Modeling Schrödinger's Cat

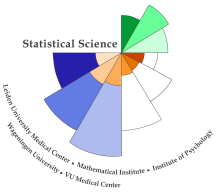
An Investigation into the Intermittent-Error Model

Judith ter Schure

First Supervisor: Drs. S. Scholtus (CBS)
Second Supervisor: Prof. dr. P.D. Grünwald (Leiden University/CWI)

MASTER THESIS

Defended on March 9, 2017



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Content

Preface

Many thanks to *Sander Scholtus*, my first supervisor at Statistics Netherlands, for his invitation into his research and his advice and support during my days on the Statistics Netherlands' methodology floor in The Hague. I have enjoyed the fruitful Friday's sessions and I thank him for his commitment when I needed to share my discoveries mid-November. Not the least because there was always enough time for all my urgent and not so urgent questions. I was happy to see he was just as interested as I was.

Thanks to *Statistics Netherlands* for having me as an intern. I am especially grateful for the opportunity to get involved in the larger context of a 2000 persons' institution through lunch lectures and a Lean Six Sigma course. I will remember *my colleagues from methodology and process development* for the interesting discussions over lunch and thank them for inviting me to the yearly team outing at Pitch and Put Leidschenveen. I would like to specifically mention Arnout van Delden, who has helped me dig into historical documents to compose a story to introduce my Master Thesis' presentation.

I am very grateful to *Peter Grünwald* for being my second supervisor, which gave my research a general statistical basis. I appreciate the nice conversations we had, especially since there were so many other matters requiring his attention. I wish to thank everyone involved in the Master Track *Statistical Science for the Life and Behavioural Sciences* for providing such a great all-round statistical education.

Thanks to *Marnick van de Zande* for his comments on lay-out and writing, and his support in the process of this research. And to *Kim Ménage*, for her help on the specifics of the English language.

Finally, special thanks to *Wikipedia*, one of the best inventions since my birth, and always the best starting point into the 'unknown' unknown[1]. And to conclude, special thanks to *Stackexchange*, for the very clear and detailed answers to my known unknowns[2].

[1] So much to explore, for example, starting from: *https://en.wikipedia.org/wiki/Normal_distribution*
[2] Some examples of known unknowns resolved by Stackexchange:
http://stats.stackexchange.com/questions/7439/how-to-change-data-between-wide-and-long-formats-in-r
http://math.stackexchange.com/questions/77306/why-does-substitution-work-in-antiderivatives/77356#77356?newreg=41311f349b644e2fa61e2fd35bba9845
http://english.stackexchange.com/questions/2120/which-is-correct-dataset-or-data-set
http://tex.stackexchange.com/questions/29664/latex-error-unknown-graphics-extension-eps

Introduction

Almost every informed debate on societal issues benefits from decent and recent statistical information on the state of society. This information needs to have the status of facts among all members in the discussion, to prevent alternative-fact-arguments. Especially numerical information, which is very prone to mislead, needs to be provided by official institutions. In most countries around the world the institutions that compile these *official statistics* are operating at the national level and denoted by *National Statistical Institutes (NSIs)*. In The Netherlands, the NSI which produces official statistics is *Statistics Netherlands* (Dutch: Centraal Bureau voor de Statistiek (CBS)). Statistics Netherlands describes its mission as follows:

> *"The mission of CBS is to publish reliable and coherent statistical information which responds to the needs of Dutch society. The responsibility of CBS is twofold: firstly, to compile (official) national statistics and secondly to compile European (community) statistics. […] The information published by CBS deals with subjects directly affecting the lives of Dutch citizens. These include economic growth, consumer prices, crime but also leisure." (CBS, 2016a)*

Traditionally, National Statistical Institutes rely on their own surveys for data. But carrying out surveys is expensive and time-consuming. Moreover, the quality of survey data in The Netherlands has declined in recent decades due to selective nonresponse and decreasing willingness of both individuals and businesses to respond to Statistics Netherland's surveys (Bakker B. F., 2009). Furthermore, a general objective has developed among NSIs to lower the respondent burden (e.g. (Van Delden & De Wolf, 2013), (Guarnera & Varriale, 2016), (Zhang, 2012)).

In this digital age, register data administered by other institutions can be easily shared with statistical institutes. In The Netherlands, the usage of public administrative registers for the statistical purposes of Statistics Netherlands is permitted and regulated since 2004 by the *CBS law* (CBS, 2003). Examples of public institutions maintaining registers in The Netherlands are the Tax and Customs Administration (Dutch: Belastingdienst), the Chamber of Commerce (Dutch: Kamer van Koophandel (KvK)) and the Employee Insurance Agency (Dutch: Uitvoeringsinstituut Werknemersverzekeringen (UWV)). An *administrative register* can be defined as an (attempted) complete list of records that describes a population (Bakker B. F., 2009), with a description in terms of variables that are continuously updated (Daas & Van Delden, 2014, p. 4). Registers therefore have the potential to provide decent and recent information.

More than half of Statistics Netherlands' resources for social and spatial statistics already originates from administrative registers (Bakker B. F., 2009, p. 3). Currently, Statistics Netherlands operates on 184 different registers (source: CBS internal register catalog, retrieved 2016/9/22). Register use has proven to be very cost-effective. The use of registers and already existing survey data during the 2001 Dutch census saved the Dutch government 290 million euros (Bakker B. F., 2009, p. 3). Not only is a shift to administrative registers for official statistics beneficial financially, registers also promise complete information on a population. On the contrary, survey data only describes the subgroup of the population that was sampled and responded to the survey.

Administrative registers thus contain valuable information, but the data sets are often not maintained for statistical purposes. For example, the Employee Insurance Agency maintains the so-called Insurance Policy Administration (Dutch: Polisadministratie). This register covers information on

employment, unemployment and pensions (Van Delden A. , 2013). Its purpose is to aid the process of unemployment benefits. Therefore, information of less administrative importance, but of large statistical interest might be more limited, such as the distinction between part-time and fulltime employment (Van Delden A. , 2013). Also, administrative register data is not without errors (e.g. (Belastingdienst, 2007)). The quality of register data is often unknown since it is not beneficiary for the register owner to publicly announce quality problems (Bakker B. F., 2009). Thus, the available information in registers can be restricted in a statistical sense and linkage to other sources as well as thorough editing might be required.

Fortunately, in case of some official statistics, such as those in economic publications, multiple agencies maintain registers with the required variables. Statistics Netherlands can obtain information on the same population from multiple sources that describe theoretical identical values. With multiple sources available, errors can be identified in terms of discrepancies between values on the same unit and patterns among larger numbers of units can indicate systematic deviations between sources. Still, when two sources report different values, either one of them can be error free or incorrect, or both of them can be incorrect. Obtaining error-free values requires a model to describe the observed values in relation to the True Value of interest. This True Value is unknown for most units and the situation is in a way like Schrödinger's cat thought experiment: The unit still needs to reveal itself to an observer to be either incorrect (dead) or error-free (alive). A model is needed to be that observer.

This study investigates the performance of such a model: The Intermittent-Error Model. A combination of sample survey and register data is available at Statistics Netherlands that describes the same businesses in terms of theoretically the same financial variables. Each source contains errors and combining sources can lead to error detection as well as the possibility to obtain True Values. This study investigates the state of yearly financial business data in which the Intermittent-Error Model could be implemented and assesses the Intermittent-Error Model's merits and potential drawbacks for Statistics Netherlands.

Research questions

This study aims to answer the following main question:

What are the advantages and disadvantages of implementing the Intermittent-Error Model to assess and improve the quality of financial business data at Statistics Netherlands?

To answer the main question, the following sub questions are discussed:

1. What concepts are needed to assess the quality of Statistics Netherlands' business data?
    a. What data sources are available to Statistics Netherlands in general and specifically with regard to business data?
    b. What error types possibly occur in Statistics Netherlands' business data?
    c. What errors apparently occur in a Statistics Netherlands' case study data set?
    d. How are errors in business data currently detected and corrected by Statistics Netherlands?
2. What is the Intermittent-Error Model?
    a. How is the Intermittent-Error Model defined?
    b. What are the Intermittent-Error Model's assumptions?
    c. How are the Intermittent-Error Model parameters estimated?
3. How can the Intermittent-Error Model's performance be assessed?
    a. Why does the Intermittent-Error Model assume a 'True Value'-distribution?
    b. How does the assumption of a 'True Value'-distribution influence the Intermittent-Error Model's True Value estimates?
    c. What measures can assess the Intermittent-Error Model's fit and soundness?
4. What is the Intermittent-Error Model's performance on a case study data set?
    a. What is the general case study model fit?
    b. How does the case study model fit relate to apparent errors in the case study data set?
    c. What is the case study model fit with regard to the fit measures/soundness measures?
    d. What is the influence of transformations on the case study model fit?
    e. What is the stability of the case study model fit?
5. What are the Intermittent-Error Model's merits in detecting and correcting error types occurring in Statistics Netherlands business data?
    a. How does the Intermittent-Error Model definition relate to possible error types?
    b. What are the Intermittent-Error Model's merits for error detection?
    c. What are the Intermittent-Error Model's merits for error correction?
    d. What are the Intermittent-Error Model's merits for source assessment?
    e. What are the Intermittent-Error Model's merits for selection in manual editing?

Outlook

The core of this report covers exactly 100 pages. Fortunately, of those 100 pages, only 56 contain plain written text. These 56 pages of text are complemented with a wide variety of figures, plots, tables and summaries, to encourage the reader to digest it all.

Concepts to assess data quality from Statistics Netherlands' perspective are discussed in Chapter 1, in response to research sub question 1, with Section 1.1, 1.2 and 1.4 devoted to questions 1a, 1b and 1d respectively. Section 1.2 introduces a case study data set, for which research sub question 1c is answered in Section 1.3. The same case study data set is used to answer research sub question 4 in Chapter 4. Chapter 2 discusses the Intermittent-Error Model, in response to research sub question 2. How the Intermittent-Error Model's performance can be assessed is discussed in Chapter 3, in response to sub question 3, with Section 3.1 and 3.2 answering sub questions 3a and 3b about the True Value distribution. Section 3.3 introduces specific Fit/Soundness measures to assess the Intermittent-Error Model in response to sub question 3c and these are used in Chapter 4 to assess the Intermittent-Error Model's performance on the case study data set, in Section 4.1 and 4.3. Further model fit assessment is carried out in relation to the errors already apparent in Chapter 1, in Section 4.2 in response to sub question 4b, the influence of transformations (sub question 4c) is inspected in Section 4.4 and the stability of the model fit (sub question 4e) is investigated in Section 4.5. The final research questions, 5a - 5e, discussing the Intermittent-Error Model's merits in general, are answered in Chapter 5. Sections 1.5, 2.4, 3.4, 4.6 and 5.6 conclude each chapter with a summary in response to the five sub questions. Chapter 6 provides a general conclusion to the study and Chapter 7 pays special attention to insights and suggestions for future research.

All statistical analyses in this project are performed in the R-software environment. The R code used for the analyses and plots is available upon request: judithterschure@gmail.com.

1 Concepts needed to assess business data quality

To investigate the Intermittent-Error Model's performance on financial business data from various sources, concepts are needed to assess (source) data quality and the way data can be processed to produce economic statistics. Quality of data depends on the type of the source from which the data originates, the possible error types that can occur in the source and combination of sources and the means to detect and correct those errors. Section 1.1 discusses the various data sources available to Statistics Netherlands in general and specifically with regard to financial business data. A system to describe the origin of various error types is discussed in Section 1.2, by means of examples with regard to the case study data set that is central to this study. This case study data set is further introduced in Section 1.3, with special attention to the apparent errors that occur in its data. Section 1.4 describes the way errors are currently detected and corrected in Statistics Netherlands' business data and Section 1.5 concludes this first chapter with a summary.

1.1 Statistics Netherlands' data sources

The cost-effectiveness of National Statistical Institute's shift from sample survey data to administrative register data mainly results from the shift in collection effort. Sample survey data is denoted by *primary data*, since it is collected by the NSI itself. Administrative register data is available elsewhere, thus collected by others, and denoted by *secondary data* ((Hox & Boeije, 2005) as referred to by Daas & Van Delden (2014). This distinction between primary and secondary data is from the National Statistical Institute's point of view. For example, data that originates from a tax register is primary to the Tax and Customs Administration which is responsible for its collection. The distinction between primary and secondary data is revisited in Section 1.2. Figure 1.1 (on page 2) shows a subdivision of secondary data sources from the NSI's perspective in *statistical sources*, *administrative sources* and *organic sources*. This figure and the description in the paragraphs below originate from (Daas & Van Delden, 2014).

Statistical sources
Secondary statistical sources contain data collected by other survey-oriented organizations, such as market research institutes and government research bodies. This type of secondary data is most similar to data from primary sources since the collection purpose is statistical.

Administrative sources
The collection purpose of administrative sources is to aid the administration of the owner organization and not to produce official statistics. Organizations collecting data on a large (sub)population can be both public and private. Examples of public administrative sources in The Netherlands are the Chamber of Commerce's Trade Register (Dutch: Handelsregister), the National Medical Registration and the Municipal Population Register (Dutch: Gemeentelijke Basisadministratie persoonsgegevens (GBA)). Examples of private administrative sources are databases with prices of supermarket products, mobile phone call-detail records, and data collected by smart electricity and gas meters.

Organic sources
A considerable part of organic sources can be identified as *Big Data*. Organic sources and big data are created in a more unstructured way, and for other purposes than the other main categories of statistical sources. Therefore, an organic source often lacks a clear definition of a target population.

Examples of organic sources are dwelling prices collected from websites, GPS data and collections of social media messages.

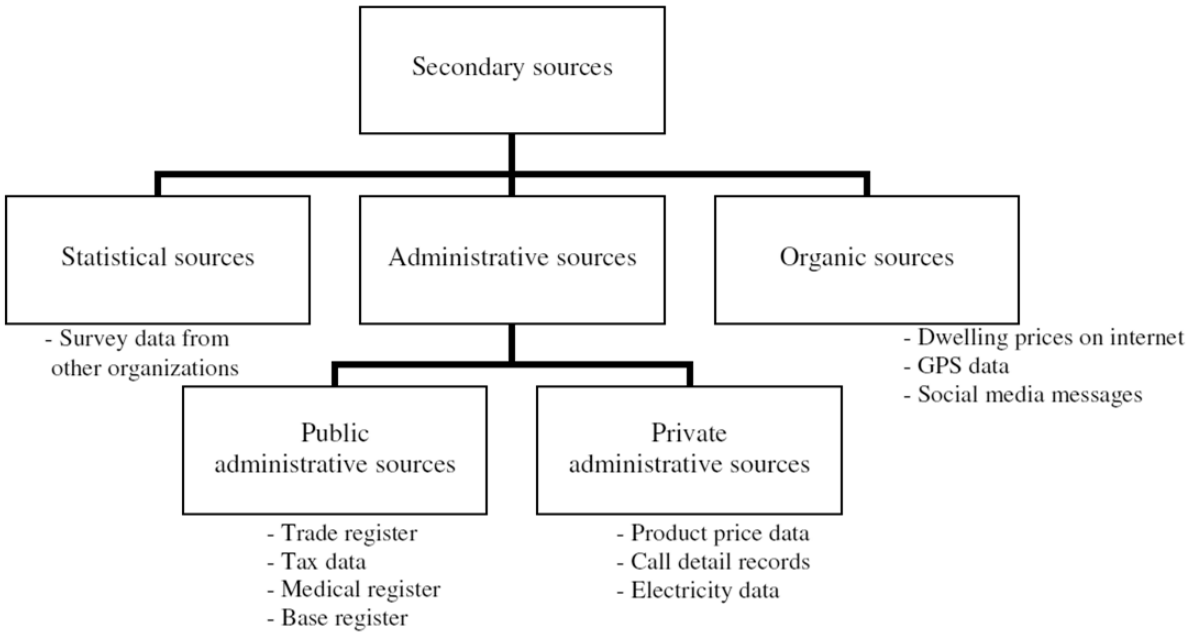


Figure 1.1 "Distinguished categories of secondary sources with examples" [from: **(Daas & Van Delden, 2014, p. 5)**]

Base Registers

Statistics Netherlands cooperates with public administrative register owners to maintain *base registers*. Base registers not only have an administrative function but are also meant to give a complete view of the population of interest (Van Delden A. , 2013). An example register that is used as a base register by Statistics Netherlands is the Municipal Population Register (Dutch: Gemeentelijke Basisadministratie persoonsgegevens (GBA)). Section 1.2.3.1 introduces the base register for businesses that is central to economic statistics.

Statistics Netherlands' data sources for financial business data

The sources discussed in this study are both primary and secondary. The primary source is a yearly sample survey of businesses carried out by Statistics Netherlands. Secondary sources are registers maintained by the Chamber of Commerce and the Tax and Customs Administration, thus public administrative sources.

1.2 Possible error types in Statistics Netherlands' business data

The data under study concerns businesses and the main focus is on data collected to compile yearly economic statistics, such as turnover per economic activity. Therefore, the starting point is to collect turnover data on businesses and a classification by economic activity. The case study data set is characterized by the fact that three different sources report turnover values on the same businesses (so-called *multi-source data* (Guarnera & Varriale, 2016)). The data set is composed of units that are especially defined for its statistical purpose, denoted by *statistical business units* (Dutch: Bedrijfseenheden (BE), also known as *enterprises* within the European Community). The variables on which data is obtained for each statistical business unit are described in more detail in Section 1.3.2. The case study data set is related to the so-called *Structural Business Statistics (SBS)* (Dutch: Productiestatistieken) of the year 2012. The Structural Business Statistics are publishable economic statistics manufactured by Statistics Netherlands since the 1970s with the objective of a general industry profit and loss account. The Structural Business Statistics data is provided to Eurostat (the statistical institute for official statistics of the European community) and the data is intensively used to acquire overarching statistical reports such as the National Accounts (Dutch: Nationale Rekeningen) (CBS, 2016b). Published SBS values are currently based on sample survey data only, but information from Tax and Customs Administration's registers on the same units is maintained for research into future usage to improve the quality and cost-effectiveness of the SBS and other official statistics.

Macro data, micro data, linking, micro integrating and editing

The aggregated values on groups of units that National Statistical Institutes publish, such as economic statistics, are denoted by *macro data*. The data collected to produce these statistics is at the individual level, such as businesses, and is denoted by *micro data* (e.g. (CBS, 2016b)). This study investigates the quality of micro data when multiple sources are available. When combining data from various registers, or combining register data with survey data, records can be *linked* at the micro level when information on identical units can be identified in separate data sets. Records need to be *micro integrated* when variables in separate data sets have incompatible values for some or all units. *Micro integration* is defined by Bakker (2011, p. 6) as: *"A method where data from statistical units is matched at individual level with the goal of compiling better information than is possible when using the separate sources. The quality relates to validity, reliability and consistency (cross-theme and longitudinal)."* So micro integration corrects the micro data from multiple sources to obtain a combined data set of certain quality. Within such a combined data set, individual records can be *edited* when they do not agree to certain decision rules (such as the relation between costs, turnover and profit for businesses). The process of micro integration is vital to the situation with multi-source data and will be further discussed in this section in relation to error occurrences in the micro-integration process. Editing will be further discussed in Section 1.4, with regard to Statistics Netherlands' current practices of error detection and correction.

Error types

Neither sample surveys, nor public administrative registers provide perfect micro data for statistical purposes. Reasons for imperfect data are for example that available values differ from the ideal measures or that the measured objects differ from target units. These reasons are referred to as *error sources* ((Groves, et al., 2004) as referred to in: (Zhang, 2012)). Error sources originate at different stages in the data *life cycle*, which are for example collection, processing and production. A Two-Phase Life-Cycle Model is proposed by Zhang (2012) as a framework of error types corresponding to error sources in micro data. In this section the *Zhang Two-Phase Life-Cycle Model* will form the basis to describe possible error types in the various sources that constitute the business data under study.

1.2.1 Structure of the case study data

The case study data set contains financial data on businesses that originate from multiple sources. The population exists of so-called statistical business units, which are defined with regard to the purpose of compiling economic statistics to publish. The population definition originates from the Chamber of Commerce as well as the Tax and Customs Administration. The financial values are available from Statistics Netherlands own Structural Business Statistics survey as well as two Tax and Customs Administration's registers.

Population of Statistical Business Units

The definition of the statistical business unit concerns the following four characteristics (Konen, 2012, pp. 4-5). *Autonomy:* The business needs to be autonomous with regard to (among others) finance, production and sales. This enables grouping subsidiary companies within their parent company. *External orientation:* The sales market needs to be outside the company. *Describable:* The company needs to have its own accountancy. *Continuity:* A certain degree of continuing economic activity needs to be present.

Each statistical business unit is characterized according to its economic activities. The statistical classification of economic activities in the European Community is denoted by *NACE* (*Nomenclature des Activités Économiques dans la Communauté Européenne*) (Dutch: Standaard Bedrijfsindeling (SBI)) (Eurostat, 2008). NACE codes are hierarchical, dividing all economic activity into classes at different levels. Table 1.1 shows an example of a NACE code build-up. To create *functional statistics*, all economic activities of a certain company are relevant, and therefore a company can have multiple NACE codes. To produce *institutional statistics*, each statistical business unit needs to be uniquely classified by its main economic activity and can only have one NACE code (Konen, 2012). The Structural Business Statistics, and therefore the case study data set, provide institutional statistics, so each business unit is characterized by only one NACE code in this study's data set.

Table 1.1 Example NACE code build-up for *H52.2.9* **(Eurostat, 2010) (CBS, 2016d)**

Section	*H*	Transportation and storage
Division	*H52*	Warehousing and support activities for transportation
Group	*H52.2*	Support activities for transport
Class	*H52.2.9*	Other transportation support activities *Forwarding agencies, ship brokers and charterers; weighing and measuring*

Another characteristic of the statistical business unit is its locations of economic activity. If a business unit is economically active at multiple locations, it contains multiple local businesses. Knowledge of these locations is important to define local statistical units in production of regional economic statistics.

Very large or complex business units are assigned to the so-called *TopX*. There are approximately 9000 TopX businesses of which 494 are present in the case study data set. These TopX companies are very important with regard to macro data, thus the aggregated values that need to be published. According to (Konen, 2012) often a '10-90' rule holds: 10% of the companies are responsible for 90% of the economic activity (in terms of turnover, or employment). Therefore, obtaining quality data from these companies is of extra importance, and as a result they are treated separately in data set construction and editing.

The population framework of statistical business units is denoted by the *General Business Register (GBR)* (Dutch: Algemeen Bedrijvensregister (ABR)). This register is based on information from the Chamber of Commerce and the Tax and Customs Administration, and maintained by Statistics

Netherlands in collaboration with these two administrative institutions. Since its purpose is to define a population of businesses, it is a base register. From 2014 onwards, the GBR is based on information from only the Chamber of Commerce (CBS, 2016f). But since the case study data set regards 2012, the situation before 2014 is described.

Financial business data sources
Statistics Netherlands' system to link monthly, quarterly or yearly administrative register data to the GBR's exiting population framework is denoted by *BaseLine*. The GBR only contains *identification variables* and *structural variables* (e.g. name and address (location), NACE group, Size Class, legal structure) on the statistical units. BaseLine adds the yearly financial values, such as turnover, purchases and costs.

The financial data values in the case study data set contain 2012 information from three primary sources: The *Value Added Tax (VAT)* (Dutch: Belasting Toegevoegde Waarde (BTW) register), the *Profit Declaration Register (PDR)* (Dutch: Winstbelasting (WIA)/Vennootschapsbelasting (VPB) register) and the original sample survey carried out by Statistics Netherlands to produce the Structural Business Statistics. The first two originate from the Tax and Customs Administration, and are both linked to the GBR in BaseLine (Rademakers, 2005). The third consists of survey data on a sample of statistical business units from the same GBR, and is linked to the BaseLine output.

The data to produce the Structural Business Statistics is acquired by a stratified sampling procedure based on the sampling framework defined by the GBR. Strata are defined by NACE group and Size Class, and within each NACE group the probability of a business being sampled relates to the Size Class, with smaller Size Classes corresponding to smaller sampling probabilities (since the smaller Size Classes' turnover values are much more homogeneous than the larger Size Classes) (Meijers & Smeets, 2011). All businesses with more than 50 employees receive a survey (CBS, 2016e).

VAT data is monthly or quarterly collected by the Tax and Customs Administration, and quarterly values are processed by Statistics Netherlands to produce the *Short-Term Business Statistics*. Statistics Netherlands aims to base the entire Short-Term Business Statistics on the VAT data. However, the TopX enterprises need to be excluded and data on some NACE groups is still collected with a sample survey since previous research showed that their perceived 'true' survey results did not agree with the data provided by the VAT register (Van Delden & De Wolf, 2013). For the case study data set, the quarterly VAT data is combined to yearly values and thus comparable to the values from the Structural Business Statistics sample survey. The data from the Profit Declaration Register is added in due time. These values are based on yearly tax reports, but cannot be yearly linked since the Tax and Customs Administration allows companies to request postponement in reporting on Profits Tax (Belastingdienst, 2016b).

1.2.2 The Zhang Two-Phase Life-Cycle Model

The Zhang Two-Phase Life-Cycle Model is an extension of an earlier model proposed by Groves et al. (2004) that describes sources of errors in sample survey data collected by National Statistical Institutes themselves (primary source data). The Zhang Two-Phase Life-Cycle Model extends this original model with sources of error characteristic to administrative register data not collected by National Statistical Institutes (secondary source data) and sources of error that originate in combining multiple sources. The Zhang Two-Phase Life-Cycle Model is displayed graphically in Figure 1.2 and Figure 1.4 (on page 8 and 12 respectively).

Two phases
Figure 1.2 displays the first and Figure 1.4 displays the second phase of the Zhang Two-Phase Life-Cycle Model. These two phases relate to the *primary usage* and *secondary usage* of data. Therefore,

the first and primary phase can also describe secondary source data. Namely, the first phase describes errors that can occur with regard to the intended purpose of the data of the institution that originally collects the data. This can be administrative with regard to, for example, a Tax and Customs Administration's register or statistical with regard to an NSI's own sample survey. The second phase describes the reuse of elsewhere collected data, or combination of source data, for statistical purposes. Therefore, the end node of the first phase (in Figure 1.2, page 8) is denoted by *single-source (primary) micro data*, in which the adjective 'primary' refers to the first and intended purpose of the data set and not a primary source in as discussed in Section 1.1. Both data collected for statistical purposes and data collected for administrative purposes can be described by the first phase of the Zhang Two-Phase Life-Cycle Model.

The second phase describes stages of error occurrence in the statistical reuse of data originally collected for another purpose. The input to the second phase can originate from a single source or multiple sources that are combined with linkage and/or micro integration.

Measurement Errors and Representation Errors

The main structure of the Zhang Two-Phase Life-Cycle Model is provided by the separated left-hand and right-hand side of the flow represented by the broad arrows. These two sides indicate a classification of errors into Measurement Errors and Representation Errors. Those two main error categories are defined by Bakker (2011) as following: *Representation Errors* are errors that *"involve an incomplete and/or selective description of the population about which statements are being made*. *Measurement Errors* are errors *"that incorrectly describe the characteristics of the population"*. Thus Representation Errors concern the erroneously inclusion or exclusion of units in the population, such as nonexistent or incorrectly represented businesses in the case study data set. Measurement Errors concern incorrect values to variables belonging to those units, such as yearly turnover. The larger category of Measurement Errors needs to be distinguished from one of the six individual errors part of the category that is also denoted by *measurement error*. The first corresponds to a category including six different types of errors that are displayed at the left sides of the two phases of the Zhang Two-Phase Life-Cycle Model in Figure 1.2 and Figure 1.4. This report stresses the distinction between the category and the individual error by emphasis on the category with capital letters and abbreviated (cat) for category, thus using *Measurement Error (cat)* for the category of error types and *measurement error* for the individual error type.

Objects and Units

Objects and units represent the records at the micro level, the individuals on which the variables are measured. The distinction between objects and units follows from a distinction in purpose of the data, with 'object' used for primary purposes and 'unit' for secondary purposes. For instance, a register maintained by the Chamber of Commerce consists of individual businesses from a legal perspective, which are *objects*. When a mother and daughter company report on some tax together, they form one *object* from the Tax and Customs Administration's perspective. While in secondary use of the tax register data, the object might be split again in two *units* because from a statistical perspective they belong to different sectors of economic activities. Thus to obtain the *statistical business units* on which published economic statistics are based (Konen, 2012), a register of objects (phase one) needs to be restructured into a data set of units (phase two). This is a form of micro-integration, and indicated by the middle square in Figure 1.4 of phase two.

Target and Actual

The flow from top to bottom in the Zhang Two-Phase Life-Cycle Model also represents a distinction. At the top of each phase the ideal or preferred data is considered, which is denoted by the *target*, both at the representation (*target set* (phase one) / *target population* (phase two)) and at the measurement (*target concept*) level. More to the bottom of each phase emerges the *actual* data, at both sides of Figure 1.2 and Figure 1.4.

1.2.3 Possible error types in the case study data set

The population of businesses is described by the General Business Register (GBR), and yearly financial data on the statistical business units in the population is linked in BaseLine. Therefore, from the perspective of the Zhang Two-Phase Life-Cycle Model, the construction of case study data set involves executing the two phases twice. Namely, once to create the GBR and once with regard to the collection and linkage of yearly financial business data from other sources. Error types that can occur in the first (GBR) execution are described in Section 1.2.3.1, and error types that occur in the second (BaseLine) execution in Section 1.2.3.2. Section 1.2.3.3 discusses error types in which the two executions interact.

1.2.3.1 The GBR: creating the population framework

Statistics Netherlands maintains the General Business Register (GBR) in collaboration with the Chamber of Commerce and the Tax and Customs Administration. The GBR is a base register, since it defines a statistical population of businesses. The GBR is used as a sampling framework for economic sample surveys and also defines the set of unique statistical units for administrative register data to be linked to (Konen, 2012). The GBR is continuously updated and accompanied by an external framework describing a list of changes to statistical units, characteristics of units, links between units and quality information (Aelen, 2005). Due to this external framework, historical situations are reproducible and changes to the GBR can be explained and accounted for. The GBR's units are the statistical business units previously described, and these are characterized by the identification and structural variables such as names, addresses (location), NACE groups, Size Class and legal structure. This section describes (a simplified version of) the production of the General Business Register from the perspective of the Zhang Two-Phase Life-Cycle Model.

The GBR - Phase One: Data sources

The General Business Register is based on the Chamber of Commerce's *Business Register (BR)* (Dutch: Handelsregister (HR)) and the Tax and Customs Administration's *Relations Register (RR)* (Dutch: Beheer van Relaties (BvR)) (Konen, 2012). The Chamber of Commerce and Tax and Customs Administration cooperate with Statistics Netherlands to maintain the General Business Register (Aelen, 2005). The combination of the two sources for the GBR was needed in the past due to registration exemptions of some businesses with the Chamber of Commerce, for example agricultural companies and some 'liberal professions' (KvK, 2016b).

The GBR's data originates from public administrative register data, and therefore the objectives of the Business Register (BR) and Relations Register (RR) are nonstatistical. As a result, the original structure of the data might be different from the targeted statistical business units. The objects in these original sources represent *legal units* (Chamber of Commerce) and *fiscal units* (Tax and Customs Administration) and are in their original sources already accompanied by various identification variables and structural variables, such as names, addresses, ownership relations (BR and RR), NACE group and legal structure (BR) (Aelen, 2005, p. 3) (Konen, 2012, p. 9).

Phase one of the Zhang Two-Phase Life-Cycle Model concerns data collection with regard to its primary objective. The primary aim of the Chamber of Commerce's Business Register is to facilitate trade: *"Entrepreneurs and private individuals can use the Business Register to find out about business contacts, for example existing or potential customers and suppliers. This fosters legal certainty in commercial transactions."* (KvK, 2016a) The purpose is therefore closely related to that of Statistics Netherlands. The Chamber of Commerce also mentions among the three goals of its register: "*Government bodies can obtain information about companies and legal entities from the Business Register, and no longer need to ask the entrepreneur for this. This reduces the administrative burden."* (KvK, 2016a) The primary objective of the Tax and Customs Administration's Relations Register is to register all companies obligated to pay Value Added Tax (VAT) (Dutch:

Belasting Toegevoegde Waarde (BTW)), payroll taxes, insurance premium tax, dividend tax and Profits Tax (Dutch: Vennootschapsbelasting/Winstbelasing) (Belastingdienst, 2016a). Therefore, the purpose is administrative and not statistical.

The GBR - Phase One: Representation Errors

The error sources concerning Representation Errors in phase one are displayed at the right-hand side of Figure 1.2.

Originally the terms 'target set' and 'accessible set' relate to concepts in surveys. The *target set* in a sample survey contains the units of the target statistical population, whereas the *accessible set* corresponds to the available sampling framework. A difference between the two produces a *frame error*. Such differences can also occur in administrative data from the two sources that produce the GBR. For example, a 'black market' business might not be willing to register with either source, while both the Chamber of Commerce and the Tax Agency would want to include this business. Therefore, it is part of the target set, but since the two sources are unable to include the business it is not in the accessible set, thus constituting a frame error.

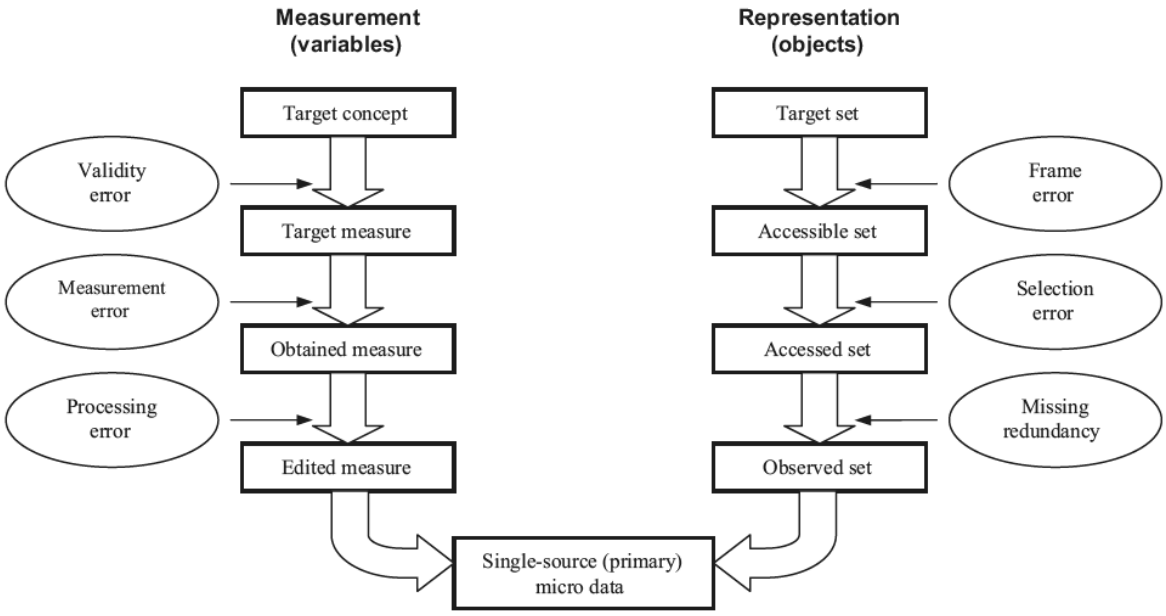


Figure 1.2 Phase One of "Two-phase (primary and secondary) life cycle of integrated statistical micro data from a quality perspective. Data concept (square); Error Type (oval) Source of error indicated by narrow arrow; Input, flow and/or processing of data indicated by broad arrow."
The ovals of error types in these figures point to the stages of error sources, not to when the detection of errors might occur. The data concepts (squares) that are horizontally at same level are not necessarily related. [from: **(Zhang, 2012, p. 43)**]

Selection errors occur due to the difference between the actually *accessed set* and the accessible set, as a result of the reporting/recording process. In case of survey sampling data, the selection error is known as the sampling error, thus introduced by the fact that not all businesses are sampled. With regard to administrative registers an instance of selection errors occurs due to delayed reporting, for example if a company registers with the Chamber of Commerce mid-January but has already employed economic activity in the previous year.

In case of sample survey data, the difference between accessed set and *observed set* is known as the unit nonresponse or missing values. With regard to administrative register data, objects that are in the

accessed set might not be in the observed set because the register owner considers them inadmissible for administrative reasons. This results in *missing redundancy*. So missing redundancy only occurs in administrative register data if the register owner deletes an object or considers it not worthy to add, which is unlikely for the Chamber of Commerce's Trade Register since its purpose is to describe all possible *"business contacts"* (KvK, 2016a). It is also unlikely for the Chamber of Commerce as well as the Tax and Customs Administration's Relations Register to delete an object due to the collaboration with Statistics Netherlands to define the General Business Register.

The GBR - Phase One: Measurement Errors (cat)

The error sources concerning Measurement Errors (cat) in phase one are displayed at the left-hand side of Figure 1.2 (on page 8).

The *target concept* of a variable may be an abstract construct, while a *target measure* is concrete and observable. For example, information on NACE group is obtained from the Chamber of Commerce. The NACE code is the target measure to obtain a categorization per economic activity that is the target concept. Since NACE codes are agreed on periodically by the entire European statistical community, the available categories might not capture future economic activities entirely[3]. (Within countries also extensions of the NACE categories are formulated to describe specific economic activity of national importance. The so-called SBI (Standaard Bedrijfsindeling) implements such extensions with regard to The Netherlands (CBS, 2016d)). The incapability to capture a concept with a measure produces the *validity error*. This error occurs mainly in variables that are not of direct administrative importance, since otherwise they would fail to serve administrative purposes. The NACE code variable, for example, does not necessarily aid the Chamber of Commerce, but is administered mainly to provide to Statistics Netherlands.

The errors arising between the target measure and the *obtained measure* are referred to as *measurement errors*, in the classical statistical sense. These can occur, for example, if the businesses reporting data to the administration misinterpret the definition of a variable (Groen, 2012, p. 180), measure the value of a variable invalidly, or estimate a value instead of measuring it (Hoogland, Van der Loo, Pannekoek, & Scholtus, 2010, p. 4). With regard to the two sources that constitute the GBR, measurement errors can occur with regard to all identification and structural variables that are obtained from the Chamber of Commerce and Tax and Customs Administration. A measurement error on a NACE class (for example if a company misreported its main economic activity to the Chamber of Commerce) can be very influential in the production of statistics on subpopulations by economic activities. Statistics Netherlands therefore started cooperating with the Chamber of Commerce to specify the right NACE code for each newly registered business by offering them a software tool to determine the code (Konen, 2012). Unfortunately, erroneously assigned NACE groups still occur[4], also because when economic activities of businesses change these changes are not always updated at the Chamber of Commerce (Van Delden A. , Scholtus, De Wolf, & Pannekoek, 2014).

Any process needed to obtain an administrative data set can produce *processing errors*. These processes can be automatically, and the error be a software bug. The process can be manual work, such as filing written number into a computer or registering information heard over the phone. And in a sample survey primary phase, these processes can also involve editing or imputing data to obtain the *edited measures*. With regard to the identification and structural variables processed by the Chamber of Commerce and Tax and Customs Administration, these can be erroneously registered location or company name. But when these values are supplied by the business itself through online registering procedures, errors will mainly originate with the business itself and are therefore measurement errors.

[3] [4] During the course of this Master Thesis research, I founded my own statistical consulting business 'Significant Help'. When registering with the Chamber of Commerce, no Statistics Netherlands tool was used to decide on NACE group. I was assigned to the NACE group: "Technical exploration and development" (Dutch: "Technisch speur- en ontwikkelingswerk").
The Tax and Customs Administration still applies the NACE categories that originate from the eighties, and mapped my Chamber of Commerce NACE group to: "Technical, medical and natural scientific testing stations, laboratories" (Dutch: "Technische, medische en natuurwertenschappelijke proefstations, laboratoria").

The GBR - Phase Two: Data sources

When entering phase two, the input data is transformed from object to unit structure. The objects in the original sources represent *legal units* (Chamber of Commerce) and *fiscal units* (Tax and Customs Administration). Although linkage and micro integration are generally interconnected processes, in the flow of the diagram they are somewhat disconnected. The top part of the right-hand side of phase two in Figure 1.4 (on page 12) represents the matching and linking of units, the bottom part the micro integration. In this section, terms 'matching' and 'linking' are not used synonymously and also might have different meanings than elsewhere. *'Matching'* is used to describe the recognition of base units that need to be combined. *'Linking'* is used to describe the actual combining.

The GBR - Phase Two: Creating Statistical Business Units

The statistical business units are created with information on legal units from the Chamber of Commerce and information on fiscal units from the Tax and Customs Administration. For complex companies, legal and fiscal units can be quite different from the statistical business units of interest. Since companies use legal and fiscal constructions to limit their tax obligations and avoid company liability for separate projects, many legal or fiscal units can compose one statistical business unit.

Figure 1.3 (on page 11) shows nine ([A] -[I]) possible co-occurrences of legal units, fiscal units and statistical business units. The legal units (all small circles) form the *base units*, whose population is visualized by the largest brown/pink circle. The largest blue circle contains all legal units that are of interest to compose statistical business units. The blue ovals indicated by [A], [B] and [F] show *composite statistical business units* consisting of multiple legal units. In [A], one legal unit is not contained in a purple oval, or itself purple, and therefore exempted from paying certain taxes and reporting to the Tax and Customs Administration. An example of this situation in The Netherlands is the VAT tax exemption for certain small companies (Belastingdienst, 2016f) and agricultural companies (Belastingdienst, 2016e). The composite statistical business unit [B] is complex, but all composing legal units do report to the Tax and Customs Administration (all are purple or contained in a purple oval). [C] indicates a base legal unit that composes a statistical business unit, but does not report to the Tax and Customs Administration for certain taxes. The purple oval [D] shows two legal units composing one fiscal unit, that also occurs within [B]. This situation arises if the Tax and Customs administration allows two businesses to report taxes together because of ownership relations, for example if one company is main shareholder in the other company (Belastingdienst, 2016d). Such combined tax reporting can even differ between VAT and profit tax (Konen, 2012, p. 22) (Belastingdienst, 2016c). This specific duo [D] is not entirely contained in one statistical business unit, since one legal base unit constitutes its own statistical unit. An example of this situation occurs when one of the two companies has its own external sales market, which could evoke a separate statistical business unit from Statistics Netherlands' perspective. Fortunately, these complex situations mainly occur in the TopX, which is one of the reasons why these companies are mostly dealt with manually by so-called *profilers*. Outside of the TopX, most companies defined by a legal unit also constitute a fiscal unit and a statistical business unit [E] or only a few individual legal units are combined into a composite statistical business unit that all report to the Tax and Customs Administration [F]. Some square base units are located outside the population of Chamber of Commerce's legal units, which visualizes over-coverage of the Tax and Customs Administration's Relations Register of fiscal units [G], [H], and 'black market' companies that are not registered at all [I]. [G] and [H] occur, for instance, if parts of international companies are only represented fiscally in The Netherlands. Another example concerns companies that use The Netherlands for the transition of goods. Such companies are manually identified and excluded from the statistical business unit population (Van Delden & De Wolf, 2013). Example institutions that are automatically excluded are pension funds, which are not regarded as statistical business units, but are registered with the Tax and Customs Administration (Aelen, 2005).

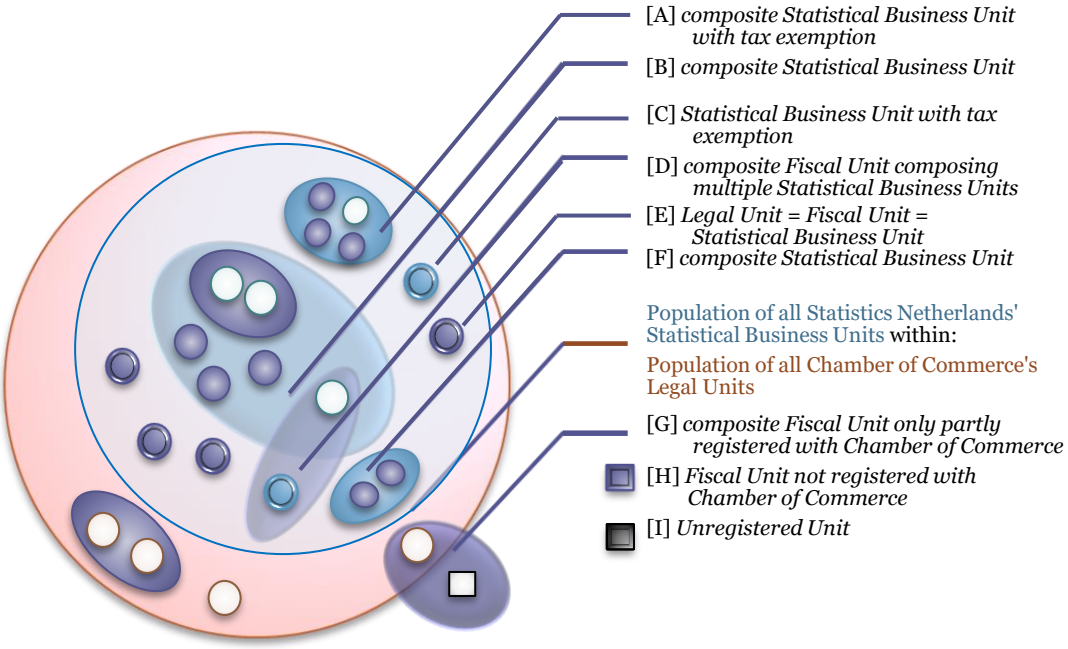


Figure 1.3 Graphical representation of nine ([A] - [I]) co-occurrences of Chamber of Commerce's Legal Units (*smallest circles within brown/pink outer circle*), Statistics Netherlands' Statistical Business Units (*blue circles/ovals within the outer blue circle*) and Tax and Customs Administration's Fiscal Units (*purple circles/ovals*).
Legal units that are not purple, or contained in a purple oval, represent legal units that do not report to the Tax and Customs Administration with regard to certain taxes. Units outside the pink circle are not registered with the Chamber of Commerce.

The GBR - Phase Two: Representation Errors

The error sources concerning Representation Errors in phase two are displayed at the right-hand side of Figure 1.4 (on page 12).

The *target population* is the set of statistical units that the statistics should ideally cover. *Linked sets* are sets of legal units and fiscal units that could be matched to constitute a statistical business unit (in this flow chart the actual linkage takes place in the next step). Statistical business units that are in the target population, but cannot be constructed from recognized combinations from the input sources, or units that could be matched but are not part of the target population form the (under and over) *coverage errors*.

Identification errors also result from the process of statistical unit construction, but they are not due to the process of linking, but due to the input variables used for linkage. Different data sources can contain conflicting information, for example in case a company has changed its name and reported to the Chamber of Commerce but this name change did not reach the Tax and Customs Administration. Information on the original legal and fiscal units needs to be combined with other information to obtain *aligned sets* (for example by using information from the company's websites). Obtaining aligned sets is a form of micro integration. Errors that occur in this process of alignment are *identification errors*. As is clear from Figure 1.3 sometimes many decision rules need to be applied to construct a statistical business unit, and these decision rules might be very dependent on input from the original sources. Therefore, identification errors might not be uncommon in the GBR.

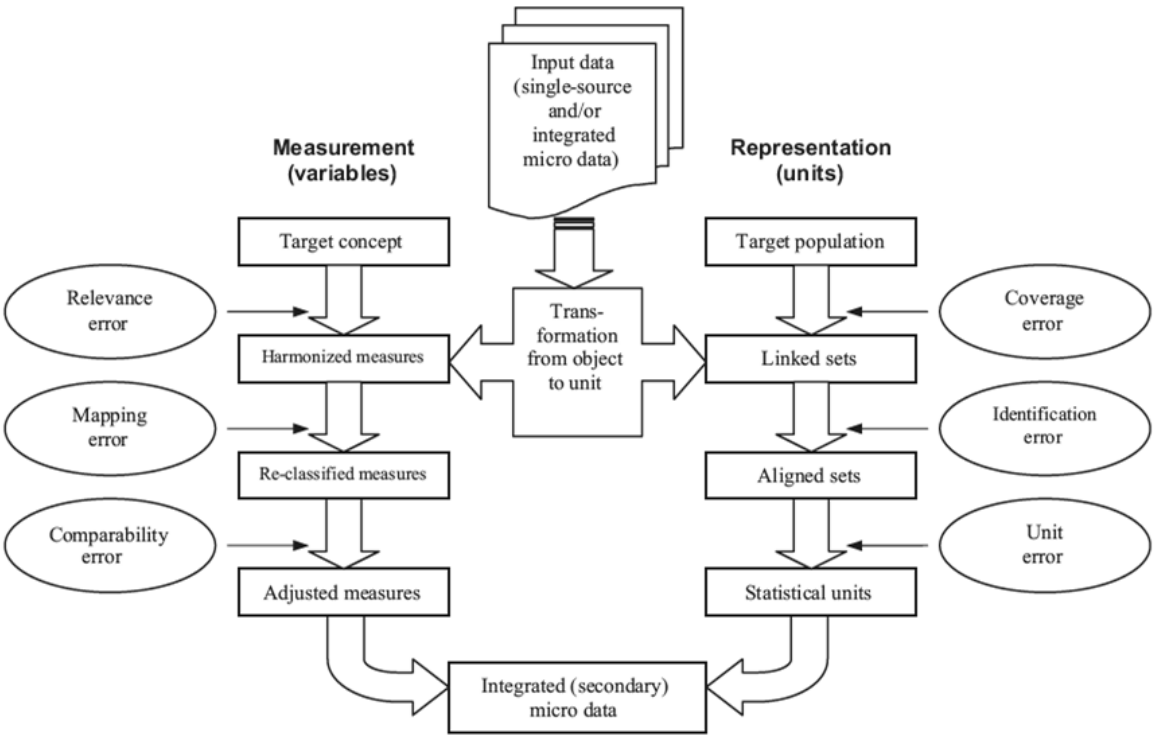


Figure 1.4 Phase Two of "Two-phase (primary and secondary) life cycle of integrated statistical micro data from a quality perspective. Data concept (square); Error Type (oval) Source of error indicated by narrow arrow; Input, flow and/or processing of data indicated by broad arrow."
The ovals of error types in these figures point to the stages of error sources, not to when the detection of errors might occur. The data concepts (squares) that are horizontally at same level are not necessarily related at each level. [from: **(Zhang, 2012, p. 43)**]

The actual creation of the *statistical units* from the aligned sets can result, if carried out erroneously, in a *unit error*. The GBR is a continuously updated register with regard to the identification variables and structural variables (on the Measurement side of the second phase). With regard to the unit construction, it is updated each month, along with the external framework containing the list of changes (Konen, 2012). An automatic procedure handles the changes due to, for example, mergers, or newly established and abolished companies. At macro level a manual procedure is carried out to decide if these changes are plausible with respect to historical reference data (Aelen, 2005). Also, changes to TopX businesses are carried out manually. A manual procedure can produce errors and erroneously automatically constructed units can slip through the macro level check and result in unit errors.

The GBR - Phase Two: Measurement Errors (cat)

The error sources concerning Measurement Errors (cat) in phase two are displayed at the left-hand side of Figure 1.4.

When constructing composite statistical business units in the GBR from the base legal units and fiscal units, identification variables and structural variables on the base units, such as NACE groups, Size Classes and hierarchical structure of companies, are also combined. Name, location and legal form values follow from the business part with the largest number of employees, with the number of employees provided by the Tax and Customs Administration (Konen, 2012). The Size Class of a composite unit comes from combining Size Class information (number of employees) from the original fiscal units that are combined. For businesses without employees, information on the composing legal units from the Chamber of Commerce is used (if no employees, the number of owners constitutes the Size Class) (Aelen, 2005). The main NACE group is also defined by the business part with the largest number of employees involved, and is established per hierarchical NACE code level

(Aelen, 2005). So, with regard to the level coding in Table 1.1 (on page 4) by first establishing the Section with the highest number of employees, then the Division, the Group and by concluding with the Class.

To obtain *harmonized measures*, no modification occurs to the data, but a kind of conceptual alignment of metadata from different sources is carried out to approximate the *target concept*. An example of harmonizing is the assignment of the main NACE group from multiple NACE groups involved in the base units composing a composite statistical business unit. Harmonization is obtaining a common standard. If the best common standard does not correspond to the target concept, a *relevance error* occurs. For example, if the main NACE group cannot be approached by the standard procedure if multiple business parts active in other economic areas are equally important in terms of number of employees.

Obtaining *re-classified measures* has a similar purpose to obtaining harmonized measures, but does modify the data. The input measures of various sources are mapped to standard ones. If source classifications are erroneously mapped to standard ones, *mapping errors* occur. For example, when the number of employees reported by the Tax and Customs Administration is incorrect, the Size Class cannot be correctly determined by combining this with information of number of owners from the Chamber of Commerce.

Adjusted measures are obtained by all further data editing such as imputation of suspicious and missing values. In contrast to obtaining edited measures in the first phase, the data records become suspicious in the second phase as a result of combining the input data sources. This is also a form of micro integration. Each source itself might be consistent and complete and not constitute a quality problem from the register owner's perspective, but problems still arise by the combination of sources for statistical purposes. For example, the Tax and Customs Administration might have a longer period in which a business needs to report a changed company name than the Chamber of Commerce. As a result, the two registers are inconsistent, but neither of them lacks quality from the owner's point of view. When combining sources such inconsistencies might evoke editing of values, which results in *comparability errors* when erroneously executed.

1.2.3.2 BaseLine: linking yearly financial data to the population framework

The population of Statistical Business Units, the GBR, in which errors occur in the GBR execution of the Zhang Two-Phase Life-Cycle Model, is used as sampling frame for the *Structural Business Statistics (SBS)* sample survey. With regard to the SBS sample survey data collection, the first phase is executed again. In this first phase execution, the possible error sources are exactly those in the original model for sample survey errors on which the Zhang Two-Phase Life-Cycle Model is based (Groves, et al., 2004). In the second phase, also yearly values from the Tax and Customs Administration's *VAT register* and *Profit tax register (PDR)* are linked to the GBR framework in BaseLine. Both tax registers have their own phase one at the hands of the original register owner and in relation to the administrative purpose of the register.

Since the population of businesses is defined by the already existing GBR, in the second phase of the BaseLine execution only Measurement Errors (cat) occur. Thus the representation side of the Zhang two-phase lifecycle model is not discussed with regard to phase two in the presentation below. With regard to the sample survey data there is no second phase, since the data is already collected on the GBR's statistical business units and not further combined or transformed.

BaseLine - Phase One: Representation Errors
The error sources concerning Representation Errors in phase one are displayed at the right-hand side of Figure 1.2 (on page 8).

Only *selection errors* and *missing redundancy* are likely to occur in the data from the SBS sample survey. The sampling framework is created in the GBR execution of the Two-Phase Life-Cycle Model, therefore frame errors do not originate in the BaseLine execution. Selection errors occur in terms of sampling errors, and missing redundancy in terms of nonresponse in the sample survey.

Yearly financial data from the Value Added Tax register and Profit Declaration Register might contain frame errors, selection errors and missing redundancy. Businesses might avoid paying either VAT or profit tax and are therefore not listed in the VAT register or the PDR, resulting in a *frame error*. Delayed reporting to the VAT register or PDR results in a *selection error*. However unlikely, when an object in the VAT register or PDR is deleted by the Tax and Customs Administration for administrative purposes, the error of *missing redundancy* occurs.

BaseLine - Phase One: Measurement Errors (cat)
The error sources concerning Measurement Errors (cat) in phase one are displayed at the left-hand side of Figure 1.2 (on page 8).

The sample survey data can contain, *validity errors*, *measurement errors* and *processing errors*. Validity errors are assumed to be less influential, since a financial target concept can be described as a target measure quite clearly in a survey. Small differences between the two occur for example if the description of the target measure is too elaborate to communicate exactly to the business, and therefore a simplified measure is used. An example of measurement error is a reported turnover value in euros instead of €1000, as is requested in the SBS survey. A processing error occurs, for example, if a value is correctly supplied on a paper survey but erroneously registered in a computer (when a 7 is mistaken for a 1, for example), or if an error occurs in editing or imputation by Statistics Netherlands.

Also, only *measurement errors* and *processing errors* are to be expected in the Tax and Customs Administration's registers, but no validity errors. Administrative sources are generally regarded to be validity error free (Zhang, 2012, p. 47), since target measures are formulated to aid the administrative purposes. Also, a Tax and Customs Administration can ask for elaborately defined financial target concepts since businesses are obliged to report taxes and therefore nonresponse concerns are of a different order than they are for Statistics Netherlands. The absence of validity errors is limited to this phase one definition, that is related to the administrative use of financial variables. A distinction with the statistical target concept is encountered in the second phase of statistical reuse of the administrative data. Measurement errors might occur if a business, for example, reports an estimate of its turnover instead of a measurement. Processing errors might occur at the Tax and Customs Administration in the form of a software bug, or when tax reports are processed from paper to computer.

BaseLine - Phase Two: Measurement Errors (cat) (only administrative register data)
The error sources concerning Measurement Errors (cat) in phase two are displayed at the left-hand side of Figure 1.4 (on page 12).

In case of fiscal exemptions of some legal units, such as graphically depicted by co-occurrences [A] and [C] in Figure 1.3 (on page 11), the difference between turnover concept with regard to the VAT or profit tax and the statistical turnover concept needs to be harmonized. For example, by imputing the turnover for the tax exempted business part. If harmonization results in a target concept different from the harmonized measure, a *relevance error* occurs. This not only the case for composite statistical business units, since also the turnover concept for a simple unit can differ between Tax and Customs Administration and Statistics Netherlands, denoted by *reporting differences* (Groen, 2012). Such situation was previously researched at Statistics Netherlands with regard to the VAT data in comparison to survey collected data (Aelen, et al., 2011) (Van Delden, Pannekoek, Banning, & De Boer, 2016). VAT turnover values are influenced by regulations with regard to, for example, international sales and sales of VAT-exempted goods, such as second hand cars (that are purchased VAT-free) (Lammertsma, 2016). In the latter example businesses are allowed to report their profit to

the Tax and Customs Administration and not their turnover, which is not required to be indicated in the tax report. For some NACE groups, a significant deviation between the values from the survey and the values from the VAT register was found, resulting in a proposed correction on VAT values when used to compile Short-Term Business Statistics (Van Delden & De Wolf, 2013). Such correction is another example of harmonization, which can result in a relevance error if the correction is estimated erroneously or the necessity to correct is erroneously applied or discarded.

If not the concept of the correction, but the correction itself, is carried out erroneously on a specific unit, a *mapping error* occurs. These errors occur mainly as a result of errors in the original sources. An example of such mapping errors occurs in VAT turnover values that are quarterly reported (VAT reporting can be monthly, quarterly or yearly in The Netherlands). Some quarterly reported VAT turnover show patterns that are not likely to represent real turnover values, such as three out of four quarters with exactly the same turnover (even three quarters with zero turnover, unlike the fourth quarter) (Van Delden A. , 2013, p. 12). Since the Tax and Customs Administration is mainly interested in yearly paid VAT taxes, these reported values are not corrected with regard to the aims of the original register owner (in phase one). But mapping errors occur in the statistical reuse of the data for quarterly economic statistics (such as the Short-Term Business Statistics). However, these patterns are less problematic for the yearly Structural Business Statistics.

Although a harmonization procedure can include imputation, *comparability errors* arise with regard to editing and imputation in the combined data set. Such editing or imputation follows from errors or missing values that are detected when multiple sources are combined. The Intermittent-Error Model is one way to detect and correct errors in multi-source data. This study investigates, among other things, if the Intermittent-Error Model's performance could lead to comparability errors when applied to the case study data set.

1.2.3.3 Mixed-execution errors

The error types distinguished in the previous section are sometimes interconnected across phases and executions. Take for example, over-coverage of inactive businesses in the GBR. A company deregistering with the Chamber of Commerce is also erased from the GBR, but many inactive businesses will not deregister. An inactive business will not report to the Tax and Customs Administration since it does not have to pay taxes anymore, but these missing values will initially be considered delayed reporting. Thus with regard to the Chamber of Commerce's Trade Register and the Tax and Customs Administration's Relations Register a *frame error* occurs (GBR execution, phase one, representation side). The frame error influences the GBR as a *coverage error* in the maintenance of the statistical business units that constitute the business population of the GBR (GBR execution, phase two, representation side). Therefore, these missing values might be imputed in the BaseLine execution by Statistics Netherlands (which is of course erroneous imputation, but not a comparability error since the origin of the error (error source) lies in erroneous representation of the population). Even if a company reports to Statistics Netherland in the SBS survey to no longer exist, the GBR is not immediately updated. Since already published statistics were based on previous versions of the GBR, not immediately updating the GBR assures consistency across publications.

A second example of an error that can occur across executions deals with *subpopulation representation errors*. The population of statistical business unit can be subdivided into subpopulations per NACE group. Data on these subpopulations can have characteristics specific to the NACE group and are therefore also often used in research, such as in the application of the Intermittent-Error Model (as discussed in Chapter 4). An erroneously assigned NACE group can result from a *validity error* or *measurement error* (GBR execution, phase one, measurement side), or *relevance error* or *mapping error* (GBR execution, phase two, measurement side) with regard to composite statistical business units, which yields an *identification error* (GBR execution, phase two, representation side) with regard to the incorrect NACE group subpopulation and the true NACE group subpopulation. These errors can be very influential for economic statistics per economic activity (Van

Delden, Scholtus, & Burger, 2016). In addition, NACE groups are used to decide whether or not to correct turnover values obtained from the VAT register for the Short-Term Business Statistics. If businesses are assigned to incorrect NACE groups, *relevance errors* occur with regard to turnover values (BaseLine execution, phase two, measurement side) when such businesses are included in the harmonization research to define the correction, and *mapping errors* occur when the business' turnover value is erroneously corrected.

1.3 Apparent errors in a case study data set

This section introduces the case study data set in more detail. It discusses how the case study data set was selected (Section 1.3.1), what variables are available in the case study data set (Section 1.3.2) and shows examples of contradictions among variables (Section 1.3.3) and some typical (possibly erroneous) records in the data set (Section 1.3.4).

1.3.1 Case study data selection

The data originates from the 2012 state of the GBR for a specific set of NACE groups. With regard to yearly financial business data, the data from the Structural Business Statistics sample survey is available for a sample of the businesses in these NACE groups. The year 2012 also allows quite complete Profit Declaration Register data from BaseLine to be linked to the survey data, since delayed reported Profit tax was linked in the years that followed 2012. The Structural Business Statistics are generally published 16 months after the end of the year (CBS, 2016e), the PDR data was used which was available by October 2016. The VAT data was already collected in BaseLine on quarterly basis to produce the Short-Term Business Statistics of the year 2012, and was also available as annual VAT Turnover values.

Selection of the data set NACE groups
Prior to the use of VAT data to publish the Short-Term Business Statistics, research was carried out into deviations between the VAT turnover concept and the turnover concept in the SBS sample survey. Too large deviations between these two are undesirable, since as a result also the obtained published (quarterly) Short-Term Business Statistics based on VAT data would deviate from the (yearly) Structural Business Statistics. Consistency among published statistics is one of the main quality measures applied by Statistics Netherlands (Bakker B. F., 2011, p. 6).

The research into the VAT and SBS turnover concepts was carried out in two ways. Firstly, by expert investigation of tax definitions and exemptions in various NACE groups in comparison to the SBS definition (Lammertsma, 2016). This showed, among others, deviating VAT turnover definitions for businesses trading in second-hand cars (see Section 1.2.3.2, Phase Two, Measurement Errors (cat)). But also SBS turnover definitions vary across NACE groups. For example, in turnover values regarding construction, trade and industry, turnover includes transport costs by third parties, while this is not the case with regard to retail, transportation and commercial services (Aelen, et al., 2011).

Secondly, data driven research was carried out to compare obtained VAT turnover values to SBS values. This research showed that NACE groups had to be divided into three sets. One for which the VAT data was found to be reliable in comparison to the SBS data, one for which a systematic deviation was found that could be corrected, and one group for which the VAT data was found to be completely unreliable (Van Delden, Pannekoek, Banning, & De Boer, 2016), (Van Delden & De Wolf, 2013). The possible correction in the second set was perceived to be linear at the macro level of NACE group subpopulations (Van Delden & De Wolf, 2013). These linear deviations of a source concept in comparison to the target concept are referred to as *intercept bias* (Scholtus, Bakker, & Van Delden, 2015) and *slope bias*. When aggregated values, such as total turnover per NACE group, are based on a source with intercept/slope bias, a biased estimate of the statistic is produced. Correcting for this bias is a form of harmonization.

In the Short-Term Business Statistics, the VAT data is only used in NACE groups for which both the expert's opinion and the data driven research found that the VAT data was reliable, or it was found that the systematic deviation could be reliably corrected. The case study data set contains eight quite randomly chosen NACE groups for which VAT turnover values are currently not used to produce the Short-Term Business Statistics.

Edited values
The SBS survey turnover values were used to publish the Structural Business Statistics of the year 2012 and are therefore edited at the micro level (see Section 1.4). The VAT data of the NACE groups in the case study data set was not used for publication and are therefore also not edited. VAT turnover values for business with less than 4 quarters reported were imputed to yearly values. The turnover values from the Profit Declaration register were collected and linked by Statistics Netherlands for research. These were not yet used for any publication, and therefore not edited.

1.3.2 Case study data variables overview

Table 1.2 shows the variables in the case study data set and the sources they originate from. The GBR column shows the identification and structural variables that accompany the statistical business units in the GBR population framework. The SBS column shows some relevant variables obtained from the Structural Business Statistics sample survey, although in a simplified form, since the SBS survey is quite elaborate and detailed and not all details are relevant to obtain turnover totals. The SBS values are only available on a sample of the GBR units. The variables in the VAT and PDR columns are obtained from the Tax and Customs Administration and linked to the GBR in BaseLine. The VAT values are originally collected to produce the Short-Term Business Statistics. Therefore, only *Turnover* is available since the Short-Term Business Statistics only published turnovers and developments in turnover. From the PDR register also *Costs* and *Purchases* are obtained for possible future research, but these are not of interest to this study.

Table 1.2 Sources and Variables in the case study data set.
GBR: General Business Register; SBS: Structural Business Statistics survey; VAT: Value Added Tax register; PDR: Profit Declaration Register
The variables shown in bold face in the lower part are used to fit the Intermittent-Error Model.

Statistics Netherlands		**Tax and Customs Administration**	
GBR	SBS	VAT	PDR
Statistical Business Unit ID	*Manually Edited*	*VAT Quarters*	*Coverage Percentage*
NACE group	*Automatically Edited*	*Response Percentage*	*Fill Percentage*
Size Class	*Weight*		*Unlock Code*
TopX			
Legal form			
Number of Employees	*Number of Employees*		
	Turnover	***Turnover***	***Turnover***

GBR
The units in the GBR are all assigned a random identification number, the *Statistical Business Unit ID*, that is used to link the units to information from other sources. Thus real identification information (name, identification number with the Chamber of Commerce/Tax and Customs Administration) is already removed and specific companies are harder to identify in a processed data set like the case study data set. The case study data set consists of 37 462 statistical business units, from eight NACE groups. These eight *NACE groups* are shown in Table 1.3. At the broadest level, the units belong to the groups G and H, which are *'Wholesale and retail trade; repair of motor vehicles and motorcycles'* and *'Transportation and storage'* respectively. All NACE codes, but the first, represent groups as prescribed by Eurostat. *G45.1.1.2* resulted from an additional distinction applied by Statistics Netherlands with regard to imported and non-imported car sales, which is an important distinction with regard to second-hand car sales. The eight NACE codes shown in Table 1.3 represent

the groups for which macro statistics at the business sector level are published by Statistics Netherlands. In total, publications are produced for 276 different NACE groups (Meijers & Smeets, 2011, p. 4). NACE group *G45.1.1.2* has a very high frequency in the case study data set, with almost 50% of the total businesses in the eight NACE groups categorized by *G45.1.1.2*. NACE group *H52.1.0* and *H50.1.0* occur the least in the case study data set, with only 764 (~2%) and 943 (~3%) occurrences respectively (see Table 1.7 on page 24 for a complete overview of NACE group occurrences).

Table 1.3 Overview of NACE groups in the case study data set

NACE CODE	DESCRIPTION
G45.1.1.2	Sale and repair of passenger cars and light motor vehicles (no import of new cars)
G45.1.9	Sale and repair of trucks, trailers and caravans
G45.2.0	Specialized repair of motor vehicles
G45.4.0	Sale and repair of motorcycles and related parts
H50.1.0	Sea and coastal passenger water transport and ferry-services
H50.3.0	Inland passenger water transport and ferry services
H52.1.0	Warehousing and storage
H52.2.9	Forwarding agencies, ship brokers and charterers; weighing and measuring

There are ten *Size Classes* defined within Statistics Netherlands, which are numbered 0 to 9. Their meaning is shown in Table 1.4. The values on *Number of Employees* on which these *Size Class*es are based are also a structural variable in the GBR. One employee corresponds to a fulltime equivalent occupation, so could contain multiple people. Therefore, '0 employees' means that no fulltime equivalent occupation is worked, but not necessarily that the business is empty. Moreover, the 0 *Size Class* categorization is also used for businesses with unknown number of employees (De Wolf & Van Delden, 2011, pp. 8-9). Therefore, many more businesses are assigned to *Size Class* 0 then there should be, and these businesses have much larger turnover values than expected based on the small size. Figure 1.5 (on page 20) and Figure 1.6 (on page 21) show the frequencies of the *Size Class*es within each NACE group, and the proportions of the *Size Class*es that belong to the *TopX* (see Section 1.2.1).

Table 1.4 Overview of *Size Classes* in the data set **(CBS, 2016b, p. 30)**

0	Statistical Business Units with 0 or unknown number of employees
1	Statistical Business Units with 1 employee
2	Statistical Business Units with 2-5 employees
3	Statistical Business Units with 5-10 employees
4	Statistical Business Units with 10-20 employees
5	Statistical Business Units with 20-50 employees
6	Statistical Business Units with 50-100 employees
7	Statistical Business Units with 100-200 employees
8	Statistical Business Units with 200-500 employees
9	Statistical Business Units with 500 or more employees

More than 70% of the statistical business units in this GBR subpopulation belong to *Size Class* 1 or 2, as shown in Figure 1.5 and Figure 1.6 (on page 20 and 21), and are thus small businesses with 1 to 5 employees. Overall, 1,3% of these statistical business units is characterized as TopX. Of these, most (in absolute terms) occur in NACE groups *H52.2.9* and *G45.1.1.2* (132 (4%) and 123 (0.7%) respectively) and few (in absolute and relative terms) occur in NACE group *G45.4.0* (4 out of 1763 (0.2% of units in this NACE group)).

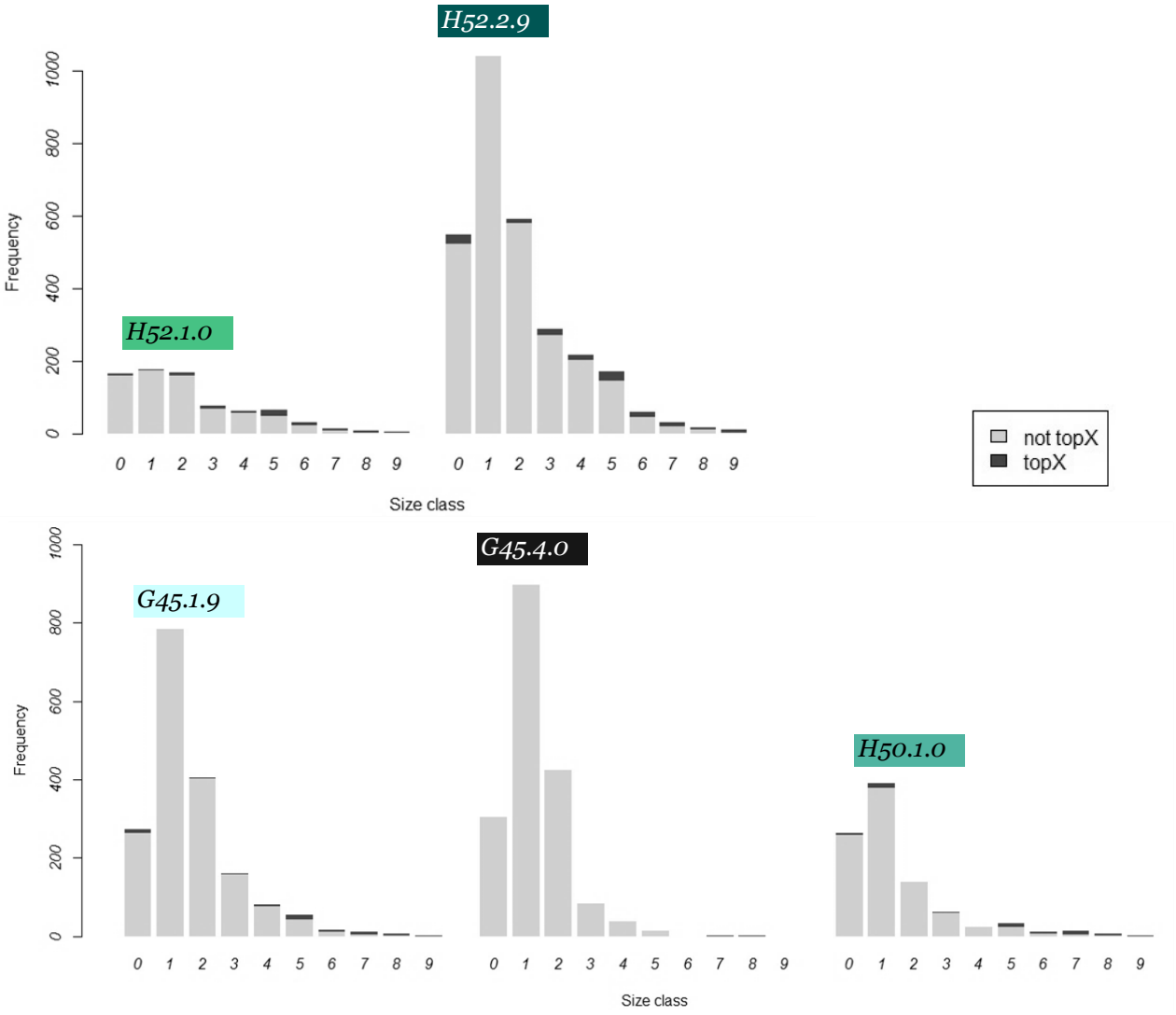


Figure 1.5 Histograms of *Size Class* frequencies and TopX occurrences within *Size Classes*, for the NACE groups with less than 3000 businesses in the GBR: *H52.1.0, H52.2.9, G45.1.9, G45.4.0* and *H50.1.0*. The NACE groups with larger number of businesses are shown in Figure 1.6 (on page 21).

Statistical Business Statistics sample survey

The Statistical Business Statistics sample survey provides turnover values on a sample of 2829 statistical business units, which is 8% of the total amount of units within these NACE groups in the GBR. Since *Size Class* 0 businesses are either very small or unknown, none of these are sampled to produce the SBS. Instead these businesses' contribution to economic statistics is estimated from known average values.

All records are automatically checked and corrected for the most basic systematic errors, such as turnover values reported in euros instead of €1000, and falsely reported negative values. On top of that, inconsistences, such as purchases, turnover and profit that are not in agreement, can be automatically detected and corrected with an optimization procedure. The SBS records are accompanied by variables indicating whether they are *manually* or *automatically edited*. Automatically edited means that an automatic optimization procedure has run that checks and corrects inconsistencies. Manually editing means that the record entered a procedure in which it is automatically checked, and scrutinized by a person if a detected error or suspicion record is deemed important enough to be corrected manually. Therefore, both the manually edited and the automatically edited variable do not necessarily mean that the value has been changed. (For more details, see Section 1.4.) Automatically edited does mean that the record is regarded consistent, either with or without correcting a value. Manually edited means one of two things: (1) inconsistencies or

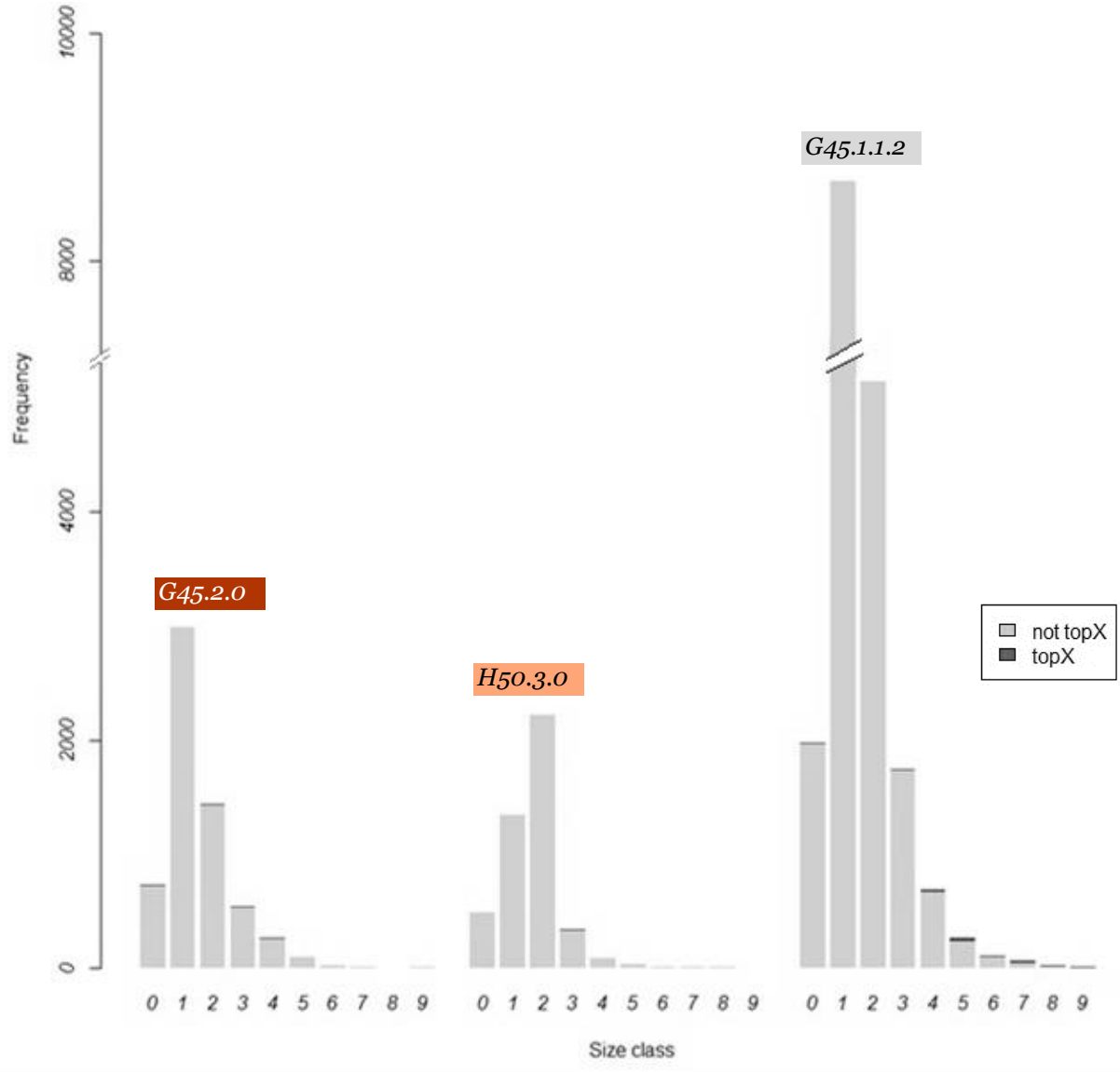


Figure 1.6 Histograms of *Size Class* frequencies and TopX occurrences within *Size Classes*, for the NACE groups with more than 3000 businesses in the GBR: *G45.2.0*, *H50.3.0* and *G45.1.1.2*.

unlikely values in the records are corrected or (2) inconsistencies or unlikely values in the record are not influential enough to spend resources on correcting the record (influence with regard to the published economic statistics that will be based on the data).

For 35% of the records, automatically editing obtained records of sufficient quality. These records are shown in the first 7 bars in the histogram of Figure 1.7 (on page 22) and as shown none of them belongs to *Size Class* 8 or 9. On 2% of the records the automatic procedure detected an inconsistency, tried to alter the values but was not successful. These records are depicted by the last few bars in the histogram of Figure 1.7 and all occur in *Size Class*es 1, 2, 3, 4, 5 and 7. The rest of the records were directly manually edited or successfully automatically edited but deemed influential enough to also enter the manually editing process. All four groups depicted in Figure 1.7 occur in equal proportions across NACE groups. With regard to *Size Class*es, the larger the business unit, the smaller the probability to be automatically edited and the larger the probability to be manually edited. But since there are so much more smaller *Size Class* business units than larger ones (see Figure 1.5 and Figure 1.6), the smaller ones are more often manually edited in absolute numbers than proportionally, as shown in Figure 1.7. All *Size Class* 8 and 9 business units are manually edited.

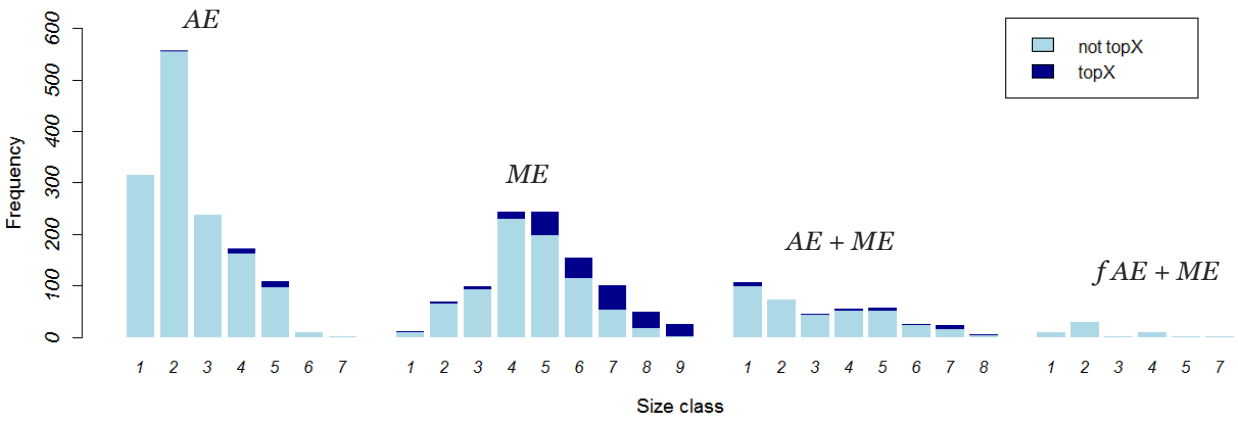


Figure 1.7 Histograms of *Size Class* frequencies and *TopX* occurrences within *Size Classes* for SBS records by edit procedure. *AE: Automatically Edited; ME: directly Manually Edited; AE + ME: Automatically Edited and additionally Manually Edited; fAE + ME: failed Automatically Edited and thus Manually Edited.*

Weights are included that were used to calculate population totals from the obtained sample. The Structural Business Statistics are sample survey based, and therefore the aggregated values result from a weighted combination of the obtained sample and imputed values. These weights are partly based on the stratified sampling procedure and the unit nonresponse, with extra weighing by number of employees and legal form (CBS, 2016b) (Aelen, et al., 2011).

VAT Register

The variable *VAT Quarters* indicates how many complete quarterly VAT tax returns were provided by the statistical business unit. In 13% of the records this variable is missing, 74% of statistical business units reported on all four quarters, 5,5% on only three, 3,5% on 2 and 4% on 1. The VAT Turnover values are increased to yearly values by imputation.

The *Response Percentage* indicates for which proportion of the statistical business units (often constituted by multiple legal units with individual tax returns) information on turnover is available from the VAT register. Differences have to do with fiscal units being exempted from VAT taxes and nonresponse/delayed response with the Tax and Customs Administration (CBS, 2016c) (see Figure 1.3 on page 11). For 6% of the records for which VAT information is available this value is missing, 88% of records have a *Response Percentage* of 100%, 3% of records have *Response Percentage* of 0%, which occur only in *Size Classes* 0-7 (in most NACE groups in less than 5% of records except *H50.1.0* (18% of records) and *H52.1.0* (8% of records)). In *Size Class* 9 all non-NA *Response Percentages* are 100%, in the other *Size Classes* the percentages do not vary much around the mean of 94%. All NACE groups are quite comparable, except NACE group *H50.1.0* for which only 78% of records have a 100% *Response Percentage*. Some values are clearly erroneous: Two values are negative (both in NACE group *H50.1.0*, *Size Class* 0 and 1), and three values are larger than 100% (in NACE group *H50.1.0* and *G45.1.1.2*, *Size Class* 0 and 1).

Profit Declaration Register

The variable *Coverage Percentage* indicates which proportion of a statistical business unit is obliged to pay Profit Tax (some legal units constituting the statistical business unit might be exempted). This proportion is measured with regard to the number of employees in the tax paying part of the company in comparison to the total number of employees. 98% of the records has a *Coverage Percentage* of 100%. The *Fill Percentage* describes which proportion of those tax paying units actually appears in the Profit Declaration Register at the moment it was linked to the GBR. Nonappearance can have to do with delayed reporting. Or, for example, if one fiscal unit is part of two statistical business units, the

information on the fiscal unit from the Profit Declaration Register cannot be used for either of the two statistical business units (co-occurrence [D] in Figure 1.3 on page 11). 99% of the records has a *Fill Percentage* of 100% and 1% has a *Fill Percentage* in the range (16%-98%).

The *Unlock Code* indicates whether the linkage to the GBR was successful, and assigns one of seven possible codes shown in Table 1.5. PDR *Unlock Code*s are available on all units in the SBS sample. Most statistical business units have *Unlock Code* A and C, as shown in the third column of Table 1.5. Table 1.5. also shows how often certain *Coverage Percentages* occur with each *Unlock Code*. Records with *Unlock Codes B*, *D*, *W*, *X* and *Z* are most problematic, and occur in all *Size Class*es and NACE groups. 80% of business units with *Unlock Code W* are classified as TopX, none with *Unlock Code B* and only one with *Unlock Code D*. No PDR turnover values are available for units with *Unlock Code W*, *X* or *Z*.

Table 1.5 Overview of PDR *Unlock Codes* with percentages of occurrence within the GBR, and occurrences of *Coverage Percentages* within *Unlock Codes*.

Unlock Code	*Description*	*Co-occurrence in Figure 1.3 (on page 11)*	*Percentage of GBR units*	*Coverage Percentage within Unlock Code*		
				0%	> 0% < 100%	100%
A	One PDR reporting fiscal unit constitutes one statistical business unit	[E]	83%		0.1%	99.9%
B/D	One or multiple PDR reporting fiscal unit contributes to one statistical business unit which also consists of other legal units that are exempted from reporting to the PDR	[A]	2%		100%	
C	Multiple PDR reporting fiscal units constitute one statistical business unit	[F]	6%		100%	
W	One PDR reporting fiscal unit constitutes multiple statistical business units	[D]	1%		4%	96%
X/Z	All legal units constituting the statistical business unit are exempted from reporting PDR/ No legal unit constituting the statistical business unit is linked to the PDR	[C]	8%	100%		

Variables available to fit the Intermittent-Error Model

Figure 1.8 (on page 24) shows a graphical representation of the three sources on the target variable *Turnover* and covariate *Number of Employees*. The VAT data provides *Turnover* values for 97% of all statistical business units in the case study data set. The sample survey data for the Statistical Business Statistics (SBS) provides values for 8% of the statistical business units, of which 0.7% (21 records out of 2829) are not covered by the VAT data. The Profit Declaration Register (PDR) provides data on 78% of the statistical business units in the case study data set, of which 0.4% (112 records out of 29252) are not covered by the VAT data. So of 1035 units (3% of the population) no turnover value from either of the four sources is available. Per NACE group these percentages range between 2,5% and 3,5% with the largest amount in NACE group *G45.2.0*. For 2388 units (6% of the population) a turnover value is available from all three sources.

Table 1.6 (on page 24) shows the weighted total SBS turnover and the summed VAT and PDR total turnover. The weighted SBS turnover is by far the largest estimate of the total turnover in the eight NACE groups in the year 2012. This can be partly explained since the SBS value is an estimate for the turnover in the entire GBR population, which is larger than the subpopulations covered by VAT and PDR values. The PDR total turnover is larger than the VAT total, even though the number of businesses for which PDR turnover is available is smaller than those for which VAT turnover is available. So the PDR might structurally overestimate the turnover values, or the VAT underestimates them. This assumption is confirmed by the totals for a subgroup of the businesses for which the three sources all report turnover values (also shown in Table 1.6) for which a difference in sampled businesses (large ones in the PDR register, small ones in the VAT) is rejected as sole explanation.

The occurrences of NACE groups within sources is shown in Table 1.7. NACE group *G45.1.1.2* is by far the largest NACE group in the data, with almost 50% of the statistical business units in the eight NACE groups' GBR population belonging to this NACE group. In the subgroup of the data with values on turnover from all three sources (the overlapping part of Figure 1.8), 37% of the units belong to NACE group *G45.1.1.2*. This is mainly due to the more limited occurrence of this NACE group in the SBS data, due to the stratified sampling approach. Within this stratified sampling procedure small businesses have a smaller probability to be sampled, and since NACE group *G45.1.1.2* contains many small businesses (see Figure 1.6, on page 21), less businesses within this NACE group are sampled in comparison to its GBR population.

Table 1.6 Weighted total SBS *Turnover*, total VAT *Turnover* and total PDR *Turnover*

SBS total *Turnover*	**VAT total *Turnover***	**PDR total *Turnover***
68 742 424 000	59 130 635 000	59 478 799 000
SBS total *Turnover* NACE group *G45.1.1.2*, 885 overlapping businesses	**VAT total *Turnover*** NACE group *G45.1.1.2*, 885 overlapping businesses	**PDR total *Turnover*** NACE group *G45.1.1.2*, 885 overlapping businesses
11 636 718 000	8 903 517 000	11 472 401 000

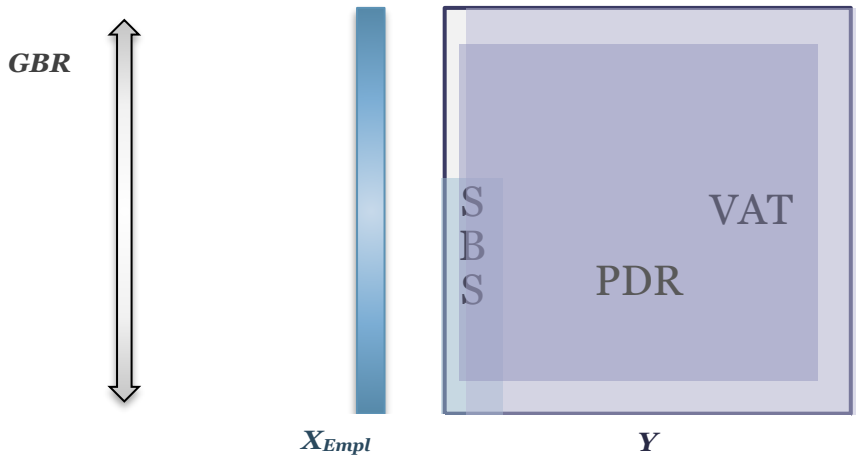


Figure 1.8 Graphical representation of three data sources on the target variable *Y* (*Turnover*).
Outer box: GBR population; Light purple rectangular area: VAT observations; Dark purple square area: PDR observations; Light blue rectangular area: SBS observations.
The covariate X_{Empl} (Number of Employees) is provided by the GBR itself, and available on all statistical business units.

Table 1.7 NACE group occurrences in data set components. The percentages represent the division of NACE groups within data source, thus add up to 100% row wise.

	G45.1.1.2	*G45.1.9*	*G45.2.0*	*G45.4.0*	*H50.1.0*	*H50.3.0*	*H52.1.0*	*H52.2.9*
GBR	18 680	1790	6054	1763	943	4500	764	2968
	50%	5%	16%	5%	2%	12%	2%	8%
VAT	18 142	1752	5852	1701	909	4352	738	2865
	50%	5%	16%	5%	2%	12%	2%	8%
PDR	15 054	1439	4654	1326	492	3524	536	2227
	51%	5%	16%	4%	2%	12%	2%	8%
SBS	966	196	284	77	215	388	196	507
	35%	7%	10%	3%	7%	14%	7%	18%
Overlap	885	166	247	64	128	333	148	417
	37%	7%	10%	3%	5%	14%	6%	18%

1.3.3 Apparently contradicting variables in the case study data set

In the previous section some variables were discussed that are related. Therefore, also possible errors can be discerned from contradicting values on these variables. This section discusses the variables on *Number of Employees* and *Turnover* in relation to each other.

GBR *Size Class* and *Number of Employees*

Figure 1.9 shows deviations between *Number of Employees* and *Size Class* in the GBR, with horizontal lines that mark the borders of the *Size Class*es, as defined by Table 1.4 (on page 19). Only values are shown for the data that is used to fit the Intermittent-Error Model, thus the businesses for which *Turnover* values are available from three sources. *Size Class* 0 businesses do not occur in this data, because no *Size Class* 0 businesses are sampled for the SBS. Yet, there are businesses for which the GBR reports 0 employees, even though they belong to *Size Class* 1 or 2. Table 1.8 (on page 26) shows for which NACE groups this is the case and also for how many the opposite occurs: businesses with number of employees larger than 0 that do occur in the GBR classified by *Size Class* 0. Because these businesses are classified by *Size Class* 0 they are not part of the data set to fit the Intermittent Error Model on, because no SBS turnover value is available.

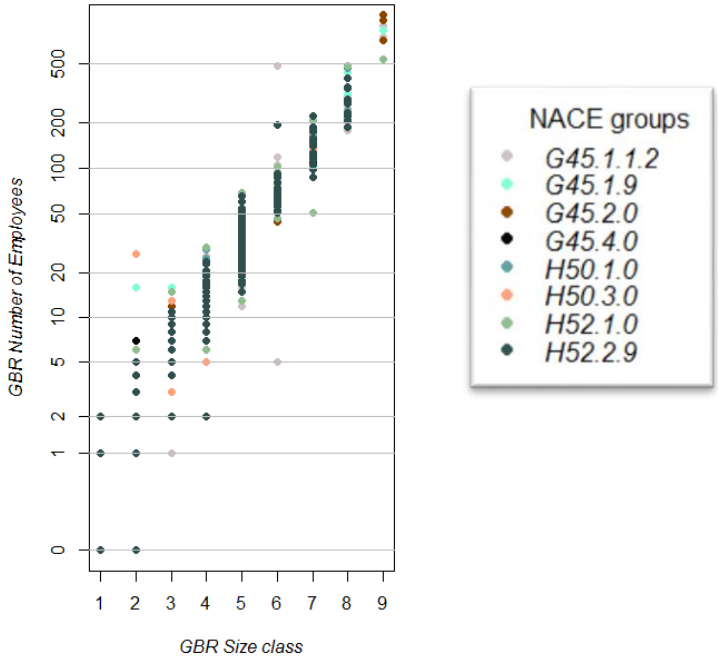


Figure 1.9 GBR *Number of Employees* plotted against GBR *Size Classes* on a logarithmic y-axis for the data for which turnover values are available from all three sources (overlapping part in Figure 1.8 on page 24). *Grey horizontal lines represent the borders of the Size Classes described in Table 1.4 (on page 19).*

Table 1.8 Non-agreeing *Size Class* 0 and *Number of Employees* in GBR and 'Overlap', which is the subset of businesses for which turnover values are available from all three sources (overlapping part in Figure 1.8).

		G45.1.1.2	*G45.1.9*	*G45.2.0*	*G45.4.0*	*H50.1.0*	*H50.3.0*	*H52.1.0*	*H52.2.9*
GBR	*Size Class* 0, *Number of Employees* > 0	412	31	176	43	60	91	10	67
	Size Class > 0, *Number of Employees* 0	258	23	117	28	12	55	13	52
	Total	18 680	1790	6054	1763	943	4500	764	2968
Overlap	*Size Class* > 0, *Number of Employees* 0	7	1	0	1	0	0	0	4
	Total	885	166	247	64	128	333	148	417

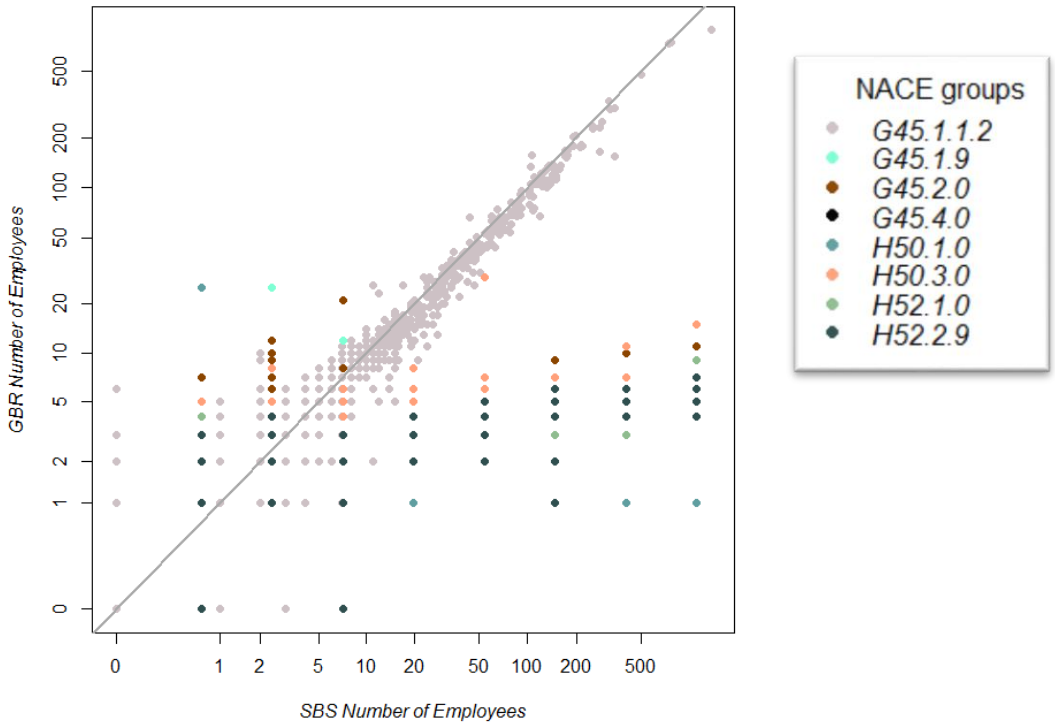


Figure 1.10 *Number of Employees* from the GBR plotted against *Number of Employees* from the SBS survey on a logarithmic x- and y-axis. *The grey line represents the values for which GBR Number of Employees = SBS Number of Employees.*

GBR *Number of Employees* and SBS *Number of Employees*

Figure 1.10 shows the values for *Number of Employees* reported in the GBR, plotted against the *Number of Employees* from the SBS survey. This shows that also these sources do not agree and that certain patterns in reporting these numbers in the SBS survey might be responsible for that. Since the GBR is a description of the entire population and should describe all businesses in equal ways it is not directly updated after new information comes in from a survey. Doing so would lead to inconsistencies between sampling procedures used to publish statistics before and after such update.

Especially the businesses in the bottom right corner of this figure show very large deviations and some patterns among reported SBS values and NACE groups.

SBS, VAT and PDR *Turnover*

Figure 1.11 shows all Turnover values from case study data for which information is available from the SBS, PDR and VAT (overlapping area in Figure 1.8, 2388 business units) by NACE group. Only for 58 business units (3%) three equal turnover values are reported by the three sources. For 194 business units (8%) the SBS value equals the VAT value and for 759 of the records (32%) the SBS coincides with the PDR.

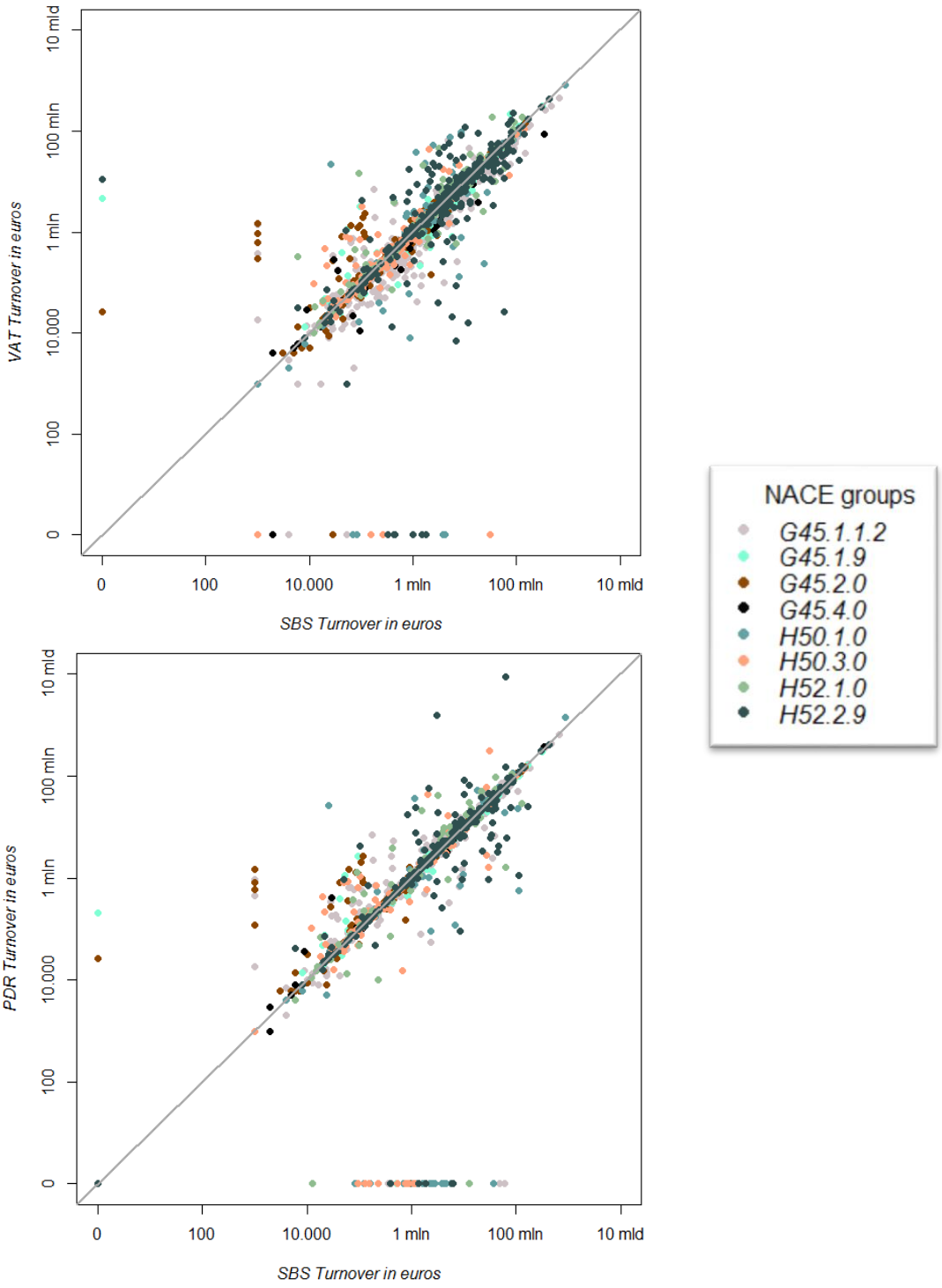


Figure 1.11 VAT and PDR *Turnover* values plotted against SBS *Turnover* values on a logarithmic x- and y-axis. One business with a negative VAT *Turnover* value is excluded (see Section 4.1.2). *The grey line represents the values for which VAT/PDR Turnover = SBS Turnover.*

1.3.4 Case study example data records

This section discusses some example data records that are shown in Table 1.9 (on page 29).

1. The first example record illustrates VAT *reporting differences* (*relevance error*) (BaseLine execution, phase two, measurement side) in a record from the largest NACE group *G45.1.1.2*. It emphasizes that even though the PDR *Coverage Percentage* is smaller than 100%, the PDR *Turnover* can provide the largest value and possibly the correct value since it agrees with the SBS value. It therefore might be that the 20% of the business not covered (the *Coverage Percentage* is 80%) did not generate any turnover. Or that the business incorrectly reported a turnover value in the SBS survey that only covers 80% of the business. The PDR *Unlock Code* is A, which would mean that the statistical business unit build-up is simple and that no tax exemption takes place. This is inconsistent with the PDR *Coverage Percentage* which is smaller than 80%. This disagreement might be a *mapping error* (BaseLine execution, phase two, measurement side) in either the *Unlock Code* or the *Coverage Percentage* that might have occurred at Statistics Netherlands when producing this variable from tax information.
2. The second example record illustrates unreliable €0 *Turnover* values obtained from the SBS and PDR. Since the VAT *Turnover* value is so much larger than €0, this difference is very unlikely due to reporting differences. Such errors can result from incorrect reporting (*measurement error*) or processing (*processing error*) (BaseLine execution, phase one, measurement side), but can also result from erroneous linking two different business units at Statistics Netherlands, being an *identification error* or *unit error* (GBR execution, phase two, representation side).
3. The third example shows that even though the VAT *Response Percentage* is 0%, the VAT *Turnover* can be a large value. This is possibly a *processing error* on VAT *Response Percentage* (by Statistics Netherlands) or a value that was erroneously imputed (also *processing error*) by the Tax and Customs Administration (BaseLine execution, phase one, measurement side). When the value results from incorrect imputation by Statistics Netherlands, this is a second phase error denoted by *comparability error* (BaseLine execution, phase two, measurement side).
4. The fourth example shows a record with a SBS *Turnover* value that is not plausible considering the available VAT and PDR turnover values. Since turnovers are required to be reported in the SBS survey in €1000, an erroneous 1 occurred somewhere in the process. This is either a *measurement error* (BaseLine execution, phase one, measurement side), or *processing error* (BaseLine execution, phase one, measurement side), since no phase two exists in relation to the SBS survey data. For example, he value could have occurred from a bug in the automatic editing procedure (processing error).
5. The fifth example indicates a situation similar to the first example in which the PDR *Coverage Percentage* is not 100% but the PDR turnover value is still the largest of the three. This example is interesting since it has PDR *Unlock Code* B, which should make the turnover value less trustworthy. *Unlock Code* B indicates that the statistical business unit contains a legal unit that is exempted from reporting to the PDR. This example indicates that it might be that the exempted part of the business only produces a marginal part of the total turnover of the company, or even none at all, and is therefore rightfully exempted. Since the Turnover values from the three sources do not agree, some sort of *measurement errors* or *reporting differences* (*relevance errors*) might be at work. This is the case for most businesses as shown in Section 1.3.3.
6. The negative VAT value most likely occurred at the Tax and Customs Administrations, which sometimes allows businesses to report negative turnover to correct an overestimation of turnover in a previous tax report. Therefore, this negative value could be characterized as a *relevance error*, since it can be attributed to a difference between the statistical target concept and the available administrative information.
7. The seventh example shows a negative VAT *Response Percentage* which is a *processing error* since the variable is created at Statistics Netherlands (and thus a first phase error).

8. The eighth example shows a business for which the *Number of Employees* does not agree with the *Size Class*, and is very unlikely considering the large *Turnover* values. This is a *mapping error* or *comparability error* (GBR execution, phase two, measurement side). The cause of this error could also be erroneous linking two different business units at Statistics Netherlands, being an *identification error* or *unit error* (BaseLine execution, phase two, representation side).
9. The ninth example is similar to the fourth example, only this record is manually edited. If the improbable €1000 SBS *Turnover* occurred during manually editing it is a *processing error*.
10. The tenth example shows an imputed VAT *Turnover* value since the *VAT Quarters* value is smaller than 4. The imputation might also introduce errors, as shown by the disagreeing *Turnover* values from the three sources. Since this concerns processing when the VAT data is reused at statistics Netherlands, an error in imputation is denoted by *comparability error* (BaseLine execution, phase two, measurement side).

Table 1.9 Example data records. *These are not representations of actual statistical business units. The examples do resemble typical patterns that occur in the data.*

Example number	GBR *NACE group*	GBR *Size Class*	GBR *Number of Employees*	GBR *TopX*	SBS *Automatic edited*	SBS *Manually edited*	*VAT Quarters*	VAT *Response Percentage*	PDR *Unlock Code*	PDR *Coverage Percentage*	PDR *Fill Percentage*	SBS *Turnover (in €1000)*	VAT *Turnover (in €1000)*	PDR *Turnover (in €1000)*
1	*G45.1.1.2*	7	102	yes	no	yes	4	100%	A	80%	100%	58 000	40 000	58 000
2	*H52.2.9*	4	10	no	yes	no	4	100%	A	100%	100%	0	11 200	0
3	*G52.1.0*	3	5	yes	NA	NA	4	0%	W	100%	NA	NA	9 220	NA
4	*G45.2.0*	2	3	no	yes	no	NA	100%	A	100%	100%	1	295	120
5	*G45.1.9*	5	45	no	yes	no	NA	100%	B	75%	100%	7 660	7 631	7 775
6	*H50.3.0*	1	1	no	yes	yes	NA	100%	A	100%	100%	1	- 4	0.5
7	*H50.1.0*	1	1	no	NA	NA	NA	-4%	W	100%	NA	NA	283	NA
8	*G45.1.1.2*	2	0	no	yes	no	2	100%	A	100%	100%	245	70	245
9	*G45.2.0*	4	15	no	yes	yes	4	100%	A	100%	100%	1	1 480	1 450
10	*G45.1.1.2*	5	36	no	yes	no	3	100%	A	100%	100%	7 005	5 800	4 600

1.4 Current error detection and correction by Statistics Netherlands

How are errors detected? Representation Errors can only be detected if a reference file exists that describes the target population, or if such reference file can be constructed from multiple sources by micro integration (Bakker B. F., 2011). With regard to Measurement Errors (cat) there are three main procedures to detect errors. One way to is to carry out additional research, by collecting the same data again, or thoroughly checking an *audit sample*. Another form of additional research is *Response Analysis Surveys (RAS)*. With RAS a sample of the units is contacted, on which information was collected previously, to gather information about the data collection process such as record-keeping practices and understanding of instructions and definitions (Groen, 2012). A second way to detect Measurement Errors (cat) is from inconsistencies within data from one source, such as within values on purchases, costs and turnover from the same business. With regard to the range of values (some cannot be negative), or by comparing data to historical reference data or data from similar units. This is the most common approach among NSIs and what is mainly denoted by *editing*. A third means to detect values suspicious with regard to Measurement Errors (cat) is to compare values on the same units from different sources in multi-source data, such as the case study data set. Of course, when related values are inconsistent, or when values on the same variable do not coincide, it might be still unknown which one is erroneous.

Deductively editing systematic Measurement Errors (cat)
Systematic errors, such as turnover values reported in euros instead of €1000 or incorrectly negative values, are always directly corrected (mainly by automatic procedures). The process of correcting these errors is called deductive since the error is corrected in the same way for all records in which it occurs. Namely by dividing by 1000 and rounding, or taking absolute values, which can be implemented as a simple decision rule.

Selectively editing Measurement Errors (cat)
Editing focuses mainly on Measurement Errors (cat). Representation Errors are much harder to identify since often no undisputable target population definition is available. Correcting only Measurement Errors (cat) will never obtain perfect statistics for publishing since Representation Errors are not accounted for and the procedure might introduce new errors. For NSIs also considerations in terms of costs (time and resources) come into play. Therefore, editing, mainly manual editing, is limited to the most influential data records (Hoogland, 2005, p. 5). A procedure of *selective editing* is carried out which identifies the most influential records and most influential errors, with regard to the macro level statistics that need to be published based on the data records. Records that are internally inconsistent are always edited. Other possible errors are recognized by comparing data values of units to historical reference data of the same units, or by comparison to current reference data from similar units, to median values or data on the same unit from a different source. Selection for editing is carried out by assigning scores to possible errors, so called plausibility indices. These plausibility indices represent a combination of influence and risk of the error. With regard to business statistics, the *influence* has to do with the *Size Class* of the company since a small percentage change in values for a large company can be of large influence. The *risk* includes the expected size of the error with regard to the reference value.

Automatic and manually editing Measurement Errors (cat)
The check and editing process at Statistics Netherlands is either carried out automatically by software procedures (*automatic editing*) or handled by an editor (*manually editing*). Automatic editing is carried out with regard to the before mentioned systematic errors as well as nonsystematic errors. For systematic errors the procedure of detection and correction is deterministic and is therefore always carried out successfully. An example of nondeterministic automatic editing is carried out on to records with internal inconsistency with regard to turnover, costs and profit. Since each two involved variables

can be corrected in infinitely many ways, the procedure is an optimization problem rather than a deductive decision rule (Hoogland, Van der Loo, Pannekoek, & Scholtus, 2010). As a result, the procedure can also result in an error, or not converge within the time available and in this latter case the record still requires manual editing.

All main influential businesses are manually checked ('edited') (Hoogland, Van der Loo, Pannekoek, & Scholtus, 2010, p. 8). Although instructions and software is used to aid editors in selecting the units for editing, they may also rely on their own knowledge and experience in deciding which suspicious records to follow up on. The software indicates the plausibility index for each suspected error, and distinguishes between 'hard' errors, which all need to be handled by an editor and suspicious records which obtain plausibility indices (Boersma, 2009, p. 20). Although consistency is always pursued, not all suspicious records can be scrutinized. The philosophy is that small non-systematic Measurement Errors (cat) get neutralized by each other when aggregated into publishable macro data (Hoogland, 2005, p. 5). An example of a suspicious record is one in which the turnover value related to the main economic activity of a business is smaller than the turnover related to some other economic activity (Boersma, 2009, p. 26). An example of a hard error is a value that is outside the allowed range. When the main influential businesses are checked, time constraints regulate which records with smaller plausibility indices are also manually checked by editors.

Comparing aggregated statistics ready for publication to historical values is denoted by *macro editing*. When statistics become suspicious at macro level, data values that are responsible for the detected macro level error, such as outliers, are still corrected at micro level.

Editing Representation Errors

Statistics Netherlands maintains many *base registers* in cooperation with government institutions that are continuously updated (such as the General Business Register discussed in Section 1.2.3.1). Base registers define the population of units as apparent to multiple government institutions. To maintain consistency in productions based on this framework, Representation Errors encountered at Statistics Netherlands during editing are not always corrected in a statistical production process based on a base register (Aelen, 2005). When micro data for a yearly statistic is edited, quarterly estimates on the same population framework might already been published and consistency among published values might require not to alter the population framework. Each year, the General Business Register experiences a Representation Error investigation on small businesses by the Economic Demographics (Dutch: Economische Demografie) sample survey. This survey investigates over-coverage of small businesses which are in fact inactive. From this survey, probability estimates for inactiveness are calculated for certain types of small businesses. These estimates are used to correct the total number of active business units (Aelen, 2005, p. 12) (Aelen, et al., 2011). Errors in for example NACE classification are sometimes detected by editors in the manual editing procedure, but apart from that no systematic approach exists to detect Representation Errors in NACE group subpopulations.

Data editing in relation to the Structural Business Statistics

The published macro data for the Structural Business Statistics is completely based on values obtained from the sample survey. Even though turnover values from the VAT register are linked to the GBR population in BaseLine, these are only used in macro editing of the survey values. Therefore, not all SBS and VAT turnover values are compared at the micro level. Possibly because turnover is only one value among many measured in the SBS survey. The PDR turnover (and other) values are not immediately available since the Tax and Customs Administration allows companies to delay their tax report. As a result, for some businesses these values are not accessible yet before the publishing deadline and comparison in the editing procedure is not possible. Thus the PDR turnover values are also not used to edit the SBS data. The case study data set gives an indication of automatically and manually edited values, but these only indicate whether the record has undergone an editing process. Many records are checked but not changed because no error is detected, or because of time constraints in the process of selectively editing.

1.5 Summary

In response to research sub question 1 *What concepts are needed to assess the quality of Statistics Netherlands' business data?*

Financial business data can be collected either by Statistics Netherlands (a *primary source*) or externally (a *secondary source*). Secondary sources in The Netherlands which provide data on businesses are the Chamber of Commerce and the Tax and Customs Administration. The Chamber of Commerce's register is used to define a statistical population framework of existing businesses and the Tax and Customs Administration's registers are used to complement that population framework as well as provide monthly, quarterly and yearly financial data on these businesses.

Secondary source data is less under control of Statistics Netherlands than primary data. Since the data is not collected and processed for statistical purposes, the data variable definitions and detected errors might be different from those in primary source data. For example, error detection in tax reports might focus on taxable yearly turnover values, and therefore quarterly reported values might still contain errors. When these quarterly values are used by Statistics Netherlands to produce quarterly economic statistics, the reuse of the data might impose problems (*phase two errors)* even though the quality of the data was not a problem with regard of the original purpose of the data. Of course, also errors can be present that pose a problem to the original register holder as well, or that were introduced by data processing carried out by the original owner of the data (*phase one errors*).

The creation of a statistical population framework of businesses from units in administrative registers involves combining and splitting of information on administrative units, so-called *micro-integration*. Also, yearly financial data needs to be *linked* to the population framework to produce economic statistics of a certain year. This process of micro-integrating the sources for statistical purposes and linking yearly data can also produce phase two errors. Errors occur both in over-coverage or under-coverage of the businesses that constitute the population (category of *Representation Errors*) and in incorrect values representing the financial state of these businesses in a certain year (category of *Measurement Errors*).

The case study data set was used to publish economic statistics per economic activity in 2012, part of the so-called *Structural Business Statistics*. In this study, the main variable of interest is yearly *Turnover*. Due to very complicated legal and fiscal structures of companies, not only might businesses be misrepresented, also a classification of economic activity of the businesses might be incorrectly measured or processed, incorrectly combined for complex businesses, or not updated. As a result, subpopulations per economic activity might be misrepresented (Representation Errors). Moreover, *Turnover* values can be incorrectly measured because a company reported an estimation instead of a measurement. Also a correct value can be incorrectly processed by either the Tax and Customs Administration (phase one) or Statistics Netherlands (phase two). Due to fiscal regulations and exemptions, the concept of *Turnover* from the point of view of the Tax and Customs Administration might deviate from that of Statistics Netherlands. This is not a problem for data collected by Statistics Netherlands itself in sample surveys, but sample survey data contains other errors due to sampling and nonresponse.

The case study data set contains a semi-random sample of all businesses in eight economic activity classes, for which previous research indicated that the definition of *Turnover* in the Tax and Customs Administration's VAT register deviated from that of Statistics Netherlands. The yearly economic statistics of 2012 were based on data from the *Structural Business Statistics sample survey (SBS). Turnover* values from two tax registers are linked to the population framework, the *Value Added Tax (VAT)*, and the *Profit Declaration Register (PDR)*.

Some errors in this case study data set were already apparent from descriptive statistics. One example is the deviations of the variable *Number of Employees* reported in the surveys from that part of the population framework (constructed with information from the Chamber of Commerce and Tax and

Customs Administration). Also, the data set contains codes and percentages with regard to the (un)successful linkage of Tax and Customs Administration data to the population framework. These codes and percentages showed that not all businesses in the population were 100% covered by the data obtained from the external registers.

The VAT and PDR do not provide turnover data on all businesses in the population framework, and therefore the total turnover in these sources underestimate the total turnover in the economic activity classes in the data set. Since the PDR covers a smaller proportion of the businesses in the population, but reports a higher total turnover, either the VAT underestimates turnover values, or the PDR overestimates it. This points at different *Turnover* concepts applied in these registers. With regard to individual businesses, the VAT, PDR and SBS survey often do not agree on turnover value. Differences between sources of a factor 1000 indicate severe errors that are not only due to mistakes in reporting by the businesses. These might have also occurred during processing or by misrepresented business units for which values from different businesses was linked to the same unit. Only for 3% of the business units three equal turnover values were reported from the three sources.

Error detection and correction at Statistics Netherlands mainly focuses on the category of Measurement Errors. The procedures are partly *automatic* and partly *manual*. Measurement Error detection results from the recognition of values outside the allowed range (negative turnover values) inconsistencies within businesses (turnover, costs and profit not in agreement) and notable deviations from (historical) reference values. Since the editing procedure involves costly and time consuming manual labor, the process is selective. In *selective editing* only the severest and most influential (with regard to the economics statistics that need to be published) suspected errors are scrutinized.

2 The Intermittent-Error Model

When multiple sources on the same variable are available (*multi-source data*), erroneous records can be identified in case the various sources do not agree on the variable of interest, such as in the example data records in Section 1.3.4. Therefore, the detection procedure for erroneous records is clear, but which value is erroneous and how to correct the record might not be immediately clear. Various approaches, such as structural equation models (SEMs) are suggested to simultaneously identify errors and estimate the 'True Value' (Scholtus, Bakker, & Van Delden, 2015). The Intermittent-Error Model is proposed by Guarnera & Variale ((2015) and (2016)), and is characterized by an intermittent error mechanism. No error distribution is defined for all values, but the assumed error producing process sometimes does and sometimes does not result in values with error. Therefore, only a proportion of the data is assumed to be affected by errors, which is in accordance with the editing process assumption within Statistics Netherlands that most data values are correct. After all, the procedure of selectively editing leaves many values uncorrected by assuming that editing can also introduce new errors.

This chapter discusses the Intermittent-Error Model from the perspective of its application on the case study data set. Section 2.1 gives the Intermittent-Error Model definition. Section 2.2 discusses the model's assumptions and Section 2.3 describes the estimation of the Intermittent-Error Model from the perspective of this study's investigation. Section 2.4 concludes this chapter with a summary.

2.1 The Intermittent-Error Model definition

The Intermittent-Error Model is defined for numerical data on continuous data variables. The existence of a latent *True Value* is assumed for each unit, which is imperfectly measured by the multiple sources that are affected by intermittent errors. With regard to the case study data set introduced in Section 1.2.1 and 1.3, a graphical presentation of the Intermittent-Error Model is given in Figure 2.1 (on page 37).

The following model definition is based on the presentation of the Intermittent-Error Model by Robinson (2016) and Scholtus, Bakker & Robinson (2016 (preprint)). The variable names are brought in agreement with the R implementation of the model by Robinson (2016), except for the True Value, for which η_i is used instead of y^*.

Observed values and source error distribution

Let η_i denote the True Value of the variable of interest (turnover in the case study data set) for unit i. For each observed variable $y_{g,i}$ from source g that measures η_i, a 0-1-indicator z_g is introduced such that $z_{g,i} = 0$ if the measurement is error-free ($y_{g,i} = \eta_i$) and $z_{g,i} = 1$ otherwise. For units with $z_{g,i} = 1$, the erroneous measurement is assumed to be linearly related to the True Value η_i, introducing *intercept bias* (a_g) and *slope bias* (b_g - 1) specific to each source g:

$$y_{g,i} = \begin{cases} \eta_i & \text{if } z_{g,i} = 0 \\ a_g + b_g \eta_i + \epsilon_{g,i} & \text{if } z_{g,i} = 1 \end{cases} \tag{2.1}$$

a_g and b_g are constants, characteristic for each source g, and $\epsilon_{g,i}$ follows a normal distribution with mean zero and variance σ_g^2. In a single formula, this can be expressed as follows (as also shown in the

dark grey area in Figure 2.1 (on page 37) for the case study data set):

$$y_{g,i} = (1 - z_{g,i})\,\eta_i + z_{g,i}\,(a_g + b_g\eta_i + \epsilon_{g,i}) \tag{2.2}$$

The probability of observing an error on y_g in each source is represented by the parameter $\pi_g = \mathbb{P}(z_g = 1) = \mathbb{E}(z_g)$. This value describes the expected proportion of errors in measurements from source g. A value of π_g closer to 1 indicates that more errors occur in source g. The above model formulation is more general than that of Guarnera & Varriale ((2015) and (2016)), which assumes that $a_g = 0$ and $b_g = 1$ for all sources g.

True Value distribution

The proposed model does not only model the error producing process for the different sources, but also assumes a distribution on the True Value. It is assumed that η_i is normally distributed with its expected value linearly related to covariates ($\mathbb{E}(\eta_i \mid \boldsymbol{x_i}) = \boldsymbol{\beta' x_i}$) and that η_i arises by adding unexplained error term ϵ_i with variance σ^2. This True Value distribution is presented graphically in the light grey area of Figure 2.1 (on page 37) for the case study data set with one covariate *Number of Employees*. In general, the True Value has the following form:

$$\eta_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \epsilon_i \tag{2.3}$$

Note that, in the presence of covariates, σ^2 represents the unexplained variance in η. In general, the total variance of η is given by:

$$\sigma_\eta^2 = \boldsymbol{\beta}'\,\boldsymbol{\Sigma}_{xx}\,\boldsymbol{\beta} + \sigma^2 \tag{2.4}$$

where $\boldsymbol{\Sigma_{xx}}$ denotes the variance-covariance matrix of the vector of covariates $\mathbf{X}$.

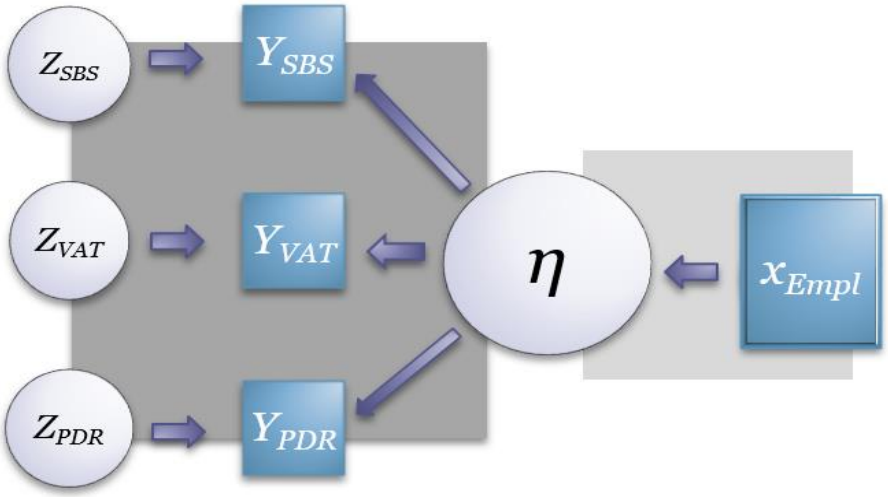
Complete data log likelihood

Table 2.1 explains the variables composing the complete data with regard to the case study data set ($y_{SBS,i}$, $y_{VAT,i}$, $y_{PDR,i}$, η_i). The indicators $z_{SBS,i}$, $z_{VAT,i}$ and $z_{PDR,i}$ can be derived from these (e.g. $z_{SBS,i} = 1$ if $y_{SBS,i} \neq \eta_i$ and $z_{SBS,i} = 0$ otherwise).

Table 2.1 Explaining table for case study variables in the complete data

$y_{SBS,i}$	*Turnover* value from the Structural Business Statistics survey	observed
$y_{VAT,i}$	*Turnover* value from the VAT register	observed
$y_{PDR,i}$	*Turnover* value from the Profit Declaration Register	observed
η_i	Turnover True Value	observed for some units, latent for others
$x_{Empl,i}$	*Number of Employees* from the GBR	observed

The complete data log likelihood follows from the fact that the Intermittent-Error Model is a special case of a finite mixture model, which is further discussed in Section 2.3.4 with regard to the current research questions. In a more general form, the complete data log likelihood is shown in Expression (2.5).



$$y_{SBS,i} = (1 - z_{SBS,i})\,\eta_i + z_{SBS,i}\,(a_{SBS} + b_{SBS} * \eta_i + \epsilon_{SBS,i})$$

$$y_{VAT,i} = (1 - z_{VAT,i})\,\eta_i + z_{VAT,i}\,(a_{VAT} + b_{VAT} * \eta_i + \epsilon_{VAT,i})$$

$$y_{PDR,i} = (1 - z_{PDR,i})\,\eta_i + z_{PDR,i}\,(a_{PDR} + b_{PDR} * \eta_i + \epsilon_{PDR,i})$$

$$\eta_i = \beta_0 + \beta_{Empl} * x_{Empl,i} + \epsilon_i$$

(Special cases of Expression (2.2) and (2.3))

Figure 2.1 Graphical representation of the Intermittent-Error Model for the case study data set.
Latent variable (circle); Observed variable (square); Modeled relations indicated by arrows.
Y_{SBS}: Turnover value from the Structural Business Statistics Survey;
Y_{VAT}: Turnover value from the Value Added Tax Register;
Y_{PDR}: Turnover value from the Profit Declaration Register;
X_{Empl}: Covariate Number of Employees (measured without error).

For n observations, each available from G sources $y_1, \ldots, y_G$, with error patterns specified by $z_1, \ldots, z_G$, the complete data log likelihood function is defined by the following expression (Scholtus, Bakker, & Robinson, 2016 (preprint)):

$$\ell_c(\theta) = C + \sum_{g=1}^{G}\left[n \log(1 - \pi_g) + n_g \log\left(\frac{\pi_g}{1 - \pi_g}\right)\right] - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\eta_i - \beta' x_i)^2 - \sum_{g=1}^{G}\left[\frac{n_g}{2}\log\sigma_g^2 + \frac{1}{2\sigma_g^2}\sum_{i=1}^{n} z_{g,i}(y_{g,i} - a_g - b_g\eta_i)^2\right] \quad (2.5)$$

$n_g = \sum z_{g,i}$ denotes the number of observations with an error on y_g, thus the number of errors per source g. C denotes a constant term that does not depend on any unknown parameters. $\boldsymbol{\theta}$ denotes a vector containing all distinct parameters of the model. These are, for the case study data set (further explained in Table 2.2 (on page 38):

$$\boldsymbol{\theta} = (\pi_{SBS}, \pi_{VAT}, \pi_{PDR}, \beta_o, \beta_{Empl}, \sigma^2, a_{SBS}, b_{SBS}, \sigma^2_{SBS}, a_{VAT}, b_{VAT}, \sigma^2_{VAT}, a_{PDR}, b_{PDR}, \sigma^2_{PDR}) \quad (2.6)$$

The log likelihood given in Expression (2.5) cannot be directly optimized since for some records the True Value η_i and the corresponding $z_{g,i}$, are latent. But the model parameters can still be estimated

from Expression (2.5) by using maximum likelihood estimation for incomplete data. Guarnera and Varriale worked out an EM (Expectation - Maximization) algorithm for the model with $G = 3$ observed variables, under the restriction that $a_g = 0$ and $b_g = 1$. Robinson (2016) gives a detailed description of this algorithm, including an extension to estimate a_g and b_g. The EM estimation procedure is discussed in Section 2.3.

Observations with and without errors

The Intermittent-Model characteristic indicator z_g, assumes a non-zero probability (namely $1 - \pi_g$) that an observed value $y_{g,i}$ is equal to the True Value η_i and therefore error-free. Since the error distribution described in Expression (2.1) and (2.2) is continuous, $y_{g,i} = \eta_i$ occurs with probability zero, if $y_{g,i}$ contains an error, because the probability of two independent continuous error distributions generating the same error is negligible. So if two (or more) different sources report the same measurement ($y_{g,i} = y_{h,i}$, for $g \neq h$), it must hold that these sources measure the corresponding True Value ($y_{g,i} = y_{h,i} = \eta_i$). Therefore, the True Value η_i is not latent for all units, and some error-free values can be recognized from the observed data. The recognition of η_i values from the data is an important aspect of the estimating procedure described in Section 2.3.

Table 2.2 Explaining table for the case study parameters (in $\boldsymbol{\theta}$) of the Intermittent-Error Model

π_{SBS}	expected proportion of errors present in *Turnover* values from the Structural Business Statistics survey
π_{VAT}	expected proportion of errors present in *Turnover* values from the VAT register
π_{PDR}	expected proportion of errors present in *Turnover* values from the Profit Declaration Register
β_0	the intercept of the linear regression model belonging to the conditional True Value distribution $\mathbb{E}(\eta_i \mid x_{Empl,i}) = \boldsymbol{\beta_0} + \beta_{Empl} * x_{Empl,i}$
β_{Empl}	the slope parameter for regression covariate x_{Empl} of the linear regression model belonging to the conditional True Value distribution $\mathbb{E}(\eta_i \mid x_{Empl,i}) = \boldsymbol{\beta_0} + \beta_{Empl} * x_{Empl,i}$
σ^2	variance of η unexplained by covariate x_{Empl}
a_{SBS}	intercept bias on *Turnover* values from the Structural Business Statistics survey
$b_{SBS} - 1$	slope bias on *Turnover* values from the Structural Business Statistics survey
σ^2_{SBS}	variance of *Turnover* values from the Structural Business Statistics survey after correction for intercept bias and slope bias
a_{VAT}	intercept bias on *Turnover* values from the VAT register
$b_{VAT} - 1$	slope bias on *Turnover* values from the VAT register
σ^2_{VAT}	variance of turnover values from the VAT register after correction for intercept bias and slope bias
a_{PDR}	intercept bias on turnover values from the Profit Declaration Register
$b_{PDR} - 1$	slope bias on turnover values from the Profit Declaration Register
σ^2_{PDR}	variance of turnover values from the Profit Declaration Register after correction for intercept bias and slope bias

2.2 Assumptions of the Intermittent-Error Model

Source independence
The occurrence of errors (z_{SBS}, z_{VAT} and z_{PDR}) and the error size (ϵ_{SBS}, ϵ_{VAT} and ϵ_{PDR}) in each source are considered independent across sources. Therefore, the covariance between observations completely follows from the True Value distribution they have in common. The observations y_{SBS}, y_{VAT} and y_{PDR} are independent conditional on the True Value η.

Independent error occurrence and size within source
Error occurrences and error size (z_{SBS}, ϵ_{SBS}), (z_{VAT}, ϵ_{VAT}), (z_{PDR}, ϵ_{PDR}) are also considered independent within sources.

Linear model assumptions True Value distribution and source errors
Both the True Value distribution (when only one covariate is considered) and the source error distributions are described by simple linear models. Therefore, they have simple linear model assumptions, such as linearity, normally distributed errors, constant error variance and error-free covariate values. Also, in both the True Value distribution and the source error distributions, the errors sizes are uncorrelated with the dependent variable: $Cor(\eta, \epsilon) = 0$ and $Cor(y_g, \epsilon_g) = 0$.

η_is Missing At Random (MAR)
The missingness of η_i describes whether η_i is measured with error and therefore itself not observed (latent), indicated by the z_is. This missingness does not depend on the value of η_i, but is described by the model parameters (the πs) based on all available data other than η_i itself. This assumption allows for the EM estimation procedure to be used, since the expected values of η_i for an EM iteration are based on the parameters of the iteration that are maximized for all available data.

2.3 Estimation of the Intermittent-Error Model parameters

The Intermittent-Error Model is estimated by the Expectation-Maximization (EM) Algorithm, which is an iterative method for finding maximum likelihood estimates of parameters when the model depends on unobserved latent variables. The procedure alternately estimates the values for the latent variables (Expectation Step (E-step)) and maximizes the unknown parameters (Maximization Step (M-step)). A full account of the EM estimation procedure for the Intermittent-Error Model is provided by Robinson (2016). This section addresses the main characteristics of the procedure, with regard to the case study data set and in relation to the Research Questions. All expressions in this section are special cases of expressions derived by Robinson (2016).

The intermittent error assumption of the Intermittent Error Model allows some measurements to be error-free. A specific part of these error-free measurements can be recognized from the data, namely if two or more sources report the same measurement ($y_{g,i} = y_{h,i} = \eta_i$ $(g \neq h)$ for some unit i) (as described in Section 2.1). This fact indicates that not all True Values (η_i) occurrences are latent, which is heavily used in the Intermittent-Error Model's estimation procedure. The different roles in the estimation procedure of known and unknown True Values are discussed in Section 2.3.1. Section 2.3.2 discusses the E-step for observations with known True Values and Section 2.3.3 discusses the E-step for observations with unknown True Values. Section 2.3.4 discusses the Log likelihood and Q-function, which is maximized to obtain updated parameter values in the M-step. An example of the analytical expression of an updated parameter is given in Section 2.3.5.

2.3.1 Observations with and without error

Each statistical business unit's triplet of turnover values ($y_{SBS,i}$, $y_{VAT,i}$, $y_{PDR,i}$) has one of $2^3 = 8$ possible error patterns:

$$\left(z_{SBS,i}, z_{VAT,i}, z_{PDR,i}\right) \in \{(0,0,0), (0,0,1), (0,1,0), (1,0,0), (0,1,1), (1,0,1), (1,1,0), (1,1,1)\} \quad (2.7)$$

z_g is equal to 1 in the presence of an error in the turnover value from source g and $z_g = 0$ if the observation from source g is error-free. With regard to four of these error patterns, the statistical business units with the patterns can be recognized from the data. These are (0,0,0) (three sources with equal observations), (0,0,1) (SBS and VAT with equal observations), (0,1,0) (SBS and PDR with equal observations) and (1,0,0) (VAT and PDR with equal observations). For the remaining observations, it can be inferred that ($z_{SBS,i}$, $z_{VAT,i}$, $z_{PDR,i}$) must have one of the four remaining patterns, so at least two of the three observed values must be erroneous, but it is unknown which ones.

The triplets of observations ($y_{SBS,i}$, $y_{VAT,i}$, $y_{PDR,i}$) can be divided into disjoint sets that correspond to the error patterns: (S_{000}, …, S_{111}). Of these, the observations belonging to sets S_{000}, S_{001}, S_{010}, S_{100} can be identified from the data, since if, and only if, two (or three) sources report the same value, the triplet belongs to a set with two (or three) error-free observations. For the remaining units it is known that they belong to the set $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$, but it is indefinite which exact set they belong to.

The probability of a triplet belonging to one of the eight sets follows from the error proportions π_{SBS}, π_{VAT} and π_{PDR} (e.g. $\pi_{SBS} = \mathbb{P}(z_{SBS} = 1) = \mathbb{E}(z_{SBS})$) as indicated by Expression (2.8). These probabilities to belong to a set can be regarded mixing weights (m_{000}, …, m_{111}), when the Intermittent-Error Model is approached as a special case of a finite mixture model. Mixing weights occur in the complete data log likelihood and the conditional expected log likelihood, the Q function for the EM algorithm in Expressions (2.20), (2.22) and (2.23). Of these mixing weights m_{000}, m_{001}, m_{010} and m_{100} are known from the data, as discussed in the next section (2.3.2). m_{011}, m_{101}, m_{110} and m_{111} are unknown and estimated with regard to the parameters of each iteration (t) of the EM-algorithm.

$$\begin{aligned}
m_{000} &= (1 - \pi_{SBS}^{(o)}) * (1 - \pi_{VAT}^{(o)}) * (1 - \pi_{PDR}^{(o)}) \\
m_{001} &= (1 - \pi_{SBS}^{(o)}) * (1 - \pi_{VAT}^{(o)}) * \pi_{PDR}^{(o)} \\
m_{010} &= (1 - \pi_{SBS}^{(o)}) * \pi_{VAT}^{(o)} * (1 - \pi_{PDR}^{(o)}) \\
m_{100} &= \pi_{SBS}^{(o)} * (1 - \pi_{VAT}^{(o)}) * (1 - \pi_{PDR}^{(o)}) \\
m_{011}^{(t)} &= (1 - \pi_{SBS}^{(t)}) * \pi_{VAT}^{(t)} * \pi_{PDR}^{(t)} \\
m_{101}^{(t)} &= \pi_{SBS}^{(t)} * (1 - \pi_{VAT}^{(t)}) * \pi_{PDR}^{(t)} \\
m_{110}^{(t)} &= \pi_{SBS}^{(t)} * \pi_{VAT}^{(t)} * (1 - \pi_{PDR}^{(t)}) \\
m_{111}^{(t)} &= \pi_{SBS}^{(t)} * \pi_{VAT}^{(t)} * \pi_{PDR}^{(t)}
\end{aligned} \quad (2.8)$$

2.3.2 Expectation step: Identified triplets with error-free observations

For the triplets belonging to sets S_{000}, S_{001}, S_{010} and S_{100}, the True Value η_i is observed, and therefore no expected value given the three observations and the parameters $\mathbb{E}(\eta_i \mid y_{SBS,i}; y_{VAT,i}; y_{PDR,i}; x_{Empl,i}; \boldsymbol{\theta}^{(t)})$ needs to be estimated. The observed value for η_i can be used directly in the maximization process to obtain parameter estimates.

Also, since no other observations belong to sets S_{000}, S_{001}, S_{010} and S_{100} than immediately observed, the mixing weights m_{000}, m_{001}, m_{010} and m_{100} are known from the data. Therefore, only $m_{011}^{(t)}$, $m_{101}^{(t)}$, $m_{110}^{(t)}$ and $m_{111}^{(t)}$ need to be estimated by the EM algorithm, and are thus dependent on the parameter estimates in each iteration (t).

The known η_i values and mixing weights m_{000}, m_{001}, m_{010} and m_{100} provide a starting point for the EM estimation procedure and are therefore used to specify the starting values for parameters $\boldsymbol{\theta}$**.** The parameters in $\boldsymbol{\theta}$ (explained in Table 2.2 on page 38) are: π_{SBS}, π_{VAT}, π_{PDR}, β_o, β_{Empl}, σ^2, a_{SBS}, b_{SBS}, σ^2_{SBS}, a_{VAT}, b_{VAT}, σ^2_{VAT}, a_{PDR}, b_{PDR}, σ^2_{PDR}.

$\pi_{SBS}^{(o)}$, $\pi_{VAT}^{(o)}$ and $\pi_{PDR}^{(o)}$
Since the mixing weights for the identified two/three error-free observations triplets are known, by rearranging the expressions for m_{000}, m_{001}, m_{010} and m_{100} in (2.8), initial values for the error proportions ($\pi_{SBS}^{(o)}$, $\pi_{VAT}^{(o)}$ and $\pi_{PDR}^{(o)}$) of the three sources can be calculated (Robinson, 2016, pp. 19-20).

$\beta_o^{(o)}$, $\beta_{Empl}^{(o)}$ and $\sigma^{2(o)}$
Since for triplets in sets S_{000}, S_{001}, S_{010} and S_{100} the True Value η_i is known, it can be regressed on the covariate $x_{Empl,i}$ to obtain starting values for the βs and σ^2 ($\beta_o^{(o)}$, $\beta_{Empl}^{(o)}$ and $\sigma^{2(o)}$) that constitute the True Value conditional distribution (the right hand side in Figure 2.1 (on page 37)).

$a_{SBS}^{(o)}$, $b_{SBS}^{(o)}$, $\sigma^2_{SBS}^{(o)}$, $a_{VAT}^{(o)}$, $b_{VAT}^{(o)}$, $\sigma^2_{VAT}^{(o)}$, $a_{PDR}^{(o)}$, $b_{PDR}^{(o)}$, $\sigma^2_{PDR}^{(o)}$
The triplets in set S_{001} contain two observations $y_{SBS,i}$ and $y_{VAT,i}$ that are identical and therefore equal to η_i, and one observation $y_{PDR,i}$ that contains an error. In set S_{010} $y_{SBS,i}$ and $y_{PDR,i}$ are equal to η_i and $y_{VAT,i}$ contains an error. And in set S_{100} $y_{SBS,i}$ contains an error and $y_{VAT,i}$ and $y_{PDR,i}$ equal η_i. Therefore, when S_{001}, S_{010} and S_{100} contain two or more units, for each source the erroneous value can be regressed on η_i to obtain initial values of the intercept bias, slope bias, and variance after correction for intercept bias and slope bias.

2.3.3 Expectation step: Indefinite triplets

Error proportions
During each iteration (t) of the EM-algorithm, probabilities are assigned to the indefinite triplets to belong to sets S_{011}, S_{101}, S_{110} or S_{111}, conditional on not belonging to the first four sets, the data, and the current parameter estimates. The probability, $\tau_{011,i}^{(t)}$, of unit i belonging to set S_{011} at iteration (t) is defined and calculated as follows ($\tau_{101,i}^{(t)}$, $\tau_{110,i}^{(t)}$ and $\tau_{111,i}^{(t)}$ are analogous):

$$\tau_{011,i}^{(t)} = \mathbb{P}\left(\left(z_{SBS,i}, z_{VAT,i}, z_{PDR,i}\right) = (0,1,1) \,\middle|\, y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i}; \mathbf{x}_{Empl,i};\ i \in S_{011} \cup S_{101} \cup S_{011} \cup S_{111};\ \boldsymbol{\theta}^{(t)}\right) \tag{2.9}$$

$$\tau_{011,i}^{(t)} = \frac{m_{011}^{(t)} * f_{011}\left(y_{SBS,i}\,; y_{VAT,i}\,; y_{PDR,i} \,\middle|\, \mathbf{x}_{Empl,i}, \boldsymbol{\theta}^{(t)}\right)}{\sum_{klm \in \{011,101,110,111\}} \left[m_{klm}^{(t)} * f_{klm}\left(y_{SBS,i}\,; y_{VAT,i}\,; y_{PDR,i} \,\middle|\, \mathbf{x}_{Empl,i}, \boldsymbol{\theta}^{(t)}\right)\right]} \tag{2.10}$$

with $m_{011}^{(t)}$ the mixing weight corresponding to error pattern (0, 1, 1) (computed with the error proportions (1 - $\pi_{SBS}^{(t)}$), $\pi_{VAT}^{(t)}$ and $\pi_{PDR}^{(t)}$ and therefore following from the parameter estimates in $\boldsymbol{\theta}^{(t)}$ in the t^{th} iteration of the EM algorithm), and f_{011} the probability density function of the three observed

values conditional on the current parameter estimates $\boldsymbol{\theta}^{(t)}$, the value of $x_{Empl,i}$ and the error pattern (0, 1, 1). The probability density functions f_{011}, f_{101}, f_{110} and f_{111} are trivariate normal densities with mean $\mathbb{E}(y_{SBS,i}\,;\, y_{VAT,i}\,;\, y_{PDR,i} \mid x_{Empl,i}\,;\, \boldsymbol{\theta}^{(t)})$ (shown in more detail in Expression (2.11) and (2.12)) and joint covariance matrix $\sum^{(t)}$ specific to each error pattern ($\sum_{011}^{(t)}$, $\sum_{101}^{(t)}$, $\sum_{110}^{(t)}$ and $\sum_{111}^{(t)}$) (shown in Expressions (2.13) and (2.14)).

$$f_{klm}(y_{SBS,i}\,;\, y_{VAT,i}\,;\, y_{PDR,i} \mid \mathbf{x}_{Empl,i},\, \theta^{(t)}) =$$

$$N\left(\begin{pmatrix} y_{SBS,i} \\ y_{VAT,i} \\ y_{PDR,i} \end{pmatrix}, \begin{pmatrix} k * a_{SBS}^{(t)} + {b_{SBS}^{(t)}}^{k} * (\beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i}) \\ l * a_{VAT}^{(t)} + {b_{VAT}^{(t)}}^{l} * (\beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i}) \\ m * a_{PDR}^{(t)} + {b_{PDR}^{(t)}}^{m} * (\beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i}) \end{pmatrix}, \sum_{klm}^{(t)}\right)$$

(2.11)

For example, for error pattern (0,1,1):

$$f_{011}(y_{SBS}\,;\, y_{VAT}\,;\, y_{PDR} \mid \mathbf{x}_{Empl,i},\, \theta^{(t)}) =$$

$$N\left(\begin{pmatrix} y_{SBS,i} \\ y_{VAT,i} \\ y_{PDR,i} \end{pmatrix}, \begin{pmatrix} 0 * a_{SBS}^{(t)} + {b_{SBS}^{(t)}}^{0} * (\beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i}) \\ 1 * a_{VAT}^{(t)} + {b_{VAT}^{(t)}}^{1} * (\beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i}) \\ 1 * a_{PDR}^{(t)} + {b_{PDR}^{(t)}}^{1} * (\beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i}) \end{pmatrix}, \sum_{011}^{(t)}\right)$$

$$= N\left(\begin{pmatrix} y_{SBS,i} \\ y_{VAT,i} \\ y_{PDR,i} \end{pmatrix}, \begin{pmatrix} \beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i} \\ a_{VAT}^{(t)} + b_{VAT}^{(t)} * (\beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i}) \\ a_{PDR}^{(t)} + b_{PDR}^{(t)} * (\beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i}) \end{pmatrix}, \sum_{011}^{(t)}\right)$$

(2.12)

These covariance matrices are defined as follows:

$$\sum_{klm}^{(t)} =$$

$$\begin{pmatrix} {b_{SBS}^{(t)}}^{2k} * {\sigma^2}^{(t)} + k * \sigma_{SBS}^{2\,(t)} & {b_{SBS}^{(t)}}^{k} * {b_{VAT}^{(t)}}^{l} * \sigma^{2(t)} & {b_{SBS}^{(t)}}^{k} * {b_{PDR}^{(t)}}^{m} * \sigma^{2(t)} \\ {b_{SBS}^{(t)}}^{k} * {b_{VAT}^{(t)}}^{l} * \sigma^{2(t)} & {b_{VAT}^{(t)}}^{2l} * {\sigma^2}^{(t)} + l * \sigma_{VAT}^{2\,(t)} & {b_{VAT}^{(t)}}^{l} * {b_{PDR}^{(t)}}^{m} * \sigma^{2(t)} \\ {b_{SBS}^{(t)}}^{k} * {b_{PDR}^{(t)}}^{m} * \sigma^{2(t)} & {b_{VAT}^{(t)}}^{l} * {b_{PDR}^{(t)}}^{m} * \sigma^{2(t)} & {b_{PDR}^{(t)}}^{2m} * {\sigma^2}^{(t)} + m * \sigma_{PDR}^{2\,(t)} \end{pmatrix}$$

(2.13)

For example, for error pattern (0,1,1):

$$\sum_{011}^{(t)} =$$

$$\begin{pmatrix} b_{SBS}^{(t)\,2*0} * \sigma^{2(t)} + 0 * \sigma_{SBS}^{2\,(t)} & b_{SBS}^{(t)\,0} * b_{VAT}^{(t)\,1} * \sigma^{2(t)} & b_{SBS}^{(t)\,0} * b_{PDR}^{(t)\,1} * \sigma^{2(t)} \\ b_{SBS}^{(t)\,0} * b_{VAT}^{(t)\,1} * \sigma^{2(t)} & b_{VAT}^{(t)\,2*1} * \sigma^{2(t)} + 1 * \sigma_{VAT}^{2\,(t)} & b_{VAT}^{(t)\,1} * b_{PDR}^{(t)\,1} * \sigma^{2(t)} \\ b_{SBS}^{(t)\,0} * b_{PDR}^{(t)\,1} * \sigma^{2(t)} & b_{VAT}^{(t)\,1} * b_{PDR}^{(t)\,1} * \sigma^{2(t)} & b_{PDR}^{(t)\,2*1} * \sigma^{2(t)} + 1 * \sigma_{PDR}^{2\,(t)} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^{2(t)} & b_{VAT}^{(t)} * \sigma^{2(t)} & b_{PDR}^{(t)} * \sigma^{2(t)} \\ b_{VAT}^{(t)} * \sigma^{2(t)} & b_{VAT}^{(t)\,2} * \sigma^{2(t)} + \sigma_{VAT}^{2\,(t)} & b_{VAT}^{(t)} * b_{PDR}^{(t)} * \sigma^{2(t)} \\ b_{PDR}^{(t)} * \sigma^{2(t)} & b_{VAT}^{(t)} * b_{PDR}^{(t)} * \sigma^{2(t)} & b_{PDR}^{(t)\,2} * \sigma^{2(t)} + \sigma_{PDR}^{2\,(t)} \end{pmatrix}$$

(2.14)

Expected True Value

In case a unit i belongs to S_{011}, S_{101} or S_{110}, one of the observed turnover values ($y_{SBS,i}$; $y_{VAT,i}$; $y_{PDR,i}$) is error-free and therefore equal to the True Value η_i. In case unit i belongs to S_{111} the estimation relies more on the True Value distribution, shown at the right hand side of Figure 2.1 (on page 37). By combining the expected values of the four sets, an expected value for η_i given $i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$ can be calculated:

$$\mathbb{E}(\eta_i \mid y_{SBS,i};\, y_{VAT,i};\, y_{PDR,i};\, \mathbf{x}_{Empl,i};\, \boldsymbol{\theta}^{(t)};\, i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}) =$$

$$y_{PDR,i} * \tau_{110,i}^{(t)} + y_{VAT,i} * \tau_{101,i}^{(t)} + y_{SBS,i} * \tau_{011,i}^{(t)} + \mathbb{E}(\eta_i \mid y_{SBS,i};\, y_{VAT,i};\, y_{PDR,i};\, \mathbf{x}_{Empl,i};\, \boldsymbol{\theta}^{(t)};\, i \in S_{111}) * \tau_{111,i}^{(t)}$$

(2.15)

The conditional expected value $\mathbb{E}(\eta_i \mid y_{SBS,i};\, y_{VAT,i};\, y_{PDR,i};\, x_{Empl,i};\, \boldsymbol{\theta}^{(t)}\,;\, i \in S_{111})$, in case all observations contain an error, is as follows (derivation by Robinson (2016, p. 30 and 43)):

$$\mathbb{E}(\eta_i \mid y_{SBS,i};\, y_{VAT,i};\, y_{PDR,i};\, x_{Empl,i};\, \boldsymbol{\theta}^{(t)};\, i \in S_{111}) =$$

$$\beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i} +$$

$$\begin{pmatrix} b_{SBS}^{(t)} * \sigma^{2(t)} & b_{VAT}^{(t)} * \sigma^{2(t)} & b_{PDR}^{(t)} * \sigma^{2(t)} \end{pmatrix} \sum_{111}^{(t)\,-1} \begin{pmatrix} y_{SBS,i} - a_{SBS}^{(t)} - b_{SBS}^{(t)} * (\beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i}) \\ y_{VAT,i} - a_{VAT}^{(t)} - b_{VAT}^{(t)} * (\beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i}) \\ y_{PDR,i} - a_{PDR}^{(t)} - b_{PDR}^{(t)} * (\beta_0^{(t)} + \beta_{Empl}^{(t)} * x_{Empl,i}) \end{pmatrix}$$

(2.16)

Expected Squared True Value

Characteristic to EM algorithm estimation is the necessity to also calculate expected squared values of unobserved variables, since in general for a random variable ζ from $\mathrm{Var}(\zeta) = \mathbb{E}(\zeta^2) - [\mathbb{E}(\zeta)]^2$ follows $\mathbb{E}(\zeta^2) \neq [\mathbb{E}(\zeta)]^2$ but $\mathbb{E}(\zeta^2) = [\mathbb{E}(\zeta)]^2 + \mathrm{Var}(\zeta)$. With regard to the Intermittent-Error model, a squared expected value for η is needed with regard to triplets in S_{111} when optimizing the Q-function (Expression (2.23) - (2.26)). Therefore, during the E-step, also $\mathbb{E}(\eta_i^2 \mid y_{SBS,i};\, y_{VAT,i};\, y_{PDR,i};\, x_{Empl,i};\, \boldsymbol{\theta}^{(t)}\,;\, i$

$\in S_{111}$) is estimated using $[\mathbb{E}(\eta_i \mid y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i};\ x_{Empl,i};\ \boldsymbol{\theta}^{(t)};\ i \in S_{111})]^2$ and $\mathrm{Var}(\eta_i \mid y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i};\ x_{Empl,i};\ \boldsymbol{\theta}^{(t)};\ i \in S_{111})$ shown in Expression (2.17) (idea of derivation by Robinson (2016), but there formula (6.34) contains an error).

$$\mathbf{Var}(\boldsymbol{\eta_i} \mid \boldsymbol{y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i}; x_{Empl,i};\ \theta^{(t)};\ i \in S_{111}}) =$$

$$\sigma^{2\,(t)} - \left(\sigma_{SBS}^{2\ (t)}\ \ \sigma_{VAT}^{2\ (t)}\ \ \sigma_{PDR}^{2\ (t)}\right) \sum_{111}^{(t)\ -1} \begin{pmatrix} \sigma_{SBS}^{2\ (t)} \\ \sigma_{VAT}^{2\ (t)} \\ \sigma_{PDR}^{2\ (t)} \end{pmatrix}$$

(2.17)

2.3.4 Log likelihood and Q-function

Complete data log likelihood:
The general form of the complete data log likelihood function of a finite mixture model with H components and n i.i.d. observations is (McLachlan & Peel, 2000):

$$\ell(\boldsymbol{\theta}) = \sum_{h=1}^{H} \sum_{i=1}^{n} z_{h,i} \{\log m_h + \log f_h(y_i \mid \boldsymbol{\theta})\}$$

(2.18)

with $z_{h,i} = 1$ if observation $\boldsymbol{y_i}$ belongs to component h and $z_{h,i} = 0$ otherwise and m_h and $f_h(.)$ denoting the mixing weight and density function of component h respectively. For the Intermittent-Error Model we use the following fact (based on the source independence assumption, see Section 2.2):

$$f_h(y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i};\ \eta_i \mid \theta) = f_h(y_{SBS,i} \mid \eta_i) * f_h(y_{VAT,i} \mid \eta_i) * f_h(y_{PDR,i} \mid \eta_i) * f_h(\eta_i)$$

(2.19)

The following log likelihood is obtained for parameters in $\boldsymbol{\theta} = \pi_{SBS}, \pi_{VAT}, \pi_{PDR}, \beta_o, \beta_{Empl}, \sigma^2, a_{SBS}, b_{SBS}, \sigma^2_{SBS}, a_{VAT}, b_{VAT}, \sigma^2_{VAT}, a_{PDR}, b_{PDR}, \sigma^2_{PDR}$:

$$\ell(\boldsymbol{\theta}) =$$

$$\begin{aligned}
&\sum_{i \in S_{000}} \{\log m_{000} + \log\delta(y_{SBS,i} - \eta_i) + \log\delta(y_{VAT,i} - \eta_i) + \log\delta(y_{PDR,i} - \eta_i) \\
&\qquad + \log N(\eta_i;\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)\} \\
&+ \sum_{i \in S_{001}} \{\log m_{001} + \log\delta(y_{SBS,i} - \eta_i) + \log\delta(y_{VAT,i} - \eta_i) + \log N(y_{PDR,i};\ a_{PDR} + b_{PDR} * \eta_i;\ \sigma^2_{PDR}) \\
&\qquad + \log N(\eta_i;\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)\} \\
&+ \sum_{i \in S_{010}} \{\log m_{010} + \log\delta(y_{SBS,i} - \eta_i) + \log N(y_{VAT,i};\ a_{VAT} + b_{VAT} * \eta_i;\ \sigma^2_{VAT}) + \log\delta(y_{PDR,i} - \eta_i) \\
&\qquad + \log N(\eta_i;\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)\} \\
&+ \sum_{i \in S_{100}} \{\log m_{100} + \log N(y_{SBS,i};\ a_{SBS} + b_{SBS} * \eta_i;\ \sigma^2_{SBS}) + \log\delta(y_{VAT,i} - \eta_i) + \log\delta(y_{PDR,i} - \eta_i) \\
&\qquad + \log N(\eta_i;\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)\} \\
&+ \sum_{i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}} \{\tau_{011,i}\ [\log m_{011} + \log\delta(y_{SBS,i} - \eta_i) + \log N(y_{VAT,i};\ a_{VAT} + b_{VAT} * \eta_i;\ \sigma^2_{VAT}) \\
&\qquad + \log N(y_{PDR,i};\ a_{PDR} + b_{PDR} * \eta_i;\ \sigma^2_{PDR}) + \log N(\eta_i;\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)] \\
&\qquad + \tau_{101,i}\ [\log m_{101} + \log N(y_{SBS,i};\ a_{SBS} + b_{SBS} * \eta_i;\ \sigma^2_{SBS}) + \log\delta(y_{VAT,i} - \eta_i) \\
&\qquad + \log N(y_{PDR,i};\ a_{PDR} + b_{PDR} * \eta_i;\ \sigma^2_{PDR}) + \log N(\eta_i;\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)] \\
&\qquad + \tau_{110,i}\ [\log m_{110} + \log N(y_{SBS,i};\ a_{SBS} + b_{SBS} * \eta_i;\ \sigma^2_{SBS}) + \log N(y_{VAT,i};\ a_{VAT} + b_{VAT} * \eta_i;\ \sigma^2_{VAT}) \\
&\qquad + \log\delta(y_{PDR,i} - \eta_i) + \log N(\eta_i;\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)] \\
&\qquad + \tau_{111,i}\ [\log m_{111} + \log N(y_{SBS,i};\ a_{SBS} + b_{SBS} * \eta_i;\ \sigma^2_{SBS}) + \log N(y_{VAT,i};\ a_{VAT} + b_{VAT} * \eta_i;\ \sigma^2_{VAT}) \\
&\qquad + \log N(y_{PDR,i};\ a_{PDR} + b_{PDR} * \eta_i;\ \sigma^2_{PDR}) + \log N(\eta_i;\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)]\ \}
\end{aligned}$$

(2.20)

where $N(\,.\,;\mu\,;\sigma^2)$ denotes the univariate normal density and $\delta(\,.\,)$ denotes the Dirac's delta function with mass on zero. In case of complete data this expression can be simplified to previous Expression (2.5).

If the complete data is available, η_i is available for each observation, also m_{011}, m_{101}, m_{110} and m_{111} are known, and therefore the likelihood of observations in set $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$ would have the same form as that of S_{000}, S_{001}, S_{010} and S_{100}. Expression (2.20) already has some of the form that emerges when, for $i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$, the specific error pattern is not known and for each of those triplets i an error pattern probability $\tau_{klm,i}$ is estimated. For complete data, in this representation, exactly one of the $\tau_{011,i}$, $\tau_{101,i}$, $\tau_{110,i}$ and $\tau_{111,i}$ is 1 and the others are zero, since the error patterns are known, and therefore the likelihood of observations in set $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$ obtains the same form as those for S_{000}, S_{001}, S_{010} and S_{100} in the expression above.

Conditional expected log likelihood: Q-function

For incomplete data, a conditional expected log likelihood is formulated. In sets S_{000}, S_{001}, S_{010} and S_{100} two or three of the observations are known to be equal to the True Value η_i. Therefore, in the Q-function η_i is replaced by one of the observed values. For the triplets belonging to set $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$, η_i is latent and therefore replaced by its conditional expectation, given the observed values and current parameter estimates $\mathbb{E}(\eta_i | y_{SBS,i}; y_{VAT,i}; y_{PDR,i}; x_{Empl,i}; \boldsymbol{\theta}^{(t)}; i \in S_{011} \cup S_{101} \cup S_{011} \cup S_{111})$, as described in Expression (2.15). This conditional expectation arises by, for each of these is, weighing the likelihoods of belonging to sets S_{000}, S_{001}, S_{010} and S_{100} with their corresponding τs ($\tau^{(t)}_{011,i}$, $\tau^{(t)}_{101,i}$, $\tau^{(t)}_{110,i}$ and $\tau^{(t)}_{111,i}$, definition in repeated in general in Expression (2.21) below). Therefore, the Q-function is related to the parameter values at time point (t) ($\boldsymbol{\theta}^{(t)}$), and when maximized, supplies parameter values for time point ($t+1$) ($\boldsymbol{\theta}^{(t+1)}$). When an observation is assumed to belong to set S_{011} (weighted with $\tau^{(t)}_{011,i}$), $y_{SBS,i}$ is assumed to measure η_i, and similarly with $y_{VAT,i}$ and $y_{PDR,i}$, when assumed to belong to set S_{101} and S_{110} respectively. Since no source measures η_i for observations in set S_{111}, an estimated $\mathbb{E}(\eta_i | y_{SBS,i}; y_{VAT,i}; y_{PDR,i}; x_{Empl,i}; \boldsymbol{\theta}^{(t)}; i \in S_{111})$ (see Expression (2.16)) is inserted in the Q-function.

$$\tau^{(t)}_{klm,i} = \mathbb{P}\left((z_{SBS,i}, z_{VAT,i}, z_{PDR,i}) = (k,l,m) \,\middle|\, y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i};\ x_{Empl,i};\ i \in S_{011} \cup S_{101} \cup S_{011} \cup S_{111};\ \theta^{(t)}\right) \tag{2.21}$$

Expression (2.22) implements these changes in (2.20) to obtain the Q-function. For convenient comparison to the complete data log likelihood, with regard to the identified two/three error-free observations, the Dirac's delta function is still present in the formula. However, these can be omitted since for example $\log\delta(y_{VAT,i} - y_{SBS,i}) = \log\delta(0)$ for error pattern (0, 0, 1) and therefore does not contain any parameters in case of observations in set S_{000} ($y_{VAT,i}$ and $y_{SBS,i}$ are the same when they are both error-free), and analogous in the other sets. To acquire analytical expressions for the parameters that optimize this Q-function, (Robinson, 2016, pp. 32-34) gives a rewritten version of the Q-function that combines the parts in the expression with regard to the occurring parameters. This rewritten Q-function is given in Expression (2.23), (2.24), (2.25) and (2.26). Since its purpose is acquiring optimal parameter estimates, all constants in the univariate normal densities are left out and so are the Dirac's delta functions.

$$
\begin{aligned}
&Q_{\theta^{(t)}}\left(\theta^{(t+1)}\right) = \\
&\sum_{i \in S_{000}}\{\log m_{000} + \log\delta(y_{SBS,i} - y_{SBS,i}) + \log\delta(y_{VAT,i} - y_{SBS,i}) + \log\delta(y_{PDR,i} - y_{SBS,i}) \\
&\qquad + \log N(y_{SBS,i};\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)\} \\
&+ \sum_{i \in S_{001}}\{\log m_{001} + \log\delta(y_{SBS,i} - y_{SBS,i}) + \log\delta(y_{VAT,i} - y_{SBS,i}) + \log N(y_{PDR,i}; a_{PDR} + b_{PDR} * y_{SBS,i}; \sigma^2_{PDR}) \\
&\qquad + \log N(y_{SBS,i};\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)\} \\
&+ \sum_{i \in S_{010}}\{\log m_{010} + \log\delta(y_{SBS,i} - y_{SBS,i}) + \log N(y_{VAT,i}; a_{VAT} + b_{VAT} * y_{SBS,i}; \sigma^2_{VAT}) + \log\delta(y_{PDR,i} - y_{SBS,i}) \\
&\qquad + \log N(y_{SBS,i};\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)\} \\
&+ \sum_{i \in S_{100}}\{\log m_{100} + \log N(y_{SBS,i}; a_{SBS} + b_{SBS} * y_{VAT,i}; \sigma^2_{SBS}) + \log\delta(y_{VAT,i} - y_{VAT,i}) + \log\delta(y_{PDR,i} - y_{VAT,i}) \\
&\qquad + \log N(y_{VAT,i};\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)\} \\
&+ \sum_{i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}}\{\tau^{(t)}_{011,i}[\log m_{011} + \log\delta(y_{SBS,i} - y_{SBS,i}) + \log N(y_{VAT,i}; a_{VAT} + b_{VAT} * y_{SBS,i}; \sigma^2_{VAT}) \\
&\qquad + \log N(y_{PDR,i}; a_{PDR} + b_{PDR} * y_{SBS,i}; \sigma^2_{PDR}) + \log N(y_{SBS,i};\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)] \\
&\qquad + \tau^{(t)}_{101,i}[\log m_{101} + \log N(y_{SBS,i}; a_{SBS} + b_{SBS} * y_{VAT,i}; \sigma^2_{SBS}) + \log\delta(y_{VAT,i} - y_{VAT,i}) \\
&\qquad + log N(y_{PDR,i}; a_{PDR} + b_{PDR} * y_{VAT,i}; \sigma^2_{PDR}) + \log N(y_{VAT,i};\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)] \\
&\qquad + \tau^{(t)}_{110,i}[\log m_{110} + \log N(y_{SBS,i}; a_{SBS} + b_{SBS} * y_{PDR,i}; \sigma^2_{SBS}) \\
&\qquad + \log N(y_{VAT,i}; a_{VAT} + b_{VAT} * y_{PDR,i}; \sigma^2_{VAT}) + log\delta(y_{PDR,i} - y_{PDR,i}) \\
&\qquad + \log N(y_{PDR,i};\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2)] \\
&\qquad + \tau^{(t)}_{111,i}[\log m_{111} \\
&\qquad + \mathbb{E}(\log N(y_{SBS,i}; a_{SBS} + b_{SBS} * \eta_i;\ \sigma^2_{SBS}) \mid y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i};\ x_{Empl,i}; \theta^{(t)}; i \in S_{111}) \\
&\qquad + \mathbb{E}(\log N(y_{VAT,i}; a_{VAT} + b_{VAT} * \eta_i;\ \sigma^2_{VAT}) \mid y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i};\ x_{Empl,i}; \theta^{(t)}; i \in S_{111}) \\
&\qquad + \mathbb{E}(\log N(y_{PDR,i}; a_{PDR} + b_{PDR} * \eta_i;\ \sigma^2_{PDR}) \mid y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i};\ x_{Empl,i}; \theta^{(t)}; i \in S_{111}) \\
&\qquad + \mathbb{E}(\log N(\eta_i;\ \beta_0 + \beta_{Empl} * x_{Empl,i};\ \sigma^2) \mid y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i};\ x_{Empl,i}; \theta^{(t)}; i \in S_{111})]\}
\end{aligned}
\tag{2.22}
$$

$$
\begin{aligned}
&Q_{\theta^{(t)}}\left(\theta^{(t+1)}\right) = \\
&\sum_{i \in S_{000}} \log m_{000} + \sum_{i \in S_{001}} \log m_{001} + \sum_{i \in S_{010}} \log m_{010} + \sum_{i \in S_{100}} \log m_{100} \\
&+ \sum_{i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}}\left\{\tau^{(t)}_{011,i} \log m_{011} + \tau^{(t)}_{101,i} \log m_{101} + \tau^{(t)}_{110,i} \log m_{110} + \tau^{(t)}_{111,i} \log m_{111}\right\} \\
&- \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2} V^{(t)}_{\sigma} - \frac{n^{(t)}_{\sigma SBS}}{2}\log\sigma^2_{SBS} - \frac{1}{2\sigma^2_{SBS}} V^{(t)}_{\sigma SBS} - \frac{n^{(t)}_{\sigma VAT}}{2}\log\sigma^2_{SBS} - \frac{1}{2\sigma^2_{VAT}} V^{(t)}_{\sigma VAT} \\
&- \frac{n^{(t)}_{\sigma PDR}}{2}\log\sigma^2_{PDR} - \frac{1}{2\sigma^2_{PDR}} V^{(t)}_{\sigma PDR}
\end{aligned}
\tag{2.23}
$$

$n_{\sigma SBS}$, $n_{\sigma VAT}$ and $n_{\sigma PDR}$ represent the expected number of erroneous measurements in each source and the Vs are squared losses (Robinson, 2016). Their expressions are shown in (2.24), (2.25) and (2.26).

$$
\begin{aligned}
n^{(t)}_{\sigma SBS} &= \sum i \in S_{100}\, z_{SBS,i} + \sum i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}\left(\tau^{(t)}_{101,i} + \tau^{(t)}_{110,i} + \tau^{(t)}_{111,i}\right) \\
n^{(t)}_{\sigma VAT} &= \sum i \in S_{010}\, z_{VAT,i} + \sum i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}\left(\tau^{(t)}_{011,i} + \tau^{(t)}_{110,i} + \tau^{(t)}_{111,i}\right) \\
n^{(t)}_{\sigma PDR} &= \sum i \in S_{001}\, z_{PDR,i} + \sum i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}\left(\tau^{(t)}_{011,i} + \tau^{(t)}_{101,i} + \tau^{(t)}_{111,i}\right)
\end{aligned}
\tag{2.24}
$$

$$V_{\sigma}^{(t)} =$$

$$\begin{aligned}
&\sum_{i \in S_{000}} \left(y_{SBS,i} - \beta_0 - \beta_{Empl} * x_{Empl,i}\right)^2 \\
&+ \sum_{i \in S_{001}} \left(y_{SBS,i} - \beta_0 - \beta_{Empl} * x_{Empl,i}\right)^2 \\
&+ \sum_{i \in S_{010}} \left(y_{SBS,i} - \beta_0 - \beta_{Empl} * x_{Empl,i}\right)^2 \\
&+ \sum_{i \in S_{100}} \left(y_{VAT,i} - \beta_0 - \beta_{Empl} * x_{Empl,i}\right)^2 \\
&+ \sum_{i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}} \Big\{ \tau_{011,i}^{(t)} \left(y_{SBS,i} - \beta_0 - \beta_{Empl} * x_{Empl,i}\right)^2 \\
&\qquad + \tau_{101,i}^{(t)} \left(y_{VAT,i} - \beta_0 - \beta_{Empl} * x_{Empl,i}\right)^2 \\
&\qquad + \tau_{110,i}^{(t)} \left(y_{PDR,i} - \beta_0 - \beta_{Empl} * x_{Empl,i}\right)^2 \\
&\qquad + \tau_{111,i}^{(t)} \Big[\mathbb{E}\left(\eta_i^2 \middle| y_{SBS,i}; y_{VAT,i}; y_{PDR,i}; x_{Empl,i}; \theta^{(t)}; i \in S_{111}\right) \\
&\qquad - 2\left(\beta_0 + \beta_{Empl} * x_{Empl,i}\right) * \mathbb{E}\left(\eta_i \middle| y_{SBS,i}; y_{VAT,i}; y_{PDR,i}; x_{Empl,i}; \theta^{(t)}; i \in S_{111}\right) \\
&\qquad + \left(\beta_0 + \beta_{Empl} * x_{Empl,i}\right)^2 \Big] \Big\}
\end{aligned}$$

(2.25)

$V_{\sigma SBS}$ is shown in Expression (2.26). $V_{\sigma VAT}$ and $V_{\sigma PDR}$ are similar, but regard the weights (τs) of group S_{011}, S_{110} and S_{111} and S_{011}, S_{101} and S_{111} respectively.

$$V_{\sigma SBS}^{(t)} =$$

$$\begin{aligned}
&\sum_{i \in S_{100}} \left(y_{SBS,i} - a_{SBS} - b_{SBS} * y_{VAT,i}\right)^2 \\
&+ \sum_{i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}} \Big\{ \tau_{101,i}^{(t)} * \left(y_{SBS,i} - a_{SBS} - b_{SBS} * y_{VAT,i}\right)^2 + \tau_{110,i}^{(t)} \left(y_{SBS,i} - a_{SBS} - b_{SBS} * y_{PDR,i}\right)^2 \\
&\qquad + \tau_{111,i}^{(t)} \Big[{y_{SBS,i}}^2 - 2 y_{SBS,i} * \left(a_{SBS} + b_{SBS} * \mathbb{E}\left(\eta_i \middle| y_{SBS,i}; y_{VAT,i}; y_{PDR,i}; x_{Empl,i}; \theta^{(t)}; i \in S_{111}\right)\right) \\
&\qquad + {a_{SBS}}^2 + 2 a_{SBS} * b_{SBS} \\
&\qquad * \mathbb{E}\left(\eta_i \middle| y_{SBS,i}; y_{VAT,i}; y_{PDR,i}; x_{Empl,i}; \theta^{(t)}; i \in S_{111}\right) \\
&\qquad + {b_{SBS}}^2 * \mathbb{E}\left(\eta_i^2 \middle| y_{SBS,i}; y_{VAT,i}; y_{PDR,i}; x_{Empl,i}; \theta^{(t)}; i \in S_{111}\right) \Big] \Big\}
\end{aligned}$$

(2.26)

2.3.5 Maximization step

Differentiating the Q-function in Expression (2.23), and equating the result to zero gives familiar analytical expressions for the updated parameters in the linear expressions of the three sources ($a_{SBS}^{(t+1)}$, $b_{SBS}^{(t+1)}$, $\sigma^2{}_{SBS}^{(t+1)}$, $a_{VAT}^{(t+1)}$, $b_{VAT}^{(t+1)}$, $\sigma^2{}_{VAT}^{(t+1)}$, $a_{PDR}^{(t+1)}$, $b_{PDR}^{(t+1)}$, $\sigma^2{}_{PDR}^{(t+1)}$) and the True Value distribution ($\beta_0^{(t+1)}$, $\beta_{Empl}^{(t+1)}$, $\sigma^{2(t+1)}$). These are described by Robinson (2016, pp. 36-37, 45-46). The obtained updated πs are also very intuitive, as shown below in the example of Expression (2.27).

$$\pi_{SBS}^{(t+1)} = \frac{\sum_{i \in S_{100}} z_{SBS,i} + \sum_{i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}} \left(\tau_{101,i}^{(t)} + \tau_{110,i}^{(t)} + \tau_{111,i}^{(t)}\right)}{n}$$

(2.27)

2.4 Summary

In response to research sub question 2 *What is the Intermittent-Error Model?*

The Intermittent-Error Model is a model that enables the estimation of *True Values* on a variable when *multi-source data* is available that does not agree on theoretically the same variable for all records. The Intermittent-Error Model was proposed by Guarnera & Varriale ((2015) and (2016)) and is distinguished from other approaches (such as Structural Equation Models) by its intermittent error mechanism. The model assumes that each measurement from each source g has a certain probability π_g to contain an error ($z_{g,i} = 1$) and therefore also a probability ($1 - \pi_g$) to be error-free ($z_{g,i} = 0$). Since the model also assumes that the occurrences of errors are independent across sources, and the error distributions are continuous, some error-free measurements can be identified from the data. When two sources measure the same value on a certain unit, the value is always considered to be error-free. After all, the probability that two independent continuous error distributions produce the same error is negligible.

With regard to the case study data set, each statistical business unit's triplet of turnover values ($y_{SBS,i}$, $y_{VAT,i}$, $y_{PDR,i}$) has one of $2^3 = 8$ possible error patterns:

$$(z_{SBS,i}, z_{VAT,i}, z_{PDR,i}) \in \{(0,0,0), (0,0,1), (0,1,0), (1,0,0), (0,1,1), (1,0,1), (1,1,0), (1,1,1)\}$$

The units with one of the first four error patterns can be identified from the data and by definition only identifiable observation triplets have one of the first four patterns. Therefore, the remaining units have one of the latter four error patterns. Observation triplets in these last four error patterns have unknown True Value and error pattern (since two or more observations contain an error, all represent different values and the error-free observation (if any) cannot be identified from the observation triplet). Therefore, probabilities (τ_{011}, τ_{101}, τ_{110} and τ_{111}) to belong to each of the four remaining groups are estimated. The model assumes that the occurrence of the errors is unrelated to the True Value and completely determined by the π_g parameter, therefore expected True Values can be calculated. By means of the expected error probabilities (the τs) and expected True Values, maximum likelihood parameter estimates can be obtained by the Expectation-Maximization (EM) algorithm.

The EM estimation procedure uses values obtained from the units with known error patterns as starting values and iteratively estimates expected values for the probabilities to belong in the remaining four sets ($\tau_{011}^{(t)}$, $\tau_{101}^{(t)}$, $\tau_{110}^{(t)}$ and $\tau_{111}^{(t)}$) and the expected conditional True Value $\mathbb{E}(\eta_i | y_{SBS,i}; y_{VAT,i}; y_{PDR,i}; x_{Empl,i}; \boldsymbol{\theta}^{(t)}; i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111})$ for each unit (with the parameters in $\boldsymbol{\theta}$ described in Table 2.3 on page 49). By iterating Expectation and Maximization step, maximum likelihood estimates are obtained for the parameters in Table 2.3 (on page 49) that relate to the measured values $y_{SBS,i}$, $y_{VAT,i}$ and $y_{PDR,i}$, in the following way:

$$y_{SBS,i} = (1 - z_{SBS,i})\,\eta_i + z_{SBS,i}\,(a_{SBS} + b_{SBS} * \eta_i + \epsilon_{SBS,i})$$

$$y_{VAT,i} = (1 - z_{VAT,i})\,\eta_i + z_{VAT,i}\,(a_{VAT} + b_{VAT} * \eta_i + \epsilon_{VAT,i})$$

$$y_{PDR,i} = (1 - z_{PDR,i})\,\eta_i + z_{PDR,i}\,(a_{PDR} + b_{PDR} * \eta_i + \epsilon_{PDR,i})$$

The as and bs in these expressions capture systematic deviations in turnover concept of the source measurement with regard to the True Value. Also, in the estimation procedure the following conditional distribution on the True Value is used (further investigated with regard to Research sub question 3a in Section 3.1):

$$\eta_i = \beta_0 + \beta_{\text{Empl}} * x_{\text{Empl,i}} + \epsilon_i$$

Table 2.3 Explaining table for the case study parameters (in $\boldsymbol{\theta}$) of the Intermittent-Error Model

π_{SBS}	expected proportion of errors present in *Turnover* values from the Structural Business Statistics survey
π_{VAT}	expected proportion of errors present in *Turnover* values from the VAT register
π_{PDR}	expected proportion of errors present in *Turnover* values from the Profit Declaration Register
β_o	the intercept of the linear regression model belonging to the conditional True Value distribution $\mathbb{E}(\eta_i \mid x_{Empl,i}) = \boldsymbol{\beta_o} + \beta_{Empl} * x_{Empl,i}$
β_{Empl}	the slope parameter for regression covariate x_{Empl} of the linear regression model belonging to the conditional True Value distribution $\mathbb{E}(\eta_i \mid x_{Empl,i}) = \boldsymbol{\beta_o} + \beta_{Empl} * x_{Empl,i}$
σ^2	variance of η unexplained by covariate x_{Empl}
a_{SBS}	intercept bias on *Turnover* values from the Structural Business Statistics survey
b_{SBS} - 1	slope bias on *Turnover* values from the Structural Business Statistics survey
σ^2_{SBS}	variance of *Turnover* values from the Structural Business Statistics survey after correction for intercept bias and slope bias
a_{VAT}	intercept bias on *Turnover* values from the VAT register
b_{VAT} - 1	slope bias on *Turnover* values from the VAT register
σ^2_{VAT}	variance of turnover values from the VAT register after correction for intercept bias and slope bias
a_{PDR}	intercept bias on turnover values from the Profit Declaration Register
b_{PDR} - 1	slope bias on turnover values from the Profit Declaration Register
σ^2_{PDR}	variance of turnover values from the Profit Declaration Register after correction for intercept bias and slope bias

3 How to assess the Intermittent-Error Model

How do you know whether the Intermittent-Error Model fits a data set? McLachlan and Peel state about finite mixture models in general that *"The problem of assessing model fit is not straightforward for mixture models at least for multivariate data."* (McLachlan & Peel, 2000, p. 84). This is the case because the assumptions that need to be tested are used to assign the observations to the mixture components, which are the sets determined by the error patterns in the Intermittent-Error Model. Therefore, testing the assumptions based on expected values that follow from the component assignments would paint a too positive picture of the model fit. Fortunately, with regard to the Intermittent-Error Model, the error pattern assignments are known for some of the observations. This fact forms the basis of the proposed *fit measures* in Section 3.3.1. Apart from fit measures, also other ways to assess the sensibility of the model fit are proposed in Section 3.3.2 which are denoted by *soundness measures*. But first Section 3.1 and 3.2 zoom in on the component of the Intermittent-Error Model that is least straight forward: the distributional assumptions on the True Value. Section 3.1 discusses the need for these assumptions in the estimation procedure of the Intermittent-Error Model, and as a result their theoretical influence. Section 3.2 shows the results of some simulations to assess the practical influence of the conditional True Value distributional. Section 3.4 concludes this chapter with a summary.

3.1 The conditional 'True Value'-distribution

> *"According to this approach, all the available information is used and 'weighted' according to its reliability and a prediction of 'true' values of some numeric variable of interest is obtained conditional on all the available information." (Guarnera & Varriale, 2016, p. 537)*

The citation from Guarnera & Varriale shown above indicates the intended structure of the Intermittent-Error Model. The multiple sources are evaluated with regard to the error proportions (π_g), error size parameters (σ_g^2) and systematic deviations (a_g and b_g), in order to 'weigh' the multiple observations into a True Value estimate. Therefore, it seems that the original focus is on the left side of Figure 2.1 (on page 37). So what is the necessity of assuming the right hand side of the figure? And how do the assumptions on the True Value distribution influence the model estimation procedure?

True Value distribution in the estimation procedure

With regard to the observation triplets in sets S_{000}, S_{001}, S_{010} and S_{100} the True Value η_i is observed and no distributional assumptions are required. With regard to set S_{111}, the True Value distribution's role is clearest, since none of the observed values represent the True Value η_i. With regard to the observation triplets in sets S_{011}, S_{101}, S_{110}, η_i is known, but it is unknown which triplets belong to these sets. It is only known whether $i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$, and therefore probabilities of belonging to each set (τ_{011}, τ_{101}, τ_{110} and τ_{111}) are estimated and used for weighing. These τs are the weights that Guarnera and Varriale refer to in the above citation. To obtain these weights, error pattern likelihoods are calculated, and these likelihoods need assumptions on the distribution of True Values.

So the conditional True Value distribution plays a role in the error pattern likelihoods to obtain τ_{klm}, denoted by $f_{klm}(y_{SBS,i}\,; y_{VAT,i}\,; y_{PDR,i} \mid x_{Empl,i};\, \boldsymbol{\theta})$ (expression (2.11) and (2.13) on page 42) and the

expected value of the conditional True Value distribution is used to obtain an estimate of the True Value under the assumption that all measurements are erroneous ($i \in S_{111}$) (expression (2.16) on page 43). Expression (2.16) shows that the *Conditional expectation w.r.t. covariate Number of Employees,* shown below, is only a part of the estimate of the True Value in set S_{111}. Also for each individual triplet of observations, this expectation for $i \in S_{111}$ is weighted with the expectations for sets S_{011}, S_{101} and S_{110} for which the expectation is one of the observed measurements:

Estimated True Value

$$\mathbb{E}(\eta_i \mid y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i}; x_{Empl,i};\ \boldsymbol{\theta}^{(t)}\ ; i \in S_{011} \cup S_{101} \cup S_{110} \cup S_{111}) =$$

$$y_{PDR,i} * \tau_{110,i}^{(t)} + y_{VAT,i} * \tau_{101,i}^{(t)} + y_{SBS,i}\ * \tau_{011,i}^{(t)} + \mathbb{E}(\eta_i \mid y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i}; x_{Empl,i};\ \boldsymbol{\theta}^{(t)};\ i \in S_{111}) * \tau_{111,i}^{(t)}$$

(3.1)

Conditional expectation w.r.t. covariate Number of Employees
(The conditional expectation of the True Value with regard to the covariate *Number of Employees.)*

$$\mathbb{E}(\eta_i | x_{Empl,i}) = \beta_0 + \beta_{Empl} * x_{Empl,i}$$

(3.2)

Thus the expectation from the conditional True Value distribution only influences the estimated True Value for observations triplets with high τ_{111}. For other triplets in $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$, $\mathbb{E}(\eta_i \mid y_{SBS,i}\ ;\ y_{VAT,i}\ ;\ y_{PDR,i}\ ; x_{Empl,i}\ ;\ \boldsymbol{\theta}^{(t)};\ i \in S_{111})$ is outweighed by the observed values and therefore also $\mathbb{E}(\eta_i | x_{Empl,i})$ plays a negligible role.

3.2 Influence of 'True Value'-distribution on estimated True Values

Since it might be no optimal predicting covariates are available to constitute the conditional True Value distribution, the question is what the influence is of covariates with bad fit. Two scenarios are proposed that can be considered 'bad'. The first is denoted by *Noise covariates*, and is characterized by covariates that are completely unrelated to the expected True Value. The second is denoted by *Misleading covariates*, and is characterized by covariates that perfectly predict the expected True Value for triplets for which the True Value is observed (triplets in sets S_{000}, S_{001}, S_{010} and S_{100}) and for which the conditional True Value distribution does not play a role, and with pure noise covariate values for the remaining triplets.

Multiple simulations were run to observe the effect of Noise covariates and Misleading covariates. The basis of the simulated data set were 400 simulated values for the covariate x_{Empl}, which were drawn from a mixture of two Poisson distributions with $\lambda_1 = 5$ and $\lambda_2 = 50$, to obtain values that relate to the target covariate *Number of Employees*. These covariate values were log transformed (as would be the case for real data, see Section 4.1.1).

Parameter values were defined that were similar to estimates found on real data (see Section 4.1.3). These parameter values are shown in the bottom row of Table 3.1 (on page 55). With regard to the log transformed covariate values, and the defined β_0 and β_{Empl}, conditional expectations for the True Values were obtained. These were used to generate the η_is with the defined unexplained error variance σ^2. $z_{SBS,i}$, $z_{VAT,i}$ and $z_{PDR,i}$ values were generated from Bernoulli distributions with the defined π_{SBS}, π_{VAT} and π_{PDR} error probabilities, and for all units the observation triplets ($y_{SBS,i}$, $y_{VAT,i}$, $y_{PDR,i}$) were generated using the zs and defined $a_{SBS}/a_{VAT}/a_{PDR}$, $b_{SBS}/b_{VAT}/b_{PDR}$ and $\sigma^2_{SBS}/\ \sigma^2_{VAT}/\ \sigma^2_{PDR}$ following the formulas given in Section 2.1 and shown for the case study data set in Figure 2.1 (on page 37)).

Noise covariates/Misleading covariates

The Intermittent-Error Model was fitted to the perfect data set obtained with the above procedure, but also for data sets in which only the covariates were replaced by new values. Therefore, the distributions on the observed *Turnover* values are the same in all simulations. Figure 3.1 (on page 54) shows the empirical distribution of an example simulation. As a result of the mixture of Poissons for the generated covariate values, also the observed values are not normally distributed. However, the data does meet all assumptions described in Section 2.2, except for those on the conditional True Value distribution. So all source error variances are normally distributed, and for the perfect data set also the variance unexplained by the covariate is normally distributed.

To obtain simulated data sets with Noise covariates, new covariate values were generated from a uniform distribution (with lower bound 0 and upper bound 200), that as a result, were completely unrelated to the simulated observed values. To obtain simulated data sets with Misleading covariates, for observation triplets in sets S_{000}, S_{001}, S_{010} and S_{100}, covariates were generated that perfectly predicted the expected True Value, therefore leaving zero σ^2 as starting value for the EM estimation procedure. On the contrary, for the remaining sets, for which the estimation procedure needs to estimate the True Value, complete noise covariates were generated. Parameter estimates for a typical simulated data set in each of the three procedures are shown in Table 3.1.

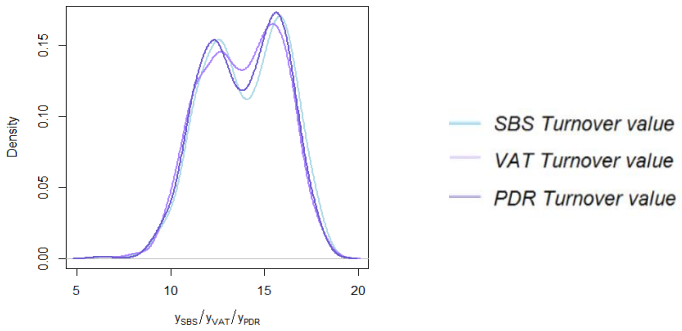


Figure 3.1 Empirical density functions of simulated $y_{SBS,i}$, $y_{VAT,i}$, $y_{PDR,i}$

Figure 3.3 (on page 55) shows three typical simulated data sets, each with its own simulation approach. The observation triplets are plotted for the observations for which τ_{111} was larger than the other τs, thus for which the Intermittent-Error Model assumes that all three sources have measured the turnover value erroneously. The number of triplets with high τ_{111} vary between 20 and 40 within each simulation procedure, as shown in Figure 3.2.

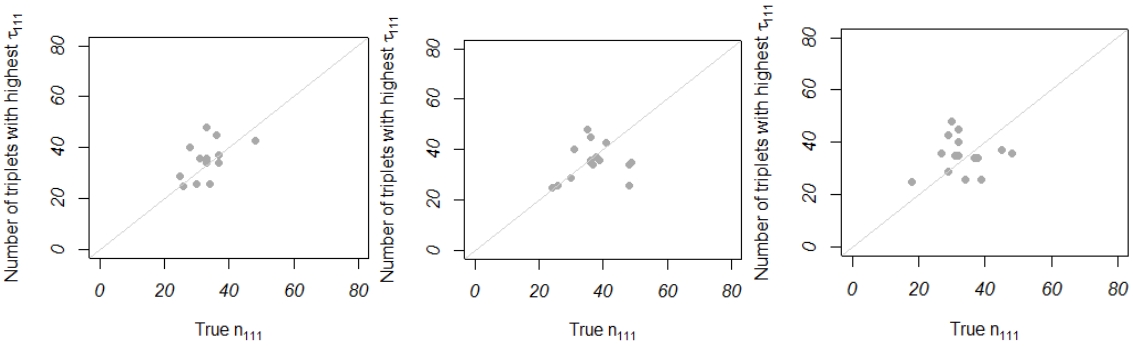


Figure 3.2 Estimated number of triplets for which the estimated τ_{111} is higher than τ_{011}, τ_{101} and τ_{110} plotted against the true number of triplets with error pattern (1, 1, 1) for 25 simulated samples.
Left: perfect covariates, Middle: noise covariates, Right: misleading covariates
The grey lines represent the values for which Number of triplets with highest τ_{111} = n_{111}.

'Estimated True Value' and 'Conditional expectation w.r.t. covariate Number of Employees'

The *Estimated True Value* represents the weighted estimate given that the exact error pattern is unknown. The expression for this value was restated in Expression (3.1). By *conditional expectation w.r.t. covariate Number of Employees* the value is meant that does not consider the measurements, but follows directly from the True Value distribution's estimated parameters. The expression for this value was given in Expression (3.2).

Influence of covariates on estimated True Values

The upper part of Figure 3.3 shows *Conditional expectations w.r.t. covariate Number of Employees* that form reliable expectations of the True Value because the model assumptions are met. The lower part shows conditional expectations that are useless, but in which case the model is misled. The middle part shows conditional expectations that are useless, but for which the model is also enabled to recognize the uselessness. Table 3.1 (on page 55) shows the parameter estimates of the Intermittent-Error Model on the typical case study data sets that were used for Figure 3.3. The parameter estimate β_{Empl} of zero for the Noise covariates shows that the model recognizes that there was no relation between the pure noise covariate and the True Value, and as a result the model estimated a large

unexplained variance σ^2. Since this estimated variance in the True Value distribution is large, the estimation with regard to the covariate has little influence on the estimated True Value. As shown in the middle part of Figure 3.3 the estimated True Value is often very close to the measured PDR *Turnover* which has the smallest estimated source error variance σ^2_{PDR}. The same holds for the misleading covariates, which show a smaller σ^2 as a result of the perfect fit on the triplets in sets S_{000}, S_{001}, S_{010} and S_{100}, the influence of the True Value distribution on the estimated True Value is limited, as shown in the lower part of Figure 1.1. Since the variance of the erroneous PDR measurement σ^2_{PDR} is still much smaller, the estimated True Values stay in the range of the measurements. Moreover, the estimations are quite comparable to the estimations in the perfect model fit case shown in the upper part of the picture. Therefore, it can be concluded that the influence of the True Value distribution on the estimated True Values is limited.

Table 3.1 Parameter estimates for typical simulated data sets

	π_{SBS}	π_{VAT}	π_{PDR}	β_0	β_{Empl}	σ^2	a_{SBS}	b_{SBS}	σ^2_{SBS}	a_{VAT}	b_{VAT}	σ^2_{VAT}	a_{PDR}	b_{PDR}	σ^2_{PDR}
Perfect	0.14	0.89	0.62	10.11	1.45	1.02	-1.00	1.06	1.17	0.15	0.98	0.31	0.26	0.96	0.06
Noise	0.19	0.88	0.65	13.89	0.00	4.63	-2.01	1.13	0.79	-0.24	1.00	0.30	-0.06	0.99	0.06
Misleading	0.11	0.89	0.64	10.42	1.35	1.87	-0.43	1.01	1.60	0.04	0.98	0.27	0.27	0.96	0.07
Simulation parameters	0.10	0.90	0.65	10.00	1.50	1.00	0.90	0.90	1.50	-0.30	1.00	0.25	0.15	0.99	0.01

Figure 3.3 Typical simulated turnover values from three sources ($y_{SBS,i}$, $y_{VAT,i}$, $y_{PDR,i}$), *Estimated True Values* and *Conditional expectations w.r.t. covariate Number of Employees* of triplets for which the estimated τ_{111} is higher than τ_{011}, τ_{101} and τ_{110}.
Upper: Perfect fit covariate; Middle: Pure noise covariate; Lower: Misleading covariate
The True Value estimates are scattered to make them visible for most units. The vertical lines facilitate observing the five values on the same statistical business unit. The horizontal axis contains a random index for each triplet.

3.3 Fit/Soundness measures

The Intermittent-Error Model describes erroneously measured values that are related to True Values under the linear model assumptions described in Section 2.2. Also, a linear model on the True Value distribution is used to describe likelihoods of error patterns for observations triplets that do not have immediately identified error patterns. In addition, the linear conditional True Value distribution is used to obtain expected values for the True Value when all measurements are considered erroneous.

Whether the assumptions of the linear models for True Value and source errors fit the data can be assessed with *fit measures*. Assessment of the sensibility of the predictions on True Values and the likelihoods on error patterns can be considered a *soundness measure* of the model fit. This section discusses in what way the model fit can be diagnosed with fit and soundness measures. Section 3.3.1 describes the applicability of linear model diagnostics to the Intermittent-Error Model. Section 3.3.2 discusses soundness measures, with Section 3.3.2.1 focusing on the proposed True Values and Section 3.3.2.2 investigating the error pattern likelihoods.

3.3.1 Fit measures

Both the parameters of the True Value distribution (right side of Figure 2.1, on page 37) and the parameters of the source error distributions (left side of Figure 2.1, on page 37) assume simple linear models with Ordinary Least Squares (OLS) fit. The fit of these assumed models can be assessed with fit measures generally applied in simple linear model diagnostics. However, since for many records the True Value η_i is unknown, either the dependent variable of the linear model (in the True Value distribution) or the covariate of the linear model (in the source error distributions) is missing. In these cases, η_i is estimated using the assumption of the model and therefore, the model fit cannot be assessed with general linear model diagnostics. Fortunately, the True Value is not unknown for every record, and the η_is are assumed to be Missing At Random (MAR) (see Section 2.2). This MAR assumption indicates that the missingness of the True Value is independent from the value of the True Value and the missing True Values are described in the same way by the model as the available ones. Thus, the True Values that are available do not deviate from those that are missing, and can therefore be assessed with regard to the linear model fit of the True Value and source error distributions.

This section deals with the observed triplets in sets S_{000}, S_{001}, S_{010} and S_{100}, for which simple linear model fit measures are proposed to assess the True Value model fit and the fit of the source error models. While residuals analysis in linear models are generally carried out on (externally) studentized residuals, the fit measures for the Intermittent-Error Model use raw residuals. After all, only for the starting values a measure of leverage can be applied to the observations in these sets, while the final residuals follow from linear model fit on all observations. Since the model fit on all observations follows from the iterative EM algorithm, no straightforward measure of leverage of the observations in the starting values sets can used to correct the residuals.

3.3.1.1 True Value distribution: Simple Linear Model fit measures

For records in sets S_{000}, S_{001}, S_{010} and S_{100} a True Value is known, and the value of the covariate *Number of Employees* is known (since that is known for all records). Therefore, the model fit of the linear model, with parameter estimates on all data, can be assessed using the data records in these four sets by using traditional linear model diagnostics in these sets. Traditional linear model diagnostics that are used are those to detect nonlinearity, non-normality and non-constant error variance.

Nonlinearity
Nonlinearity can be investigated by plotting the known True Values in sets S_{000}, S_{001}, S_{010} and S_{100} against the covariate *Number of Employees*, and investigating the trend in the plotted data for linearity. The trend in the data can be visualized by a lowess curve.

Non-normality
Various tests exist to test for non-normality in linear model residuals. The most general start to assess non-normality is to plot the quantiles in the observed residuals against the theoretic quantiles from the normal distribution. Any severe deviations from linearity in this QQ-plot show deviations from normality. When enough observations are available also an empirical density plot can be produced. This plot not only shows heavy tails of the distribution, as the QQ-plot does, but also reveals multiple modes.

Non-constant error variance
The error variance is considered non-constant when the size of the residuals varies along with the size of the covariate. A frequently observed pattern is, for example, that the residuals are larger for covariates that are larger. Patterns like this can be observed by plotting the residuals against the covariates or fitted values.

3.3.1.2 Source error distributions: Simple Linear Model fit measures

Since the source error distributions assume similar simple linear models as the True Value distribution does, the same fit measures can be used. However, while the True Values from all four sets S_{000}, S_{001}, S_{010} and S_{100} can be used to assess the model fit with regard to the covariate, to assess the model fit of the erroneously measured values against the True Value only S_{001}, S_{010} and S_{100} can be used. After all, the records in these sets are the only ones that contain a known True Value as well as a known erroneously measures value. Since the linear relation between measured value and True Value is estimated for each source separately, fit measures also apply to the records in these three groups separately. Therefore, the three fit measures discussed in Section 3.3.1.1 need to be carried out three times. It depends on the amount of data available in each group whether that is insightful enough to carry out completely.

3.3.2 Soundness measures

Although the observed triplets in sets S_{000}, S_{001}, S_{010} and S_{100} play an important role in the estimation of the parameters, it is the assignment of the observations to the other four sets S_{011}, S_{101}, S_{110} and S_{111} that is the main purpose of the Intermittent-Error Model. After all, under the assumption that identical measurements are measurement error-free, no model would be needed to obtain True Values for the records in the first four sets. To assess soundness of the Intermittent-Error Model, the probabilities for observation triplets to be assigned to S_{011}, S_{101}, S_{110} or S_{111} (the τs) need to be investigated. Especially with regard to the observations with high τ_{111}, the Intermittent-Error Model soundness depends on the proposed True Value. The soundness measures in this section investigate how these proposed True Values relate to the measured values and whether the model is able to propose sensible error pattern probabilities (the τs), which is related to the error pattern likelihoods.

3.3.2.1 Estimated True Value in relation to measured values

For the records with high probability to belong to set S_{1111}, with τ_{111} higher than the other τs, the proposed True Value is more driven by the True Value distribution than in the cases in which it is assumed that one of the measured values equals the True Value (S_{011}, S_{101} and S_{110}). Therefore, whether the Intermittent-Error Models estimates sensible True Values is mainly of interest in set S_{1111}.

Investigate estimated True Value in relation to measurements and in size
Therefore, a way to measure the soundness of the estimated True Values is to compare them with the measured values and the value that follows with regard to the covariate in the True Value distribution. This was carried out for the simulated data set in Section 3.2. To investigate whether one of the measurements or the True Value distribution is leading in the estimation of the True Value can indicate whether the estimation procedure is sound. Also, the size of the proposed values in relation to the measurements can be of interest, when the estimated values are considered to be used to obtain aggregated values for publishable economic statistics.

Opposing estimated True Value and measurements
A considerable deviation from the measured values can be defined as a True Value estimation that is larger than the largest measurement or smaller than the smallest measurement. An estimated True Value that opposes the measurements in this way either indicates that the intercept and slope bias are severe or that the model heavily leans on its True Value distribution. Both scenarios warrant caution whether the parameter estimates give a good description of the situation. Therefore, the amount of estimates that are outside the range of measurements is considered as soundness measure.

3.3.2.2 Error pattern likelihoods

The procedure to update the τs in the Intermittent-Error Model EM algorithm estimates the likelihood $f_{klm}(y_{SBS,i}\,; y_{VAT,i}\,; y_{PDR,i} \mid x_{Empl,i};\ \theta^{(t)})$ of the three observations to have the error patterns sets S_{011}, S_{101}, S_{110} or S_{111} (see Section 2.3.3). For some triplets in some of the iterations of the EM algorithm, this likelihood is so small that the *mvtnorm()* R function returns 0. When this happens for all four sets, no τ values can be estimated since the numerator in the expression for τ becomes zero (see Expression (2.10) on page 41). For these situations, a bypass is used by letting the *Mahalanobis distance* decide for which set the observed measurements are closest to the expected value of the distribution (Robinson, 2016, pp. 46-47).

The fact that the likelihood of the three measurements for all four error patterns is so small that it is not estimable by *mvtnorm()* indicates that the model does not fit the data very well. Therefore, the soundness of the model depends on how many records and how many iterations use this bypass.

3.4 Summary

In response to research sub question 3 *How can the Intermittent-Error Model's performance be assessed?*

Assessment of the Intermittent-Error Model fit is not straightforward. Since error patterns are assigned, and expected values calculated based on the assumptions of the model, these assumptions cannot be directly tested. Model diagnostics using the expected values and therefore error pattern assignments would paint a too positive picture about the model assumptions.

Fortunately, some records are assigned to error patterns based only on the assumption of independent errors conditional on the True Value. For these records the error pattern is certain and the True Value is known and can be used for diagnostics on other model assumptions. Fit measures are proposed that are standard for simple linear models: plots on dependent vs independent variable to assess nonlinearity, QQ-plots and empirical density plots to assess non-normality and residuals vs covariate plots to assess non-constant error variance.

These linear model diagnostics can be carried out for the source error distributions as well as conditional True Value distribution. The conditional True Value distribution plays a role in estimating the True Value for observation triplets for which all observations are perceived erroneous and in assigning error probabilities to observations for which the error pattern is unknown. Simulations showed that the number of triplets with high τ_{111} varied across data sets simulated from the same distribution but that the predictive power of the covariates did not influence how large this variation was. Also, it was shown that pure noise covariates or misleading covariates (perfect fit for records on which starting values are based, pure noise for the others) did not have a large influence on the estimated True Values for records with large τ_{111}.

To assess the soundness of the model, measures were proposed with regard to the estimated error pattern probabilities and True Values. Estimated True Values outside the range spanned by the observed values (thus larger than the largest or smaller than the smallest) were considered a sign of limited soundness. Also the occurrence of 0 error pattern likelihoods, for which a bypass was implemented in the estimation procedure, was considered a sign of bad model fit and as a result untrustworthy estimates.

4 The Intermittent-Error Model's performance on a case study data set

> *"All models are wrong but some are useful." (Box, 1979)*

The above quotation from British statistician George E. P. Box is a stronger version of a previous mentioning of the same idea by Box & Draper (1987): *"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."*

Chapter 5 investigates in general how useful the Intermittent-Error Model is for Statistics Netherlands. This chapter investigates in what way the case study data is described by the Intermittent Error Model, and what its performance is in detecting error patterns and estimating True Values. In other words: How wrong is the Intermittent-Error Model with regard to the case study data?

Section 4.1 discusses the case study data fit in general and Section 4.2 with regard to what was already known about the data set from Section 1.3. Section 4.3 discuss the model fit with regard to the fit/soundness measures developed in Section 3.3. In Section 4.1.1 the need to transform the data is discussed, for which various solutions can be proposed but only one is applied. Section 4.4 discusses the influence of the chosen transformation on the model fit, by comparison to the model fit with another transformation. Section 4.5 discusses the stability of the case study model fit and Section 4.6 concludes this chapter with a summary.

4.1 General case study model fit

The Intermittent-Error Model estimates an expected error proportion per source, and whether there are systematic deviations in turnover concept between sources, in the form of intercept and slope bias with regard to the estimated True Value. These source characteristics are expected to vary across NACE groups and therefore the Intermittent-Error Model is fitted for each NACE group separately. This section discusses the Intermittent-Error Model fit for all eight NACE groups. When discussing all groups becomes too extensive, NACE group *G45.1.1.2*, the largest NACE group, is used as a running example.

4.1.1 Input data and transformations

Covariate *Number of Employees* from GBR

The conditional True Value distribution needs to meet simple linear regression assumptions, as discussed in Section 2.2. One of these assumptions is that the measurements on the covariates are error-free. Theoretically, identification and structural variables from the GBR are most suitable to be used as covariates in the True Value distribution, since these define the population framework and do not originate from a source of measurement from which also erroneous turnover values are obtained (like the SBS survey). However, previous research has found that the identification and structural variables from the GBR do contain errors, such as misclassified NACE groups (Van Delden, Scholtus, & Burger, 2016) and *Size Class*es (De Wolf & Van Delden, 2011). Also, Section 1.3.2 showed large inconsistencies within GBR variables *Size Class* and *Number of Employees*. Unfortunately, *Number of Employees* is the best available covariate, even though it does not meet the error-free assumption.

Log Turnover values

Financial variables, such as business turnover, often show unequal spread of values and positive skew. For the turnover values from the three sources, this is shown in the left side of Figure 4.1. Unequal spread and positive skew have the same origin, namely the fact that financial values are often bounded below (cannot obtain negative values). Skewness in the dependent variable of an ordinary least squares model can result in inappropriate parameter estimates. Since the central tendency of skewed values is often not adequately described by the mean of the values, also the parameters that describe conditional means in a least squares model, do not describe the conditional central tendency of the dependent variable in an adequate way. The Intermittent-Error Model assumes normal errors in the erroneously measured values as well as in the conditional True Value distribution. Therefore, parameter estimates are obtained by minimizing squared losses (as shown in Expression (2.25) and (2.26) in Section 2.3.4). Skewness in the turnover values would therefore lead to inappropriate parameter estimates.

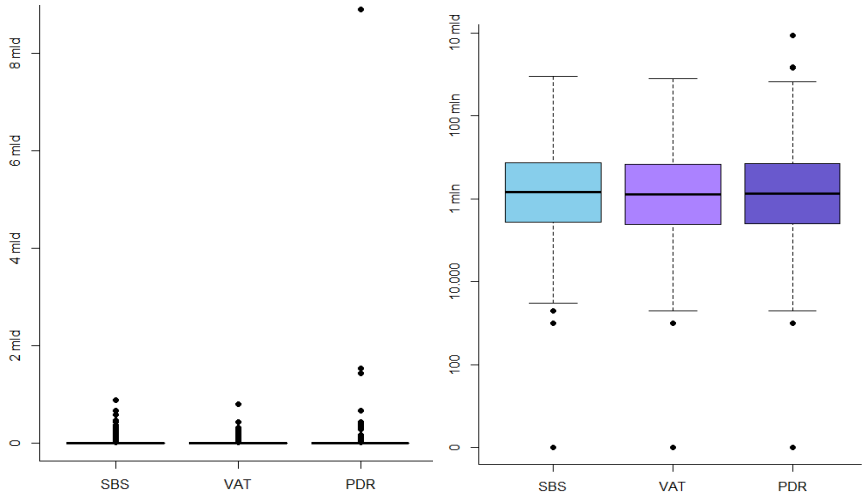


Figure 4.1 Boxplots of *Turnover* values from the three sources in euros on the original scale (left) and log transformed on a logarithmic y-axis (right).

A log transformation is often used to correct for positive skew. A logarithm compresses the large values and spreads out the small values, and therefore corrects for positive skew as well as unequal spread, as shown in Figure 4.1 for the turnover values. Thus fitting the Intermittent-Error Model on log transformed turnover values results in more appropriate parameter estimates in the squared loss components of the model, and therefore results in better estimates for the expected error proportions in each source. Since the Intermittent-Error Model is fitted for each NACE group separately, Figure 4.2 shows boxplots of log transformed turnover values for all eight NACE groups separately.

Some turnover values are 0 and the logarithm of 0 is not defined. Therefore, no direct log transformation of the turnover data can be carried out. As an alternative, all turnover values are increased with a small value (0.01) before they are log transformed. With regard to the True Value distribution, the relation between turnover and number of employees is considered linear and not log linear. Therefore, also the values on the covariate *Number of Employees* are log transformed after a small value is added (also 0.01), since also 0 employees occur in the data.

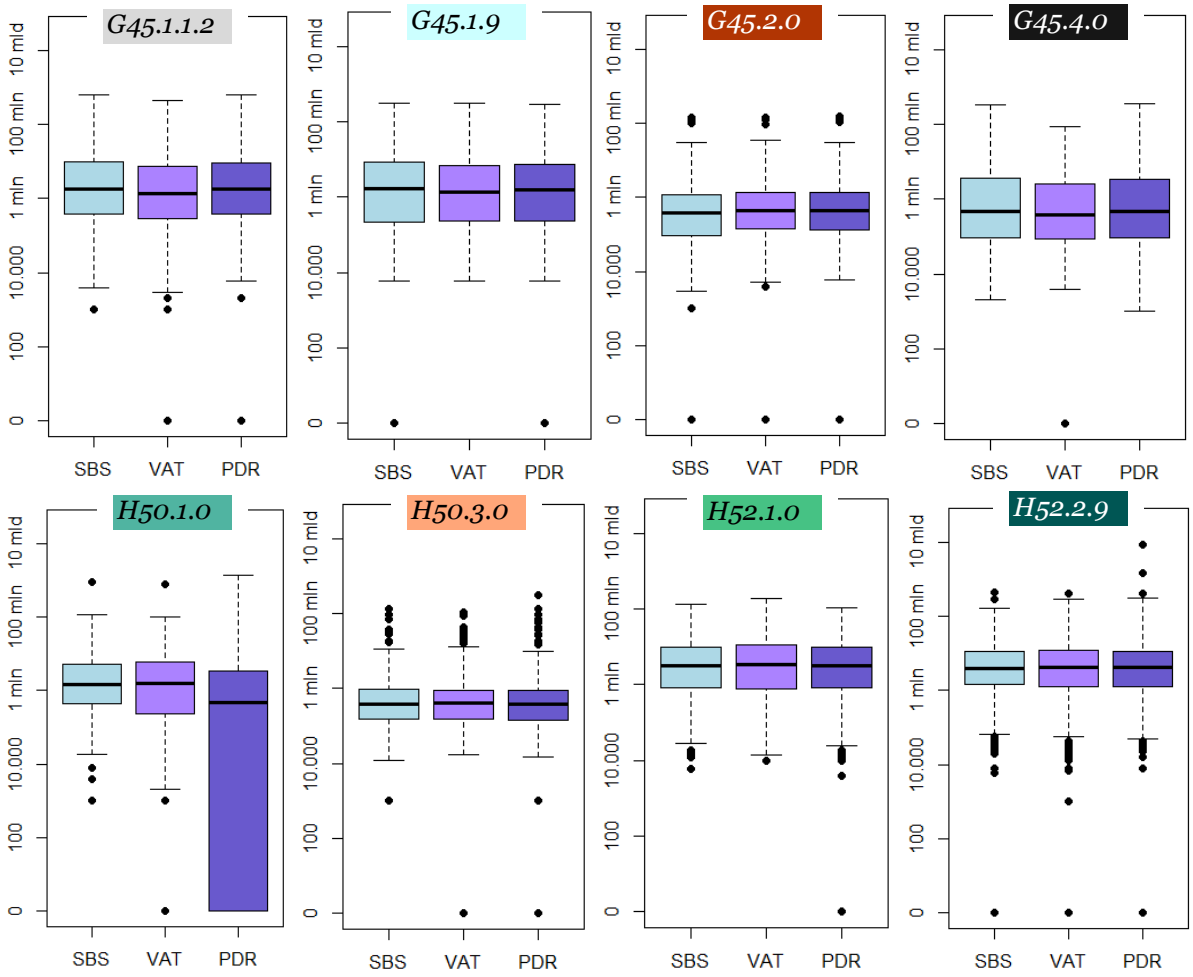


Figure 4.2 Boxplots for each NACE group of log transformed *Turnover* values from the three sources in euros on a logarithmic y-axis.

4.1.2 Outliers

Three types of outliers appear in NACE group *G45.1.1.2*, shown in Figure 4.3 (on page 64) that represent the typical outliers in the entire data set. The first type contains *zero outliers*, in which case either the SBS, VAT or PDR Turnover value is 0, while the other two are clearly not. These errors are not likely measurement errors, but have occurred in some other way in data collection and processing (as further discussed in Section 5.1). NACE group *G45.1.1.2* contains 8 outliers of this type. The second outlier type distinguished in the data is the *€1000 outlier*, for which case records have value €1000 for SBS Turnover and a much larger value for VAT and PDR Turnover. These are notable since all values in the SBS survey are reported in multiples of €1000 and thus the exact value 1 is reported for these statistical business units. These units are visible in Figure 4.3 as the three dots stacked on top of each other on the left side of the two plots. NACE group *G45.1.1.2* contains 3 outliers of this type. The third error type occurs as a *negative value* from one of the sources, of which a (non-existing) example was shown in Table 1.9 (on page 29). The occurrences of the three outlier types for the other NACE groups are shown in Table 4.1 (on page 64).

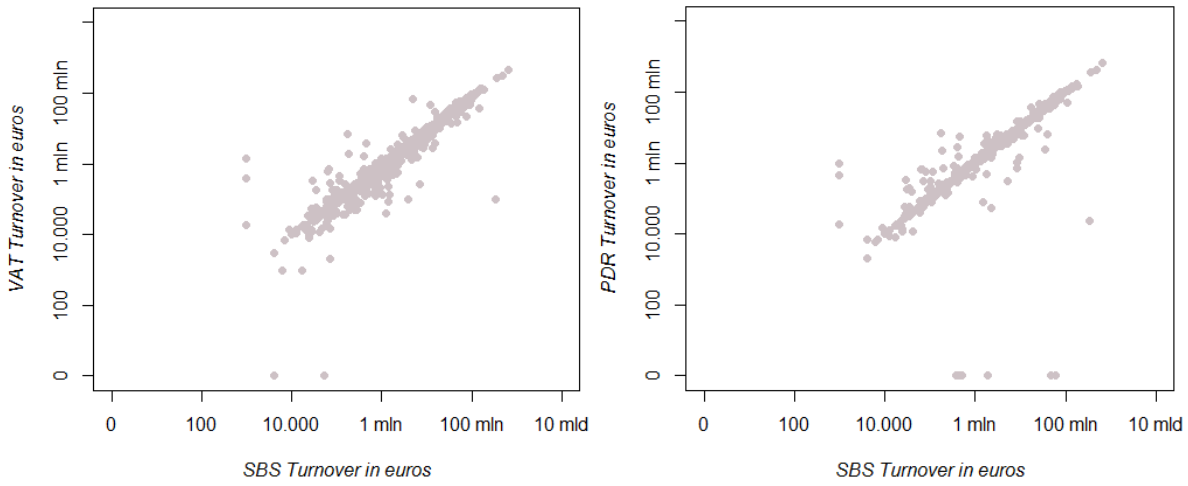


Figure 4.3 VAT and PDR *Turnover* values plotted against SBS *Turnover* values for NACE group *G45.1.1.2*. All in euros on a logarithmic x- and y-axis.

Table 4.1 Outlier occurrences per NACE group

	G45.1.1.2	*G45.1.9*	*G45.2.0*	*G45.4.0*	*H50.1.0*	*H50.3.0*	*H52.1.0*	*H52.2.9*
zero SBS outlier		1	1					
zero VAT outlier	2		1	1	3	2		5
zero PDR outlier	6	1	1		33	10	2	6
zero SBS & PDR outlier								1
zero VAT & PDR outlier					3	1		
€1000 SBS outlier	3		4					
negative VAT value						1		
Total outliers	11	2	7	1	39	14	2	12
Total triplets	885	166	247	64	128	333	148	417
Outlier proportion	1%	1%	3%	2%	30%	4%	1%	3%

Table 4.1 shows that the outlier occurrence in NACE group *H50.1.0* is very high, especially considering the total number of businesses in this NACE group. Especially, the many zero PDR outliers influence the distribution of the turnover values shown in Figure 4.2 (on page 63). Figure 4.4 (on page 65) shows a boxplots of the log transformed turnover values when these outliers are excluded.

4.1.3 Parameter estimates

When the eleven outliers are included, the estimation procedure needs 76 iterations to fit the model for the NACE group *G45.1.1.2* while excluding the outliers results in an estimation procedure of 40 iterations. (Both with convergence criteria: [1] Number of iterations maximum 200, [2] Stop if difference between parameter values in consecutive iterations $< 1*10^{-10}$.) Similar results are obtained from the other NACE groups. Table 4.2 shows the parameter estimates for the model fit with and without the outliers.

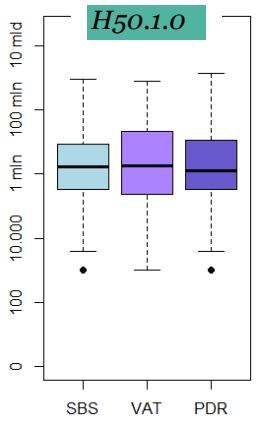


Figure 4.4 Boxplots for NACE group *H50.1.0* of *Turnover* values from the three sources in euros on a logarithmic y-axis, outliers excluded that are shown in Table 4.1 (on page 64).

Table 4.2 Intermittent-Error Model parameter estimates for all NACE groups fitted with and without outliers.

NACE groups: G45.1.1.2, G45.2.0, H50.1.0, H52.1.0, G45.1.9, G45.4.0, H50.3.0, H52.2.9

	π_{SBS}	π_{VAT}	π_{PDR}	β_0	β_{Empl}	σ^2	a_{SBS}	b_{SBS}	σ^2_{SBS}	a_{VAT}	b_{VAT}	σ^2_{VAT}	a_{PDR}	b_{PDR}	σ^2_{PDR}
Without outliers	0.12	0.91	0.64	11.97	1.24	0.98	0.89	0.93	1.28	-0.28	1.00	0.25	0.13	0.99	0.01
With outliers	0.11	0.90	0.65	11.89	1.21	3.63	11.96	0.07	6.25	5.13	0.63	2.60	0.11	0.99	0.01
Without outliers	0.57	0.82	0.22	11.82	1.21	1.30	-0.09	1.01	0.00	1.03	0.92	0.35	2.57	0.83	0.60
With outliers	0.60	0.83	0.17	11.81	1.15	3.70	-0.14	1.01	0.00	6.49	0.56	1.87	9.17	0.28	14.11
Without outliers	0.22	0.74	0.51	10.93	1.28	0.52	2.01	0.80	1.07	-0.10	1.00	0.11	-0.08	1.01	0.00
With outliers	0.22	0.70	0.54	10.90	1.25	1.89	8.83	0.20	8.46	3.98	0.70	2.81	-0.07	1.01	0.00
Without outliers	0.11	0.87	0.67	10.79	1.74	0.96	-0.99	1.04	0.92	0.67	0.93	0.25	0.14	0.99	0.00
With outliers	0.13	0.87	0.67	11.10	1.54	1.12	0.44	0.93	0.93	-3.16	1.20	2.26	0.16	0.99	0.00
Without outliers	0.66	0.82	0.22	12.31	1.10	2.66	0.20	0.98	0.02	1.90	0.87	2.18	7.64	0.45	4.71
With outliers	0.69	0.83	0.27	12.20	0.99	10.94	11.25	0.25	3.59	-3.12	1.19	7.82	4.62	-0.57	17.38
Without outliers	0.22	0.78	0.39	11.40	1.25	0.66	1.42	0.88	1.23	0.47	0.97	0.21	0.15	0.99	0.00
With outliers	0.69	0.83	0.27	12.20	0.99	10.94	11.25	0.25	3.59	-3.12	1.19	7.82	7.82	-0.57	17.38
Without outliers	0.40	0.85	0.62	11.59	1.22	1.16	2.10	0.84	1.07	2.20	0.86	0.61	0.11	0.99	0.00
With outliers	0.73	0.94	0.10	11.68	1.21	0.88	-0.20	1.00	0.27	0.72	0.95	0.33	-2.66	0.96	36.15
Without outliers	0.64	0.89	0.14	12.82	0.90	1.57	0.00	1.00	0.10	1.20	0.92	1.04	4.79	0.69	3.45
With outliers	0.15	0.88	0.64	12.42	0.95	6.88	13.20	0.13	4.79	9.65	0.35	8.65	-0.03	1.00	0.08

Sensitivity towards outliers

The parameter estimates with inclusion and exclusion of the outliers show that the model is quite sensitive towards outliers. This lack of robustness towards outliers is characteristic to models that assume normal errors, as the Intermittent-Error Model does. In many NACE groups the source error variances (σ^2_{SBS}, σ^2_{VAT} and σ^2_{PDR}) are inflated when the outliers are included. With regard to the bottom three NACE groups in Table 4.2, the estimates for expected error proportions show a large shift when the outliers are excluded. This is curious since these NACE groups do not necessarily have a high proportion of outliers. Especially noteworthy is NACE group *H52.1.0* which only has 2 outliers on a total of 148 records, and still shows a large shift in the expected error proportions for SBS and PDR. This is a first sign of the Intermittent-Error Model's instability with regard to the case study data set, which is further investigated in Section 4.5.

Zero source error variances

The zero source error variances that occur in Table 4.2 are only zero because values are rounded to two digits. Therefore, these values are actually not zero, but they are very small. These zero variances occur for σ^2_{SBS} or σ^2_{PDR} when π_{SBS} is larger than π_{PDR} and π_{PDR} is larger than π_{SBS} respectively and therefore many more records for which y_{SBS} and y_{PDR} are very close contribute to the variance (Section 5.2 discusses these records in more detail). The likelihood function of normal finite mixture models is known to have local maxima at the edge of the parameter space which can be regarded as spurious solutions (McLachlan & Peel, 2000, p. 99). This is a second sign of the Intermittent-Error Model's instability with regard to the case study data set, further investigated in Section 4.5.

Interaction between error probabilities π_g and error size σ^2_g

In all NACE groups the probability to be measured erroneously is highest for the VAT turnover. Whether the SBS values contain errors more often or the PDR values do, varies among NACE groups. Except for NACE group *H52.1.0*, the errors are considered more severe (larger source error variance) when the expected error proportion in the SBS is smaller, and the same holds for the PDR. This is shown by the fact that for NACE groups in which π_{SBS} is smaller than π_{PDR}, σ^2_{SBS} is larger than σ^2_{PDR}, and the other way around.

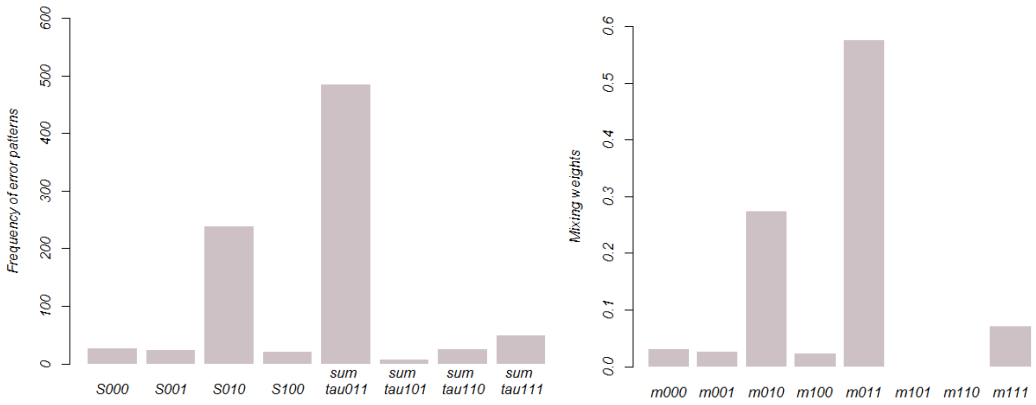


Figure 4.5 Set assignments and mixing weights for NACE group *G45.1.1.2* model fit without outliers.
Left: Number of triplets assigned to sets S_{000}, S_{001}, S_{010} and S_{100} and for triplets assigned to set $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$, the sum of the probabilities $\tau_{klm,i} = \mathbb{P}\left((z_{SBS,i}, z_{VAT,i}, z_{PDR,i}) = (k,l,m) \,\middle|\, y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i}; x_{Empl,i};\ i \in S_{011} \cup S_{101} \cup S_{011} \cup S_{111}; \boldsymbol{\theta}\right)$
Right: Mixing weights when triplets in set $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$ are assigned to the set with the largest τ.

Assigned error patterns and mixing weights

With regard to NACE group *G45.1.1.2* the estimated π_{VAT} value shown in Table 4.2 indicates that the number of errors in the VAT Turnover values is estimated to be large. Figure 4.5 shows the occurrence of the observed and expected error patterns. Error-free VAT Turnover values mainly occur with an

error-free SBS value, error-free PDR value or both, but they are very unlikely to occur as the only error-free observation.

Figure 4.6 shows the observed and estimated error patterns for all NACE groups fitted without outliers. Table 4.3 (on page 68) specifies the number of records in the identified sets S_{000}, S_{001}, S_{010} and S_{100}, in comparison to the unidentified set $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$.

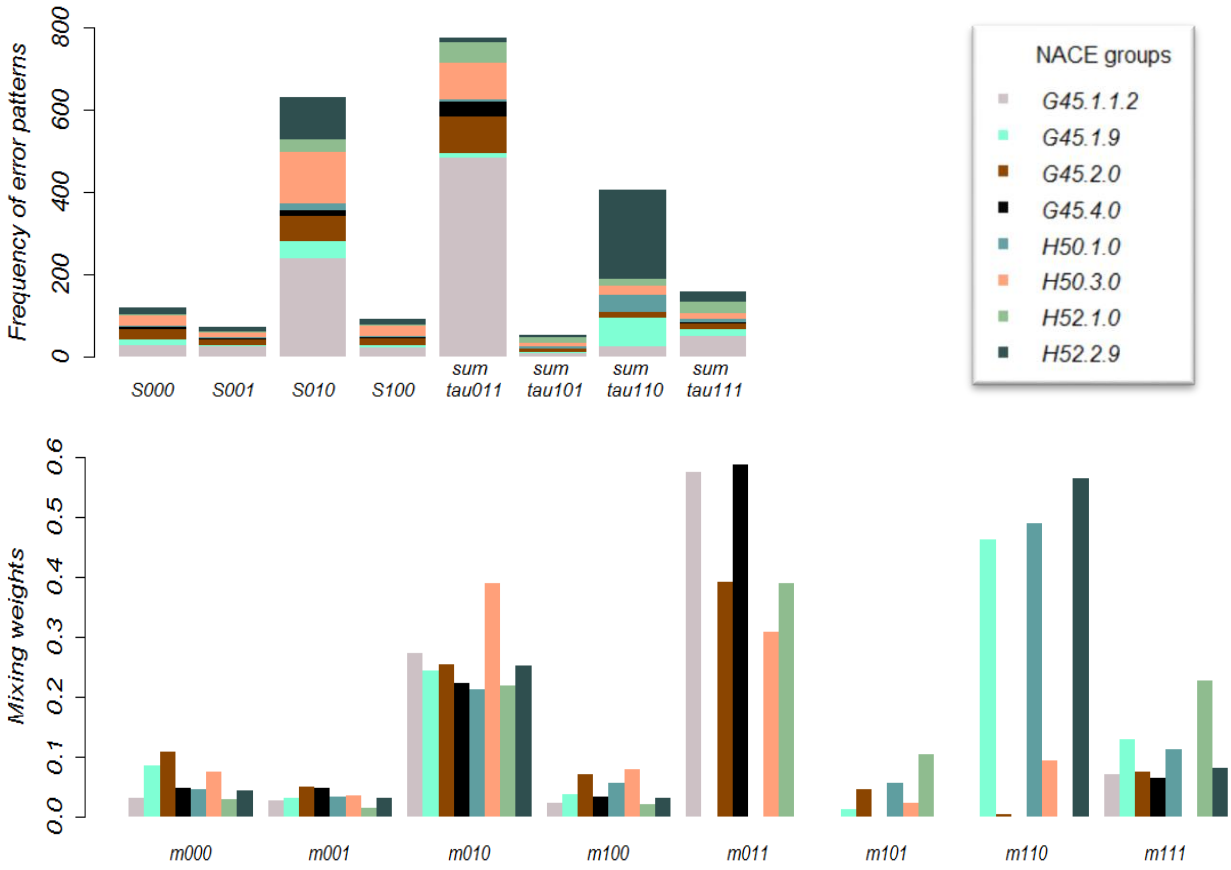


Figure 4.6 Set assignments and mixing weights for all NACE groups, model fit without outliers.
Upper: Number of triplets in each NACE group assigned to sets S_{000}, S_{001}, S_{010} and S_{100} and for triplets assigned to set $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$, the sum of the probabilities $\tau_{klm,i} = \mathbb{P}\left((z_{SBS,i}, z_{VAT,i}, z_{PDR,i}) = (k,l,m) \,\middle|\, y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i}; x_{Empl,i};\ i \in S_{011} \cup S_{101} \cup S_{011} \cup S_{111}; \boldsymbol{\theta}\right)$
Lower: Mixing weights when triplets in set $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$ are assigned to the set with the largest τ.

The lower part of Figure 4.6 shows that with regard to the records that are assigned to S_{011} and S_{110} because τ_{011} and τ_{110} are larger than the other τs, in some NACE groups none contribute to m_{011} and many to m_{110} and the other way around. This is in contrast to the mixing weights that are determined completely by the known True Values (m_{000}, m_{001}, m_{010} *and* m_{100}), which show contributions of records from all NACE groups to each mixing weight. One possible explanation for this characteristic is that the assignment to the group with the largest τ is an incorrect approach because for these records τ_{011} and τ_{110} are very close together. However, the upper part of Figure 4.6 does not support that hypothesis since for example, the missing *G45.1.1.2* bar for m_{110} in the lower part of Figure 4.6 is hardly contributing to the sum of τ_{110} in the upper part of Figure 4.6, and the same holds for the missing *H52.2.9* bar for m_{011} and the sum of τ_{011}. Therefore, τ_{011} and τ_{110} are not likely very close together for these records, and the shift from large contributions to either m_{011} or m_{110} in these NACE groups is the estimated Intermittent-Error Model's result. Section 5.2 discusses these records in more detail.

Table 4.3 Error pattern set sizes by NACE group

	G45.1.1.2	*G45.1.9*	*G45.2.0*	*G45.4.0*	*H50.1.0*	*H50.3.0*	*H52.1.0*	*H52.2.9*
S_{000}	27	14	26	3	4	24	4	17
S_{001}	23	5	12	3	3	11	2	12
S_{010}	239	40	61	14	19	124	32	102
S_{100}	23	6	17	2	5	25	3	12
$S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$	565	99	124	41	59	135	105	262
Total	874	164	240	63	90	319	146	405
Percentage $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$	64%	60%	52%	65%	65%	42%	72%	65%

4.2 Case study model fit in relation to apparent errors

Do the NACE group parameter estimates agree with what was already known about the case study data set from other available variables discussed in Section 1.3.2? Only the fact that the VAT turnover values show reporting differences (a_{VAT} deviating from zero for all NACE groups, and b_{VAT} deviating from one for some) is consistent with what was already known about these NACE groups' turnover values. Namely that all NACE groups subject to reporting differences, which was concluded previously either from expert's opinion on the tax reporting procedure or from previous research into reported values (see Section 1.3.1). However, SBS values seem to show just as much intercept (a_{SBS}) and slope bias (b_{SBS}) (with NACE group *H52.2.9* as an exception), while no such bias would be expected from Statistics Netherlands perspective on the True Value. In some NACE groups the intercept bias is extremely severe, which does not seem to be a good fit, although it comes with a very small slope estimate. NACE group *H50.1.0* shows an intercept bias of 7.64, which is a multiplication with 2080 on the original scale. Such large reporting differences for all observations in a NACE group seem very unlikely.

Table 1.6 (on page 24) showed that the PDR estimate of the total turnover was much larger than the VAT estimate, even though PDR values were available for a smaller part of the GBR population. However, no overestimation of the erroneous PDR turnover values (a_{PDR} larger than zero and b_{PDR} larger than one) is shown by the estimated model parameters. Namely, large PDR intercept bias (a_{PDR}) is compensated by a negative slope bias (b_{PDR} smaller than 1). Some underestimation of erroneous VAT turnover is shown by negative slope bias (b_{VAT} smaller than one), but these values are not considerably smaller than one in comparison to occurrences of SBS estimates (b_{SBS}) across NACE groups.

With regard to differences between NACE groups, the many businesses in NACE group *H50.1.0* with VAT *Response Percentages* smaller than 100% were considered remarkable in Section 1.3.2. This characteristic of untrustworthy VAT *Turnover* values might be related to the fact that NACE group has the largest estimated VAT source error variance (σ^2_{VAT}) when outliers are not included. However, the VAT expected error proportion (π_{VAT}) is not notably higher for this NACE group than for others. Also, the fact that NACE group *H45.4.0* contains very few TopX businesses, and might therefore have less well checked SBS turnover values is not visible in the parameter estimates. Neither is the fact that NACE group *G45.1.1.2* stands out because it contains so many *Size Class* 1 and 2 businesses, which are generally not manually scrutinized (it's π_{SBS} estimate is smallest of all NACE groups).

So, to conclude, known information about the sources from the selection process and descriptive statistics on the case study data set was not consistently recognized in the Intermittent-Error Model's parameter estimates. Also known information on the way turnover values from the Tax and Customs Administration were combined for statistical business units, and information on the SBS editing process is not apparent from the model fit.

4.3 Case study model fit with regard to fit/soundness measures

This section studies the model fit with regard to the fit and soundness measures proposed in Section 3.3. The model fit and soundness measures assess the Intermittent-Error Model fitted on the NACE group data with the outliers from Table 4.1 (on page 64) excluded.

4.3.1 Model fit with regard to the conditional 'True Value' distribution

Nonlinearity

Figure 4.7 shows the known True Values of records belonging to sets S_{000}, S_{001}, S_{010} and S_{100} plotted against their value for covariate *Number of Employees*. The fitted regression lines are shown, with β_0 and β_{Empl} estimates for each NACE group as indicated by Table 4.2 (on page 65). The deviations from the linear regression line seem larger for smaller values on *Number of Employees*. But since σ^2 (the variance of η unexplained by covariate x_{Empl}) may vary across NACE groups, this presumption needs to be investigated for each NACE group individually.

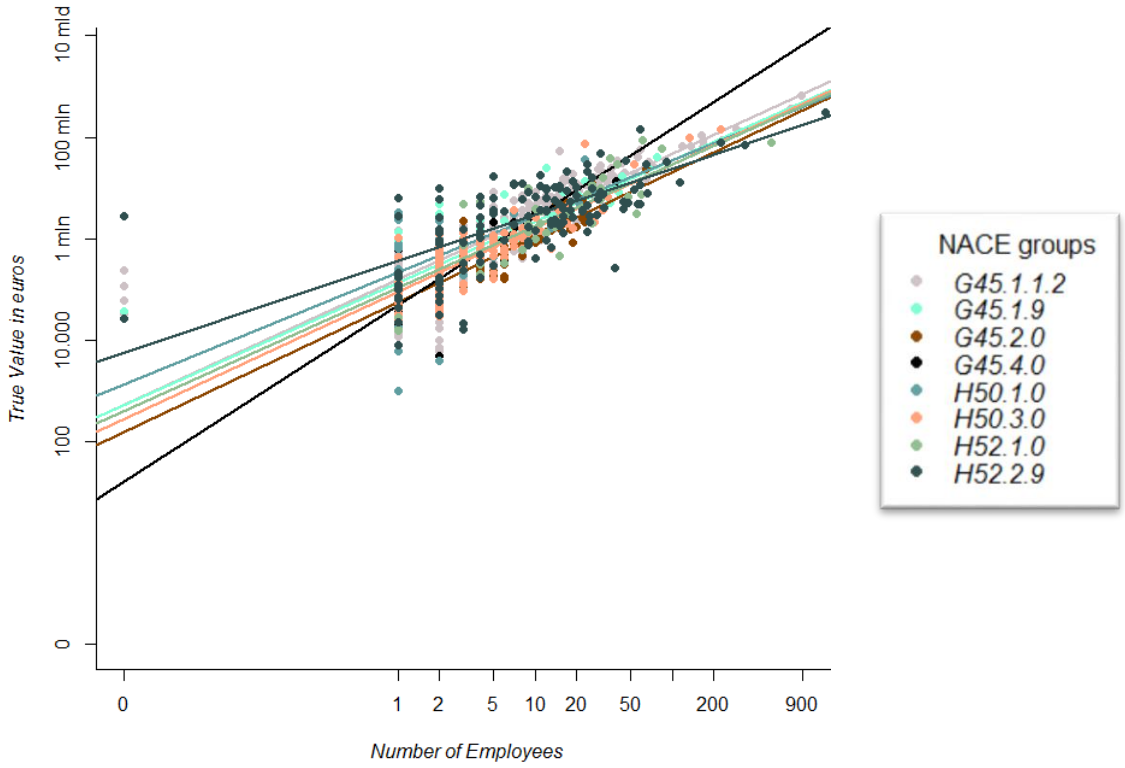


Figure 4.7 True Value plotted against covariate *Number of Employees* for observations in sets S_{000}, S_{001}, S_{010} and S_{100}, on a logarithmic x- and y-axis.
Fitted regression lines are shown, with β_0 and β_{Empl} estimates for each NACE group as shown in Table 4.2 (on page 65).

Figure 4.8 shows the plot of Figure 4.7 for only the NACE group *G45.1.1.2*. The relation between the covariate *Number of Employees* and the True Value seems linear except for the businesses with 0 employees. (These businesses might have an incorrect number of employees in the GBR, as discussed in Section 1.3.3.)

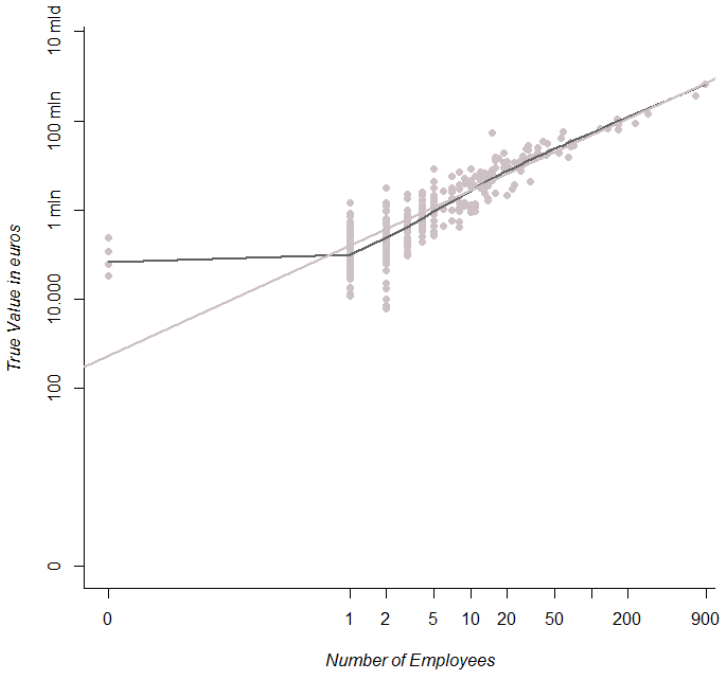


Figure 4.8 NACE group *G45.1.1.2* True Values plotted against covariate *Number of Employees* for observations in sets S_{000}, S_{001}, S_{010} and S_{100} (for which the True Value is known), on a logarithmic x- and y-axis. *Fitted regression line is shown in the NACE group's light grey color, with β_0 and β_{Empl} estimates as shown in Table 4.2 (on page 65). A lowess curve is shown in dark grey, fitted with $\delta = 0.34$ (0.01*diff(range(x))).*

Figure 4.8 also gives some idea of the residuals, showing that the log transformation overcorrects the non-constant spread available in the turnover values on the original scale. The log transformed turnover values show larger variance among small businesses (with small number of employees) in comparison to larger ones. This non-constant error variance is further investigated under 'Non-constant error variance' below.

Non-normality

Figure 4.9 shows a Normal QQ-plot and an empirical density plot on the *G45.1.1.2* residuals. Heavy tails are visible in both plots, which is an indication of non-normal error. The red dotted lines represent 95% confidence envelopes around the theoretical normal quantiles, which exclude the tails of the empirical distribution on both sides.

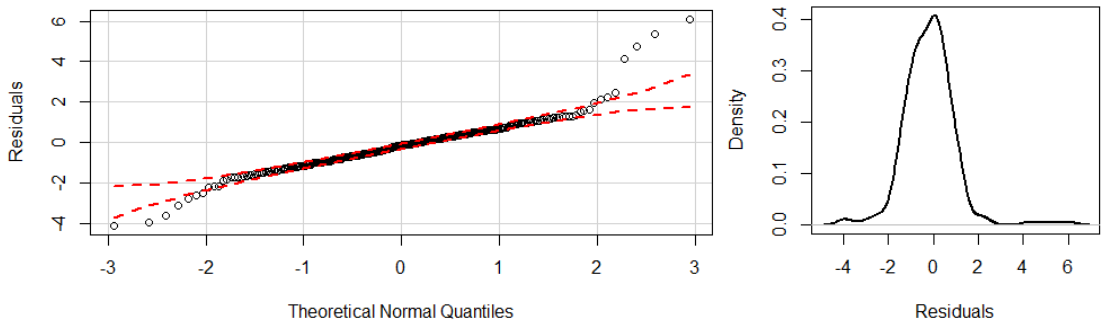


Figure 4.9 NACE group *G45.1.1.2* residuals from log transformed True Values modeled against covariate *Number of Employees* with parameter estimates β_0 and β_{Empl} as shown in Table 4.2 (on page 65). Residuals from observations in sets S_{000}, S_{001}, S_{010} and S_{100} (for which the True Value is known).
Left: QQ-plot with 95% confidence envelopes shown by the red dotted lines.
Right: Empirical density function.

Non-constant error variance

Figure 4.10 investigates if the error variance is constant across covariate and fitted values by plotting the residuals against the covariate values. As was already presumed from Figure 4.8 all largest residuals occur for small fitted values and all smallest residuals for large fitted values.

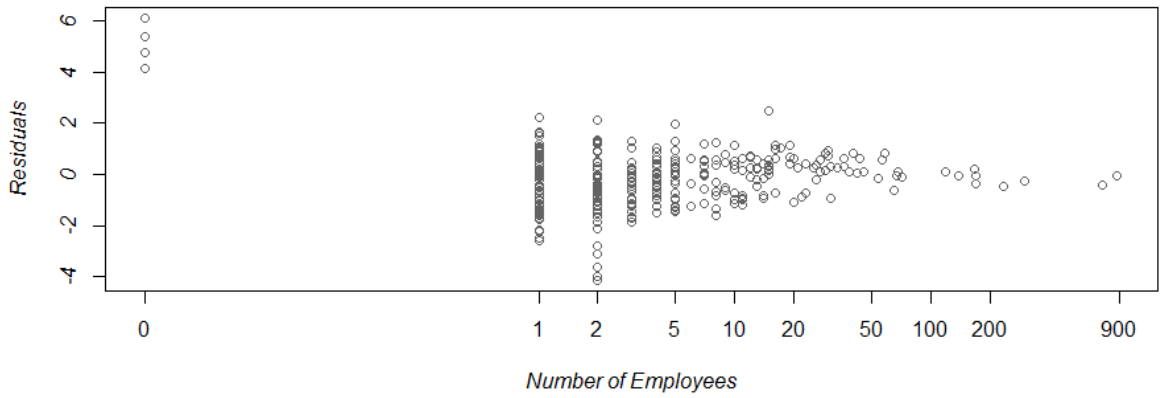


Figure 4.10 NACE group *G45.1.1.2* residuals plotted against covariate *Number of Employees*. Residuals from log transformed True Values modeled against covariate *Number of* Employees with parameter estimates β_0 and β_{Empl} as shown in Table 4.2 (on page 65), from observations in sets S_{000}, S_{001}, S_{010} and S_{100} (for which the True Value is known).

4.3.2 Model fit with regard to the source error distributions

The same fit measures that were applied in Section 4.3.1 can be applied to the known erroneous measurement and True Value in sets S_{001}, S_{010} and S_{100}. Figure 4.11 shows the values in these sets for all NACE groups. In many NACE groups, the number of records in these sets is small, in particular those in set S_{001} shown in the right plots. For some NACE groups, the parameter estimates a_{PDR} and b_{PDR} do not relate well to the observations in set S_{001}, which can be partly explained by the relatively few observations in this set (see Table 4.3 on page 68).

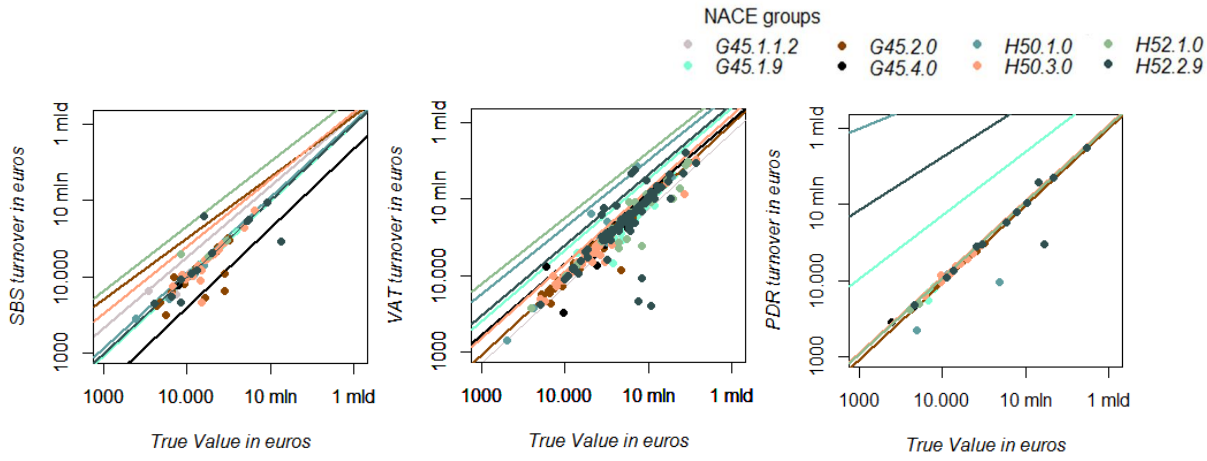


Figure 4.11 Log transformed erroneously measured SBS, VAT and PDR *Turnover* values plotted against known log transformed True Values, for observations in sets S_{100} (left), S_{010} (middle) and S_{001} (right), on a logarithmic x- and y-axis.
Fitted regression lines are shown, with a_{SBS}, b_{SBS} (left), a_{VAT}, b_{VAT} (middle) and a_{PDR}, b_{PDR} (right) estimates for each NACE group as shown in Table 4.2 (on page 65).

Nonlinearity and non-constant error variance

Figure 4.12 shows the erroneous measurements in sets S_{100}, S_{010} and S_{001} plotted against the known True Values for NACE group *G45.1.1.2*. As shown only very few observations are in set S_{100} and S_{001} which makes it difficult to assess the model assumptions. Since NACE group *G45.1.1.2* is the largest NACE group, this problem is even more severe in other NACE groups, for which the number of items in the sets S_{100}, S_{010} and S_{001} were shown in Table 4.3 (on page 68). Figure 4.12 shows no problems with the linearity assumption and constant error variance for NACE group *G45.1.1.2*. But these assumptions are difficult to judge with regard to NACE groups with less observations, especially for the PDR Turnover values plotted in the right most window in Figure 4.11 (on page 71).

Non-normality

Figure 4.13 and Figure 4.14 (on page 73) show QQ-plots and Density plots for the source error residuals from the three sources for observations in sets S_{100}, S_{010} and S_{001}. Especially set S_{100} and S_{001} have very few observations, and therefore these plots need careful conclusions. The VAT source error residuals show many residuals close to the expected value of 0, which create the sharp peak in Figure 4.14. The larger residuals are very large, as shown by the heavy tails in Figure 4.13. Therefore, with regard to the VAT source errors there are reasons to assume that the normality assumption is not met. From this follows the presumption that the same holds for the SBS and PDR source errors.

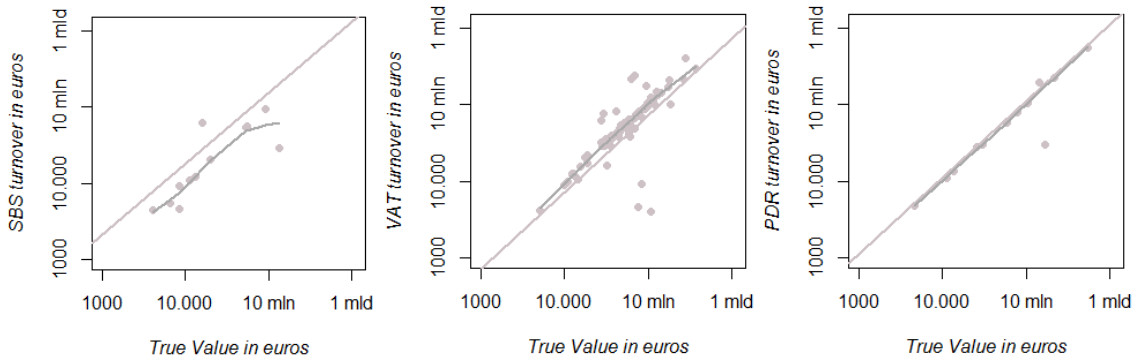


Figure 4.12 NACE group *G45.1.1.2* erroneously measured log transformed SBS, VAT and PDR *Turnover* values plotted against known log transformed True Values, for observations in sets in sets S_{100} (left), S_{010} (middle) and S_{001} (right), on a logarithmic x- and y-axis.
*Fitted regression lines are shown in the NACE group's light grey color, with a_{SBS}, b_{SBS} (left), a_{VAT}, b_{VAT} (middle) and a_{PDR}, b_{PDR} (right) as shown in Table 4.2 (on page 65). Lowess curves are shown in dark grey with δ = 0.01*diff(range(x)).*

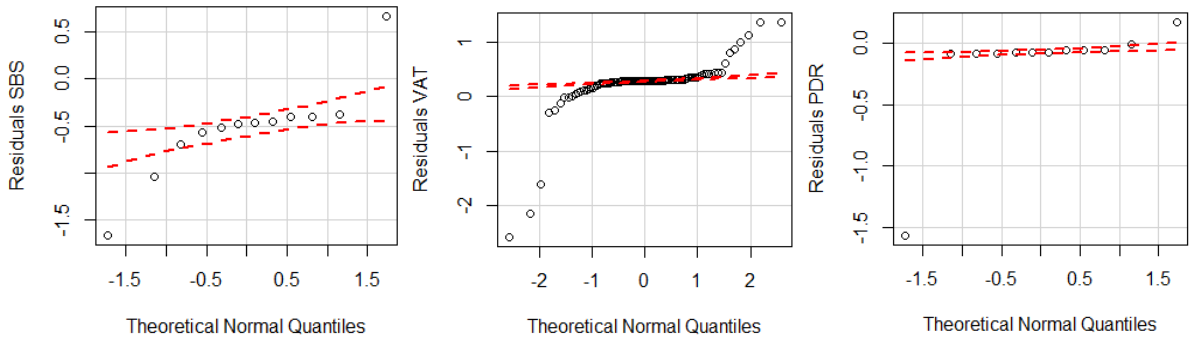


Figure 4.13 NACE group *G45.1.1.2* residuals QQ-plot. Residuals from erroneously measured log transformed *Turnover* values modeled against True Values with parameter estimates a_{SBS}, b_{SBS} *(left),* a_{VAT}, b_{VAT} *(middle) and* a_{PDR}, b_{PDR} *(right)* as shown in Table 4.2 (on page 65). Residuals from observations in sets in sets S_{100} (left), S_{010} (middle) and S_{001} (right). *95% confidence envelopes are shown by the red dotted lines.*

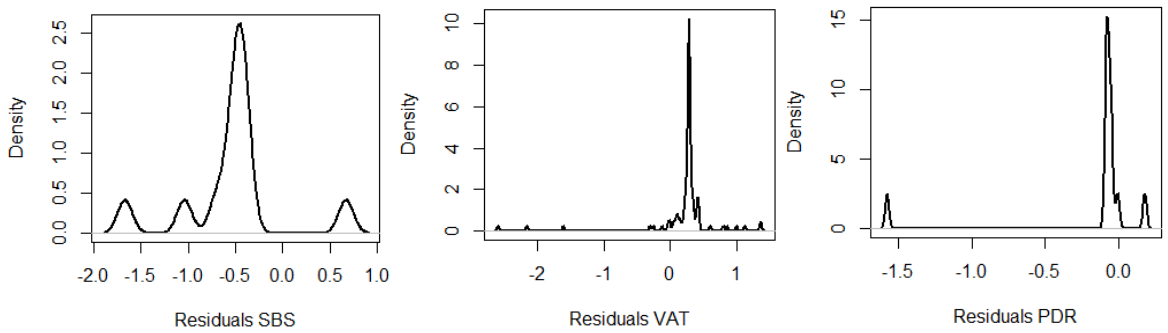


Figure 4.14 NACE group *G45.1.1.2* residuals empirical density plots. Residuals from log transformed erroneously measured values modeled against True Values with parameter estimates a_{SBS}, b_{SBS} *(left),* a_{VAT}, b_{VAT} *(middle) and* a_{PDR}, b_{PDR} *(right)* as shown in Table 4.2 (on page 65). Residuals from observations in sets in sets S_{100} (left), S_{010} (middle) and S_{001} (right).

4.3.3 Model soundness with regard to the estimated True Values

Figure 4.15 shows the 62 business units in NACE group *G45.1.1.2* for which τ_{111} was large, by plotting their measurements on the three sources, the estimated True Value for the observations and the conditional expectation w.r.t. the covariate *Number of Employees*. These are values for which τ_{111} was larger than the other τs, thus for which the Intermittent-Error Model assumes that all three sources have measured the turnover value erroneously. It is shown that these estimated True Values are close to the measured values from the PDR, which has the smallest error variance of the three sources (see Table 4.2 on page 65). These estimated True Values seem quite reasonable, since the PDR value is often somewhere in between the measurements from the SBS and VAT.

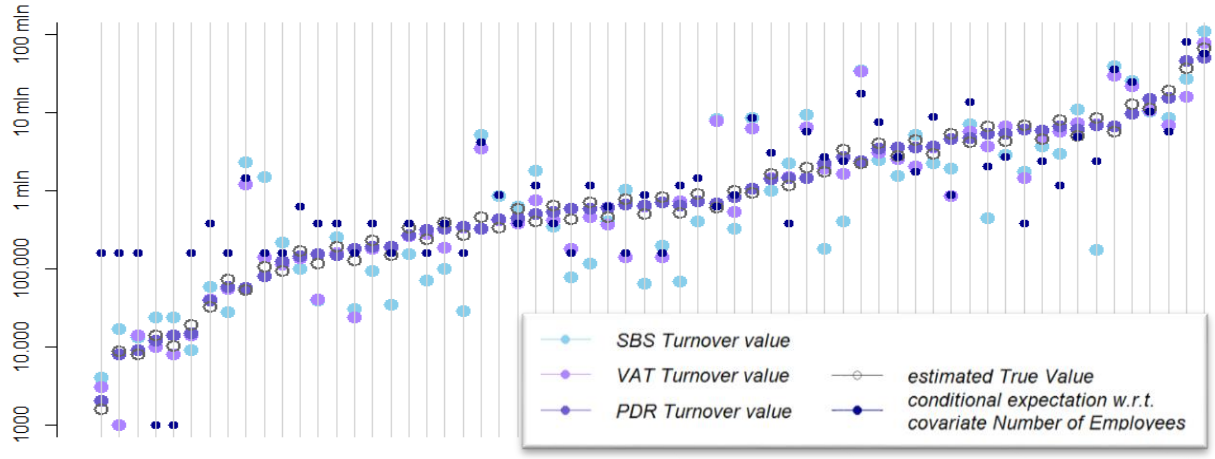


Figure 4.15 NACE group *G45.1.1.2* measured *Turnover* values, estimated True Values and *Conditional expectations w.r.t. covariate Number of Employees* in euros for of triplets for which the estimated τ_{111} is higher than τ_{011}, τ_{101} and τ_{110}, on logarithmic y-axis.
For definitions of estimated True Value and conditional expectation w.r.t. covariate Number of Employees, see Expression (3.1) and (3.2) (on page 52) respectively.
The True Value estimates are scattered to make them visible for most units. The vertical lines facilitate observing the five values on the same statistical business unit. The horizontal axis contains an index assigned to each triplet with regard to the sorted True Value estimates.

Influence Conditional expectation w.r.t. covariate *Number of Employees*

Figure 4.15 shows that the conditional expectation (blue dots) do not influence the estimated True Values very much. The figure shows that the conditional expectation is often a very crude estimate, outside of the interval spanned by the measurements from the three sources.

Opposing estimated True Value and measurements

In NACE group *G45.1.1.2* it occurs two times that the estimated True Value is larger than any of the three measured turnovers values. In both cases τ_{011} is much larger (~0.98) than τ_{101}, τ_{110} and τ_{111}, so these records are not shown in Figure 4.15. For these records the SBS *Turnover* value is the largest and the True Value is probably estimated to be even larger due to the estimated intercept and slope bias in the SBS measurements (a_{SBS} and b_{SBS}). This does not necessarily point at an unsound model fit.

4.4 The influence of transformations

As discussed in Section 4.3.1 with regard to NACE group *G45.1.1.2*, the log transformation might over-correct the larger variances among large turnover values and smaller variances among smaller turnover values. Similar effects are shown in the other NACE groups as depicted in Figure 4.7 (on page 69). Therefore, other transformations were considered.

4.4.1 Cube root transformation

A transformation that has similar effect as the log transformation is the cube root transformation. Just like the logarithm, the cube root compresses the large values and spreads out the small values, and therefore corrects for positive skew as well as unequal spread present in the turnover values on the original scale. In contrast to the log transformation, the cube root transformation is also capable to directly transform zero, and even negative values (of which one occurs among the outliers, as shown in Table 4.1 (on page 64)). Figure 4.16 shows *Turnover* values from the three sources for both transformations. These boxplots show that the cube root transformation might not correct the skew enough, since still a positive skew is visible. As a result, many large turnover values are considered outliers among the cube root transformed turnover values.

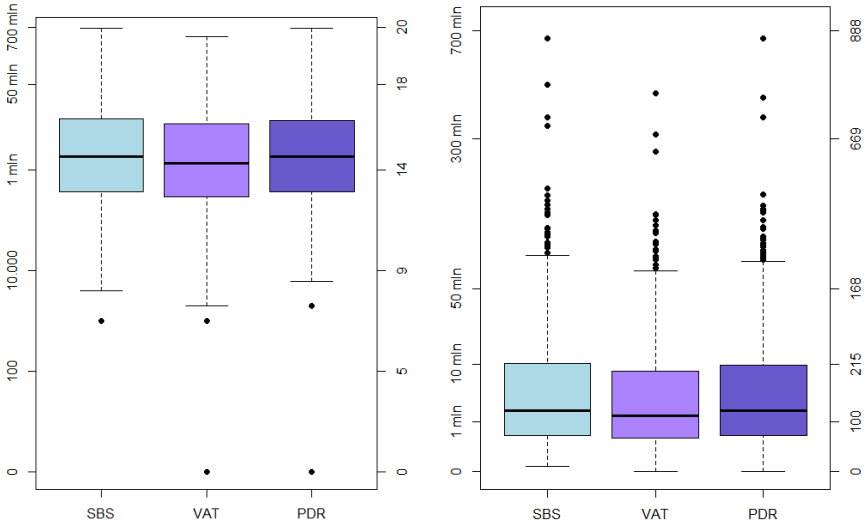


Figure 4.16 NACE group *G45.1.1.2* boxplots of *Turnover* values from the three sources in euros, log transformed (left) and cube root transformed (right).
The left y-axis is on the original scale of the turnover values, the right one on the transformed scale.

4.4.2 General case study model fit comparison

Table 4.4 shows the parameter estimates for the Intermittent-Error Model fit on NACE group *G45.1.1.2* for the log transformation as well as the cube root transformation. Since the data is on a different scale, also all simple linear model parameters are on a different scale and therefore not comparable. However, the πs can be directly compared and show a large shift. This shift is further investigated in Figure 4.17 (on page 76). This figure shows that the shift from low to high for π_{SBS} and high to low for π_{PDR}, is a shift from records with highest τ_{011} to highest τ_{110} or τ_{111}.

Table 4.4 Intermittent-Error Model parameter estimates for NACE group *G45.1.1.2* without the outliers described in Section 4.1.2, fitted with log and cube root transformation.

	π_{SBS}	π_{VAT}	π_{PDR}	β_0	β_{Empl}	σ^2	a_{SBS}	b_{SBS}	σ^2_{SBS}	a_{VAT}	b_{VAT}	σ^2_{VAT}	a_{PDR}	b_{PDR}	σ^2_{PDR}
Log	0.12	0.91	0.64	11.97	1.24	0.98	0.89	0.93	1.28	-0.28	1.00	0.25	0.13	0.99	0.01
Cube root	0.59	0.91	0.22	-40.09	89.37	1359.01	-0.09	1,00	5.61	1.98	0.90	423.91	12.35	0.94	1154.61

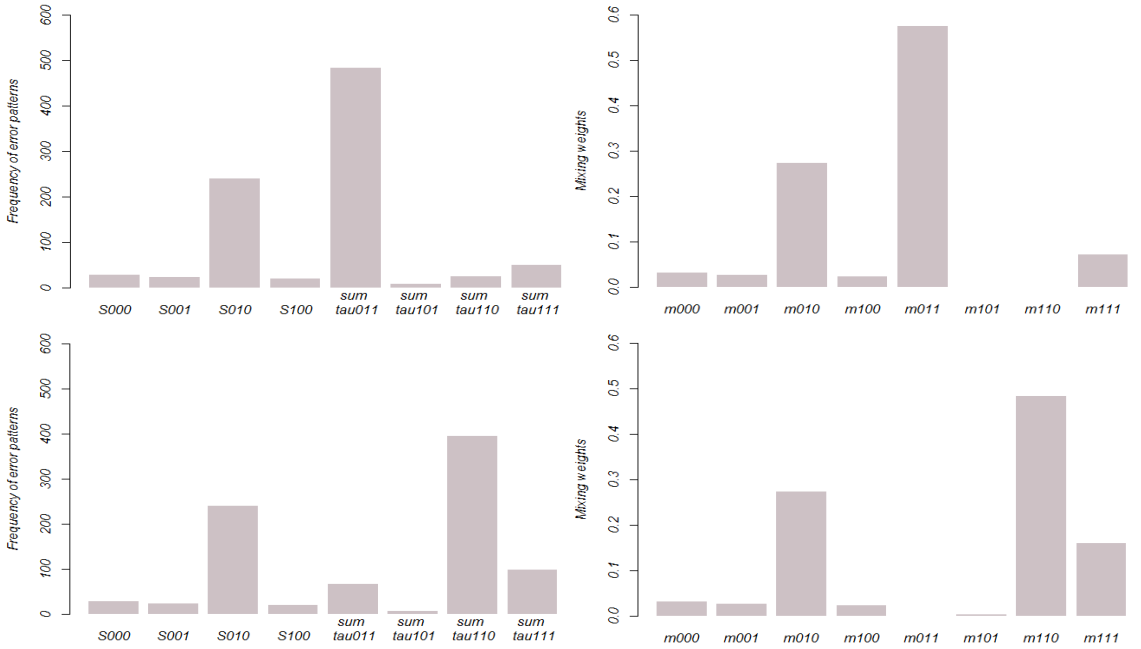


Figure 4.17 Left: Number of triplets in NACE group *G45.1.1.2* assigned to sets S_{000}, S_{001}, S_{010} and S_{100} and for triplets assigned to set $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$, the sum of the probabilities $\tau_{klm,i} = \mathbb{P}\left((z_{SBS,i}, z_{VAT,i}, z_{PDR,i}) = (k,l,m) \,\middle|\, y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i};\ x_{Empl,i};\ i \in S_{011} \cup S_{101} \cup S_{011} \cup S_{111}; \boldsymbol{\theta}\right)$
Right: Mixing weights when triplets in set $S_{011} \cup S_{101} \cup S_{110} \cup S_{111}$ are assigned to the set with the largest τ.
Upper: log transformation, Lower: cube root transformation

4.4.3 Case study model fit comparison with regard to fit/soundness measures

4.4.3.1 Model fit comparison with regard to the conditional 'True-Value' distribution

With regard to the observations with known error patterns in sets S_{000}, S_{001}, S_{010} and S_{100}, the True Value distribution fit can be assessed since the True Value is known and not estimated using this distribution. Since the observations in these sets are the same regardless of the transformation, the influence of the transformation on the True Value model fit can be directly compared. The model fit is compared for NACE group *G45.1.1.2*.

Nonlinearity

Figure 4.19 shows the True Values plotted against the covariate Number of Employees for the observations for which the True Value is known. Both show that the observations with 0 number of employees are problematic for the linear relationship.

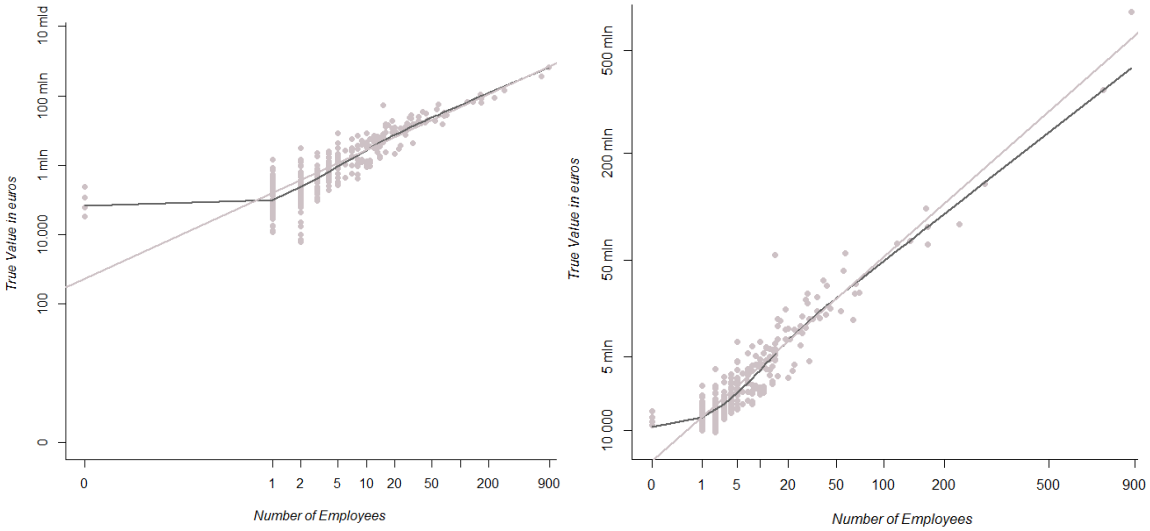


Figure 4.18 NACE group *G45.1.1.2* True Values plotted against covariate *Number of Employees* for observations in sets S_{000}, S_{001}, S_{010} and S_{100} (for which the True Value is known).
Left: logarithmic x- and y-axis, Right: cube root x- and y-axis
*Fitted regression lines are shown in the NACE group's light grey color, with β_0 and β_{Empl} estimates as shown in Table 4.4 (on page 76). Lowess curves are shown in dark grey, fitted with δ = 0.01*diff(range(x)) for each transformation.*

Non-normality

Figure 4.19 shows QQ-plots and density functions for both transformations. These show that both have heavy tails, although the cube root transformation only with regard to the right tail. Both have many residuals outside the 95% confidence envelope.

Non-constant error variance

Figure 4.20 shows the Residuals plotted against the Covariate *Number of Employees*, to assess whether the residual variance seems constant across values of this covariate. The cube root transformation seems to obtain more constant spread across residuals than the log transformation does.

Overall fit True Value distribution on observations in S_{000}, S_{001}, S_{010} and S_{100}

In terms of R-squared, the cube root transformation shows a better fit, with an R-squared of 0.90 in comparison to 0.70 for the log transformation. This can be explained from the fact that the log transformation somewhat *overcorrects* for the unequal spread among turnover values. As was shown in Figure 4.8 (on page 70) and is shown again in Figure 4.18 (on page 77), a log transformation obtains larger spread among turnover values for businesses with small number of employees and smaller spread among large businesses.

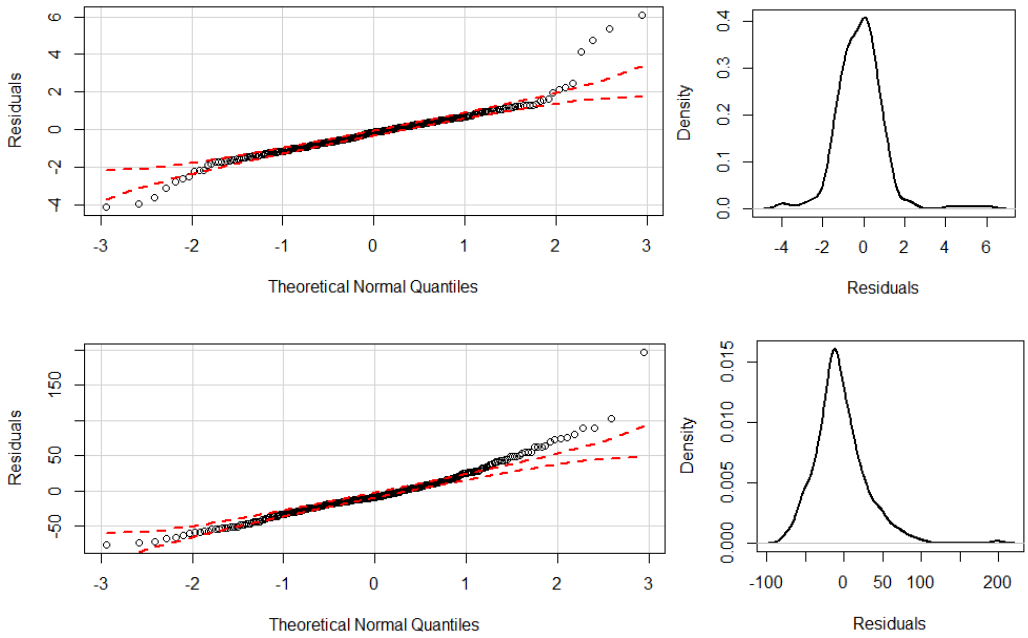


Figure 4.19 NACE group *G45.1.1.2* residuals from True Values modeled against covariate *Number of Employees* with parameter estimates β_0 and β_{Empl} as shown in Table 4.4 (on page 76). Residuals from observations in sets S_{000}, S_{001}, S_{010} and S_{100} (for which the True Value is known). *Left: QQ-plot with 95% confidence envelopes are shown by the red dotted lines. Right: Empirical density function*
Upper: log transformation, Lower: cube root transformation

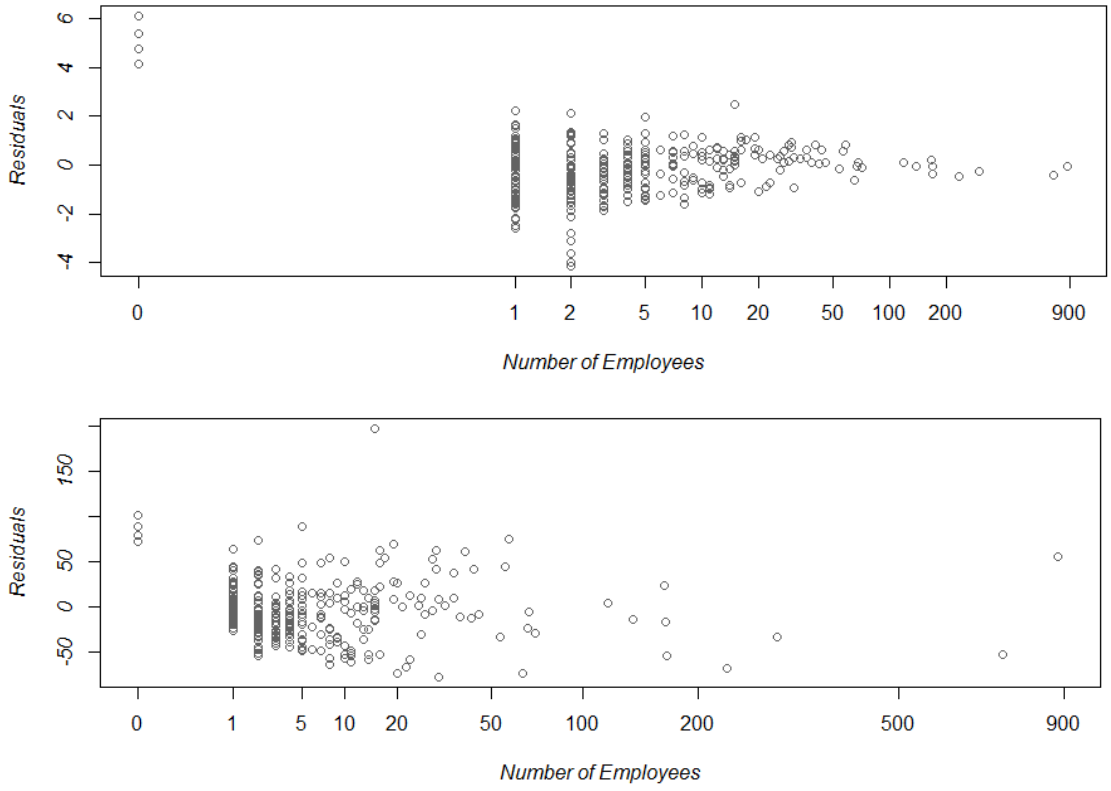


Figure 4.20 NACE group *G45.1.1.2* residuals plotted against covariate *Number of Employees*. Residuals from True Values modeled against covariate *Number of* Employees with parameter estimates β_0 and β_{Empl} as shown in Table 4.4 (on page 76), from observations in sets S_{000}, S_{001}, S_{010} and S_{100} (for which the True Value is known).
Upper: log transformation, logarithmic x-axis, Lower: cube root transformation, cube root x-axis

4.4.3.2 Model fit comparison with regard to the source error distributions

Nonlinearity and non-constant error variance
Figure 4.21 shows the erroneous measurements in sets S_{100}, S_{010} and S_{001} plotted against the known True Values for NACE group *G45.1.1.2*. Both the transformations show a linear relation and quite constant spread of residuals.

Non-normality
Figure 4.22 (on page 80) and Figure 4.23 (on page 80) show QQ-plots and Density plots for both transformations, which indicate that the heavy tails already observed for the log transformation might be even worse for the cube root transformation.

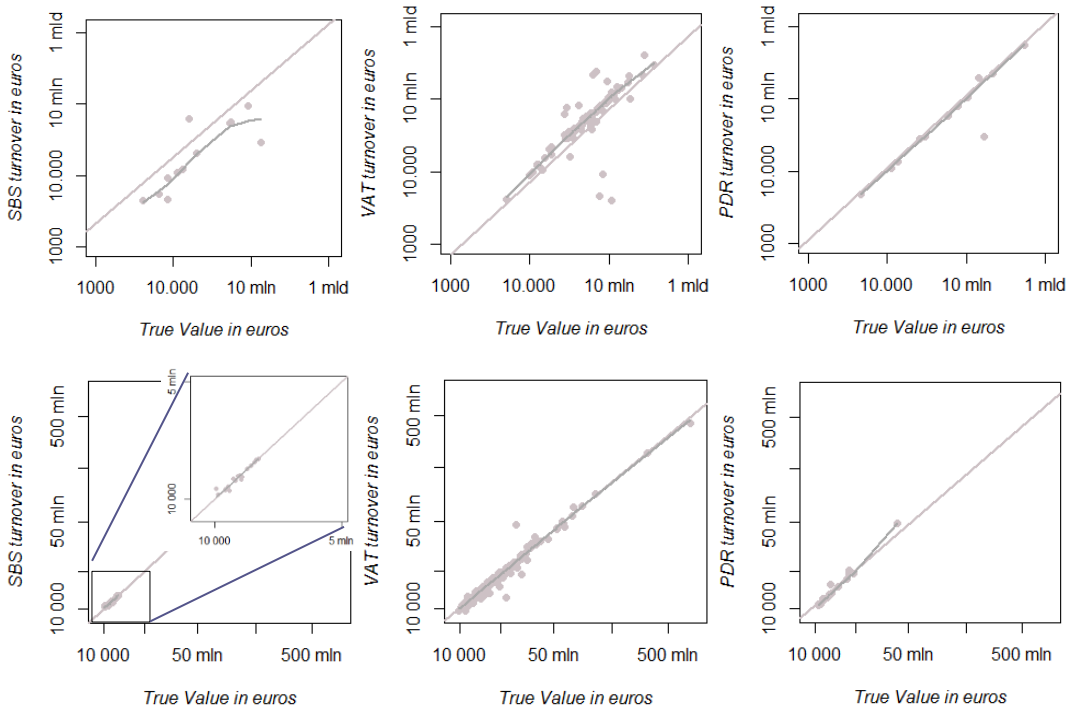


Figure 4.21 NACE group *G45.1.1.2* erroneously measured SBS, VAT and PDR *Turnover* values plotted against known True Values, for observations in sets in sets S_{100} (left), S_{010} (middle) and S_{001} (right).
Upper: logarithmic x- and y-axis, Lower: cube root x- and y-axis
Fitted regression lines are shown in the NACE group's light grey color, with a_{SBS}, b_{SBS} (left), a_{VAT}, b_{VAT} (middle) and a_{PDR}, b_{PDR} (right) as shown in Table 4.4 *(on page 76). Lowess curves are shown in dark grey with δ = 0.01*diff(range(x)).*

Overall fit True Value distribution on source error distributions
The R^2 for the SBS source error distribution on the observed True Values are 0.89 and 0.94 for the log and cube root transformation respectively. For the VAT source error distribution this is 0.95 and 0.97 and for the PDR 0.97 and 0.98.

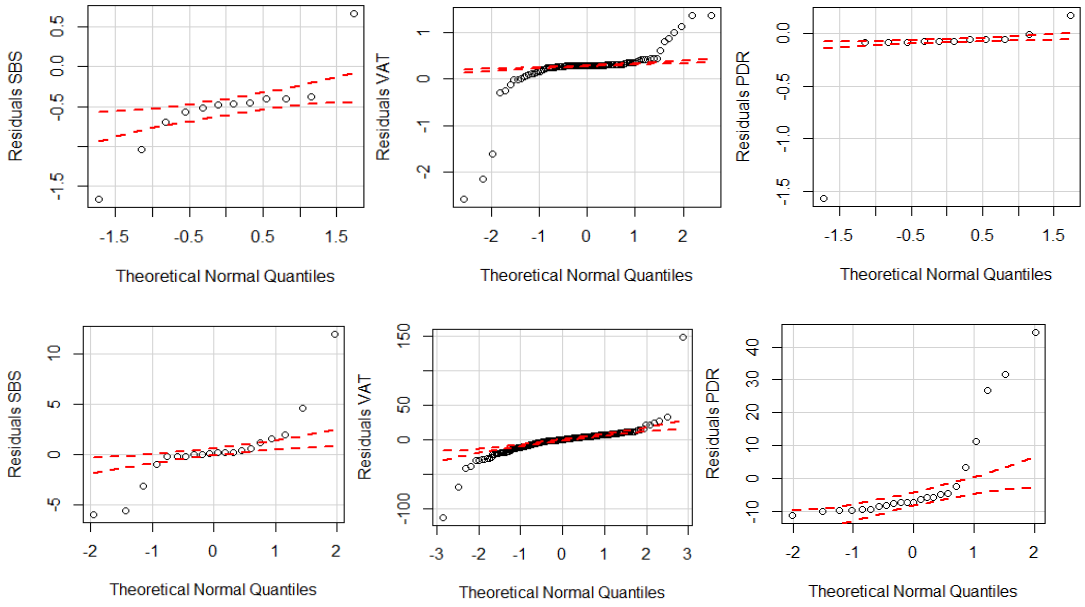


Figure 4.22 NACE group *G45.1.1.2* residuals QQ-plot. Residuals from erroneously measured *Turnover* values modeled against True Values with parameter estimates a_{SBS}, b_{SBS} *(left),* a_{VAT}, b_{VAT} *(middle) and* a_{PDR}, b_{PDR} *(right)* as shown in Table 4.4 (on page 76). Residuals from observations in sets in sets S_{100} (left), S_{010} (middle) and S_{001} (right).
95% confidence envelopes are shown by the red dotted lines.
Upper: log transformation, Lower: cube root transformation

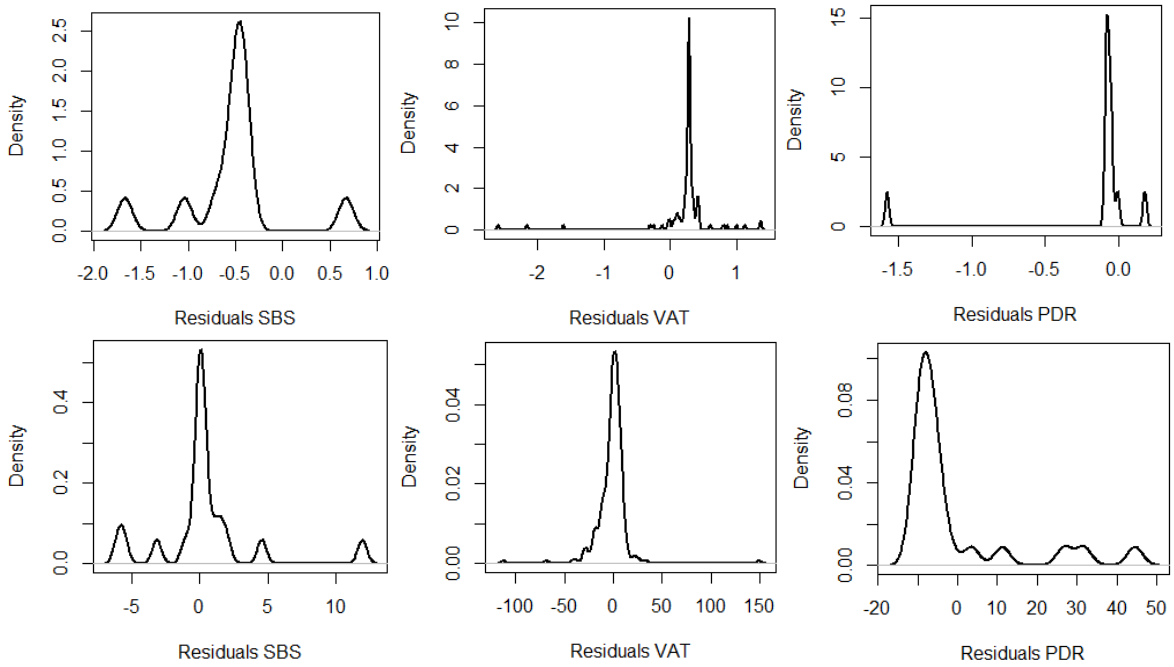


Figure 4.23 NACE group *G45.1.1.2* residuals empirical density plots. Residuals from erroneously measured *Turnover* values modeled against True Values with parameter estimates a_{SBS}, b_{SBS} *(left),* a_{VAT}, b_{VAT} *(middle) and* a_{PDR}, b_{PDR} *(right)* as shown in Table 4.4 (on page 76). Residuals from observations in sets in sets S_{100} (left), S_{010} (middle) and S_{001} (right).
Upper: log transformation, Lower: cube root transformation

4.4.3.3 Model soundness comparison with regard to the estimated True Values

When the data is log transformed, for 64 observations triplets τ_{111} is higher than the other τs and therefore the estimated True Values mainly result from the hypothesis that all measured values contain errors. When the data is transformed with the cube root transformation, this is the case for 136 observations triplets (shown in Figure 4.24), of which 52 also occur among the 64 log transformation τ_{111} triplets. The estimated True Values for these 52 triplets are shown in Figure 4.25 for both transformations.

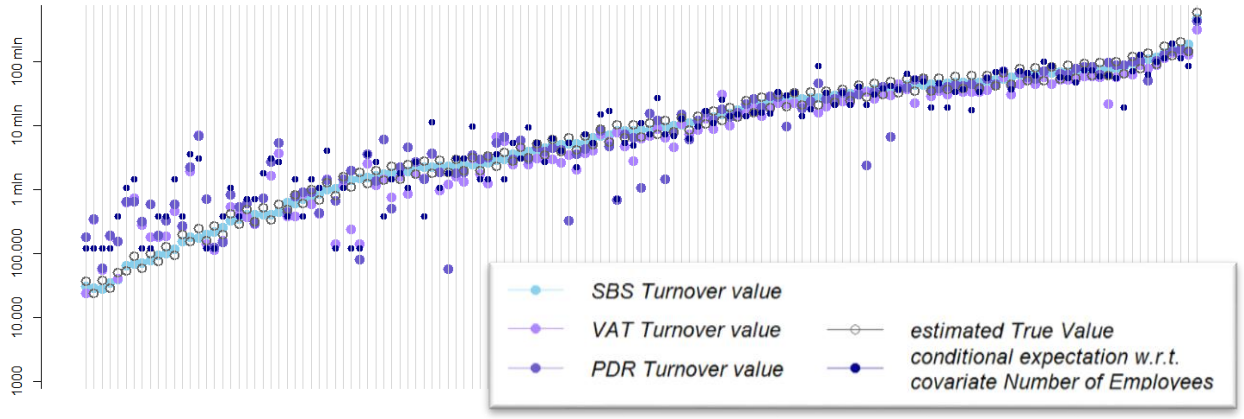


Figure 4.24 Measured turnover values, *Estimated True Values* and *Conditional expectations w.r.t. covariate Number of Employees* in euros for NACE group *G45.1.1.2* of triplets for which the estimated τ_{111} is higher than τ_{011}, τ_{101} and τ_{110}, on a logarithmic y-axis, estimated by the Intermittent-Error Model with cube root transformation on the data.
For definitions of estimated True Value and conditional expectation w.r.t. covariate Number of Employees, see Expression (3.1) and (3.2) (on page 52) respectively.
The True Value estimates are scattered to make them visible for each unit. The vertical lines facilitate observing the five values on the same statistical business unit. The horizontal axis contains an index assigned to each triplet with regard to the sorted True Value estimates.

In Figure 4.24 as well as in Figure 4.25, the *conditional expectation w.r.t. covariate Number of Employees* does not seem to influence the estimated True Value much. As already pointed out in Section 4.3.3 these conditional expectations are very crude estimates, since they are often smaller than the smallest measurement or larger than the largest measurement.

Figure 4.25 shows that in the cases where, with the highest probability, is assumed that all measurements are erroneous, the log transformation results in estimates close to the PDR measurement and the cube root transformation results in estimates close to the SBS measurement. This is in agreement with the estimated source error variances, which are estimated in case of log transformations as $\sigma^2_{PDR} < \sigma^2_{SBS}$ and in case of cube root transformation as $\sigma^2_{SBS} < \sigma^2_{PDR}$ (see Table 4.4 on page 76). This is in interaction with the estimated source error proportions π_{SBS} and π_{PDR}, which are interchanged by the two transformations. So when a higher proportion of records in a source is assumed to be erroneous, the error variance of the source is estimated to be smaller and the other way around. A shift of expected True Value estimations of more than a factor ten indicates a severe consequence of the model's indicisiveness about π_{SBS} and π_{PDR}.

Opposing estimated True Value and measurements

In 4 cases the model fitted with log transformation obtains estimated True Values that are larger than the largest measurement or smaller than the smallest measurement. These all occur with τ_{011} being the largest probability. With regard to the cube root transformation, such cases occur 32 times and all with τ_{111} being the largest probability (thus shown in Figure 4.24, although not identifiable since the

True Value dots are scattered). This is an indication of unsoundness, since the estimated True Values deviate a considerably from the average measurement.

Error pattern likelihoods

For both transformations the situation with 0 error pattern likelihoods did not occur.

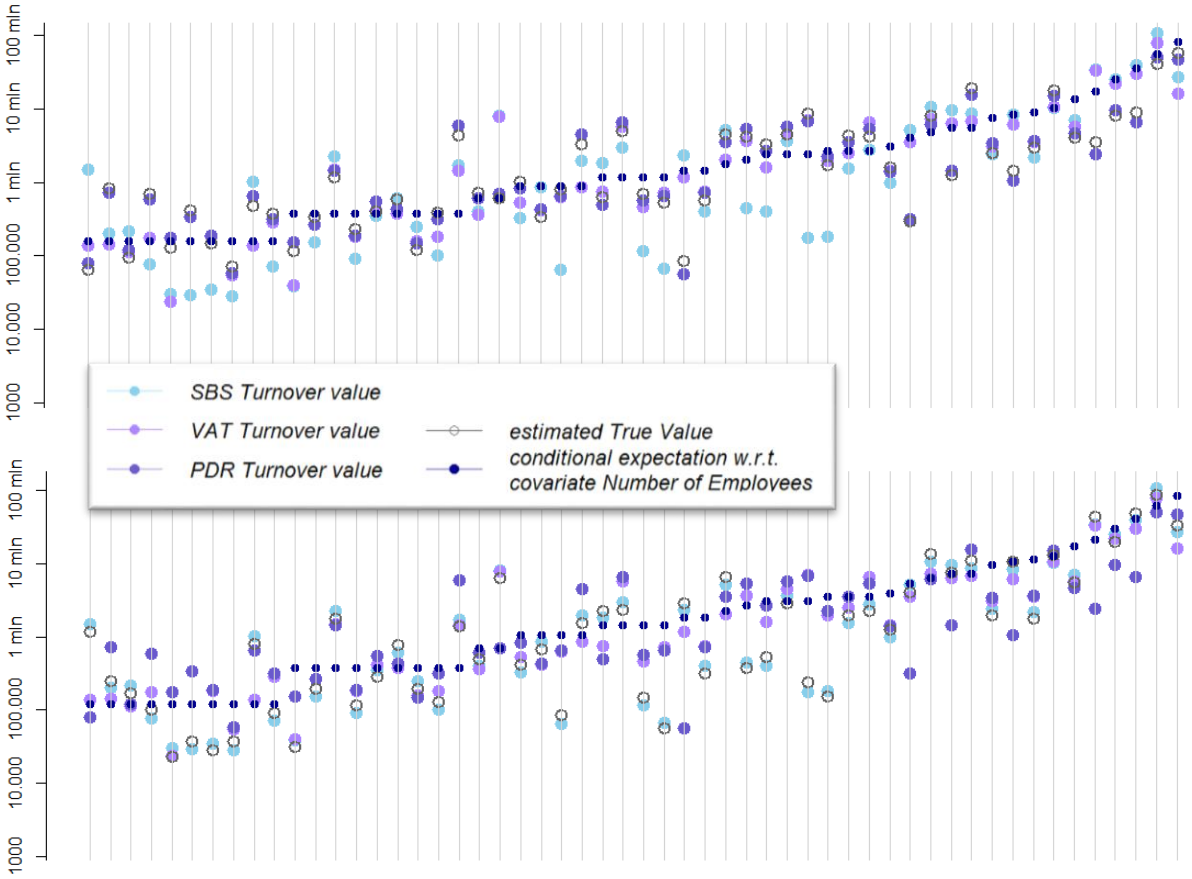


Figure 4.25 Measured turnover values, *Estimated True Values* and *Conditional expectations w.r.t. covariate Number of Employees* in euros for NACE group *G45.1.1.2* of triplets for which the estimated τ_{111} is higher than τ_{011}, τ_{101} and τ_{110}, on a logarithmic y-axis.
For definitions of estimated True Value and conditional expectation w.r.t. covariate Number of Employees, see Expression (3.1) and (3.2) respectively.
Upper: log transformation, Lower: cube root transformation
The True Value estimates are scattered to make them visible for each unit. The vertical lines facilitate observing the five values on the same statistical business unit. The horizontal axis contains an index assigned to each triplet with regard to the sorted Number of Employees.

Total estimated True Value

Table 4.5 (on page 83) shows the total turnovers estimated by the Intermittent-Error Model for the two transformations, in comparison to the measured values. (These are totals for the businesses in NACE group *G45.1.1.2* for which all three measurements were available, in contrast to the totals given in Table 1.6 (on page 24) that provided totals for the entire GBR population covered by the eight NACE groups.)

Table 4.5 NACE group *G45.1.1.2* total estimated True Value for log and cube root transformation, in comparison with the total measured *Turnover* in this NACE group on the businesses for which values from all three sources are available (overlapping part of Figure 1.8 (on page 24)).

Total ***estimated True Value Log transformation***	*Total* ***estimated True Value Cube root transformation***	Total *Turnover* **SBS**	Total *Turnover* **VAT**	Total *Turnover* **PDR**
11 522 259 000	11 655 029 000	11 636 718 000	8 903 517 000	11 472 401 000

4.4.4 How to decide on transformation?

Both transformations do not seem to fit the data very well in terms of the fit statistics on the conditional True Value distribution and source error distributions. Table 4.6 sums up the differences in characteristics previous discussed, for which the column with the better performance is marked blue. From these known characteristics it seems impossible to choose between the two transformations.

Table 4.6 Characteristics following from fit and soundness measures for log and cube root transformation. For each characteristic the better transformation is marked blue.

Log transformation	**Cube root transformation**
Intuitive interpretation	Not interpretable
Conditional True Value distribution $R^2 = 0.70$	Conditional True Value distribution $R^2 = 0.90$
Conditional True Value distribution: Non-constant error variance	Conditional True Value distribution: Quite constant error variance
Source error distribution SBS: $R^2 = 0.89$ Source error distribution VAT: $R^2 = 0.95$ Source error distribution PDR: $R^2 = 0.97$	Source error distribution SBS: $R^2 = 0.94$ Source error distribution VAT: $R^2 = 0.97$ Source error distribution PDR: $R^2 = 0.98$
2 estimated True Values outside range of measurements	32 estimated True Values outside range of measurements
Total estimated True Value in between total *Turnover* from the three sources	Total estimated True Value larger than the largest total *Turnover* from the three sources

4.5 Stability of the case study model fit

The estimated π_{SBS}, π_{VAT} and π_{PDR} showed a large shift when only two outliers on a total of 148 records were included for NACE group *H52.1.0* (Section 4.1.3). Also, for the records with high τ_{111} (Section 4.4.3.2) and parameter estimates (Section 4.4.2), another transformation of the turnover values resulted in completely different estimated True Values. For some NACE groups measurement error variances were obtained that were very close to 0. These were signs that the Intermittent-Error Model fit on the cases study data is unstable.

To investigate whether this instability was inherent to the model fit, and not due to certain outliers or an unfortunately chosen transformation, a bootstrap was carried out on the largest NACE group *G45.1.1.2* Intermittent-Error Model fit. The estimation procedure of the Intermittent-Error Model is iterative, and therefore not very fast. Also, since the primary aim was to investigate stability, and not to obtain trustworthy parameter standard errors, an initial 100 bootstrap samples were considered sufficient. Thus 100 bootstrap samples were obtained, from the original 874 records excluding the outliers from Table 4.1 (on page 64). The bootstrap samples were of size 874, just like the original data set that was fitted for this NACE group and for which the model fit was previously discussed in this chapter. Figure 4.26 shows histograms of the parameter estimates for π_{SBS}, π_{VAT} and π_{PDR} that were obtained from the 100 bootstrap samples. The same 100 bootstrap samples were fitted twice, once log transformed and once cube root transformed. Results for both are shown in Figure 4.26.

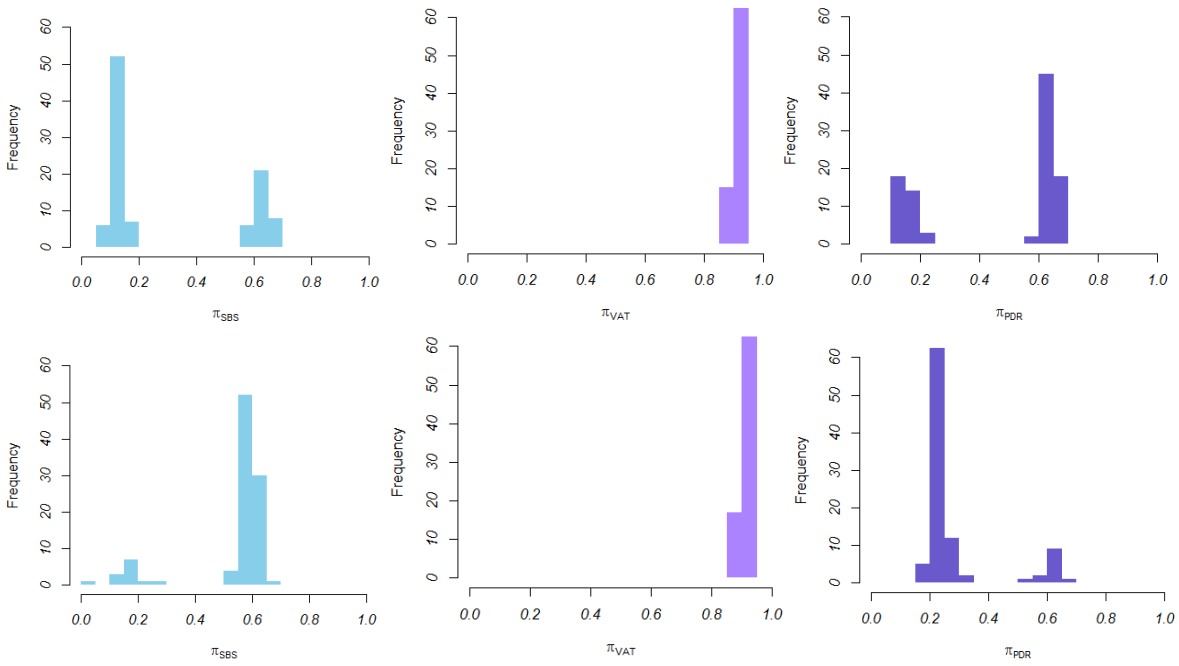


Figure 4.26 NACE group *G45.1.1.2* histograms of parameter estimates for π_{SBS}, π_{VAT} and π_{PDR} from the Intermittent-Error Model's fit on the same 100 bootstrap samples of original size 874. ***Upper: log transformation, Lower: cube root transformation***

Figure 4.26 shows that the estimates for π_{SBS} and π_{PDR} are very unstable, showing a shift in a considerable number of records for which the SBS measurement is considered erroneous in one bootstrap sample, while the PDR measurement is considered erroneous in another bootstrap sample. The histograms in Figure 4.26 also show that this instability occurs in for the cube root transformed turnover values just as well as for the log transformed ones, although the instability is less severe for the cube root transformation. The cube root points out the SBS measurements as erroneous in considerably more bootstrap samples than it does the PDR measurements. However, the situation with zero error pattern likelihoods, discussed in Section 3.3.2.2 as sign of bad fit, occurred for none of

the log transformation bootstrap model fits, but in 7 out of 100 cube root model fits. Therefore, the fact that the cube root transformation seems more stable could also be artificial, by means of the bypass programmed in the estimation procedure in case of four 0 likelihoods.

Figure 4.27 shows the shifts in source error variances that accompany the shifts in expected error proportion, as discussed in Section 4.1.3. The variances of the measurements on log scale and cube root scale are of a different order, but show the same ambivalence between a very small (only slightly larger than 0, as was discussed in Section 4.1.3) and a much larger variance. These very small variances occur when the expected error proportion is high and therefore many (mainly also those with small deviations between True Value and erroneous measurement) records contribute to the variance. The much higher variances occur when the number of records contributing to the variance is small (small expected error proportion) and therefore the records with severe deviations have much larger influence. Such small error variances might result from spurious solutions, as discussed in Section 4.1.3, but the starting values ensure that quite a few records are assigned to sets S_{001} and S_{100} (see Table 4.3 on page 68) supplying trustworthy starting values for these variances as shown in Figure 4.29 (on page 86). Therefore, these very small error variances could also be a genuine maximum likelihood solution.

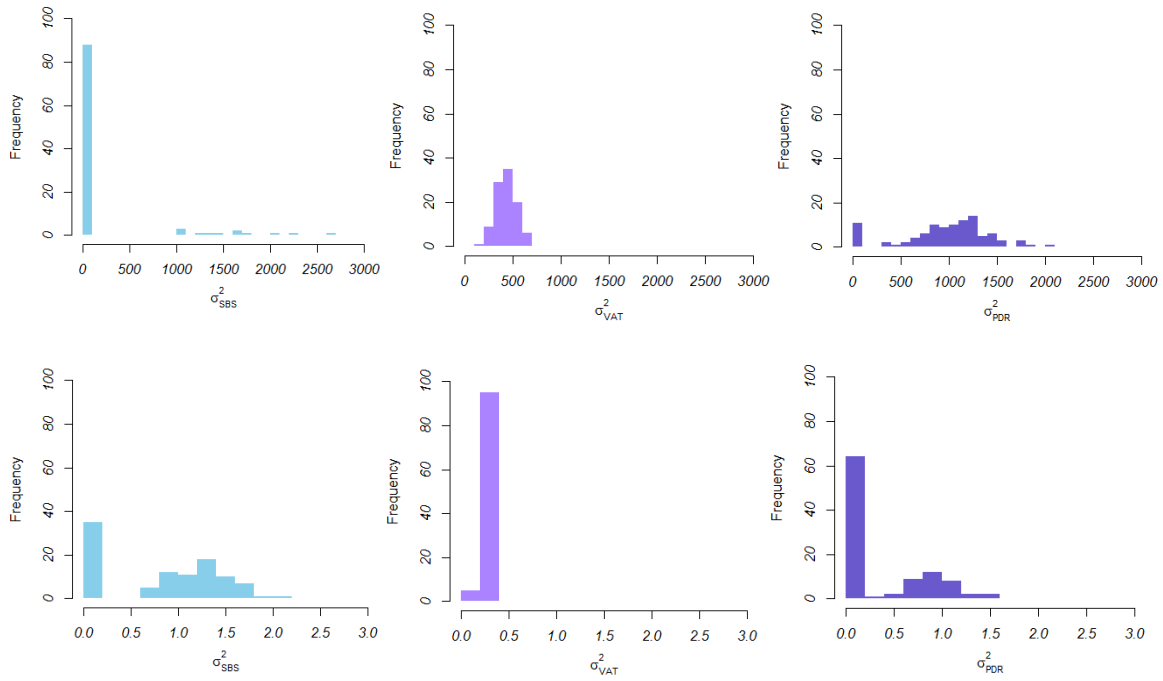


Figure 4.27 NACE group *G45.1.1.2* histograms of parameter estimates for σ^2_{SBS}, σ^2_{VAT} and σ^2_{PDR} from the Intermittent-Error Model's fit on the same 100 bootstrap samples.
Upper: log transformation, Lower: cube root transformation

Figure 4.28 and Figure 4.29 (on page 86) show the EM algorithms first 30 and 40 iterations on the expected error proportion and source error variance parameters for a sample of 10 randomly chosen data sets out of the original 100 bootstrap samples. Even though the bootstrap samples have about the same starting values for these parameters, the parameters show very different iteration paths, due to the observations that vary across bootstrap samples. Since, the original size bootstrap samples drawn with replacement ensure that most bootstrap samples have a significant proportion of units in common.

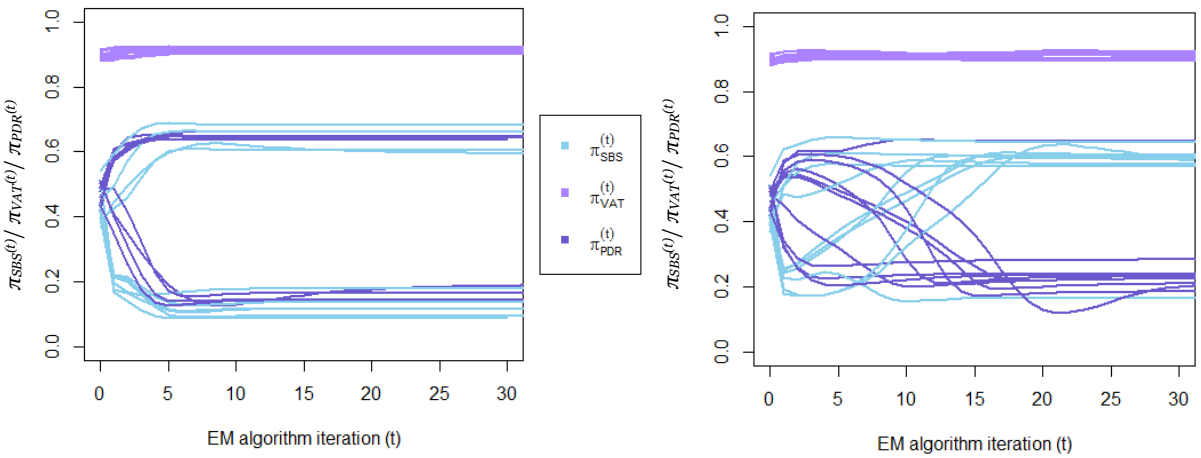


Figure 4.28 $\pi_{SBS}^{(t)}/\ \pi_{VAT}^{(t)}/\ \pi_{PDR}^{(t)}$ estimations in various iterations of the EM-algorithm for 10 randomly chosen data sets out of the original 100 bootstrap shown in Figure 4.26.
Left: log transformation, Right: cube root transformation
The same 10 samples are shown for the two transformations and therefore the parameter estimates have the same starting values. In both cases the parameter estimates don not deviate after 30 iterations.

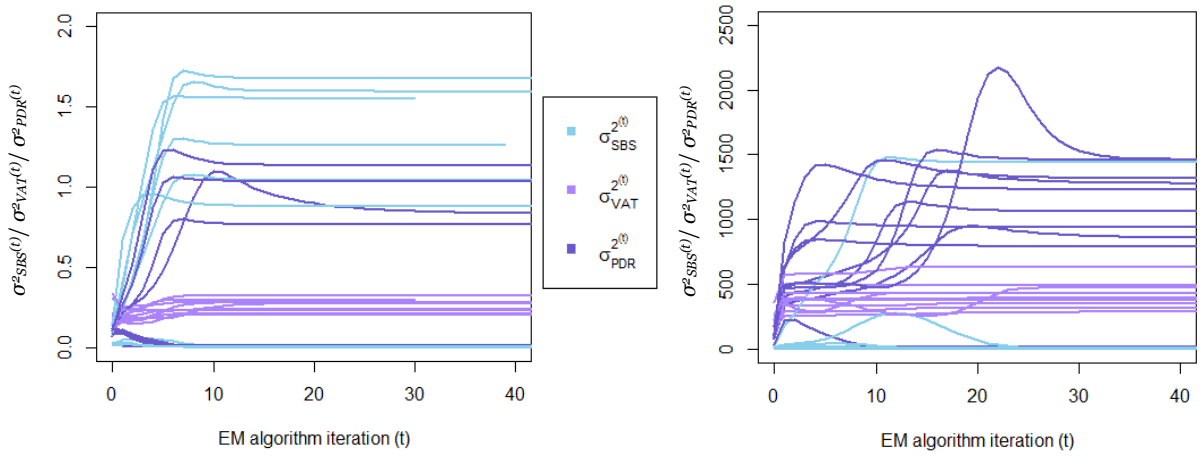


Figure 4.29 $\sigma^2_{SBS}{}^{(t)}/\ \sigma^2_{VAT}{}^{(t)}/\ \sigma^2_{PDR}{}^{(t)}$ estimations in various iterations of the EM-algorithm for 10 randomly chosen data sets out of the original 100 bootstrap shown in Figure 4.26.
Left: log transformation, Right: cube root transformation
The same 10 samples are shown for the two transformations, but the starting values are different since the transformations result in different models and variances on a different scale. In both cases the parameter estimates do not deviate after 40 iterations. Moreover, some of the samples show convergence already in case of the log transformed data in the left plot (where some lines end before 40 iterations).

4.6 Summary

In response to research sub question 4 *What is the Intermittent-Error Model's performance on a case study data set?*

Plots of the measured turnover values showed outlying measurements that do not agree with the normal source error assumption. These outliers were categorized as either *zero outliers*, *€1000 outliers* or *negative values*. The zero values were accompanied by clearly nonzero measurements from other sources, and the €1000 outliers, which were originally 1 values in the SBS survey (which required reporting in multitudes of €1000), also came along with much larger values from the tax registers. Intermittent-Error Model parameter estimates showed that including or excluding these outliers had a very large impact on the parameter estimates in three of the eight NACE group model fits, for which the outliers constituted only 1-4% of the total records.

The business turnover values in the case study data set show unequal spread and positive skew. The Intermittent-Error Model assumes normal source errors and normal unexplained variance in the conditional True Value distribution and therefore minimizes squared loss to obtain parameters of central tendency. Therefore, to obtain appropriate parameters, the unequal spread and positive skew in the turnover values needs to be corrected by a transformation. A *log transformation* (with small additive constant to deal with zeros) and *cube root transformation* were considered.

Since the transformations completely change the model with regard to the original scale, not all parameter estimates have the same meaning for both transformations. Therefore, large deviations were expected among the simple linear model parameters of the True Value distribution and source error distributions. However, the parameter estimations for π_{SBS} and π_{PDR} also deviated considerably across transformations. A large shift in π_{SBS} and π_{PDR} was observed from the SBS measurements containing errors on very few businesses according to the log transformation model and the PDR on more than half, to completely the other way around for the cube root transformation.

For both transformations fit and soundness measures were inspected that were proposed in Section 3.3. While linearity of the conditional True Value distribution was met for both transformations, the log transformation showed an overcorrection of the original unequal spread in the True Value distribution resulting in too large residual variance for a smaller number of employees. Both transformations showed deviations from normality with regard to the source error distributions. Soundness measures on the estimated True Value showed a large shift among the two transformations. Not only does the cube root transformation assign many more records (136 in comparison to 64) τ_{111} as the largest τ, also the estimated True Values showed deviations of a factor 10. The cube root transformed model estimated a total estimated True Value in one subpopulation that was €132 770 000 (1%) larger than that from the log transformation. The total estimated True Value was even €18 311 000 larger than the largest measured total from the SBS survey. Plots showed that for the records with τ_{111} the largest τ, the estimated True Value was more often in between the measured ones for the log transformation than it was for the cube root transformation. Also it occurred only 4 times that the estimated True Value was outside the range of measured turnovers, in contrast to 32 times for the cube root transformation. With so many characteristics in which either the log transformation or the cube root transformation performed the best, it seemed impossible to choose the best transformation.

A bootstrap was carried out to further investigate the stability of the parameter estimates, with bootstrap samples of the original size of the data set, excluding outliers. For both transformations, in about half of the bootstrap samples an error proportion in the range 0.1-0.3 was estimated for SBS and PDR values, while in the other half the estimation ranged between 0.5 and 0.7. Thus even within transformations and when outliers were excluded, the Intermittent-Error Model fit was unstable.

5 Intermittent-Error Model's merits for Statistics Netherlands

How wrong does a model need to be, to not be useful? Chapter 4 showed that the Intermittent-Error Model is not robust towards outliers, unstable across transformations and even unstable within transformations, as shown by the bootstrap samples.

Originally, four ways were considered in which the Intermittent-Error Model could be useful in processing Statistics Netherlands' business data. The first is *error detection*, meaning the detection of the measurement that is incorrect when multiple measurements disagree. The second is *error correction*, by proposing a True Value even when all available measurements can be considered erroneous. The third is *source assessment*. Like previously carried out research on the VAT register for the Short-Term Business Statistics (discussed in Section 1.3.1), the Intermittent-Error Model could be applied to reconsider the VAT register, or investigate the PDR register, for their trustworthiness across NACE groups to produce economic statistics. Also a source can be assessed in terms of its reporting differences with regard to the True Value, and the Intermittent-Error Model can be used to correct deviations in target value definitions. The fourth way is to order suspicious records in terms of severity and influence of errors, to *aid the manual process of selective editing*.

Firstly, Section 5.1 discusses the Intermittent-Error Model's definition in relation the possible error types in Statistics Netherlands' business data. Then, each of the four ways by which the Intermittent-Error Model could be applied is discussed in its own section (5.2 - 5.5). Section 5.6 concludes this chapter with a summary.

5.1 The Intermittent-Error Model definition and possible error types

The Intermittent-Error Model is able to estimate which observations are incorrect when multiple observations on a variable disagree, and to propose a True Value. To achieve this, an independent intermittent error mechanism is assumed, and distributional assumptions are made on the True Value. For incorrect observations, a parameter for intercept bias and slope bias is estimated for each source, and the remaining error is assumed to be normally distributed. In its application on Statistics Netherlands' business data, the Intermittent-Error Model operates on the *Turnover* variable, and therefore deals with *Measurement Errors (cat)* that originated on the left side of the Zhang Two-Phase Life-Cycle Model (Figure 1.2 and Figure 1.4 on page 8 and 12). So which specific error types can be handled by the Intermittent-Error Model?

The *Turnover* values were discussed in Section 1.2.3.2, with regard to the BaseLine execution of the Zhang Two-Phase Life-Cycle Model in which yearly financial data is linked to the already existing population framework. Since the errors are assumed to be normally distributed, not many large deviations from the True Value are expected. Therefore, the errors that the Intermittent-Error Model is able to correct might originate in slight deviations in reporting by the businesses themselves, such as estimation of components of the turnover value instead of strictly calculating it. This error source is part of the original data collection in phase one, and denoted by *measurement error* (for the distinction between the category Measurement Error (cat) and the individual error measurement error see the introduction to Section 1.2.2). Another possible source of such slight misspecifications of *Turnover* values is the processing of the *Turnover* values. This processing mainly occurs with regard to the values from the SBS survey, which underwent a procedure of automatic and manually editing at Statistics Netherlands. For example, the automatic editing procedure that assures consistency across multiple values from the same survey (purchases, costs, turnover), might introduce slight deviations from the true value, which are *processing errors* (also phase one, since the SBS values do not enter a second phase).

The Intermittent-Error Model not only assumes random normally distributed errors, but also systematic deviations from the 'True' turnover concept in terms of intercept and slope bias. These deviations in concept mainly occur because data collected for a different purpose is reused to produce statistics, which is the case for the *Turnover* values from the VAT register and PDR. This reuse of data that was collected elsewhere is described by phase two of the Zhang Two-Phase Life-Cycle Model. Deviations between the obtained concept and the target concept are denoted by *relevance errors*. Thus by estimating the intercept and slope bias in each source, the Intermittent-Error Model corrects for relevance errors by proposing True Values, which is a form of harmonization. In case the Intermittent-Error Model introduces new errors by doing this, either *relevance errors* occur because the harmonization procedure is incorrect, or *mapping errors* occur if the procedure is correct but the input data, such as the *Number of Employees* values, are incorrect in a way that causes the output re-classified measures to be erroneous.

While the Intermittent-Error Model does not model *Representation Errors* it can suffer from them. When the data shows over- or under-coverage of the true population, biased parameter estimates are proposed. Since the Intermittent-Error Model is fitted for each NACE group separately, over- or under-coverage can also occur due to NACE group misclassification, which causes misrepresentation of NACE group subpopulations. In this situation, the model might detect patterns (source intercept and slope bias, expected error proportions) within NACE groups that do not describe the real structure of the NACE group.

Errors in outliers

In the procedure of fitting the Intermittent-Error Model to the case study data set, described in Section 4.1, certain observations were classified as outliers and were subsequently excluded. These outliers were distinguished based on large deviations between the three source observations, by scrutinizing plots of values from the three sources. The errors that occurred in these outliers were incorrect zero SBS, zero VAT or zero PDR values, incorrect €1000 SBS values, and negative VAT values. For these outliers it was assumed that the occurring deviations between values cannot be explained by systematic deviations in *Turnover* concept or normally distributed errors. So which error types do we assume the Intermittent-Error Model cannot handle?

In general, the errors that were excluded are unlikely to be *measurement errors*, since a business reporting a crude 0 or 1 measurement/estimation would be unlikely to report an accurate measurement to another source. These errors are much more likely to have occurred in processing the data, being *processing errors* when they occurred in the first phase at the Tax and Customs Administration or Statistics Netherlands (phase one for the SBS turnovers) or *comparability error* in processing in phase two for statistical reuse of the administrative tax data. The negative VAT value most likely occurred at the Tax and Customs Administrations, which sometimes allows businesses to report negative turnover to correct an overestimation of turnover in a previous tax report. Therefore, this negative value could be characterized as a *relevance error*, since it can be attributed to a difference between the statistical target concept and the available administrative information.

The errors in the outliers can also have originated in the creation of the GBR population framework (GBR execution of the Zhang Two-Phase Life-Cycle Model). Different businesses might be erroneously linked together which produces *identification errors*. Also, the software that creates and maintains the units might contain a bug, or the profiler handling the software might have made a mistake, which both create *unit errors*. The large deviations in the values from the three sources that were observed in the outliers might thus have originated from businesses that are not the same, or a 1 or 0 might be falsely introduced for units that were created or maintained in an unstable manner. So the very severe errors that originate in the GBR definition cannot be handled by the Intermittent-Error Model. Could these be excluded based on already known characteristics?

Some available variables in the case study data set from the GBR and BaseLine, described in Section 1.3.2, might point to such Representation Errors or the absence of them. None of the outliers was characterized by the GBR as part of the TopX, which supports the idea that the way TopX businesses

are scrutinized prevents such outliers. For most zero PDR or zero VAT outliers, the variables VAT *Response Percentage* and PDR *Coverage/Fill Percentage* and *Unlock Code* could not explain the outliers. Almost all zero PDR outliers had PDR *Unlock Code* A (86%) or C (8%). And almost all had PDR Coverage and *Fill Percentage* 100%. For most zero VAT outliers the VAT *Response Percentage* was missing, even though not for all the variable *VAT Quarters* was missing (10 out of 14 reported 4 *VAT Quarters*). The outliers were the only VAT turnover values for which values were available from all three sources and the VAT *Response Percentage* was missing. Therefore, the missing VAT *Response Percentage* can predict some of the zero VAT outliers, although not all: 2 out of 14 had VAT *Response Percentage* 100%. Editing at Statistics Netherlands might have introduced the errors in the SBS (0/€1000) outliers or at least did not correct them. Most of the SBS outliers were only automatically edited but one €1000 outlier entered the manual editing procedure (but it could be that it was not scrutinized by an editor because of the selective procedure). So to sum all characteristics up, the outliers did not deviate from other records with regard to the available variables and could therefore not be distinguished in another way than to compare the *Turnover* values from the three sources.

Errors in GBR *Number of Employees*

The Intermittent-Error Model requires covariates for its conditional True Value distribution for which not many possible variables are available. Only the identification and structural variables included in the GBR were considered for this role, since a variable obtained from one of the turnover sources cannot theoretically be considered error-free, and might favor the source it comes from. The GBR variable *Number of Employees* is the variable that the GBR *Size Class* is based upon, which is one of the foundations of the sampling procedure for the Structural Business Statistics. Section 1.3.3 showed however, that this variable might contain errors since it showed disagreements with the *Size* Class variable as well as it large deviations from the *Number of Employees* reported in the SBS survey. These deviations could have originated from erroneous measurement or processing (*measurement*/*processing errors*) in the Tax and Customs Administration's Relations Register or in building the GBR at Statistics Netherlands (*mapping*/*comparability* error). Or two different businesses might be incorrectly combined into one (*identification error*) and have caused the disagreeing numbers. When the difference between GBR and SBS results from different definitions of *Number of Employees,* the errors could have originated as *relevance errors*.

So the GBR *Number of Employees* is the best covariate available and its use is already accepted at Statistics Netherlands. However, it is known that covariates in linear models that contain errors bias the parameter estimates, which could also be the case with regard to the Intermittent-Error Model's parameters. Fortunately, Section 3.2 investigated the role of the covariate in the conditional True Value distribution and concluded that its influence was limited.

5.2 Merits with regard to error detection

The Intermittent-Error Model assigns each record to set S_{000}, S_{001}, S_{010} or S_{100}, or assigns four error pattern probabilities τ_{011}, τ_{101}, τ_{110} and τ_{111} to the record. Therefore, for each turnover value it is known (with a certain probability) whether the measurement contains an error. Unfortunately, Chapter 4 showed that the assignments of τs was unstable for certain records, mainly when either τ_{011} or τ_{110} was high. So these assignments cannot be used to determine which turnover value contains an error.

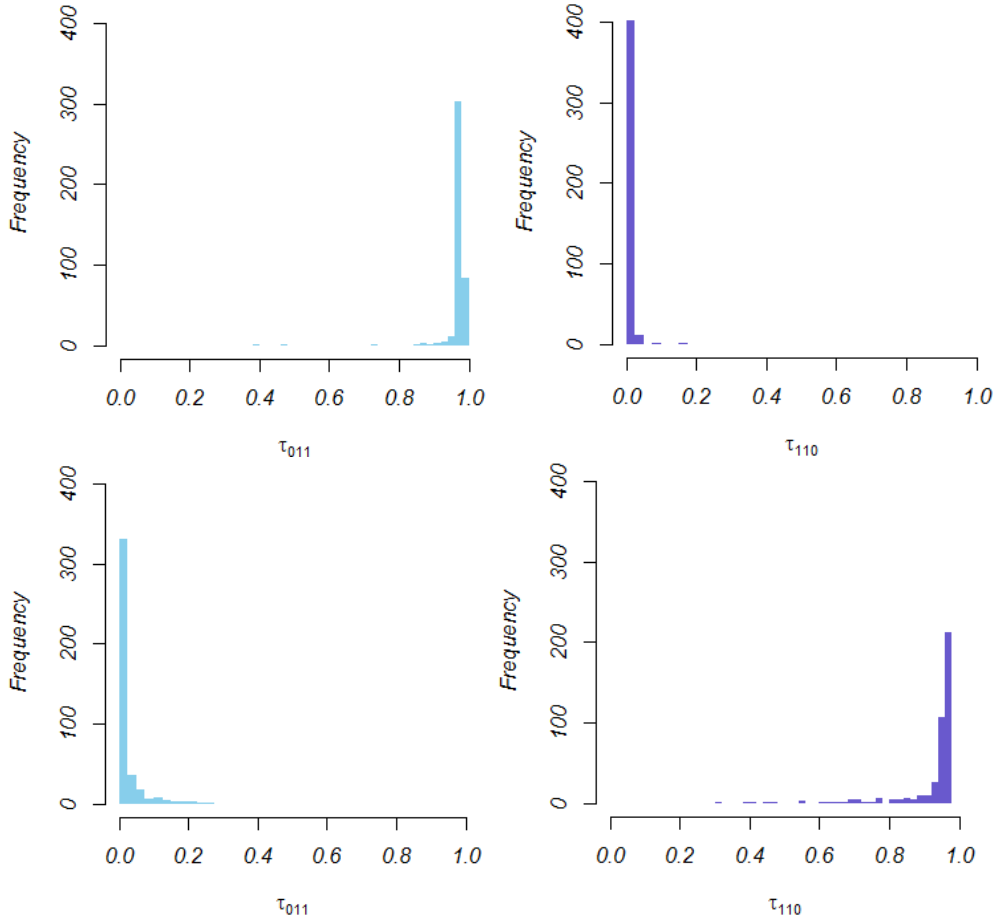


Figure 5.1 NACE group *G45.1.1.2* values for τ_{011} and τ_{110} for records that switch when data is cube root transformed instead of log transformed from being most likely in S_{011} (τ_{011} bigger than τ_{101}, τ_{110} and τ_{111}) to being most likely in S_{110} (τ_{110} bigger than τ_{011}, τ_{101} and τ_{111}) with $\boldsymbol{\tau_{klm,i}} = \mathbb{P}\left((z_{SBS,i}, z_{VAT,i}, z_{PDR,i}) = (k,l,m) \,\middle|\, y_{SBS,i};\ y_{VAT,i};\ y_{PDR,i};\ x_{Empl,i};\ i \in S_{011} \cup S_{101} \cup S_{011} \cup S_{111}; \boldsymbol{\theta}\right)$
Upper: log transformation, Lower: cube root transformation

Figure 5.1 shows the τ_{011} and τ_{110} estimates for the records in NACE group *G45.1.1.2* for which τ_{011} was the largest τ when the model was fitted on log transformed data, and τ_{110} was the largest τ when the model was fitted on cube root transformed data. Most τ values are very close to either 0 or 1, and therefore do not indicate that the model is uncertain about the error pattern of these records. Whereas the unstable estimation of π_{SBS} and π_{PDR} in the bootstrap samples with both transformations (see Section 4.5) showed that the model should be uncertain about the error patterns of these records. Figure 5.2 (on page 93) shows a sample of the records for which the assigned error pattern probabilities are unstable, showing measurements that are very close together and estimated True Values that are approximately the same regardless of the error pattern with the highest probability. So even though the estimated true values for these records might be trustworthy, the most probable error patterns according to the τs cannot be used to identify which measurements are erroneous.

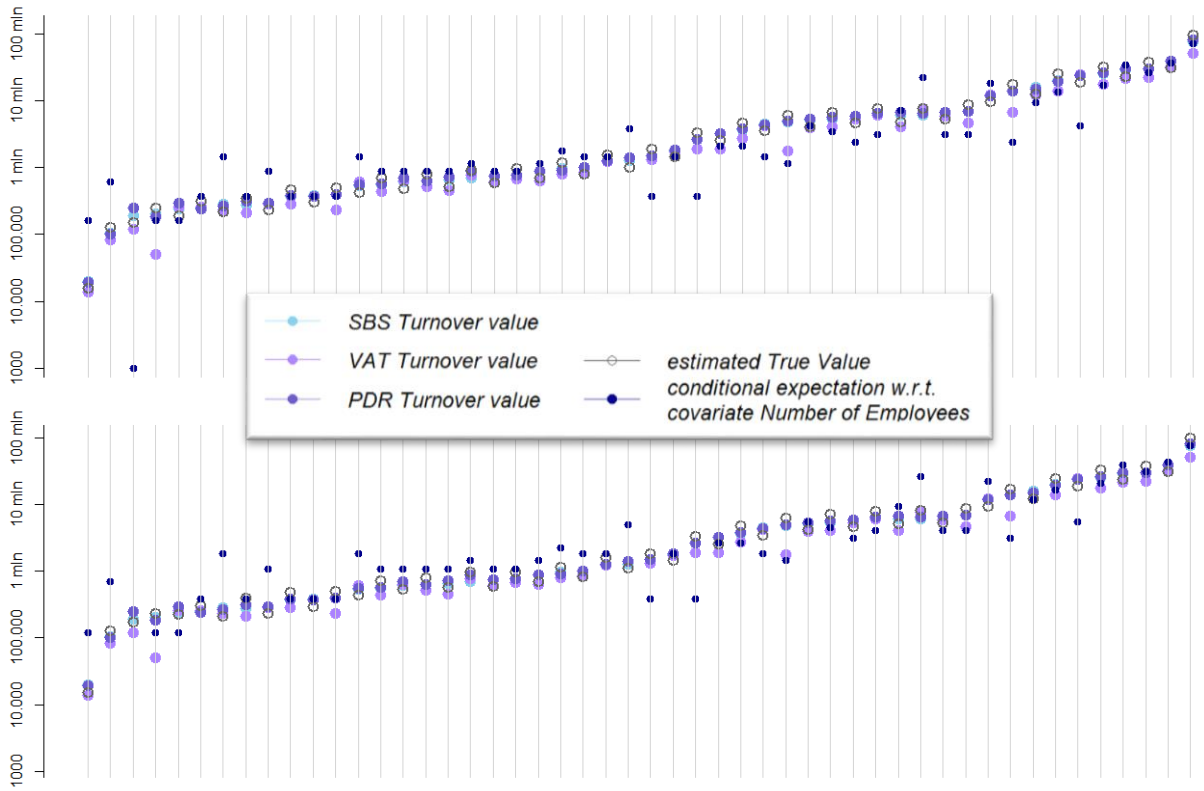


Figure 5.2 NACE group *G45.1.1.2* measured turnover values, *Estimated True Values* and *Conditional expectations w.r.t. covariate Number of Employees* in euros on a logarithmic y-axis. Sample of 50 records from total of 415 records that switch when data is cube root transformed instead of log transformed from being most likely in S_{011} (τ_{011} bigger than τ_{101}, τ_{110} and τ_{111}) to being most likely in S_{110} (τ_{110} bigger than τ_{011}, τ_{101} and τ_{111}).
For definitions of estimated True Value and conditional expectation w.r.t. covariate Number of Employees, see Expression (3.1) and (3.2) (on page 52)respectively.
Upper: log transformation, Lower: cube root transformation
The True Value estimates are scattered to make them visible for each unit. The vertical lines facilitate observing the five values on the same statistical business unit. The horizontal axis contains an index assigned to each triplet with regard to the sorted True Value estimates.

5.3 Merits with regard to error correction

Figure 5.2 showed that even though the detection of erroneous measurement (SBS or PDR) is unstable, the estimated True Value is quite stable and can therefore be used as corrected value. Therefore, for these records an error can be corrected even though it is not known which measurement contained an error. Is it possible to use the Intermittent-Error Model's estimated True Values in this way for all records? Unfortunately, the answer is no. Figure 4.25 (on page 82) showed that for records with high τ_{111}, the estimated True Values are very unstable. These deviations of estimated True Value across transformations are also very influential. Table 4.5 (on page 83) showed that the difference in total estimated True Values was €132 770 000 on the original scale, which is 1% of the total. That seems a small percentage, but since these totals are used to obtain economic statistics and especially economic growth, this can lead to very misleading perceived trends.

The unstable estimated True Values across transformations resulted mainly from the change in source error variance, as was discussed in Section 4.4.3.3. Large σ^2_{SBS} is related to small π_{SBS} values and the same holds for the PDR as argued in Section 4.1.3. Since Section 4.5 also showed unstable π_{SBS} and π_{PDR} estimates within transformations accompanied by unstable σ^2_{SBS} and σ^2_{PDR} estimates, stable True Values are not regarded a matter of choosing the right transformation.

5.4 Merits with regard to source assessment

5.4.1 Source expected error proportions and error variances

π_{SBS}, π_{VAT}, π_{PDR}, σ^2_{SBS}, σ^2_{VAT} and σ^2_{PDR} can be considered parameters that assess the quality of the three sources in terms of error proportions and error size. With regard to the VAT register, these values seem trustworthy: They are quite robust towards outliers (shown in Table 4.2 (on page 65)), insensitive to the chosen transformation (shown in Table 4.4 (on page 76) for NACE group *G45.1.1.2*) and also seem stable within transformations (shown in Figure 4.26 and Figure 4.27 (on page 84 and 85)). Therefore, it can be concluded that this source does not have a high quality for these NACE groups, which was also concluded in previous research (see Section 1.3.1).

The estimates for the SBS and PDR are not generally robust towards outliers and unstable across and within transformations. Therefore, the Intermittent-Error Model cannot be used to assess which is the better source. And also not, when the SBS is considered the best source on the basis of external information, whether the PDR might perform about equally well.

5.4.2 Systematic deviation (intercept bias/slope bias)

When systematic deviations in turnover concept exist, so-called *reporting differences*, the intercept bias (a_{SBS}, a_{VAT} and a_{PDR}) and slope bias (b_{SBS}, b_{VAT} and b_{PDR}) that relate the measurements to the True Value can be considered to correct these (*harmonization*). One advantage of the Intermittent-Error Model over previous VAT research into systematic deviations (see Section 1.3.1) could be that only the erroneous records are used to estimate the systematic deviations. Since certain tax regulations and exemptions often do not apply to all businesses in a NACE group, investigating the deviations in turnover concept for only a subgroup of the businesses might be reasonable.

However, this approach does require that the businesses for which the systematic deviations need to be estimated are correctly selected. So with regard to the VAT data, which showed stable behavior, the selection of erroneous records might be considered correct and systematic deviations are correctly estimated by a_{VAT} and b_{VAT}. However, this is not the case for the SBS and PDR. Moreover, the estimations for a_{VAT} and b_{VAT} are not very robust towards outliers, as shown in Table 4.2 (on page 65). Therefore, estimating systematic deviations in this way either needs careful data set selection or might be less reliable than the robust estimations carried out in previous research.

5.5 Merits with regard to selection in manual editing

Manual editing at Statistics Netherlands is part of a selective procedure (see Section 1.4) by which only the most influential records are scrutinized by editors. Whether a record contains an error that is influential because it is large, or because the business is very influential to the intended aggregated values, is decided by an automatic procedure. Can the Intermittent-Error Model play a role in that procedure?

The Intermittent-Error Model can give estimations on the size of the errors it detects, for example in terms of the distance between the estimated True Value and the measured values. Figure 4.25 and Figure 5.2 showed that, with regard to NACE group *G45.1.1.2*, these distances can be quite large for records with high τ_{111}, while they are rather small in records with high τ_{011} or τ_{110}. Therefore, a selection procedure based on the Intermittent-Error Model would select the records with high τ_{111} to be first in line for manual editing. Unfortunately, mainly for these records, the estimated True Values are unstable across transformations, and as was argued in Section 5.3, most likely also within transformations. Therefore, a selection ordering based on the Intermittent-Error Model would be unstable as well.

5.6 Summary

In response to research sub question 5: *What are the Intermittent-Error Model's merits in detecting and correcting error types occurring in Statistics Netherlands business data?*

The Intermittent-Error Model's definition contains a detection of systematic deviations between observed values and True Value for each source g, (*intercept bias* a_g and *slope bias* b_g) and assumes normally distributed errors. Therefore, it enables the detection of errors that occur as a result of deviating target concept and measured concept, so-called *relevance errors*, and error types that are not likely to create many large deviations, such as errors in measurement (the business might have reported an estimated value instead of a measurement) and most not too severe mistakes in processing. Businesses with severe deviations in turnover values were considered outliers with regard to the Intermittent-Error Model's error concept, and excluded from the model fit. These outliers, error that the Intermittent-Error Model is considered not to handle, might also have their origins in more severe incorrect processing at the external register or Statistics Netherlands. Or these errors might not be category Measurement Errors at all, but are related to misrepresentation of the business units, being Representation Errors. An example is the linkage of two values that belong to separate businesses.

With regard to error detection the Intermittent-Error Model assigns either an error pattern (S_{000}, S_{001}, S_{010} or S_{100}) or four error pattern probabilities (τ_{011}, τ_{101}, τ_{110} and τ_{111}) to each record. The assigned error pattern probabilities were shown to be unstable, and therefore cannot be used to detect which observation is erroneous for all businesses.

With regard to error correction the Intermittent-Error Model proposes True Values for each record. These estimated True Values were shown to be unstable for records with large τ_{111}, and therefore cannot be used to correct errors in these records.

With regard to source assessment the Intermittent-Error Model estimates an expected error proportion (π_g) and intercept (a_g) and slope bias (b_g) for each source. These expected error proportions were shown to be unstable and therefore also the intercept and slope bias are unstable since they depend on the proportion of records classified as erroneous. Thus the Intermittent-Error Model fit cannot be used to assess sources' quality.

With regard to selection in manual editing the Intermittent-Error Model proposes True Values for each record which can be used as a measure of severity of the errors in the record. For example, by calculating an average distance of the estimated True Value to the measurements. However, the proposed True Values were shown to be unstable for records with large τ_{111}, which are the measurements that contain the largest deviations across measurements. Therefore, the True Values cannot be used to order the records in terms of severity of errors.

6 Conclusion

In response to the main research question:

What are the advantages and disadvantages of implementing the Intermittent-Error Model to assess and improve the quality of financial business data at Statistics Netherlands?

A shift has taken place among National Statistical Institutes (NSIs) from collecting data with sample surveys, designed by the institute and carried out on a well-defined population, to obtaining data from already existing external registers. As a result, NSIs benefit from the fact that register owners such as the Tax and Customs Administration have the means to obtain values on an entire population of businesses and are able to require complete information due to business' tax reporting obligations.

But external data is not always perfect for statistical purposes. Variable definitions can vary with the purpose of the register. For example, the purpose of a VAT tax register is to obtain information on the turnover and purchases of businesses in relation to the obligation to pay tax on the value added to a product, which is in terms of difference between prices of purchases and sales. Because businesses trading in second hand cars do not have a VAT tax component in the purchase of their cars, tax exemptions exist to the turnover value required for reporting. As a result, some of the reported turnover values do not agree with the statistical turnover concept that Statistics Netherlands is interested in, and cannot be used to publish economic statistics. Such variable definitions can also vary among registers maintained by the same register holder. For example, the register maintained to collect profit tax adopts a different turnover concept for these second hand car businesses than the VAT register does.

When register data is used to compile official statistics, the data collection and processing is outside the control of the NSI, and therefore errors in the obtained data are difficult to recognize. Fortunately obtaining already collected register data is costless and well-regulated for Statistics Netherlands, and therefore theoretically identical variables can be obtained on the same businesses from multiple sources. Multiple source values that describe the same business enable assessment of the available information on each business individually. When sources disagree, the presence of errors is recognized. However, the detection which exact value is erroneous and how it should be corrected is not always straightforward. Obtaining error-free values from multiple disagreeing sources requires a model to describe the observed values in relation to the True Values of interest.

The Intermittent-Error Model (IEM) is such a model. It is characterized by the intermittent nature of the error producing process, which distinguishes it from other approaches like Structural Equation Models that assume errors on all available values. The IEM allows measurements from source g to be erroneous with a certain probability (π_g) as well as being error-free (1 - π_g). Since a continuous error distribution is assumed on the erroneous values, error-free measurements can be identified from the data whenever two or more sources report the same value on the same unit (since the probability that two continuous error distributions produce the same error is negligible). By modelling the error distributions on the observations as well as a True Value distribution, the IEM is able to assign an error pattern to each combination of observations, such as (0, 1, 0) for three sources of which the second contains an error. Also, the IEM is able to propose an expected True Value in case all sources contain an error (error pattern (1, 1, 1) for three sources). The IEM assumes normally distributed errors on the observations, but is also able to capture systematic deviations in variable definitions. By estimating a deviation from the True Value in terms of intercept (a_g) and slope (b_g) for each source g, deviations such as in the VAT definition of turnover for second-hand cars can be accounted for.

The assumed normal distributions on the errors unexplained by the systematic deviations beg the question whether errors in multi-source data are well described by normal distributions. This study's investigation into 2012 financial business data from the VAT register, the Profit Declaration register

and a Statistics Netherlands survey showed that possible errors can be of a variety of different types in terms of origination. Not only do errors arise by incorrect reporting of values by businesses or incorrect processing by register owners or Statistics Netherlands, the defined population of business units might also contain errors. When data from multiple sources is linked to the units in the population framework, incorrect linkage can result in large deviations in values from different sources.

Defining a population of business units from in multiple sources is challenging. Companies are sometimes composed of multiple registered legal and fiscal units, to be able to shield reliability in risky projects or minimize tax obligation. Therefore, constructing a statistical population of business units from these fiscal and legal units is a complicated process that might introduce errors in the data. When separate companies are incorrectly combined into one business unit, the deviations in these units' values can become very large and might not meet the normality assumption on the source errors. Such large deviations in values were in fact observed in the 2012 case study data set of financial business data.

The Intermittent-Error Model was fitted to this case study data set twice, once including the severest deviations in turnover values from different sources and once excluding them, since they were considered outliers with regard to the normal error assumption. The estimated model parameters showed large sensitivity towards the inclusion of these outliers, even in subsets of businesses for which outliers composed only 1% of the data records. Further research into the stability of the model fit, by a nonparametric bootstrap, showed that the expected error proportions in some of the sources (π_{SBS} and π_{PDR}) were unstable in general. In about half of the bootstrap samples an error proportion in the range 0.1-0.3 was estimated for one source, while in the other half the estimation ranged between 0.5 and 0.7. These very different parameter estimates were obtained because the algorithm shifted the error pattern of a large number of businesses between (0, 1, 1) and (1, 1, 0). For these businesses, it does not follow directly from the data which one of the three values is correct and the unstable estimated error proportions showed that the IEM's decision was uncertain. As a result of different error pattern assignments, the source error variances were also unstable which in turn caused unstable True Value estimates.

Since the estimated error proportions and True Values are unstable, the Intermittent-Error Model cannot be used to assess and improve the quality of financial business data at Statistics Netherlands. Because the assigned error patterns are unstable for some businesses, uncertainty arises with regard to which of the source observations is erroneous and which is error-free. Meanwhile, the estimated systematic deviations in variable definition per source (a_g and b_g) relate to these assigned error patterns and as a result, are also unstable. So the model estimates cannot be used to estimate either source quality or the quality of records for all individual businesses. The estimated True Values were extremely unstable for businesses that were likely to have errors on all available observations, producing differences in estimated values of more than a factor 10. Thus the True Values proposed by the Intermittent-Error Model are not suitable to improve all observed data.

Hence, in its current form the application of the Intermittent-Error Model on financial business data at Statistics Netherlands is as puzzling as Schrödinger's cat: After you have observed the parameter estimates you can still contest the interpretation to the extent that certain observations were incorrect (dead) or error-free (alive) to begin with.

7 Future research

7.1 Designing a stable model

Future Research into an improved version of the Intermittent-Error Model needs to solve the instability of the parameter estimates and error pattern assignments. Three reasons for this instability can be considered: (1) The EM-algorithms Q-function creates local maxima. (2) The model fits badly and is therefore indifferent to some component assignments, thus the model fit is creating multiple modes in the likelihood function/Q-function. (3) The parameter estimates for the expected error proportions have very large standard errors.

Simulations with data that was perfect under the Intermittent-Error Model assumptions showed that the EM estimation procedure was able to estimate the true parameters. The distributions on the parameter estimates seemed normal and simulation's expected values were close to the true parameters. Therefore, (1) cannot be the sole reason for the model's instability but still can play a role with regard to data with less perfect fit, since the EM algorithm is known to be a 'greedy' algorithm likely to find local maxima (Figueiredo & Jain, 2002). But the starting values in the Intermittent-Error Model's estimation procedure are not arbitrary. These starting values are based on the observation triplets for which the True Value is known and therefore follow directly from the data. In case the Q-function creates local maxima, these starting values can be considered to be the best possible starting point to reach the right maximum.

Robinson (2016, pp. 47-48) implemented a method to obtain standard errors on the Intermittent-Error Model parameters. This method is based on the observed Fisher information matrix and adapted to be accurate for incomplete data. The approach was described by Little & Rubin (2002) and applies a 'supplemented' EM algorithm (SEM) that corrects the parameter variances from the Fisher information matrix by a measure of EM algorithm convergence. Robinson (2016, p. 49) showed that, in a perfect fit data simulation, 95% confidence intervals based on these standard errors were able to capture the true parameter value for about 95% of the parameters. Robinson's estimations on real data showed standard errors of around 0.01 for the π_g parameters which is in disagreement with hypothesis (3) of large variances causing the instability of the π_g parameter estimates. When the SEM approach was used to estimate standard errors on the case study data fit parameter estimates, the result was not at all in agreement with the variability in parameter estimates found in the bootstrap samples. This indicates that the main assumption of the SEM approach, which is sufficient agreement with the model assumptions, was not met. Therefore, the main hypothesis for the model's instability is the insufficient model fit with regard to the model assumptions (2) (described in Section 2.2).

Fit statistics on the case study model fit showed deviations from normality in the conditional True Value distributions as well as the source error distributions. However, simulations with noisy covariates and non-normal True Values (but normally distributed source errors) showed that insufficient model fit with regard to the True Value distribution does not result in unstable parameter estimates. Therefore, the robustness towards non-normal source errors should be the main focus of future research into model improvements.

7.2 Reconsidering the meaning of the assumed 'True Value'

The Intermittent-Error Model, in the way it is currently implemented, assumes a latent True Value and allows for incorrect measurements on this True Value in terms of random error as well as systematic deviations. Since also the values from the Structural Business Statistics survey carried out by Statistics Netherlands are allowed to show systematic deviations from the True Value, the question

arises what this True Value actually represents. The parameter estimates on the intercept and slope bias for the SBS values (a_{SBS} and b_{SBS}) showed parameter values across NACE groups for the log transformed data that were similar to the estimates for the VAT and PDR (see Table 4.2 on page 65). This indicates that the model does not recognize that the nature of reporting differences in relation to the assumed True Value are more specific to administrative register data. A version of the Intermittent-Error Model can be researched that does not assume intercept and slope bias on the SBS values ($a_{SBS} = 0$ and $b_{SBS} = 1$, but allowing for random $\epsilon_{SBS,i}$) and accordingly redefines the meaning of the assumed True Value. This might create a certain preference for the True Value to be close to the SBS values, since the SBS measurements have less uncertain parameters that relate them to the True Value. As a result of this more limited approach to the True Value, the overall distribution on the data is more restricted and the error pattern assignments might become more stable.

7.3 Narrowing the model's use

The data contained many severe errors, that are not likely normally distributed measurement errors. In general, some error types that can be encountered in financial business data are not in agreement with the normal distribution, like severe processing errors, comparability errors, identification errors and unit errors. When these can be corrected by another procedure, the Intermittent-Error Model could be applied to handle the remaining errors. Within the current editing procedures at Statistics Netherlands, data with only measurement errors might be available on TopX businesses. This was shown by the absence of TopX businesses among the outliers and the known thorough procedure for these businesses. Whether the Intermittent-Error Model fit on only TopX businesses meets the model assumptions should be further investigated.

7.4 Expanding the possibilities of the implementation

Especially when the use of the model is narrowed, the amount of data to fit the model needs to be expanded. In its current form, the Intermittent-Error Model can only be fitted on businesses for which values are available from three sources. This resulted in data that constituted only a small subset of the total businesses in the GBR, mainly due to the SBS sampling approach (see Figure 1.8 on page 24). Therefore, a large improvement to the applicability of the Intermittent-Error model would be to also be able to deal with records for which values are available from two sources. There is no restriction in the model definition with regard to two sources, only the current estimation procedure needs to be updated to be able to handle both three and two sources among the records. This update would naturally fit within the same EM-estimation approach since the two out of three available sources constitute another missing data problem.

7.5 Allowing for correlated errors

The Intermittent-Error Model assumes that measurements are independent conditional on the True Value, and therefore that the source errors are uncorrelated. This assumption might be unreasonable since a business might look up an estimated turnover value that it previously reported to pay VAT taxes, and use it to respond to a Statistics Netherlands survey. Therefore, the error in the VAT turnover value, as a result of the business estimating the value and not supplying a precise value, is related to the error in the SBS turnover. Future research can focus on certain relaxations of the source independence assumption (described in Section 2.2).

8 References

Aelen, F. (2005). *Startcursus Methodologie - Module 2: Bedrijvenregister*. Den Haag/Heerlen: CBS (internal document).

Aelen, F., Ariel, A., Boonstra, H., Van der Loo, M., Pannekoek, J., Tennekes, M., & De Wolf, P.-P. (2011). *Methodologie Nieuwe Opzet Productiestatistieken.* Den Haag/Heerlen: CBS (internal document).

Bakker, B. F. (2009). *(Oratie) Trek alle registers open!* Amsterdam: Vrije Universiteit.

Bakker, B. F. (2011). *Statistical Methods - Micro Integration*. The Hague/Heerlen: Statistics Netherlands.

Belastingdienst. (2007). *Integrale probleemanalyse loonaangifteketen*. Belastingdienst en UWV.

Belastingdienst. (2016a). *Business*. belastingdienst.nl. Retrieved 9 19, 2016, from http://www.belastingdienst.nl/wps/wcm/connect/bldcontenten/belastingdienst/business/

Belastingdienst. (2016b). *Uitstel aangifte vennootschapsbelasting*. belastingdienst.nl. Retrieved 9 28, 2016, from http://www.belastingdienst.nl/wps/wcm/connect/bldcontentnl/belastingdienst/zakelijk/winst/vennootschapsbelasting/uitstel_aangifte_vennootschapsbelasting/uitstel_aangifte_vennootschapsbelasting

Belastingdienst. (2016c). *Fiscale Eenheid BTW*. belastingdienst.nl. Retrieved 10 14, 2016, from http://www.belastingdienst.nl/wps/wcm/connect/bldcontentnl/belastingdienst/zakelijk/btw/hoe_werkt_de_btw/voor_wie_geldt_de_btw/fiscale_eenheid/fiscale_eenheid

Belastingdienst. (2016d). *Fiscale eenheid vennootschapsbelasting*. belastingdienst.nl. Retrieved 10 14, 2016, from http://www.belastingdienst.nl/wps/wcm/connect/bldcontentnl/belastingdienst/zakelijk/winst/vennootschapsbelasting/fiscale_eenheid_vennootschapsbelasting/fiscale_eenheid_vennootschapsbelasting

Belastingdienst. (2016e). *Ondernemers die onder de landbouwregeling of veehandelsregeling vallen*. belastingdienst.nl. Retrieved 10 14, 2016, from http://www.belastingdienst.nl/wps/wcm/connect/bldcontentnl/belastingdienst/zakelijk/btw/administratie_bijhouden/wie_moeten_een_btw_administratie_bijhouden/ondernemers_die_onder_de_landbouwregeling_of_veehandelsregeling_vallen/ondernemers_die_onder_de_land

Belastingdienst. (2016f). *Kleineondernemersregeling*. belastingdienst.nl. Retrieved 10 14, 2016, from http://www.belastingdienst.nl/wps/wcm/connect/bldcontentnl/belastingdienst/zakelijk/btw/hoe_werkt_de_btw/kleineondernemersregeling/kleineondernemersregeling

Boersma, J. (2009). *PS6064gaafmaakinstructies versie 8.0*. Den Haag/Heerlen: CBS (internal document).

Box, G. E. (1979). Robustness in the strategy of scientific model building. In R. L. Launer, & G. N. Wilkinson, *Robustness in Statistics* (pp. 201–236). New York: Academic Press.

Box, G. E., & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. John Wiley & Sons.

CBS. (2003). *Wet op het Centraal Bureau voor de Statistiek*. Staatsblad van het Koninkrijk der Nederlanden, 516.

CBS. (2016a). *About Us - Organisation*. cbs.nl. Retrieved 11 11, 2016, from https://www.cbs.nl/en-gb/about-us/organisation

CBS. (2016b). *Documentatierapport Productiestatistieken 2014*. CBS. Retrieved 9 12, 2016, from https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/ps-autohandel-2009-2014

CBS. (2016c). *Documentatierapport Aangifte Omzetbelasting (BTW)*. CBS. Retrieved 9 12, 2016, from https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/microdatabestanden/btw-2007-2014

CBS. (2016d). *SBI 2008 versie 2016 Engels*. Den Haag/Heerlen: CBS. Retrieved 10 5, 2016, from https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/activiteiten/standaard-bedrijfsindeling--sbi--/sbi-2008-standaard-bedrijfsindeling-2008

CBS. (2016e). *Productie Statistieken*. cbs.nl. Retrieved 10 5, 2016, from https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/productiestatistiek

CBS. (2016f). *Populatie van actieve bedrijven in Nederland*. cbs.nl. Retrieved 11 22, 2016, from https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/populatie-van-actieve-bedrijven-in-nederland

Daas, P., & Van Delden, A. (2014). Collection and Use of Secondary Data. In *Memobust Handbook on Methodology of Modern Business Statistics*. Eurostat.

De Wolf, P.-P., & Van Delden, A. (2011). *Methodenbeschrijving KS en SB voor het MKB, Versie 2*. Den Haag/Heerlen: CBS (internal document).

Eurostat. (2008). *Statistical Classification of Economic Activities in the European Community*. Retrieved 9 19, 2016, from https://www.scribd.com/document/49709130/Eurostat-NACE

Eurostat. (2010). *List of NACE codes*. Retrieved 9 2016, 19, from http://ec.europa.eu/competition/mergers/cases/index/nace_all.html

Figueiredo, M. A., & Jain, A. K. (2002). Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on pattern analysis and machine intelligence*, 381-396.

Groen, J. A. (2012). Sources of error in survey and administrative data: The importance of reporting procedures. *Journal of Official Statistics*.

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

Guarnera, U., & Varriale, R. (2015). Estimating and editing for data from different sources - An approach based on latent class model. *Emerging methods and data revolution* (pp. 1-9). Budapest, Hungary: UNECE Work Session on Statistical Data Editing.

Guarnera, U., & Varriale, R. (2016). Estimation from contaminated multi-source data based on latent class models. *Statistical Journal of the IAOS*, 537-544.

Hoogland, J. (2005). *Startcursus Methodologie - Module 7: Gaafmaken*. Den Haag/Heerlen: CBS (internal document).

Hoogland, J., Van der Loo, M., Pannekoek, J., & Scholtus, S. (2010). *Methodenreeks: Thema: Controle en Correctie*. Den Haag/Heerlen: CBS.

Hox, J. J., & Boeije, H. R. (2005). Data collection, Primary vs. Secondary. In *Encyclopaedia of Social Measurement Vol. 1* (pp. 593–599).

Konen, R. (2012). *Methodenreeks - Thema: Statistische eenheden in de institutionele statistieken*. CBS.

KvK. (2016a). *What does the Business Register contain?* kvk.nl. Retrieved 9 19, 2016, from https://www.kvk.nl/english/business-register/what-does-the-business-register-contain/

KvK. (2016b). *Inschrijven bij de Kamer van Koophandel*. kvk.nl. Retrieved 9 21, 2016, from https://www.kvk.nl/inschrijven-en-wijzigen/inschrijven-bij-de-kamer-van-koophandel/moet-ik-mijn-bedrijf-inschrijven/

Lammertsma, A. (2016). *PS-omzet en DRT-omzet: systematische verschillen o.b.v. BTW-regelingen?* Den Haag/Heerlen: CBS (internal document).

Little, R., & Rubin, D. (2002). *Statistical Analysis with Missing Data*. Wiley Interscience.

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models*. Wiley Interscience.

Meijers, R., & Smeets, M. (2011). *Steekproefontwerp Productiestatistiek en Statistiek Investeringen en Lease 2010*. Den Haag/Heerlen: CBS (internal document).

Rademakers, F. (2005). *BaseLine: Ontsluiting - Dekking - Vulling - Tijd*. Den Haag/Heerlen: CBS (internal document).

Robinson, S. (2016). *Modelling Measurement Errors in Linked Administrative and Survey Data*. Den Haag/Heerlen: CBS/Leiden University Master Thesis. Retrieved from http://www.math.leidenuniv.nl/en/theses/year/2016/

Scholtus, S., Bakker, B. F., & Robinson, S. (2016 (preprint)). Assessing the Quality of Business Survey Data before and after Automatic Editing. In S. Scholtus, *Editing and Estimation of Measurement Errors in Statistical Data (PhD Thesis)*. Den Haag/Heerlen: CBS.

Scholtus, S., Bakker, B. F., & Van Delden, A. (2015). *Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables*. Den Haag/Heerlen: CBS.

Van Delden, A. (2013). *Startcursus Methodologie - Module 2: Registers en andere secundaire bronnen*. Den Haag/Heerlen: CBS (internal document).

Van Delden, A., & De Wolf, P.-P. (2013). A production system for quarterly turnover levels and growth rates based on VAT data. *New Techniques and Technologies for Statistics (NTTS) conference*. Brussels.

Van Delden, A., Pannekoek, J., Banning, R., & De Boer, A. (2016). Analysing correspondence between administrative and survey data. *Statistical Journal of the IAOS*, 569-584.

Van Delden, A., Scholtus, S., & Burger, J. (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics*, 619-642.

Van Delden, A., Scholtus, S., De Wolf, P.-P., & Pannekoek, J. (2014). *Methods to assess the quality of mixed-source estimates*. The Hague/Heerlen: SBS.

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