text{SBS}}, \quad \text{where} \quad I_{\text{SBS}} = \begin{cases} 1 & \text{if} \quad i \in \text{SBS} \\ 0 & \text{otherwise} \end{cases}$$

$$y_3^i = 1 - I_{\text{ICT}}, \quad \text{where} \quad I_{\text{ICT}} = \begin{cases} 1 & \text{if} \quad i \in \text{ICT} \\ 0 & \text{otherwise} \end{cases}$$

$$y_4^i = 1 - I_{\text{DET}}, \quad \text{where} \quad I_{\text{DET}} = \begin{cases} 1 & \text{if} \quad i \in \text{DET} \\ 0 & \text{otherwise} \end{cases}$$

$$y_5^i = 1 - I_{\text{ARB}}, \quad \text{where} \quad I_{\text{ARB}} = \begin{cases} 1 & \text{if} \quad i \in \text{ARB} \\ 0 & \text{otherwise} \end{cases}$$

$$y_6^i = 1 - I_{\text{LOSA}}, \quad \text{where} \quad I_{\text{LOSA}} = \begin{cases} 1 & \text{if} \quad i \in \text{LOSA} \\ 0 & \text{otherwise} \end{cases}$$

$$y_7^i = 1 - I_{\text{CVTS}}, \quad \text{where} \quad I_{\text{CVTS}} = \begin{cases} 1 & \text{if} \quad i \in \text{CVTS} \\ 0 & \text{otherwise} \end{cases}$$

$$y_8^i = 1 - I_{\text{KRED}}, \quad \text{where} \quad I_{\text{KRED}} = \begin{cases} 1 & \text{if} \quad i \in \text{KRED} \\ 0 & \text{otherwise} \end{cases}$$

$$y_9^i = 1 - I_{\text{LON}}, \quad \text{where} \quad I_{\text{LON}} = \begin{cases} 1 & \text{if} \quad i \in \text{LON} \\ 0 & \text{otherwise} \end{cases}$$

$$y_{10}^i = 1 - I_{\text{PPRIJ}}, \quad \text{where} \quad I_{\text{PPRIJ}} = \begin{cases} 1 & \text{if} \quad i \in \text{PPRIJ=prod-prijs} \\ 0 & \text{otherwise} \end{cases}$$

$$y_{11}^i = 1 - I_{\text{TOER}}, \quad \text{where} \quad I_{\text{TOER}} = \begin{cases} 1 & \text{if} \quad i \in \text{TOER} \\ 0 & \text{otherwise} \end{cases}$$

$$y_{12}^i = 1 - I_{\text{WIJN}}, \quad \text{where} \quad I_{\text{WIJN}} = \begin{cases} 1 & \text{if} \quad i \in \text{WIJN} \\ 0 & \text{otherwise} \end{cases}$$

$$y_{13}^i = 1 - I_{\text{GRAAN}}, \quad \text{where} \quad I_{\text{GRAAN}} = \begin{cases} 1 & \text{if} \quad i \in \text{GRAAN} \\ 0 & \text{otherwise.} \end{cases}$$

For example, the conditional probability for an establishment with 5 employees which belongs to the universe of SBS and DETHAN of entering cluster 3 in the overlapping population classification is $.59$, whereas the conditional probability for the same establishment of entering cluster 4 in the enterprise-based classification is $.91$.

The table displayed above, the information criteria and goodness-of-fit measures lead us to conclude that the target population criteria are well explained by both classifications, showing perhaps a slightly better performance of the enterprise-based criterion.

In this few pages some classification issues have been addressed, especially focusing on the *target population* component of an output-variable set. Of course, there are many other ways of classifying variables, using other variable components (such as time, sampling units etc.) or a combination of them. One could also set a *multi-step* classification, using a priority list of criteria. For

example, variables could be classified first by sampling units definition (group variables focusing on establishments and group of variables focusing on statistical units), then in each cluster a new classification is operated by another component, for example the time-frequency (annual variables, quarterly, monthly etc.) and so forth... It is clear that a priority list is constructed mainly on the basis of the constraints which are considered more binding, in which statistical methods cannot help much.

Classification by target population criteria, in particular by the enterprise-based approach, has the main advantage to allow the construction of criteria which include each new-born business into the appropriate universe for the target variables. On the other hand, this system would not be enough flexible with respect to changes in the target population criteria. Suppose for instance that for some variables a different industry breakdown is required by the statistical authority. Then a new classification by target population is required for all businesses to create the universes for the variables. If the creation of new variables or changes in their breakdown industry happen often, the enterprise-based criterion require several updates and moreover does not allow for comparisons across the periods of change.

4 Description of the survey structure

Several issues must be taken into account in the survey architecture. Some of them depend on the way businesses are organised and on how they behave in the market. Therefore, several studies on Belgian firms have been considered in order to have a picture of the micro- structure of the business from a decisional and organizational viewpoint (approach within firm), the existing connections and interactions between firms (approach between firms) and the impact of macroeconomic policy on them. In this section are considered the main principles leading to a possible configuration of the IBS.

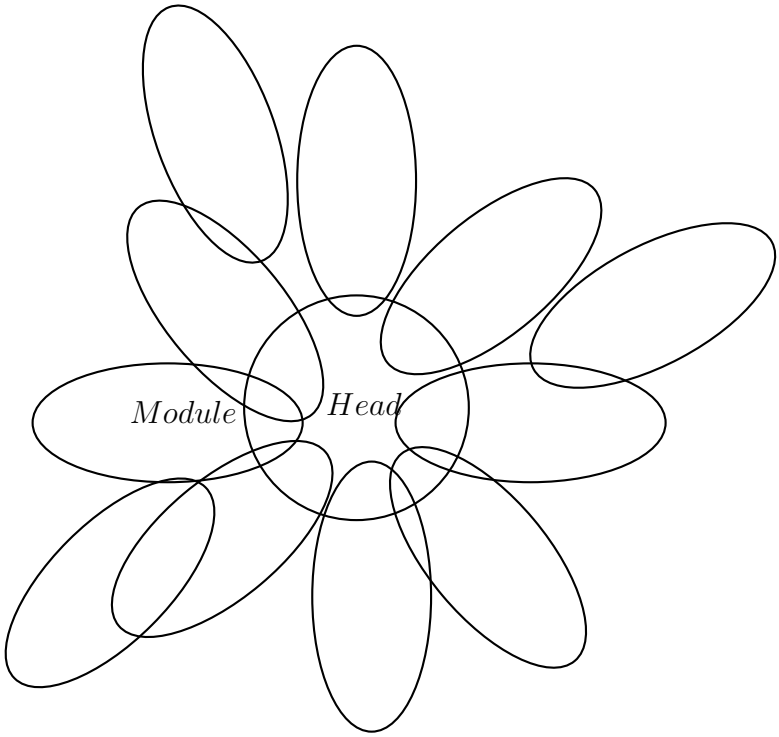
The next step is to conceive a new coordinated and coherent structure for business surveys. One possible approach could be to focus on the most rich and complete (in terms of information which is of common interest for other surveys) survey already existing (like the SBS, for instance) and to set it as the *head-survey* which is the building block from which the other ones (called *satellites* or *petals* of a flower) stem. The head-survey is based on 5 categories (or modules) of key-variables concerning the **activity**, the **expenditure**, the **investment**, the **employment** and the **innovation** of the firm.

Satellite surveys are conceived as external modules (like the petals of a flower) of those composing the head-survey. They consitute a flexible tool for specific needs like

- different surveying periodicity;
- information on variables not contained in the head-survey;
- information on variables relevant only for subsets of the universe, such as some specific sectors of economic activity (for example construction), or some specific typology of enterprises (selected by a given criterion, like the size-class).

Of course, integrating the existing business surveys into a unique, coordinated and centralized structure does not necessarily imply a simplification in the methodology. For instance, the sampling design(s), the extraction of the sample, the data collection and the estimation step might

incur an increase of complexity, which means a transformation of the actual *horizontal* complexity (many business surveys living independently) into a *vertical* complexity.



5 The structure of IBS

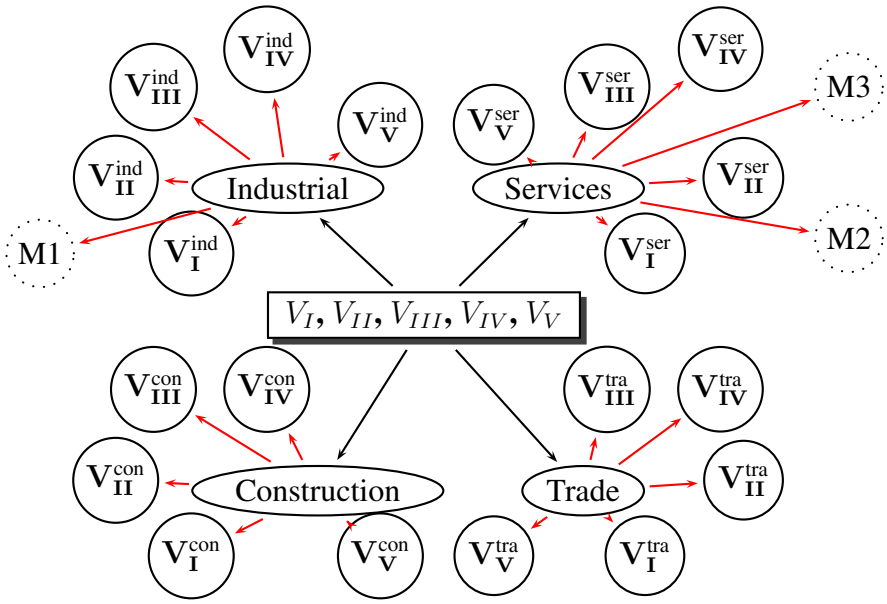
The structure of the IBS is made of a central core composed by I groups of variables V_1, V_2, ..., V_I, each of which contains highly correlated variables, or at least linked by a common economic concept. In the example below $I = 5$ and the choice of the groups of variable is

- V_1: production and activity variables (like value added, turnover,...);
- V_2: financial structure variables (like liquidity, solvability,...);
- V_3: employment variables (hours worked, salaries,...);
- V_4: investment variables;
- V_5: innovation (R&D expenditure).

Each group of variables might correspond to a single questionnaire to administrate to businesses. Questionnaires are adapted according to the NACE sector to which businesses belong. Four main sectors of activity are distinguished: Industry, Construction, Trade and Services. When groups V_1, V_2, ..., V_I are considered in each sector, they do not contain necessarily the same variables since they may differ across sectors. For instance, some variables concerning production prices and quantities might belong to the group V_1 containing production and activity variables for

the sector of Industry and Construction, but they will not be in the V_1 group for Trade and Services sectors.

Therefore, when the core of variables of interest are applied to each sector, the outcome is given by $I \times J$ groups V_i^j of variables, where $j = 1, \ldots, J$ is the index for the sector (in the graphics $j = \text{ind}$, con , tra , ser stands for industry, construction, trade and services respectively) and $i = 1, \ldots, I$ indicates the group definition (in the example: production and activity, financial structure, employment, investment and innovation), as the graphics illustrates.



Therefore, the basic structure of the IBS is composed by $I \times J$ modules obtained crossing groups of variables with sectors of activity. Such construction allows for a certain degree of flexibility in terms of

- surveying **time** constraints which very often depend on the sector of activities and on the survey subject;
- **sampling design** which would embody differences and peculiarities of firms over sectors, allowing for an ad-hoc optimal stratification depending on the group of variables of interest;
- reduction of **questionnaire size** (not one big questionnaire administrated to firms of all sectors, but light modules spread across firms of the same sector) by a customization according to sector and survey variables;
- **ease of management** of the modules and of task fragmentation.

The frequency of the survey is annual. Of course, some special cases may arise, like variables required with a higher frequency or *una tantum* surveys. In this event the structure of IBS allows for external modules which could be added to the cluster of modules concerning the sector of interest (in the picture are indicated with M1-M3).

5.1 Definition of the statistical unit and the sampling frame

In this view, one of the most important requirements is to clearly define the statistical unit. The majority of business surveys focuses on the **enterprise** (economic definition, see EUROSTAT) as statistical unit. Some surveys (PRODCOM, labor costs) require the **establishment** as the statistical unit.

Basically, two possible directions might be followed.

The easiest one is to fix a common definition of statistical unit for both the head-survey and all the survey-modules. The main advantage is the gain in the *consistency between surveys* (thanks to this standardization comparisons are allowed). The mains disadvantage is that the *consistency within surveys* is not straightforward. For instance, if the common statistical unit is set as the legal entity then a survey interested in establishments will need adjustments and revisions of definitions/variable/statistical methods.

The second way is to create a *multilevel* definition of the statistical unit according to three criteria contained in the **Council Regulation (EEC) N. 696/93** of 15 March 1993, i.e. the *institutional*, the *geographical* and the *activity* criteria. This allows for definitions of statistical units which range from the establishment (or local unit), which would be the *primitive* unit, to statistical units obtained by combining primitive ones (like legal units, enterprises and enterprise groups[5]). One of the advantage of this approach is that it is possible to obtain a specific universe for each survey as a collection of clusters of primitive statistical units. If N primitive units constitute the universe, then $2^N - 1$ non-empty clusters can be formed (considering all possible ways of grouping the units). Therefore, it is necessary to have the information which allow to identify the clusters at each definition level. Furthermore, this approach would allow to respect the EUROSTAT requirements on definitions of statistical unit (especially the economic definition of the enterprise would be applied). On the other hand, the lack of a national register of all the establishments (though some legal entities must declare the address of their establishments) and of the enterprises (but only of institutional units) represent a limit to this method.

IDEALLY: Adopt a statistical structure which is a realistic, accurate and almost untouched picture of the structure of a business -the structure of the business as viewed by the business itself (see Hunsberger, Beaucage and Pursey, 2005). In other words, to reduce the survey burdens it is important to collect the data where they are supposed to be. A study of the structure of the business is needed. It would be nice to link classes of economic variables to components of the business, in order to locate the information in the business. This approach might produce good results depending on the way of collecting data. In case of virtual collection (via web-surveys) such an approach does not necessarily involve a reduction of the response burden for businesses.

To choose between the two approaches is necessary to

- find a common definition of statistical unit satisfactory for all business surveys and estimate the impact of uniforming all business surveys to such definition;
- know, or at least have an idea of the size of the population according to the definitions of statistical units required by business surveys. For instance, it would be useful to know the size of the population of establishments, the size of the population of institutional units, the size of the population of enterprises and the size of the population of the enterprise

[5] enterprises can be combinations of legal entities and legal entities can be combinations of local units, nested in a hierarchical structure.

groups. Comparing the discrepancy between these numbers might help (roughly) to clarify the decision on the direction to follow. Also the rate of entities for which local unit≡legal unit≡enterprise, local unit≡legal unit, legal unit≡enterprise etc. could be helpful.

- promote a feasibility study (joint with business demographists) for the creation of registers for the 4 populations. For surveys focusing on establishments it might be necessary to tackle with the problem of coverage. Alternatives to stratification which do not require the knowledge of the target population might be a solution to take seriously into account.

The problem of under-coverage/over-coverage of the target population is an important issue, especially the over-coverage (enterprises in bankruptcy or non-active do not disappear contextually from the registers). Improvements of the data quality in the KBO register might reduce the coverage bias. It would be necessary to construct a measure of the coverage error.

5.2 Auxiliary information

The choice of the auxiliary variables to use should be more depending on the type of variable of interest (activity, expenditure, investment, employment and innovation). For instance, it might be logical to consider the turnover a good auxiliary variable for activity-type variables, but it is not maybe the case for investment-type variables.

Moreover, auxiliary variables need to be statistically reliable, i.e.

- coherent (homogeneous construction, same reference year),
- consistent with their definition,
- referring to the same universe as the variables of interest (same definition of statistical unit),
- exhaustive for the target population,
- available according to the required deadlines and with high updating frequencies,
- of good quality (issues of missing observations and outliers),

since they might be used for the sampling design and for the calibration.

5.2.1 The case of SBS: the turnover

In the Structural Business Survey the main auxiliary variable is the turnover (from VAT register). This variable, obtained as the difference between total transactions and credit notes summed over the year, might take also zero or negative values. This can lead not only to definition problems (non-active enterprise according to Eurostat) but also in the sampling design, in the extrapolation and in the calibration.

For instance, let $x_t^{(i)}$ be the yearly turnover at time t for unit i and $x_{t-1}^{(i)}$ the yearly turnover at time $t-1$. If $x_t^{(i)} \cdot x_{t-1}^{(i)} \leq 0$ or $x_{t-1}^{(i)} = 0$ (even worse), it does not make sense the extrapolation of a variable (like number of hours worked) done scaling the same variable of the previous period by the ratio $x_t^{(i)}/x_{t-1}^{(i)}$.

On top of this, since turnover is obtained by summing monthly (sometimes quarterly) values of the VAT declaration, if there are some gaps for few months, this is not taken into account (simple sum is made over the available data points on the reference year). This cause a problem of under/over-estimation of the yearly turnover. An **interpolation** technique should be used to fill in the gaps in order to obtain coherent yearly data.

At present, the detection of anomalies is made by the statistician by comparing aggregate turnover by sector with a trend index by sector. Corrections are made either after a contact with the enterprise (for big-size enterprises) or directly by the statistician (for more evident mistakes, like scale errors). This practice of 'subjective cleaning' should be replaced by a more theoretically rigorous and less time-consuming approach. An automatic check and correction for **anomalous data-points** should be performed either at this stage by the *outlier-rejection* procedure (see Cook, 1977), or at the sampling design construction step and at the estimation step by the *bounded-influence function* approach (see Hampel et al., 1981) or by the *high breakdown-point* approach (see Rousseeuw and Leroy, 1987).

Furthermore, **seasonal** adjustments should be made for monthly and quarterly turnover, if required. Both the raw series and the adjusted one should be available in the database.

Starting from April 2007 the Belgian tax system will not require anymore a VAT declaration for each legal unit belonging to the same enterprise or group, but only a consolidated one for the whole. Therefore, only external flows will be recorded and no trace of intra-units flows will be available anymore. As a consequence, the turnover variable will refer to the consolidated activity of the enterprise or group which do not coincide with the statistical unit as defined in the SBS at present. On the other hand, turnover could help to reconstruct the link between the register of the legal units and the register of the enterprises. For surveys with statistical units other than enterprises, the turnover variable (from VAT) should require either an adjustment (according to the definition of the statistical unit and its relation with the enterprise) or replaced by other variables.

First issue to consider is the request of statistics coming from final users, such as public institutions at international, European, national and regional level and private users (academics, businesses, citizens...). Such demands are characterized by specific requirements in terms of variables, survey units, coverage, precision, industry breakdown, surveying and delivering time... information which are gathered and contained in the concept of **metadata**. Metadata are organized through rules which feature the production process steps. Therefore, given that the statistical production process is structured by metadata, it is straightforward that the integration process should assume the same output-driven perspective.

5.3 Statistical methods

Once the output requirements have been set, some information (*input*) needs to be provided in order to satisfy the final demand. The input information, or *input variables*, are submitted to the production process and treated according to specified methodological rules. Such input information can be obtained by direct surveying or by administrative records or other external sources. Several treatments of the input data are possible, they range from the identical transformation to the model-based estimation and forecasts.

Since one of the main objective of integration is to reduce statistical response burdens for businesses, an improved and more extensive use of statistical methods could help in decreasing the volume of survey questionnaires. Therefore, one of the main feature of the IBS is to increase the

use of model-based estimation and forecasts whenever possible in the transformation process to produce the required output.

5.4 Administrative records and simplification

The use of administrative sources is fundamental in the effort of reducing surveying burdens for businesses. Not only this could be achieved by simplifying questionnaires, but also decreasing the sample size by dropping from the survey some businesses for which administrative entries can be provided. The availability of *reliable* administrative records, in terms of coverage, quality and coherency, is the main requirement for simplifying questionnaires and limiting the sample size for the IBS.

5.5 Sectors of activity

One relevant factor which contributes to characterize and differentiate the enterprise structure and behavior is the NACE industry to which it belongs. For instance, firm productivity and wage structure, the strategy and timing of price revision, the size of the enterprise (in terms of number of workers), the investment and financial structure, the reaction to monetary policy actions, the R&D strategies, the geographical dispersion and other characteristics can be very different depending on the sector of activity. Several studies based on empirical findings on Belgian firms confirm this intuition.

Lallemand, Plasman and Rycx (2007) considers the wage structure and labor productivity stressing the fact that the wage bargaining lies at the sectoral level in Belgium.

Aucremanne and Druant (2005) shows that the price-setting behavior of a Belgian firm and its adjustment frequency is very different across sectors of activity. In particular, optimal price-setting is more recurrent in manufacture (45% of manufactory firms) than in services, where the rule of thumb is practiced (37% of firms in services); the price-review occurs mostly once per year, in January for the 43% of the firms and in December for the 9%. On average price are revised once every 12 months for services, every 7 months for construction, every 9.6 months for industry and every 9 months for retail.

Van Gastel (1999) provides a breakdown of large, medium and small enterprises (defined according to the number of workers) into the main sectors of activity (manufacture, construction, services and retail) showing that for the construction sector small companies are prevailing, whereas for some services, such as Banking, Insurance and distribution of daily food, large companies are prevailing.

Butzen, Fuss and Vermeulen (2001) investigates some of the effects of monetary policy on firm's investment behavior for each sector and for large and small size firms separately. Their main finding is that for manufacturing and construction firms react negatively to user cost changes and positively to value added growth and cash flow changes, whereas the investment dynamics for services is different accordingly to the size of the firm (smoother for large firms than for small ones). In sum, the results support the hypothesis of a differential impact across sectors and sizes of the interest rate and credit channel.

R&D activities are at the root of regional economic development and growth. It is common sense that R&D expenditures differ highly between sectors. This is confirmed by Teirlinck and Spithoven (2005) who studies the principal determinants for regional R&D activities and detect

successful and less successful R&D regions combining the differences in R&D performance by economic sector with a spatial dispersion of sector activity between districts.

The need to make a distinction between sectors of activity in the IBS is confirmed also by the different employment structure of the enterprises according to the kind of activity.

The distribution of and the frequency distribution of businesses belonging to the same universe with at least one employee (enterprises with ONSS number).

Comparisons are made on the basis of Also coefficients of variation for each sector (columns CV tot and CV ONSS) are provided.

The exercise of considering the percentage of enterprises with employees on the total number for each 5-digits NACE and the turnover share they represent on the total turnover of the considered sector (Structural Business Survey universe in 2005) shows how the employment structure of enterprises in manufacture sector differs from that of construction, trade and services. For the manufacture sector in almost every $5-$digits industry the enterprises with employees represent at least the 90% of the total turnover of the firms belonging to the same industry. This is not always the case for services, nor for the trade and construction sectors.

Such a result is a useful insight in the survey design for the manufacture sector for which sampling could be limited to ONSS-registered firms only (and for the non-surveyed ones some extrapolations could be done by VAT records, estimation...). Of course, such a strategy is not necessarily convenient for the other main sectors.

Furthermore, several existing surveys are focusing on $1-$digit NACE sector only, as it is the case of trade survey, PRODCOM, and even the SBS is composed also by 4 modules corresponding each to a $1-$digit NACE sector.

Therefore, it would be useful to distinguish 4 main modules for the IBS, one of each corresponding to a NACE $1-$digit sector, i.e. **manufacture**, **construction**, **trade** and **services**.

5.6 Geographical location

In this paragraph we discuss the importance of geography in the distribution of businesses across economic activities according to their economic size (measured in terms of number of employees and turnover). To evaluate the relevance of the geographical component it is used the 2-way and 3-way analysis of variance (ANOVA) of the turnover. The idea is to compare the percentage of variance (of turnover) explained by the NACE and the economic size (benchmark model) with respect to the model in which geography is introduced as additional factor. Such comparative analysis is performed for each of the 4 main sectors of economic activities: manufacture and mineral industries, construction, retail trade and services.

From the results of such analysis some interesting conclusions can be drawn[6].

- The impact of geography location of businesses is different according to the industry. Indeed, it seems that geography is more important as a determinant for turnover variability for manufacture, mineral and construction industries. The highest impact of geography is for the construction industry.

- For construction, mineral and manufacture industries it seems that the decrease in explained variance due to a going up in the NACE breakdown can be compensated by an increase due

[6]Results are available from the author under request.

to a going down in the geographical breakdown.

6 Sampling design

The choice of the sampling technique is highly dependent on the main objectives of integration. It is perhaps useful to recall them briefly.

- The reduction of statistical response burden for enterprises, especially for small and medium ones;
- reduction of costs in terms of budget and human resources for the maintainance of the current business surveys;
- increase the quality of business surveys in terms of precision of the estimates and quality of the collected data;
- enhancement of the image of Statistics Belgium offering high quality and modern statistics to institutional partners (National Bank of Belgium, Bureau Federal du Plan...), Academies and citizens.

6.1 Stratification

Stratified sampling has been proven to be the most efficient surveying technique under some basic assumptions (see Tillé, 2001) and it is currently in use at Statistics Belgium for the majority of the business surveys. Nevertheless, the stratification criteria need to be reviewed in order to fix new thresholds for the size criteria of the enterprises.

The main principles of stratification should consist of three dimensions: sector of activity (NACE classification), geographical location (NUTS classification) and economic size of the business. The number of digits for NACE and NUTS level should be fixed in order to ensure on the one hand homogeneous strata (in terms of the stratification variable, like the economic size for instance) and on the other hand *sufficiently populated* strata.

Of course, the choice of the NACE and NUTS digits should be consistent with the required breakdown of the output variables in order to get a better precision of the estimates on survey domains. However, in practice the choice of the NACE and NUTS breakdown is based on a trade-off between the required breakdown of the output variable and the one which would ensure sufficiently populated strata.

Once the bi-dimensional strata are constructed (geography and sector of activity), for each stratum enterprises are grouped according to their economic size in three classes:

1. Large enterprises: for them an exhaustive sampling is required, which means that all businesses belonging to this class are surveyed.
2. Medium and Small enterprises in the sampling frame: for them a simple random sampling is required.
3. Medium and Small out of the sampling frame: they do not enter the sampling frame (i.e. 0-probability of being sampled).

It is clear that in every stratum each business has not the same sampling probability which depends on the size class to which the business belongs. However, within each size class businesses have the same sampling probability.

Now, the key issue is how to group businesses in size classes. In the Structural Business Survey (SBS) currently the thresholds which distinguish large, medium and small enterprises are based on turnover and number of employees and are fixed following common conventions and practices (*ad-hoc* criterion).

It is necessary to review those thresholds on the basis of solid statistical and economic arguments, using perhaps other variables to better represent the economic size of a firm and applying optimization algorithms and basic statistical tools (like empirical distribution function).

The expected change is a reduction of the number of firms considered *large* and therefore belonging to the group of exhaustive sampling. As for the medium and small firms, a part of them (the ones with very low economic impact in their sector of activity) will not sustain any response burden. The tools suggested are presented in the following subsections.

6.2 Determination of strata bounds, sample size and allocation

The first step is the determination of the businesses to drop in the sampling process, which are the ones entering the take-0 stratum. Given a NACE and geographical location, a possible criterion is to choose a quantile of the cumulative distribution function of the auxiliary variable and to drop those businesses which are below it. In other words, this would disregard the enterprises which do not represent a relevant share of the auxiliary information used for the stratification design, for a given NACE and geographical location.

For example, if the auxiliary variable is the turnover and the take-0 quantile is chosen at 5%, then the take-0 businesses belonging to a given NACE and geography are those for which the turnover summed up does not represent more than the 5% of the total turnover for the given NACE and geography. For those dropped businesses some estimation techniques combined with administrative data could be used to make some inference.

If the 5% quantile of the cdf of turnover (in logarithms) is chosen, the number of take-0 enterprises in NACE sections 451, 452, 453, 454 and 455 are displayed below.

NACE section	5% quantile	take-0 enterprises ($\leq 5\%$ quantile)	enterprises in the sample
451	11.80	139	2634
452	12.20	1238	23517
453	11.97	978	18576
454	12.33	1150	21845
455	12.93	4	74

Once the sample target has been re-defined by dropping the statistical units which are out-of-scope according to an economic criterion, the definition of the strata bounds should follow. This is achieved by a modified Lavallée-Hidiroglou (1988) (HL) algorithm for log-linear and heteroscedastic linear regression relationship between the auxiliary variable and the target variable, using the Neyman allocation.

The method allows for the simultaneous determination of the minimum sample size, the strata bounds and the sample allocation which satisfy a desired statistical precision. Furthermore, the fact that the stratification variable is only a proxy for the target variable is taken into account by modeling a log-linear or heteroscedastic linear relationship between them, as suggested by Rivest (2002).

The method is fully illustrated by Rivest (2002) for the case of the sample mean survey estimator. Since the IBS output variables are totals, we derive in the following section the modifications needed for the sample total survey estimator.

Consider a stratified random sampling scheme with L strata for a variable of interest Y in a target population U of size N. Then, denoting by $N_h$the size, S_h the random sample with size n_h and $a_h = \frac{n_h}{N_h}$ the sampling fraction of stratum h, $h = 1, \ldots, L$, the survey estimator for the total $\hat{t}_{\text{ystrat}} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{k \in S_h} y_k$ has variance estimated by

$$\hat{\text{Var}}(\hat{t}_{\text{ystrat}}) \quad = \quad \sum_{h=1}^{L} N_h \frac{(1 - a_h)}{a_h} s_{yh}^2 \tag{1}$$

where

$$s_{yh}^2 = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \hat{\overline{y}}_h)^2,$$

and $\hat{\overline{y}}_h$ is the sample mean of Y within stratum h.

In the procedure we fix the L-th stratum as a take-all stratum, i.e. all the enterprises belonging to it are exhaustively sampled. As for the enterprises in the remaining $L - 1$ strata, a random sampling is performed. In other words, for the take-all stratum $n_L = N_L$, whereas for $h < L$, the sample size n_h in the take-some stratum can be written as $(n - N_L)a_h$.

Therefore, by straightforward calculations (1) can be rewritten as

$$\hat{\text{Var}}(\hat{t}_{\text{ystrat}}) \quad = \quad \frac{1}{n - N_L} \sum_{h=1}^{L-1} \frac{N_h^2 s_{yh}^2}{a_h} - \sum_{h=1}^{L-1} N_h s_{yh}^2 \tag{2}$$

from which, solving for n, we obtain

$$n_{\hat{t}_{\text{ystrat}}} \quad = \quad N_L + \frac{\sum_{h=1}^{L-1} \frac{N_h^2 s_{yh}^2}{a_h}}{\text{Var}(\hat{t}_{\text{ystrat}}) + \sum_{h=1}^{L-1} N_h s_{yh}^2}, \tag{3}$$

which is the sample size required for a given (supposed known) variance of the sample total estimator.

Expression (3) can also be written as

$$n_{\hat{t}_{\text{ystrat}}} \quad = \quad N_L + \frac{\sum_{h=1}^{L-1} \frac{W_h^2 s_{yh}^2}{a_h}}{(c\overline{Y}/N)^2 + \sum_{h=1}^{L-1} \frac{W_h}{N} s_{yh}^2}, \tag{4}$$

where $W_h = \frac{N_h}{N}$, c is the target coefficient of variation (precision level, which often ranges between 1% to 10% in business surveys) and $\overline{Y}$ the mean of Y.

The idea is to find the optimal stratum boundaries $b_1, \ldots, b_{L-1}$ which minimize $n_{\hat{t}_{\text{ystrat}}}$ subject to a requirement on the precision of $\hat{t}_{\text{ystrat}}$ such as $\text{Var}(\hat{t}_{\text{ystrat}}) = c^2\overline{Y}^2$.

Now, it is known that there exists a discrepancy between the auxiliary variable X used for stratification and the survey variable Y. Therefore, the strategy suggested by Rivest (2002) is to recover such discrepancy by the use of a regression model.

In the business survey literature, the relationship existing between Y and X is often modeled by a log-linear or a linear heteroscedastic regression model.

In what follows we consider variables X and Y as continuous random variables and we denote by $f(x), x \in \mathbb{R}$ the density of X. The data $x_1, \ldots, x_N$ are considered as N independent realizations of the random variable X.

Since stratum h consists of the population units with an X-value in the interval $(b_{h-1}, b_h]$, the stratification process uses the values of $E(Y|b_h \geq X > b_{h-1})$ and $\text{Var}(Y|b_h \geq X > b_{h-1})$, the conditional mean and variance of Y given that the unit falls in stratum h, for $h = 1, \ldots, L-1$.

6.2.1 Log-linear model

This model considers the regression relationship between Y and X expressed by

$$\log Y = \alpha + \beta_{\log} \log X + \varepsilon,$$

where ε is assumed to be a 0-mean random variable, normally distributed with variance $\sigma^2_{\log}$ and independent from X, whereas α and $\beta_{\log}$ are the parameters to be estimated.

The conditional moments of Y are obtained using the basic properties of the lognormal distribution, and they are

$$E(Y|b_h \geq X > b_{h-1}) = \exp^{\alpha+\sigma^2_{\log}/2} E(X^{\beta_{\log}}|b_h \geq X > b_{h-1}) \tag{5}$$

and

$$\text{Var}(Y|b_h \geq X > b_{h-1}) = \exp^{\alpha+\sigma^2_{\log}/2}\{e^{\sigma^2_{\log}} E(X^{2\beta_{\log}}|b_h \geq X > b_{h-1}) - E(X^{\beta_{\log}}|b_h \geq X > b_{h-1})^2\}. \tag{6}$$

6.2.2 Linear heteroscedastic model

This model is often used in the sampling literature because it allows for a non constant variance across the observations. Its classical formulation is

$$Y = \beta_{\text{lin}} X + \varepsilon$$

where ε is assumed to be a 0-mean random variable with conditional variance given by $\sigma^2_{\text{lin}} X^\gamma$, for some $\gamma > 0$. The conditional expectation and the conditional variance of Y are

$$E(Y|b_h \geq X > b_{h-1}) = \beta_{\text{lin}} E(X|b_h \geq X > b_{h-1}) \tag{7}$$

and

$$\text{Var}(Y|b_h \geq X > b_{h-1}) = \beta^2_{\text{lin}}\{\text{Var}(X|b_h \geq X > b_{h-1}) + \frac{\sigma_{\text{lin}}{}^2}{\beta_{\text{lin}}} E(X^\gamma|b_h \geq X > b_{h-1})^2\}, \tag{8}$$

respectively. Note that when $\gamma = 2$, the conditional mean and variance of Y are proportional to those for the log-linear model with $\beta_{\log} = 1$ and $\sigma^2_{\log} = (1 + (\sigma_{\text{lin}}/\beta_{\text{lin}})^2)$.

6.3 Neyman Allocation

The allocation rule considered in the IBS is the Neyman allocation rule, which is based on

$$a_h = \frac{n_h}{N_h} = \frac{W_h s_{yh}}{\sum_{k=1}^{L-1} W_k s_{yk}} \tag{9}$$

which means that the number of units sampled in a stratum is proportional to the relative dispersion of the survey variable within the stratum compared to the overall dispersion, all being ponderated by the relative weight of the stratum.

If Neyman allocation rule is applied to (4), the following holds

$$n_{\hat{t}_{\text{ystrat}}} \quad = \quad N_L + \frac{(\sum_{h=1}^{L-1} W_h s_{yh})^2}{(c\overline{Y}/N)^2 + \sum_{h=1}^{L-1} \frac{W_h}{N} s_{yh}^2}. \tag{10}$$

Now, supposing that a log-linear relationship exists between the survey variable Y and the auxiliary one X, writing the first two conditional moments of Y given $b_{h-1} < X \leq b_h$ in terms of

$$W_h = \int_{b_{h-1}}^{b_h} f(x)dx, \quad \phi_h = \int_{b_{h-1}}^{b_h} x^{\beta} f(x)dx, \quad \text{and} \quad \psi_h = \int_{b_{h-1}}^{b_h} x^{2\beta} f(x)dx,$$

where β and σ are the parameters of the log-linear model in the previous section, expression (10) can be rewritten as

$$n_{\hat{t}_{\text{ystrat}}} \quad = \quad N_L + \frac{(\sum_{h=1}^{L-1} (e^{\sigma^2} \psi_h W_h - \phi_h^2)^{1/2})^2}{(c \sum x_i^{\beta}/N)^2 + \sum_{h=1}^{L-1} \frac{(e^{\sigma^2} \psi_h - \phi_h^2/W_h)}{N}}. \tag{11}$$

6.4 Robust stratification

The presence of outliers can strongly bias the sampling design described above. In particular, it could induce a wrong computation of the number of statistical units to sample, usually overestimating it.

For instance, suppose that in the stratification variable X some outliers arise. Outliers are observations arbitrarily far from the majority of the data. They are often due to mistakes, like editing, measurement and observational errors. Intuitively, when outliers are present in a given stratum for the stratification variable X they affect both the location and scale measures for X. Therefore, it is clear that a higher dispersion than the 'true' one will be observed in that stratum.

Such a situation will bias the outcome of the HL method. For instance, the sample size would be bigger than it should be, given the fact that observations seem to be more distant (in average) than they are in the reality. Moreover, the strata bounds and the sample allocation would be both biased. This is clear when we consider the Neyman allocation, for example, which is based on within-stratum dispersion. Since the principle is to survey more units in the strata in which the auxiliary variable is more dispersed within the stratum, outliers might have the effect of increasing enormously and unduly the sample size in each stratum.

For this reason we build a robust version of the HL method, applying the following strategy. In the first step, robust estimators are used to estimate the parameters of both the log-linear and heteroscedastic regression models (S-estimator for regression is the choice for IBS). Then, the

estimated robust parameters are plugged in the HL objective function used for the computation of the strata bounds, the minimum sample size and the sample allocation which satisfy a fixed statistical precision.

6.5 An example: the sector of Construction

For illustration purposes Rivest (2002)' algorithm presented above is applied to the sector of construction with 3-digits NACE detail for the group of variables concerning production and activity of businesses. In the exercise, the target variable considered is the value added (source: NBB) and the stratification variables is the turnover (source: VAT). The universe used is the Structural Business Survey (SBS)'s universe in 2005.

The NACE detail chosen for the exercise is the 3-digit one, which for construction is described by the following subsections

- 451: site preparation (demolition and wrecking of buildings, Test drilling and boring);
- 452: building of complete constructions or parts thereof; civil engineering (general construction of buildings and civil engineering works, erection of roof covering and frames, construction of motorways, roads, airfields and sport facilities, construction of water projects, other construction worked involving special trades);
- 453: building installation (installation of electrical wiring and fittings, insulation work activities, plumbing, other building installation);
- 454: building completion (plastering, joinery installation, floor and wall covering, painting and glazing, other building completion);
- 455: renting of construction or demolition equipment with operator.

In the first step of the procedure, the parameters σ^2 and β of the log-linear model are estimated twice: once by means of the classical LS estimator and a second time by means of the robust S-estimator for regression for each NACE 3-digits section (sections 451, 452, 453, 454, 455)[7].

There are some differences in the regression estimates between the non-robust and the robust approach. The log-linear regression coefficients are generally underestimated in the non-robust approach for all sections of the construction sector, which is the effect of the presence of $3\% - 4\%$ outliers in each NACE-3 section. Taking into account the bias induced by outliers and weighting it out help to improve the quality of the regression fit, as a higher R^2 in the robust regressions shows.

Now, plugging in the estimated log-linear parameters in the HL algorithm with Neyman allocation, we obtain for both the classical case and robust case the stratification design, i.e. the sample size, allocation and strata bounds. The number of strata has been set to 5, one is the *take-all* stratum and the other 4 are the *take-some*.

The exhaustive stratum usually contains 2 enterprises and the most populated stratum is the third one. It is clear from the results obtained that strata bounds are not very different in the classical and robust log-linear estimates (exception of section 455). This is not the case for the sample size, which for the same precision level in the robust approach is up to the half of the

[7]The resulting estimates are available under request.

one obtained in the non-robust case. An explanation for this is because Rivest (2002)' algorithm is based on sample moments which enters mostly the computation of strata bounds, with lower impact in the sample size selection. Sample size is highly sensitive to the log-linear regression parameters, especially the standard deviation (which is normally much lower in the robust approach than in the classical one).

In the future, it is recommended to use robust sample first and second-order moments to correct also the strata bounds for possible outliers in the sample. A robustified version of Rivest's algorithm is therefore suggested.

6.6 Multiple survey variables

In what it has been seen above, stratification has been made considering as survey variable only the value added. Validity of this approach can be discussed since the stratification is of course not optimal with respect to the whole set of survey variables entering the same group.

Two approaches are proposed to improve the stratification. The first one is based on the principal component analysis (PCA) on the survey variables belonging to the group, whereas the second is more oriented to the achievement of an optimal stratification in the HL stream line.

6.6.1 Principal component approach

The aim is to extract a new variable (the first principal component) able to summarize the maximum common information contained in the group variables. The PCA technique is generally used to reduce multidimensional datasets to lower dimensions for analysis.

Mathematically speaking, PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. Such a tool can be used for dimensionality reduction in a data set by retaining those characteristics of the data thata contribute most to its variance, by keeping lower-order principal components and ignoring higher-order ones. Such low-order components often contain the "most important" aspects of the data.

To illustrate the method, we run PCA on two variables gross operating margin (*marge brute d'exploitation* 70/61) and operating income (*benefice d'exercice*, code 70/67) for the NACE 3-digits construction sections (source: Annual Balance Sheets of enterprises, Centrale des Bilans, NBB). Only the first principal component is retained.

Then, for each NACE section $451-455$ the linear heteroscedastic regression model is estimated for the first principal component using turnover as regressor.

Comparing the robust and non-robust approaches, we observe that in the non-robust approach the regression coefficient is always overestimated. This is due to the detection of about 10% of outliers. In general, the regression fit is better in the robust model, as the higher R^2 shows.

The heteroscedasticity is modeled by a value of $\gamma = 2$, which allows for conditional moments proportional to the log-linear model. Therefore, with the appropriated choice of $\beta_{\text{log}} = 1$ and σ_{log} as function of β_{lin} and σ_{lin}, the modified HL algorithm is applied to the heteroscedastic linear model.

Comparing the non-robust versus the robust sampling scheme, we remark the same general features as those noticed in the single survey variable case, such as an important reduction of the

sample size in the robust design, almost the same strata bounds with a very few populated *take-all* stratum.

Making comparisons with the sampling schemes obtained in the single survey variable case, we observe that for sections $452 - 455$ in the multiple survey variable case with the principal component approach the sample sizes are lower than in the single variable one. Moreover, there are some differences in the strata bounds computed, which is more accentuated in sections $453 - 455$.

6.6.2 Multiple survey optimal stratification

The main drawback of the procedure presented above is that it is not possible to say something about the precision of each survey variable estimates belonging to the group for which PCA is performed. This is due to the fact that optimality is achieved with respect to the first principal component.

Therefore, we might want to find a way to keep trace of the optimal stratification of each variable entering the group and to look for the optimal stratification in the multiple setting.

One possible approach is to base the optimality criterion on a notion of relative efficiency as a weighted average of the ratio between the variance of the survey estimator in the multiple variable scheme and the variance in the optimal stratification-single variable case.

In symbols, considering a group of Q survey variables, the loss function for the estimator of totals which we want to minimize is

$$\sum_{q=1}^{Q} H_q \frac{\text{Var}(\hat{t}_q)}{\text{Var}(\hat{t}_q^*)},$$

where $\text{Var}(\hat{t}_q^*)$ is the variance of the estimator of the total for variable q in the optimal stratification-single variables case (for instance, the stratification obtained by the HL algorithm), whereas $\text{Var}(\hat{t}_q)$ is the variance of the estimator of the total for survey variable q in a general sampling plan. The weights H_q, $q = 1, \ldots, Q$ sums to 1 and take into account the relative importance of the survey variables entering the group. We call this loss function the *Sampling Relative Efficiency*, or in short SRE.

The same function has taken different names in the literature. For example, Holmberg (2003) calls it "*Anticipated Overall Relative Efficiency Loss*", or *ANOREL*.

Given the knowledge of $V_q = \text{Var}(\hat{t}_q^*)$, and replacing the $\text{Var}(\hat{t}_q)$ by its usual formula, the SRE loss function can be rewritten as

$$SRE_Q(\hat{t}_{ystrat}) = \sum_{q=1}^{Q} \frac{H_q}{V_q} \sum_{h=1}^{L} N_h (\frac{N_h}{n_h} - 1) s_{y_q h}^2 \tag{12}$$

where $s_{y_q h}^2$ is the sampling variance for variable Y_q computed in stratum h.

Setting L as the *take-all* stratum, so that $n_L = N_L$, the sample size n_h for $h < L$ can be written as $(n - N_L)a_h$, with the usual notation introduced above. Then, solving (12) for n, after straightforward calculations we obtain

$$n_{\hat{t}_{\text{ystrat}};Q} = N_L + \frac{\sum_{q=1}^{Q} \frac{H_q}{V_q} \sum_{h=1}^{L-1} \frac{N_h^2 s_{y_q h}^2}{a_h}}{\text{SRE}_Q(\hat{t}_{\text{ystrat}}) + \sum_{q=1}^{Q} \frac{H_q}{V_q} \sum_{h=1}^{L-1} N_h s_{y_q h}^2}. \tag{13}$$

Now, considering the Neyman allocation rule in (9), expression (13) becomes

$$n_{\hat{t}_{\text{ystrat};Q}} = N_L + \frac{\sum_{q=1}^{Q} \frac{H_q}{V_q} (\sum_{h=1}^{L-1} W_h s_{y_q h})^2}{\text{SRE}_Q(\hat{t}_{\text{ystrat}})/N^2 + \sum_{q=1}^{Q} \frac{H_q}{V_q} \sum_{h=1}^{L-1} \frac{W_h s_{y_q h}^2}{N}}. \tag{14}$$

The above equation means that, fixing the precision for each of the $q = 1, \ldots, Q$ variables estimates, the overall sampling relative efficiency loss SRE, together with the relative importance H_q of each $q = 1, \ldots, Q$ survey variable, it is possible to find the minimum sample size, the strata bounds and the optimal allocation (in the sense of Neyman) which satisfy the given precision requirements.

In practice, such a procedure cannot be apply since the survey variables Y_q, $q = 1, \ldots, Q$ are not known.

Therefore, the use of some auxiliary information X is necessary. Given the discrepancy existing between the stratification variable(s) X_q and the Y_q, $q = 1, \ldots, Q$, we could model it by a log-linear regression relationship.

Then, with the notation already introduced in the previous subsections for the log-linear model setting, we want to find the strata bounds b_h, $h = 1, \ldots, L-1$ which minimize the sample size $n_{\hat{t}_{\text{ystrat};Q}}$ for a given overall and relative precision, allowing for Neymann allocation with auxiliary information linked to the Q survey variables by a log-linear regression relationship.

Mathematically speaking, we need to solve for b_h, $h = 1, \ldots, L-1$ the following equation

$$\sum_{q=1}^{Q} \frac{H_q}{V_q} \left((\frac{\partial n}{\partial \psi_{q,h}} - \frac{\partial n}{\partial \psi_{q,h+1}}) b_h^{2\beta_q} + (\frac{\partial n}{\partial \phi_{q,h}} - \frac{\partial n}{\partial \phi_{q,h+1}}) b_h^{\beta_q} \right) + (\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}}), \tag{15}$$

where

$$W_h = \int_{b_{h-1}}^{b_h} f(x)dx, \quad \phi_{q,h} = \int_{b_{h-1}}^{b_h} x^{\beta_q} f(x)dx, \quad \text{and} \quad \psi_{q,h} = \int_{b_{h-1}}^{b_h} x^{2\beta_q} f(x)dx,$$

with β_q the log-linear regression coefficient in

$$\log Y_q = \alpha_q + \beta_{q,\log} \log X_q + \varepsilon_q, \quad q = 1, \ldots, Q.$$

Unfortunately, we are not able to derive an expression in closed form for equation (15). The b_h $h = 1, \ldots, L-1$ which solve equation (15) can be obtained using numerical methods, like Newton-Raphson, Hill climbing, etc.

Note that this procedure for determining the optimal sample size, strata bounds and sample allocation in the multiple survey variable situation can be used also to simplify the IBS structure, considering only one big group of survey variables for each sector (at the place of I groups for each sector). Therefore, only J modules would be kept (one per sector) at the place of $I \times J$ modules, providing one stratification scheme for each sector $j = 1, \ldots, J$ and avoiding complicated sample coordination.

6.7 Multi-level precision

One main concern about the overall procedure is the global precision attained at the sector level for all the modules put together and the precision for each module at the national level. To answer to such questions we need to put some sample size constraints depending on the desired precision at the aggregate level.

For instance, given the matrix $\mathbf{n}$ of sample sizes for each group of survey variables $i = 1, \ldots, I$ and for each sector $j = 1, \ldots, J$, which is obtained by the modified HL algorithm with Neyman allocation choosing the $I \times J$ matrix $\mathbf{c}$ of precisions

$$\mathbf{n} = \begin{pmatrix} n_1^1 & \ldots & n_1^j & \ldots & n_1^J \\ n_2^1 & \ldots & n_2^j & \ldots & n_2^J \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ n_i^1 & \ldots & n_i^j & \ldots & n_i^J \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ n_I^1 & \ldots & n_I^j & \ldots & n_I^J \end{pmatrix} \qquad \mathbf{c} = \begin{pmatrix} c_1^1 & \ldots & c_1^j & \ldots & c_1^J \\ c_2^1 & \ldots & c_2^j & \ldots & c_2^J \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_i^1 & \ldots & c_i^j & \ldots & c_i^J \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_I^1 & \ldots & c_I^j & \ldots & c_I^J \end{pmatrix} \tag{16}$$

we might require that for a fixed *global sector* precision $c_\cdot^j$ across all survey variables (which entails a minimum size requirement $n_\cdot^j$)

$$\sum_{i=1}^{I} n_i^j \geq n_\cdot^j \quad j = 1, \ldots, J. \tag{17}$$

Also, one might want to guarantee a precision c_i of the estimates for a group i of survey variables at the national level, which would implies a minimum size requirement n_i. Therefore the following constraint should be imposed

$$\sum_{j=1}^{J} n_i^j \geq n_i \quad i = 1, \ldots, I. \tag{18}$$

In the example of Construction we have the simple case in which $I = 1$ and $J = 5$ (for illustrative purposes we consider the $3-$digits NACE sections $451 - 455$ as different sectors of activity), therefore if the constraint of type (18) is applied requiring a precision of 5% at national level for the group of survey variables concerning production ($i = 1$), the following needs to be satisfied

$$\sum_{j=1}^{5} n_1^j \geq 499.$$

In Table 7 is displayed the outcome of the modified HL algorithm for the overall sector of construction.

Now, summing the sample sizes obtained by robust HL stratification at 5% precision over the NACE 3-digits sections $(451 - 455)$, the total sample size obtained is of 767 businesses. In this case, the constraint in (18) is satisfied.

6.8 Empirical comparison of stratification designs

To judge about the quality of the stratification design, some ANOVA type analysis are performed in the sector of construction, for each subsection $451 - 455$. This is done for evaluating how well the variance of the stratification variable, which is the turnover in our example, is explained by the factors used for grouping enterprises into strata, such as the kind of economic activity (NACE), the geography and the economic size. Therefore, in the following analysis the dependent variable is the turnover and the factors are the 5-digits NACE economic activity section, the geography at municipal level (or at regional level whenever municipal is not computationally feasible) and the size class to which enterprises belong.

Three different economic criteria which define the size classes are considered in the analysis.

The first one is the size class criterion used currently in the Structural Business Survey (SBS) (results displayed in Tables 1.1-1.5), which distinguishes 6 size classes according to the turnover (VAT declaration) and the ONSS size following the scheme

- class 5: enterprises with ONSS class > 4;
- class 4: enterprises not belonging to class 5 and with ONSS class $= 4$ or turnover > 200 millions of euro;
- class 3: enterprises not belonging to class 4 and with ONSS class $= 3$ or turnover between 100 and 200 millions of euro;
- class 2: enterprises not belonging to class 3 and with ONSS class $= 2$ or turnover between 50 and 100 millions of euro;
- class 1: enterprises not belonging to class 2 and with ONSS class $= 1$ or turnover between 20 and 50 millions of euro;
- class 0: enterprises with ONSS class $= 0$ and turnover < 20 millions of euro.

The second size class criterion is based on the modified HL stratification method, as presented in the previous paragraphs for which the value added is the survey variable and the turnover is the stratification variable. The discrepancies existing between the two variables is modeled by the *robust* log-linear regression equation. Results of the ANOVA analysis are reported in Tables 2.1-2.5.

The third size class criterion is also based on the modified HL stratification method, but this time the target variable is the first principal component for the gross operating margin and the operating income, with the turnover as stratification variable. This time the model chosen is the linear heteroscedastic one, with robust S-estimator for regression parameters. Results are displayed in Tables 3.1-3.5.

Comparing the results, we observe that in general the modified HL stratification method performs better than the ad-hoc stratification currently in use for the SBS. This is shown by the higher R^2 which are always above the 87% in the optimal stratification, and by the significance of factors like the class of economic activity (NACE) and geography (at municipal or regional level) and their interactions with the economic size class. Indeed, in the SBS scenario, geography, NACE and interactions are not significant in all 3-digits NACE Construction subsections.

The optimal stratification achieved by both single-variable and multiple-variable approaches (log-linear and heteroscedastic linear models) produces comparable results in terms of analysis of variance of turnover. They show evidence of a stratification design which is far better than the ad-hoc one used in SBS.

7 Panel: the good compromise between reduction of statistical burden and high data quality

A panel survey is a survey in which similar measurements are made on the same sample at different points in time.

From a statistician's point of view, the essence of a panel survey is that repeated measurement on a single sample provides the following advantages:

- reduced sampling variation in the measurement of change;
- the ability to analyse behaviour in general, and change in particular, at the individual respondent level;
- the ability to 'train' respondents to perform relatively complex tasks within the data collection process, such as completing a special diary (in this respect some general information-training sessions on web-surveys/questionnaires... could be organized by Statistics Belgium for businesses);
- the ability to collect a greater range of data than is possible at a single contact (a follow-up of the business statistics through time);
- the spreading of costs over a long time period and a potentially large user base.

Against these advantages must be set some features of panel research which act against accuracy in estimation:

- the initial cooperation rate is lower than for a single contact;
- the sample undergoes over attrition over time (i.e. increasing levels of non-response with each successive 'wave' of the panel);
- there may be response conditioning and/or behavioural conditioning.

A way to compensate and minimize these factors is to use rotating panels.

7.1 Rotating panels and *statistical holidays*

In a rotating panel survey, part of the sample is changed each period (for example every year). Panel with a rotational design aims to be the most cost effective and efficient for both satisfying the cross-sectional and longitudinal requirements. The use of a rotational panel allows introduction in the sample of new firms subgroups each year, and as a result the cross-sectional data derived from this design will have a wider representativity than the data derived from a pure panel.

The use of rotating panels would allow for

- increase in the statistical accuracy and data quality;
- lower burden over businesses which would be in the sample for a limited number of years;
- potential positive attitude of enterprises to cooperate with Statistics Belgium on a short period (possible training sessions, contacts...).

A rotating panel of SMEs over a 9-years horizon is an example of panel horizon. A given rotating group is interviewed for 3 years, it leaves the sample during the following 6 years, and then returns for other 3 consecutive years. Under this system $2/3$ of the sample is common from year to year. This procedure provides a substantial amount of year-to-year overlap in the sample, thus providing better estimates of change and reducing discontinuities in the data series without burdening any specific group of firms with an unduly long period of inquiry.

The period during which an enterprise does not enter the sample is called a *statistical holidays* period for the enterprise.

Intuitively, the lower the sample frequency in a given stratum, the higher the number of statistical holidays which the enterprises belonging to that stratum can receive. Therefore, to reduce the statistical burdens for enterprises allowing for longer statistical holidays, it is very important to build a good stratification. Indeed, when the stratification is done using variance optimizing methods like the one seen in the previous sections, the advantage is to obtain more homogeneous strata in terms of within-stratum variability and more heterogeneity between-strata with respect to the survey variables. This allows for lower sample frequencies than general ad-hoc stratification criteria (like for instance the one used in SBS).

The intuition given is confirmed by the following formula which could be used for determining the maximum number of periods of statistical holidays (vac_h) for enterprises belonging to stratum h with sampling frequency f_h, when the rotation period during which the enterprise stays in the sample is set to $p_{r,h}$.

$$\text{vac}_h \leq \lfloor (\frac{1}{f_h} - 1) p_{r,h} \rfloor, \quad h = 1, \ldots, L. \tag{19}$$

This inequality can be easily rewritten as

$$f_h \leq \frac{p_{r,h}}{p_{r,h} + \text{vac}_h} \quad h = 1, \ldots, L \tag{20}$$

and it represents the upper bound for the sampling frequency of stratum h which guarantees the overall panel horizon $p_{r,h} + \text{vac}_h$ with rotation period $p_{r,h}$.

For instance, in the case of a rotating panel of a 9-years horizon, in which the rotation period $p_{r,h} = 3$ and the statistical holidays last $\text{vac}_h = 6$ years, the sampling fraction in each stratum required for this kind of rotation must not be greater than $1/3$.

8 Conclusions and Agenda

In this document two main aspects of survey integration have been considered: the operational structure of IBS and the sampling design.

The advantages of integration for both businesses and the NIS have been pointed out throughout the paper. It is clear that the integration process involves also some operational costs for the Administration, which need to be evaluated according to the survey domain.

The description of a possible structure for IBS is given in the paper according to the *variable-oriented* approach, presented in Section 2. The structure of IBS is the byproduct of several empirical evidences. The importance of the sectors of activity and of the geographical location of businesses, joint to the variable-oriented approach lead to a modular structure of the IBS, in which 5 main clusters of variables grouped by themes are crossed with the 4 main sectors of activity, i.e. industry, construction, trade and services.

From such a structure, it is evident that integration goes further beyond a mere coordination of surveys, but it involves a complete new organization and clustering of surveys.

Sampling is performed using a 3-dimensional stratification by NACE, geography and economic size-class. Take-all, take-some and take-0 strata are considered. Thresholds for the take-0 strata are obtained by selecting the $\%5$ quantile of the cumulative distribution function of the annual turnover (VAT source). Thresholds defining the economic size strata are obtained by the generalized HL algorithm (see Rivest, 2002) applied on annual turnover (source VAT) as auxiliary variable. Sample frequencies are obtained by Neyman allocation.

Robustness issues are considered to cope with errors frequently present in the VAT data.

A new method for constructing optimal sampling plans with respect to a *group* of survey variables is proposed. Such sampling plan is optimal in terms of minimizing the sample size required for a fixed relative precision level. The notion of Sampling Relative Efficiency is introduced and it plays the role of objective function in the optimization problem.

An empirical application to the sector of Construction illustrates the performance of the sampling strategy hereby proposed, showing clearly many advantages compared to the *ad-hoc* stratification currently used in surveys like SBS.

Further work is required for the numerical implementation of the multiple variable survey optimal stratification. The future steps in the project are also the construction of a criterion which selects a good combination of survey modules to submit to enterprises without burdening them too much and to retrieve a maximum of complete and coherent information at the same time.

For instance, one could measure the existing overlap between survey modules in terms of the universe and apply those rates to the sample selected in each stratum.

Other issues like the choice of the sampling algorithm and the implementation of a DEMO version of 6 surveys are also in the agenda for future work.

9 References

Aucremanne, L., Druant, M. (2005). Price-setting behaviour in Belgium: what can be learned from an ad hoc survey? *ECB Working Paper Series* n. **448**.

Butzen, P., Fuss, C. (2001). The interest rate and credit channels in Belgium: an investigation with micro-level firm data. *ECB Working Paper Series* n. **107**.

Lallemand, T., Plasman, R., Rycx, F. (2007). Wage structure and firm productivity in Belgium. *NBER Working Paper Series* n. **12978**.

Holmberg, A. (2003). Utilisation des données auxiliaires pour déterminer le plan d'échantillonage dans une enquête comportant plusieurs variables clés. *Acts Symposium Statistics Canada 2003* n. **11-522-XIF**.

Rivest, L.P. (2002). A generalization of Lavallée and Hidiroglou algorithm for stratification in business surveys. *Techniques d'enquêtes* **28**, 207-214.

Teirlinck, P., Spithoven, A. (2005). Spatial inequality and location of private R&D activities in Belgian districts. *Royal Dutch Geographical Society* **96**, 558-572.

Tillé, Y. (2001). *Théorie des sondages*. Paris, Dunod.

Van Gastel, G. (1999). Small and medium sized enterprises in business tendency surveys: some critical remarks. *National Bank of Belgium*. Internal discussion paper.

Ward, J.H.J. (1963). Hierarchical grouping to optimize an objective function. *Journal of The American Statistical Association* **58**, 236-244.

Williams, W.T., Lance, G.N. (1965). Computer programs for monothetic classification. *The Computer Journal* **8**, 246-249.

Table 1.1

Stratification using SBS size criterion on turnover

SBS size classes 0-5

NACE 451

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	707	7268.33532	10.28053	6.54	<.0001
Error	2047	3216.40252	1.57128		
Corrected Total	2754	10484.73784			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.693230	8.383144	1.253506	14.95269

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	2	66.984917	33.492458	21.32	<.0001
size_class_ese	2	5871.799068	2935.899534	1868.48	<.0001
commune	237	562.952177	2.375326	1.51	<.0001
NACE*size_class_ese	4	9.485188	2.371297	1.51	0.1969
NACE*commune	206	369.983434	1.796036	1.14	0.0894
size_class_e*commune	188	267.871438	1.424848	0.91	0.8065
NACE*size_cl*commune	68	119.259097	1.753810	1.12	0.2435

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	2	5.8483792	2.9241896	1.86	0.1558
size_class_ese	2	657.6877072	328.8438536	209.28	<.0001
commune	237	351.4347319	1.4828470	0.94	0.7145
NACE*size_class_ese	2	1.7849433	0.8924717	0.57	0.5668
NACE*commune	201	312.2864877	1.5536641	0.99	0.5315
size_class_e*commune	188	248.5465514	1.3220561	0.84	0.9375
NACE*size_cl*commune	68	119.2590966	1.7538102	1.12	0.2435

Table 1.2

Stratification using SBS size criterion on turnover

SBS size classes 0-5

NACE 452

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	916	61494.92470	67.13420	42.95	<.0001
Error	24493	38280.44095	1.56291		
Corrected Total	25409	99775.36566			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.616334	8.182649	1.250165	15.27825

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	9	6326.28245	702.92027	449.75	<.0001
size_class_ese	2	52099.59540	26049.79770	16667.5	<.0001
region	49	232.25772	4.73995	3.03	<.0001
NACE*size_class_ese	18	1152.61034	64.03391	40.97	<.0001
NACE*region	371	848.10105	2.28599	1.46	<.0001
size_class_es*region	93	220.35470	2.36941	1.52	0.0010
NACE*size_cla*region	374	615.72305	1.64632	1.05	0.2311

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	9	59.370695	6.596744	4.22	<.0001
size_class_ese	2	3855.279802	1927.639901	1233.36	<.0001
region	49	76.044730	1.551933	0.99	0.4871
NACE*size_class_ese	18	560.743301	31.152406	19.93	<.0001
NACE*region	370	580.357343	1.568533	1.00	0.4711
size_class_es*region	93	167.562334	1.801746	1.15	0.1494
NACE*size_cla*region	374	615.723053	1.646318	1.05	0.2311

Table 1.3

Stratification using SBS size criterion on turnover

SBS size classes 0-5

NACE 453

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	523	32719.61195	62.56140	39.78	<.0001
Error	18996	29876.04352	1.57275		
Corrected Total	19519	62595.65547			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.522714	8.383827	1.254095	14.95850

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	4	2438.88083	609.72021	387.68	<.0001
size_class_ese	2	28712.05322	14356.02661	9127.95	<.0001
region	49	295.03846	6.02119	3.83	<.0001
NACE*size_class_ese	8	494.54636	61.81830	39.31	<.0001
NACE*region	194	417.88332	2.15404	1.37	0.0005
size_class_es*region	75	151.71250	2.02283	1.29	0.0487
NACE*size_cla*region	191	209.49726	1.09684	0.70	0.9995

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	4	10.260477	2.565119	1.63	0.1633
size_class_ese	2	4593.577776	2296.788888	1460.36	<.0001
region	49	53.727858	1.096487	0.70	0.9467
NACE*size_class_ese	8	197.067548	24.633443	15.66	<.0001
NACE*region	194	263.956925	1.360603	0.87	0.9120
size_class_es*region	75	91.194831	1.215931	0.77	0.9269
NACE*size_cla*region	191	209.497260	1.096844	0.70	0.9995

Table 1.4

Stratification using SBS size criterion on turnover

SBS size classes 0-5

NACE 454

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	842	29841.05773	35.44069	28.30	<.0001
Error	22222	27830.39877	1.25238		
Corrected Total	23064	57671.45650			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.517432	7.465446	1.119098	14.99037

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	8	1675.32042	209.41505	167.21	<.0001
size_class_ese	2	26642.44502	13321.22251	10636.7	<.0001
region	49	213.00006	4.34694	3.47	<.0001
NACE*size_class_ese	14	166.78985	11.91356	9.51	<.0001
NACE*region	383	666.15607	1.73931	1.39	<.0001
size_class_es*region	63	125.76614	1.99629	1.59	0.0019
NACE*size_cla*region	323	351.58018	1.08848	0.87	0.9558

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	8	13.570804	1.696351	1.35	0.2112
size_class_ese	2	3953.911301	1976.955650	1578.56	<.0001
region	49	51.682415	1.054743	0.84	0.7756
NACE*size_class_ese	14	72.452057	5.175147	4.13	<.0001
NACE*region	383	400.730067	1.046293	0.84	0.9912
size_class_es*region	63	81.356154	1.291368	1.03	0.4084
NACE*size_cla*region	323	351.580180	1.088484	0.87	0.9558

Table 1.5

Stratification using SBS size criterion on turnover

SBS size classes 0-5

NACE 455

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	56	367.2001050	6.5571447	6.89	<.0001
Error	16	15.2259903	0.9516244		
Corrected Total	72	382.4260953			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.960186	5.780343	0.975512	16.87638

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	0	0.0000000	.	.	.
size_class_ese	2	279.3874752	139.6937376	146.80	<.0001
commune	46	69.8224845	1.5178801	1.60	0.1547
NACE*size_class_ese	0	0.0000000	.	.	.
NACE*commune	0	0.0000000	.	.	.
size_class_e*commune	8	17.9901453	2.2487682	2.36	0.0681
NACE*size_cl*commune	0	0.0000000	.	.	.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	0	0.00000000	.	.	.
size_class_ese	2	61.89191403	30.94595702	32.52	<.0001
commune	46	72.31472958	1.57205934	1.65	0.1369
NACE*size_class_ese	0	0.00000000	.	.	.
NACE*commune	0	0.00000000	.	.	.
size_class_e*commune	8	17.99014530	2.24876816	2.36	0.0681
NACE*size_cl*commune	0	0.00000000	.	.	.

Table 2.1

Stratification for value added using HL/Neymann on turnover

Classes 1-5

NACE 451

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1002	9795.06775	9.77552	24.83	<.0001
Error	1752	689.67009	0.39365		
Corrected Total	2754	10484.73784			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.934222	4.195989	0.627413	14.95269

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	2	66.984917	33.492458	85.08	<.0001
NO_STRATE	4	9189.010991	2297.252748	5835.81	<.0001
commune	237	109.910510	0.463757	1.18	0.0415
NACE*NO_STRATE	7	17.217399	2.459628	6.25	<.0001
NACE*commune	206	95.320392	0.462720	1.18	0.0535
NO_STRATE*commune	429	242.929511	0.566269	1.44	<.0001
NACE*NO_STRA*commune	117	73.694028	0.629863	1.60	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	2	2.432369	1.216184	3.09	0.0458
NO_STRATE	4	1331.383149	332.845787	845.54	<.0001
commune	237	110.664124	0.466937	1.19	0.0354
NACE*NO_STRATE	6	8.641465	1.440244	3.66	0.0013
NACE*commune	185	88.515594	0.478463	1.22	0.0314
NO_STRATE*commune	429	235.458387	0.548854	1.39	<.0001
NACE*NO_STRA*commune	117	73.694028	0.629863	1.60	<.0001

Table 2.2

Stratification for value added using HL/Neymann on turnover

Classes 1-5

NACE 452

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1308	87910.40488	67.20979	136.52	<.0001
Error	24101	11864.96078	0.49230		
Corrected Total	25409	99775.36566			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.881083	4.592425	0.701642	15.27825

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	9	6326.28245	702.92027	1427.82	<.0001
NO_STRATE	4	79672.22243	19918.05561	40459.1	<.0001
region	49	41.86486	0.85438	1.74	0.0011
NACE*NO_STRATE	28	593.89481	21.21053	43.08	<.0001
NACE*region	371	366.95218	0.98909	2.01	<.0001
NO_STRATE*region	147	168.72962	1.14782	2.33	<.0001
NACE*NO_STRAT*region	700	740.45854	1.05780	2.15	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	9	67.788785	7.532087	15.30	<.0001
NO_STRATE	4	5000.771376	1250.192844	2539.49	<.0001
region	49	35.178428	0.717927	1.46	0.0199
NACE*NO_STRATE	27	241.924162	8.960154	18.20	<.0001
NACE*region	371	393.380212	1.060324	2.15	<.0001
NO_STRATE*region	147	126.428193	0.860056	1.75	<.0001
NACE*NO_STRAT*region	700	740.458537	1.057798	2.15	<.0001

Table 2.3

Stratification for value added using HL/Neymann on turnover

Classes 1-5

NACE 453

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	895	54129.30818	60.47967	133.04	<.0001
Error	18624	8466.34729	0.45459		
Corrected Total	19519	62595.65547			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.864745	4.507372	0.674235	14.95850

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	4	2438.88083	609.72021	1341.24	<.0001
NO_STRATE	4	50834.54400	12708.63600	27956.1	<.0001
region	49	42.78377	0.87314	1.92	0.0001
NACE*NO_STRATE	12	142.43254	11.86938	26.11	<.0001
NACE*region	194	112.68614	0.58086	1.28	0.0056
NO_STRATE*region	149	202.34042	1.35799	2.99	<.0001
NACE*NO_STRAT*region	483	355.64048	0.73632	1.62	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	4	10.314569	2.578642	5.67	0.0001
NO_STRATE	4	8752.179058	2188.044764	4813.19	<.0001
region	49	54.734555	1.117032	2.46	<.0001
NACE*NO_STRATE	12	48.606969	4.050581	8.91	<.0001
NACE*region	194	184.389542	0.950462	2.09	<.0001
NO_STRATE*region	149	154.820867	1.039066	2.29	<.0001
NACE*NO_STRAT*region	483	355.640479	0.736316	1.62	<.0001

Table 2.4

Stratification for value added using HL/Neymann on turnover

Classes 1-5

NACE 454

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1535	50411.59407	32.84143	97.39	<.0001
Error	21529	7259.86242	0.33721		
Corrected Total	23064	57671.45650			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.874117	3.873824	0.580701	14.99037

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	8	1675.32042	209.41505	621.02	<.0001
NO_STRATE	4	47807.84923	11951.96231	35443.3	<.0001
region	49	28.97706	0.59137	1.75	0.0009
NACE*NO_STRATE	25	92.97761	3.71910	11.03	<.0001
NACE*region	383	167.48245	0.43729	1.30	<.0001
NO_STRATE*region	147	104.80985	0.71299	2.11	<.0001
NACE*NO_STRAT*region	919	534.17746	0.58126	1.72	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	8	8.971896	1.121487	3.33	0.0008
NO_STRATE	4	7883.564876	1970.891219	5844.64	<.0001
region	49	38.504614	0.785808	2.33	<.0001
NACE*NO_STRATE	24	53.638644	2.234943	6.63	<.0001
NACE*region	383	244.845326	0.639283	1.90	<.0001
NO_STRATE*region	147	84.970886	0.578033	1.71	<.0001
NACE*NO_STRAT*region	919	534.177455	0.581259	1.72	<.0001

Table 2.5

Stratification for value added using HL/Neymann on turnover

Classes 1-5

NACE 455

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	61	374.8898617	6.1457354	8.97	0.0002
Error	11	7.5362336	0.6851121		
Corrected Total	72	382.4260953			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.980294	4.904578	0.827715	16.87638

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	0	0.0000000	.	.	.
NO_STRATE	4	345.5556948	86.3889237	126.09	<.0001
commune	46	25.2853059	0.5496806	0.80	0.7146
NACE*NO_STRATE	0	0.0000000	.	.	.
NACE*commune	0	0.0000000	.	.	.
NO_STRATE*commune	11	4.0488611	0.3680783	0.54	0.8412
NACE*NO_STRA*commune	0	0.0000000	.	.	.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	0	0.00000000	.	.	.
NO_STRATE	4	87.38179596	21.84544899	31.89	<.0001
commune	46	25.55134227	0.55546396	0.81	0.7068
NACE*NO_STRATE	0	0.00000000	.	.	.
NACE*commune	0	0.00000000	.	.	.
NO_STRATE*commune	11	4.04886105	0.36807828	0.54	0.8412
NACE*NO_STRA*commune	0	0.00000000	.	.	.

Table 3.1

Stratification for the First principal component using HL/Neymann on turnover

Classes 1-5

NACE 451

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1002	9795.01163	9.77546	24.83	<.0001
Error	1752	689.72620	0.39368		
Corrected Total	2754	10484.73784			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.934216	4.196159	0.627439	14.95269

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	2	66.984917	33.492458	85.08	<.0001
NO_STRATE	4	9188.840530	2297.210133	5835.23	<.0001
commune	237	109.447584	0.461804	1.17	0.0457
NACE*NO_STRATE	7	16.788950	2.398421	6.09	<.0001
NACE*commune	206	95.132259	0.461807	1.17	0.0557
NO_STRATE*commune	429	244.124666	0.569055	1.45	<.0001
NACE*NO_STRA*commune	117	73.692728	0.629852	1.60	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	2	2.435919	1.217959	3.09	0.0456
NO_STRATE	4	1331.806495	332.951624	845.74	<.0001
commune	237	109.805571	0.463315	1.18	0.0425
NACE*NO_STRATE	6	8.632801	1.438800	3.65	0.0013
NACE*commune	185	88.497270	0.478364	1.22	0.0316
NO_STRATE*commune	429	236.707652	0.551766	1.40	<.0001
NACE*NO_STRA*commune	117	73.692728	0.629852	1.60	<.0001

Table 3.2

Stratification for the First principal component using HL/Neymann on turnover

Classes 1-5

NACE 452

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1306	87908.01210	67.31088	136.71	<.0001
Error	24103	11867.35355	0.49236		
Corrected Total	25409	99775.36566			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.881059	4.592698	0.701684	15.27825

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	9	6326.28245	702.92027	1427.66	<.0001
NO_STRATE	4	79662.84847	19915.71212	40449.5	<.0001
region	49	41.74334	0.85190	1.73	0.0012
NACE*NO_STRATE	28	591.81425	21.13622	42.93	<.0001
NACE*region	371	368.99734	0.99460	2.02	<.0001
NO_STRATE*region	147	167.13639	1.13698	2.31	<.0001
NACE*NO_STRAT*region	698	749.18986	1.07334	2.18	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	9	65.379261	7.264362	14.75	<.0001
NO_STRATE	4	5011.433458	1252.858365	2544.60	<.0001
region	49	35.579335	0.726109	1.47	0.0171
NACE*NO_STRATE	27	245.065291	9.076492	18.43	<.0001
NACE*region	371	398.319043	1.073636	2.18	<.0001
NO_STRATE*region	147	125.261828	0.852121	1.73	<.0001
NACE*NO_STRAT*region	698	749.189864	1.073338	2.18	<.0001

Table 3.3

Stratification for the First principal component using HL/Neymann on turnover

Classes 1-5

NACE 453

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	898	54071.72745	60.21350	131.54	<.0001
Error	18621	8523.92802	0.45776		
Corrected Total	19519	62595.65547			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.863826	4.523038	0.676579	14.95850

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	4	2438.88083	609.72021	1331.97	<.0001
NO_STRATE	4	50771.74219	12692.93555	27728.4	<.0001
region	49	45.48332	0.92823	2.03	<.0001
NACE*NO_STRATE	12	144.28754	12.02396	26.27	<.0001
NACE*region	194	115.35725	0.59462	1.30	0.0034
NO_STRATE*region	149	198.80131	1.33424	2.91	<.0001
NACE*NO_STRAT*region	486	357.17501	0.73493	1.61	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	4	9.854122	2.463531	5.38	0.0003
NO_STRATE	4	8564.732305	2141.183076	4677.53	<.0001
region	49	52.761620	1.076768	2.35	<.0001
NACE*NO_STRATE	12	46.379054	3.864921	8.44	<.0001
NACE*region	194	189.545105	0.977037	2.13	<.0001
NO_STRATE*region	149	151.501017	1.016785	2.22	<.0001
NACE*NO_STRAT*region	486	357.175006	0.734928	1.61	<.0001

Table 3.4

Stratification for the First principal component using HL/Neymann on turnover

Classes 1-5

NACE 454

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1536	50344.71932	32.77651	96.31	<.0001
Error	21528	7326.73718	0.34034		
Corrected Total	23064	57671.45650			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.872957	3.891715	0.583383	14.99037

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	8	1675.32042	209.41505	615.32	<.0001
NO_STRATE	4	47744.21023	11936.05256	35071.5	<.0001
region	49	26.96385	0.55028	1.62	0.0041
NACE*NO_STRATE	25	95.00197	3.80008	11.17	<.0001
NACE*region	383	167.04417	0.43615	1.28	0.0002
NO_STRATE*region	147	102.18798	0.69516	2.04	<.0001
NACE*NO_STRAT*region	920	533.99070	0.58042	1.71	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	8	5.330320	0.666290	1.96	0.0475
NO_STRATE	4	7839.673143	1959.918286	5758.79	<.0001
region	49	36.107921	0.736896	2.17	<.0001
NACE*NO_STRATE	24	54.738042	2.280752	6.70	<.0001
NACE*region	383	242.900991	0.634206	1.86	<.0001
NO_STRATE*region	147	85.991241	0.584974	1.72	<.0001
NACE*NO_STRAT*region	920	533.990701	0.580425	1.71	<.0001

Table 3.5

Stratification for the First principal component using HL/Neymann on turnover

Classes 1-5

NACE 455

Dependent Variable: turnover

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	61	374.8898617	6.1457354	8.97	0.0002
Error	11	7.5362336	0.6851121		
Corrected Total	72	382.4260953			

R-Square	Coeff Var	Root MSE	logomzet Mean
0.980294	4.904578	0.827715	16.87638

Source	DF	Type I SS	Mean Square	F Value	Pr > F
NACE	0	0.0000000	.	.	.
NO_STRATE	4	345.9380760	86.4845190	126.23	<.0001
commune	46	24.9029246	0.5413679	0.79	0.7258
NACE*NO_STRATE	0	0.0000000	.	.	.
NACE*commune	0	0.0000000	.	.	.
NO_STRATE*commune	11	4.0488611	0.3680783	0.54	0.8412
NACE*NO_STRA*commune	0	0.0000000	.	.	.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
NACE	0	0.00000000	.	.	.
NO_STRATE	4	87.38179596	21.84544899	31.89	<.0001
commune	46	25.15236336	0.54679051	0.80	0.7185
NACE*NO_STRATE	0	0.00000000	.	.	.
NACE*commune	0	0.00000000	.	.	.
NO_STRATE*commune	11	4.04886105	0.36807828	0.54	0.8412
NACE*NO_STRA*commune	0	0.00000000	.	.	.

Informations

La Direction générale Statistique et Information économique relève du SPF Economie, PME, Classes moyennes et Energie. Une de nos missions est de répondre aux besoins des autorités, des entreprises et des citoyens par une information chiffrée sur la situation réelle du pays dans différents domaines d'actualité

Où trouver l'information statistique et économique?

Sur nos sites Internet http://statbel.fgov.be (statistiques) et http://economie.fgov.be (économie)

Dans cinq grandes villes du pays, la Direction générale Statistique et Information économique met à la disposition du public :

- ◊ Des annuaires et des publications spécialisées ainsi qu'une sélection de disquettes et de cédéroms.
- ◊ Une salle de lecture où il est possible de consulter nos publications, ainsi que celles d'autres ministères ou d'institutions belges et internationales.

Toutes nos bibliothèques sont accessibles les jours ouvrables de 8h30 à 16h30 (Bxl) ou de 9h à 12h et de 13h à 16h (autres).

Bruxelles City Atrium C

Rue du Progrès 50, 1210 Bruxelles
tél. 02/277.55.03 – 02/277.55.04 fax 02/277.55.19
e-mail : info@economie.fgov.be

Train (B) :	Gare du Nord
Métro (M) :	ligne 2, station Rogier
Trams :	3, 52, 55, 56, 81, 90 arrêts Rogier ou Nord
Bus STIB :	38, 58, 61 arrêts Rogier ou Nord
Bus De Lijn :	318, 351, 358, 410, 526, 554 arrêt Nord

Anvers

Italiëlei 124 - bus 85, 2000 Antwerpen
tél. 03/229.07.00 fax 03/233.28.30
e-mail : info.antwerpen@economie.fgov.be

Train (B) :	Centraal Station
Métro (M) :	arrêt Opera
Tram-Bus :	accès facile (Fr. Rooseveltplaats)

Charleroi

Tour Biarent, Bd Audent 14/5, 6000 Charleroi
tél. 02/277.80.37 fax 02/277.57.03
e-mail : info.charleroi@economie.fgov.be

Train (B) :	Charleroi Sud, 20 min depuis la gare (Place Buisset, Rue du Collège, Place Charles II, Boulevard Tirou, rue de la Montagne)
Bus :	arrêt Tirou
Autoroute :	petite ceinture de Charleroi - sortie Gare du Sud

Gand

Gaston Crommenlaan 6 bus 0901,9050 Gent
tél. 02/277 86 96 fax 02/277 54 06
e-mail : info.gent@economie.fgov.be

Train (B):	Gent St. Pieters
Tram-Bus :	40, 43 arrêt Theresianenstraat
Autoroute:	accès aisé par autoroute E40 (sortie N° 13 - Gent - West/Drongen)

Liège

Bd de la Sauvenière 73-75, 4000 Liège
tél. 02/277.55.78 fax 04/222.49.94
e-mail : info.liege@economie.fgov.be

Train (B):	Gare des Guillemins ou Gare du Palais
Tram-Bus :	(Guillemins) 1 et 4 arrêt Sauvenière
Parking (P):	Neujean (à 20 m - même trottoir) Mercure (en face)

Achevé d'imprimer
par l'imprimerie de la
Direction générale Statistique
et Information économique
B-1000 Bruxelles

Septembre 2008