SAMPLING METHODOLOGY FOR ECONOMIC STATISTICS

Statistics South Africa (Stats SA) has now developed a new Business Sampling Frame (BSF), based on Value Added Tax (VAT) database obtained from the South Africa Revenue Services (SARS). A new sample design approach for economic surveys at Stats SA, based on the BSF, and using the enterprise as sampling unit, has now been developed. This has been preceded by extensive discussions on the general methodology (delimitation, stratification, allocation, sample drawing and estimation) to be used in the sample design of an economic survey. All the economic surveys at Stats SA use enterprise as sampling unit. This document will serve as a guideline in sampling methodology.

i. The purposes of this document

Firstly, this is a reference document for subject matter areas in the design of samples in Economic Statistics. Survey areas use the BSF for drawing samples. This document will be used as a guideline in the process of drawing samples annually from the sampling frame. Every survey area will be responsible to draw a sample for their specific survey and consult with Quality Methodology with any uncertainties and clarifications needed.

Secondly the document will be a part of the support function that Quality and Methodology should play within Economic Statistics.

Thirdly, Stats SA is committed to quality work and quality management according to SQM principles. In this context, the document serves as a tool for standardization and stabilization of a part of the sample methodology process. Standardization and stabilization are necessary for continuous quality improvement through improvements of the sample methodologies. **Continuous process improvements** mean that this document is a "live" document and will be improved and updated to reflect the current best methods in sampling methodology.

Contents

Page

0 Summary

Statistics South Africa (Stats SA) has developed the Business Sampling Frame (BSF), which will be used as a sampling frame for all economic surveys at Stats SA.

The BSF at Stats SA contains the following types of statistical units:

1. Enterprise Units (EN);
2. Kind of Activity Units (KAU): the single activity part of an enterprise; and
3. Geographical Units (GEO): the single activity part of an enterprise that is operating at a single geographical location.

The main source of information for the BSF is the Value Added Tax (VAT) database from the South African Revenue Services (SARS). This means that the coverage for enterprises in the BSF is good, at least for enterprises with annual turnover equal to or exceeding R300 000 and belonging to the formal sector, because the VAT-data are only available on this level.

Samples are drawn for economic surveys from the BSF using the enterprise as the sampling unit, and it is possible to produce reliable estimates on a national level. All the economic surveys at Stats SA uses the enterprise as sampling unit. However, the present status of the BSF does not allow the drawing of reliable economic surveys using either the Kind of Activity Unit (KAU, i.e. the single activity part of one enterprise) or the Geographical Unit (GEO, i.e. the single activity part of one enterprise operating from only one geographical location) as sampling unit. There is under-coverage in the current Geographical (GEO) frame, which means that it is impossible to use it as a sampling frame. The enterprise could be used as the sampling unit and the GEO as the observational unit. This would mean that all selected enterprises would be required to report on all their GEO units. Thus sampling could be done on the enterprise level and reporting could be done on the geographical level (the observational units) of the enterprises, drawn in a specific sample where an observational unit is a unit for which data are reported that does not equate to the sampling unit on the sampling frame. In many instances, especially in the case of the smaller enterprises (i.e. enterprises with smaller turnovers), the enterprise is also the KAU and/or the GEO. Large enterprises are frequently also complex enterprises, i.e. enterprises with more than one KAU and/or GEO. This is especially possible because the vast majority of complex enterprises fall in size group 1 (the largest enterprises in terms of turnover within each industry) and all size group 1 enterprises are included in the sample (completely enumerated).

The use of the same frame for many surveys makes it possible to compile comparable statistics, which, inter alia, is vital for the National Accounts. Surveys used by the National Accounts to compile the Gross Domestic Product should use similar definitions of population units and compatible variables in the survey design.

The sampling design includes all the steps that should be considered in the design of an economic survey.

CHAPTER 1

BACKROUND, CONCEPTS AND DEFINITONS

1. Background

1.1 Introduction

The objective of this document is to assist all production areas in the drawing of samples from the BSF.

Stats SA has developed a Business Sampling Frame (BSF), based mainly on the Value Added Tax (VAT) database obtained from the South Africa Revenue Services (SARS). VAT-data is delivered monthly from South African Revenue Services (SARS) to Stats SA. All business units (enterprises) are legally required to register for VAT when their turnover for a period of twelve months equals or exceeds R300 000. Consequently, the coverage of enterprises in the BSF can be assumed to be adequate for sampling purposes for Stats SA's economic surveys, using the enterprise as the sampling unit to produce reliable estimates on a national level, for enterprises in the formal sector with an annual turnover of R300 000 or more.

All the economic surveys at Stats SA use the enterprise as a sampling unit. However, the present status of the BSF does not allow the drawing of reliable economic surveys using either Kind of Activity Unit (KAU, i.e. the single activity part of one enterprise) or the Geographical Unit (GEO, i.e. the single activity part of one enterprise operating from only one geographical location) as a sampling unit. In many instances, especially in the case of the smaller enterprises (i.e. enterprises with smaller turnovers), the enterprise is also the KAU and/or the GEO. Large enterprises are frequently also complex enterprises, i.e. enterprises with more than one KAU and/or GEO. The BSF may be used as a sampling frame for economic surveys with the aim of producing regional (e.g. provincial) estimates, by considering the selected enterprises as clusters of GEO's, and by requiring the selected enterprises to report on all their GEO's (i.e. considering the GEO as an observational unit), is probably a viable option for the interim period till all GEO's on the BSF is appropriately delineated.

A new sample design approach for economic surveys at Stats SA, based on the BSF, and using the enterprise as a sampling unit, has been developed. This has been preceded by extensive discussions on the general methodology (delimitation, stratification, allocation, sample drawing and estimation) to be used in the sample design of an economic survey with consultants form the Australian Bureau of Statistics and from Statistics Sweden.

The Division: System of Registers provides a snapshot of the Business Sampling Frame annually to Quality and Methodology. It will be the responsibility of every survey area to verify the quality of the sampling frame and to draw samples from.

This document will focus on the design of surveys, assesses the allocation of the total sample to the industry sector SIC code levels of interest, the determination of the turnover cut-off points to be used for subdividing the enterprises into non-overlapping groups, per industry Standard Industrial Classification (SIC) code level (primary strata), into the measure of size (MOS) intervals based on turnover, called secondary strata (viz. the cross-classification of the industry strata (primary strata) with the measure of size intervals (secondary strata)). The industry strata (primary strata) are defined as the pre-specified SIC code levels (i.e. 1-digit and/or 2-digits and/or 3-digits) of the different industry sectors to be taken into account in the design of the sample, i.e. in respect of which an "adequate" or "sufficient" representation in the sample is required. Enterprises within each primary stratum will generally be grouped into 4 non-overlapping size groups (secondary strata) based on the measure of size (MOS). The size group cut-offs will be the same for all enterprises in each primary strata. The SIC code levels of these strata (primary strata) vary between different economic surveys, depending on various factors, such as the purpose of the survey and the size of the sample. The number of (measure of) size intervals or groups per industry primary stratum is usually taken as a value between 4 and 6. However, Stats SA has decided to use 4 size groups per industry primary stratum in its economic surveys, constructed in such a way that size group 1 contains the largest enterprises in that stratum. All enterprises in these first size groups of the different strata are included in the sample.

This document also assesses the calculation of the (theoretical) relative standard error (RSE) values per industry sector SIC code level of interest, as well as for the total economy (using turnover as study variable), the method or procedure to be used in the drawing of the samples and sub-samples and the estimation of population parameters or characteristics as well as their standard errors from the sample values obtained. The estimation of "domains of interest", where the domains of interest are different than the strata specified in the sample design, is also assessed. A domain of interest could be a subdivision of primary stratum SIC code or a combination of these.

1.2 Terms of reference

The main parts of the BSF system are the frame itself including the maintenance processes, and the survey management and quality management subsystems. Sampling and survey management form the usage side of the BSF. A comprehensive, user-oriented document[1] to address this issue has been prepared, to be used in conjunction with this document.

This document will concentrate on sampling methodology, and the actual design and drawing of samples. It will provide a basis for the standardisation of the sampling and sampling methodology, including the standardisation and integration of the SAS programs, using the programs developed for the purposes of sampling within Economic Statistics as a basis to provide and understanding in the standardisation and integration of sampling procedures. Either the Swedish procedure CLAN or the SAS version 8 procedure "surveymeans" can be used to estimate the values of population parameters together with their standard errors and 95% confidence intervals in the case of a single survey. Both procedures will be discussed in this document. For a detailed explanation on the use of CLAN, see the CLAN manual[2].

[1] Sample Maintenance Manual for Economic Statistics, 2003.

[2] Claes Andersson and Lennart Nordberg: A User's Guide to Clan 97. Statistics Sweden.

1.3 Business Sampling Frame at Statistics South Africa

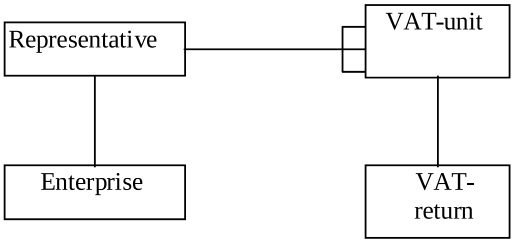
The main purpose of the Business Sampling Frame at Stats SA is to serve as a sampling frame, to draw probability samples of enterprises within the formal business sector.

The BSF contains the following types of statistical units:

1. Enterprise Unit (EN)
2. Kind of Activity Unit (KAU): the single activity part of an enterprise
3. Geographical Unit (GEO): the single activity part of an enterprise that is operating at a single geographical location

The main source of information for the BSF is the Value Added Tax (VAT) database from SARS. The VAT-unit is the tax object and could be an entire enterprise or a part of an enterprise. The relation between enterprise and the VAT-unit is mostly one-to-one, but in some cases it is one-to-many. The SARS collects VAT by sending out a VAT-return to a representative. A representative can be responsible for filling in the VAT-return and to pay VAT for more than one VAT-unit. There is one VAT-return for each VAT-unit.

Figure 1 The relation between the enterprise and the VAT-unit



The VAT-information is on the enterprise level and is delivered monthly to Stats SA.

VAT-information is used for:

A. Classification of enterprises
B. Deaths and births of the enterprises
C. Information on annual turnover of enterprises

Stats SA use the 1993 edition of the South African Standard Industrial Classification[3] (SIC'93) nomenclature. An older version of the SIC nomenclature is used at SARS for classification of the VAT-units. The translation from this code to the more recent version used by Stats SA is not straightforward in all cases. The majority of the codes have a straightforward translation down to the three-digit level.

Every enterprise with an expected annual turnover of R300 000 or more is required to register for VAT. It is also possible for enterprises with an annual turnover of less than R300 000 to register voluntarily.

[3] See appendix 3

1.4 The Administrative unit and the Statistical units

There are also disadvantages of using administrative sources. We become dependent on the legislation relevant to and the practices applied at the source organizations. The unit types defined and used by the source organizations differ from the unit types ideal for statistical use. The deviation has two dimensions.

Firstly, a difference in **delineation** forces us in some instances to bundle several VAT units together to form an "enterprise" (statistical unit), or to split an administrative unit up to form (sometimes parts of) "kind-of-activity units" and "geographical units" (statistical units). A statistical unit is a unit about which statistics are tabulated, compiled or published. The statistical units are derived from and linked to the South African Revenue Service (SARS) administrative data. An example of the difference in delineation in the BSF is:

> A company X has two administrative units A and B. Although unit A consists of two branches, operating from two different addresses with two separate bookkeeping systems, it is registered only as one unit at the administrative unit. The activities of these two branches can be the same or be different. For statistical purposes these will be registered as two branches, since there are two addresses from which it operates.

Secondly, the **survival rules** for an administrative unit can be different from the survival of a statistical unit. Thus, even if there is a direct and "one-to-one" link between an administrative and a statistical unit, the survival rules make them different. An administrative unit can for instance survive a move, a name change and a change of activity. This is impossible for a statistical unit. An example:

> An owner of a company X sells this company to another, new owner with no prior business. The activity, physical addresses of company X and the employment remains the same. The only change is the ownership and perhaps a change in legal and trading names. From a statistical point of view, has anything happened with this company that should influence our statistical output? Most likely not! But in the administrative systems the new owner will correspond to completely new units In this case the existing statistical units will be re-linked to new administrative units.

The delineation specifications and the survival rules (sometimes called "continuity rules") are the building blocks of unit definitions. They specify the entities that are of a specific unit type and when they should be included into and excluded from that unit type.

2. Definitions

2.1 Business Sampling Frame

The Business Sampling Frame is essentially a list of all employing businesses operating in South Africa. The prime purpose of the Business Sampling Frame is to provide a comprehensive source of business names and addresses from which selections can be made for inclusion in Stats SA's economic surveys. The Business

Sampling Frame provides an integrated medium for recording the hierarchical structure of Statistical Units.

2.2 Statistical unit

A statistical unit is a unit about which statistics are tabulated, compiled or published. The statistical units are derived from and linked to the South African Revenue Service (SARS) administrative data. For publications purposes, the statistical unit in the survey is the enterprise. An enterprise is defined as a legal unit or a combination of legal units that includes and, directly controls all functions necessary to carry out its production activities.

2.3 Sampling frame

A list of all sampling units in the survey population available for selection at the time of sample selection. It should be comprehensive, complete and up-to-date to avoid bias.

2.4 Sampling unit

A sampling unit is a statistical unit on the frame, which is available for selection. A sampling unit is characterised by a unique identifier (e.g. a Enterprise Number [EN] on the Business Sampling Frame), a unit identification name (e.g. legal entity name) and attributes describing the unit (e.g. turnover, industry).

2.5 Enterprise

An enterprise is a legal unit or a combination of legal units that includes and, directly controls all functions necessary to carry out its production activities. The coverage for enterprises in the BSF is adequate for sampling purposes since it includes (theoretically) all enterprises with an annual turnover equal to or exceeding R300 000 and belonging to the formal sector.

2.5.1 National Statistics (enterprise level)

It is possible to draw new samples from the BSF using the enterprise as a sampling unit, and produce reliable estimates, at national level. Turnover can be used as a measure of size in the surveys. It would be better for some of the surveys (labour statistics for example) to use number of employees as a measure of size. But at the moment this type of information is not in the BSF. Turnover and number of employees are generally correlated, but of course, this varies by economic activity.

It is also possible for enterprises with an annual turnover of less than R300 000 to register voluntarily. An annual turnover of R300 000 is used as a lower cut-off limit in business surveys, as there is no reliable source of information for enterprises below this limit[4]. The enterprises with an annual turnover below R300 000 would have a minimal impact on the estimates. It is estimated that the total contribution of these

[4] All releases and reports published should document the population that is covered by the survey.

smaller enterprises with a turnover below R 300 000, to the total turnover of all enterprises in South Africa to be about 1%.

A decision was made to standardise sampling methodology across all economic surveys within Stats SA. Only one measure of size will be used in all economic surveys namely turnover.

2.6 Kind of activity unit (KAU)

Every enterprise in the BSF has a KAU assigned. At the moment there is no enterprise that has been delineated into two or more KAUs, since VAT-data provide information on primary economic activity, but not on secondary economic activities. There is no other source of information on secondary economic activities of enterprises. In order to delineate an enterprise into more than one KAU, GEOs.

2.6.1 Functional Statistics (KAU level)

All surveys in Economic Statistics present their data by industrial activity. If the enterprise is used as a sampling unit, the enterprise's primary activity classification is used to design the survey. This means that various secondary activities are either erroneously included (enterprises whose secondary activities are included) or else excluded (enterprises with other primary activities, whose secondary activities should have been included) in the sampling frame. Functional statistics are statistics based on economic activity and not on the institutional conditions. To be able to obtain functional statistics on the enterprise level, the KAU unit should be used as sampling or observational unit. However, if the KAU is used as sampling unit in functional statistics it could make it possible for one and the same Enterprise to be part of the populations of different sectors of industry.

Another way to obtain functional statistics would be to use the GEO units as sampling units. However, at the present time the current units in the GEO frame cannot be used as sampling units, as described below. Even if it could be possible to use the GEO frame, it is doubtful whether the information demanded in the economic surveys could be provided on the GEO level. For example if an enterprise with two GEOs, namely an IT-department and a manufacturing company exist on the BSF. The IT-department within a manufacturing company is restructured into a company of its own. The only "market" of this new company is the rest of the original company. Thus there will be two companies in the administrative system. The two companies would be combined as one single enterprise on the BSF in the process to create a statistical unit. If only the IT-company (GEO) is selected in a survey, it does not operate on the open market and does not necessarily sell its services at market prices. Will all data asked for in the survey be available at the IT-company? It is likely that some information for the IT-company still has to be collected from the other, main part of the original company. Thus, there is an obvious risk for lack of data availability. If only the rest of the original company (GEO) is selected, its investments in IT development may be purchased at non-market prizes. Thus it will be impossible for the two GEOs to provide information of the other when surveyed by Stats SA.

Until enterprises with important secondary activities are delineated into KAU units, approximate functional statistics could be obtained in another way: assume that a survey on manufacturing is to be conducted. If enterprise is used as the sampling unit, then the survey is designed by the primary economic activity of the enterprises. Non-manufacturing enterprises with one or several GEO units in manufacturing would not be included in the sampling population for this survey. But it would be possible to include all known non-manufacturing enterprises with one or more GEO units in manufacturing in the population and treat them in the same way as the other enterprises. The information on GEO units would come from the GEO frame. As there is under-coverage in the GEO frame, it is impossible to be sure that the whole activity is covered, but non-manufacturing enterprises with a large manufacturing unit will probably be covered by this approach.

2.7 Geographical unit (GEO)

The BSF contains a large number of GEO units. Currently, GEOs cannot be used as sampling units. The current units in the BSF called GEO units should probably not have been identified as GEO units if the definition had been used when they were created. The Division: Systems of Registers are currently surveying some of largest enterprises in terms of turnover (usually these large enterprises is also more complex enterprises) to determine their branches, in order to improve the quality of GEOs on the BSF. In many cases, observational units are implemented as GEO units in the BSF.

2.7.1 Regional statistics (GEO level)

It is important to concentrate on the most important issues, which, in this situation, is to produce reliable estimates at the national level. The most important issue for the National Accounts is to be able to calculate the Gross Domestic Product with high precision at a national level.

There are nine provinces in South Africa. To be able to calculate reliable estimates at a provincial level, GEOs should be used as sampling units. Unfortunately there is presently no comprehensive information on multi-location enterprises. The information in the BSF on multi-location enterprises degree of coverage is unknown.

The units in the current GEO frame cannot be used as sampling units. The most important reasons are the under-coverage and the unknown quality on the GEO units in the BSF. Another reason is that if samples are drawn from the current GEO frame, it will be impossible to stratify the surveys in respect of size. The GEO units belonging to multi-location enterprises, which are the interesting ones because they can operate in different provinces (regions), do not have any information on size e.g. turnover. Until a measure of size is available for all GEOs it will be difficult to draw an effective sample using GEO units as sampling units.

Until the GEO frame has a high (er) quality, enterprise should be used as the sampling unit, not only for national, but also for regional economic statistics. Then the enterprise can be used as the sampling unit and GEO as an observational unit, in other words cluster sampling. This means that all selected enterprises will be required to report on all their GEOs.

Most of the multi-location enterprises would fall into the sizegroup 1 of completely enumerated strata, or into strata with medium sized enterprises. A multi-location enterprise is here defined as an enterprise in the BSF with more than one GEO assigned. Strata with medium sized enterprises will have quite a large sampling fraction. This means that most multi-location enterprises will be covered with this approach. The data could then be presented both on a national and on a provincial level. The provinces would be used as domains, and if most multi-location enterprises were to be included in the sample, estimates would not be too far off the mark.

2.8 Observational units

An observational unit is a unit for which data are reported that does not equate to the statistical unit on the sampling frame, i.e. an aggregation or dissection of statistical units. These units could not be used as sampling units in the new BSF but they could be used as observational units. That is, the enterprise could be used as the sampling unit and the selected enterprises could then be required to report on their observational units.

Examples of such units are:

1. **Establishment**: the smallest economic unit that functions as separate unit; and
2. **Branch**: the part of the enterprise for which data are collected.

CHAPTER 2

SURVEY DESIGN, DRAWING OF THE SAMPLE AND ESTIMATION OF POPULATION PARAMETERS

1. Survey design

When designing a survey, it is fundamental to keep the main purpose of the survey in focus. The resources should be allocated in such a way so that the required estimates could be produced with the best attainable precision.

For all economic surveys, required sample sizes are calculated under different conditions. For Stats SA’s purposes a stratified simple random sample will be drawn from the sampling population for each survey. But the required sample size and relative precision obtained could differ from survey, to survey because of budget constraints, and the level of presentation. A relative standard error of one percent in the estimates is a considerably high precision. This will result into obtaining in reality, when estimating the study variable, in a much larger relative standard error (RSE). At the moment only one size measure is available in the BSF. Therefore it is necessary to use measure of size (MOS) as a variable, such as annual turnover, both for stratification and for allocation.

Having decided per industry primary stratum, the turnover cut-off points of the size groups have to be determined next. There are 4 size groups per industry primary stratum in most of the economic surveys, constructed in such a way that size group 1 contains the largest enterprises in that stratum. All enterprises in these first size groups of the different strata are included in the sample. The lower turnover cut-off point of size group 1 of industry primary strata is obtained directly from the cumulative turnover values of the ordered (according to turnover) enterprises in the industry primary strata. Considering 4 size groups per primary stratum, taking L as the lower limit of size group 4, then two further turnover cut-off points has to be determined to differentiate between size groups 2 and 3. Two possible approaches to obtain this “in between” turnover cut-off value are:

(a) to use the rounded values of $L\times\sqrt[3]{(U/L)}$ and $L\times\sqrt[3]{(U/L)^{2}}$ where U denotes the lower turnover cut-of point of size group 1 – this approach is based on the assumption of an exponential (J-shaped) turnover distribution; and

(b) to divide the total turnover of all enterprises with turnover between L and U into three approximately equal intervals. The rounded turnover value of the enterprise “nearest” to these division points in the ascending ordering of enterprises with turnover between L and U is then to be used as the “in between” turnover cut-off values – percentile method.

Thus, considering now R 300 000 as lower limit of size group 4, in which case two “in between” turnover cut-off points have to be determined. These two “in between” turnover cut-off points will be the rounded values of $300000\times\sqrt[3]{(U/300000)}$ and $300000\times\sqrt[3]{(U/300000)^{2}}$ in approach (a) and the rounded off turnover values of the enterprises “nearest” to the two division points in the ascending ordering of the enterprises with turnovers between 300000 and U which divide the total turnover of all enterprises with turnovers in the range 300000 to U into three approximately equal intervals in approach (b).

The total turnover of the enterprises in size group 1, calculated as a proportion of the total turnover of all enterprises in a primary stratum, should be relatively high, preferably not less than 60%. An annual turnover of R300 000 is used as a lower cut-off limit (U4 cut-off value) in the surveys.

The preferred method for sample designs of economic surveys within Stats SA is the percentile method. The cut-off of the upper U1 cut-off value is calculated through the use of a percentile method whereby the cut-off values for size group 2 and 3 are calculated programmatically. Percentiles are the cumulative value of the turnover of the population, where every percentile represents a 100th of the population.

The stratification for each survey is presently done on the two-digit sic-code level to divide the population into more homogenous subgroups. Due to the present quality defects on the BSF currently, the most reliable information exist on a 2-digit and thus it is possible to stratify on a 2-digit siccode level. Furthermore, the survey results are mainly to be presented on the one- and two-digit siccode level.

1.1 Delimitation of the survey

The sampling frame for all surveys include at least all operating enterprises with a one-digit SIC level (minimum requirement) as well as invalid siccodes. The sampling frame will be created from the Business Sampling Frame annually and should include all active, birthed (newly created enterprises) and reactivated units [life_cycle_code], with a turnover greater than or equal to R300 000.

1.2 Stratification of the population

Stratified sampling is often used:

- Because it yields better precision in the estimates at a national level if the population is divided into more homogenous sub-populations or strata, compared to the case where the population is not divided in this way.
- To ensure adequate representation of subpopulations in the sample.

1.2.1 Economic activity

The most important siccode levels of study in economic surveys of Stats SA are five-digit SIC level activities, but it is not possible to stratify at this level because the 5-digit SIC codes are not known for all the enterprises in the frame. If the 5-digit SIC code is specified in the frame, the information is not reliable for all enterprises contained in the frame. Therefore, the stratification should be done at the one-digit or two-digit SIC level, where the classification in the BSF is more reliable. The industry strata (primary strata) are defined as the pre-specified SIC code levels (i.e. 1-digit and/or 2-digits and/or 3-digits) of the different industry sectors to be taken into account in the design of the sample. Enterprises within each primary stratum will generally be grouped into 4 non-overlapping size groups (secondary strata) based on the measure of size (MOS).

The five-digit SIC, can be used as the basis for forming a domain, if there is questions in the questionnaire, which make it possible to classify the enterprises into the correct five-digit SIC level category.

A decision was made by Stats SA to include all enterprises without a valid two-digit SIC classification in the sampling frame in the attempt to collect the correct classification information on these enterprises when the questionnaires are dispatched to the respondents. For these ‘invalid’ classified enterprises a sperate domain would be created and the re-classified information would be used when the estimates are calculated.

1.2.2 Size groups

Within each primary stratum (1-digit or 2-digit SIC codes, for example) the enterprises are ordered from largest turnover to lowest turnover. The turnover cut-off points is used for subdividing the enterprises into non-overlapping groups, per industry Standard Industrial Classification (SIC) code level (primary strata), into the measure of size (MOS) intervals based on turnover, called secondary strata (viz. the cross-classification of the industry strata (primary strata) with the measure of size intervals (secondary strata)). The industry strata (primary strata) are defined as the pre-specified SIC code levels (i.e. 1-digit and/or 2-digits and/or 3-digits) of the different industry sectors to be taken into account in the design of the sample. Enterprises within each primary stratum will generally be grouped into 4 non-overlapping size groups (secondary strata) based on the measure of size (MOS). The size group cut-offs will be the same for all enterprises in each primary strata. The SIC code levels of these strata (primary strata) vary between different economic surveys, depending on various factors, such as the purpose of the survey and the size of the sample. The number of (measure of) size intervals or groups per industry primary stratum is usually taken as a value between 4 and 6. However, Stats SA has decided to use 4 size groups per industry primary stratum in its economic surveys, constructed in such a way that size group 1 contains the largest enterprises in that stratum with lower cut-off value U1. All enterprises in these first size groups of the different strata are included in the sample.

1.3 Determining of cut-off values

If samples are drawn among enterprises with an annual turnover of less than R300 000 it is not possible to calculate estimates for **all** enterprises with an annual turnover of less than R300 000. It is only possible to calculate estimates for the enterprises with an annual turnover of less than R300 000 **in the BSF**. It should however be noted that enterprises with an annual turnover of less than R300 000 contribute very little to the economy. This means that a cut-off limit of R300 000 would be appropriate for all surveys within Economic Statistics.

Steps followed in determining the cut-off values (see chapter 3 for the discussion of the practical application in detail):

- Sort the enterprises within each primary stratum, from the largest turnover to the smallest turnover, where size group 1 constitutes of the enterprises with the largest turnovers.

- Determine the cut-off for every primary stratum, the lower cut-off value U1 of size group 1 for a series of proportions of total turnover contribution by the large enterprises in size group 1, ranging from 0.50 to 0.95 by 0.025 or 0.05. The number of enterprises in size group 1 should also be determined.
- Keeping the total sample size in mind, decide on preliminary sample size per stratum and the sample size to be allocated to size group 1 per primary stratum (leaving "sufficient" sample size to be allocated to the other three size groups).
- Calculate size group cut-off values for U2 and U3 for each strata:

Size group 1:	Turnover ≥ U1
Size group 2:	U2 ≤ turnover < U1
Size group 3:	U3 ≤ turnover < U2
Size group 4:	R 300 000 ≤ turnover < U3

Note: Take an initial decision on the values of U1 for different cut-off strata based on: the proportions of total turnover contribution by enterprises in size group 1 and the number of enterprises included in these size groups 1 – using the percentile method. Previous similar surveys and results obtained should also be used as guideline when an initial decision is made.

See appendix 1 for the SAS program. Percentiles are the cumulative value of the turnover of the population, where every percentile represents a 100th of the population.

Table 1 Summary table for size group 1 of cut-off points for the survey Economic Activity Survey (EAS), as been decided upon (cf. table 2 and table 3)

Sic0	**Siccode**	**U1 – (Lower boundary of size group 1 enterprises)**
2	Mining	44 026 267
3	Manufacturing	74 361 74
4	Electricity	25 800 494
5	Construction	26 840 045
7	Transport, storage and communication	55 081 824
8	Financial Intermediation, insurance, real estate and business services	37 182 180
9	Community, social and personal services	18 747 666
61	Wholesale trade	110 736 554
62	Retail trade	55 032 591
63	Motor trade	58 134 094
64	Hotels and restaurants	11 536 617

This is an extract of the program that uses the percentile method to assist in the decision of the U1 cut-off points (see appendix 1 for the SAS program).

Note that if enterprises are sorted in increasing order, the SAS program becomes.

```
proc univariate data=frame1A noprint;
        var turnover;
        output out=percentiles pctlpre=P_          pctlpts=5 to 50 by 2.5;
        weight turnover;
        by sic0;
    run;
```

where:

sic0 is the chosen primary stratum and
frame 1A is the sampling frame of the specific survey containing the created chosen primary stratum (sic0).

Note if enterprises are sorted in decreasing order then, the SAS program becomes.

```
    proc univariate data=frame1A noprint;
        var turnover;
        output out=percentiles pctlpre=P_          pctlpts=50 to 95 by 2.5;
        weight turnover;
        by sic0;
    run;
```

Table 2 – Determining U1 cut-off values for size group 1 per sector – Turnover and number of enterprises (counts) by percentile

Note: Enterprises sorted by ascending Turnover.

[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Where P_5 is the 5th percentile and P_7_5 is the 7.5th percentile etc.

Table 3 – Determining U1 cut-off values for size group 1 per sector – Turnover and number of enterprises (counts) by percentile

Note: Enterprises sorted by ascending Turnover

[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Where **cnt950** for SIC 2 (Mining) is the number of enterprises falling in the 5th percentile with a cumulative turnover of R 44 026 267 million and **cnt875** for SIC 4 (Electricity) is the number of enterprises falling in the 7.5th percentile with a cumulative turnover of R 25 800 494 million, etc. **counts_select** is total number of enterprises falling in this cut-off point for the specified industry. This will differ for every economic survey.

1.4 Decide sample sizes in strata, allocation

Neyman optimum allocation is used to allocate per primary stratum, the available sample size to the size groups 2, 3 and 4. To be able to use optimum allocation there must be a variable available for the whole population that is correlated with the study variable (the variable to be measured in the survey). Turnover is available for the whole population of enterprises in the BSF, and turnover is likely to be correlated with many variables in the questionnaires used in economic surveys. For other kind of business surveys, where turnover is not correlated with the variables in the questionnaire, it might be better to use another method for allocation, for example proportional allocation.

If the purpose of an economic survey is only to obtain the best possible (optimal) precision of an estimate of a parameter of the population for the economy as a whole, then Neyman's optimum allocation of the sample to primary as well as secondary strata should be used. Note that no sampling is done in size groups 1.

It should be emphasised that the allocated sample sizes to the industry stratum[5] obtained should only be considered as a guideline. The final decision whether or not an industry stratum is adequately represented in the sample is given by the resulting value of the (theoretical) relative standard error (RSE) (using turnover as study variable) relative to the (theoretical) RSE values obtained for the other industry strata considered.

1.4.1 Sample allocation methods[6]

Notation: Consider a population with H strata containing N elements (sampling units).

Let

$N_h = \text{number of population elements in h-th stratum}$

$T_h = \text{population total of an auziliary variable z in h-th stratum}$

$\overline{T}_h = \text{population average of variable z in h-th stratum}$

$S_h = \text{population standard deviation of variable z in h-th stratum}$

$$CV_h = \text{population coefficient of relative variation of z in h-th stratum} \\ = S_h / \overline{T}_h$$

$n = \text{total sample size and } n_h = \text{sample size allocated to the h-th stratum.}$

The auxiliary variable z indicates the MOS variable, and is here defined as the turnover of the enterprise. The word population refers to the BSF. The word stratum

[5] Stratum = Neyman stratum || Size group.

[6] Lehtonen, R. and Pahkinen, E.J. Practical Methods for Design and Analysis of Complex Surveys. John Wiley & Sons, New York. 1994.

could refer to an industry sector with SIC code consisting of a given number of digits or to a (measure of) size interval or to a combination of these. According to the power rule,

for $0 \le \delta \le 1$,

$$n_h = n\left(T_h^{\delta} \times CV_h\right) / \left(\sum_{h=1}^{H} T_h^{\delta} \times CV_h\right)$$

Special cases

1) Proportional to CV allocation:
   If $\delta = 0$, then

$$n_h = n\,(CV_h \,/\, \sum_{h=1}^{H} CV_h)$$

2) Neyman (optimal) allocation:
   $\delta = 1$. Then $CV_h = S_h / \overline{T}_h = N_h S_h / T_h$ and thus

$$n_h = n(N_h S_h \,/\, \sum_{h=1}^{H} N_h S_h)$$

If, furthermore, all S_h = a, constant value, then
$n_h = nN_h / N$, i.e. proportional allocation.

Remarks

1) Proportional allocation is a necessary condition for obtaining a self-weighting sample and guarantees an equal relative share of the sample in all strata. The problem, however, is that it is not an optimal allocation procedure and provides less efficient estimates than generally expected, the only exception being when the variation in the size variable z is equal in all strata.
2) Neyman's allocation procedure is an optimal procedure (i.e. provides the most efficient estimates) under stratified sampling and should be used whenever feasible. There are, however, situations where it is not feasible or desirable to use Neyman's procedure and where the general power allocation procedure provides more acceptable estimates.
3) The power allocation procedure is appropriate for surveys where there are numerous small strata and also a need for relative precise estimates at each stratum level. A suitable choice in practice for the power δ is often $^1/_2$ or $^1/_3$. These choices can be viewed as a compromise between Neyman's allocation (i.e. δ=1) and proportional allocation procedure (i.e. δ=0), which leads to approximately equal relative precision for all strata.

The purpose of an economic survey is to obtain the best possible precision of an estimate of a parameter of the population for the economy as a whole as well as acceptable precision levels at certain domains, thus Stats SA decided to use Neyman's allocation of the sample to all size group (secondary) strata within each primary strata except for size group 1 where all enterprises will be completely enumerated. In the next section the Neyman Optimal allocation will be discussed in detail.

1.4.1.1 Optimum allocation (Neyman allocation)

How large a sample is it necessary to draw to achieve a specified precision in the estimates at a certain level? Given the budget constraints, is this sample size realistic? First it has to be decided on what level the allocation should be conducted, in other words, what is the main purpose of the survey?

If the most important issue is to estimate total Manufacturing then the allocation should be conducted on that level (one-digit level). When the relative precision is decided in terms of the relative standard error (coefficient of variation) for that level, the total required sample size is calculated under the relative condition that the Neyman optimum allocation method over size (secondary) strata will be used. Initially the size groups that are to be completely enumerated are given to the allocation program.

A minimum sample size in each size group is also given. To be able to impute non-response with the average in the stratum there must be some respondents in each stratum. If too few units are selected then it is possible that no one responds. In all surveys within Economic Statistics, the sample size in each size group consists of a minimum size of 10 enterprises (or all enterprises if these are less than 10 in the population in that size group).

The relative standard error (RSE) at a certain level could be expressed as follows:

$$\frac{\boldsymbol{SE\ (t)}}{\boldsymbol{t}} \leq \boldsymbol{\alpha}$$

where ***α*** = required RSE
SE (t) = the standard error of T
t = estimated value of T.
Relative precision = *1.96 * RSE* at 95% confidence level.

Neyman allocation over size strata is given by[7]:

$$n_h = \frac{nW_hS_h}{\sum_{h=1}^{H} W_hS_h}$$

1.4.2 A SAS-macro for Neyman allocation

A SAS-macro, called Neymann, has been developed at Statistics Sweden. The macro performs Neyman (optimum) allocation in order to give some guidance when deciding on the sample size in each secondary stratum per primary strata. It is very important that the variable used in the calculations is correlated with the study variable. Otherwise, it might be better to use another method for the allocation, for example proportional allocation, or to combine the Neyman allocation with other strategies.

[7] See Cochran.

The SAS-macro calculates the required sample size in each stratum to achieve a specified relative precision. In fact, each time the user runs the macro, the required sample sizes are calculated for ten beforehand-specified relative precisions. The precision should be specified for the strata. A stratum could be a one-digit or two-digit siccode, for example the stratification could be done on the two-digit siccode. For the specified precision, the program uses the same precision for all specified strata.

1.4.2.1 Population or stratum level

The Neyman allocation macro can base the calculations on a variable available for the whole population or sub-population, for example turnover. Then the input to the macro should be the entire population or sub-population, and each enterprise should have information on turnover (population level).

But, it is also possible to base the calculations on sample variance (see page 103 in Särndal, Swensson, Wretman 1992) in each stratum (stratum level). For example, in SEE, the size stratification is based on turnover, because number of employees is not available for the whole population. But in the questionnaires, sent out to the enterprises in the sample, Stats SA ask for number of employees. This information could be used to calculate estimated total number of employees and sample variance in each stratum.

1.4.2.2 Neyman stratum level

Stratification of a finite population U = $\{1,\ldots,k,\ldots N\}$ means a partitioning of U into H subpopulations, called strata and denoted U_1, ….., U_h, …, U_H, where
U_h = $\{$ k: k belongs to stratum h $\}$. By stratified sampling is meant that a probability sample s_h is drawn from U_h according to a design p_h (·) (h=1, .., H) and that the selection within one stratum is independent of the selection within all other strata.

The Horvitz-Thompson estimator for the total number of employees (t_h) in stratum h is given by:

$$\hat{t}_h = \frac{N_h}{n_h}\sum_{s_h} y_k$$

where y_k denotes number of employees on enterprise number k in the sample, N_h total number of enterprises in stratum h, and n_h total number of enterprises included in the sample in stratum h.

If the allocation variable is known for the whole population, the stratum variance, S^2, regarding this variable is used in the Neyman allocation formula.

The allocation variable should be known for all units in the population and is taken as the MOS. The allocation variable should be correlated with the study variable, for example number of employees. Below is an example where the allocation variable is turnover (MOS) and the study variable is number of employees. It should be noted that the study variable could be any variable of interest.

The stratum variance is given by:

$$S^2_{yU_h} = \frac{1}{N_h - 1}\sum_{U_h}\left(y_k - \bar{y}_{U_h}\right)^2$$

where y_k now denotes number of employees for enterprise number k in the population, N_h total number of enterprises included in the population in stratum h and $\bar{y}_{U_h}$ means the population mean in stratum h.

If the desired allocation variable is not known for the whole population then survey data regarding this variable could be used. The stratum variance is then estimated by the sample variance, and the sample variance in stratum h is given by:

$$s^2_{ys_h} = \frac{1}{n_n - 1}\sum_{s_h}\left(y_k - \bar{y}_{s_h}\right)^2$$

where y_k denotes number of employees on enterprise number k in the sample, n_h total number of enterprises included in the sample in stratum h and $\bar{y}_{s_h}$ denotes the sample mean in stratum h.

Using estimates in the allocation means to replace t with $\hat{t}$ and the stratum variance with the sample variance in the formulas described on page 20 in Lindblom, 2000. This is done by giving the Neyman macro estimated totals and sample variances for each stratum as input, as illustrated in example 4 below.

When using estimated totals and sample variances referring to number of employees the given sample sizes could be interesting to compare with the given sample sizes using turnover. It is not possible to stratify on number of employees, but at least it is possible to base the allocation on a variable (**highly) correlated** with the study variable. If estimates are used as input it is advisable to use several occasions from the survey. Calculate averages of the estimates in each stratum, and use the averages as input to the allocation. In strata where the given sample sizes differ greatly (using turnover versus number of employees) this could be an indicator of low correlation between number of employees and turnover, and could be considered when deciding sample sizes.

1.4.2.3 Using the macro

Initially strata (primary and secondary strata) to be completely enumerated must be specified, as well as a minimal sample size in each stratum (which could be zero). In the Neyman macro the parameter ”villkor” is used to specify, size groups to be completely enumerated. It is also possible to specify (combination of industry and size group) strata to be completely enumerated. If one does not use the parameter villkor then there is no group of enterprises that is specified to be completely enumerated *in beforehand*. But the Neyman allocation could, of course, give sample sizes, which are equal to the total number of enterprises in the stratum

The allocation could initially give a sample size larger than the total number of enterprises in a stratum. Then the allocation program considers that stratum as completely enumerated and recalculates the sample sizes in the remaining strata to be included in the survey.

The macro needs certain parameters specified:

%Macro neymann (*niva* =, *dsn* =, *stratum* =, *allvar* =, *villkor* = (dummy=0),
bg = 1, *startpre* = 0, *inc* = 0.1, *lillmin* = 0, *popvar* =, *storan* =);

Parameters to be specified to the macro:

Parameter	**Mandatory**	**Default value**	**Explanation**
Niva	No	Blank (means stratum level)	pop (means the entire population)
Dsn	Yes	No	Name of the input data-set
Stratum	Yes	No	Stratum identification (primary strata by secondary stratum)
Allvar	Yes	No	Name of the allocation variable (if niva = pop) or the estimated total of the allocation variable in each strata (if niva = blank).
Villkor	No	dummy = 0 means that no strata is specified to be completely enumerated in beforehand	Villkor specifies strata (primary and secondary strata) to be completely enumerated
Bg	No	bg = 1 means one strata for the whole population	Desired strata should be specified (primary stratum)
Startpre	No	0	Start precision minus first increment
Inc	No	0.1 percent	The size of the increment (in percent)
Lillmin	No	0	Minimum sample size in each strata
Popvar	Yes if niva is blank	No	Name of the variable containing the sample variance in each strata
Storan	Yes if niva is blank	No	Name of the variable containing the number of

			enterprises in each strata

Clarification of table above:

niva = means that the calculations are based on the entire population.
dsn = means the data-set containing the population.
stratum = is the variable containing the stratum identification i.e. primary*secondary strata (in program stratum = sic3 || sizegrp.
allvar = the variable containing the allocation variable.
villkor = specify the size group that should be completely enumerated (On the input data-set there must be variable called sizegroup containing this information).
bg = the variable containing the desired primary strata (in program is bg = sic0).
Startpre = 0 means that the start precision is 0. But the first precision sample sizes are calculated for the startpre+inc.
inc = means that for every iteration the relative standard error increases with a percentage point specified.
Lillmin = specifies the minimum sample size in each strata.

1.4.2.4 Two examples on how to use the macro

<u>Example 3:</u> Population level

%neymann(niva = pop, *dsn* = frame_SEE, *stratum* = stratum, *allvar* = turnover, *villkor* = (sizegroup = ‘1’), *bg* = sic2, *startpre* = 0, *inc* = 0.2, *lillmin = 10);*

In example 3 the following parameters are given:

niva = pop means that the calculations are based on the entire population.
dsn = frame_SEE means that the data-set containing the population is called frame_SEE.
stratum = stratum means that the variable containing the stratum identification is called stratum (primary*secondary strata).
allvar = turnover means that the variable containing the allocation variable is called turnover.
villkor = (sizegroup = ‘1’) means that size group one should be completely enumerated (On the input data-set there must be variable called sizegroup containing the information).
bg = sic2 means that the variable containing the desired primary strata is called sic2.
Startpre = 0 means that the start precision is 0. But the first precision sample sizes are calculated for the startpre+inc, in this example 0+0.2 = 0.2
inc = 0.2 means that for every iteration the relative standard error increases with 0.2 percentage points.
Lillmin = 10 means that a minimum sample size of five enterprises is specified in each strata.

Note that the given primary strata could vary in the allocation. If the primary stratification is done on the two-digit level then, for example, the primary strata’s could be the two-digit siccode in manufacturing and the one-digit siccode for the rest of the industries. Bg could be primary strata or it could be a domain consisting of several primary strata. But, as mentioned before, bg could not be domains cutting through strata.

Create the strata (bg) for example like this:
If manufacturing then bg = substring(sic-code,1,2); else bg = substring(sic-code,1,1);

Example 4: Stratum level (primary and secondary strata)

Create a SAS dataset (for example called survey_SEE) from the SEE survey containing the following variables for **each stratum**:
- variable containing stratum identification (primary and secondary strata)
- variable containing total number of enterprises in the population
- variable containing estimated total number of employees
- variable containing sample variance regarding number of employees

Lets say that the variable containing the stratum (primary and secondary strata) identification is called stratum, the variable containing estimated total number of employees is called tot_employees and the variable containing sample variance is called var_employees and the variable containing number of enterprises in the population is called numb_pop. The rest is similar to the population level.

Then the macro can be used like this:

%neymann(*niva* =   , *dsn* = survey_SEE, *stratum* = stratum, *allvar* = tot_employees, *villkor* = (sizegroup='1'), *bg* = sic2, *startpre* = 0, *inc* = 0.2, *lillmin* = 5, *popvar* = var_employees, *storan* = numb_pop);

1.4.2.5 Output from the macro

The macro will create an output work data-set called ut, see table 5. It contains the variables:

Bg = primary strata
Stratum = stratum (primary and secondary strata)
Storan = number in population
Allvar = sum of the allocation variable
Nhsh = number in population multiplied with the standard deviation in the population based on the allocation variable
N1 = required sample size for the start precision
.
.
N10 = required sample size for the stop precision

The macro prints, in the output window, the total required sample sizes required for each precision as shown in table 4. N1 is the required sample size for the start precision and N10 is the required sample size for the stop precision.

Table 4 Required sample sizes

N1	58 446
N2	29 716
N3	19 197
N4	14 384
N5	11 813
N6	10 282
N7	9 294
N8	8 630
N9	8 155
N10	7 818

Table 5 The data-set ut

bg	stratu	stora	allvar	NhSh	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10
30	302	391	7936023328	3262456888	388	179	94	57	38	27	20	15	12	10
30	303	426	2663989179	719833091	86	40	21	12	8	6	5	5	5	5
30	304	1972	2621485848	1895818364	225	104	55	33	22	15	12	9	7	6
31	312	306	5980719370	2398563402	306	144	75	44	29	21	15	12	9	7
31	313	425	2718886127	738202675	106	44	23	14	9	6	5	5	5	5
31	314	2067	2674402460	1991234091	285	120	62	37	24	17	13	10	8	6
32	322	302	5605317386	2409951123	302	270	162	104	71	51	39	30	24	19
32	323	455	2892196535	767925959	236	86	52	33	23	16	12	9	8	6
32	324	2505	3213202919	2330461388	718	261	157	100	69	50	37	29	23	19
33	332	442	8430172175	3450195841	270	93	45	26	17	12	8	6	5	5
33	333	603	3866375050	1015198860	79	27	13	8	5	5	5	5	5	5
33	334	2032	2853144967	2042212978	160	55	26	15	10	7	5	5	5	5
34	342	112	1984846558	791584301	112	112	80	53	37	27	20	16	13	10
34	343	179	1163261348	301239214	131	51	30	20	14	10	8	6	5	5
34	344	969	1199806590	887614522	386	151	90	60	42	30	23	18	14	12
35	352	771	14183072203	6206692222	771	435	227	136	89	63	47	36	29	23
35	353	1393	8753252462	2342685266	434	164	86	51	34	24	18	14	11	9

The macro also calculates the actual relative precisions (as a %), that the given sample sizes will produce. It will not be exactly the ones specified, mainly because the macro does not calculate the sample size as an integer. But when the sample is to be drawn it is not possible to include 8.7 enterprises in a sample, so the calculated sample sizes have to be rounded off. Actual precision is the precision achieved after the rounding, see table 6.

The macro will create a work output dataset called pres_bg, see table 6. It contains the variables:

Bg = primary strata
Erpr1 = achieved precision for the start precision
.
.
.

Erpr10 = achieved precision for the stop precision

Table 6 Actual precisions

bg	erpr1	erpr2	erpr3	erpr4	erpr5	erpr6	erpr7	erpr8	erpr9	erpr10
30	0.200	0.400	0.600	0.802	0.998	1.199	1.373	1.569	1.752	1.902
31	0.200	0.400	0.600	0.802	1.008	1.206	1.399	1.559	1.756	1.991
32	0.200	0.401	0.598	0.799	0.995	1.198	1.397	1.602	1.791	2.011
33	0.200	0.401	0.601	0.797	0.993	1.159	1.380	1.511	1.608	1.608
34	0.200	0.401	0.599	0.799	0.998	1.208	1.407	1.605	1.808	1.986
35	0.200	0.400	0.600	0.799	1.001	1.201	1.402	1.604	1.790	2.005

2. The drawing of samples

SAS programs have been developed which can serve as examples on programs for defining and stratifying the population for a specific business survey.

2.1 Sequential simple random sampling

2.1.1 The JALES technique

The basic idea in the JALES[8] sampling technique is to associate a *permanent, independent and unique* random number, uniformly distributed over the interval (0,1), to every unit in the register. The population is ordered in ascending order according to the size of these random numbers. The *n* first units from randomly selected starting value on the list constitute the desired sample. It can be shown (Ohlsson 1992) that this procedure - sequential simple random sampling without replacement - is equivalent to simple random sampling without replacement.

For every unit persisting in the register the same random number is used on each sample occasion. Every new business is assigned a new random number while discontinued (closed-down) businesses are withdrawn from the register with their assigned random number. Annually a new sequential simple random sample will be drawn, using the permanent random numbers. In this way we always get a simple random sample from the up-dated register. However, a large overlap with the latest sample can be expected if the same starting point is used, since ongoing businesses have the same random numbers on all occasions. This enables good precision in estimates of change over time.

By the symmetry of the uniform distribution, we could just as well take the last *n* units preceding the randomly selected starting point to obtain the simple random sample. We can, in fact, select the first n units to the left, or to the right, of any fixed point in (0,1). If there are not enough units to the right (or left) of our starting point a, we simply continue the selection to the right (or left) of the point 0 (or 1). See

[8] This technique was developed at Statistics Sweden in the early 70' s by Johan Atmer and Lars-Erik Sjoberg from whom 'JALES' is an acronym

Appendix 2, for starting points of all surveys within Economic Statistics, using the BSF as sample frame.

2.2 Sample co-ordination

Co-ordination among surveys and over time is obtained by using the JALES technique, which is based on the use of random numbers permanently associated with the sampling units (see 2.3.1 – 2.3.2.3). The method is used by Statistics Sweden and several other countries, (Ohlsson 1992). The use of the same frame for many surveys makes it possible to compile comparable statistics, which is vital for the National Accounts. Surveys used by the National Accounts to compile the Gross Domestic Product (GDP) should be using similar definitions of population units and compatible variables as those used in the survey design.

To increase the precision in estimates of change over time, the positive co-ordination design ensures that subsequent samples for the same survey are overlapping, although each sample is drawn from an up-to-date register. Negative co-ordination between surveys is used to spread the response burden among the businesses. Positive co-ordination between surveys is used to facilitate comparisons of variables in various surveys.

Drawing co-ordinated samples for business surveys has three main purposes:

- to obtain positive co-ordination of samples over time
- to obtain samples that are co-ordinated - negatively or positively - among different surveys
- to promote the use of the same frame for many surveys.

2.2.1 Positive and negative sample co-ordination

In order to coordinate two samples with desired sample sizes n_1 and n_2, choose two arbitrary constants a_1 and a_2 in (0,1). The units with the n_1 random numbers closest from a_1 in one direction (left or right) are included in the first sample. The second sample includes the ones with the n_2 random numbers closest from a_2, in the same or the opposite direction as the first sample.

The maximal positive co-ordination of two surveys is obtained by using the same starting point and direction for both. Negative co-ordination of two surveys can be achieved by choosing different starting points (well apart) and using the same direction. An alternative way is to choose the same starting point (or two close ones) but different directions. There are not always enough units to obtain complete negative co-ordination, but this technique at least reduces the number of units that the surveys have in common.

Negative co-ordination is a very effective tool to distribute the response burden among small enterprises. It is important to spread the response burden among these enterprises, because they often do not have the capacity to fill in many questionnaires. Among large and medium sized enterprises, there is no room for spreading the response burden because the number of these enterprises are small, and they have a large impact on the estimates in terms of turnover. Large enterprises are always included in the samples and medium sized enterprises could have a heavy burden, especially in industries with few enterprises. For medium sized enterprises other methods such as more efficient stratification and allocation, or the use of administrative sources etc. could be used to reduce the respondent burden.

2.2.2 Co-ordination when surveys are stratified

In practice there are several strata in a survey, and a sequential simple random sample is drawn in each of them. For a particular survey the same direction and starting point is used in all strata. If the starting points are different then the surveys will be negatively co-ordinated, because a random number which is small (large) in one stratum is also small (large) in another stratum.

2.2.3 Co-ordination when surveys are based on different kinds of units

It is possible to coordinate surveys based on different kind of sampling units, e.g. enterprise and geographical unit. This can be done if the units are co-ordinated through their random number, e.g. if the random numbers are assigned to the geographical units. A single-location enterprise is given the same random number as its geographical unit. A multiple-location enterprise is given the random number of one of its geographical units (the main one, in some respect). For the majority of the enterprises, single-location enterprises, the co-ordination between the units is perfect. For the multiple-location enterprises the co-ordination is less efficient, because it is only possibly to coordinate a multiple-location enterprise with *one* of its geographical units.

Co-ordination between surveys, based on different kinds of units, is most important for small enterprises, because it is important to spread the respondent burden. The majority of small enterprises are single-location enterprises, and for them the co-ordination between surveys based on different kinds of units is perfect. For multiple-location enterprises the co-ordination is less efficient, but on the other hand the majority of the multiple-location enterprises are large. Large enterprises are almost always included in the samples so there is less opportunity for distributing the respondent burden among them. In other words, it is not so important that the co-ordination for multiple-location enterprises is less efficient.

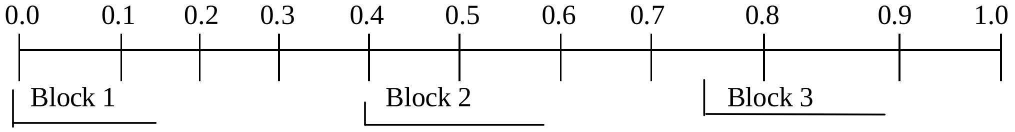
2.3 Co-ordinating in practice

2.3.1 Blocks

In practice, to coordinate, the same starting point and direction is used for several surveys. Surveys with the same starting point and direction are said to be in the same block. Consequently, one survey cannot be negatively co-ordinated with *every* other survey in the same block. Surveys in the same block are always positively co-ordinated. If this is not desirable, then surveys covering different industries or different size groups should be put together in the same block and surveys which are not to be positive co-ordinated, in different blocks. Then negative co-ordination is achieved automatically in these surveys. (The questionnaires for two positively co-ordinated surveys should be sent out on different dates.) It is advisable to think about how many blocks are needed, and where to put them on the random line, before

starting to draw co-ordinated samples. It is possible to change the design afterwards, but it always means less precision in the estimates of change over time for a couple of years.
Stats SA decided upon the use of three blocks where one block with starting point 0.0 and sampling direction to the right was decided upon, another block with starting point 0.4 and sampling direction to the right, and a third block with starting point 0.75 and direction to the right, see figure 2. The starting points are then well apart so the negative co-ordination will be achieved for small enterprises in surveys in different blocks. The response burden will then probably be evenly distributed over the "random line". See Appendix 2, for starting points of all surveys within Economic Statistics, using the BSF as sample frame.

Figure 2



2.3.2 A SAS-macro for drawing co-ordinated samples

A SAS macro, called draw_sample, for drawing co-ordinated samples was developed by Annika Lindblom of Statistics Sweden. In each stratum the macro draws a simple random sample from a specified population using the JALES-technique. This means using the permanent random number (PRN), a specified starting point and sample direction. In this macro, enterprises with PRN equal to the starting point are included in the sample if the direction to move to the left is specified. If the direction to move to the right is specified, then the sample selection starts with the enterprise with PRN immediately greater than the starting point. It could be the other way around, but the starting point should not be included in both directions.

The macro needs a few parameters specified:
%draw_sample(*dsn* = , *stratid* =, *sampsize* =, *direction* =, *stpoint* =, *prn* =);

> **dsn** = name of the data set containing the population from which the sample is to be drawn.
> **stratid** = name of the variable containing the stratum information.
> **sampsize** = name of the variable containing the desired sample size in each stratum.
> **direction** = sample drawing direction, ‘left’ or ‘right’.
> **stpoint** = starting point on the random line.
> **prn** = name of the variable containing the permanent random number (PRN).
> **stratum** = sic0||sizegrp.

The output from the macro is a SAS work data-set called sample, which contains the desired sample.

Below is an example on how to use the macro: (the macro is, in this example, called draw_samp.sas and it is stored in d:\methodology\sas programs)

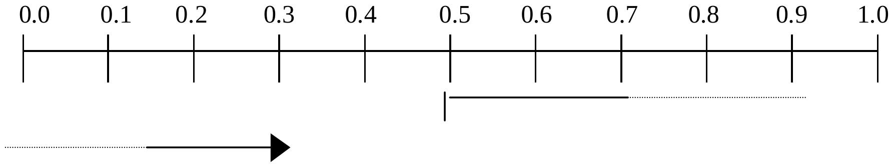
```
%include d:\methodooldgy\sas programs \draw_samp.sas;

%draw_sample(dsn = frame_SEE, stratid = stratum, sampsize = samplesize,
direction = right, stpoint = 0.5, prn = rnd_nbr);
```

Sample direction to the right means sorting the population of enterprises in ascending order, with respect to their random number, and starting the sample selection at the specified starting point. Sample direction to the left means sorting the population of enterprises in descending order, i.r.o. their random number, and starting the sample selection at the specified starting point.

There is, aspects namely the ”round the corner” problem, that needs special consideration.

2.3.2.1 The “round the corner” problem

Figure 3



Take for example, a starting point of 0.5 and the sampling direction to the right, then there may not be sufficient enterprises to select to the right of the starting point. Then the sample selection must continue to the right of the point 0.0, and this is a problem that should be solved, (Figure 2 illustrates the use).

2.3.2.2 Sample direction to the right

One simple way to solve this problem, used by the macro, is to duplicate every enterprise in the population, in such a way that the duplicated enterprises be given the original random number plus one. The enterprises in the population are sorted, within strata, in ascending order, i.r.o. their random number. This means that the original enterprises come, first and then the duplicated ones.

Then, when all enterprises to the right of the starting point are included in the sample the user continues the sample selection among the duplicated ones. In this way the user will have exactly the same enterprises included in the sample as if he or she had been able to continue the sample selection from the beginning of the original enterprises (as illustrated in example 1 and example 2 below).

Example 1: (within one specific stratum)

Original and duplicated population sorted with respect to ascending random number:

Table 7			**Table 8**	
Original population			Original plus duplicated population	
Enterprise	Random number		Enterprise	Random number
Ent 1	0.1		Ent 1	0.1
Ent 2	0.2		Ent 2	0.2
Ent 3	0.3		Ent 3	0.3
Ent 4	0.4		Ent 4	0.4
Ent 5	0.5		Ent 5	0.5
Ent 6	0.6		Ent 6	0.6
Ent 7	0.7		Ent 7	0.7
Ent 8	0.8		Ent 8	0.8
Ent 9	0.9		Ent 9	0.9
Ent 10	1.0		Ent 10	1.0
			Ent 1	1.1
			Ent 2	1.2
			Ent 3	1.3
			Ent 4	1.4
			Ent 5	1.5
			Ent 6	1.6
			Ent 7	1.7
			Ent 8	1.8
			Ent 9	1.9
			Ent 10	2.0

Assume that the starting point is 0.5 (immediately greater than), the sample direction is to the *right*, and number of enterprises included in the sample should be seven. This sample should then include Ent 6, Ent 7, Ent 8, Ent 9, Ent 10, Ent 1, Ent 2, as indicated in table 7. The user is required to go ”around the corner”.

2.3.2.3 Sample direction to the left

Assume that the starting point is 0.5 (included), the sample direction is to the *left* and the number of enterprises included in the sample should be seven. To be able to use sample direction to the left, the enterprises should be sorted, by strata, in descending order with respect to their random number.

Example 2: (one specific stratum)

Original and duplicated population sorted with respect to descending random number:

Table 9			Table 10	
Original population			Original plus duplicated population	
Enterprise	Random number		Enterprise	Random number
Ent 10	1.0		Ent 10	2.0
Ent 9	0.9		Ent 9	1.9
Ent 8	0.8		Ent 8	1.8
Ent 7	0.7		Ent 7	1.7
Ent 6	0.6		Ent 6	1.6
Ent 5	0.5		Ent 5	1.5
Ent 4	0.4		Ent 4	1.4
Ent 3	0.3		Ent 3	1.3
Ent 2	0.2		Ent 2	1.2
Ent 1	0.1		Ent 1	1.1
			Ent 10	1.0
			Ent 9	0.9
			Ent 8	0.8
			Ent 7	0.7
			Ent 6	0.6
			Ent 5	0.5
			Ent 4	0.4
			Ent 3	0.3
			Ent 2	0.2
			Ent 1	0.1

This sample should include Ent 5, Ent 4, Ent 3, Ent 2, Ent 1, Ent 10, Ent 9, as indicted in table 9, in order to go ”around the corner”. The user is required to duplicate the enterprises and sort them with respect to descending random number.

The sample selection starts at 1.5 (included) and continues among the original enterprises, as indicted in table 10. This sample will include Ent 5, Ent 4, Ent 3, Ent 2, Ent 1, Ent 10, Ent 9.

2.4 Rotation

It has to be noted that rotation is not yet implemented at Statistics South Africa but will be implemented when historical data becomes available on the enterprises sampled this year. Every survey is attached with a specific starting point and every enterprise is allocated an 11-digit random number, which will be needed in future to implement rotation.

Due to co-ordination over time, a selected unit may have to participate in a survey with a fixed starting point for many years. On the other hand, a unit (randomly) not

included in the sample has the advantage of not having to participate for many years. It is therefore advisable to have a system of rotation, whereby the selected units rotate out of the sample after a certain number of occasions. The number of occasions the units should participate in a survey is a balance between how many occasions the units need to participate, and what the decrease in the precision is of the estimates over time that is acceptable. Negative co-ordination, together with rotation spreads, the response burden as much as possible, and after a certain number of occasions the size group 3 and 4 enterprises in a survey will be completely new. Thus negative co-ordination and rotation will have an impact on the small enterprises. There is also another reason for rotation: units within size strata are of different sizes and a representative sample should be evenly distributed over the size strata. The sample that first was drawn for a survey could randomly include too many small (large) units in respect to the even distribution over the size strata. When estimates are calculated from such a sample, the estimates will be too low (high), it introduces an element of bias. This will be avoided, at least in the long run, if rotation is used.
Rotation works only if there is room for rotation. By room is meant that if a unit rotates out of a survey sample after the certain number of years, it should not immediately rotate into another sample. Such room is only available among small enterprises, possibly in size group 3 and 4. It should be noted that only small business can be rotated, large enterprises are all included in surveys.

2.4.1 Methods of rotation

Let us assume that we want the units to be included in the sample for five occasions, and in strata with inclusion probability less than 0.10 (i.e. $n/N < 0.10$, where n is the sample size and N is the total population), there will be room for rotation. There are two possible methods for rotation:

- Shifting the starting points
- Shifting the random numbers

If a unit with inclusion probability less than 0.10 is to be excluded from of a survey sample after five occasions, then the starting point must be shifted by 0.02 with every occasion. The same is achieved by shifting the random number with 0.02 with every occasion. After five occasions the starting point has shifted 0.1, or the unit has moved 0.1 on the “random line” and should be out of the sample.

The disadvantage of shifting the starting points (or random numbers) each occasion with the same length is that the rotation will vary considerably among strata. For example, in strata with a very small inclusion probability, the majority of the units in the sample will be renewed after one occasion. This is a disadvantage for the estimates of change over time. To achieve the same rotation in each stratum there must be individual shifts for each stratum. But to use individual shifts will, in the long run, destroy the positive and negative co-ordination among the surveys. The co-ordination among the surveys is maintained only if the same length of the shift is used for all surveys.

2.4.2 Random rotation groups

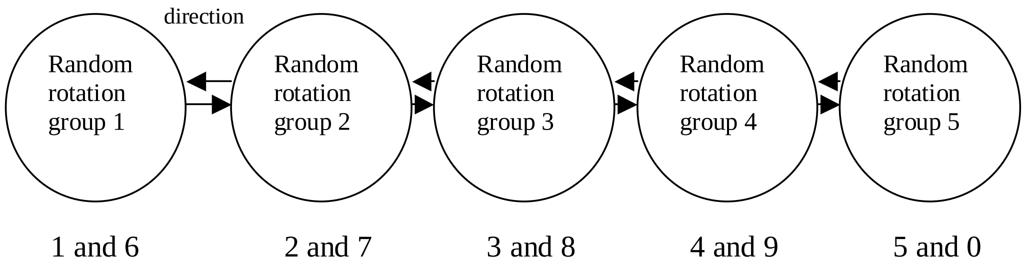
Grouping all units into five rotation groups can solve the problem of varying rotation. The starting points (or random numbers) are then shifted by 0.10 in one rotation group for every occasion. That is, all units in rotation group one will shift starting point (or random number) the first occasion. The second occasion all units in rotation group two will shift and so on. This method makes the rotation in all strata with inclusion probability less than 0.10, on the average 1/5. This method also ensures the negative and positive co-ordination between surveys.

The random number is used to decide the rotation group into which a unit should be assigned, e.g. the last digit in the random number, i.e. the units are randomly assigned to a rotation group.

Co-ordination among surveys based on different kind of units can also be managed in this way. A single-location enterprise and its geographical unit will be assigned to the same rotation group, because they have the same random number. To use the random rotation group method together with shifting starting points would mean using two starting points in each stratum: one initial starting point and one where 1/5 of the units move each occasion. Two starting points for each stratum are not easy to monitor in a complex system. It is easier to shift the random numbers. This method of shifting random numbers is used by Statistics Sweden and this method will also be applied in all economic surveys at Stats SA.

In figure 4 below, the random rotation line is divided into five groups where the random rotation group 1 will contain the enterprises with a permanent random number (PRN) ending on a 1 or 6, and random rotation group 2 will contain the enterprises with a PRN ending on a 2 or 7, etc. It is possible to start at any of the rotation groups as long as a logical sequence is followed. New enterprises birthed during that specific year gets assigned a PRN and the enterprise enter into a rotation random group (RRG), however the enterprise should be fixed for that specific period of rotation even if it falls into the group that should be rotated.

Figure 4



Every enterprise in the BSF is randomly assigned to one of 5 rotation groups (e.g. by using the last digit in the random number, e.g. digits 1 and 6 for the first rotation group, digits 2 and 7 for the second rotation group, etc.). Births are assigned to the appropriate rotation groups as they enter the sampling frame. After the first occasion all enterprises in rotation group 1 are shifted (either the starting point or the random numbers, preferably the random numbers) 0.2 in one direction, after the second occasion all units in rotation group 2 are shifted 0.2 in the same direction, etc. It can then be expected that the vast majority, if not all, enterprises (with inclusion probability less than, say 0.1) in a RRG will be out of the sample after 5 occasions, resulting in an expected rotation rate of 0.2. (Note: For enterprises with larger inclusion probabilities the rotation will be slower, for enterprises with an inclusion probability equal to 0.2 it is to be expected that 50% of these enterprises will be out of the sample after 5 occasions.) It is recommended that the RRG method be uniformly applied to all enterprises in those strata that are to be sampled, although the rotation will be very slow in strata with large inclusion probabilities.

2.5 Randomness

The JALES-technique ensures a simple random sample, but does not guarantee that a specific enterprise will be included, only in a certain number of surveys. It is difficult to make this kind of guarantee without destroying randomness. The same applies to the method for rotation described above; it is impossible to guarantee that an enterprise will be excluded from the sample in a certain number of occasions. One way to be able to give this kind of guarantee is to give an enterprise a zero inclusion probability. But then there are other problems, for instance calculations of the point and standard error estimates for the whole population of enterprises exclude this enterprise.

3. Using information from the surveys on correct economic activity as domains and estimation of population parameters

3.1 Using information from the surveys on correct economic activity as domains

3.1.1 Currently used methods

Currently, at Stats SA, the survey areas consider the incorrectly classified enterprises as:

1. Non-responses and they impute them from the average in stratum; or
2. Use the incorrectly classified enterprises as correctly classified and they handle them like all the other enterprises in the industry.

The non-response method is based on the assumption that over-coverage and under coverage evens out, imputing the over coverage is a way to compensate for the under coverage. There are, of course, enterprises incorrectly classified into other industries that should be correctly classified into the industry where the over- coverage was found. The method to keep the incorrectly classified enterprises in the original industry and handle them as the other enterprises in the industry is also based on the assumption that over coverage and under coverage even themselves out.

The method to impute an incorrectly classified enterprise with the average in the stratum is often preferable, if the enterprise is wrongly classified in an industry. For example, if a manufacturing enterprise is incorrectly classified into retail trade, it is likely that it is impossible to keep the original information from that enterprise in retail trade.

These current methods should, however, change as discussed below.

3.1.2 Using domains

The classification in the BSF, in terms of economic activity, could sometimes be incorrect. In the majority of the economic surveys the enterprises included in the samples are asked, in the questionnaires, about their main activity. This information makes it possible for the survey areas to classify the enterprises in the samples into the correct economic activity.

Whether to use the correct information on economic activity as domains or not is a balance between smaller coverage error and larger variance. In practice there are probably a few important incorrectly classified enterprises having a large impact on the coverage error. And there are, as well, a lot of incorrectly classified enterprises having a small impact on the coverage error, contributing mainly to a larger variance. One way to take this balance into account is to combine the methods for handling incorrectly classified enterprises. We suggest that Stats SA creates a new variable, which is to be used as information on domain activity. For the majority of the enterprises in the sample the domain activity is the same as the stratum activity from the BSF. But for important incorrectly classified enterprises the domain activity is the information collected from the survey. Stats SA should use this new variable to give information on economic activity, when the result is to be presented. Note that the estimation program should be able to handle domains cutting through strata, in order to have correct variance estimates. (For technical details on domain estimation, see chapter 10 in Särndal, Swensson, Wretman 1992).

With both methods discussed in 3.1.1, correct information on economic activity is thrown away. To avoid this, it is possible and desirable to use the information on economic activity collected in the questionnaires as domains. This information could be used in the same way as all the other variables collected in the questionnaires, such as like gender, province etc. introduce a new domain activity variable. If the reported activity corresponds to the activity indicated on the BSF, then the domain activity will be the same. However, if these activities differ, then the domain activity is taken as the reported activity. *If the correct information on economic activity is used as domains the precision, in terms of coverage, in the point estimates will increase. But everything has a price, and in this case the price will be larger variances. This is due to the fact that the domains based on the correct information on economic activity will cut through the original strata, based on the information from the BSF.*

3.1.3 Economic activities not covered by the survey

The method of using the correct information on economic activity as domains is straightforward, if the survey covers the whole economy, but if the survey excludes certain industries, then this method is problematical. If, for example, a survey covers all industries except agriculture, it is impossible, with information from the survey, to find enterprises incorrectly classified into agriculture. This means that it is not correct to use information from the survey to classify enterprises into agriculture. It is impossible to use information from the survey to classify enterprises from agriculture into other activities. This would lead to under coverage in the industries covered by the survey. If the correct economic activity for an enterprise is not covered by the survey then it is recommended to use one of the methods currently in use, as described in section 4.1.1, i.e. to impute the enterprise with the average in the stratum or keep it in the incorrect activity.

3.1.4 Non-response

It is very important that the large enterprises in the completely enumerated strata answer the questionnaire, because they have a large impact on the estimates on a national level. But if it is impossible to get an answer from one of these enterprises, then the imputation must be done individual. Enterprises in size groups 2, 3 and 4 have relatively smaller or small impacts on the estimates and therefore they can be imputed by the average in the respective stratum. This is achieved by replacing the number of selected enterprises in the stratum is replaced with the number of responding enterprises in the stratum, with the weight (N_h/n_h). This will impute the non-responding enterprises by the average in the stratum. This method works best when the enterprises in the stratum are more similar in MOS. This is one reason to have two size groups for enterprises with a smaller turnover.

3.1.5 Feed back from the surveys

Feedback from survey areas should not be used in the BSF maintenance procedures. If the technique with permanent random numbers is used to obtain a large overlap between two subsequent samples for one survey, then the enterprises in the BSF, which are included in samples, are updated, and the enterprises only included in the BSF are not. The large overlap between two subsequent samples means that you receive almost the same sample the next time a sample is drawn, and that this sample is always updated. In turn, this means that the estimates will be biased, because the sample is no longer representative of the whole population. For size group 1 enterprises the sample and the population is the same feedback could be used by survey areas to the BSF but only as an indicator to prioritise future quality improvements in the BSF through profiling of these large enterprises.

> In the case of large and complex in structure enterprises a personal visit is made to the group enterprise to determine the structure of the enterprises residing within the group. This process is referred to as ***profiling*** and delineation and briefly encompasses the finding and linking of the smallest set of legal units in the administrative registers together to form an autonomous and complete production unit from which statistically relevant information can be collected whereas delineation is the division of the enterprise into statistical sub-units.

4. Estimation of population parameters in the case of a single sample

CLAN is generally used to calculate point and standard error estimates. CLAN is a very flexible SAS program and can be used to calculate point- and standard error estimates for all parameters such as sums, means and ratios, see Andersson, Nordberg 1998. CLAN can handle domains cutting through strata, which is important at Stats SA. There is not information on all desired classification variables in the BSF, which means that it is impossible to stratify on all domains, like detailed level of siccode.

Currently at Stats SA, principally point and standard error estimates for totals and averages are calculated. For this kind of ”simple” parameters either CLAN or version 8.2 SAS surveymeans procedure could be used. But if point and standard error estimates concerning more complicated parameters should be calculated, then CLAN is preferable.

Either the Swedish procedure CLAN or the SAS version 8.2 procedure “surveymeans” can be used to estimate the values of population parameters together with their standard errors and 95% confidence intervals in the case of a single survey. Both procedures are discussed below. For a detailed explanation on the use of CLAN, see the CLAN manual[9].

4.1 The SAS (version 8.2) procedure “surveymeans”

This procedure produces estimates of survey population means, totals and ratios with their 95% confidence limits from sample survey data. The procedure also produces variance estimates, confidence limits, and other descriptive statistics. The sample design is taken into account in the calculation of these estimates.

The following statements are available in Proc Surveymeans:

BY variables; (The data set must be pre-sorted i.r.o. these variables)
CLASS variables;
CLUSTER variables;
DOMAIN variables ;
STRATA variables / options;
VAR variables;
WEIGHT variable;
RATIO variables;

The procedure uses the OUTPUT DELIVERY SYSTEM (ODS) to place results in output data sets.

The general steps for analysing an economic series sample (assuming independent simple random samples drawn from the industry primary strata by size group strata) are as follows:

```
Proc surveymeans data=aaa  all;  /* all = all statistics */
  Strata sic sizegrp;  /*or stratum (if stratum = sic0 || sizegrp);*/
  Domain turnover;
  Var turnover;
  Weight samplingweight;
  Ods output DOMAIN=Estimates;
  Run;
```

The data set aaa must contain the weight variable. “Sic” indicates the siccode strata (primary strata) of the industry primary strata to be analysed separately and

[9] Claes Andersson and Lennart Nordberg: A User’s Guide to Clan 97. Statistics Sweden.

“sizegroup” indicates the (measure of) size groups per industry secondary strata to be analysed separately. The variable “sic” should consist of the same number of characters or digits for all siccodes considered. ”Estimates” is the name of the output SAS file.

4.2 Swedish procedure CLAN

The example below illustrates the CLAN output of a sample drawn for the EAS survey. It was drawn as a sequential simple random sample without replacement, using the permanent random numbers in the BSF. The starting point was 0.4 and the direction was to the right. The total turnover for different activities in the BSF was estimated by this sample.

Point and standard error estimates were calculated for various domains. Sometimes strata and domains coincide. The standard errors calculated on the study variables in the survey will of course be **much** larger since the calculations below are based on a 100 % response rate and on the same variable as the strata were created with.

Table 10 Coefficient of Variation (CV) ≡ (RSE) and Relative Precision (RP) per sector

SIC	row	ptturnr	stturnr	pmean2	smean2	CV	RP%
2	1	110647914527	544646453	110647915	544646	0.004922	0.964778281
3	2	544869393997	5265386299	17568498	169774	0.009664	1.894060716
4	3	5967595723	50316964	23966248	202076	0.008432	1.652612781
5	4	75661212742	793870546	5467246	57365	0.010492	2.05651775
61	5	327939698732	3377067019	17287280	178021	0.010298	2.018374532
62	6	187451523742	1825671796	6817163	66395	0.009739	1.908929119
63	7	133055371031	1322978321	12172296	121030	0.009943	1.948840914
64	8	23259473398	238909462	3120820	32055	0.010271	2.013212157
7	9	130732546490	1388059388	18802322	199635	0.010618	2.081039859
8	10	769074541318	7898067954	12807023	131523	0.01027	2.012836514
9	11	156699237524	1744602664	9586397	106730	0.011133	2.182155622

Where CV and RP% is calculated in the SAS programs. Below the extract of the SAS program:

```
data dut;
      set dut(drop=col);
      cv=stturnr/ptturnr;
      rp=(cv*1.96)*100;
```

See appendix 2 for the SAS program used in the estimation of “imputed” size group 1 cases after the survey has been conducted.

The standard error will then be calculated as follows:

SE = RSE (as calculated in CLAN without imputed values)*(CLAN total + imputed).

CHAPTER 3

PRACTICAL APPLICATION

In this chapter actual steps will be described and elaborated upon which should be followed in the design of a survey (the design of the questionnaire and the finalisation thereof is not discussed as part of this manual). Below is a summary of the actual sampling process.

The sampling approach for Economic surveys: the following steps were followed in the sampling process:

- Decide on the amount of money available for the survey (probably the most important determinant); document the budget constraints of the specific survey, although a good precision for every sample design should be achieved.
- Initially, there is no pre-decision made in respect of the total sample size, a guideline i.r.o the sample size is the amount of money allocated for the specific survey.
- Decide on primary stratum industry levels for which an acceptable precision is required.
- Decide on the number of SIC digits per industry for the determination of size group cut-off points (secondary stratum).
- The determination of the turnover cut-off points to be used for subdividing the enterprises into non-overlapping groups, per industry Standard Industrial Classification (SIC) code level into the measure of size (MOS) intervals based on turnover, called strata (viz. the cross-classification of the industry strata (primary strata) with the measure of size intervals (secondary strata)). The industry strata (primary strata) are defined as the pre-specified SIC code levels (i.e. 1-digit and/or 2-digits and/or 3-digits) of the different industry sectors to be taken into account in the design of the sample, i.e. in respect of which an "adequate" or "sufficient" representation in the sample is required. Enterprises within each primary stratum will be grouped into 4 non-overlapping size groups (secondary strata) based on the measure of size (MOS). The SIC code levels of these strata (primary strata) vary between different economic surveys, depending on the purpose of the survey. The measure of size intervals or groups per industry secondary strata constructed in such a way that size group 1 contains the largest enterprises in that strata. All enterprises in these first size groups of the different strata's are included in the sample.
- Divide the total turnover of all enterprises with turnover between L and U into three approximately equal intervals. The rounded turnover value of the enterprise "nearest" to these division points in the ascending ordering of enterprises with turnover between L and U is then to be used as the "in between" turnover cut-off values – percentile method. The cut-off of the upper U1 cut-off value is calculated through the use of a percentile method.

- Keeping the total sample size in mind, decide on preliminary sample size per stratum and the sample size to be allocated to sizegroup 1 per primary stratum (leaving "sufficient" sample size to be allocated to the other size groups).
- Determine size group (turnover) cut-off values U2 and U3 for each primary stratum, viz.

Size group 1:	Turnover $\geq$ U1
Size group 2:	U2 $\leq$ turnover < U1
Size group 3:	U3 $\leq$ turnover < U2
Size group 4:	R300 000 $\leq$ turnover < U3.

NOTE: An initial decision has to be taken on the values of U1 for different primary stratum based on: the proportions of total turnover contribution by enterprises in size group 1 and the number of enterprises included in these size groups 1. Each primary stratum (siccode) has its own U1, U2 and U3 values.

- Run a SAS program which determines, per primary stratum, the lower cut-off value U1 of size group 1 for a series of proportions of total turnover contribution by the (large) enterprises in size group 1, ranging from 0.50 to 0.95 by 0.05 or 0.025. The number of enterprises in these size groups 1's are also obtained in each case.
- Take an initial decision on the values of U1 cut-off point for the different primary stratum industry levels based on:
    - The proportions of total turnover contribution by enterprises in size group 1.
    - The number of enterprises included in these size groups 1.
    - And previous experience of the number of enterprises that were included in a specific survey for a given budget, allowing enough enterprises for inclusion in the other size groups.
- Adjust turnover cut-off values and/or the allocated sample sizes (value of N/n) for those primary stratum industry levels for which unsatisfactory cut-off values for size group 1 were obtained.
- Run a "combined" SAS program, which
    - Calculates the size group cut-off values U2 and U3 per cut-off industry level programmatically, using the percentile method.
    - Group all enterprises into the 4 size groups per primary stratum industry levels.
- Increase all sizegroup sample sizes to a minimum population size of 10 (or take all enterprises). If the size of the allocated sample to a size group is less than, say, 10, it is increased to a minimum value of 10 with the view to take the possibility of some non-response in the actual survey into account.
- Subtracted the size group 1 enterprises from the sample and determine remainder of allocated sample (per primary stratum industry level sizes) to its size groups, using Neyman optimal allocation. *Note:* at this point in time the number of enterprises in size group 1 is known (determined). The number of enterprises available per primary stratum for allocation to size group 2, 3 and 4 should now be determined. The proportion allocation to size groups should also be checked to ensure that an even spread in all size groups are achieved in the allocation. Apply the Swedish macro "%neymann" to determine per primary stratum industry level the sample sizes needed for a series of 10 pre-specified relative precision (RP)

values, starting at 0.1% or any other starting precision values, increasing each unit with the same increment as a percentage (%).

- Decide on the preliminary sample size to use for each primary stratum industry level, based on the RP values.
- Adjust the various parameters until, for an affordable sample size, satisfactory RP values were obtained for primary stratum industry levels as well as for all, pre-determined domains. Note: the above assume 100% response.
- Finally, from all secondary strata (primary strata industries by size groups) draw a simple random sample (SRS) of the required size using the JALES sampling technique. The JALES sampling technique should be used to draw the sample using the allocated starting point to the specific survey (see appendix 2).
- After the survey has been conducted use the Swedish macro "%CLAN" to calculate the estimated RP values. Calculate relative precision (RP) of total turnover estimates per primary stratum industry level by using CLAN.

Guideline criteria to use

- 60% of total turnover (if obtainable) per primary stratum to be contributed by sizegroup 1 enterprises.
- ± 50% (but not more than 70%) of enterprises of primary stratum in size group 1.
- Minimum number per size group 2, 3 and 4 should be 10 or use the population size if less.

References

Ohlsson (1992) SAMU - *The system for Co-ordination of Samples from the Business Register at Statistics Sweden*. R&D Report, Statistics Sweden, 1992:18

Lindblom (2000) *Drawing Samples from the Business Frame at Statistics South Africa*. Mission report RSATAT 2000:9, Annika Lindblom, November 13, 2000

Lindblom (2001) *Sample Co-ordination and Survey Design for Three Business Surveys*. Mission Report RSATAT 2001:7, Annika Lindblom, August 14, 2001

Särndal, Swensson, Wretman (1992) *Model-Assisted Survey Sampling*, New York: Springer-Verlag

Andersson, Nordberg (1997) *A User's Guide to CLAN 97*: A SAS-program for computation of point- and standard error estimates in sample surveys

Stoker, D.J. (2002) Workshop at the Economic Summit, 2002

Stoker, D.J (2002) Development of SAS programs, U1 values a combined "procedure", 2002

Stoker, D.J (2002) New design of the business surveys of Statistics South Africa, Statistics South Africa, 2002

Appendix 1

Note the macros for Neyman allocation, draw and CLAN is not included in this document because it is standardised for all survey areas.

SAS program[10]

```
   *********************************************************
  *     Drawing of SRS using Annika's approach sample      *
  * The EAS sample redesign using frame300 w.o. duplicates *
  *        Calculating the U1 cut-off values               *
  *      based on the calculation of percentiles           *
  *                   24 July 2002                         *
  **********************************************************;

Libname s 'D:\sample design 2002\BSF020724\';
Libname s 'D:\sample design 2002\';

data frame1;
 set s.BSF020724;  by sic1 sic2 sic3;
  /* Note that enterprises with "invalid" codes are not deleted */
 if sic2 in ('91','92') then delete;
 drop sic_code;
run;

data frame1A;
  set frame1; by sic1 sic2 sic3;
  if sic1='0' or sic1='1' or sic1='2' or sic1='3' or sic1='4' or
       sic1='5' or sic1='6' or sic1='7' or sic1='8' or sic1='9'
    then sic0=sic1+0;
  if sic2='61' or sic2='62' or sic2='63' or sic2='64'
    then sic0=sic2+0;
run;
proc sort data=frame1A; by sic0 turnover; run;

proc univariate data=frame1A noprint;
  var turnover;
  output out=percentiles pctlpre=P_  pctlpts=5 to 50 by 2.5;
  weight turnover;
  by sic0;
run;

data global;
  merge frame1A percentiles; by sic0;
run;


proc means data=global N sum noprint;
  var turnover;
  output out=count0 N=cnt0 sum=sum0;
  by sic0;
run;


data count95;
  set global; by sic0;
  if turnover>=P_5;
```

[10] Developed by D.J Stoker.

```
  proc univariate noprint;
    var turnover;
      output out=c95 N=cnt95;
    by sic0;
run;

data count90;
  set global; by sic0;
  if turnover>=P_10;
  proc univariate noprint;
    var turnover;
      output out=c90 N=cnt90;
    by sic0;
run;

data count85;
  set global; by sic0;
  if turnover>=P_15;
  proc univariate noprint;
    var turnover;
      output out=c85 N=cnt85;
    by sic0;
run;

data count80;
  set global; by sic0;
  if turnover>=P_20;
  proc univariate noprint;
    var turnover;
      output out=c80 N=cnt80;
    by sic0;
run;

data count75;
  set global; by sic0;
  if turnover>=P_25;
  proc univariate noprint;
    var turnover;
      output out=c75 N=cnt75;
    by sic0;
run;

data count70;
  set global; by sic0;
  if turnover>=P_30;
  proc univariate noprint;
    var turnover;
      output out=c70 N=cnt70;
    by sic0;
run;

data count65;
  set global; by sic0;
  if turnover>=P_35;
  proc univariate noprint;
    var turnover;
      output out=c65 N=cnt65;
    by sic0;
run;

data count60;
```

```
  set global; by sic0;
  if turnover>=P_40;
  proc univariate noprint;
    var turnover;
      output out=c60 N=cnt60;
    by sic0;
run;

data count55;
  set global; by sic0;
  if turnover>=P_45;
  proc univariate noprint;
    var turnover;
      output out=c55 N=cnt55;
    by sic0;
run;

data cnt50;
  set global; by sic0;
  if turnover>=P_50;
  proc univariate noprint;
    var turnover;
      output out=c50 N=cnt50;
    by sic0;
run;


data s.global_cnt;
  merge percentiles c95 c90 c85 c80 c75 c70 c65 c60 c55 c50;
    by sic0;
run;

PROC EXPORT DATA= s.global_cnt
            OUTFILE= "D:\sample design 2002\global.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;

  **********************************************************
  *                                                        *
  *     Drawing of SRS using Annika's approach sample      *
  * Reading the U1 cut-off values from an external file    *
  *                     5 July 2002                        *
  *                                                        *
  **********************************************************;

Libname BSF 'D:\Sample design 2002\BSF020828\';

data frame1;
 set BSF.BSF020724;  by sic1 sic2 sic3;
 if sic2 in ('91','92') then delete;
 run;

data frame1A;
  set frame1; by sic1 sic2 sic3;
  if sic1='0' or sic1='1' or sic1='2' or sic1='3' or sic1='4' or
      sic1='5' or sic1='6' or sic1='7' or sic1='8' or sic1='9'
    then sic0=sic1+0;
  if sic2='61' or sic2='62' or sic2='63' or sic2='64'
    then sic0=sic2+0;
run;
proc sort data=frame1A; by sic0 turnover; run;
```

```
PROC IMPORT OUT=work.cut_off
     DATAFILE= "D:\Sample design 2002\eas2002\U1_values.xls"
     DBMS=EXCEL2000 REPLACE;
  GETNAMES=YES;
RUN;

proc sort data=BSF.cutt_off; by sic0; run;

data frame1B;
  merge frame1A BSF.cutt_off; by sic0;
  if turnover<U1;
run;


**** PERCENTILE METHOD ****

proc univariate data=frame1B noprint;
  var turnover;
  output out=triads pctlpre=P_  pctlpts=33 to 66 by 33;
  weight turnover;
  by sic0;
run;

data frame1C;
  merge frame1A triads work.cut_off; by sic0;
  U3=P_33;
  U2=P_66;
run;


**** EXPONENTIAL METHOD ****

/*
data cutpnts;
  set BSF.cutt_off;
  U3=300000*1.5*(U1/300000)**(1/3);
  U2=300000*1.5*(U1/300000)**(2/3);
run;
proc sort data=cutpnts; by sic0; run;

data frame1C;
  merge frame1A cutpnts; by sic0; run;
run;
*/

Data frame2;
 set frame1C; by sic0;
     if 300000 <= turnover < U3 then sizegrp = '4';
     else if U3 <= turnover < U2 then sizegrp = '3';
     else if U2 <= turnover < U1 then sizegrp = '2';
     else if turnover >= U1 then sizegrp = '1';
run;
proc sort data=frame2; by sic1 sic2 sic3; run;


 /*
  Proc means data=frame2;
    class sic3 sizegrp;
    var turnover;
    output out=help N=Count mean=Average stddev=Sdev
                      min=Minimum max=Maximum;
  run;

 PROC EXPORT DATA=HELP
      OUTFILE= "D:\ekonstat\eas2002\Sic3A_1_short.xls"
      DBMS=EXCEL2000 REPLACE;
 RUN;
 */

   %include 'D:\methodology\sas programs\neymann.sas';

   Data frame3;
     set frame2;
     stratum = sic3 || sizegrp;
   run;
   proc sort data=frame3; by stratum; run;

  %neymann(niva=pop, dsn = frame3, stratum = stratum,
           allvar=turnover,
           villkor = (sizegrp in ('1')), bg = sic0, startpre = 0,
           inc = 0.2, lillmin = 10);

   data result (keep = stratum bigN sampsize);
     set ut;
       rename n5 = sampsize storan = bigN;
   run;

   Data frame4;
     merge frame3(in=a) result(in=b); by stratum;
       if a and b;
   run;


  Proc sort data=frame4; by random_nbr; run;

  %include 'd:\methodology\sas programs\draw_macro_feb02.sas';


  %draw_sample(dsn = frame4, stratid = stratum, sampsize = sampsize,
               direction = right, stpoint = 0.4, prn = random_nbr);

  %INCLUDE 'D:\methodology\sas programs\clan97.sas';

 %MACRO FUNCTION(r,c);
   IF &r =  1 THEN rv = (sic1 = '2');
   IF &r =  2 THEN rv = (sic2 = '30');
   IF &r =  3 THEN rv = (sic2 = '31');
   IF &r =  4 THEN rv = (sic2 = '32');
   IF &r =  5 THEN rv = (sic2 = '33');
   IF &r =  6 THEN rv = (sic2 = '34');
   IF &r =  7 THEN rv = (sic2 = '35');
   IF &r =  8 THEN rv = (sic2 = '36');
   IF &r =  9 THEN rv = (sic2 = '37');
   IF &r = 10 THEN rv = (sic2 = '38');
   IF &r = 11 THEN rv = (sic2 = '39');
   IF &r = 12 THEN rv = (sic1 = '3');
   IF &r = 13 THEN rv = (sic1 = '4');
   IF &r = 14 THEN rv = (sic3 = '501');
   IF &r = 15 THEN rv = (sic3 = '502');
   IF &r = 16 THEN rv = (sic3 = '503');
   IF &r = 17 THEN rv = (sic3 = '504');
   IF &r = 18 THEN rv = (sic3 = '505');
```

```
   IF &r = 19 THEN rv = (sic1 = '5');
   IF &r = 20 THEN rv = (sic2 = '61');
   IF &r = 21 THEN rv = (sic2 = '62');
   IF &r = 22 THEN rv = (sic2 = '63');
   IF &r = 23 THEN rv = (sic2 = '64');
   IF &r = 24 THEN rv = (sic1 = '6');
   IF &r = 25 THEN rv = (sic1 = '7');
   IF &r = 26 THEN rv = (sic1 = '8');
   IF &r = 27 THEN rv = (sic1 = '9');

   %TOT(tturnr, turnover, RV);
   %TOT(numbcr, 1, RV);
   %TOT(number, 1, 1);
   %DIV(mean2, tturnr, numbcr);
   %ESTIM(tturnr, tturnr);
   %ESTIM(mean2, mean2);
 %MEND;
 %CLAN(DATA = sample, STRATID = stratum, NPOP = bigN,
              NRESP = sampsize, MAXROW = 27, MAXCOL = 1);
 RUN;
data dut;
      set dut(drop=col);
      cv=stturnr/ptturnr;
      rp=(cv*1.96)*100;

proc print data=dut;
title;
run;
 PROC EXPORT DATA= work.dut
    OUTFILE= "D:\Sample design 2002\eas2002\ExampleB_clan_short.xls"
    DBMS=EXCEL2000 REPLACE;
 RUN;

 PROC EXPORT DATA= work.sample
    OUTFILE= "D:\Sample design 2002\eas2002\samplesizeB.xls"
    DBMS=EXCEL2000 REPLACE;
 RUN;
```

Appendix 2

Using CLAN to estimate the results after the survey has been conducted. Calculating estimates before re-classification and after re-classification. Estimation of “imputed” size group 1 cases after the survey has been conducted.

SAS program[11]

```
/*******************************************************************
* QFS 2002                                                         *
*                                                                  *
* Estimation, using CLAN97.                                        *
*                                                                  *
* Includes imputation.                                             *
*                                                                  *
********************************************************************/

options nocenter nodate nonumber;
options orientation=landscape;

libname qfs 'd:\qfs';

%let a=200206; *Specify the period! --------------------------------
---;
%let b=sicRep1; * Specify the reporting variable! -------------------
---;
*Use sicRep1 for: Estimates BEFORE reclassification.;
*Use sicRep2 for: Estimates AFTER reclassification.;
%let c=before; *Choose one: Before or after! ------------------------
---;
%let d=estim1; *Specify the file (estim1 or estim2)! ----------------
---;
*Use estim1 if b=sicRep1, and estim2 if b=sicRep2.;

/*
data test;
input sicRep1 stratum popSize capCnt qfsTurn db_int;
cards;
3 31 2 2 10000 .
3 31 2 2 20000 500
3 32 20 2 6000 .
3 32 20 2 4000 100
;
run;
*/
data samcap;
      set qfs.x&a(drop=weight);
*Impute for non-respone (only size group 1 units)--------------------
---;
      if cap ne 'y' and sizeGrp='1' then do;
            qfsTurn=round(bfTurnover/4); *Use the BF turnover;
            *Assumes that the other items are zero???;
            impute='y';
            label impute='Impute status';
      end;
run;
/*
```

[11] Developed by Quality and Methodology.

```
title;
proc freq data=samcap;
      tables sizeGrp*impute / norow nocol;
run;
*/
proc means data=samcap noprint;
      where (cap='y' or impute='y') and sample='ori';
      class stratum;
      var qfsTurn;
      output out=count(rename=(_freq_=capCnt)) n=nturn nmiss=nmturn;
run;

data samcap2;
      merge samcap
            count(keep=stratum capCnt _type_);
      by stratum;
      if _type_=1;
      drop _type_;
      label capCnt='Captured units count';
      *if capCnt<=0; *Test for total non-response--------------------
-----;
run;

data test;
      set samcap2;
      where sizeGrp='1' and sample='ori' and capCnt ne samSize;
      if stratum=lag(stratum) then delete;

proc print data=test noobs n label;
var sicStr samsize capCnt;
sum samsize capCnt;
title "QFS-&a: The count is not the same as the sample size - only
      size group 1";
run;

proc datasets;
      delete count samcap test;
run;

/*Creating 3 different datasets, a data set containing only the
imputed values and a dataset containing only the original sampled
units without any imputations*/

data orirec oriimp add;
      set samcap2;
      where cap='y' or impute='y'; *Captured or imputed units;
      if sample='add' then
            output add; *Units added after sample selection;
      else if (usi='21' or impute='y') and sizeGrp='1' then
            output oriimp; *The original sampled units - imputed;
      else
            output orirec; *The original sampled units - not imputed;
run;

%include 'd:\qfs\CLAN97.sas';

%macro function(r,c);
      if &r =  1 then rv = (&b = '2');
      if &r =  2 then rv = (&b = '3');
      if &r =  3 then rv = (&b = '4');
      if &r =  4 then rv = (&b = '5');
      if &r =  5 then rv = (&b = '6');
      if &r =  6 then rv = (&b = '7');
      if &r =  7 then rv = (&b = '8');
      if &r = 8 then rv = (&b = '9');
      %tot(turn, qfsTurn, rv);
      %tot(int, db_int, rv);
      %tot(tax, db_tax, rv);
      %tot(crnet, cr_net, rv);
      %tot(dbnet, db_net, rv);
      %tot(cpxn, cpxn_tot, rv);
      %tot(bvland, bv_land, rv);
      %*tot(num, 1, rv); *Number of units per row;
      %*tot(numAll, 1, 1); *Total number of units;
      %*div(average, emp, num);
      %estim(turn, turn);
      %estim(int, int);
      %estim(tax, tax);
      %estim(crnet, crnet);
      %estim(dbnet, dbnet);
      %estim(cpxn, cpxn);
      %estim(bvland, bvland);
      %*estim(num, num);
      %*estim(numAll, numAll);
      %*estim(average, average);
%mend;

%clan(data=orirec, stratid=stratum, npop=popSize, nresp=capCnt,
      maxrow=8, maxcol=1);

data out;
      set dut(drop=col);
      rse=sTurn/pTurn*100;
      if row=1 then &b='2';
      if row=2 then &b='3';
      if row=3 then &b='4';
      if row=4 then &b='5';
      if row=5 then &b='6';
      if row=6 then &b='7';
      if row=7 then &b='8';
      if row=8 then &b='9';
      drop row;

proc print data=out noobs n label;
*var &b pTurn sTurn rse pInt pTax pCrnet pDbnet pCpxn pBvland;
var &b pTurn sTurn rse;
sum pTurn;
title "QFS-&a: Estimates &c reclassification - the original sampled
      units (not imputed)";
run;

proc summary data=oriimp;
      class &b;
      var qfsTurn db_int db_tax cr_net db_net cpxn_tot bv_land;
      output out=out2 sum=iTurn iInt iTax iCrnet iDbnet iCpxn
                          iBvland;
run;

proc print data=out2 noobs n label;
var &b _type_ _freq_ iTurn;
title "QFS-&a: Estimates &c reclassification - the original sampled
      units (imputed)";
```

```
run;


proc summary data=add;
      class &b;
      var qfsTurn db_int db_tax cr_net db_net cpxn_tot bv_land;
      output out=out3 sum=aTurn aInt aTax aCrnet aDbnet aCpxn
                          aBvland;
run;

proc print data=out3 noobs n label;
var &b _type_ _freq_ aTurn;
title "QFS-&a: Estimates &c reclassification - units added after
      sample selection";
run;

data &d;
      retain &b period pTurn sTurn rse iTurn aTurn cilTurn fTurn se
             ciuTurn;
      label period='Period'
            pTurn='Point estimate of turnover'
            sTurn='SE of turnover'
            rse='RSE of turnover'
            iTurn='Imputed turnover'
            aTurn='Addisional turnover'
            cilTurn='Lower CI of final turnover'
            fTurn='Final turnover'
            se='SE of final turnover'
            ciuTurn='Upper CI of final turnover';
      merge out
            out2(where=(_type_=1))
            out3(where=(_type_=1));
      by &b;
      drop _type_ _freq_;
      period="&a";
      fTurn=sum(pTurn,iTurn,aTurn);
      /*Σ the turnover of the original dataset, the imputed
      dataset.*/
      se=(rse/100)*(fTurn);
      /*The standard error will then be calculated as follows:
      SE = RSE (as calculated in CLAN without imputed values)*(CLAN total +
      imputed).*/
      cilTurn=round((fTurn-(1.96*se))/1000000);
      ciuTurn=round((fTurn+(1.96*se))/1000000);
      rse=round(rse,.1);
      array x{6} pTurn sTurn iTurn aTurn fTurn se;
      do i=1 to 6;
            x{i}=round(x{i}/1000000);
      end;
      drop i;
      fInt=round((sum(pInt,iInt,aInt))/1000000);
      fTax=round((sum(pTax,iTax,aTax))/1000000);
      fCrnet=round((sum(pCrnet,iCrnet,aCrnet))/1000000);
      fDbnet=round((sum(pDbnet,iDbnet,aDbnet))/1000000);
      netProf=sum(fCrnet,-fDbnet);
      fCpxn=round((sum(pCpxn,iCpxn,aCpxn))/1000000);
      fBvland=round((sum(pBvland,iBvland,aBvland))/1000000);
      array y{6} sInt sTax sCrnet sDbnet sCpxn sBvland;
      do i=1 to 6;
            y{i}=round(y{i}/1000000);
      end;
      drop pInt iInt aInt pTax iTax aTax pCrnet iCrnet aCrnet pDbnet
           iDbnet aDbnet pCpxn iCpxn aCpxn pBvland iBvland aBvland i;

proc summary data=&d;
      class period;
      var fTurn fInt fTax fCrnet fDbnet netProf fCpxn fBvland;
      output out=total(where=(_type_=1)) sum=fTurn fInt fTax fCrnet

       fDbnet netProf fCpxn fBvland stderr=fTurnSe;
run;

data total;
      set total(drop=_type_ _freq_);
      se=round(fTurnSe);
      drop fTurnSe;
      &b='0'; *Total (all industries);

data qfs.&d;
      set qfs.&d(where=(period ne "&a"))
          &d
          total;
      by &b period;
      label preDif='Precision of difference between turnover
                    estimates'
            turnDif='Difference between turnover estimates'
              turnDifP='%-Difference between turnover estimates';
      preDif=round((sqrt(se**2+(lag(se))**2))*1.96);
      turnDif=fTurn-lag(fTurn);
      turnDifP=round((((fTurn/lag(fTurn))-1)*100),.1);
      array a{5} fInt fTax netProf fCpxn fBvland;
      array z{5} fIntP fTaxP netProfP fCpxnP fBvlandP;
      do i=1 to 5;
            z{i}=round((((a{i}/lag(a{i}))-1)*100),.1);
      end;
      drop i;
      if &b ne lag(&b) then do;
            array xx{8} preDif turnDif turnDifP fIntP fTaxP netProfP
                        fCpxnP fBvlandP;
            do i=1 to 8;
                  xx{i}=.;
            end;
      end;
      length sig $15.;
      label sig='Significance status of the difference between
                turnover estimates';
      if abs(turnDif)>abs(preDif) then sig='Significant';
      else if &b=lag(&b) then sig='Not significant';
      if turnDif>0 and preDif=. then sig='Not available';
      if &b='0' then do;
            cilTurn=round(fTurn-(1.96*se));
            ciuTurn=round(fTurn+(1.96*se));
      end;

proc print data=qfs.&d noobs n label;
var &b period sTurn rse fTurn se preDif--sig;
title "QFS-&a: Estimates &c reclassification";
run;

/*
proc export
```

```
      data=qfs.&d
      outfile="d:\qfs\&d..xls"
      dbms=excel2000 replace;
run;
*/
proc datasets;
      delete test samcap2 orirec oriimp add _ds0 dut out out2 out3
             total &d;
run;
```

Appendix 3

Economic Statistics starting points for sampling purposes 2002

Economic Surveys (Monthly and Quarterly)

NAME OF SURVEY	SERIES	SCOPE	INDUSTRIES	FREQUENCY	STARTING POINTS
Civil cases for Debt (CCDS)	0041	153	Source: Magistrates' offices	Monthly	Not applicable
Liquidations and insolvencies (LIS)	0043	Not applicable	Source: Registrar of Companies and Close Corporations	Monthly	Not applicable
Survey of Employment and Earnings (SEE)	0275	10 183	All industries, except mining and quarrying and agriculture	Quarterly	0.00
Survey of Average Monthly Earnings (AME)	0276	7 719	All industries, except mining and quarrying and agriculture	Quarterly	0.30
Mining: Production and sales (MPSS)	2041	Not Available	Source: Department of Minerals and Energy	Monthly	Not applicable
Manufacturing: Production and sales (MPS)	3041.2	3 000	Manufacturing Industry	Monthly	0.60
Manufacturing: Utilisation of Production Capacity (UPCS)	3043	1 500	Manufacturing Industry	Quarterly	0.60
Manufacturing: production and sales – Products manufactured (MPS-Products)	3051.1 - 3051.4	3 000	Manufacturing Industry	Monthly	0.60
Generation and consumption of electricity	4141	23	Electricity Industry	Monthly	0.85
Selected Building Statistics (BSS)	5041	128	Construction Industry	Monthly	0.85

Wholesale Trade Sales (WTSS)	6141	804	Wholesale Trade Industry	Monthly	0.85
Retail Trade Sales (RTSS)	6242	2 936	Retail Trade Industry	Monthly	0.85
Motor Trade (MTS)	6343	599	Motor Trade Industry	Monthly	0.85
Bed Breakfast establishments and other Short-stay accommodation	6410	New	Accommodation Industry	Monthly	0.85
Restaurants, Bars and Canteens	6420	New	Accommodation Industry	Monthly	0.85
Trading Statistics of Hotels (TSHS)	6441	410	Hotel Trading Industry Source: Business Address Register (BAR)	Monthly	Not applicable
Land Freight Transport (LFTS)	7142	499	Land Freight Industry	Monthly	0.85
Quarterly Financial Statistics Survey (QFS)	8042	3 273	All industries, except agriculture (SIC1), Financial intermediation (SIC81), Insurance and pension funding (SIC82), Public administration (SIC91) and Education (SIC92)	Quarterly	0.75
Local government financial statistics	P9144/5	284	Local Government Institutions. Source: Administrative	Quarterly	Not applicable
Remuneration and turnover based on levies	P9149	53	Local Government (categories A & C). Source: Administrative	Quarterly	Not applicable

Economic Surveys (Annual and Periodic)

NAME OF SURVEY	SERIES	SCOPE	INDUSTRIES	FREQUENCY	STARTING POINTS
Survey of Employment by Occupation, Gender and Race (SEOGR)	0201	This survey has been discontinued	All industries	Annual	This survey has been discontinued
Census of Agriculture (CAGR)	1101	55 193	Agriculture Industry	6 Yearly	0.00
Personnel Services Large Sample Survey (PSLSS)	1109	New	Personnel Services Industry	6 Yearly	New
Mining Large Sample Survey (MILSS)	2001	800	Mining Industry	6 Yearly	0.60
Manufacturing Large Sample Survey (MLSS)	3001	12 321	Manufacturing Industry	5 Yearly	0.50
Construction large sample Survey (CLSS)	5001	6 000	Construction Industry	6 yearly	0.80
Selected Annual Building Statistics (ABSS)	5011	203	Construction Industry	Annual	0.85
Wholesale and Retail Trade Large Sample Survey (WRTLSS)	6101	18 132	Wholesale and Retail Industry	6 Yearly	0.20
Motor Trade Large Sample Survey (MTLSS)	6301	3 151	Motor Trade Industry	6 Yearly	0.85
Accommodation Large Sample Survey (ALSS)	6401	New	Accommodation Industry	6 Yearly	New
Transport Large Sample Survey (TLSS)	7101	2 521	Transport Industry	6 Yearly	0.85
Post and Telecommunication Census (PTC)	7501	315	Post and Telecommunication Industry	6 Yearly	0.85

Economic Activity Survey (EAS)	8001	7 345	All industries, except Agriculture (SIC1), Financial intermediation (SIC81), Insurance and pension funding (SIC82), Public administration (SIC91) and Education (SIC92).	Annual	0.40
Business Services Large Sample Survey	8801	New	Business Services Industry	6 Yearly	New
Actual and Expected Capital Expenditure by The Public Sector	P9101	612	Public Sector. Source: Administrative.	Annually	Not applicable
Financial statistics of extra-budgetary accounts and funds	P9102	128	General government. Source: Administrative.	Annual	Not applicable
Financial statistics of universities and technikons	P9103	36	General government. Source: Administrative.	Annual	Not applicable
Financial Census of Municipalities	P9114/5	284	Government (Municipalities)	Annual	Not applicable
Non-financial census of local government *	P9118	284	Local government institutions. Source: Administrative.	Annual	Not applicable
National government expenditure	P9119.2	35	General government. Source: Administrative.	Annual	Not applicable
Provincial government expenditure	P9120	117	General government. Source: Administrative.	Annual	Not applicable

* Currently donor funded

Appendix 4

Major divisions, divisions and major groups

Division	Major group	Title of category
		MAJOR DIVISION 3: MANUFACTURING
30		Manufacture of food products, beverages and tobacco products
	301	Production, processing and preserving of meat, fish, fruit, vegetables, oils and fats
	302	Manufacture of dairy products
	303	Manufacture of grain mill products, starches and starch products and prepared animal feeds
	304	Manufacture of other food products
	305	Manufacture of beverages
	306	Manufacture of tobacco products
31		Manufacture of textiles, clothing and leather goods
	311	Spinning, weaving and finishing of textiles
	312	Manufacture of other textiles
	313	Manufacture of knitted and crocheted fabrics and articles
	314	Manufacture of wearing apparel, except fur apparel
	315	Dressing and dying of fur; manufacture of articles of fur
	316	Tanning and dressing of leather; manufacture of luggage, handbags, saddlery and harness
	317	Manufacture of footwear
32		Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials; manufacture of paper and paper products; publishing, printing and reproduction of recorded media
	321	Sawmilling and planing of wood
	322	Manufacture of products of wood, cork, straw and plaiting materials
	323	Manufacture of paper and paper products
	324	Publishing
	325	Printing and service activities related to printing
	326	Reproduction of recorded media
33		Manufacture of coke, refined petroleum products and nuclear fuel; manufacture of chemicals and chemical products; manufacture of rubber and plastic products
	331	Manufacture of coke oven products
	332	Petroleum refineries/synthesisers
	333	Processing of nuclear fuel
	334	Manufacture of basic chemicals
	335	Manufacture of other chemical products

	336	Manufacture of man-made fibres
	337	Manufacture of rubber products
	338	Manufacture of plastic products
34		Manufacture of other non-metallic mineral products
	341	Manufacture of glass and glass products
	342	Manufacture of non-metallic mineral products n.e.c.
35		Manufacture of basic metals, fabricated metal products, machinery and equipment and of office, accounting and computing machinery
	351	Manufacture of basic iron and steel
	352	Manufacture of basic precious and non-ferrous metals
	353	Casting of metals
	354	Manufacture of structural metal products, tanks, reservoirs and steam generators
	355	Manufacture of other fabricated metal products; metalwork service activities
	356	Manufacture of general purpose machinery
	357	Manufacture of special purpose machinery
	358	Manufacture of household appliances
	359	Manufacture of office, accounting and computing machinery
36		Manufacture of electrical machinery and apparatus n.e.c.
	361	Manufacture of electric motors, generators and transformers
	362	Manufacture of electricity distribution and control apparatus
	363	Manufacture of insulated wire and cable
	364	Manufacture of accumulators, primary cells and primary batteries
	365	Manufacture of electric lamps and lighting equipment
	366	Manufacture of other electrical equipment n.e.c.
37		Manufacture of radio, television and communication equipment and apparatus and of medical, precision and optical instruments, watches and clocks
	371	Manufacture of electronic valves and tubes and other electric components
	372	Manufacture of television and radio transmitters and apparatus for line telephony and line telegraphy
	373	Manufacture of television and radio receivers, sound or video recording or reproducing apparatus and associated goods
	374	Manufacture of medical appliances and instruments and appliances for measuring, checking, testing, navigating and other purposes, except optical instruments
	375	Manufacture of optical instruments and photographic equipment
	376	Manufacture of watches and clocks
38		Manufacture of transport equipment
	381	Manufacture of motor vehicles
	382	Manufacture of bodies (coachwork) for motor vehicles; manufacture of trailers and semi-trailers

	383	Manufacture of parts and accessories for motor vehicles and their engines
	384	Building and repairing of ships and boats
	385	Manufacture of railway and tramway locomotives and rolling stock
	386	Manufacture of aircraft and space craft
	387	Manufacture of transport equipment n.e.c.
39		Manufacture of furniture; manufacturing n.e.c.; recycling
	391	Manufacture of furniture
	392	Manufacture n.e.c.
	395	Recycling n.e.c.
		MAJOR DIVISION 4: ELECTRICITY, GAS AND WATER SUPPLY
41		Electricity, gas, steam and hot water supply
	411	Production, collection and distribution of electricity
	412	Manufacture of gas; distribution of gaseous fuels through mains
	413	Steam and hot water supply
42	420	Collection, purification and distribution of water
		MAJOR DIVISION 5: CONSTRUCTION
50		Construction
	501	Site preparation
	502	Building of complete constructions or parts thereof; civil engineering
	503	Building installation
	504	Building completion
	505	Renting of construction or demolition equipment with operators
		MAJOR DIVISION 6: WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES, MOTOR CYCLES AND PERSONAL AND HOUSEHOLD GOODS; HOTELS AND RESTAURANTS
61		Wholesale and commission trade, except of motor vehicles and motor cycles
	611	Wholesale trade on a fee or contract basis
	612	Wholesale trade in agricultural raw materials, livestock, food, beverages and tobacco
	613	Wholesale trade in household goods
	614	Wholesale trade in non-agricultural intermediate products, waste and scrap
	615	Wholesale trade in machinery, equipment and supplies
	619	Other wholesale trade
62		Retail trade, except of motor vehicles and motor cycles; repair of personal household goods
	621	Non- specialised retail trade in stores
	622	Retail trade in food, beverages and tobacco in specialised stores
	623	Other retail trade in new goods in specialised stores
	624	Retail trade in second-hand goods in stores
	625	Retail trade not in stores
	626	Repair of personal and household goods

63		Sale, maintenance and repair of motor vehicles and motor cycles; retail trade in automotive fuel
	631	Sale of motor vehicles
	632	Maintenance and repair of motor vehicles
	633	Sale of motor vehicle parts and accessories
	634	Sale, maintenance and repair of motor cycles and related parts and accessories
	635	Retail sale of automotive fuel
64		Hotels and restaurants
	641	Hotels, camping sites and other provision of short-stay accommodation
	642	Restaurants, bars and canteens
		MAJOR DIVISION 7: TRANSPORT, STORAGE AND COMMUNICATION
71		Land transport; transport via pipelines
	711	Railway transport
	712	Other land transport
	713	Transport via pipelines
72		Water transport
	721	Sea and coastal water transport
	722	Inland water transport
73	730	Air transport
74	741	Supporting and auxiliary transport activities; activities of travel agencies
75		Post and telecommunications
	751	Postal and related courier activities
	752	Telecommunications
		MAJOR DIVISION 8: FINANCIAL INTERMEDIATION, INSURANCE, REAL ESTATE AND BUSINESS SERVICES
81		Financial intermediation, except insurance and pension funding
	811	Monetary intermediation
	819	Other financial intermediation n.e.c.
82	821	Insurance and pension funding, except compulsory social security
83		Activities auxiliary to financial intermediation
	831	Activities auxiliary to financial intermediation, except insurance and pension funding
	832	Activities auxiliary to insurance and pension funding
84		Real estate activities
	841	Real estate activities with own or leased property
	842	Real estate activities on a fee or contract basis
85		Renting of machinery and equipment, without operator, and of personal and household goods
	851	Renting of transport equipment
	852	Renting of other machinery and equipment

	853	Renting of personal and household goods n.e.c.
86		Computer and related activities
	861	Hardware consultancy
	862	Software consultancy and supply
	863	Data processing
	864	Data base activities
	865	Maintenance and repair of office, accounting and computing machinery
	869	Other computer related activities
87		Research and development
	871	Research and experimental development of natural sciences and engineering
	872	Research and experimental development of social sciences and humanities
88		Other business activities
	881	Legal, accounting, bookkeeping and auditing activities; tax consultancy; market research and public opinion research; business and management consultancy
	882	Architectural, engineering and other technical activities
	883	Advertising
	889	Business activities n.e.c.
		MAJOR DIVISION 9: COMMUNITY, SOCIAL AND PERSONAL SERVICES
91		Public administration and defence activities
	911	Central government activities
	912	Regional services council activities
	913	Local authority activities
92		Education
	920	Educational services
93		Health and social work
	931	Human health activities
	932	Veterinary activities
	933	Social work activities
94		Other community, social and personal service activities
	940	Sewage and refuse disposal, sanitation and similar activities
95		Activities of membership organisations n.e.c.
	951	Activities of business, employers’ and professional organisations
	952	Activities of trade unions
	959	Activities of other membership organisations
96		Recreational, cultural and sporting activities
	961	Motion picture, radio, television and other entertainment activities
	962	News agency activities
	963	Library, archives, museums and other cultural activities
	964	Sporting and other recreational activities

99	990	Other service activities