VRIJE UNIVERSITEIT

Editing and Estimation of Measurement Errors in Administrative and Survey Data

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Sociale Wetenschappen
op dinsdag 6 maart 2018 om 9.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Sander Scholtus

geboren te Den Haag

promotoren:	prof.dr. B.F.M. Bakker
	prof.dr. H.B.G. Ganzeboom
copromotor:	prof.dr. C.H. Elzinga

And the stories you hear, you know they never add up
I hear the natives fussing at the data chart

Pavement – *Frontwards* (1992)

La solution était évidente, aussi évidente que le problème avait semblé insoluble tant qu'il ne l'avait pas résolu

Georges Perec – *La Vie mode d'emploi* (1978)

Table of contents

Chapter 1

Setting the Problem

1.1 Introduction

It is the task of national statistical institutes (NSIs) and other statistical agencies to produce statistical information on social and economic aspects of society, for users such as policymakers, researchers, journalists and the general public. These official statistics have an impact on everyday life in many different ways: they are used by policymakers to make informed decisions and evaluate their effects, by the government to allocate resources, by interest groups and unions as a basis for negotiations, and by journalists and "fact checkers" to substantiate the points that they are trying to make. It is therefore important that the quality of official statistics is high.

Data that are collected for the production of official statistics or, more generally, for statistical analyses nearly always contain measurement errors. NSIs, other statistical agencies and academic researchers have therefore developed methods to handle error-prone data. Two broad classes of approaches can be distinguished: methods that aim to reduce the effects of measurement errors by adjusting the data themselves and methods that try to correct for measurement errors at the analysis stage.

The first strategy is widely used in official statistics, where it is known as *data editing* (De Waal et al., 2011); other terms are also used, such as *data cleaning* and *data validation*. The second approach involves *estimating* a model for measurement errors, either in a separate step or as part of the analysis model itself. This strategy is more commonly adopted by researchers working outside official statistics. This is not to say that these approaches are mutually exclusive. In fact, most researchers apply at least some basic form of data editing, for instance to detect outlying observations. So far, applications of error modelling in the production of

official statistics have been rare, but one example is provided by Van Delden et al. (2016).

In fact, a third approach should also be mentioned: designing the data collection methodology to *prevent*, as much as possible, measurement errors from occurring in the first place. Much research has been done both within and outside official statistics to find strategies that are likely to yield correct responses to questionnaires and other data collection instruments; see, e.g., Presser et al. (2004) and Brancato et al. (2006). Obviously, it is important to understand which design choices for data collection methodology work and which do not. In practice, however, it seems inevitable that some measurement errors do occur whenever data are collected. In this thesis, we will focus on methods that can be used to address the problem of measurement error after the data have been collected.

The aim of this thesis is to contribute to the development of editing and estimating methods for dealing with measurement errors, with a particular focus on their extension and application to large data sets from administrative sources. We will also point out commonalities of the two approaches, by discussing the implicit measurement error models behind certain data editing methods. This might help to make the use of those editing methods more acceptable to academic researchers. Conversely, we will discuss how to extend some existing measurement error models so that they can be applied in an official-statistics context.

In particular, this thesis aims to make progress on the following points. Firstly, current methods for automatic data editing have limited applicability, because they are based on rather restrictive assumptions. We will develop two new methods for automatic editing of survey or administrative data that relax some of these assumptions. In this way, the practical applicability of automatic editing increases. By increasing the use of automatic methods for data editing at NSIs, the use of other, more costly and labour-intensive methods for data editing that involve manual work can be reduced. Moreover, the focus of these manual methods could then be shifted to the most difficult cases, where their contribution is most likely to be of use. Thus, increasing the use of automatic data editing at NSIs can lead to statistical production processes that are more efficient and yield statistical output of higher quality.

Secondly, applications of measurement error estimation in the social sciences usually focus on the effect of errors on bivariate and multivariate relations (e.g., correlations or regression coefficients). In official statistics, univariate statistics such as population totals and means are often of interest (Skinner et al., 1989). While, in principle, existing measurement error models can be used to estimate the effect of errors on these univariate statistics, the design choices that are traditionally

made for these models in social-science applications are not suitable in this context (e.g., the use of arbitrary reference indicators to obtain model identification). In this thesis, we will apply two different models to study the effects of measurement errors on official statistics: a structural equation model and a finite mixture model. In addition, we will discuss alternative design choices that are appropriate in this context. In this way, an important obstacle is removed for the wider use of measurement error models in official statistics. As will be shown below, models for measurement errors can be used in official statistics to assess the suitability of new data sources. Moreover, they can be used to improve the quality of statistical output and to gain better insight into the accuracy of statistics.

Thirdly, it is known that in practice automatic data editing methods do not resolve all measurement errors in a data set. Traditionally, the quality of automatic editing methods has been evaluated by comparing them to manual editing, under the assumption that manually-edited data are error-free. This assumption is unlikely to hold in practice. To evaluate the effects on statistical output of measurement errors that remain in the data after editing, and also to compare the amount of measurement error before and after editing, we will use a measurement error model which does not require the assumption that manually-edited data are always error-free. We will also discuss how such a model could be used in practice to improve the way official statistics are produced, both in terms of efficiency and in terms of accuracy of the statistical output.

The remainder of this introductory chapter is organised as follows. Section 1.2 provides some background information and reviews some terminology: types of data sources, types of errors in statistics, types of variables and types of measurement scales. The editing and estimation approaches to measurement errors are introduced briefly in Section 1.3 and Section 1.4, respectively. Having established this context, we give an outline of the rest of this thesis in Section 1.5.

1.2 Types of data and types of errors

1.2.1 Survey data and administrative data

Statistical agencies and empirical researchers need data to generate statistical information. In the past, the required data were usually not yet available and had to be created by conducting a *survey*, often by means of a questionnaire. We refer to Bethlehem (2009), De Leeuw et al. (2008) and Groves et al. (2009) for a detailed introduction into survey methodology based on questionnaires.

In somewhat idealised terms, the survey process consists of the following steps. First, a research question is formulated. This question is made more precise by

defining a target population and a parameter of interest. These concepts are then made operational: the target population is translated into a surveyable population of units and the parameter of interest is translated into one or more measurable properties of these units. For each operational property, one or more questions are constructed to measure it for a unit in the survey population, by eliciting a response from either the unit itself or from someone else about the unit. These questions are collected in a questionnaire. In theory, the survey objective could then be fulfilled by obtaining responses to this questionnaire about each unit in the survey population. In practice, a complete enumeration (or *census*) of a survey population is rarely taken because this is expensive, time-consuming and burdensome to society. Moreover, statistical practice has shown that, for most purposes, sufficiently accurate estimates of the parameters of interest can be obtained by surveying only a fraction of the population (a *sample*), if this sample is selected randomly from the survey population. Ideally, responses to the questionnaire are obtained for all units in the sample. Sample survey theory can then be used to obtain estimates of parameters of interest from the sample data, as well as measures of accuracy for these estimates (bias, variance, and confidence intervals). The first satisfactory treatment of sample survey theory was given by Neyman (1934); good overviews of the subject can be found in Cochran (1977), Särndal et al. (1992) or Knottnerus (2003).

In practice, complications arise for all of the above-mentioned steps in the survey process. For instance, surveys are hardly ever conducted with a single research question in mind; rather, the aim is to answer multiple research questions or even to construct a general-purpose data set about units in a population, which may then be analysed by various researchers to answer various questions. At the operationalisation stage, some compromises usually have to be made with respect to the definitions of the target population and parameters of interest, in order to obtain a survey population that can effectively be sampled and variables that can effectively be measured. For instance, the ideal target population of a demographic survey might include undocumented immigrants, but these are very difficult to sample and therefore often excluded from the survey population (Bethlehem, 2009).

Various *survey modes* can be used to obtain information from potential respondents (Bethlehem, 2009). An interviewer can visit the respondent and record his/her answers face-to-face, or an interview can be conducted by telephone. In the Netherlands and other developed countries, these interviews are nowadays nearly always done with the aid of a computer (e.g., a laptop or tablet in the case of face-to-face interviewing). In this case, the above survey modes are commonly referred to as Computer-Assisted Personal Interviewing (or CAPI) and Computer-Assisted

Telephone Interviewing (or CATI), respectively. Alternatively to these interview-based modes, a so-called self-administered questionnaire can be filled in by the respondent without an interviewer being present. This can be done on paper (mail survey) or online (commonly referred to by the misnomer Computer-Assisted Web Interviewing or CAWI).

Modern surveys often use a mixture of modes to improve response rates and decrease costs (De Leeuw, 2005). In fact, in any given survey only a subset of the units in the original sample will respond. Nonresponse occurs for various reasons: some units cannot be contacted, other units are contacted but refuse to co-operate, still other units are willing but unable to co-operate, etc. See, e.g., Bethlehem et al. (2011) for an overview of the non-response problem, the effects it can have on the accuracy of survey estimates, and methods that attempt to reduce or correct for nonresponse.

Traditional surveys can be expensive and burdensome to respondents. Over the past decades they have also suffered from decreasing response rates (or increasing efforts being required to obtain the same response rates) in many countries (Stoop, 2005), which makes them less attractive as a data collection method. At the same time, the arrival of the digital age means that all kinds of non-survey data are now generated on a regular basis by businesses, governmental agencies and other institutes for their own administrative purposes. NSIs and other statistical agencies have started to use these administrative data more and more for official statistics, first as auxiliary information to improve estimates based on survey data, and subsequently as a replacement for survey data (Zhang, 2012). A similar trend occurs for academic research in the social sciences (Bakker and Kuijvenhoven, 2010).

Statistics Netherlands is now using administrative data in the production of several important statistics, both in the social and economic domain. For instance, in the two most recent population censuses in the Netherlands (2001 and 2011) nearly all variables were obtained from administrative sources (Schulte-Nordholt et al., 2004, 2014). For the estimation of the 2011 census tables, sample survey data were used only for the variables occupation and educational attainment. For the quarterly economic statistics on turnover (which in turn are an important component of the Gross Domestic Product), a new production system was introduced in 2011 which uses data of value-added tax (VAT) declarations submitted by businesses to the Netherlands' tax authorities, supplemented by a small census survey of the largest and most complex businesses (Van Delden and De Wolf, 2013).

The advantages of using administrative rather than survey data for statistics are fairly obvious [see, e.g., Wallgren and Wallgren (2014)]. Using data that are already available in administrative registers is more efficient than collecting new

data by means of questionnaires. It also removes the burden on individual persons or businesses to provide responses to questionnaires. At the same time, this makes the quality of statistical estimates less dependent on the willingness and ability of individual potential respondents. In some cases, administrative data can be used to obtain information about phenomena that are difficult to observe accurately in surveys, such as criminal behaviour (Bakker and Kuijvenhoven, 2010). Finally, since administrative data sources often contain a complete enumeration of their target population and are available over time, these data may provide opportunities to do detailed statistical analyses that would be difficult, costly or impossible with survey data based on relatively small samples, such as estimates for small sub-populations and certain types of longitudinal analyses.

Administrative data can also have disadvantages. Most of these have to do with the fact that a statistical agency has less control over the way these data are collected than with survey data. In fact, it is important to realise that administrative data are originally collected and processed by a register owner for some non-statistical purpose (Bakker and Daas, 2012; Wallgren and Wallgren, 2014). In many cases, the concepts and definitions that are used by the register owner differ to some extent from the intended definitions for statistical purposes. In the above VAT example, the original purpose of the tax declaration data is to levy taxes on turnover. Thus, the tax authorities attempt to measure the total amount of taxable turnover for each unit as well as possible. This is not necessarily identical to the total amount of turnover that is needed for economic statistics; some economic activities may be exempt from taxes but relevant to the size of the economy from a statistical point of view, and vice versa (Van Delden et al., 2016).

By re-purposing an administrative data set for the production of statistics, a statistical agency is forced to work with administrative concepts, at least initially. In some (probably quite rare) applications, the administrative concept coincides with the statistical concept, and the administrative source can be seen as a "gold standard". One example is provided by Mittag (2013) who used administrative records of a food stamp program in the United States of America in a study that focussed on amounts of food stamps received.

A related potential problem with administrative data is that the register owner will aim to process these data in a way that is optimal for its own internal purposes (Bakker and Daas, 2012). Thus, the quality of different variables in the same administrative data set may vary according to how relevant these variables are to the register owner (Wallgren and Wallgren, 2014). We refer to Van Delden et al. (2014) for some examples in the Netherlands.

Another related problem is that, since the administrative data are primarily

used for a different purpose than making statistics, individual units may have an incentive to be registered in a certain way. For instance, in the VAT data set, most businesses will aim to report their turnover in a way that minimises the amount of tax to be paid, which makes under-reporting more likely than over-reporting (Van Delden et al., 2014; Wallgren and Wallgren, 2014).

It differs by country which administrative data sources exist and to what extent statistical agencies and researchers are given access to these data (Wallgren and Wallgren, 2014). For the most recent round of world-wide population censuses (circa 2011), administrative data were used as a direct data source (in some cases supplemented by survey data) in Austria, Belgium, Denmark, Finland, Germany, Israel, Latvia, Lithuania, the Netherlands, Norway, Slovenia, Sweden, and Switzerland (UN/ECE, 2014). In other countries, notably the United Kingdom and the United States of America, the use of administrative data for statistical purposes has been limited so far, due to a lack of suitable registers and/or legal difficulties in obtaining access to existing sources.

Administrative data sources have not replaced traditional surveys completely, and it is unlikely that this will happen in the future. Many "subjective" concepts that are of interest in the social sciences, such as attitudes, cannot be found in any register. Surveys therefore remain necessary to collect information about these concepts. In practice, to obtain a data set with all variables that are needed for a particular application, it is often necessary to combine multiple administrative data sources, or to combine administrative and survey data. The problem of linking multiple data sources together and performing statistical inference on these linked data is known as *data integration* (Zhang, 2014).

1.2.2 Errors in statistics

Statistical statements based on real data are subject to many different types of error. Groves (1989) and Bethlehem (2009) developed taxonomies of errors in estimates based on survey data. Bakker (2011b) and Zhang (2012) extended Groves' taxonomy to administrative data and integrated data, respectively.

With the exception of Bethlehem, these authors all made a distinction between errors along the *representation* side and errors along the *measurement* side. Representation errors occur when the set of observed units deviates from the intended target population. These include: *coverage error*, because the survey population as listed in the sampling frame does not include all units in the target population (*undercoverage*) or includes some units that do not belong to the target population (*overcoverage*); *sampling error*, because the sample includes only a subset of the survey population; *nonresponse error*, because not all sampled units are observed.

For probability samples, the sampling error takes the form of stochastic estimation uncertainty. The properties of this uncertainty are well understood and its size can be controlled under the sampling design. The effects of the other representation errors on estimates are more difficult to quantify, but it is reasonable to assume that all statistical estimates are affected by them to some extent. In the case of integrated data, *linkage error* caused by mislinks or missed links of units between different data sets is also an important source of representation error.

In the remainder of this thesis, we will focus on measurement errors. The term "measurement error" will be given a more precise meaning when we consider measurement error models in Section 1.4. For now, we use the intuitive notion of a measurement error: a deviation of an observed value in a data set from its "true" value.

First of all, measurement errors can arise due to differences between the definition of a conceptual variable of interest and the way it is operationalised. In particular, this often happens with administrative data because the administrative concept differs from the statistical concept (e.g., taxable turnover for VAT purposes versus turnover from statistically relevant economic activities). But definitional errors can also occur for surveys, because the ideal conceptual variable of interest might be too difficult to measure in a questionnaire and is therefore replaced by a simpler, approximating concept.

In surveys, measurement errors also occur during data collection. A response to a survey question can be seen as the result of a complex cognitive process (Tourangeau et al., 2000; Bavdaž, 2010). Respondents can misunderstand a question, make a mistake in their answer or deliberately provide an erroneous answer. The latter problem arises in particular for questions about "sensitive" subjects, such as fraud or drug addiction. Even for non-sensitive subjects, it has been found that responses to survey questions are affected by the way these questions are formulated: different wordings of a question or (for a multiple-choice question) different sets of possible answer categories generally produce different response distributions (Saris and Gallhofer, 2007; Bethlehem, 2009). Under the assumption that a unique true response exists for each respondent (see also the next subsection), this finding implies that different question wordings are subject to different amounts of measurement error. Measurement errors can even arise after data collection during further data processing – for instance, when paper questionnaires are scanned and digitised by optical character recognition (De Waal et al., 2011).

For some administrative sources, the data collection and storage process is very similar to that of surveys. For instance, the above-mentioned VAT declarations are submitted by businesses to the tax authorities on a regular basis (monthly, quar-

terly or yearly), using a paper form or a web form. This data collection process is comparable to that of a mail or CAWI survey; in fact, a tax form looks very similar to a survey questionnaire. It therefore seems plausible that similar measurement errors occur during data collection (Bakker, 2011a). Wallgren and Wallgren (2014, pp. 29–30) suggested that measurement errors in administrative data are influenced more by accounting principles and legislation than by cognitive processes, although they did not deny that such processes exist also for administrative reporting.

Other administrative sources are less similar to surveys, because data are generated and stored continuously rather than at fixed intervals. These registers are longitudinal and "event-driven": the data of each unit remain fixed until a new event occurs. As an example, consider the Base Registration of Persons (the population register of the Netherlands) which is maintained by municipalities. Here, the "events" that prompt changes in the register are events that occur in the lives of persons living in the Netherlands (births, deaths, marriages, moves, etc.). Persons report these events to the municipality where they are registered and the register is updated. So-called administrative delays, in which an event is registered some time after it actually occurred, are an important source of measurement error in this situation (Bakker, 2011a). The registration process itself could be compared to that of a CAPI survey, with municipal civil servants acting as "interviewers" for their residents. An important difference with survey data is that a version of the data set can be obtained at any desired time point. However, the "event-driven" nature of the register implies that data at closely-spaced time points will be correlated strongly.

For yet other administrative sources, the data collection process is not comparable to that of any survey mode. Some NSIs – including Statistics Netherlands – are currently experimenting with a new form of data collection, by which data are extracted directly from the accounting systems of businesses (Snijkers et al., 2016). A crucial feature of this approach are the links between the statistical variables of interest and the variables in the accounting data. In the Netherlands, these links are established through the so-called Reference Classification System of Financial Information (in Dutch: *Referentie Grootboek Schema*) and have to be set up manually by the reporting unit the first time the system is used. This manual translation between statistical and administrative concepts is now the main – if not the only – source of measurement error: any erroneous link will systematically affect all future data extractions. On the other hand, this approach leaves virtually no room for random response errors in the individual data extractions, as these are entirely automated.

Representation and measurement errors can have either a systematic or a random effect on a statistical estimator $\hat{Y}$, or both. The net effect of all errors can be summarised in terms of the *mean squared error* (MSE) of the estimator, which is defined as its expected squared deviation from the true parameter value Y:

$$\text{MSE}(\hat{Y}) = E(\hat{Y} - Y)^2.$$

The MSE can be decomposed into *bias* and *variance*:

$$\begin{aligned} \text{MSE}(\hat{Y}) &= \{E(\hat{Y}) - Y\}^2 + E\{\hat{Y} - E(\hat{Y})\}^2 \\ &\equiv \{\text{bias}(\hat{Y})\}^2 + \text{var}(\hat{Y}). \end{aligned} \tag{1.1}$$

The bias component captures the net effect of all systematic error contributions and the variance captures the net effect of all random error contributions.

In practice for traditional sample surveys, often only the sampling error is taken into account when estimating the MSE, as this is relatively easy to do. It is then tacitly assumed that the sampling error dominates the contributions of all other representation errors and measurement errors. For census surveys and for estimators based on administrative data or integrated data, this assumption is not tenable. Even for many sample surveys, it is adopted more out of convenience than because of its plausibility. Ideally, all types of error should be taken into account in the MSE, which then becomes a measure of the so-called *total survey error*. The problem of quantifying the total survey error in practice is discussed by, among others, Groves (1989) and Biemer and Lyberg (2003). In particular, the contribution of measurement errors in the observed data to the bias and variance of an estimator can be evaluated as part of the total survey error.

1.2.3 True scores and true values

An observed variable is the result of applying some measurement procedure to each unit. This measurement procedure might involve, for instance, obtaining a response to a question on a survey form, or obtaining a value from an administrative report. Biemer (2011) distinguished between the *true score* and the *true value* of an observed variable for a given unit. The true score is the "average" value that we would expect to find for a unit under the measurement procedure (in a sense that will be made more precise in Section 1.4). This true score might be different from the true value of the concept that we are trying to measure for that unit. As a simple example, suppose that we measure the height of a person by a very crude procedure: using a stick that is known to be exactly one metre long, we simply count the number of whole "stick lengths" that best approximate the

person's height. Under this procedure, the true score of a given person might be "2 metres", whereas his actual height (true value) may be, for instance, "1.85 metres". Under different measurement procedures that aim to measure the same concept for a given unit, the true scores may vary but the true value remains the same.

Biemer's true values and true scores correspond to what are known as, respectively, *platonic true scores* and *operational* or *classical true scores* in the literature on measurement in sociology and psychology [e.g., Sobel and Arminger (1986); Biemer and Stokes (1991)]. It is often posited that platonic true scores exist only for some variables, namely those that can (in theory) be measured "objectively" by some other means than asking questions. Thus, for instance, a person's height clearly has a platonic true score which is defined independently of the way we choose to measure it. On the other hand, a person's intelligence (at the moment) can be measured only "subjectively" by means of a questionnaire (an intelligence test). For the latter type of variable, only an operational true score can be meaningfully defined, according to these authors.

Borsboom et al. (2003) objected to this distinction on philosophical grounds. In fact, most measurement models that are used in the social sciences (and in particular all models that will be considered in this thesis) are so-called *reflective models*, in which the outcome for an observed variable is explained by the underlying unobserved concept that is being measured. (This can be contrasted to *formative models*, in which a concept is constructed from several observed variables.) Borsboom et al. (2003) argued that the use of a reflective model is consistent only with the philosophical stance that the concepts that are being measured really exist. In particular, this implies that each unit must have a true value that exists independently of the measurement procedure, even for a variable such as intelligence that cannot be observed directly.

In most of the applications in this thesis, we will assume that we are dealing with variables for which true values not only exist but can also be observed, at least in principle. In particular, this assumption pervades the literature on data editing, although it is not always made explicit: that the true value exists and *could* be obtained for any unit if a sufficient editing effort were made – although in practice it is not feasible to make this effort for all units, given time and budget constraints. This assumption seems reasonable for many "factual" variables that are encountered in official statistics, such as turnover, income, etc.

1.2.4 Measurement levels

According to measurement theory, different levels of measurement or types of scale can be distinguished. Usually, five basic levels of measurement are defined: nom-

inal, ordinal, interval, ratio and absolute (Stevens, 1946, 1959). The measurement level determines which meaningful conclusions about a variable of interest can be inferred from the observed values (Sarle, 1997).

In the literature on data editing, a somewhat less formal distinction is often made between *categorical* variables and *numerical* (or *continuous*) variables; see, e.g., De Waal et al. (2011). This convention will be followed in the parts of this thesis on data editing (in particular Chapters 3–5). Categorical and numerical variables are conceptualised as having a finite and infinite number of possible values, respectively. [In fact, De Waal et al. (2011) assume that numerical variables are real-valued and treat integer-valued variables as a separate type.] Of course, data are always observed and processed with finite precision. In practice, the categorical variables are the ones that have a very limited number of possible values.

Unfortunately, this informal distinction between categorical and numerical variables does not correspond exactly to a distinction between traditional measurement levels. In practice, most numerical variables will be measured at the interval level or higher and most categorical variables will be nominal or ordinal variables, but it is possible to find counter-examples (Sarle, 1997).

The methods to be discussed in this thesis nearly always assume that one is dealing with numerical variables and/or variables that can be measured at the interval level or higher. The main exception occurs in Chapter 4 which deals explicitly with categorical as well as numerical variables. In particular, the measurement error models that will be considered in Chapters 6–7 are all designed for interval-or-higher-level variables. Measurement error models for nominal and ordinal variables are discussed, for instance, by Biemer (2011).

1.3 The editing approach

We will now review the two approaches for handling error-prone data that were introduced in Section 1.1: the editing approach and the estimation approach. We start with the editing approach.

Traditionally, NSIs and other statistical agencies have always spent much effort on checking and, if necessary, adjusting the data that they collect for measurement errors. Over time, a more or less standardised process for data editing has evolved (EDIMBUS, 2007; De Waal et al., 2011).

A prerequisite of any data editing method is that it is – in principle – possible to identify records that contain erroneous values. This requires some form of external information with which the observed data can be compared. An important type of external information consists of restrictions that would be expected to hold if the

data were error-free. These restrictions are often called *edit rules*, or *edits* for short. Examples of edit rules include:

- univariate range restrictions – for instance: "Turnover ≥ 0" (for businesses) or "Age (in years) $\in \{0, 1, \ldots, 120\}$" (for persons);
- balance edits – for instance: "Turnover $-$ Costs $=$ Profit";
- ratio edits – for instance: "Total Staff Costs / Number of Employees $\leq$ 100,000 Euros".

A distinction can be made between *hard edit rules* and *soft edit rules*. A hard edit must be satisfied by any error-free record; in other words, if a record does not satisfy a hard edit then we know that this record must contain at least one error. By contrast, if a record does not satisfy a soft edit then this means that the data are implausible (suspicious) but not necessarily incorrect. Of the above examples, "Turnover ≥ 0" and "Turnover $-$ Costs $=$ Profit" are hard edits (they have to hold by definition), while the other two are soft edits. For instance, it is unlikely for a person to be more than 120 years old, but not impossible.

It is straightforward to check each record in a data set against a given set of edit rules. Nowadays, such edit checks are nearly always automated. It is much less straightforward to work out how to adjust a given record of data that does not satisfy all edit rules. The end result should be an adjusted record that satisfies at least all hard edits, but there are typically many possible adjustments that could be made and it is usually not obvious which of these will correct the actual underlying errors. The adjustments can be done manually by subject-matter experts (known as *editors*) or automatically by a dedicated software program. Manual editing is more flexible than automatic editing, as editors can use external sources of information (including re-contacts with the respondent) to improve the quality of the data. On the other hand, manual editing is much more time-consuming and expensive, and it is also less reproducible as it may depend on subjective decisions by the editors.

Most NSIs nowadays apply some form of *selective editing* (Di Zio and Guarnera, 2014). This means that a selection procedure is used to identify the records in a data set that are most likely to contain important errors (i.e., errors that would have a noticeable impact on statistical output). These selected records are then edited manually. The selection procedure itself can be automated (e.g., by computing an "error score" for each record and selecting all records with a score above a certain threshold value) or manual (e.g., by letting an editor select outlying observations in a scatter plot).

The non-selected records are either not edited or they can be edited automatically. A possible advantage of editing the non-selected records automatically is that it leads to a data set in which all records are at least consistent with all hard edit rules (De Waal and Scholtus, 2011). This is particularly important if the micro-data themselves are part of the statistical output.

Regarding methods for automatic editing, a distinction is often made in the data editing literature between so-called *systematic errors* and *random errors* (De Waal et al., 2011). In this context, these terms have a somewhat different meaning than the bias and variance decomposition in formula (1.1). Here, the terms "systematic" and "random" refer to the mechanisms that cause the errors rather than their effects on an estimator. In fact, different definitions exist. According to UN/ECE (2000), a systematic error is "an error reported consistently over time and/or between responding units." Alternatively, EDIMBUS (2007) defines a systematic error as "a type of error for which the error mechanism and the imputation procedure are known." The term random error is used simply for any error that is not systematic.

Automatic methods for correcting systematic errors usually take the form of relatively simple IF-THEN rules. The IF statement describes a condition which identifies a particular systematic error, while the THEN statement describes the adjustment that should be made to the data to correct that error. Such deductive correction rules can be designed to formalise and automate correction strategies that are used by subject-matter experts for certain errors that are known to occur frequently. Some examples will be given in Chapter 3.

The problem of finding random errors in a record is known as the *error localisation problem* in the data editing literature. Automatic editing methods for random errors at NSIs are usually based on a formulation of the error localisation problem as a mathematical optimisation problem (De Waal et al., 2011). That is to say, the data in a record are "minimally" adjusted under the restriction that they have to satisfy all edit rules. Different methods can be obtained by choosing different minimisation criteria. The most widely used error localisation methods are based on the paradigm of Fellegi and Holt (1976) which minimises the number of values in a record that are changed. The underlying assumption is that most values are reported correctly to begin with; therefore, it would appear that by making a record consistent with the edit rules while leaving as many values unchanged as possible, we have the best chance of correcting the actual errors in that record. This heuristic argument can be made more precise; see Section 2.4.

For reasons of efficiency and timeliness, it is desirable to limit the amount of manual editing involved in a data editing process as much as possible (Pannekoek et al., 2013). In fact, if efficiency were the only criterion, the ideal editing process

would involve only automatic editing. However, it is generally assumed that data that have been edited automatically are of lower quality than data that have been edited manually by subject-matter experts (EDIMBUS, 2007).

At Statistics Netherlands at the beginning of the 21st century, a series of evaluation studies were conducted in which the same original data sets were edited both manually and automatically (Van der Pijll and Hoogland, 2003; Bikker, 2003). A number of systematic differences were found between the two edited data sets. For instance, editors sometimes find that respondents have interchanged the correct responses to two related questions (e.g., costs and revenues); they resolve this error by interchanging the responses again. This solution is virtually never chosen during automatic error localisation, because according to the Fellegi-Holt paradigm it is suboptimal to change two variables when it would suffice to change one. Some significant differences could also be seen in the estimated population totals based on the two edited data sets. These results indicate that the current automatic editing methodology alone is not sufficient to obtain edited data of adequate quality for publication purposes, and that selective manual editing of the most important errors remains necessary.

1.4 The estimation approach

We now turn to the estimation approach. Under this approach, given a data set that contains measurement errors which might invalidate statistical outcomes based on the data, one uses a model to estimate the amount of measurement error and somehow correct the statistical outcomes directly, without trying to identify individual errors.

If the data have been collected to estimate a statistical model of substantive interest, one possible approach is to extend the analysis model with a model for the measurement errors and to estimate both models simultaneously. In the econometrics literature, such models are known as *errors-in-variables models*. There exists an extensive literature on this subject; see, e.g., Durbin (1954), Fuller (1987), Bound et al. (2001) and Carroll et al. (2006). In this thesis, we will focus on measurement error models that can be estimated separately from any particular statistical analysis based on the data. These models can be useful to evaluate the overall measurement quality of the observed variables.

Consider an observed variable y which is the outcome of some measurement procedure. For instance, it might be obtained as a response to a question on a survey form or from an administrative data source. Perhaps the simplest possible

measurement error model for y is defined by the following formula:

$$y = T + \epsilon. \tag{1.2}$$

Here, ϵ denotes a random measurement error and T denotes the true score of the observed variable. It is assumed that ϵ is uncorrelated to T in the population. By definition, the random measurement error ϵ has an expected value of zero for each respondent. That is to say, the true score T for a given respondent is the average value that we would observe for y if the measurement procedure were repeated infinitely often for this respondent under identical circumstances. Here, "identical circumstances" implies in particular that at each repetition the respondent would have no memory of any of his/her previous responses. Clearly, this could never be achieved in practice, and in this sense the true score T is an unobservable or *latent* variable. It is a theoretical construct, but, as it turns out, a useful one.

Model (1.2) was originally developed in the field of educational testing and later extended to other applications in psychometrics. There exists an extensive theory based on variants of this model, known as the classical test theory (Lord and Novick, 1968).

As noted above in Section 1.2.3, the true score of an observed variable need not be equal to the true value of the variable of interest, because the measurement procedure might be systematically biased. Within the context of psychometrics for which the classical test theory was originally developed, researchers have generally focussed on analysing true scores rather than true values, because true values cannot be defined objectively for variables that arise in this context (e.g., intelligence). However, with applications in official statistics, there is often a natural interest in true values. In this thesis, we will therefore focus mostly on models that relate observed variables to the underlying true values of the variables of interest. In fact, the same modelling techniques can be applied to true scores or true values – the difference lies in the interpretation of the model parameters (Sobel and Arminger, 1986).

Let F denote the true value of the variable of interest that is measured by the observed variable y. Like the true score T, the true value F is treated as a latent variable. Later in this thesis, we will consider applications where F is assumed to be – in principle – observable, but not necessarily observed for all respondents. The relationship between the true value and the true score can be modelled in different ways. A relatively simple, linear model is given by:

$$T = a + bF + U, \tag{1.3}$$

where a and b denote intercept and slope parameters and U denotes a disturbance

term. The disturbances U represent systematic errors in the measurement procedure; unlike ϵ, the value of U for a given respondent is considered fixed under (hypothetic) repeated application of the measurement procedure. In the special case $U = 0$, the relationship between T and F is perfectly linear, and (1.2) and (1.3) define the following model for y:

$$y = a + bF + \epsilon. \tag{1.4}$$

The parameters a and b in (1.3) and (1.4) describe the relation between the true scale of the variable of interest and the scale of the observed variable. In the special case $a = 0$ and $b = 1$, the observed variable provides unbiased measurements of the variable of interest. If $a \neq 0$, the observed variable contains an *intercept bias*, i.e., a systematic deviation which is the same for all respondents. If $b \neq 1$, the observed variable contains a systematic deviation which is proportional to the true value. That is to say, the absolute size of this systematic deviation is larger for respondents with larger true values.

In the literature on measurement error models, the measurement quality of an observed variable is often described by its reliability and validity. These properties can be defined in terms of the (common) variation in the distributions of y, T and F in a population of potential respondents. Somewhat different definitions are given in the literature. In this thesis, we will follow the terminology of Saris and Andrews (1991) and Scherpenzeel and Saris (1997). See also Lord and Novick (1968, Chapter 2) or Chapter 2 of this thesis for more details about the assumed sampling distribution of y, T and F.

The *reliability* of y is the square of the correlation between the observed variable and its true score: $\mathrm{R}(y) = \rho^2(y, T)$. It is a measure of the absence of random measurement error. A value of $\mathrm{R}(y)$ close to 1 indicates that y is an accurate measure of whatever it is that it is measuring. This does not guarantee that y is also an accurate measure of the variable of interest, because it could still be affected by systematic measurement error in the form of U in (1.3). Note for instance that the crude "stick-length" measurement procedure of Section 1.2.3 probably yields measurements that are very reliable: basically, this procedure attempts to round the height of a person to the nearest metre, which in most cases leaves hardly any room for random error.

The *true-score validity* of y is equal to the square of the correlation between its true score and its true value: $\mathrm{TV}(y) = \rho^2(T, F)$. The term "true-score validity" is due to Saris and Andrews (1991); Biemer (2011) used the alternative term *theoretical validity*. Whatever terminology is used, a value of $\mathrm{TV}(y)$ close to 1 indicates that y is strongly related to the variable of interest *after correction for*

random measurement error. This implies that the influence of systematic measurement errors on y is small. In the example of height measurement by stick lengths, the true-score validity is probably quite low: the heights of persons rounded to the nearest metre are only weakly correlated to their true heights.

Finally, the *indicator validity* of y is equal to the square of the correlation between the observed variable and its true value: $\text{IV}(y) = \rho^2(y, F)$. Biemer (2011) called this the *empirical validity*. This quantity captures the joint effect of random and systematic measurement errors on y. In fact, it will be seen in Chapter 2 that $\text{IV}(y) = \text{TV}(y) \times \text{R}(y)$.

Being squared correlations, $\text{R}(y)$, $\text{TV}(y)$ and $\text{IV}(y)$ all take on values between 0 and 1. Note that the simplified model (1.4) has $\text{TV}(y) = \rho^2(T, F) = 1$ and $\text{R}(y) = \text{IV}(y)$. Thus, the distinction between reliability and (indicator) validity is relevant only for variables that contain systematic as well as random measurement errors.

For a single observed variable y, the model given by (1.2) and (1.3) is not identified, because there is no unique way to separate the observed y into latent factors F, U and ϵ. To obtain an identified model, we need to have multiple observations on the same set of respondents: different measures for the same variable of interest and/or measures for different variables of interest and/or observations from different waves of a panel study.

As an example, consider a situation where J different variables of interest $F_1, \ldots, F_J$ are measured each by K different observed variables $y_{1j}, \ldots, y_{Kj}$ ($j = 1, \ldots, J$). Suppose that each of these $K \times J$ observed variables can be modelled by an instance of the simple model (1.4):

$$y_{kj} = a_{kj} + b_{kj}F_j + \epsilon_{kj}, \tag{1.5}$$

with the additional assumption that the ϵ_{kj} are mutually uncorrelated in the population. For the purpose of estimating the reliability and validity, the joint model is identified in the presence of at least two variables of interest with at least two measures each (i.e., for $J \geq 2$ and $K \geq 2$). The model is also identified for a single variable of interest with at least three measures (i.e., for $J = 1$ and $K \geq 3$). The identification of the intercept and slope parameters themselves introduces an additional complication (see the end of Section 2.3.2). Several other examples of designs that lead to identified measurement error models will be discussed in Chapter 2.

Model (1.5) is an example of a *structural equation model*. Structural equation modelling provides a framework that can be used to estimate a wide variety of measurement error models. [See, e.g., Bollen (1989) for a general introduction.]

Authors who discussed applications of structural equation modelling to estimate reliability and validity include Heise and Bohrnstedt (1970), Jöreskog (1971), Andrews (1984), Saris and Andrews (1991) and Scherpenzeel and Saris (1997).

Once a measurement error model has been estimated, the results can be used in different ways. In questionnaire design, there is a tradition of using structural equation models to evaluate how choices made in the design of questions affect the reliability and validity of the obtained responses (Saris and Gallhofer, 2007; Alwin, 2007). Questions in future surveys could then be designed to optimise the reliability and validity of measurement. The studies listed at the end of the previous paragraph, starting with Andrews (1984), were all conducted with this aim in mind. In a data-integration context, measurement error modelling can also provide useful information: if several potential data sources (existing surveys or administrative data sets) are available for the production of statistical results, we can estimate the reliability and validity of the variables in these sources and compare them, in order to choose the best source.

In addition, an estimated measurement model can be used to correct statistical outcomes for the biasing effects of measurement errors in a data set. A well-known example concerns the estimation of correlation coefficients. Suppose that we are interested in the correlation between two variables F_1 and F_2 which are measured by y_{11} and y_{12}, respectively, and suppose that model (1.5) holds for these observations. Then it can be shown that the observed correlation $\rho(y_{11}, y_{12})$ is related to the true correlation $\rho(F_1, F_2)$ in the following way:

$$\begin{aligned}\rho(y_{11}, y_{12}) &= \rho(F_1, F_2)\rho(y_{11}, F_1)\rho(y_{12}, F_2) \\ &= \rho(F_1, F_2)\sqrt{\mathrm{IV}(y_{11})\mathrm{IV}(y_{12})}.\end{aligned}$$

This follows as a special case of Formula (4) in Scherpenzeel and Saris (1997); a direct proof proceeds along the lines of Lord and Novick (1968, pp. 69–70). Under this simple model, measurement errors always cause the correlation to be attenuated towards zero. Under more complex models, the correlation can be either attenuated or inflated by measurement errors (Bound et al., 2001). Having estimated the indicator validities of y_{11} and y_{12}, we can easily obtain an error-corrected estimate of the true correlation $\rho(F_1, F_2)$:

$$\frac{\rho(y_{11}, y_{12})}{\sqrt{\mathrm{IV}(y_{11})\mathrm{IV}(y_{12})}}.$$

Univariate statistics may also be biased in the presence of measurement errors. From the estimated parameters of an error model, it is possible to obtain predictions of the true values of the variables of interest, in order to correct individual variables

for systematic deviations in scale, such as intercept bias; see Meijer et al. (2012) and Chapter 6 of this thesis. More generally, these predicted true values can be used as a basis for imputation, either to handle non-response or to facilitate the estimation process (Boeschoten et al., 2016).

All measurement error models that fit within the structural equation modelling framework assume that the random measurement errors follow a continuous distribution. This implies that an observed value of a variable that contains errors has zero probability of being equal to the underlying true value. This assumption may not always be reasonable. In some cases, it may be more realistic to suppose that errors are "intermittent" (Di Zio and Guarnera, 2013): an observed value has a certain probability of being error-free (i.e., exactly equal to the underlying true value) and otherwise it contains an error from a continuous distribution. Measurement models with this property are known as *contamination models* (Bound et al., 2001). For instance, a contaminated version of model (1.5) is given by:

$$y_{kj} = (1 - z_{kj})F_j + z_{kj}(a_{kj} + b_{kj}F_j + \epsilon_{kj}), \tag{1.6}$$

where z_{kj} denotes a 0-1-indicator such y_{kj} contains an error if $z_{kj} = 1$ and no error if $z_{kj} = 0$. The probability of observing an error in y_{kj} is denoted by $\pi_{kj} = P(z_{kj} = 1)$. This probability affects the indicator validity of y_{kj} (see Section 2.3.3) and is also an interesting measurement quality parameter in its own right. Applications of model (1.6) to administrative data are discussed by Guarnera and Varriale (2015, 2016) and Robinson (2016) and in Chapter 7 of this thesis.

1.5 Outline of the rest of this thesis

The remaining chapters of this thesis are organised as follows. Chapter 2 provides a detailed review of existing work on the editing and estimation approaches to measurement errors. Reading this chapter first may be helpful to put the new results in the other chapters into context, but in principle each chapter can be read independently of the others.

Chapters 3–5 focus on new methods for automatic editing. In Chapter 3, we look at deductive correction methods for systematic errors. Correcting systematic errors in a separate step at the beginning of a data editing process can improve the efficiency of data editing as well as the quality of the edited data. This is true because, if a systematic error can be corrected accurately by a deductive rule, it does not have to be treated later on by a human editor or a more complex algorithm for automatic error localisation. This means that editors and more complex algorithms

can focus attention on cases with more complicated error structures, where their contribution is more likely to be worthwhile.

With the above aims of improving efficiency and quality in mind, two new deductive methods are developed for correcting two errors that are known to occur in data of the so-called Structural Business Statistics (SBS) at Statistics Netherlands: sign errors and rounding errors. Both methods require an algorithm that is more complex than a simple IF-THEN rule, but they are still relatively easy and cheap to implement. Theoretical properties of the algorithms are investigated. By way of illustration, both algorithms are applied to real data from the Netherlands' SBS of 2007.

In Chapters 4 and 5, we focus on error localisation for random errors. Two generalisations of the Fellegi-Holt paradigm are proposed that aim to improve the quality of automatically-edited data. Both generalisations address a different limitation of the existing Fellegi-Holt paradigm.

Chapter 4 posits the idea that some of the systematic differences that have been found between manual and automatic editing may be explained by the fact that human editors make use of soft edits as well as hard edits, whereas the Fellegi-Holt paradigm for automatic editing assumes that only hard edit rules occur. Under the Fellegi-Holt paradigm, soft edits have to be either ignored or treated as hard edits during automatic error localisation. We propose a new formulation of the error localisation problem that can distinguish between hard and soft edit rules. The new approach involves solving a minimisation problem that is a generalisation of the problem of Fellegi and Holt, with an extra term that measures the extent to which soft edit rules are violated. The new problem can be solved by an extension of the existing error localisation algorithm of De Waal and Quere (2003). To test the new method, a simulation study is conducted with synthetic data.

The Fellegi-Holt paradigm tacitly assumes that errors independently affect one variable at a time. By contrast, human editors often make adjustments to the data that involve more than one variable at a time. It is in fact likely that respondents often make errors that affect several variables simultaneously. In Chapter 5 we therefore introduce a generalised error localisation problem in which the assumption is relaxed that errors affect one variable at a time. This problem is based on a new minimisation criterion which involves the number of required edit operations rather than the number of changed values. Here, each edit operation is a well-defined elementary adjustment that can be made to a record to correct one particular error, which might involve changing the values of one, two, or more variables simultaneously. It is suggested that these edit operations be chosen to mimic the operations made by human editors as well as possible. The Fellegi-Holt-based

error localisation problem is in fact a special case of the new problem, obtained by restricting the set of admissible edit operations to one particular class (i.e., operations that impute a new value for a single variable). An algorithm is developed for solving the new error localisation problem. This algorithm is used in a simulation study with synthetic data to compare the new approach to Fellegi and Holt's original error localisation problem.

Chapters 6 and 7 focus on applications of measurement error models. In Chapter 6, we use a measurement error model to estimate the quality of administrative and survey data for official statistics. It is shown how both the indicator validity and intercept bias of administrative and survey variables can be estimated through structural equation modelling. In particular, the indicator validity can be used as a measure to decide whether the administrative concept is sufficiently related to the true variable of interest to be of use. In cases where the validity is high but significant intercept bias occurs, a correction formula can be derived from the model by predicting the true value of the variable of interest from the observed value. To fully identify the model, we take a random subsample of our original observations and attempt to measure the true values for these units (an *audit sample*). The inclusion of an audit sample is necessary for the estimation of the true intercept bias and true correction formulas for the observed variables, but not for indicator validity.

The methodology is applied to real data at Statistics Netherlands to estimate the validity and intercept bias of VAT turnover for short-term statistics (monthly or quarterly statistics on the development of the economy). Structural equation models are fitted to linked data from three administrative sources (VAT, the Profit Declaration Register and the General Business Register) and one survey (SBS). Additional data for an audit sample are obtained by re-editing the survey data. The results of the structural equation models are compared to earlier results by Van Delden et al. (2016) based on robust linear regression. We also simulate an application of the estimated correction formulas from the model to publication figures for the short-term statistics.

In Chapter 7, we use measurement error modelling to gain insight into the quality of edited data. The indicator validity and bias of observed variables in a data set of the Netherlands' SBS before and after automatic editing are evaluated and compared. We analyse the data using two different models: a structural equation model and a contamination model. The latter model seems more appropriate for the data at hand, but its current formulation does have some limitations that require further development.

Finally, a summary of the results of this thesis follows in Chapter 8. We also draw some conclusions from these results, about their potential application to im-

prove the quality of official statistics and about questions for future research.

Note: Some chapters are based on articles or reports that have been published previously; if so, this is indicated at the start of the chapter. If a chapter features co-authors then their names and contributions are indicated similarly at the start of the chapter. The remaining chapters (1, 2, and 8) have been written specifically for this thesis with Sander Scholtus as the single author.

Chapter 2

Editing and Estimation of Measurement Errors

2.1 Introduction

In this chapter, we give a more detailed overview of the two approaches for handling measurement errors in statistical data that were introduced in Section 1.3 and Section 1.4: editing and estimation. The main aim of this chapter is to provide some context for the new methodological research that will be discussed in the rest of this thesis. In addition, we take the opportunity to mention here some features of the two approaches of which a discussion did not fit naturally into the other chapters – in particular, some general thoughts on their application to administrative data. The editing approach is discussed in Section 2.2; the estimation approach in Section 2.3. In Section 2.4, we briefly discuss a connection between the two approaches by examining measurement error models that occur in the literature on data editing. Some conclusions follow in Section 2.5.

2.2 The editing approach

2.2.1 Methods for statistical data editing

We begin this section by reviewing different types of data editing methods that have been developed over time: manual editing, selective and macro-editing, and automatic editing.

Manual editing

In the traditional data editing process, all data were edited by hand. Human editors would check each record of data against a set of edit rules (see Section 1.3), identify

the records that contain inconsistencies with respect to these rules and then try to correct errors in these records by making adjustments. In the past, respondents were often re-contacted to clarify implausible responses.

As noted in Section 1.3, a distinction can be made between hard and soft edit rules. Traditionally, the data were edited manually until all inconsistencies with respect to hard edits had been resolved and all inconsistencies with respect to soft edits had been resolved or explained. As a result, the data editing process was very time-consuming and labour-intensive. It has been estimated that, by the late 1980s, NSIs and other statistical agencies would spend up to 20 to 40 per cent of their survey costs on data editing (Federal Committee on Statistical Methodology, 1990; Granquist, 1995, 1997).

In the last decades, many innovations have been introduced to improve the efficiency of data editing (see below). Nevertheless, manual editing is still a part of nearly every data editing process, although it is now focussed on a subset of the data. Nowadays, editors are usually supported by a computer; the term *interactive editing* is then also used (Scholtus, 2014c).

The role of re-contacts has also changed over time. In current surveys, subject-matter experts often try to edit the data without contacting the respondent. Re-contacts are avoided because they add to the response burden of a survey. Moreover, in particular for business surveys, editors can often find much of the information they need about responding units on the internet or in dedicated databases.

Granquist (1997) argued that, if re-contacts are used, their aim should not just be to correct individual errors in the data, but rather to find out the causes of these errors. In this way, the editing process can reveal deficiencies of the data collection design – such as question formulations that are difficult to answer correctly – that might be improved in future surveys. In the long run, reducing the number of measurement errors during data collection is more useful than trying to correct these errors later on. According to Granquist (1997), "editing should highlight, not conceal, serious problems in the survey vehicle."

Selective editing and macro-editing

Electronic computers have been used to support the data editing process since the 1950s (Nordbotten, 1963; Stuart, 1966). Somewhat unexpectedly, the introduction of (mainframe) computers often actually led to an increase in the costs and time spent on the data editing process. The reason for this was that the computer could be used to check the data against more – and more complex – edit rules than was possible before. In particular, this led to a proliferation of soft edit rules. However, all records that did not satisfy these edits were still checked and, if necessary,

adjusted manually by editors (Granquist, 1995).

One way to improve the efficiency of a data editing process is by improving the formulation of the edit rules, in particular the soft ones (Granquist, 1997). By varying, for instance, the upper bound of a ratio edit, one obtains different risks of false positives (when the edit is failed by an observation that is correct) versus false negatives (when the edit is satisfied by an observation that does contain an error). The formulation of an effective soft edit involves balancing these two risks and is therefore a statistical problem (Di Zio et al., 2005b). Granquist (1995, 1997) noted that a recurring problem with surveys in the 1980s was that soft edits yielded too many false positives.

A more fundamental improvement is based on the realisation that, for nearly every application in official statistics, it is unnecessary and in fact undesirable to edit a data set until all edit violations have been either resolved or explained (Granquist and Kovar, 1997; De Waal et al., 2011). The main output of NSIs consists of relatively simple descriptive statistics of a population (totals, means, ratios) that are not particularly sensitive to small errors in individual observations. In fact, the effects of individual errors might well cancel out when the data are aggregated to the population level. Moreover, statistical estimates are subject to other sources of error besides measurement errors, as discussed in Section 1.2.2. It is not very useful to focus attention on correcting measurement errors if the total survey error is dominated by other types of error (e.g., undercoverage). In particular, for sample-based estimates the effect of measurement errors might be small in comparison to the sampling error, once the most important errors have been corrected.

This notion that it is not necessary to edit all data in every detail – which can be found implicitly already in Nordbotten (1955) – became widely acknowledged in the 1980s and 1990s. Studies were done that showed empirically that accurate estimates of population means and totals could be obtained by editing only the most important errors in a data set; see, e.g., Granquist (1995, 1997) and Granquist and Kovar (1997). This prompted the development of new data editing methods called *selective editing* and *macro-editing* which aim to identify (select) the records that are likely to contain the most important errors. The selected records are then edited manually. Different selection methods and tools were discussed by, among others, Hidiroglou and Berthelot (1986), Granquist (1990), Lawrence and McKenzie (2000), Hedlin (2003), Hoogland (2006), Arbués et al. (2013), Di Zio and Guarnera (2013) and Norberg (2016). See also Di Zio and Guarnera (2014) for an overview. These methods all focus on numerical variables and are used mainly for business surveys.

We will not discuss methods for selective editing in detail here. One feature

that should be mentioned is that, traditionally, selective editing is done essentially by cut-off sampling: all records that satisfy a certain criterion are selected for manual editing, and no other record is. Ilves and Laitila (2009) noted that this non-probability sampling approach makes it impossible to estimate the total amount of measurement error that remains in the non-selected part of the data. Thus, the quality of selective editing cannot be evaluated during regular production – this would require a separate study in which some of the non-selected records are edited as well. To overcome this problem, Ilves and Laitila (2009) suggested the use of a probabilistic selection criterion instead. For instance, in addition to the records that satisfy the original selection criterion, a small subset of the other records could be selected for manual editing by probability proportional to size sampling, with the inclusion probability of each record based on the expected influence of its errors. In that case, sampling theory could be used to estimate the effects of the remaining measurement errors in the non-selected records on statistical output.

Automatic editing

Most NSIs nowadays use selective and/or macro-editing to restrict manual editing to those observations that are likely to contain influential errors (i.e., errors that are likely to affect the statistical output in a significant way). The remaining observations can be left as they are, or they can be edited automatically. Methods for automatic editing have been developed since the 1960s (Freund and Hartley, 1967). In automatic editing, a software program checks the data against the edit rules and also makes adjustments to the data to obtain consistency with these rules.

De Waal and Scholtus (2011) argued that it can be useful in practice to edit the non-selected records in a data set automatically, even though the errors within these records should have a limited impact on statistical output. Firstly, it is often desirable to ensure that all data are at least consistent with the hard edit rules, because inconsistencies with respect to these rules (e.g., totals that are not equal to the sum of their parts) might lead statistical users to reject the data. Automatic editing provides an efficient way to obtain consistency with the hard edits. Secondly, automatic editing of the non-selected records also provides an inexpensive way to check whether the selection criteria for influential errors are working correctly: if they are, only relatively small errors should be found during automatic editing. Furthermore, Pannekoek et al. (2013) noted that if automatic editing methods can be improved to the point where they yield data of reasonable quality, this means that less data would have to be selected for manual editing to obtain the same overall quality. Thus, the efficiency of the data editing process could be improved further.

Automatic editing will be discussed in detail in Chapters 3–5 of this thesis. For

now, it is useful to note that different automatic editing methods exist for systematic and random errors (as defined in Section 1.3). Systematic errors are corrected deductively, sometimes by simple algorithms and more often by means of so-called *correction rules* which take the form of IF-THEN rules. The idea behind such a rule is that a certain type of error is made in the same way by different respondents, can be recognised in the data, and once it has been recognised it is obvious how to correct it. In practice, an error is called systematic when such a correction procedure is possible.

The best-known example of a systematic error is the so-called unit of measurement error or "thousand error". Often, business surveys contain an instruction to report amounts that are rounded to the nearest multiple of 1,000 units. Invariably, some respondents ignore or misread this instruction and report amounts in single units, which the NSI then interprets as being 1,000 times too large. This error is often relatively easy to detect, for instance by comparing the observed values to reference values from a different source or from a previous survey, either numerically or graphically (in a scatter plot). Different methods for detecting thousand errors are discussed by Di Zio et al. (2005a, 2007), Al-Hamad et al. (2008) and De Waal et al. (2011). Once the error has been detected, correcting it is straightforward.

De Waal and Scholtus (2011) distinguished *generic* and *subject-related* systematic errors. A systematic error is called generic if it occurs in essentially the same way for a variety of different variables and data sources. Examples include the above-mentioned unit of measurement error and simple typing errors where, for instance, a respondent writes "379" instead of "397" (Scholtus, 2009). By contrast, a subject-related systematic error affects a specific variable in a specific data source. The correction of such an error requires subject-matter knowledge. An example, given by De Waal and Scholtus (2011), is that businesses in the construction industry often incorrectly report the specification of staff by department, with too many employees classified as "working in other departments".

Automatic correction of generic and subject-related systematic errors is usually done at the beginning of a data editing process, even before selective or macro-editing is applied (Pannekoek et al., 2013). After this step, only random errors remain in the data. As mentioned in Section 1.3, the automatic editing problem for data that contain random errors is known as the *error localisation problem*, and most methods for solving this problem that are currently in use at NSIs are based on the paradigm of Fellegi and Holt (1976).

According to the Fellegi-Holt paradigm, each record should be made consistent with all edit rules by adjusting the smallest possible number of observed values. Since the accuracy of measurement might be different for each variable, the

paradigm is often generalised by assigning a so-called *confidence weight* w_k to each observed variable y_k. The objective is then to find a subset of variables that can be adjusted to obtain a consistent record and has the minimal sum of confidence weights among such subsets. That is to say, the expression

$$\sum_{k=1}^{K} w_k \delta_k \tag{2.1}$$

is minimised, with $\delta_k = 1$ if y_k is adjusted and $\delta_k = 0$ otherwise, under the restriction that the adjusted record must satisfy all edit rules. Note that this approach implicitly assumes that all edit rules are hard.

The error localisation problem under the Fellegi-Holt paradigm is computationally challenging in practice. Its complexity increases with the number of observed variables and the number of edit rules. Over the past decades, many NSIs have developed dedicated algorithms for solving this error localisation problem. Different solutions were suggested by, among others, Fellegi and Holt (1976), Schaffer (1987), Garfinkel et al. (1988), Kovar and Whitridge (1990), Ragsdale and McKeown (1996), De Waal (2003c), De Waal and Quere (2003), Riera-Ledesma and Salazar-González (2003, 2007), Bruni (2004), and De Jonge and Van der Loo (2014). See also De Waal et al. (2011, Chapters 3–5) for a detailed overview.

2.2.2 Administrative data

The data editing methods discussed above were originally developed for survey data. Di Zio and Luzi (2014) discussed their extension to administrative data. In broad terms, the data editing process for administrative data can be similar to that for survey data. A few important differences occur, however.

Firstly, manual editing of administrative data is complicated by the fact that these data were originally collected and processed outside the NSI. It is likely that these data have already been subjected to some form of data editing by the register owner to make them suitable for administrative purposes, but the NSI usually has little information about this. Editors may also have less information about the way the data were collected and the way administrative variables are defined, which makes it more difficult to identify and resolve errors in the data. Moreover, it is usually impossible or illegal for NSIs to re-contact individual units in an administrative data set. If serious problems are found in the data, it may be possible to contact the register owner and obtain a new version of the entire data set.

A second important difference is that administrative data usually contain a much larger number of units than survey data. This increases the importance of having an efficient data editing process. Standard selective editing approaches for

survey data may not be useful for large administrative data sets, because the selected subset of units is still so large that manual editing would be too costly or time-consuming. Automatic editing methods could therefore be very useful for administrative data, provided that these methods can yield data of sufficient quality.

2.3 The estimation approach

2.3.1 The true-score measurement error model

In Section 1.4, a number of useful measurement quality parameters were introduced from the point of view of a very simple measurement error model: intercept bias, reliability and validity. We will now examine these concepts again using a more general measurement error model that was originally proposed for survey data by Saris and Andrews (1991). This will provide a useful starting point for a review of some survey designs by which these measurement quality parameters can be estimated in practice, which will follow in Section 2.3.2. Section 2.3.3 re-examines the contamination model that was also mentioned in Section 1.4. The possibilities of applying these models to administrative data are discussed in Section 2.3.4.

As before, we use F to denote the true value of a variable of interest, that is, an attribute that we would like to measure. An attempt is made to measure F as well as possible – e.g., by asking a question on a survey form – and the resulting observed variable is denoted by y. In the presence of measurement errors, the individual values of y and F will not be identical. Saris and Andrews (1991) proposed to model the relation between y and F in the following way:

$$\begin{aligned} y &= a + bF + gM + u + \epsilon \\ &= T + \epsilon. \end{aligned} \tag{2.2}$$

In this expression, ϵ denotes a random measurement error term, and a, b and g are constants.

The remaining components of formula (2.2) aim to capture all systematic influences on the observed variable. As noted in Section 1.2.2, it is well-known that answers to survey questions can depend on other factors than the variable of interest itself; e.g., the wording of a question or the set of possible answer categories. In expression (2.2), the variable M denotes the systematic contribution of the "method" by which y is measured. This method component can include all kinds of aspects of the way the question is presented to the respondent [e.g., length of the question text, number of answer categories, survey mode; see Andrews (1984) or Scherpenzeel and Saris (1997) for more examples]. As will be seen in the next subsection,

it is assumed that one can hold the method component "fixed" while varying the variable of interest. That is to say, the same method can be used to measure different variables of interest for the same respondent and the value of M in (2.2) is then supposed to be the same in each instance. In fact, the method component is assumed to be uncorrelated to the variable of interest [see (2.4) below]. Any interaction between the variable of interest F and the method component M that might also influence y is included in the so-called "unique component" u. Finally, the factor T denotes the true score of y – i.e., the stable part that remains when the observed variable is corrected for random measurement error (Lord and Novick, 1968) – which under this model equals $a + bF + gM + u$.

To complete the formulation of the error model, we have to add the following assumptions, which state that the various error components in (2.2) have an expected value of zero and are both uncorrelated amongst themselves and uncorrelated to the variable of interest:

$$E(M) = E(u) = E(\epsilon) = 0 \tag{2.3}$$

and

$$\begin{aligned} \text{cov}(F, M) = \text{cov}(F, u) = \text{cov}(F, \epsilon) = \text{cov}(M, u) \\ = \text{cov}(M, \epsilon) = \text{cov}(u, \epsilon) = 0. \end{aligned} \tag{2.4}$$

The sampling distribution to which these expectations and covariances refer is defined as follows [cf. Lord and Novick (1968, Chapter 2)]. Consider a (finite or infinite) population of potential respondents. Each respondent i is supposed to have a hypothetical distribution of potential responses that could be obtained by repeating the measurement procedure for y infinitely many times under identical conditions. (In this hypothetical setting, the respondent is supposed to have no memory of any responses that he/she gave previously.) The random measurement error ϵ_i captures all variability within the potential response distribution of respondent i. The expected value of this response distribution, and hence the expected value of y_i, is defined to be equal to $T_i = a + bF_i + gM_i + u_i$. From this, it follows that $E(\epsilon_i) = 0$ for each respondent i under (hypothetical) repeated measurement. The expected value of y_i depends on the choice of variable of interest (through F_i and u_i) and the choice of method (through M_i and u_i). Next, imagine a sampling procedure by which one respondent i is selected at random from the population and its values y_i, F_i, M_i, u_i and ϵ_i are obtained once. (Of course, in practice only the value of y_i would be observed.) The random variables y, F, M, u and ϵ in (2.2) are defined by this sampling procedure. In particular, the expectations and covariances in (2.3) and (2.4) refer to the outcome of this random sampling of respondents.

It is worth repeating that, under this model, the values of F, M and u for a given respondent are assumed to be fixed: under repeated observation of the same respondent under identical conditions, only the value of ϵ would vary.

According to (2.2) and (2.3), the expected value of y when sampling from the population of respondents equals $E(y) = E(T) = a + bE(F)$. Thus, the parameters a and b in (2.2) capture differences in scale between the variable of interest and the observed variable. In particular, a represents the so-called *intercept bias*. A non-zero value of a can affect statistics about the population distribution of a single variable (e.g., estimates of population means and totals) but it will not have any impact on statistics about relationships between variables (e.g., regression coefficients or correlations). In social science applications, for which this error model was originally developed, only the latter type of statistics are usually of interest and the possibility of intercept bias is usually ignored. In official statistics, however, population means and totals are often part of the statistical output, so intercept bias may be an important aspect of measurement quality. The estimation of intercept bias introduces an additional complication, so we defer a discussion of this point until the end of Section 2.3.2. We first focus on some other measurement properties.

In what follows, for simplicity we assume that $b > 0$. Given the form of (2.2), this is in fact a necessary condition for y to qualify as a measure of F in the terminology of Lord and Novick (1968, p. 20): "We shall call an observable variable a *measure* of a theoretical construct if its expected value is presumed to increase monotonically with the construct."

Several useful measurement quality parameters can be defined in terms of the above error model (Scherpenzeel and Saris, 1997). Firstly, the *reliability coefficient* of y is equal to its correlation to the true score:

$$\text{RC}(y) = \rho(y, T) = \frac{\text{cov}(y, T)}{\sqrt{\text{var}(y)\text{var}(T)}} = \frac{\text{var}(T)}{\sqrt{\text{var}(y)\text{var}(T)}} = \sqrt{\frac{\text{var}(T)}{\text{var}(y)}}. \quad (2.5)$$

The third equality follows from (2.2) and assumption (2.4). Taking the square of the reliability coefficient yields the fraction of total variance in y that is explained by the stable components F, M and u:

$$\text{R}(y) = \text{RC}^2(y) = \frac{\text{var}(T)}{\text{var}(y)} = 1 - \frac{\text{var}(\epsilon)}{\text{var}(y)}. \quad (2.6)$$

This quantity is known as the *reliability* of y. It corresponds to the so-called "test-retest reliability" in the classical test theory (Lord and Novick, 1968).

Secondly, the *true-score validity coefficient* of y is equal to the correlation of

its true score to its true value:

$$\text{TVC}(y) = \rho(T, F) = \frac{b\text{var}(F)}{\sqrt{\text{var}(T)\text{var}(F)}} = b\sqrt{\frac{\text{var}(F)}{\text{var}(T)}}. \tag{2.7}$$

The second equality follows from $T = a + bF + gM + u$ and assumption (2.4). The right-most expression in (2.7) shows that $\text{TVC}(y)$ is equal to the standardised coefficient of F in a linear regression of T on F and M, with u as a disturbance term. Taking the square of this validity coefficient yields the fraction of total variance in T that is explained by the variable of interest:

$$\text{TV}(y) = \text{TVC}^2(y) = \frac{b^2\text{var}(F)}{\text{var}(T)}. \tag{2.8}$$

Heise and Bohrnstedt (1970) proposed (essentially) this quantity as a measure of the *validity* of y. Saris and Andrews (1991) suggested the term *true-score validity* to avoid confusion with other definitions of validity. Biemer (2011) used the alternative term *theoretical validity*.

Thirdly, the *indicator validity coefficient* of y is equal to its correlation to the true value:

$$\text{IVC}(y) = \rho(y, F) = \frac{b\text{var}(F)}{\sqrt{\text{var}(y)\text{var}(F)}} = b\sqrt{\frac{\text{var}(F)}{\text{var}(y)}}. \tag{2.9}$$

Taking the square of this coefficient yields the fraction of total variance in y that is explained by the variable of interest alone:

$$\text{IV}(y) = \text{IVC}^2(y) = \frac{b^2\text{var}(F)}{\text{var}(y)}. \tag{2.10}$$

Saris and Andrews (1991) called (2.10) the *indicator validity* of y; Biemer (2011) used the term *empirical validity*. Table 2.1 summarises the above measurement quality parameters for ease of reference.

Table 2.1: Overview of measurement quality parameters under the true-score model.

Name	Abbreviation	Definition	Formula
Reliability coefficient	$\text{RC}(y)$	$\rho(y, T)$	$\sqrt{\text{var}(T)/\text{var}(y)}$
True-score validity coefficient	$\text{TVC}(y)$	$\rho(T, F)$	$b\sqrt{\text{var}(F)/\text{var}(T)}$
Indicator validity coefficient	$\text{IVC}(y)$	$\rho(y, F)$	$b\sqrt{\text{var}(F)/\text{var}(y)}$

From expressions (2.5)–(2.10) it is easy to derive that

$$\text{IVC}(y) = \text{TVC}(y) \times \text{RC}(y), \tag{2.11}$$

$$\text{IV}(y) = \text{TV}(y) \times \text{R}(y). \tag{2.12}$$

Thus, the indicator validity of y summarises the joint effect of random and systematic measurement errors on y. For this reason, Saris and Andrews (1991) preferred to report separate true-score validity and reliability coefficients, as these are easier to interpret. However, it will be seen below that the true-score validity of a variable is often more difficult to estimate than its indicator validity.

All of the above measurement quality parameters take on values between 0 and 1, with values closer to 1 indicating better measurement properties. In particular, it follows from (2.12) that

$$\text{IV}(y) \leq \text{R}(y), \tag{2.13}$$

$$\text{IV}(y) \leq \text{TV}(y). \tag{2.14}$$

Property (2.13) is sometimes stated as "high reliability is necessary, but not sufficient, for high validity". It is worth noting that this holds true for indicator validity but not for true-score validity: it is in fact possible for the same measure to have low reliability and high true-score validity with respect to some variable of interest.

2.3.2 Designs for estimating measurement quality

As noted in Section 1.4, the various components of model (2.2) (i.e., the contributions of the variable of interest, the method component, the unique component and the random measurement error) cannot be separated if one observes only a single variable y. If multiple observations are available on the same respondents, then it may be possible to identify the model by treating the different components as latent factors in a structural equation model (SEM). Structural equation modelling in general will be introduced in Chapter 6. In the present subsection, we consider a number of research designs that have been used in the past to estimate the measurement quality parameters listed in Section 2.3.1 for survey data.

Multitrait-multimethod designs

The *multitrait-multimethod* (MTMM) design, originally proposed by Campbell and Fiske (1959), combines a number of variables of interest (*traits*) with a number of methods. In the basic set-up, each trait is measured once using each method for all respondents. Thus, if there are J variables of interest $F_1, \ldots, F_J$ and K methods $M_1, \ldots, M_K$, then the MTMM design requires $J \times K$ observations for each respondent. If these measurements are collected in a vector $\mathbf{y} = (y_{11}, \ldots, y_{KJ})'$, where y_{kj} denotes the observed value for variable of interest F_j measured by method M_k, then the MTMM design allows one to estimate the variance-covariance matrix of $\mathbf{y}$. Heise and Bohrnstedt (1970) and Jöreskog (1971) proposed to analyse

this matrix using an SEM. This idea was developed further and applied to versions of model (2.2) by Andrews (1984) and Saris and Andrews (1991).

In terms of the MTMM notation of the previous paragraph, (2.2) can be written as an SEM as follows:

$$y_{kj} = T_{kj} + \epsilon_{kj}, \tag{2.15}$$

$$T_{kj} = a_{kj} + b_{kj}F_j + g_{kj}M_k + u_{kj}, \tag{2.16}$$

where each combination of j and k contributes one instance of (2.15) and one instance of (2.16). In SEM terminology, (2.15) is a measurement equation and (2.16) is a structural equation. In addition to (2.3) and (2.4), it is now also assumed that $\text{cov}(\epsilon_{kj}, \epsilon_{k'j'}) = 0$ and $\text{cov}(u_{kj}, u_{k'j'}) = 0$ for all pairs of distinct measures. This latter assumption is crucial to the model but not easy to achieve in practice: as the MTMM design essentially involves asking respondents a set of K slightly altered questions about each of the J traits, the responses to questions about the same trait could easily become correlated because respondents remember, and try to remain consistent with, their previous answers. The risk of these memory effects can be reduced by dividing the different questions for each trait over multiple interviews, but this requires the additional (untestable) assumption that the true score of the respondent has not changed between interviews. Within a single interview, the risk of memory effects can be reduced by interspersing the questions about each variable of interest among large numbers of similarly formatted questions. For the latter approach, Van Meurs and Saris (1990) suggested as a rule of thumb that memory effects can be ignored provided that each pair of repeated questions is separated by at least 20 minutes of similarly formatted questions with related content.

If one is not interested in the intercept bias a_{kj}, the model given by (2.15)–(2.16) can be simplified by standardising all variables y_{kj}, T_{kj}, F_j and M_k to have mean 0 and variance 1. (The disturbance terms ϵ_{kj} and u_{kj} remain unstandardised.) This leads to the following equations:

$$y^s_{kj} = h^s_{kj}T^s_{kj} + \epsilon^*_{kj}, \tag{2.17}$$

$$T^s_{kj} = b^s_{kj}F^s_j + g^s_{kj}M^s_k + u^*_{kj}. \tag{2.18}$$

Here, y^s_{kj}, T^s_{kj}, F^s_j and M^s_k denote standardised variables and the new coefficients are given by

$$h^s_{kj} = \sqrt{\frac{\text{var}(T_{kj})}{\text{var}(y_{kj})}},$$

$$b^s_{kj} = b_{kj}\sqrt{\frac{\text{var}(F_j)}{\text{var}(T_{kj})}},$$

and

$$g_{kj}^s = g_{kj}\sqrt{\frac{\text{var}(M_k)}{\text{var}(T_{kj})}}.$$

Thus, referring back to Table 2.1, it is seen that the reliability, true-score validity and indicator validity coefficients of y_{kj} can be obtained directly from the parameters of the standardised SEM:

$$\begin{aligned}\text{RC}(y_{kj}) &= h_{kj}^s,\\ \text{TVC}(y_{kj}) &= b_{kj}^s,\\ \text{IVC}(y_{kj}) &= h_{kj}^s b_{kj}^s.\end{aligned}$$

These expressions are also given by Scherpenzeel and Saris (1997). By analogy to the true-score validity coefficient, they define g_{kj}^s as a further measurement quality parameter called the *method effect*. In addition, the quantity $1-(b_{kj}^s)^2 = (g_{kj}^s)^2 + \text{var}(u_{kj}^*)$ is known as the *invalidity* of y_{kj}. In what follows, we will focus on reliability and validity.

As it stands, the model given by (2.17)–(2.18) cannot be estimated, because each observed variable y_{kj} contributes both a random measurement error ϵ_{kj} and a unique component u_{kj} and these cannot be distinguished. (The model is not identified.) Saris and Andrews (1991) discussed three possible solutions to this problem. Firstly, the model can be made identifiable by repeating the whole MTMM design across two (or more) waves of a panel study. From a theoretical point of view, this is the best solution, as it allows one to estimate the model as formulated above without introducing any new assumptions. However, it would place an even higher burden on the respondents than the original MTMM design. Applications of this panel-based extension have therefore been rare (Saris and Andrews, 1991).

A second option is to assume that all unique components u_{kj} are zero. Under this assumption, all variation in y_{kj} that is not explained by the latent factors F_j and M_k must be due to random measurement error. This approach was advocated by Saris and Andrews (1991) and Scherpenzeel and Saris (1997), who called (2.17)–(2.18) with this additional assumption the *true-score MTMM model*. The true-score MTMM model yields estimates of the reliability, true-score and indicator validity and method effect for each of the observed variables. Saris (1990b) and Saris and Andrews (1991) argued that the assumption of zero unique components is plausible if the variables $y_{1j}, \ldots, y_{Kj}$ that are supposed to measure F_j indeed all measure the same variable of interest, i.e., all systematic differences between these observed variables are captured by the methods $M_1, \ldots, M_K$. They also pointed out that in rare cases where the above repeated MTMM design was used and the

unique component variances could be estimated, these turned out to be close to zero (Rodgers, 1989; Andrews, 1984, footnote 14).

A third option is to ignore the true scores and replace (2.17)–(2.18) by

$$y^s_{kj} = \ell^s_{kj} F^s_j + m^s_{kj} M^s_k + w_{kj}, \tag{2.19}$$

where the disturbance term w_{kj} captures both ϵ_{kj} and u_{kj}. Saris and Andrews (1991) called this the *classical MTMM model*, since it pre-dates the true-score formulation. In applications that use this approach, the reliability and validity of y_{kj} are defined as $1 - \text{var}(w_{kj}) = (\ell^s_{kj})^2 + (m^s_{kj})^2$ and $(\ell^s_{kj})^2$, respectively. This definition of validity is equivalent to indicator validity as defined by (2.10). The definition of reliability corresponds to the squared multiple correlation coefficient in the regression of y_{kj} on F_j and M_k [as suggested, e.g., by Bollen (1989)] and is *not*, in general, equivalent to (2.6). From the point of view of the true-score model, this is a drawback. Another drawback of the classical MTMM model is that it cannot be used, in general, to estimate the true-score validity. However, Saris and Andrews (1991) pointed out that if the assumption that $u_{kj} = 0$ in fact holds, the true-score and classical MTMM models are equivalent. In particular, their definitions of reliability then coincide, and the true-score validity may be derived under the classical model using expression (2.12):

$$\text{TV}(y_{kj}) = \frac{\text{IV}(y_{kj})}{\text{R}(y_{kj})} = \frac{(\ell^s_{kj})^2}{(\ell^s_{kj})^2 + (m^s_{kj})^2}.$$

Both the classical and true-score MTMM models as defined above are theoretically identified provided that data are available on $J \geq 3$ correlated traits combined with $K \geq 3$ methods (Költringer, 1990). Usually, the assumption is added that the method factors $M_1, \ldots, M_K$ are uncorrelated with each other (Scherpenzeel and Saris, 1997). In practice, if this assumption is not made, the model with $J = 3$ traits and $K = 3$ methods has often been found to be empirically underidentified (as evidenced by convergence problems during model estimation and by unstable parameter estimates). See Scherpenzeel (1995) for a detailed study of this and other practical issues that may arise with MTMM models.

The MTMM design places a heavy burden on respondents. If the above rule of thumb by Van Meurs and Saris (1990) is adhered to, collecting MTMM data for $K = 3$ different methods in a single interview requires that respondents answer similarly structured questions for more than 40 minutes. With questionnaires of this length, there is a substantial risk that the accuracy of measurement changes as the interview goes on. Saris and Gallhofer (2007, p. 219) suggested that, initially, the accuracy could be higher for questions that are asked later in the survey, because respondents have had more time to think about the subject of the survey and

Table 2.2: Example of a two-group and three-group split-ballot MTMM design with $K = 3$ methods, as suggested by Saris et al. (2004).

	First	Second
Group 1	M_1	M_3
Group 2	M_2	M_3

	First	Second
Group 1	M_1	M_2
Group 2	M_2	M_3
Group 3	M_3	M_1

have become more familiar with the interviewing process. At some point, however, respondents are likely to become tired or annoyed and the accuracy of their responses will start to decrease. Moreover, surveys with long questionnaires often suffer from high non-response rates, which means that their conclusions may be biased due to selection effects.

To improve the feasibility of MTMM studies, Saris et al. (2004) proposed the so-called *split-ballot MTMM design*. Respondents are divided at random into a number of groups and each group is assigned a questionnaire which contains all the MTMM questions for a different subset of the K proposed methods. Thus, for each respondent only some of the y_{kj} are observed, but all y_{kj} are observed for at least some of the respondents. Saris et al. (2004) showed that, for some judicious choices of this split-ballot design, it is possible to estimate all parameters of the true-score or classical MTMM model from the joint information in all groups, using the fact that the missing observations are "missing by design". Table 2.2 shows two examples of feasible split-ballot MTMM designs for $K = 3$ methods, with two and three groups of respondents, respectively. In these examples, each respondent has to answer questions for only two of the methods. Therefore, much shorter questionnaires can be used than for the full MTMM design. Revilla and Saris (2013) tested these split-ballot MTMM designs in a simulation study. They found that, for samples of small to moderate sizes, the two-group design in Table 2.2 often led to problems during estimation due to empirical underidentification. The three-group design in Table 2.2 performed well under a variety of conditions.

Since evaluating the measurement quality of survey variables by an MTMM design is costly – both in terms of resources and response burden –, attempts have been made to generalise the results of MTMM studies so that they might be used to predict the reliability and validity of variables in other surveys. Andrews (1984) was the first to perform a meta-analysis of several studies based on the classical MTMM design. He used multiple classification analysis to find a predictive model for the reliability and (indicator) validity coefficients of measured variables in six MTMM studies, as a function of a large set of characteristics of the underlying survey questions. Descriptions of similar, larger meta-analyses for the true-score

MTMM design can be found in Saris and Münnich (1995) and Scherpenzeel and Saris (1997). Further work in this area led to the development of the *Survey Quality Predictor* (SQP) computer program (Saris and Gallhofer, 2007; Saris et al., 2011). This program enables users to predict the reliability, validity and method effect of a proposed survey question based on its characteristics, without having to carry out an MTMM study themselves. SQP makes use of a database with the results of a large number of previous MTMM studies.

Repeated multi-method designs

Saris (1990a) suggested an alternative design to estimate the measurement quality parameters of Section 2.3.1, called the *repeated multi-method* (RMM) design. Here, a two-wave panel survey is used in which a single variable of interest F is measured by K different methods $M_1, \ldots, M_K$ at both points in time. A strong assumption of this approach is that a respondent's score on the variable of interest and indeed his/her true score for each method have not changed between the two points of measurement, which are supposed to be fairly close in time. If y_{tk} denotes the observed value with method M_k at time point t ($t \in \{1, 2\}$), then the true-score model (2.2) for these measurements can be written as the following SEM:

$$y_{tk} = T_k + \epsilon_{tk}, \tag{2.20}$$

$$T_k = a_k + b_k F + U_k. \tag{2.21}$$

In (2.21), the contributions of method M_k and its unique component u_k have been combined into a single disturbance term U_k, because the method effect cannot be identified under this RMM design. It is assumed that $\text{cov}(\epsilon_{tk}, \epsilon_{t'k'}) = 0$ and $\text{cov}(U_k, U_{k'}) = 0$ for all pairs of distinct measures.

After standardisation of y_{tk}, T_k and F in (2.20)–(2.21), the following equations are obtained:

$$y^s_{tk} = h^s_{tk} T^s_k + \epsilon^*_{tk},$$
$$T^s_k = b^s_k F^s + U^*_k,$$

with

$$h^s_{tk} = \sqrt{\frac{\text{var}(T_k)}{\text{var}(y_{tk})}},$$

$$b^s_k = b_k \sqrt{\frac{\text{var}(F)}{\text{var}(T_k)}}.$$

Thus, it is seen that the reliability, true-score validity and indicator validity coefficients of y_{tk} are given by h^s_{tk}, b^s_k and $h^s_{tk}b^s_k$, respectively. One would expect $h^s_{1k} = h^s_{2k}$ to hold (i.e., y_{1k} and y_{2k} should be equally reliable), but this restriction is not enforced by the model.

Saris and Andrews (1991) noted that the RMM design with a single variable of interest requires $K \geq 3$ methods for the model to be identified. This again leads to practical problems due to memory effects and a high response burden. Alternatively, two or more (correlated) variables of interest can be modelled jointly, in which case the use of $K \geq 2$ methods for each F_j is sufficient for identification. In contrast to the MTMM approach, the RMM approach does not require that the *same* set of methods is used to measure each variable of interest.

Quasi-simplex designs

The designs discussed so far all assume that it is possible to obtain multiple measurements of a variable of interest for the same respondents by different methods. This is necessary to separate the true score T into its valid (F) and invalid components [M and u in the formulation of (2.2)]. A glance at the expressions in Table 2.1 reveals that one needs this separation to estimate the true-score and/or indicator validity of an observed variable, but not to estimate its reliability. The *quasi-simplex* approach has been used as an alternative, less complicated data collection design that allows one to estimate only the reliability. This design was proposed by Heise (1969). An extensive discussion of this approach, including generalisations not considered here, can be found in Alwin (2007).

The quasi-simplex design requires a panel survey of at least three waves, where the same variable of interest F is measured by the same method M during each wave. In contrast to the RMM design, the quasi-simplex design explicitly models the development of the true score of the observed variable between the points of measurement, so it does not require the assumption that the variable of interest remains constant over time. If y_t and T_t denote the observed variable and its true score at wave t, respectively, then the quasi-simplex version of model (2.2) is as follows:

$$y_t = T_t + \epsilon_t, \tag{2.22}$$

$$T_t = c_t + d_t T_{t-1} + v_t, \tag{2.23}$$

where c_t and d_t are coefficients of the regression of T_t on T_{t-1} and v_t denotes a disturbance term. The measurement errors ϵ_t and disturbances v_t are assumed to be uncorrelated to each other and to measurement errors and disturbances at other time points.

After standarisation of all y_t and T_t, (2.22)–(2.23) are replaced by

$$y_t^s = h_t^s T_t^s + \epsilon_t^*,$$
$$T_t^s = d_t^s T_{t-1}^s + v_t^*,$$

where

$$h_t^s = \sqrt{\frac{\text{var}(T_t)}{\text{var}(y_t)}},$$
$$d_t^s = d_t \sqrt{\frac{\text{var}(T_{t-1})}{\text{var}(T_t)}}.$$

By comparison to (2.5), it is seen that h_t^s corresponds to the reliability coefficient of y_t. In fact, under the quasi-simplex design with three waves, only the reliability of y_2 is identified and one needs to make the assumption that $h_1^s = h_2^s = h_3^s$. More generally, this design can be used to estimate the reliability of y for all waves of the panel survey except the first and last one (Saris and Andrews, 1991).

Congeneric measures designs

Finally, we consider the case that a number of traits $F_1, \ldots, F_J$ are measured by various methods at a single point in time, where different methods may be used for each trait. (This is similar to the repeated multi-method design but without the repetition over time.) Let y_{kj} denote the observed value for F_j measured by its k^{th} method (say, M_{kj}). Since the lack of repetition of methods makes it impossible to separate random and systematic measurement error, we consider the following simplified version of model (2.2):

$$y_{kj} = a_{kj} + b_{kj} F_j + e_{kj}. \tag{2.24}$$

The disturbance term is now equal to $e_{kj} = g_{kj} M_{kj} + u_{kj} + \epsilon_{kj}$. It is assumed that $\text{cov}(e_{kj}, e_{k'j'}) = 0$ for all distinct measures. Jöreskog (1971) considered (2.24) as a general model for so-called *congeneric measures*. As noted in Section 1.4, this model is identified for a single variable of interest that is measured by at least three methods, or for $J \geq 2$ correlated variables of interest with two or more methods each.

Upon standardisation, (2.24) is replaced by

$$y_{kj}^s = b_{kj}^s F_j^s + e_{kj}^*, \tag{2.25}$$

with

$$b_{kj}^s = b_{kj} \sqrt{\frac{\text{var}(F_j)}{\text{var}(y_{kj})}}.$$

Note that if the true-score model of Section 2.3.1 holds, then, by virtue of assumptions (2.3)–(2.4), the simpler model (2.24) can be used to obtain a consistent estimate of b_{kj} and a_{kj}. Therefore, to estimate these parameters of the true-score model, we can always use the congeneric measures model, even if the data were originally collected by a more complicated design (e.g., the MTMM design). In fact, variations on the congeneric measures model will be used for that purpose in all applications in this thesis.

It also follows that b^s_{kj} corresponds to the indicator validity coefficient of y_{kj} given by (2.9). The true-score validity and reliability of y_{kj} cannot be estimated directly under the congeneric measures model. Through inequalities (2.13) and (2.14), the indicator validity of y_{kj} does provide a lower bound for both the reliability and true-score validity of y_{kj}. (Similar inequalities hold for the associated validity and reliability coefficients.) This is particularly informative if y_{kj} has high indicator validity; for instance, if $\text{IVC}(y_{kj}) = b^s_{kj} = 0.95$ then it follows that $0.95 \leq \text{RC}(y_{kj}) \leq 1$ and $0.95 \leq \text{TVC}(y_{kj}) \leq 1$.

Identification and intercept bias

The models discussed so far in this section are all examples of SEMs with latent variables. The parameters of such a model are not naturally identified, regardless of the number of observed variables that are included in the model. This is true because, without additional assumptions, each latent variable is defined only up to an arbitrary linear transformation (Bollen, 1989). That is to say, the observed data do not fully determine the scale (or metric) of a latent variable. To assign a metric to the latent variables and achieve model identification, some restrictions should be added on the parameters of an SEM.

For the congeneric measures model (which will be our main focus in the rest of this thesis), Little et al. (2006) discussed three types of restrictions that can be used to fix the metric of the latent variables F_j. The two devices that are used most commonly in practice are: *standardising* the latent variables or using *reference indicators*. The latter approach introduces restrictions $a_{kj} = 0$ and $b_{kj} = 1$ for one observed variable for each j, thereby giving the latent variable F_j the same metric as the observed variable y_{kj}.

For the purpose of estimating the indicator validity of the observed variables in a congeneric measures model, the intercept parameters a_{kj} can be ignored and the other parameters can be identified by any of the three methods in Little et al. (2006). The resulting values of IVC will be identical, as these depend only on standardised parameters of the model. More generally, for all SEMs considered in this chapter, the values of RC, TVC and IVC do not depend on the method of identification

that is used. In particular, identification can be achieved by standardising the latent variables, as we have done throughout this section.

If one is also interested in estimating the intercept bias of each observed variable with respect to the true value, the choice of identification device does matter. Standardisation is not an option in this case. The reference indicator approach could be used here, but only if it is reasonable to assume that the reference indicators provide unbiased measurements of the underlying true values – that is, the true intercept a_{kj} of each variable that is chosen as a reference indicator should be 0 and its true slope b_{kj} should be 1. The most obvious way to ensure this is by choosing as reference indicators "gold standard" measures of the variables of interest which – for practical purposes – can be assumed to contain no errors. According to Bielby (1986a,b), this is the only way to obtain meaningful estimates of the measurement intercept and slope parameters in applications where true values are of interest.

Sobel and Arminger (1986) noted that it is sufficient to collect "gold standard" data only for a random subsample of the units in the original data set. By formulating the identification problem as a missing-data problem, one can still use the partially observed variables as reference indicators to fix the scale of the latent variables for all units. This is very convenient, as the collection of "gold standard" data can be difficult, costly or otherwise inconvenient in practice. This approach is discussed more extensively in Scholtus (2014b) and Chapter 6 of this thesis.

2.3.3 Contamination models and other measurement models

In the literature, many variations can be found on the error models and designs that were discussed in Section 2.3.2. For instance, Kenny (1976) proposed an alternative SEM formulation that can be used to analyse an MTMM matrix, known as the *correlated uniqueness model*. Other authors have proposed measurement error models that are multiplicative rather than additive; see, e.g., Wothke and Browne (1990) and a discussion by Saris and Andrews (1991). More recently, Oberski et al. (2017) have proposed an extension of the MTMM model that is particularly aimed at evaluating the quality of administrative as well as survey variables. This model is not an SEM but it does fit within a more general class of latent-variable models that was introduced by Skrondal and Rabe-Hesketh (2004).

Under any SEM the event of observing a value that is equal to the true value (or even the true score) occurs with probability zero for any variable which contains measurement errors. That is to say, all observed values of an error-prone variable are supposed to be wrong to some extent. As noted in Section 1.4, the alternative assumption that error-free observations occur with some non-zero probability leads to so-called *contamination models* (Bound et al., 2001; Di Zio and Guarnera,

2013).

A contaminated extension of the congeneric measures model (2.24) is:

$$y_{kj} = (1 - z_{kj})F_j + z_{kj}(a_{kj} + b_{kj}F_j + e_{kj}), \tag{2.26}$$

where z_{kj} denotes a random variable on $\{0, 1\}$ that generates an error in y_{kj} when $z_{kj} = 1$. Thus, the observed variable y_{kj} has a distribution which is a mixture of two components: part of the observed values are error-free and the other observed values contain errors according to the congeneric measures model. Di Zio and Guarnera (2013) called this an *intermittent* error mechanism. Model (2.26) is an example of a so-called *finite mixture model* (McLachlan and Peel, 2000).

With this contamination model, an important new measurement quality parameter is the probability of observing an error in y_{kj}: $\pi_{kj} = P(z_{kj} = 1)$. Since the indicator validity of the observations in the error-free part of the data is equal to 1, a reasonable definition for the overall indicator validity coefficient of y_{kj} under this model is:

$$\begin{aligned}\text{IVC}(y_{kj}) &= (1 - \pi_{kj}) \times \text{IVC}(y_{kj}|z_{kj} = 0) + \pi_{kj} \times \text{IVC}(y_{kj}|z_{kj} = 1) \\ &= (1 - \pi_{kj}) + \pi_{kj} b_{kj} \sqrt{\frac{\text{var}(F_j|z_{kj} = 1)}{\text{var}(y_{kj}|z_{kj} = 1)}},\end{aligned}$$

where b_{kj} and the variances of F_j and y_{kj} should now be evaluated only in the error-prone part of the data.

Guarnera and Varriale (2015, 2016) proposed to use model (2.26) for linked administrative and survey data (with the assumption that $a_{kj} = 0$ and $b_{kj} = 1$ for all variables). They showed how to estimate this model for $J = 1$ and $K = 3$. An interesting feature of the contamination model is that a_{kj} and b_{kj} can be identified without "gold standard" data (Robinson, 2016). An application of this model will be discussed in Chapter 7 of this thesis.

2.3.4 Administrative data

The true-score measurement error model (2.2) and the designs discussed in Section 2.3.2 were originally developed for psychological tests and later extended to survey variables. Conceptually, the same model could be applied also to administrative variables. The interpretations of the variable of interest F, the true score T and the random measurement error ϵ remain the same for administrative data. In particular, the reliability and validity coefficients defined in Section 2.3.1 are useful summary statistics of the measurement quality of administrative variables.

The interpretation of the method component M in this context is less clear. The measurement procedures by which administrative data are collected usually fall

outside the influence of the researcher who uses the data (Bakker and Daas, 2012); in fact, the researcher often has only limited knowledge about these measurement procedures. Hence, it is usually not possible for a researcher to identify different "methods" behind administrative variables at a more detailed level than that of the administrative source itself. That is to say, in the context of administrative data, each register constitutes a different method, but it is not possible to distinguish different methods within the same register. This somewhat devalues the concept of a "method".

As a consequence, the role of the unique component u becomes more important for administrative data than for survey data. In particular, the assumption of the true-score MTMM model that $u = 0$ seems untenable for most administrative variables. In fact, if the administrative source itself is taken as the method, then any systematic effects caused by different measurement procedures for different variables within a single source will contribute to u in (2.2). Such systematic differences are likely to occur in practice, just as they also occur for different variables in a single survey. Recall, for instance, that some variables in an administrative source may be more important than others to the register owner.

To estimate the various components of model (2.2), multiple measurements are needed. It is less obvious how to obtain multiple, independent measures with administrative variables than with survey variables. The designs in Section 2.3.2 all assume to some extent that it is possible to choose which measurement procedures are applied to each unit in the population, that is to say, that it is possible to conduct methodological experiments as part of the data collection procedure. With surveys based on questionnaires, this is certainly possible in theory, although in practice some limitations do exist in terms of costs, available resources, bounds on response burden, etc. By contrast, methodological experiments are almost never possible with administrative data sources.

Biemer (2004) noted that an alternative way to obtain multiple measurements may be to link different data sources together. In particular, linking administrative data and (existing) survey data can be a powerful approach to obtain multiple measurements at virtually no additional costs and no additional response burden. Bakker (2012) obtained linked data from one administrative data source and one survey and applied the congeneric measures model (2.25) to estimate the indicator validity of $J = 4$ variables in both data sources. With this type of application in mind, Scholtus and Bakker (2013b) conducted a simulation study of the robustness to minor model misspecifications of validity estimates in model (2.25) when each variable of interest is measured by two methods.

Linking administrative data to survey data might also make it possible to apply

the MTMM design to administrative data. As noted above, the true-score MTMM model may be inappropriate for administrative data, but one could still use the classical MTMM model to obtain estimates of indicator validity and reliability (as defined under that model). Recall that at least three methods and three correlated variables of interest are needed to identify all parameters of the MTMM model. A direct application of the MTMM design is therefore possible if one can link (A) an administrative data source to a survey in which at least two different methods are used, or (B) two independent administrative data sources to a survey with a single method, or (C) three independent administrative data sources among themselves, provided that at least three correlated variables of interest are measured in all sources. Situations (B) and (C) rarely arise. Situation (A) could be achieved in a trivial manner by conducting a new survey which is specifically designed to evaluate the measurement quality of the administrative data, but this is unattractive from a practical point of view. It is desirable to make use of data that are already available as much as possible (Van Delden et al., 2014).

In practice, it may often be relatively easy to obtain three different measurements of several variables of interest from three different sources: one administrative data set and two existing surveys. Usually, these data would not be sufficient to apply the original MTMM design, due to a lack of overlapping units between the two surveys. The split-ballot MTMM design introduced by Saris et al. (2004) could be of interest for this situation. In particular, the available data match the two-group design in the left panel of Table 2.2 exactly, if methods M_1 and M_2 denote the two (non-overlapping) surveys and method M_3 denotes the administrative source. Unfortunately, as noted above, results on simulated data suggest that this particular split-ballot design often leads to problems during model estimation.

If one has only one administrative data set and one linked survey (with a single method), the MTMM design with $K = 2$ methods cannot be used directly as it is not identified. In this situation, one final option might be to use the SQP software of Saris and Gallhofer (2007) to obtain separate predictions of the reliability, validity and method parameters of the survey variables. By fixing the measurement parameters for the survey variables in the MTMM model to their predicted values, one could obtain an identified model which can be used to estimate the remaining parameters for the administrative variables. It remains to be seen whether this approach can yield accurate results in practice.

Finally, the RMM and quasi-simplex designs may be useful for some applications where administrative data are available over time, similar to panel surveys. An important assumption of these designs is that the measurements at different time points can be considered independent. This assumption may be reasonable

for some administrative sources that use a "survey-like" data collection method. However, as noted in Section 1.2.2, longitudinal administrative data are often generated by the (automatic or manual) registration of events. The data then remain unchanged until a new event is registered. In this case, although it is possible (and much less costly than for panel surveys) to obtain an administrative data set of measurements at several points in time, the measurement errors at different time points will be correlated strongly. The RMM and quasi-simplex designs are therefore inappropriate for administrative data of this type. In fact, even with "survey-like" longitudinal administrative data, the assumption that the measurement errors in these data are uncorrelated over time is often problematic, because the nature of administrative reporting means that errors that go unnoticed will often be repeated in the future.

In principle, it is possible to extend the quasi-simplex model to allow for correlated measurement errors over time. For categorical variables, latent class models with Markov assumptions have been used in this context; see, e.g., Langeheine (1994), Bassi et al. (2000), Biemer (2011) and Pavlopoulos and Vermunt (2015).

2.4 Models for data editing

It is interesting to examine briefly the relation between data editing and measurement error modelling. Although most data editing methods were developed heuristically, by making gradual improvements based on practical experiences, in fact some modelling assumptions about measurement errors are implicitly contained in these methods. As noted in Section 1.2.3, all editing methods suppose that true values exist and can be obtained in principle if measurement is done with sufficient care. In particular, it is assumed that the majority of values in a data set are observed correctly during data collection (i.e., errors are intermittent). For those values that were not observed correctly in the initial response, it is assumed that the true values can be obtained later by subject-matter experts.

These assumptions are not in line with the true-score error model of Section 2.3.1, but they are in line with the contamination model of Section 2.3.3. It is therefore not surprising that publications in the data editing literature that do feature explicit error models are often based on contamination-like models. For instance, Di Zio and Guarnera (2013) discussed a model-based procedure for selective editing that assumes that the observed data come from a mixture of multivariate normal distributions.

Naus et al. (1972) introduced a simple measurement error model for survey data that influenced the development of the Fellegi-Holt paradigm. Under this

model, each observed variable y_k has a certain probability π_k of containing an error, and erroneous values are supposed to be random draws from an unspecified distribution. If it is assumed that π_k does not depend on the true value of y_k and that errors in different variables occur independently, then it can be shown that maximum likelihood estimation of the error pattern in an observed record is approximately equivalent to minimising (2.1) with confidence weights given by

$$w_k = -\log\left(\frac{\pi_k}{1-\pi_k}\right).$$

A proof of this result was first given by Liepins (1980).

Although this simple model is unlikely to hold for any real data set, it does provide some motivation for the Fellegi-Holt paradigm. In particular, the derivation in Liepins (1980) suggests conditions under which the Fellegi-Holt-based error localisation problem might be expected to work well, and also provides some direction for choosing confidence weights. In Chapter 5, we will consider an extension of the model of Naus et al. (1972) to motivate a generalisation of the Fellegi-Holt paradigm.

2.5 Conclusion

To conclude this review of existing work on editing and estimation of measurement errors, we briefly describe how the contents of the remaining chapters of this thesis build on this work. Regarding the editing approach, it was discussed in Section 2.2 that automatic editing methods could be very useful for increasing the efficiency of editing processes, in particular for administrative data. However, as mentioned in Section 1.3, evaluations of automatic editing in practice have shown that with the current methodology the quality of automatically-edited data is often quite low; therefore, automatic editing is currently applied only on a limited scale as a supplement to selective manual editing. It is therefore important to develop improved methods that will allow automatic editing to be used more widely in practice.

In this thesis, the development of improved methods for automatic editing will be addressed in Chapters 3–5. In Chapter 3, we will focus on the automatic correction of systematic errors, by developing new algorithms that can correct two types of generic systematic errors that are common in data for business statistics: sign errors and rounding errors. In Chapters 4 and 5, we will focus on random errors. We will develop two extensions of the Fellegi-Holt paradigm that relax some of the assumptions of this paradigm by incorporating, respectively, soft edit rules (Chapter 4) and complex edit operations (Chapter 5). Both extensions lead to a new

formulation of the error localisation problem that is more flexible and therefore has the potential to improve the quality of automatic editing.

Regarding the estimation approach, it was noted that applications of this approach in official statistics will often relate to univariate statistics such as population totals and means. Such applications therefore require not just estimates of the validity and reliability of an observed variable, but also of its bias in terms of the intercept and slope parameters of the measurement error model. As noted above, this has implications for the identification of the model. In Chapter 6, we will describe an application of structural equation modelling for administrative data on businesses, to determine whether these data can be used for the production of economic statistics, possibly after a correction for measurement bias. To identify the SEM, we will use the device suggested by Sobel and Arminger (1986) of collecting "gold standard" data for an audit sample (a random subsample of the original sample). We will work out some adjustments to the estimation procedure that are necessary to incorporate the audit sample while also accounting for a complex sample design and non-normality of the data.

In official statistics, the editing approach is currently used with the tacit assumption that any errors that remain in the edited data have a negligible influence on published estimates. It is clear that this is not guaranteed to be the case, in particular when the editing process involves selective editing and/or automatic editing. In Chapter 7, we will therefore check this assumption by modelling the errors in a data set before and after editing. We will use two different models: an SEM and a variation on the contamination model of Guarnera and Varriale (2016). As noted in Section 2.4, the latter type of model seems more appropriate as it is more in line with the assumptions of editing methods.

An interesting feature of the application in Chapter 7 is that it combines the two approaches to handling measurement errors, editing and estimation. This combination of approaches could be developed further, as we will discuss in Chapter 8. Firstly, by estimating the residual measurement error after editing, it may become possible to correct statistical output for the effect of these errors, rather than simply making the untested assumption that this effect is negligible. Secondly, if the editing process is followed by an estimation step that corrects for residual measurement errors, it is no longer necessary to edit the data until the remaining errors have a negligible influence on the (uncorrected) estimates and it may therefore be possible to reduce the amount of manual editing. Thus, incorporating a combination of the editing and estimation approach in a regular statistical production process could lead to statistical data of a higher quality at lower costs.

Chapter 3

Algorithms for Correcting Sign Errors and Rounding Errors in Business Survey Data

The contents of this chapter have been published in *Journal of Official Statistics* as Scholtus (2011a). That version omitted Appendix 3.A. Otherwise, the chapter is identical to the article, apart from some minor textual corrections and adjustments. Elements of this article were also incorporated into Chapter 2 of De Waal et al. (2011).

3.1 Introduction

It is well-known that data collected in a survey or register contain errors. In the case of a survey, these errors may be introduced when the respondent fills in the questionnaire or during the processing of survey forms at the statistical office. It is important to resolve the errors by editing the data, because figures based on erroneous data may be biased or logically inconsistent. For the structural business statistics, all survey variables are quantitative and many (linear) relationships between them can be formulated. Thus, a set of constraints called *edit rules* is established. Two examples of edit rules are

$$\textit{profit} = \textit{turnover} - \textit{costs}$$

and

$$\begin{aligned} &\textit{number of employees} \text{ (in persons)} \\ \geq\ &\textit{number of employees} \text{ (in full time equivalent (fte))}. \end{aligned}$$

If the data in a particular record violate an edit rule, the record is found to be inconsistent and it is deduced that some variable(s) must be in error.

A distinction is often made between *systematic errors* and *random errors*. According to EDIMBUS (2007, §3.3.1), an error is systematic if it is reported consistently over time by different respondents. This type of error occurs when respondents consistently misunderstand a survey question, e.g., by reporting financial amounts in Euros rather than the requested multiples of $1,000$ Euros [this example is called a *unity measure error*, cf. Di Zio et al. (2005a)]. A fault in the data processing system might also introduce systematic errors. Since it is reported consistently by a number of respondents, an undiscovered systematic error can lead to biased aggregates. Once identified, a systematic error can be corrected deductively, because the underlying error mechanism is assumed to be known. Random errors on the other hand do not have a structural cause. An example of a random error occurs when a particular "1" on a particular survey form is accidentally keyed in as a "7" during data processing.

At Statistics Netherlands, *selective editing* is used to clean the data collected for structural business statistics (De Jong, 2002). This means that only records containing (potentially) influential errors are edited manually by subject-matter specialists, whereas the remaining records are edited automatically. For the latter step, many statistical institutes have implemented error localisation algorithms based on a generalisation of the Fellegi-Holt paradigm (Fellegi and Holt, 1976), which states that the smallest possible (weighted) number of variables should be labelled erroneous such that the record can be made consistent with every edit rule. This paradigm is based on the assumption that the data contain only random errors.

Examples of software packages for automatic editing based on the Fellegi-Holt paradigm are: GEIS (Kovar and Whitridge, 1990) and its successor Banff (Banff Support Team, 2003), SPEER (Winkler and Draper, 1997), DISCRETE (Winkler and Petkunas, 1997) and AGGIES (Todaro, 1999). At Statistics Netherlands, the software package SLICE was developed for automatic editing. SLICE also uses an error localisation algorithm based on the Fellegi-Holt paradigm; a description of this algorithm can be found in De Waal and Quere (2003) and De Waal (2003a).

A *plausibility indicator* is calculated for each record to assess whether it may contain influential errors and should be edited manually (Hoogland, 2006). The plausibility indicator is calibrated such that all records that receive a score above a certain threshold are deemed suitable for automatic editing. Only the records with the lowest scores on the plausibility indicator are edited manually. In addition to this, the data of very large companies are always edited manually, since it is considered impossible to construct meaningful aggregates unless this part of the data set is error-free.

Selective editing leads to a more efficient editing process than traditional edit-

ing (where every record is edited by hand), because part of the data stream is not reviewed by subject-matter specialists any more. However, Fellegi-Holt based algorithms for automatic error localisation are not considered suitable for editing records that contain either influential or systematic errors. In particular, the correction of systematic errors often requires changing more variables than the Fellegi-Holt paradigm suggests. For instance, a unity measure error affects all financial variables on the survey form, but it leads to few violated edit rules. Furthermore, in practice the error localisation problem becomes too complicated if many variables contain erroneous values and/or if many edit rules are violated (De Waal and Quere, 2003). For the Netherlands' structural business survey, because of the large number of edit rules involved, the error localisation problem tends to be too complex to solve with SLICE if more than, say, about 15 variables have to be changed.

To preserve the quality of the statistical output, only records that contain a limited number of non-influential random errors should be edited automatically. Ideally, the plausibility indicator filters out all records containing influential errors or too many inconsistencies. Prior to this, several types of *obvious errors* can be detected and resolved automatically in a separate step. A systematic error is called obvious if it can be detected "easily", i.e., by applying some basic, specific search algorithm. Obvious errors are easy to correct, because the underlying cause of the error is detectable. An example of such an error is the unity measure error, which can be detected by comparing the reported amounts with reference values (see Section 3.2).

It is useful to detect and correct obvious inconsistencies as early as possible in the editing process, since it is a waste of resources if subject-matter specialists have to deal with them. When obvious inconsistencies are corrected in a separate step, before the plausibility indicator is calculated, the efficiency of the selective editing process increases because more records will be eligible for automatic editing. Moreover, solving the error localisation problem becomes easier once obvious inconsistencies have been removed, since the number of violated edit rules becomes smaller.

Furthermore, since obvious inconsistencies are systematic errors, they can be corrected more accurately by a specific, deductive algorithm than by a general error localisation algorithm based on the Fellegi-Holt paradigm. The deductive algorithm uses knowledge of the underlying cause of the error, so that the corrected values are true values, assuming that the error has been detected correctly. By contrast, the Fellegi-Holt-based algorithm does not use this knowledge, and the values returned by this algorithm are consistent with respect to the edit rules but are not necessarily true values. Hence, if a certain type of systematic error is expected

to occur commonly and if a specific, trustworthy search routine is available to detect and correct it, it makes sense to apply this routine rather than to rely on the general algorithm used by SLICE. After all, if the error is left in the data to be resolved by SLICE, at best the general algorithm will detect and correct the error the same way the simple algorithm would have done, but at a much higher computational cost.

The currently implemented editing process for the structural business statistics at Statistics Netherlands contains a step during which three obvious systematic errors are treated. Section 3.2 provides a brief description of this step. Other obvious inconsistencies have been discovered by comparing raw and manually edited data from past cycles of the structural business survey. This study has resulted in several new deductive correction methods (Scholtus, 2008, 2009).

The purpose of this chapter is to present new algorithms for the detection and correction of two types of errors. Section 3.3 deals with so-called sign errors and interchanged revenues and costs. Section 3.4 describes a heuristic method for correcting rounding errors. Rounding errors are not obvious inconsistencies in the true sense of the word (they can be considered as random errors), but the efficiency of the editing process is expected to increase if these errors are also treated separately. Section 3.5 presents some results of an application of the two algorithms to real-world data. Finally, a few concluding remarks follow in Section 3.6.

Due to item non-response, the unedited data contain a substantial number of missing values. The algorithms described in this chapter assume that these missing values have been temporarily replaced with zeros. This is merely a precondition for determining which edit rules are violated and which are satisfied, and should not be considered a full imputation. When the obvious inconsistencies have been corrected, all placeholder zeros should be replaced by missing values again, to be imputed by a valid method later. Clearly, this requires that placeholder zeros can be distinguished from actually reported zeros.

3.2 Current approach at Statistics Netherlands

The currently implemented editing process for structural business statistics at Statistics Netherlands contains a step in which three kinds of obvious systematic errors are detected. These errors are treated deductively before any other correction is made in the data of the processed survey forms.

The first of these obvious inconsistencies is the unity measure error from Section 3.1: the amounts on the survey form are sometimes reported in Euros instead of in $1,000$ Euros. This particular unity measure error is also referred to as a

uniform $1,000$*-error*. It is important to detect this error because otherwise publication figures of all financial items will be overestimated. Depending on which auxiliary information is available, two methods are used to detect uniform $1,000$-errors. If the respondent is present in the VAT register, the amount of turnover in the register is compared to the reported turnover in the survey. For the other respondents, the amount of reported turnover per reported number of employees (in fte) is compared to its median in the edited data of the previous year. If a large discrepancy is found by either method, all financial amounts reported by the respondent are divided by $1,000$. This is how uniform $1,000$-errors are currently detected at Statistics Netherlands. Different methods are suggested by Di Zio et al. (2005a) and Al-Hamad et al. (2008).

The second obvious inconsistency occurs when a respondent adds a redundant minus sign to a reported value. This sometimes happens with variables that have to be subtracted, even though there already is a printed minus sign on the survey form. As a result, the value of the variable becomes incorrectly negative after data processing. The resulting inconsistency can be detected and corrected easily: the reported amount is simply replaced by its absolute value.

The third and final obvious inconsistency occurs when respondents report component items of a sum but leave the corresponding total blank. When this is detected, the total value is calculated from the reported items and filled in automatically.

3.3 Sign errors

3.3.1 The profit-and-loss account

The *profit-and-loss account* is a part of the questionnaire used for structural business statistics where the respondent has to fill in a number of balance amounts. These balance variables are denoted by $x_0, x_1, \ldots, x_{n-1}$. A final balance amount x_n called the *pretax results* is found by adding up the other balance variables. That is, the data should conform to the following edit rule:

$$x_0 + x_1 + \cdots + x_{n-1} = x_n. \tag{3.1}$$

Rule (3.1) is sometimes referred to as the *external sum*. A balance variable is defined as the difference between a revenue item and a cost item. If these items are also asked in the questionnaire, the following edit rule should hold:

$$x_{k,r} - x_{k,c} = x_k, \tag{3.2}$$

where $x_{k,r}$ denotes the revenue item and $x_{k,c}$ the cost item. Rules of this form are referred to as *internal sums*.

A statistical office may decide not to ask the revenues and costs for every balance variable in the survey, to limit the burden on respondents. To keep the notation simple but sufficiently general, it is assumed that the balance variables are arranged such that only $x_0, \ldots, x_m$ are split into revenues and costs, for some $m \in \{0, 1, \ldots, n-1\}$. Thus, the following set of edit rules is used:

$$\begin{cases} x_0 &= x_{0,r} - x_{0,c} \\ & \vdots \\ x_m &= x_{m,r} - x_{m,c} \\ x_n &= x_0 + x_1 + \cdots + x_{n-1} \end{cases} \tag{3.3}$$

In this notation the 0^{th} balance variable x_0 stands for *operating results*, and $x_{0,r}$ and $x_{0,c}$ represent *operating returns* and *operating costs*, respectively.

3.3.2 Sign errors and interchanged revenues and costs

Table 3.1 displays the structure of the profit-and-loss account from the structural business statistics questionnaire that was used at Statistics Netherlands until 2005. The associated edit rules are given by (3.3), with $n = 4$ and $m = n - 1 = 3$. Table 3.1 also displays four example records that are inconsistent. The first three example records have been constructed for this chapter with nice "round" amounts to improve readability, but the types of inconsistencies present were taken from actual records from the structural business statistics of 2001. The fourth example record contains realistic values.

In Example (a) two edit rules are violated: the external sum and the internal sum with $k = 1$. In this case, the profit-and-loss account can be made fully consistent with all edit rules by just changing the value of x_1 from 10 to -10 (see Table 3.2). This is the natural way to obtain a consistent profit-and-loss account here, since any other explanation would require more variables to be changed. Moreover, it is quite conceivable that the minus sign in x_1 was left out by the respondent or "lost" during data processing.

Two internal sums are violated in Example (b), but the external sum holds. The natural way to obtain a consistent profit-and-loss account here is by interchanging the values of $x_{1,r}$ and $x_{1,c}$, and also of $x_{3,r}$ and $x_{3,c}$ (see Table 3.2). By treating the inconsistencies this way, full use is made of the amounts actually filled in by the respondent and no imputation of synthetic values is necessary.

The two types of errors found in Examples (a) and (b) are referred to as *sign errors* and *interchanged revenues and costs*, respectively. For the sake of brevity,

Table 3.1: Structure of the profit-and-loss account in the structural business statistics until 2005, with four example records, (a) − (d).

Variable	Full name	(a)	(b)	(c)	(d)
$x_{0,r}$	Operating returns	$2,100$	$5,100$	$3,250$	$5,726$
$x_{0,c}$	Operating costs	$1,950$	$4,650$	$3,550$	$5,449$
x_0	Operating results	150	450	300	276
$x_{1,r}$	Financial revenues	0	0	110	17
$x_{1,c}$	Financial expenditure	10	130	10	26
x_1	Financial result	10	130	100	10
$x_{2,r}$	Provisions rescinded	20	20	50	0
$x_{2,c}$	Provisions added	5	0	90	46
x_2	Balance of provisions	15	20	40	46
$x_{3,r}$	Exceptional income	50	15	30	0
$x_{3,c}$	Exceptional expenses	10	25	10	0
x_3	Exceptional result	40	10	20	0
x_4	Pretax results	195	610	-140	221

the term *sign error* is also used to refer to both types. In an evaluation study at Statistics Netherlands, which compared raw data to manually edited data, it was found that a substantial number of respondents made these errors. Moreover, it was found that these errors were often not correctly identified by SLICE and hence resolved in a different way during automatic editing. In particular, it is very difficult to handle interchanged revenues and costs correctly by means of the Fellegi-Holt paradigm, because this requires changing two variables where it would actually suffice to change one. Therefore, it seems advantageous to add a separate detection step for sign errors at the beginning of the automatic editing process.

Sign errors and interchanged revenues and costs are closely related and should therefore be searched for by one detection algorithm. In the remainder of this section such an algorithm is formulated, working from the assumption that if an inconsistent record can be made to satisfy all edit rules in (3.3) by only changing signs of balance variables and/or interchanging revenue items and cost items, this is indeed the way the record should be corrected.

It should be noted that *operating returns* ($x_{0,r}$) and *operating costs* ($x_{0,c}$) differ from the other variables in the profit-and-loss account in the sense that they are

also present in other edit rules, connecting them to items from other parts of the survey. For instance, *operating costs* should equal the sum of *total labour costs*, *total machine costs*, etc. If $x_{0,r}$ and $x_{0,c}$ were interchanged to suit the 0^{th} internal sum, other edit rules might be violated. It is therefore not allowed to interchange $x_{0,r}$ and $x_{0,c}$ when detecting sign errors. Because of the way the questionnaire is designed, it seems highly unlikely that any respondent would mix up these two amounts anyway.

As stated above, a record contains a sign error if it satisfies the following two conditions:

- at least one edit rule in (3.3) is violated;
- it is possible to satisfy (3.3) by only changing the signs of balance amounts and/or interchanging revenue and cost items other than $x_{0,r}$ and $x_{0,c}$.

An equivalent way of formulating this is to say that an inconsistent record contains a sign error if the following set of equations has a solution:

$$\begin{cases} x_0 s_0 &= x_{0,r} - x_{0,c} \\ x_1 s_1 &= (x_{1,r} - x_{1,c})\, t_1 \\ &\vdots \\ x_m s_m &= (x_{m,r} - x_{m,c})\, t_m \\ x_n s_n &= x_0 s_0 + x_1 s_1 + \cdots + x_{n-1} s_{n-1} \\ (s_0, \ldots, s_n; t_1, \ldots, t_m) &\in \{-1, 1\}^{n+m+1} \end{cases} \tag{3.4}$$

Note that in (3.4) the x's are used as known constants rather than unknown variables. Thus, a different set of equations in $(s_0, \ldots, s_n; t_1, \ldots, t_m)$ is found for each record.

Moreover, once a solution to (3.4) has been found, it is immediately clear how to obtain a consistent profit-and-loss account: if $s_j = -1$ then the sign of x_j must be changed, and if $t_k = -1$ then the values of $x_{k,r}$ and $x_{k,c}$ must be interchanged. It is easy to see that the resulting record satisfies all edit rules (3.3). Since $x_{0,r}$ and $x_{0,c}$ may not be interchanged, no variable t_0 is present in (3.4).

Example. Consider (3.4) for Example (c) from Table 3.1:

$$\begin{cases} 300 s_0 &= -300 \\ 100 s_1 &= 100 t_1 \\ 40 s_2 &= -40 t_2 \\ 20 s_3 &= 20 t_3 \\ -140 s_4 &= 300 s_0 + 100 s_1 + 40 s_2 + 20 s_3 \\ (s_0, s_1, s_2, s_3, s_4; t_1, t_2, t_3) &\in \{-1, 1\}^8 \end{cases} \tag{3.5}$$

Table 3.2: Corrected versions of the example records from Table 3.1. Changes are shown in boldface.

Variable	Full name	(a)	(b)	(c)	(d)
$x_{0,r}$	Operating returns	$2,100$	$5,100$	$3,250$	$5,726$
$x_{0,c}$	Operating costs	$1,950$	$4,650$	$3,550$	$5,449$
x_0	Operating results	150	450	$\mathbf{-300}$	276
$x_{1,r}$	Financial revenues	0	$\mathbf{130}$	110	17
$x_{1,c}$	Financial expenditure	10	$\mathbf{0}$	10	26
x_1	Financial result	$\mathbf{-10}$	130	100	$\mathbf{-10}$
$x_{2,r}$	Provisions rescinded	20	20	$\mathbf{90}$	0
$x_{2,c}$	Provisions added	5	0	$\mathbf{50}$	46
x_2	Balance of provisions	15	20	40	$\mathbf{-46}$
$x_{3,r}$	Exceptional income	50	$\mathbf{25}$	30	0
$x_{3,c}$	Exceptional expenses	10	$\mathbf{15}$	10	0
x_3	Exceptional result	40	10	20	0
x_4	Pretax results	195	610	-140	221

This system has the (unique) solution $(s_0 = -1, s_1 = 1, s_2 = 1, s_3 = 1, s_4 = 1; t_1 = 1, t_2 = -1, t_3 = 1)$. This solution shows that the value of x_0 should be changed from 300 to -300 and that the values of $x_{2,r}$ and $x_{2,c}$ should be interchanged. This correction indeed yields a fully consistent profit-and-loss account with respect to (3.3), as can be seen in Table 3.2. □

An important question is: does system (3.4) always have a unique solution? Scholtus (2008) derives the following sufficient condition for uniqueness: if $x_0 \neq 0$, $x_n \neq 0$, and if the equation

$$\lambda_0 x_0 + \lambda_1 x_1 + \cdots + \lambda_{n-1} x_{n-1} = 0$$

does not have any solution $(\lambda_0, \lambda_1, \ldots, \lambda_{n-1}) \in \{-1, 0, 1\}^n$ for which at least one term $\lambda_j x_j \neq 0$, then an inconsistency in the record can be resolved by changing signs and/or interchanging revenues and costs in at most one way. It appears that this condition is usually satisfied; in the data examined at Statistics Netherlands, the condition holds for over 95 per cent of all records. In the rare case that system (3.4) has more than one solution, it makes sense to assume that most of the original values were reported correctly by the respondent and therefore choose the solution

with the smallest number of -1's among s_j and t_k. This assumption will indeed be made in the next subsection.

3.3.3 A binary linear programming problem

Detecting a sign error in a given record is equivalent to solving the corresponding system (3.4). Therefore all that is needed to implement the detection of sign errors is a systematic method to solve this system. Before addressing this point, it is convenient to write (3.4) in matrix notation to shorten the expressions. Define the $(m+2) \times (n+1)$-matrix $\mathbf{U}$ by

$$\mathbf{U} = \begin{pmatrix} x_0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ 0 & x_1 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & x_m & 0 & \cdots & 0 & 0 \\ x_0 & x_1 & \cdots & x_m & x_{m+1} & \cdots & x_{n-1} & -x_n \end{pmatrix}$$

and define the $(m+2) \times (m+1)$-matrix $\mathbf{V}$ by

$$\mathbf{V} = \begin{pmatrix} x_{0,r} - x_{0,c} & 0 & \cdots & 0 \\ 0 & x_{1,r} - x_{1,c} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_{m,r} - x_{m,c} \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Note that the bottom row of $\mathbf{V}$ consists entirely of zeros. Moreover, define $\mathbf{s} = (s_0, s_1, \ldots, s_n)'$ and $\mathbf{t} = (1, t_1, \ldots, t_m)'$. Using this notation, (3.4) can be rewritten as:

$$\begin{cases} \mathbf{Us} - \mathbf{Vt} = \mathbf{0}, \\ \mathbf{s} \in \{-1, 1\}^{n+1}, \\ \mathbf{t} \in \{1\} \times \{-1, 1\}^{m}, \end{cases} \tag{3.6}$$

where $\mathbf{0}$ denotes the $(m+2)$-vector of zeros.

The least sophisticated way of finding a solution to (3.6) would be to simply try all possible vectors $\mathbf{s}$ and $\mathbf{t}$. Since m and n are small in this situation, the number of possibilities is not very large and this approach is actually quite feasible. However, it is also possible to reformulate the problem as a so-called binary linear programming problem. This has the advantage that standard software may be used to implement the method. Moreover, it will be seen presently that this formulation can be adapted easily to accommodate possible rounding errors present in the data.

The following binary variables are introduced to reformulate the problem:

$$\begin{aligned} \sigma_j &= \tfrac{1-s_j}{2}, \qquad && j \in \{0, 1, \ldots, n\}, \\ \tau_k &= \tfrac{1-t_k}{2}, \qquad && k \in \{1, \ldots, m\}. \end{aligned}$$

Finding an optimal solution to (3.6) may be restated as follows:

$$\begin{array}{l} \text{minimise } \sum_{j=0}^{n} \sigma_j + \sum_{k=1}^{m} \tau_k \\ \text{such that:} \\ \mathbf{U}\,(\mathbf{1} - 2\boldsymbol{\sigma}) - \mathbf{V}\,(\mathbf{1} - 2\boldsymbol{\tau}) = \mathbf{0} \\ \boldsymbol{\sigma} \in \{0,1\}^{n+1}\,,\ \boldsymbol{\tau} \in \{0\} \times \{0,1\}^{m}\,, \end{array} \tag{3.7}$$

where $\mathbf{1}$ is a vector of ones, $\boldsymbol{\sigma} = (\sigma_0, \sigma_1, \ldots, \sigma_n)'$ and $\boldsymbol{\tau} = (0, \tau_1, \ldots, \tau_m)'$.

Observe that in this formulation the number of variables s_j and t_k that are equal to -1 is minimised, i.e., the solution is searched for that results in the smallest number of changes being made in the record. Obviously, if a unique solution to (3.6) exists, then this is also the solution to (3.7). The binary linear programming problem may be solved by applying a standard branch and bound algorithm. Since n and m are small, very little computation time is needed to find the solution.

3.3.4 Allowing for rounding errors

It often happens that balance edit rules are violated by a very small difference. For instance, a reported total value is just one or two units smaller or larger than the sum of the reported item values. These inconsistencies are called *rounding errors* if the absolute difference is no larger than δ units. In the examples in this chapter, δ is chosen equal to 2. In the profit-and-loss account, rounding errors can occur in two ways. Firstly the pretax results may differ slightly from the sum of the balance amounts (a rounding error in the external sum), and secondly a balance amount may just disagree with the difference between the reported revenue and cost items (a rounding error in an internal sum).

Rounding errors often occur in conjunction with other errors. In particular, a record might contain a sign error that is obscured by a rounding error. Column (d) in Table 3.1 shows an example of such a record. If the method described in the previous subsection is applied directly, the sign error will not be detected.

Fortunately, the binary linear programming problem (3.7) can be adapted to take the possibility of rounding errors into account. This leads to the following problem:

$$\begin{array}{l} \text{minimise } \sum_{j=0}^{n} \sigma_j + \sum_{k=1}^{m} \tau_k \\ \text{such that:} \\ -\boldsymbol{\delta} \leq \mathbf{U}\,(\mathbf{1} - 2\boldsymbol{\sigma}) - \mathbf{V}\,(\mathbf{1} - 2\boldsymbol{\tau}) \leq \boldsymbol{\delta} \\ \boldsymbol{\sigma} \in \{0,1\}^{n+1}\,,\ \boldsymbol{\tau} \in \{0\} \times \{0,1\}^{m}\,, \end{array} \tag{3.8}$$

where $\boldsymbol{\delta}$ is a vector of δ's and the rest of the notation is obtained as before. Problem (3.7) is obtained by taking $\delta = 0$, i.e., by assuming that no rounding errors occur.

Example. If (3.8) is set up for Example (d) from Table 3.1, with $\delta = 2$, the following solution is found: $(\sigma_0 = 0, \sigma_1 = 1, \sigma_2 = 1, \sigma_3 = 0, \sigma_4 = 0; \tau_1 = 0, \tau_2 =$

$0, \tau_3 = 0$). Recalling that $\sigma_j = 1$ if and only if $s_j = -1$ (and a similar expression for τ_k and t_k), the sign error may be removed by changing the signs of both x_1 and x_2. As can be seen in Table 3.2, this correction indeed eliminates the sign error. It does not lead to a fully consistent profit-and-loss account, however, because there are rounding errors left in the data. To remove these, a separate method is needed. This problem will be discussed in Section 3.4. □

3.3.5 Summary

The following plan summarises the correction method for sign errors and interchanged revenues and costs. The input consists of a record that does not satisfy (3.3) and a choice for δ.

1. Determine the matrices $\mathbf{U}$ and $\mathbf{V}$ and set up the binary linear programming problem (3.8).

2. Solve (3.8). If no solution is possible, then the record does not contain a sign error. If a solution is found: continue.

3. Replace x_j by $-x_j$ for every $\sigma_j = 1$ and interchange $x_{k,r}$ and $x_{k,c}$ for every $\tau_k = 1$.

If Step 3 is performed, the resulting record satisfies (3.3) barring possible rounding errors.

3.4 Rounding errors

3.4.1 Introduction

It was mentioned in the previous section that very small inconsistencies with respect to balance edit rules often occur, e.g., a total value is just one unit smaller or larger than the sum of the component items. Such inconsistencies are called rounding errors, because they may be caused by values being rounded off to multiples of $1,000$. It is not straightforward to obtain a so-called *consistent rounding*, i.e., to make sure that the rounded off values have the same relation as the original values. For example, if the terms of the sum $2.7 + 7.6 = 10.3$ are rounded off to natural numbers the ordinary way, then the additivity is destroyed: $3 + 8 \neq 10$. Several algorithms for consistent rounding are available in the literature; see e.g., Salazar-González et al. (2004). Obviously, very few respondents are even aware of these methods, let alone inclined to use them while filling in a questionnaire.

By their nature, rounding errors have virtually no influence on aggregates, and in this sense the choice of method to correct them is unimportant. However, as mentioned in Section 3.1, the complexity of the automatic error localisation problem in SLICE increases rapidly as the number of violated edit rules becomes larger, irrespective of the magnitude of the violations. Thus, a record containing many rounding errors and few "real" errors might not be suitable for automatic editing by means of a Fellegi-Holt-based approach and might have to be edited manually. This is clearly a waste of resources. It is therefore advantageous to resolve all rounding errors in the early stages of the editing process, for instance immediately after the correction of obvious inconsistencies. Given the uninfluential nature of rounding errors, it might seem like a good approach to not correct them at all during automatic editing. Using SLICE, the only way to achieve this is by replacing each balance edit rule by two inequality edit rules that bound the difference of the total amount and its items between, say, -2 and 2. Unfortunately, this would make the automatic editing much more computationally demanding, because the error localisation algorithm of SLICE can handle equalities more efficiently than inequalities (De Waal and Quere, 2003). On balance, it is actually more efficient to handle rounding errors in a separate step.

In the remainder of this section, a heuristic method is described to resolve rounding errors in business survey data. This method is called a heuristic method because it does not return a solution that is "optimal" in some sense, e.g., that the number of changed variables or the total change in values is minimised. The rationale for using such a method is that the adaptations needed to resolve rounding errors are very small, and that it is therefore not necessary to use a sophisticated and potentially time-consuming search algorithm.

Although the idea behind the method is quite simple, some results from matrix algebra are needed to explain why it works. The necessary background will be briefly summarised in Section 3.4.2.

3.4.2 Matrix theory

Recall that Cramer's Rule is a theorem named after the Swiss mathematician Gabriel Cramer (1704–1752) which states the following. Let $\mathbf{A} = (a_{ij})$ be an invertible $p \times p$-matrix. The unique solution $\mathbf{x} = (x_1, \ldots, x_p)'$ to the system $\mathbf{Ax} = \mathbf{b}$ is given by:

$$x_k = \frac{\det \mathbf{B}_k}{\det \mathbf{A}}, \qquad k = 1, \ldots, p,$$

where $\mathbf{B}_k$ denotes the matrix found by replacing the k^{th} column of $\mathbf{A}$ by $\mathbf{b}$.

An alternative way of formulating this is that for any invertible matrix $\mathbf{A}$,

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \mathbf{A}^{\dagger}, \tag{3.9}$$

where $\mathbf{A}^{\dagger}$ denotes the *adjoint matrix* of $\mathbf{A}$. The adjoint matrix is found by transposing the matrix of cofactors: $\left(\mathbf{A}^{\dagger}\right)_{ji} = (-1)^{i+j} \det \mathbf{C}_{ij}$, where $\mathbf{C}_{ij}$ is the matrix $\mathbf{A}$ with the i^{th} row and the j^{th} column removed. For a proof of (3.9), see e.g., Harville (1997, Section 13.5).

A square matrix is called *unimodular* if its determinant is equal to 1 or -1. The following property is an immediate consequence of Cramer's Rule.

Property 3.1 *If* $\mathbf{A}$ *is an integer-valued unimodular matrix and* $\mathbf{b}$ *is an integer-valued vector, then the solution to the system* $\mathbf{Ax} = \mathbf{b}$ *is also integer-valued.*

A (not necessarily square) matrix for which the determinant of every square submatrix is equal to 0, 1 or -1 is called *totally unimodular*. That is to say, every square submatrix of a totally unimodular matrix is either singular or unimodular. Clearly, in order to be totally unimodular, a matrix must have all elements equal to 0, 1 or -1. A stronger version of Property 3.1 can be proved for the submatrices of a totally unimodular matrix.

Property 3.2 *Let* $\mathbf{B}$ *be a square submatrix of a totally unimodular matrix. If* $\mathbf{B}$ *is invertible, all elements of* $\mathbf{B}^{-1}$ *are in* $\{-1, 0, 1\}$.

Proof. This is easily seen using the adjoint matrix $\mathbf{B}^{\dagger}$. Since $|\det \mathbf{B}| = 1$ and all cofactors are equal to 0, 1 or -1, the property follows immediately from Equation (3.9). □

Verifying whether a matrix is totally unimodular by directly applying the definition is usually impossible – unless the matrix happens to be very small – because the number of determinants to evaluate is simply too high. Scholtus (2008) lists some results on total unimodularity that may be used in practice to determine whether a given matrix is totally unimodular without computing determinants.

3.4.3 The scapegoat algorithm

Basic idea

When the survey variables are denoted by the vector $\mathbf{x} = (x_1, \ldots, x_p)'$, the balance edit rules can be written as a linear system

$$\mathbf{Rx} = \mathbf{a}, \tag{3.10}$$

where each row of the $r \times p$-matrix $\mathbf{R}$ defines an edit rule and each column corresponds to a survey variable. The vector $\mathbf{a} = (a_1, \ldots, a_r)'$ contains any constant terms that occur in the edit rules. Denoting the i^{th} row of $\mathbf{R}$ by $\mathbf{r}_i'$, an edit rule is violated when $|\mathbf{r}_i'\mathbf{x} - a_i| > 0$. The inconsistency is called a rounding error when $0 < |\mathbf{r}_i'\mathbf{x} - a_i| \leq \delta$, where $\delta > 0$ is small. Similarly, the edit rules that take the form of a linear inequality can be written as

$$\mathbf{Q}\mathbf{x} \geq \mathbf{b}, \tag{3.11}$$

where each edit rule is defined by a row of the $q \times p$-matrix $\mathbf{Q}$ together with a constant from $\mathbf{b} = (b_1, \ldots, b_q)'$. Initially, it is assumed that only balance edit rules are given.

The idea behind the heuristic method is as follows. For each record containing rounding errors, a set of variables is selected beforehand. Next, the rounding errors are resolved by only adjusting the values of these selected variables. Hence, the name *scapegoat algorithm* seems appropriate. The name "scapegoat algorithm" was coined by Léander Kuijvenhoven (Statistics Netherlands).

In fact, the algorithm performs the selection in such a way that exactly one choice of values exists for the selected variables such that all rounding errors are resolved. Different variables are selected for each record to minimise the effect of the adaptations on aggregates.

It is assumed that the $r \times p$-matrix $\mathbf{R}$ satisfies $r \leq p$ and $\text{rank}(\mathbf{R}) = r$, that is: the number of variables should be at least as large as the number of restrictions and no redundant restrictions may be present. Clearly, these are very mild assumptions. Additionally, the scapegoat algorithm becomes simpler if $\mathbf{R}$ is a totally unimodular matrix. At Statistics Netherlands, it was found that matrices of balance edit rules used for structural business statistics are always of this type. A similar observation is made by De Waal (2002, §3.4.1).

An inconsistent record $\mathbf{x}$ is given, possibly containing both rounding errors and other errors. In the first step of the scapegoat algorithm, all rows of $\mathbf{R}$ for which $|\mathbf{r}_i'\mathbf{x} - a_i| > \delta$ are removed from the matrix and the associated constants are removed from $\mathbf{a}$. The resulting $r_0 \times p$-matrix is denoted by $\mathbf{R}_0$ and the resulting r_0-vector of constants by $\mathbf{a}_0$. It may happen that the record satisfies the remaining balance edit rules $\mathbf{R}_0\mathbf{x} = \mathbf{a}_0$. In that case, the algorithm stops here.

It is easy to see that if $\mathbf{R}$ satisfies the assumptions above, then so does $\mathbf{R}_0$. Hence $\text{rank}(\mathbf{R}_0) = r_0$ and $\mathbf{R}_0$ has r_0 linearly independent columns. The r_0 leftmost linearly independent columns may be found by putting the matrix in row echelon form through Gaussian elimination, as described by Fraleigh and Beauregard (1995, §2.2), or alternatively by performing a QR-decomposition with column

pivoting, as discussed by Golub and Van Loan (1996, §5.4). (How these methods work is irrelevant for the present purpose.) Since the choice of scapegoat variables and hence of columns should vary between records, a random permutation of columns is performed beforehand, yielding $\tilde{\mathbf{R}}_0$. The variables of $\mathbf{x}$ are permuted accordingly to yield $\tilde{\mathbf{x}}$.

Next, $\tilde{\mathbf{R}}_0$ is partitioned into two submatrices $\mathbf{R}_1$ and $\mathbf{R}_2$. The first of these is an $r_0 \times r_0$-matrix that contains the leftmost linearly independent columns of $\tilde{\mathbf{R}}_0$, the second is an $r_0 \times (p - r_0)$-matrix containing all other columns. The vector $\tilde{\mathbf{x}}$ is also partitioned into subvectors $\mathbf{x}_1$ and $\mathbf{x}_2$, containing the variables associated with the columns of $\mathbf{R}_1$ and $\mathbf{R}_2$, respectively. Thus

$$\tilde{\mathbf{R}}_0\tilde{\mathbf{x}} = \mathbf{a}_0 \text{ becomes } (\mathbf{R}_1\ \mathbf{R}_2)\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{pmatrix} = \mathbf{a}_0.$$

At this point, the variables from $\mathbf{x}_1$ are selected as scapegoat variables and the variables from $\mathbf{x}_2$ remain fixed. Therefore the values of $\mathbf{x}_2$ are filled in from the original record to obtain the following system:

$$\mathbf{R}_1\mathbf{x}_1 = \mathbf{a}_0 - \mathbf{R}_2\mathbf{x}_2 \equiv \mathbf{c}, \tag{3.12}$$

where $\mathbf{c}$ is a vector of known constants.

By construction, the square matrix $\mathbf{R}_1$ is of full rank and therefore invertible. Thus (3.12) has the unique solution $\hat{\mathbf{x}}_1 = \mathbf{R}_1^{-1}\mathbf{c}$. In general, this solution might contain fractional values, whereas most business survey variables are restricted to be integer-valued. If this is the case, a controlled rounding algorithm similar to the one described in Salazar-González et al. (2004) can be applied to the values of $(\hat{\mathbf{x}}_1', \mathbf{x}_2')'$ to obtain an integer-valued solution to $\mathbf{R}_0\mathbf{x} = \mathbf{a}_0$. Note however that this is not possible without slightly changing the value of at least one variable from $\mathbf{x}_2$ too.

If $\mathbf{R}$ happens to be a totally unimodular matrix, this problem does not occur. In that case $\det \mathbf{R}_1$ is equal to -1 or 1, and Property 3.1 says that $\hat{\mathbf{x}}_1$ is always integer-valued. In the remainder of this chapter, it is assumed that $\mathbf{R}$ is indeed totally unimodular.

An example

To illustrate the scapegoat algorithm, a small-scale example now follows. Suppose a data set contains records of eleven variables $x_1, \ldots, x_{11}$ that should conform to

the following five balance edit rules:

$$\left.\begin{array}{rcl} x_1 + x_2 & = & x_3 \\ x_2 & = & x_4 \\ x_5 + x_6 + x_7 & = & x_8 \\ x_3 + x_8 & = & x_9 \\ x_9 - x_{10} & = & x_{11} \end{array}\right\} \tag{3.13}$$

These edit rules may be written as $\mathbf{Rx} = \mathbf{0}$, with $\mathbf{x} = (x_1, \ldots, x_{11})'$ and

$$\mathbf{R} = \begin{pmatrix} 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 \end{pmatrix}. \tag{3.14}$$

Thus $\mathbf{a} = \mathbf{0}$ here. It is easily established that $\text{rank}(\mathbf{R}) = 5$. Moreover, it can be seen that $\mathbf{R}$ is totally unimodular by repeatedly applying the following property: a matrix containing only elements from $\{-1, 0, 1\}$ is totally unimodular, if and only if the submatrix found by removing all columns or rows with less than two non-zero elements is totally unimodular [see Scholtus (2008) for a proof].

The following record is inconsistent with respect to (3.13):

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
12	4	15	4	3	1	8	11	27	41	-13

This record violates all edit rules, except for $x_2 = x_4$. In each instance, the violation is small enough to qualify as a rounding error. Thus in this example $\mathbf{R}_0$ is identical to $\mathbf{R}$.

A random permutation is applied to the elements of $\mathbf{x}$ and the columns of $\mathbf{R}$. Suppose that the permutation is given by

$$\begin{array}{lllll} 1 \to 11, & 2 \to 8, & 3 \to 2, & 4 \to 5, & 5 \to 10, \quad 6 \to 9, \\ 7 \to 7, & 8 \to 1, & 9 \to 4, & 10 \to 3, & 11 \to 6. \end{array}$$

This yields the following result:

$$\tilde{\mathbf{R}} = \begin{pmatrix} 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

It so happens that the first five columns of $\tilde{\mathbf{R}}$ are linearly independent. Thus $\mathbf{R}_1$ consists of the first five columns of $\tilde{\mathbf{R}}$, and $\mathbf{R}_2$ consists of the remaining six

columns. The scapegoat variables are those that correspond to the columns of $\mathbf{R}_1$, that is to say x_8, x_3, x_{10}, x_9 and x_4. The original values from the record are filled in for the non-scapegoat variables to calculate the constant vector $\mathbf{c}$:

$$\mathbf{c} = -\mathbf{R}_2 \begin{pmatrix} x_{11} \\ x_7 \\ x_2 \\ x_6 \\ x_5 \\ x_1 \end{pmatrix} = - \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -13 \\ 8 \\ 4 \\ 1 \\ 3 \\ 12 \end{pmatrix} = \begin{pmatrix} -16 \\ -4 \\ -12 \\ 0 \\ -13 \end{pmatrix}.$$

Thus, the following system in $\mathbf{x}_1$ is obtained:

$$\mathbf{R}_1\mathbf{x}_1 = \begin{pmatrix} 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_8 \\ x_3 \\ x_{10} \\ x_9 \\ x_4 \end{pmatrix} = \begin{pmatrix} -16 \\ -4 \\ -12 \\ 0 \\ -13 \end{pmatrix} = \mathbf{c}.$$

Solving this system yields: $\hat{x}_3 = 16$, $\hat{x}_8 = 12$, $\hat{x}_9 = 28$, $\hat{x}_4 = 4$ and $\hat{x}_{10} = 41$. When the original values of the variables in $\mathbf{x}_1$ are replaced by these new values, the record becomes consistent with respect to (3.13):

x_1	x_2	$\hat{x}_3$	$\hat{x}_4$	x_5	x_6	x_7	$\hat{x}_8$	$\hat{x}_9$	$\hat{x}_{10}$	x_{11}
12	4	16	4	3	1	8	12	28	41	−13

Observe that in this example it was not necessary to change the value of every scapegoat variable. In particular, x_4 and x_{10} have retained their original values.

On the size of the adjustments

The solution vector $\hat{\mathbf{x}}_1$ is constructed by the scapegoat algorithm without any explicit use of the original vector $\mathbf{x}_1$. Therefore, it is not completely trivial that the adjusted values remain close to the original values, which is obviously desirable. In order to demonstrate this property, two upper bounds on the size of the adjustments are now derived, under the assumption that $\mathbf{R}$ is totally unimodular.

Recall that the *maximum norm* of a vector $\mathbf{v} = (v_1, \ldots, v_p)'$ is defined as

$$|\mathbf{v}|_\infty = \max_{j=1,\ldots,p} |v_j|.$$

The associated matrix norm is [cf. Stoer and Bulirsch (2002, §4.4)]:

$$||\mathbf{A}||_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{p} |a_{ij}|,$$

with $\mathbf{A} = (a_{ij})$ any $m \times p$-matrix. It is easily shown that

$$|\mathbf{A}\mathbf{v}|_\infty \leq ||\mathbf{A}||_\infty |\mathbf{v}|_\infty \tag{3.15}$$

for every $m \times p$-matrix $\mathbf{A}$ and every p-vector $\mathbf{v}$.

Turning to the scapegoat algorithm, it holds by construction that $\mathbf{R}_1\hat{\mathbf{x}}_1 = \mathbf{c}$. The original vector $\mathbf{x}_1$ satisfies $\mathbf{R}_1\mathbf{x}_1 = \mathbf{c}^*$, with $\mathbf{c}^* \neq \mathbf{c}$. Thus

$$\hat{\mathbf{x}}_1 - \mathbf{x}_1 = \mathbf{R}_1^{-1}\left(\mathbf{c} - \mathbf{c}^*\right). \tag{3.16}$$

It follows from (3.15) and (3.16) that

$$|\hat{\mathbf{x}}_1 - \mathbf{x}_1|_\infty \leq ||\mathbf{R}_1^{-1}||_\infty |\mathbf{c} - \mathbf{c}^*|_\infty \leq r_0 |\mathbf{c} - \mathbf{c}^*|_\infty, \tag{3.17}$$

where the last inequality is found by observing that Property 3.2 implies

$$||\mathbf{R}_1^{-1}||_\infty = \max_{i=1,\ldots,r_0} \sum_{j=1}^{r_0} |\left(\mathbf{R}_1^{-1}\right)_{ij}| \leq r_0.$$

Writing $\hat{\mathbf{x}} = (\hat{\mathbf{x}}_1', \mathbf{x}_2')'$ and observing that

$$\begin{aligned}
\mathbf{c} - \mathbf{c}^* &= \mathbf{R}_1\hat{\mathbf{x}}_1 - \mathbf{R}_1\mathbf{x}_1 \\
&= \mathbf{R}_1\hat{\mathbf{x}}_1 + \mathbf{R}_2\mathbf{x}_2 - \mathbf{a}_0 - (\mathbf{R}_1\mathbf{x}_1 + \mathbf{R}_2\mathbf{x}_2 - \mathbf{a}_0) \\
&= \tilde{\mathbf{R}}_0\hat{\mathbf{x}} - \mathbf{a}_0 - (\mathbf{R}_0\mathbf{x} - \mathbf{a}_0) \\
&= -(\mathbf{R}_0\mathbf{x} - \mathbf{a}_0)
\end{aligned}$$

it is seen that $|\mathbf{c} - \mathbf{c}^*|_\infty = |\mathbf{R}_0\mathbf{x} - \mathbf{a}_0|_\infty = \delta_{\max}$, where $\delta_{\max} \leq \delta$ is the magnitude of the largest rounding error that occurs for this particular record. Plugging this into (3.17) yields

$$|\hat{\mathbf{x}}_1 - \mathbf{x}_1|_\infty \leq r_0\delta_{\max}. \tag{3.18}$$

This upper bound on the maximum difference between elements of $\hat{\mathbf{x}}_1$ and $\mathbf{x}_1$ shows that the solution found by the scapegoat algorithm cannot be arbitrarily far from the original record. The fact that (3.18) is proportional to the order of $\mathbf{R}_1$ suggests that ever larger adjustments should be expected as the number of balance edit rules increases, which is somewhat worrying. However, in practice much smaller adjustments than $r_0\delta_{\max}$ are found. For instance, in the above example with eleven variables the maximal absolute difference according to (3.18) equals 5, but actually no value was changed by more than one unit. Nevertheless, it is possible to construct a pathological example for which the upper bound (3.18) becomes exact; see Appendix 3.A.

In practice, a more interesting view on the size of the adjustments may be provided by the quantity

$$\frac{1}{r_0}\sum_{i=1}^{r_0} |\,(\hat{\mathbf{x}}_1 - \mathbf{x}_1)_i\,|$$

which measures the *average* size of the adjustments, rather than the maximum. Starting from (3.16), it is seen that

$$|\,(\hat{\mathbf{x}}_1 - \mathbf{x}_1)_i\,| = \left|\sum_{j=1}^{r_0} \left(\mathbf{R}_1^{-1}\right)_{ij} (\mathbf{c} - \mathbf{c}^*)_j\right| \leq \sum_{j=1}^{r_0} |\left(\mathbf{R}_1^{-1}\right)_{ij}|\,|(\mathbf{c} - \mathbf{c}^*)_j\,|.$$

Using again that $|\mathbf{c} - \mathbf{c}^*|_\infty = \delta_{\max}$ yields

$$\frac{1}{r_0}\sum_{i=1}^{r_0} |\,(\hat{\mathbf{x}}_1 - \mathbf{x}_1)_i\,| \leq \frac{\delta_{\max}}{r_0}\sum_{i=1}^{r_0}\sum_{j=1}^{r_0} |\left(\mathbf{R}_1^{-1}\right)_{ij}| \equiv \gamma(\mathbf{R}_1)\delta_{\max}, \tag{3.19}$$

where $\gamma(\mathbf{R}_1) = \frac{1}{r_0}\sum_{i=1}^{r_0}\sum_{j=1}^{r_0} |\left(\mathbf{R}_1^{-1}\right)_{ij}|$.

This upper bound on the average adjustment size can be evaluated before the scapegoat algorithm is applied to an actual data set. Namely, suppose that a set of balance edit rules (3.10) is given. Restricting oneself to the case $r_0 = r$, $\gamma(\mathbf{R}_1)$ can be computed for various invertible $r \times r$-submatrices of $\mathbf{R}$ to assess the magnitude of the upper bound in (3.19). It can be shown [see Scholtus (2008)] that there exist exactly $\det(\mathbf{R}\mathbf{R}')$ of these submatrices. In practice, this number is very large and it is infeasible to compute $\gamma(\mathbf{R}_1)$ for all matrices $\mathbf{R}_1$. In that case, a random sample of reasonable size can be taken, by repeatedly performing the part of the scapegoat algorithm that constructs $\mathbf{R}_1$.

Example. For the 5×11-matrix from the above example, $\det(\mathbf{R}\mathbf{R}') = 121$, so $\mathbf{R}$ has 121 invertible 5×5-submatrices. Since this number is not too large, it is possible to evaluate $\gamma(\mathbf{R}_1)$ for all these matrices. The mean value of $\gamma(\mathbf{R}_1)$ turns out to be 1.68, with a standard deviation of 0.39. Since $\delta_{\max} = 1$ in this example, according to (3.19) the average adjustment size is bounded on average by 1.68. □

Section 3.4.4 examines the adjustments in a real-world example. These turn out to be quite small.

Critical variables

In addition to balance edit rules, business survey variables usually have to satisfy a large number of edit rules that take the form of linear inequalities. For instance, it is very common that most variables are restricted to be nonnegative. The scapegoat algorithm as described above does not take this into account. A nonnegative

variable might therefore be changed by the algorithm from 0 to -1, resulting in a new violation of an edit rule. The present section extends the algorithm to prevent this.

Suppose that in addition to the balance edit rules (3.10), the data also have to satisfy the inequalities (3.11). For a given record, a variable will be called *critical* if it occurs in an inequality that (almost) holds with exact equality when the current values of the survey variables are filled in:

$$x_j \text{ is a critical variable iff } 0 \leq \mathbf{q}_i'\mathbf{x} - b_i \leq \epsilon_i \text{ for some } i \text{ with } q_{ij} \neq 0, \qquad (3.20)$$

where $\mathbf{q}_i'$ denotes the i^{th} row of $\mathbf{Q}$ and ϵ_i marks the margin chosen for the i^{th} restriction. As a particular case, x_j is called critical if it must be non-negative and currently has a value between 0 and $\epsilon_{i(j)}$, with $i(j)$ the index of the row in $\mathbf{Q}$ corresponding to the nonnegativity constraint for x_j. To prevent the violation of edit rules in (3.11), no critical variable should be selected for change during the execution of the scapegoat algorithm.

A way to achieve this works as follows. Rather than randomly permuting all variables (and all columns of $\mathbf{R}_0$), two separate permutations should be performed for the noncritical and the critical variables. The permuted columns associated with the noncritical variables are then placed to the left of the columns associated with the critical variables. This ensures that linearly independent columns are found among those that are associated with noncritical variables, provided the record contains a sufficient number of noncritical variables. In practice, this is typically the case, because the number of survey variables is much larger than the number of balance edit rules.

If a record contains many critical variables, some of these might still be selected as scapegoat variables. This is not necessarily a problem, because usually not all scapegoat variables are changed by the algorithm. This is, in fact, the reason why the critical variables are also randomly permuted: it is unimportant whether a solution to (3.12) contains critical variables, provided that no inequality edit rules are violated as a result. It is therefore sufficient to build in a check at the end of the algorithm that rejects the solution if a new violation of an edit rule from (3.11) is detected. If this does happen, it seems advantageous to let the record be processed again, because a different permutation of columns may yield a feasible solution. To prevent the algorithm from getting stuck, the number of attempts should be maximised by a preset constant K. If no feasible solution has been found after K attempts, the record remains untreated.

Good values of ϵ_i and K have to be determined in practice. However, not too much effort should be put into this, because these parameters only affect a limited

number of records. In the real-world example to be discussed in Section 3.4.4, only a handful of infeasible solutions were found before the improvements of the current section were included in the algorithm.

Exceptional variables

In practice, the data may contain some variables that should not be changed by the scapegoat algorithm at all. An example of such a variable in the structural business statistics is *number of employees*. This variable occurs in a balance edit rule that is often inconsistent because of a very small violation, but this violation cannot be the result of inconsistent rounding; this variable is asked as a number, not as a multiple of $1,000$ Euros. Moreover, the impact of changing the *number of employees* to suit the balance edit rule can be considerable, particularly for very small companies. Therefore, at this stage it seems preferable to leave the inconsistency as it is, to be resolved later by either a subject-matter specialist or SLICE.

This can be achieved by removing the balance edit rules concerning these exceptional variables from $\mathbf{R}$. The variables should not be removed from $\mathbf{x}$, however, as they may also occur in edit rules in (3.11). (For instance, the *number of employees* times a constant is used to maximise the *total labour costs*.) The values of the exceptional variables therefore play a role in determining the critical variables. Note that it is not necessary to remove the exceptional variables from $\mathbf{x}$ anyway, because the columns that correspond with these variables contain only zeros in the new version of $\mathbf{R}$.

Summary

The following plan summarises the scapegoat algorithm. The input consists of an inconsistent record $\mathbf{x}$ with p variables, a set of r balance edit rules $\mathbf{Rx} = \mathbf{a}$, a set of q inequalities $\mathbf{Qx} \geq \mathbf{b}$ and parameters δ, ϵ_i $(i = 1, \ldots, q)$ and K. Edit rules concerning exceptional variables (as defined above) have been removed from $\mathbf{Rx} = \mathbf{a}$ beforehand.

1. Remove all edit rules for which $|\mathbf{r}_i'\mathbf{x} - a_i| > \delta$. The remaining system is denoted as $\mathbf{R}_0\mathbf{x} = \mathbf{a}_0$. The number of rows in $\mathbf{R}_0$ is called r_0. If $\mathbf{R}_0\mathbf{x} = \mathbf{a}_0$ holds: stop. Otherwise: determine the critical variables according to (3.20).

2. (a) Perform random permutations of the critical and noncritical variables separately. Then permute the corresponding columns of $\mathbf{R}_0$ the same way. Put the noncritical variables and their columns before the critical variables and their columns.

(b) Determine the r_0 leftmost linearly independent columns in the permuted matrix $\tilde{\mathbf{R}}_0$. Together, these columns are a unimodular matrix $\mathbf{R}_1$ and the associated variables form a vector $\mathbf{x}_1$ of scapegoat variables. The remaining columns are a matrix $\mathbf{R}_2$ and the associated variables form a vector $\mathbf{x}_2$.

(c) Fix the values of $\mathbf{x}_2$ from the record and compute $\mathbf{c} = \mathbf{a}_0 - \mathbf{R}_2\mathbf{x}_2$.

3. Solve the system $\mathbf{R}_1\mathbf{x}_1 = \mathbf{c}$.

4. Replace the values of $\mathbf{x}_1$ by the solution just found. If the resulting record does not violate any other edit rule from $\mathbf{Qx} \geq \mathbf{b}$, the algorithm outputs the adjusted record and terminates. Otherwise, return to step 2a, unless this has been the K^{th} attempt. In that case, the record is not adjusted.

In this description, it is assumed that $\mathbf{R}$ is totally unimodular.

3.4.4 A real-world application

In Section 3.5, results will be discussed of an application of the two algorithms from this study to a large data set from the Netherlands' structural business statistics. These results focus on the impact of the algorithms on the efficiency of the editing process. In the current subsection some earlier test results are presented that focus more on technical aspects of the scapegoat algorithm.

The scapegoat algorithm has been tested using data from the wholesale structural business statistics of 2001. There are $4,725$ records containing 97 variables each. These variables should conform to a set of 28 balance edit rules and 120 inequalities, of which 92 represent nonnegativity constraints. After exclusion of edit rules that affect exceptional variables, 26 balance edit rules remain. The resulting 26×97-matrix $\mathbf{R}$ is totally unimodular, as can be determined very quickly using the method of removing columns and rows mentioned above. Note that it would be practically impossible to determine whether a matrix of this size is totally unimodular just by computing all the relevant determinants.

An implementation of the algorithm in *S-Plus* was used to treat the data. The parameters used were: $\delta = 2$, $\epsilon_i = 2$ $(i = 1, \ldots, 120)$ and $K = 10$. The total computation time on an ordinary desktop PC was less than three minutes.

Table 3.3 summarises the results of applying the scapegoat algorithm. No new violations of inequalities were found. In fact, the adjusted data happen to satisfy four additional inequalities.

According to (3.18) the size of the adjustments made by the algorithm is theoretically bounded by $26 \times 2 = 52$, which is rather high. A random sample of

Table 3.3: Results of applying the scapegoat algorithm to the wholesale data.

Number of records	4,725
Number of variables per record	97
Number of adjusted records	3,176
Number of adjusted variables	13,531
Number of violated edit rules (before)	34,379
Balance edit rules	26,791
Inequalities	7,588
Number of violated edit rules (after)	23,054
Balance edit rules	15,470
Inequalities	7,584

$10,000$ invertible 26×26-submatrices of $\mathbf{R}$ was drawn to evaluate (3.19). The sample mean of $\gamma(\mathbf{R}_1)$ is 1.89, with a standard deviation of 0.27. Thus, the average adjustment size is bounded on average by $1.89 \times 2 \approx 3.8$. Note that this value of $\gamma(\mathbf{R}_1)$ is only marginally higher than the one obtained for the much smaller restriction matrix from the example with eleven variables given above.

Table 3.4 displays the adjustment sizes that were actually found for the wholesale data. These turn out to be very reasonable.

Table 3.4: Distribution of the adjustments (in absolute value).

Magnitude	Frequency
1	11,953
2	1,426
3	134
4	12
5	4
6	2

3.5 Application to the Netherlands' Structural Business Statistics of 2007

The algorithms from Sections 3.3 and 3.4 have been applied in an experiment using data from the Netherlands' structural business statistics of 2007. The data were collected by Statistics Netherlands from businesses in various sectors, including wholesale, construction and audiovisual services. The algorithms were run using the original, unedited data.

Table 3.5: Structure of the profit-and-loss account in the structural business statistics survey of 2007, with an example of interchanged revenues and costs.

Variable	Full name	Original data	Corrected data
$x_{0,r}$	Operating returns	$49,110$	$49,110$
$x_{0,c}$	Operating costs	$46,550$	$46,550$
x_0	Operating results	$2,560$	$2,560$
$x_{1,r}$	Provisions rescinded	340	$\mathbf{0}$
$x_{1,c}$	Provisions added	0	$\mathbf{340}$
x_1	Balance of provisions	-340	-340
x_2	Book profit/loss	-90	-90
x_3	Financial result	30	30
x_4	Exceptional result	0	0
x_5	Pretax results	$2,160$	$2,160$

Table 3.5 displays the structure of the profit-and-loss account in the structural business survey that was used in 2007. This differs from the examples in Section 3.3 because the questionnaire was redesigned after 2005. The associated edit rules are given by (3.3), with $n = 5$ and $m = 1$.

The results of applying the algorithm that corrects sign errors and interchanged revenues and costs are displayed in Table 3.6. As can be seen, the fraction of profit-and-loss accounts requiring editing is low: almost 90 per cent of the accounts are reported without error. This can be explained by the fact that in 2007 the majority of businesses reported by electronic questionnaire. Some of the edit rules were built into this questionnaire, so that respondents received a warning message if the reported amounts were inconsistent. In this way, many errors that would occur on a paper questionnaire could be avoided during electronic data collection (Giesen, 2007).

On the other hand, among the profit-and-loss accounts that do require editing, the fraction of accounts containing sign errors is substantial: about one in five. This means that, as far as the profit-and-loss account is concerned, using the algorithm of Section 3.3 substantially reduces the amount of work remaining for either manual editing or automatic editing by SLICE. It should also be mentioned that the majority of errors corrected by the algorithm were in fact interchanged revenues

Table 3.6: Results of applying the correction algorithm for sign errors to the 2007 data.

Total number of profit-and-loss accounts	17,258	
Without inconsistencies	15,465	(89.6%)
With inconsistencies	1,793	(10.4%)
Corrected by the algorithm	392	(21.9%)

and costs. As noted in Section 3.3, this type of error is difficult to handle using an automatic editing method based on the Fellegi-Holt paradigm.

The scapegoat algorithm was applied to the data using the same parameter settings as in the real-world application of Section 3.4.4. The results are displayed in Table 3.7. Please note that the total number of records and the number of records with/without inconsistencies are not comparable to the corresponding numbers in Table 3.6, because both records with an empty profit-and-loss account and inconsistencies outside the profit-and-loss account are not counted in Table 3.6.

Table 3.7: Results of applying the scapegoat algorithm to the 2007 data.

Total number of records	17,297	
Without inconsistencies	11,183	(64.6%)
With inconsistencies	6,114	(35.4%)
Corrected by the algorithm	1,295	(21.2%)
Number of violated balance edit rules (before)	11,584	
Number of violated balance edit rules (after)	10,113	
Number of resolved violations	1,471	(12.7%)

Again, the large percentage of records without inconsistencies is due to the use of electronic data collection. This can be seen from the marked difference between the relative number of inconsistencies in the survey data of 2001 (shown in Table 3.3) and 2007 (shown in Table 3.7). In 2001, all data collection was still done through paper questionnaires.

Of the records that do contain inconsistencies, about one in five contains at least one rounding error. Moreover, the scapegoat algorithm succeeds in resolving 1,471 of the 11,584 violations of balance edits in the original data set, i.e., about one in eight. This entails a substantial reduction of the amount of editing that remains to be done either manually or automatically by SLICE.

The figures in Table 3.7 were obtained by applying the scapegoat algorithm directly to the unedited data. In practice, it would be better to first correct sign errors and then rounding errors, because an application of the algorithm from Section 3.3 may reveal "hidden" rounding errors; cf. Example (d) in Tables 3.1 and 3.2.

3.6 Conclusion

The main purpose of this study has been to discuss the use of deductive methods for correcting obvious inconsistencies in business survey data, i.e., inconsistencies where the underlying error mechanism can be recognised easily. In particular, algorithms have been described that detect and correct two types of inconsistencies that occur in data collected for the Netherlands' structural business statistics: sign errors and rounding errors. Other errors are discussed by Scholtus (2008, 2009).

Deductive algorithms are intended to be applied at the beginning of the data editing process, before manual editing and regular automatic editing with SLICE take place. In this way, the efficiency of the editing process is expected to increase, because more records will be eligible for automatic editing. In particular, the presence of obvious errors and rounding errors may cause a record to be submitted to manual editing, because the automatic error localisation problem is too difficult to solve, when in fact the record can be handled automatically by SLICE once the obvious errors have been removed. Moreover, in many cases the use of a deductive algorithm for obvious errors also increases the quality of the edited data, because systematic errors are often handled incorrectly by SLICE. This is for example true for sign errors.

Often, the presence of errors with a structural cause in survey data signifies that many respondents find it difficult to answer correctly because of a certain aspect of the questionnaire design. An alternative way to handle systematic errors is, therefore, to try to prevent them during data collection, e.g., by improving the wording of questions or the design of answer boxes, or by adding explanatory notes, or – in the case of electronic data collection – by means of warning messages. If this approach succeeds in removing the underlying cause of the systematic error, then the need for a deductive correction algorithm vanishes.

Nevertheless, in practice, deductive correction methods can still play an important part in the editing process. Firstly, it is not always possible to prevent systematic errors through an improved form of data collection. For instance, unity measure errors have been observed for many years in data collected by statistical offices but, so far, no conclusive method has been found to stop respondents from making this type of error. Secondly, changes in the questionnaire design are costly,

because the new questionnaire has to be extensively tested, and they can adversely influence the comparability of statistics over time. Hence, they should not be implemented too often. By contrast, implementing a deductive correction algorithm is cheap and straightforward. If a new systematic error is discovered in the data, it can therefore be advantageous to correct this error deductively at first, until a major revision of the data collection strategy is due.

The algorithms for correcting sign errors and rounding errors described in this chapter have been implemented as part of the R package `deducorrect`, which is available for download from the Comprehensive R Archive Network (http://cran.r-project.org). See Van der Loo et al. (2011) for more details.

Appendix 3.A A pathological example

The upper bound (3.18) on the size of the maximal adjustment made by the scapegoat algorithm can be achieved exactly for a particular set of balance edit rules in combination with a particular input record. Scholtus (2008) provided an example for the case $\delta = 2$ which is generalised here to work for any value of δ.

Let the balance edit rules be given by $\mathbf{Rx} = \mathbf{0}$, with $\mathbf{R}$ the following $r \times (r+1)$-matrix (for some $r \geq 3$):

$$\mathbf{R} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 1 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

Let the inconsistent record be

$$\mathbf{x} = \begin{pmatrix} -(r-3)C - (2r-3)\delta \\ C + 2\delta \\ \vdots \\ C + 2\delta \\ -C \\ -C - \delta \end{pmatrix},$$

where C may be any constant. Note that $\mathbf{Rx} = (-\delta, \delta, \ldots, \delta)'$, so all edit rules are violated and all violations qualify as rounding errors. For simplicity, the example is constructed with $r_0 = r$ and $\delta_{\max} = \delta$.

If $x_1, x_2, \ldots, x_r$ are chosen as scapegoat variables, the matrix $\mathbf{R}_1$ consists of the first r columns of $\mathbf{R}$. Then $\mathbf{c} = -\mathbf{R}_2 x_{r+1} = (0, C + \delta, \ldots, C + \delta, -C - \delta)'$

and it is easy to see that

$$\mathbf{R}_1^{-1}\mathbf{c} = \begin{pmatrix} 1 & -1 & -1 & \cdots & -1 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 0 \\ C+\delta \\ \vdots \\ C+\delta \\ -C-\delta \end{pmatrix}$$
$$= \begin{pmatrix} -(r-3)(C+\delta) \\ C+\delta \\ \vdots \\ C+\delta \\ -C-\delta \end{pmatrix}.$$

The adjusted record is therefore

$$\hat{\mathbf{x}} = \begin{pmatrix} -(r-3)(C+\delta) \\ C+\delta \\ \vdots \\ C+\delta \\ -C-\delta \\ -C-\delta \end{pmatrix}$$

which, as the reader may verify, indeed satisfies all edit rules. Note that

$$\hat{x}_1 - x_1 = -(r-3)(C+\delta) + (r-3)C + (2r-3)\delta = r\delta,$$

which is equal to the upper bound given by (3.18).

It is worth noting that even in this contrived example most adjustments are quite small: apart from x_1, no scapegoat variable is adjusted by more than δ units.

Chapter 4

Automatic Editing with Hard and Soft Edits

The contents of this chapter have been published in *Survey Methodology* as Scholtus (2013). That version omitted the last paragraph of Section 4.4.1 and Footnote 1. Otherwise, the chapter is identical to the article, apart from some minor textual corrections and adjustments.

4.1 Introduction

An important part of every statistical process is *data editing*, i.e., detecting and correcting errors as well as missing values in the collected data. National statistical institutes have traditionally relied on manual editing, where the data are checked and, if necessary, adjusted by subject-matter experts. Unfortunately, this approach can be very time-consuming and expensive. Alternative methods have therefore been developed to increase the efficiency of the editing process, such as *selective editing* and *automatic editing*. This chapter focuses on the latter approach. We refer to De Waal et al. (2011) and their references for a discussion of selective editing and other forms of statistical data editing.

The goal of automatic editing is to accurately detect and correct errors as well as missing values in a data file in a fully automated manner, i.e., without human intervention. Provided that automatic editing leads to data of sufficient quality, it can be used as a partial alternative to manual editing. In practice, automatic editing implies that the data are made consistent with respect to a set of constraints: the so-called *edits*. Examples of edits include:

$$\textit{Profit} = \textit{Total Turnover} - \textit{Total Costs}; \tag{4.1}$$

and

$$\textit{Profit} \leq 0.6 \times \textit{Total Turnover}. \tag{4.2}$$

Most automatic editing methods proceed by solving two separate problems: first the *error localisation problem*, i.e., determining which variables are erroneous or missing, and subsequently the *consistent imputation problem*, i.e., determining new values for these variables that satisfy all the edits. The present chapter focuses on the error localisation problem.

With respect to the two examples of edits given above, it is interesting to note the conceptual difference that exists between them. Edit (4.1) is an example of an edit that has to hold by definition, so that every combination of values that fails this edit necessarily contains an error. Edits of this type are known as *hard edits*, *fatal edits*, or *logical edits*. Edit (4.2), on the other hand, is an example of an edit that identifies combinations of values that are implausible but not necessarily incorrect. In this example, records for which *Profit* is larger than 60% of *Total Turnover* are considered suspicious. However, it is conceivable that such a combination of values is occasionally correct. Edits of this type, which do not identify errors with certainty, are known as *soft edits* or *query edits*.

An important limitation of existing algorithms for automatic editing is that they treat all edits as hard edits. That is to say, a failed edit is always attributed to an error in the data. In manual editing, however, subject-matter specialists also make extensive use of soft edits. During automatic editing, these soft edits are either not used at all, or else interpreted as hard edits. Both solutions are unsatisfactory: in the first case some errors may be missed during automatic editing, and in the second case some correct values may be wrongfully identified as erroneous. In fact, the inability of automatic editing methods to handle soft edits partly explains why in practice many differences are found between manually edited and automatically edited data.

The object of this chapter is to present a new formulation of the automatic error localisation problem which can distinguish between hard edits and soft edits. In addition, the chapter shows how the error localisation algorithm of SLICE – the software package for automatic editing developed at Statistics Netherlands – can be adapted to solve this new error localisation problem.

The remainder of this chapter is organised as follows. Section 4.2 provides a brief summary of existing methods for solving the error localisation problem. A distinction between hard and soft edits is introduced in the error localisation problem in Section 4.3. Section 4.4 extends the theory behind the algorithm of SLICE to the case that not all edits have to be satisfied. Based on these theoretical results, an algorithm that solves the error localisation problem for hard and soft edits is outlined in Section 4.5. In Section 4.6, the new algorithm is illustrated by means of a small example. Section 4.7 mentions the first experiences with a

practical implementation of the new algorithm. Finally, some concluding remarks follow in Section 4.8.

4.2 Background

4.2.1 Edits

The problem to be discussed in this chapter entails, in its most general form, the detection of erroneous and missing values in a record containing both categorical variables $(v_1, \ldots, v_m)$ and numerical variables $(x_1, \ldots, x_p)$. These variables are supposed to satisfy a set of restrictions (edits), each of which can be written in one of the following forms:

$$\begin{aligned} \psi^k: \quad & \text{IF} \quad (v_1, \ldots, v_m) \in F_1^k \times \cdots \times F_m^k \\ & \text{THEN} \quad (x_1, \ldots, x_p) \in \{a_{k1}x_1 + \cdots + a_{kp}x_p + b_k \geq 0\} \end{aligned} \tag{4.3}$$

or

$$\begin{aligned} \psi^k: \quad & \text{IF} \quad (v_1, \ldots, v_m) \in F_1^k \times \cdots \times F_m^k \\ & \text{THEN} \quad (x_1, \ldots, x_p) \in \{a_{k1}x_1 + \cdots + a_{kp}x_p + b_k = 0\}\,. \end{aligned} \tag{4.4}$$

In these expressions, F_j^k is a subset of D_j, the domain of observed values for the categorical variable v_j, and a_{kj} and b_k are known numerical constants. The index k is used to number the edits. Note that D_j is assumed to contain all values of v_j that may be encountered in practice; this includes erroneous values. To simplify matters, we restrict the problem to edits having linear numerical conditions. This class of edits turns out to be sufficiently powerful for most practical applications [cf. De Waal (2005)].

A record $(v_1^0, \ldots, v_m^0, x_1^0, \ldots, x_p^0)$ is said to *fail* an edit if the categorical IF-condition is true (i.e., $v_j^0 \in F_j^k$ for all $j = 1, \ldots, m$), but the numerical THEN-condition is false (i.e., either $a_{k1}x_1^0 + \cdots + a_{kp}x_p^0 + b_k < 0$ or $a_{k1}x_1^0 + \cdots + a_{kp}x_p^0 + b_k \neq 0$, depending on the form of the edit). Otherwise, we say that the edit is *satisfied* by that record. It should be noted that an edit is always satisfied by any record for which the categorical IF-condition is false, regardless of the status of the numerical THEN-condition. A record is called *consistent* if it satisfies every edit.

A categorical variable v_j is said to be *involved* in an edit if and only if $F_j^k \neq D_j$, since any edit with $F_j^k = D_j$ is failed or satisfied regardless of the value of v_j. Similarly, a numerical variable x_j is said to be involved in an edit if and only if $a_{kj} \neq 0$. We may assume that $F_j^k \neq \emptyset$, where $\emptyset$ denotes the empty set. Clearly, a

degenerate edit with $F_j^k = \emptyset$ can be discarded with no loss of information, since it is never failed. The same holds for any edit with a numerical THEN-condition that is always true.

Two important special cases of (4.3) and (4.4) are edits that involve only categorical or only numerical variables. A purely categorical edit has the following form:

$$\psi^k : \text{ IF } (v_1, \ldots, v_m) \in F_1^k \times \cdots \times F_m^k \text{ THEN } \emptyset. \tag{4.5}$$

Edit (4.5) is failed by each record for which the categorical condition is true. A purely numerical edit can be written as

$$\psi^k : \ (x_1, \ldots, x_p) \in \{a_{k1}x_1 + \cdots + a_{kp}x_p + b_k \geq 0\} \tag{4.6}$$

or

$$\psi^k : \ (x_1, \ldots, x_p) \in \{a_{k1}x_1 + \cdots + a_{kp}x_p + b_k = 0\}. \tag{4.7}$$

Edits (4.6) and (4.7) are failed by each record for which the numerical conditions are false.

Edits (4.1) and (4.2) above are examples of purely numerical edits. A simple example of a purely categorical edit is:

$$\text{IF } (\textit{Age}, \ \textit{Marital Status}) \in \{\text{“} < 16\text{”}\} \times \{\text{“Married”}\} \text{ THEN } \emptyset.$$

This edit states that persons aged less than 16 years cannot be married. Finally, an example of a mixed edit is:

$$\text{IF } \textit{Age} \in \{\text{“} < 12\text{”}\} \text{ THEN } \textit{Income} = 0.$$

According to this edit, persons aged less than 12 years do not have a positive income.

4.2.2 The error localisation problem

For a given record $(v_1^0, \ldots, v_m^0, x_1^0, \ldots, x_p^0)$ and a collection of edits, it is straightforward to verify which values in the record are missing and whether any of the edits are failed. However, given that some of the edits are failed, solving the error localisation problem is much less straightforward. On the one hand, most edits involve more than one variable, and on the other hand, most variables are involved in more than one edit.

In order to solve the error localisation problem automatically, one has to adopt a formal strategy for finding erroneous values. The most commonly-used strategy is based on the paradigm of Fellegi and Holt (1976): make the record consistent

by changing the smallest possible number of original values. Other strategies have also been proposed; for instance, Little and Smith (1987) suggested a criterion based on outlier detection (without edits) and Casado Valero et al. (1996) formulated error localisation as a quadratic minimisation problem. We shall restrict attention to the Fellegi-Holt paradigm here, because of its frequent use in official statistics.

The original Fellegi-Holt paradigm is easily generalised to allow a distinction between a priori suspicious and less suspicious variables. To this end, one associates a *confidence weight* to each variable. According to the generalised Fellegi-Holt paradigm, one should search for a subset of the variables which (i) can be imputed in such a way that the imputed record satisfies all edits, and (ii) minimises the following target function:

$$D_{\mathrm{FH}} = \sum_{j=1}^{m} w_j^C y_j^C + \sum_{j=1}^{p} w_j^N y_j^N. \tag{4.8}$$

Here, w_j^C and w_j^N denote the confidence weights of the categorical and numerical variables, respectively. The binary target variables y_j^C and y_j^N describe the structure of the solution: $y_j^C = 1$ if v_j is to be imputed and $y_j^C = 0$ otherwise, and similarly $y_j^N = 1$ if x_j is to be imputed and $y_j^N = 0$ otherwise. Since variables with missing values have to be imputed with certainty, we set $y_j^C = 1$ or $y_j^N = 1$ when v_j^0 or x_j^0 is missing.

Fellegi and Holt (1976) also presented a method for solving the error localisation problem under this paradigm. This method first derives a well-defined set of logically implied edits from the original set of edits, to obtain a so-called *complete set of edits*. Next, the error localisation problem may be formulated as a straightforward set-covering problem for any record (Fellegi and Holt, 1976; Boskovitz et al., 2005). Unfortunately, especially for numerical data the complete set of edits can be extremely large in practice, so the method of Fellegi and Holt is not always computationally feasible.

Many alternative algorithms have been developed for error localisation according to the Fellegi-Holt paradigm. Besides improvements on Fellegi and Holt's original method (Garfinkel et al., 1986; Winkler, 1995), the list includes formulations based on vertex generation (Sande, 1978; Kovar and Whitridge, 1990; Todaro, 1999; De Waal, 2003c), cutting planes (Garfinkel et al., 1986, 1988; Ragsdale and McKeown, 1996), and mixed integer (Schaffer, 1987; Riera-Ledesma and Salazar-González, 2003) and integer programming (Bruni, 2004, 2005); see also De Waal et al. (2011) for an overview. Here, we shall focus on a branch-and-bound algorithm due to De Waal and Quere (2003) which, in contrast to some of the above

approaches, can handle a mix of categorical and numerical data. This algorithm has been implemented in the software package SLICE at Statistics Netherlands and has been found to be computationally feasible in practice.

4.2.3 The branch-and-bound algorithm of SLICE

A detailed description of the error localisation algorithm implemented in SLICE can be found in De Waal and Quere (2003), De Waal (2003b), and De Waal et al. (2011). Here, we only mention those aspects of the algorithm that we shall need later. For a general introduction to branch-and-bound algorithms, see e.g., Nemhauser and Wolsey (1988).

For each record, the SLICE algorithm sets up a binary tree, as illustrated in Figure 4.1. In the root node of the tree, we start with the original set of edits and we select one of the variables. From the root node, two branches are added to the tree. In the first branch, the original value of the selected variable in the record is assumed to be correct, and in the second branch this value is assumed to be erroneous. Both assumptions correspond with a transformation of the set of edits, to be outlined below, after which the selected variable is no longer involved in the edits: the selected variable has been *treated*. Next, one of the remaining variables is selected and the operation is repeated.

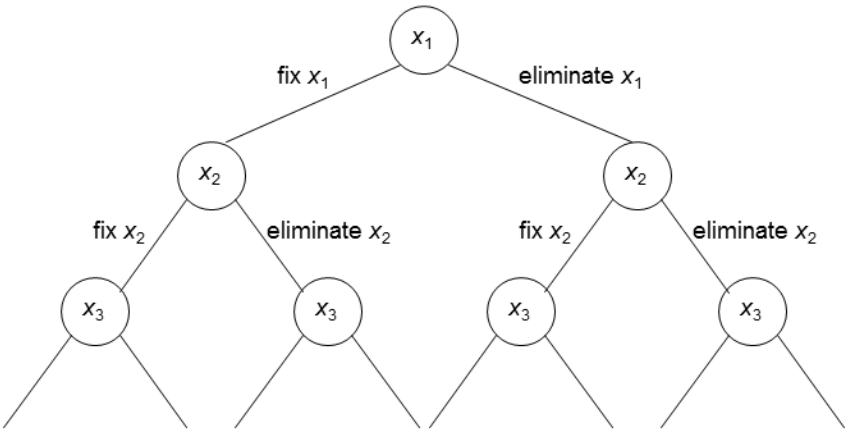


Figure 4.1: Illustration of the branch-and-bound algorithm as a binary tree.

Once all variables have been treated, the algorithm reaches an end node of the tree. It is seen that, together, the end nodes of the binary tree enumerate all possible choices of erroneous subsets of variables. The transformed set of edits corresponding to an end node does not involve any variables, so it must either be empty or consist of elementary relations such as "$1 \geq 0$" (a tautology) and "$-1 \geq$

0" (a self-contradicting statement). As will be discussed below, it is possible to satisfy the original edits by only imputing the variables that have been considered erroneous in the branch leading to an end node, if and only if the transformed set of edits for that end node contains no self-contradicting statements. Using this property, all feasible solutions to the error localisation problem may be identified. Moreover, since we are only interested in feasible solutions that minimise target function (4.8), a branch of the tree may be pruned as soon as we find that it only leads to end nodes corresponding with infeasible or suboptimal solutions.

We will now outline the transformations of the set of edits that occur, depending on whether a variable is assumed to be correct or erroneous. A variable that is assumed to be correct is removed from the edits by simply substituting the original value from the record in the edits. This is called *fixing* a variable to its original value. A variable that is assumed to be erroneous is removed from the edits by a more complex operation, called *eliminating* a variable from the edits. Numerical variables and categorical variables are eliminated by two different but equivalent methods.

To eliminate a numerical variable, say x_g, from a set of edits having the general forms (4.3) and (4.4), we generate logically implied edits by considering all pairs of edits ψ^s and ψ^t that involve x_g. We first check whether $F_j^s \cap F_j^t \neq \emptyset$ for all $j = 1, \ldots, m$; if any of these intersections yields the empty set, then the pair ψ^s and ψ^t does not generate an implied edit. If the numerical THEN-condition of one of the edits, say ψ^s, is an equality, then this equality may be solved for x_g. By substituting the resulting expression for x_g in the THEN-condition of ψ^t, we obtain the numerical THEN-condition of the implied edit. The categorical IF-condition of the implied edit is found by taking the non-empty intersections $F_j^* = F_j^s \cap F_j^t$ for $j = 1, \ldots, m$.

If the numerical THEN-conditions of ψ^s and ψ^t are both inequalities, the algorithm uses a technique called *Fourier-Motzkin elimination* [see e.g., Williams (1986)] to generate an implied edit. A pair of edits is relevant for this elimination method only if the coefficients of x_g have opposite signs, so we may assume without loss of generality that $a_{sg} < 0$ and $a_{tg} > 0$. The implied edit generated from ψ^s and ψ^t may then be written as [cf. De Waal and Quere (2003)]:

$$\begin{aligned} \psi^* : \quad & \text{IF} \quad (v_1, \ldots, v_m) \in F_1^* \times \cdots \times F_m^* \\ & \text{THEN} \quad (x_1, \ldots, x_p) \in \left\{ a_1^* x_1 + \cdots + a_p^* x_p + b^* \geq 0 \right\} \end{aligned} \tag{4.9}$$

with $a_j^* = a_{tg} a_{sj} - a_{sg} a_{tj}$, $b^* = a_{tg} b_s - a_{sg} b_t$, and $F_j^* = F_j^s \cap F_j^t$ as above. This edit does not involve x_g, since $a_g^* = 0$. In this manner, implied edits are generated by considering all pairs of edits that involve x_g. These edits are added to the set of

original edits that do not involve x_g, to find the transformed set of edits obtained by eliminating x_g.

For the elimination of categorical variables, De Waal and Quere (2003) make the simplifying assumption that these variables are only selected when all numerical variables have been treated. This assumption implies that categorical variables are always eliminated from purely categorical edits of the form (4.5). To eliminate a categorical variable, say v_g, from a set of edits of the form (4.5), a technique is used that was first described by Fellegi and Holt (1976).

Consider all minimal sets of edits T with the following properties:

$$F_g^*(T) = \bigcup_{k \in T} F_g^k = D_g \tag{4.10}$$

and

$$F_j^*(T) = \bigcap_{k \in T} F_j^k \neq \emptyset, \qquad \text{for } j = 1, \ldots, g-1, g+1, \ldots, m. \tag{4.11}$$

Here, by "minimal" we mean that property (4.10) does not hold for any set $T' \subset T$. Each of these minimal sets T generates an implied edit:

$$\text{IF } (v_1, \ldots, v_m) \in F_1^*(T) \times \cdots \times F_m^*(T) \text{ THEN } \emptyset, \tag{4.12}$$

which does not involve v_g because of property (4.10). These implied edits are added to the set of original edits that do not involve v_g, to find the transformed set of edits obtained by eliminating v_g.

It should be clear that the computational work of the algorithm lies mainly in the elimination steps. In particular, it is known that the number of implied edits under Fourier-Motzkin elimination may be exponential in the number of eliminated variables (Schrijver, 1986).

A fundamental property of both elimination techniques, for numerical and categorical variables, is exhibited by the following result. Consider a system of implied edits Ψ_1 obtained by eliminating x_g or v_g from a system of edits Ψ_0. Then the original values of the untreated variables satisfy all edits in Ψ_1, if and only if there exists a value for x_g or v_g that, together with these original values, satisfies all edits in Ψ_0. For a proof, see Theorem 8.1 in De Waal (2003b) or Theorem 4.3 in De Waal et al. (2011). The above-mentioned correspondence between end nodes without self-contradicting elementary relations and feasible solutions to the error localisation problem follows from a repeated application of this fundamental property.

4.3 An error localisation problem with hard and soft edits

In the formulation of the error localisation problem given in Section 4.2.2, which is based on the Fellegi-Holt paradigm, it is tacitly assumed that all edits are hard edits. Consequently, the only subsets of the variables that are considered feasible solutions to this problem are those which can be imputed to make the record consistent with respect to all edits. As mentioned in the introduction, this interpretation of all edits as hard edits can lead to systematic differences between automatic editing and manual editing, because it precludes a meaningful use of soft edits. In this section, we suggest a new formulation of the error localisation problem which distinguishes between hard and soft edits.

Let Ψ denote the set of edits to be used in the error localisation problem. We assume that this set can be partitioned into two disjoint subsets: $\Psi = \Psi_H \cup \Psi_S$. The edits in Ψ_H are hard edits; the edits in Ψ_S are soft edits. From now on, a subset of the variables is considered a feasible solution to the error localisation problem if it can be imputed to satisfy all edits in Ψ_H. Moreover, we want to use the status of the imputed record with respect to the edits in Ψ_S as auxiliary information in the choice of an optimal solution. This may be done by adding another term to (4.8).

More precisely, the objective of the new error localisation problem is to find a subset of the variables which (i) can be imputed so that the imputed record satisfies all edits in Ψ_H, and (ii) minimises the following target function:

$$D = \lambda D_{\mathrm{FH}} + (1 - \lambda) D_{\mathrm{soft}}, \tag{4.13}$$

where D_{soft} represents the costs associated with failed edits in Ψ_S. The parameter $\lambda \in [0, 1]$ determines the relative contribution of both terms in (4.13). The original Fellegi-Holt paradigm is recovered as a special case by choosing $\lambda = 1$. Thus, the new error localisation problem can be seen as a generalisation of the old one.

In order to use (4.13) in practice, one has to choose an expression for D_{soft}. Probably the easiest way to assign costs to failed soft edits is to associate a fixed *failure weight* s_k to each edit in Ψ_S, and to define D_{soft} as the sum of the failure weights of the soft edits that remain failed:

$$D_{\mathrm{soft}} = \sum_{k=1}^{K_S} s_k z_k, \tag{4.14}$$

with K_S the number of edits in Ψ_S and z_k a binary variable such that $z_k = 1$ if the k^{th} soft edit is failed and $z_k = 0$ otherwise. The failure weights may be chosen by subject-matter experts, analogously to the confidence weights, to express the importance that is attached to different soft edits from a subject-matter related

point of view. Alternatively, the failure weights may be based on the proportion of records that fail each soft edit in a historical data set which has been edited manually.

A drawback of using fixed failure weights is that they do not take the size of the edit failures into account: every record that fails a particular soft edit receives the same contribution to D_{soft}, namely s_k. By contrast, a human editor sees a soft edit failure as an indication that an observed combination of values is suspicious, and the degree of suspicion depends on the size of the edit failure: a small failure is ignored more easily than a large failure. Hence, it seems interesting to take the size of the edit failures into account in D_{soft}. This point will be taken up in Section 4.8, since it introduces certain additional difficulties. For now, we assume that expression (4.14) is used.

We should mention that taking soft restrictions into account by adding an appropriate term to a target function is a well-known technique in mathematical optimisation. The idea is related to other optimisation techniques, such as Lagrangian relaxation [see e.g., Nemhauser and Wolsey (1988)]. One example of a practical application with soft constraints is that of the so-called benchmarking problem for national accounts (Magnus et al., 2000). To our best knowledge, the application in the context of the error localisation problem is new.

We should also note that expression (4.13) is in some respects similar to the minimisation criterion of the Nearest-neighbour Imputation Methodology (NIM) developed by Statistics Canada for editing demographic census data (Bankier et al., 2000; Bankier and Crowe, 2009). In particular, the NIM also departs from the Fellegi-Holt paradigm by minimising a convex combination of two terms, the first measuring the amount of imputation and the second measuring the plausibility of the imputed record.

4.4 A short theory of edit failures

4.4.1 Numerical data

Having formulated a new error localisation problem, we will now show how this problem may be solved by an adapted version of the branch-and-bound algorithm of De Waal and Quere (2003). To do this, we first need to extend the fundamental property mentioned at the end of Section 4.2.3 to the case that some of the edits may be failed. For convenience, we shall first examine the case of purely numerical data. The next subsection examines the case of purely categorical and mixed data.

In the case of purely numerical data, all edits take the form (4.6) or (4.7). Moreover, the implied edit (4.9) is reduced to its numerical part. The fundamental

property given at the end of Section 4.2.3 implies in particular the following: if a given set of values for $x_1, \ldots, x_{g-1}, x_{g+1}, \ldots, x_p$ does not satisfy the implied edit (4.9), then it is impossible to find a value for x_g that satisfies ψ^s and ψ^t simultaneously. However, it is still possible in this case to find a value for x_g that satisfies one of the edits ψ^s or ψ^t. This observation, which is more or less trivial, forms the basis for the proof of Theorem 4.1 below.

Suppose that, at some point during an execution of the branch-and-bound algorithm of De Waal and Quere (2003), q numerical variables have been treated (i.e., either eliminated or fixed). We denote the current set of edits by Ψ_q, and the edits in this set by ψ_q^k. By definition, $\Psi_0 \equiv \Psi$, the original set of edits. It is possible to associate with each current edit ψ_q^k an index set B_q^k, which contains the indices of all the original edits that have been used, directly or indirectly, to derive this edit. In fact, B_q^k can be defined recursively as follows:

- For an original edit ψ_0^k, we define $B_0^k := \{k\}$.
- For an edit ψ_q^k which is derived from one other edit ψ_{q-1}^l, either by fixing a variable to its original value or by simply copying the edit, we define $B_q^k := B_{q-1}^l$.
- For an edit ψ_q^k which is derived by eliminating a variable from a set of edits ψ_{q-1}^t ($t \in T$), we define $B_q^k := \bigcup_{t \in T} B_{q-1}^t$.

Note that, for numerical data, the set T in the last item always contains exactly two edits. Larger edit sets may be encountered in the categorical case considered below.

A set B is called a *representing set* of a collection of sets $B_q^{k_1}, \ldots, B_q^{k_r}$ if it contains at least one element from each of $B_q^{k_1}, \ldots, B_q^{k_r}$; see, for instance, Mirsky (1971, p. 25). It should be noted that, in our case, a representing set B identifies a subset of Ψ_0, the set of original edits. We can now formulate the following theorem.

Theorem 4.1 *Suppose that q numerical variables have been treated and that the current set of numerical edits can be partitioned as $\Psi_q = \Psi_q^{(1)} \cup \Psi_q^{(2)}$, where the edits in $\Psi_q^{(1)}$ are satisfied by the original values of the $p - q$ remaining variables, and the edits in $\Psi_q^{(2)}$ are failed. Let B be a representing set of the index sets B_q^k for all $\psi_q^k \in \Psi_q^{(2)}$. Then there exist values for the eliminated variables that, together with the original values of the other variables, satisfy all original edits except those in B.*

Proof. The proof of this theorem is given in Appendix 4.A.1. □

Example. Suppose that there are three numerical variables (x_1, x_2, x_3) that should satisfy the following eight edits:

$$\begin{array}{lrcr}
\psi_0^1: & x_1 + x_2 + x_3 & = & 20 \\
\psi_0^2: & x_1 - x_2 & \geq & 3 \\
\psi_0^3: & -x_1 + x_2 & \geq & -6 \\
\psi_0^4: & -x_1 + x_3 & \geq & 5 \\
\psi_0^5: & x_1 - x_3 & \geq & -10 \\
\psi_0^6: & x_1 & \geq & 0 \\
\psi_0^7: & x_2 & \geq & 0 \\
\psi_0^8: & x_3 & \geq & 0
\end{array}$$

The record $(x_1^0, x_2^0, x_3^0) = (10, 1, -3)$ is inconsistent with respect to these edits. Upon eliminating x_1 from the original set of edits, we find the following updated set of edits:

$$\begin{array}{lrcrl}
\psi_1^1: & -2x_2 - x_3 & \geq & -17 & (B_1^1 = \{1, 2\}) \\
\psi_1^2: & 2x_2 + x_3 & \geq & 14 & (B_1^2 = \{1, 3\}) \\
\psi_1^3: & x_2 + 2x_3 & \geq & 25 & (B_1^3 = \{1, 4\}) \\
\psi_1^4: & -x_2 - 2x_3 & \geq & -30 & (B_1^4 = \{1, 5\}) \\
\psi_1^5: & -x_2 - x_3 & \geq & -20 & (B_1^5 = \{1, 6\}) \\
\psi_1^6: & x_2 & \geq & 0 & (B_1^6 = \{7\}) \\
\psi_1^7: & x_3 & \geq & 0 & (B_1^7 = \{8\}) \\
\psi_1^8: & 0 & \geq & -3 & (B_1^8 = \{2, 3\}) \\
\psi_1^9: & -x_2 + x_3 & \geq & 8 & (B_1^9 = \{2, 4\}) \\
\psi_1^{10}: & x_2 - x_3 & \geq & -16 & (B_1^{10} = \{3, 5\}) \\
\psi_1^{11}: & 0 & \geq & -5 & (B_1^{11} = \{4, 5\}) \\
\psi_1^{12}: & x_2 & \geq & -6 & (B_1^{12} = \{3, 6\}) \\
\psi_1^{13}: & x_3 & \geq & 5 & (B_1^{13} = \{4, 6\})
\end{array}$$

The index set B_1^k is displayed in brackets next to each edit.

By substituting the original values of x_2 and x_3 in the current set of edits, we see that ψ_1^2, ψ_1^3, ψ_1^7, ψ_1^9, and ψ_1^{13} are failed. The set $B = \{1, 4, 8\}$ is a representing set for the associated index sets B_1^k. According to Theorem 4.1, there exists a value for x_1 which, together with the original values of x_2 and x_3, satisfies the original edits apart from ψ_0^1, ψ_0^4, and ψ_0^8. That this assertion is correct can be seen by substituting $x_2^0 = 1$ and $x_3^0 = -3$ into the original set of edits; in fact, any value $x_1 \in [4, 7]$ will do. □

The importance of Theorem 4.1 is that it enables one to evaluate, at each node of the branch-and-bound algorithm, which combinations of the original edits could be satisfied by imputing the variables that have been eliminated so far, and also which edits would remain failed. In particular, if we distinguish between hard and

soft original edits, then this result makes it possible to use the branch-and-bound algorithm to find all feasible solutions to the new error localisation problem from Section 4.3, and also to evaluate, for each feasible solution, which of the soft edits remain failed, and hence to evaluate the value of D_{soft}. This idea will be elaborated in Section 4.5.

Interestingly, the above-defined sets B_q^k may also be used to identify redundant edits, i.e., edits that follow directly from a combination of the other edits. According to a result found independently by Černikov (1963) and Kohler (1967), when q variables have been eliminated by Fourier-Motzkin elimination, all edits with more than $q + 1$ elements in B_q^k are redundant; see also Williams (1986) and De Jonge and Van der Loo (2011) for a discussion of this result.

4.4.2 Categorical and mixed data

We shall now derive a similar result to Theorem 4.1 for the case of purely categorical data. At the end of this section, we shall combine the two results so that they may also be applied to mixed data.

In the case of purely categorical data, all edits take the form (4.5). Let us consider the elimination method for categorical variables described in Section 4.2.3. If a given set of values for $v_1, \ldots, v_{g-1}, v_{g+1}, \ldots, v_m$ does not satisfy the implied edit (4.12), then it is not possible to find a value for v_g that, together with the other values, satisfies all edits ψ^k with $k \in T$ simultaneously. This is true because, by property (4.11), $F_j^*(T) \subseteq F_j^k$ for all $j \neq g$ and all $k \in T$. Hence, if (4.12) is failed by $v_1, \ldots, v_{g-1}, v_{g+1}, \ldots, v_m$, then plugging these values into an original edit with $k \in T$ produces a non-degenerate univariate edit for v_g. Moreover, every possible value of v_g fails at least one of these univariate edits, because of property (4.10). Interestingly, it is still always possible in this case to find a value for v_g that satisfies all edits in T but one. This follows from property (4.10) and the fact that T is a minimal set having this property: for each $k \in T$, F_g^k must contain at least one value from D_g that is not covered by any other F_g^l with $l \in T$.

We now present the analogue of Theorem 4.1 for categorical data, using the same notation as for numerical data. In particular, the recursive definition of B_q^k is exactly the same as in Section 4.4.1.

Theorem 4.2 *Suppose that q categorical variables have been treated and that the current set of categorical edits can be partitioned as $\Psi_q = \Psi_q^{(1)} \cup \Psi_q^{(2)}$, where the edits in $\Psi_q^{(1)}$ are satisfied by the original values of the $m - q$ remaining variables, and the edits in $\Psi_q^{(2)}$ are failed. Let B be a representing set of the index sets B_q^k for all $\psi_q^k \in \Psi_q^{(2)}$. Then there exist values for the eliminated variables that, together*

with the original values of the other variables, satisfy all original edits except those in B.

Proof. The proof of this theorem is given in Appendix 4.A.2. □

For an example that illustrates the use of this theorem, see Scholtus (2011b).

Finally, we remark that Theorem 4.1 and Theorem 4.2 can be used together when the data are a mix of categorical and numerical variables. This follows from the structure of the branch-and-bound algorithm of De Waal and Quere (2003), where categorical variables are only treated once all numerical variables have been eliminated or fixed. Hence, the two results may be applied consecutively. There is a slight difference in the procedure for eliminating numerical variables, namely that implied edits are only generated from pairs of edits having an overlapping IF-condition; see Section 4.2.3. However, this does not affect the correctness of Theorem 4.1.

4.5 An algorithm for solving the error localisation problem with hard and soft edits

We shall now describe an adapted version of the branch-and-bound algorithm of De Waal and Quere (2003), which may be used to solve the error localisation problem defined in Section 4.3. The basic setup of the algorithm is the same as in Section 4.2.3. In particular, the procedures for eliminating and fixing variables are carried out the same way as in the original algorithm.

The main difference is that now in each node, the current set of edits Ψ_q is partitioned into a current set of hard edits Ψ_{qH} and a current set of soft edits Ψ_{qS}. For the root node, the partition simply follows that of the original set of edits, i.e., $\Psi_{0H} = \Psi_H$ and $\Psi_{0S} = \Psi_S$. For all other nodes, the partition can be summarised as follows: if an edit is generated only from hard edits, then it is a hard edit; if any soft edits are involved in its generation, then it is a soft edit. Furthermore, for each soft edit $\psi_{qS}^k \in \Psi_{qS}$, we construct an index set B_{qS}^k – analogous to B_q^k in Section 4.4 – which contains the indices of all the original *soft* edits ψ_{0S}^k that were involved, directly or indirectly, in its generation.

Having generated Ψ_{qH} and Ψ_{qS} for a particular node, we can fill in the original values of the variables that have not been treated yet, to check which of these edits are failed. In the old algorithm, this check could have two possible outcomes: either more variables need to be eliminated (at least one of the edits is failed), or a feasible solution has been found (none of the edits is failed). In the new algorithm, three different situations may arise.

First of all, if at least one edit in Ψ_{qH} is failed, then the variables that have been eliminated so far cannot be imputed to satisfy the original hard edits. Hence, more variables need to be eliminated. In this case, we continue the generation of branches from the current node.

A second possibility is that none of the edits in Ψ_{qH} or Ψ_{qS} is failed. This means that the variables that have been eliminated so far can be imputed to satisfy all the original edits, both hard and soft. Thus, a feasible solution has been found, for which the value of target function (4.13) equals $D = \lambda D_{\mathrm{FH}}$. If this value is smaller than or equal to the value of (4.13) for the best solution found so far, say $D_{\min}$, then the new solution is stored. Otherwise, it is discarded. Either way, it is not useful to continue the algorithm from the current node, because if more variables are eliminated, the value of D can only increase. Hence, we return to the last previous branch that has not been completely searched yet and continue the algorithm from there.

The last possibility is that the edits in Ψ_{qH} are satisfied, but that at least one edit in Ψ_{qS} is failed. In this case, the variables that have been eliminated so far can be imputed to satisfy the original hard edits, but not all the original soft edits. Hence, a feasible solution to the error localisation problem has been found, but the contribution of D_{soft} to D is non-zero. According to Theorem 4.1 or Theorem 4.2, it is possible to satisfy all original soft edits, except those in a representing set B of the index sets B^k_{qS} for all failed edits in Ψ_{qS}. Since this property is shared by all representing sets, we are free to choose B in such a way that D_{soft} is minimised, given the selection of variables to impute. If expression (4.14) is used for D_{soft}, then the optimal choice of B can be found by solving the following minimisation problem:

$$\begin{array}{l} \min \sum_{k=1}^{K_S} s_k z_k, \text{ under the conditions that:} \\ \sum_{k \in B^l_{qS}} z_k \geq 1, \text{ for all failed } \psi^l_{qS} \in \Psi_{qS}, \\ z_k \in \{0,1\}\,,\ k = 1, \ldots, K_S. \end{array} \tag{4.15}$$

This is a standard binary linear optimisation problem for which algorithms are available [see e.g., Nemhauser and Wolsey (1988)]. The solution consists of a vector $(z^*_1, \ldots, z^*_{K_S})$ of zeros and ones. The associated optimal representing set is $B^* = \{k\,:\,z^*_k = 1\}$ and the associated contribution of D_{soft} to D is precisely the minimal value of problem (4.15), say

$$D^*_{\mathrm{soft}} = \sum_{k=1}^{K_S} s_k z^*_k = \sum_{k \in B^*} s_k.$$

As in the previous case, the value $D = \lambda D_{\mathrm{FH}} + (1-\lambda) D^*_{\mathrm{soft}}$ is compared to $D_{\min}$. If $D \leq D_{\min}$, then the current solution is stored, otherwise it is discarded.

Either way, it is meaningful in this case to continue the algorithm from the current node, because eliminating more variables may lead to a lower value of the target function. This can happen because a solution that imputes more variables typically fails fewer soft edits. Therefore, we continue the generation of branches from the current node.

The correctness of this algorithm follows from the correctness of the original algorithm of De Waal and Quere (2003) and the theory presented in Section 4.4. The index sets B_q^k only have to be computed for the soft edits, because a subset of the variables is never considered a feasible solution to the error localisation problem when at least one of the hard edits remains failed. This means that, in every application of Theorem 4.1 or Theorem 4.2, all implied edits in Ψ_{qH} must be contained in $\Psi_q^{(1)}$. Finally, we note that the new algorithm reduces to the original algorithm of De Waal and Quere (2003) in the special case that no soft edits have been specified.

4.6 Example

To illustrate the algorithm of Section 4.5, we shall apply it to a small example with numerical data. This is essentially an example from De Waal (2003b) to which we have added a distinction between hard and soft edits. For a somewhat larger example involving a mix of categorical and numerical variables, see Scholtus (2011b).

In a fictitious business survey, there are four numerical variables: *total turnover* (T), *profit* (P), *total costs* (C), and *number of employees* (N). The following hard edits and soft edits have been identified:

$$
\begin{array}{lrcll}
\psi_{0H}^1: & T - C - P & = & 0 & \\
\psi_{0H}^2: & T & \geq & 0 & \\
\psi_{0H}^3: & C & \geq & 0 & \\
\psi_{0H}^4: & N & \geq & 0 & \\
\psi_{0H}^5: & 550N - T & \geq & 0 & \\
\psi_{0S}^1: & 0.5T - P & \geq & 0 & (B_{0S}^1 = \{1\}) \\
\psi_{0S}^2: & P + 0.1T & \geq & 0 & (B_{0S}^2 = \{2\})
\end{array}
$$

Consider the following unedited record:

$$(T^0, P^0, C^0, N^0) = (100; 40,000; 60,000; 5).$$

This record fails the first hard edit and the first soft edit. The confidence weights of the variables are: $(w_T, w_P, w_C, w_N) = (2, 1, 1, 3)$. We choose the failure weights of the two soft edits to be $s_1 = s_2 = 2$. Finally, we choose $\lambda = 1/2$ in expression (4.13).

Suppose that the variable P is selected first. In the branch where P is eliminated from the original edits, we obtain the following new set of edits:

$$
\begin{array}{lrcll}
\psi^1_{1H}: & T & \geq & 0 & & (\psi^2_{0H}) \\
\psi^2_{1H}: & C & \geq & 0 & & (\psi^3_{0H}) \\
\psi^3_{1H}: & N & \geq & 0 & & (\psi^4_{0H}) \\
\psi^4_{1H}: & 550N - T & \geq & 0 & & (\psi^5_{0H}) \\
\psi^1_{1S}: & -0.5T + C & \geq & 0 & (B^1_{1S} = \{1\}) & (\psi^1_{0H}, \psi^1_{0S}) \\
\psi^2_{1S}: & 1.1T - C & \geq & 0 & (B^2_{1S} = \{2\}) & (\psi^1_{0H}, \psi^2_{0S}) \\
\psi^3_{1S}: & 0.6T & \geq & 0 & (B^3_{1S} = \{1,2\}) & (\psi^1_{0S}, \psi^2_{0S})
\end{array}
$$

We have indicated in brackets from which of the previous edits each new edit is derived. The third soft edit ψ^3_{1S} is in fact equivalent to the first hard edit ψ^1_{1H}, which means that it can be discarded.

Upon substituting the original values $(T^0, C^0, N^0) = (100; 60,000; 5)$ into the current edits, it is seen that all edits are satisfied except for ψ^2_{1S}. Since all hard edits are satisfied, identifying only the original value of P as erroneous is a feasible solution to the error localisation problem. Moreover, since $B = \{2\}$ is (trivially) a minimal representing set of B^2_{1S}, it is possible to impute a value for P which satisfies all the original edits except for ψ^2_{0S}. Hence, the value of target function (4.13) for this solution is $(w_P + s_2)/2 = 3/2$.

Possibly, the current solution may be improved by eliminating another variable, say C, from the current set of edits. This yields:

$$
\begin{array}{lrcll}
\psi^1_{2H}: & T & \geq & 0 & & (\psi^1_{1H}) \\
\psi^2_{2H}: & N & \geq & 0 & & (\psi^3_{1H}) \\
\psi^3_{2H}: & 550N - T & \geq & 0 & & (\psi^4_{1H}) \\
\psi^1_{2S}: & 1.1T & \geq & 0 & (B^1_{2S} = \{2\}) & (\psi^2_{1H}, \psi^2_{1S}) \\
\psi^2_{2S}: & 0.6T & \geq & 0 & (B^2_{2S} = \{1,2\}) & (\psi^1_{1S}, \psi^2_{1S})
\end{array}
$$

Each of the two new soft edits is redundant, because both are equivalent to hard edit ψ^1_{2H}. In fact, the remaining original values $(T^0, N^0) = (100, 5)$ satisfy all the current edits. This means that P and C can be imputed to satisfy all the original edits, both hard and soft. The value of target function (4.13) for this solution equals $(w_P + w_C)/2 = 1$. Thus, the new solution improves on the previous one. Moreover, this solution cannot be improved further by eliminating more variables in the current branch of the binary tree.

If the rest of the binary tree is explored, it eventually turns out that the best solution found so far (impute P and C) is also the optimal solution. A possible consistent record obtained by imputing P and C is: $(T, P, C, N) = (100; 40; 60; 5)$. This solution has the nice interpretation that the original values of *profit* and *total costs* were overstated by a factor of $1,000$. It is of interest to note that, if only the

hard edits are used in this example, then the first solution found above (impute only P) is the optimal solution. In that case, there is only one way to obtain a consistent record: $(T, P, C, N) = (100; -59,900; 60,000; 5)$. This illustrates that, in this example at least, soft edits are important for finding imputations that are not only consistent with the hard edits, but also plausible.

4.7 Application

To test the new error localisation algorithm in practice, a prototype implementation was written using the R programming language. This prototype draws heavily on the existing error localisation functionality in R that was made available in the `editrules` package (De Jonge and Van der Loo, 2011; Van der Loo and De Jonge, 2011).

To test the prototype, an artificial data set was constructed by selecting twelve numerical variables $(x_1, \ldots, x_{12})$ from the Netherlands' structural business statistics of 2007 for the wholesale sector. We selected all records pertaining to medium-sized businesses (with 10 to 100 employees) that had been edited manually during regular production, and divided these into two data sets of 728 records each. Both of the original data sets were considered error-free. We introduced a substantial number of random errors into one of the data sets by applying the following procedure:

- in 4% of the original non-zero values, two digits were interchanged;
- in 4% of the original non-zero values, a random digit was added;
- in 4% of the original non-zero values, a random digit was omitted;
- in 4% of the original non-zero values, a random digit was replaced by another digit;
- 4% of the original non-zero values were multiplied by 25;
- 4% of the original non-zero values were divided by 25 and rounded to the nearest integer;
- 6% of the original non-zero values were replaced by zero;
- 5% of the original zero values were replaced by random integers from the set $\{1, \ldots, 1000\}$;
- 10% of the original values of x_{11} and x_{12} were multiplied by -1.

Table 4.1: The edits that were used in the test application.

hard edits:	$x_1 + x_2 = x_3$
	$x_2 = x_4$
	$x_5 + x_6 + x_7 = x_8$
	$x_3 + x_8 = x_9$
	$x_9 - x_{10} = x_{11}$
	$x_j \geq 0$ ($j = 1, \ldots, 10$ and $j = 12$)
soft edits:	$x_2 \geq 0.5x_3$
	$x_3 \geq 0.9x_9$
	$x_5 + x_6 \geq x_7$
	$x_9 \geq 50x_{12}$
	$x_9 \leq 5000x_{12}$
	$x_{11} \leq 0.4x_9$
	$x_{11} \geq -0.1x_9$
	$x_{12} \geq 1$
	$x_{12} \geq 5$
	$x_{12} \leq 100$

This procedure was carried out in such a way that at most one change could occur in each value. The second data set was left error-free and was used as reference data.

Table 4.1 shows the hard and soft edits that were applied to the test data. The hard edits were copied from the regular production system. The soft edits were identified by examining a number of univariate and bivariate distributions in the reference data.

The error localisation algorithm was applied to the data set with artificial errors using several different set-ups. Throughout, all confidence weights w_j^N were chosen equal to 1, and the parameter λ in (4.13) was chosen equal to $1/2$. We considered the following approaches:

A. The first test used only the hard edits from Table 4.1.

B. The second test used all edits from Table 4.1, with all edits interpreted as hard edits.

C. The third test used all edits from Table 4.1, with a distinction between hard and soft edits. Each soft edit received the same fixed failure weight $s_k = 1$.

D. The fourth test was similar to the third test, but with fixed failure weights that differed between soft edits. For each soft edit, s_k was calculated as the fraction of records in the reference data set that satisfied the edit. Thus, a

soft edit received a lower failure weight if it was failed more often in the reference data set, and vice versa. The rationale behind this is that all soft edit failures occurring in the reference data were caused by unusual, but correct combinations of values. By associating low weights to soft edits that are often failed in the reference data, we ensure that these edits may also be failed more easily when editing the test data.

Since the distribution of errors in our test data set was known, we could directly evaluate the performance of each automatic error localisation approach. To this end, we used several quality indicators. Consider the following 2×2 contingency table:

		detected:	
		error	no error
true:	error	TP	FN
	no error	FP	TN

The first quality indicator measures the proportion of true errors that were missed by the algorithm (proportion of false negatives):

$$\alpha = \frac{FN}{TP + FN}.$$

The second quality indicator measures the proportion of correct values that were mistaken for errors by the algorithm (proportion of false positives):

$$\beta = \frac{FP}{FP + TN}.$$

The third quality indicator measures the overall proportion of wrong decisions made by the algorithm:

$$\delta = \frac{FN + FP}{TP + FN + FP + TN}.$$

These three indicators evaluate the performance of the algorithm with respect to identifying individual values as correct or erroneous. They have been used in previous evaluation studies; see, for instance, Pannekoek and De Waal (2005).

To evaluate the performance of the algorithm from a slightly different angle, we also calculated the percentage of records for which the algorithm found exactly the right solution – that is, the solution that identifies as erroneous all erroneous values and only these. This indicator is denoted by ρ. A good editing approach should have low scores on α, β, and δ, but a high score on ρ.

Table 4.2 shows the values of the quality indicators for editing approaches A, B, C, and D. It can be seen that approach B is outperformed by the other approaches

Table 4.2: Results of automatic error localisation for the artificial data.

	quality indicators			
approach	α	β	δ	ρ
A	0.364	0.047	0.115	40%
B	0.232	0.131	0.153	37%
C	0.227	0.060	0.096	47%
D	0.253	0.037	0.083	52%

on all measures, except for the proportion of missed errors. Thus, using the soft edits as if they were hard edits does not work well for this data set; in fact, better results are achieved by approach A, which does not use the soft edits at all. It can also be seen that approaches C and D, which use the new algorithm to take the soft edits into account, yield better results than approaches A and B, which use the old algorithm. Overall, approach D appears to achieve the best results in this experiment. Compared with approach A, approach D in fact correctly identifies more errors *and* more correct values.

It should be noted that, under the old definition of the error localisation problem, approaches A and B represent the two extreme options available for using soft edits: either not using them, or using them all as hard edits. As a compromise between these options, one could also decide to use only a subset of the soft edits as hard edits and discard the others. We did not test this approach during the experiment. One might expect that it would lead to scores on the α, β, δ, and ρ measures in between those of approaches A and B.

4.8 Conclusion

In this chapter, we proposed a new formulation of the error localisation problem which can take the distinction between hard and soft edits into account. In addition, we showed that a modified version of the branch-and-bound algorithm of De Waal and Quere (2003) can be used to solve this new error localisation problem. It was suggested that this new algorithm can be used to increase the quality of automatic editing. This suggestion was confirmed by the empirical results reported in Section 4.7, although it should be stressed that these results were obtained with data containing synthetic errors. An application is currently being investigated of the new error localisation algorithm to realistic data.

It remains an open problem how the costs of soft edit failures may best be modelled, i.e., how the term D_{soft} in (4.13) should be defined. The different results with approaches C and D in Section 4.7 demonstrate that the quality of automatic

error localisation may be improved by a suitable choice of failure weights. It will be interesting to see to what extent the quality of automatic editing may be improved further by experimenting with different combinations of failure weights s_k, confidence weights w_j, and the balancing parameter λ in (4.13).

Other forms of D_{soft} than (4.14) could also be considered, including forms that depend on the sizes of the soft edit failures. As mentioned in Section 4.3, it is intuitively appealing to take the amounts by which soft edits are failed into account in the error localisation problem, so that larger soft edit failures yield higher values of D_{soft}. One interesting choice for D_{soft} could be the Mahalanobis distance of soft edit failures, as suggested by Hedlin (2003) in a different context. It should be noted that the algorithm from Section 4.5 may be used to solve the error localisation problem for all choices of D_{soft} that can be expressed as (reasonably well-behaved) functions of $z_1, \ldots, z_{K_S}$. One simply uses the appropriate expression for D_{soft} as the target function in problem (4.15). On the other hand, if D_{soft} depends explicitly on the sizes of the soft edit failures, then we have to resort to a more complex approach. In particular, the information provided by Theorems 4.1 and 4.2 is no longer sufficient, because we now need to know not only which soft edits will be failed after imputation but also the amounts by which they will be failed. An approach for solving the error localisation problem in this more complicated situation can be found in Scholtus (2011b). Scholtus and Göksen (2012) experimented with many different forms of D_{soft} in a simulation study involving both real and synthetic data.[1]

In summary, it remains to be seen how the theoretical results outlined in this chapter should be applied to obtain the best results in practice. Nevertheless, given that subject-matter experts use the conceptual difference between hard and soft edits during manual editing, it seems evident that the new error localisation algorithm has the potential to increase the quality of automatic editing.

Appendix 4.A Proofs

4.A.1 Proof of Theorem 4.1

In order to prove Theorem 4.1, it is convenient to prove first an auxiliary lemma. Suppose that Ψ_q is obtained from Ψ_{q-1} by eliminating x_g. We define, for each edit

[1] It is shown in Scholtus (2015) that the new error localisation problem given by (4.13) with D_{soft} of the form (4.14) can be re-formulated as an instance of the original Fellegi-Holt-based error localisation problem involving only hard edits, by introducing auxiliary variables and re-writing the soft edits. This result does not extend to the problem with other forms of D_{soft}. Thus, one practical advantage of using the simple form (4.14) is that in this case existing algorithms for automatic editing could be used, with some very minor modifications, to solve the new error localisation problem.

ψ_q^k, the index set A_q^k of the edit(s) in Ψ_{q-1} from which it has been derived. That is to say, we define $A_q^k := \{l\}$ if ψ_q^k is obtained by copying the edit ψ_{q-1}^l, and we define $A_q^k := \{s, t\}$ if ψ_q^k is obtained by eliminating x_g from the pair of edits $(\psi_{q-1}^s, \psi_{q-1}^t)$.

Lemma 4.1 *Consider the situation of Theorem 4.1 for $q \geq 1$, and suppose that x_g has been eliminated to obtain Ψ_q from Ψ_{q-1}. Let A be a representing set of the index sets A_q^k belonging to all $\psi_q^k \in \Psi_q^{(2)}$. Then there exists a value for x_g that, together with the original values of the variables that are involved in Ψ_q, satisfies all edits in Ψ_{q-1} except those in A.*

Proof (of Lemma 4.1). By construction, A contains all indices of failed edits from Ψ_{q-1} which do not involve x_g. Hence, the only way for the lemma to be false would be if there existed two edits that involve x_g, say ψ_{q-1}^s and ψ_{q-1}^t, with $s \notin A$ and $t \notin A$, so that it is not possible to find a value for x_g that satisfies both edits simultaneously. In this case, an implied edit in Ψ_q is generated by eliminating x_g from ψ_{q-1}^s and ψ_{q-1}^t. Moreover, by the fundamental property given at the end of Section 4.2.3, this implied edit must be failed by the original values of the other variables, i.e., the implied edit must be an element of $\Psi_q^{(2)}$. But this would contradict the assumption that A is a representing set of A_q^k for all $\psi_q^k \in \Psi_q^{(2)}$. Hence, it is impossible to find such a pair of edits, and the lemma follows. □

The proof of Theorem 4.1 now proceeds by induction on the number of treated variables q. For $q = 0$, the statement is trivial. For $q = 1$, the theorem follows as a special case of Lemma 4.1; note that $B_1^k \equiv A_1^k$. We suppose therefore that the statement has been proved for all $q \in \{0, 1, \ldots, Q-1\}$, and we consider the case $q = Q$, with $Q \geq 2$.

If Ψ_Q is obtained from Ψ_{Q-1} by fixing a variable to its original value, and B is a representing set of the sets B_Q^k for the failed edits from Ψ_Q, then by construction B is also a representing set of the sets B_{Q-1}^k for the failed edits from Ψ_{Q-1}. Thus, in this case, the statement for $q = Q$ follows immediately from the induction hypothesis.

Hence, we are left with the case that Ψ_Q is obtained from Ψ_{Q-1} by eliminating a variable, say x_g. We define, for each $\psi_Q^k \in \Psi_Q^{(2)}$, the index set A_Q^k of the edit(s) from Ψ_{Q-1} from which ψ_Q^k is derived, as above. Next, we use B to construct a set A, by applying the following procedure to each $\psi_Q^k \in \Psi_Q^{(2)}$:

- If ψ_Q^k is obtained by copying ψ_{Q-1}^l (so $A_Q^k = \{l\}$ and $B_Q^k = B_{Q-1}^l$), then we add l to A.

- If ψ_Q^k is obtained by eliminating x_g from ψ_{Q-1}^s and ψ_{Q-1}^t (so that $A_Q^k = \{s,t\}$ and $B_Q^k = B_{Q-1}^s \cup B_{Q-1}^t$), then we add s to A if B contains an element of B_{Q-1}^s, and we add t to A otherwise.

It is easy to see that this procedure produces a representing set A of the index sets A_Q^k for all $\psi_Q^k \in \Psi_Q^{(2)}$.

According to Lemma 4.1, there exists a value for x_g which, together with the original values of the $p-q$ variables that have not been treated, satisfies the edits in Ψ_{Q-1} except those in A. That is to say, Ψ_{Q-1} can be partitioned similarly to Ψ_Q as $\Psi_{Q-1} = \Psi_{Q-1}^{(1)} \cup \Psi_{Q-1}^{(2)}$, where $\Psi_{Q-1}^{(2)}$ contains the edits with indices in A. Moreover, it is not difficult to see that the above procedure implies that B is a representing set of the index sets B_{Q-1}^k for all $\psi_{Q-1}^k \in \Psi_{Q-1}^{(2)}$. Hence, the induction hypothesis establishes that, given the original values of the variables that have not been eliminated *and* given the chosen value for x_g, there exist values for the other eliminated variables that satisfy all the original edits except those in B. This shows that the statement holds for $q = Q$ and completes the proof of Theorem 4.1.

4.A.2 Proof of Theorem 4.2

To prove Theorem 4.2, we start again with an auxiliary lemma. Analogous to the numerical case, when Ψ_q is obtained from Ψ_{q-1} by eliminating v_g, we define the index set A_q^k of edits in Ψ_{q-1} from which the edit $\psi_q^k \in \Psi_q$ is derived. To be precise, we define $A_q^k := \{l\}$ if ψ_q^k is obtained by copying the edit ψ_{q-1}^l, and we define $A_q^k := T$ if ψ_q^k is obtained by eliminating a variable from the set of edits ψ_{q-1}^t ($t \in T$).

Lemma 4.2 *Consider the situation of Theorem 4.2 for $q \geq 1$, and suppose that v_g has been eliminated to obtain Ψ_q from Ψ_{q-1}. Let A be a representing set of the index sets A_q^k belonging to all $\psi_q^k \in \Psi_q^{(2)}$. Then there exists a value for v_g that, together with the original values of the variables that are involved in Ψ_q, satisfies all edits in Ψ_{q-1} except those in A.*

Proof (of Lemma 4.2). By construction, A contains all indices of failed edits from Ψ_{q-1} which do not involve v_g. Hence, the only way for the lemma to be false would be if there existed edits that involve v_g, say $\psi_{q-1}^{t_1}, \ldots, \psi_{q-1}^{t_r}$, with $A \cap \{t_1, \ldots, t_r\} = \emptyset$, so that it is not possible to find a value for v_g that satisfies these edits simultaneously, given the values of the other variables. Clearly, this could only happen if $F_g^{t_1} \cup \cdots \cup F_g^{t_r} = D_g$, since otherwise any value for v_g outside $F_g^{t_1} \cup \cdots \cup F_g^{t_r}$ would work. We may assume without loss of generality that $T' = \{t_1, \ldots, t_r\}$ is a minimal set having this property. Furthermore, it must

hold in this case that for all variables involved in Ψ_q, the original value of v_j is contained in all sets $F_j^{t_1}, \ldots, F_j^{t_r}$. In other words, T' must satisfy properties (4.10) and (4.11). This means that T' would generate an implied edit in Ψ_q which, by the fundamental property given at the end of Section 4.2.3, must be failed by the original values of the remaining variables. However, this would contradict the assumption that A is a representing set of A_q^k for all $\psi_q^k \in \Psi_q^{(2)}$. This completes the proof of Lemma 4.2. □

The proof of Theorem 4.2 is now completely analogous to that of Theorem 4.1, with Lemma 4.2 taking the role of Lemma 4.1.

Chapter 5

A Generalised Fellegi-Holt Paradigm for Automatic Error Localisation

The contents of this chapter have been published in *Survey Methodology* as Scholtus (2016). In that version, Appendix 5.B was omitted and the term "allowed edit operation" was used instead of "admissible edit operation". Otherwise, the chapter is identical to the article, apart from some minor textual corrections and adjustments.

5.1 Introduction

Data that have been collected for the production of statistics inevitably contain errors. A data editing process is needed to detect and amend these errors, at least in so far as they have an appreciable impact on the quality of the statistical output (Granquist and Kovar, 1997). Traditionally, data editing has been a manual task, ideally performed by professional editors with extensive subject-matter knowledge. To improve the efficiency, timeliness, and reproducibility of editing, many statistical institutes have attempted to automate parts of this process (Pannekoek et al., 2013). This has resulted in deductive correction methods for *systematic errors* and error localisation algorithms for *random errors* (De Waal et al., 2011, Chapter 1). In this chapter, I will focus on automatic editing for random errors.

Methods for this task usually proceed by minimally adjusting each record of data, according to some optimisation criterion, so that it becomes consistent with a given set of constraints known as *edit rules*, or *edits* for short. Depending on the effectiveness of the optimisation criterion and the strength of the edit rules, automatic editing may be used as a partial alternative to traditional manual editing. In practice, automatic editing is applied nearly always in combination with some

form of *selective editing*, which means that the most influential errors are treated manually (Hidiroglou and Berthelot, 1986; Granquist, 1995, 1997; Granquist and Kovar, 1997; Lawrence and McKenzie, 2000; Hedlin, 2003; De Waal et al., 2011).

Most automatic editing methods that are currently used in official statistics are based on the paradigm of Fellegi and Holt (1976): for each record, the smallest subset of variables is identified as erroneous that can be imputed so that the record becomes consistent with the edits. A slight generalisation is obtained by assigning so-called *confidence weights* to the variables and minimising the total weight of the imputed variables. Once this *error localisation problem* is solved, suitable new values have to be found in a separate step for the variables that were identified as erroneous. This is the so-called *consistent imputation problem*; see De Waal et al. (2011) and their references. In this chapter, I will focus on the error localisation problem.

At Statistics Netherlands, error localisation based on the Fellegi-Holt paradigm has been a part of the data editing process for Structural Business Statistics (SBS) for over a decade now. In evaluation studies, where the same SBS data were edited both automatically and manually, a number of systematic differences were found between the two editing efforts. Many of these differences could be explained by the fact that human editors performed certain types of adjustments that were suboptimal under the Fellegi-Holt paradigm. For instance, editors sometimes interchanged the values of associated costs and revenues items, or transferred parts of reported amounts between variables.

In practice, the outcome of manual editing is usually taken as the "gold standard" for assessing the quality of automatic editing. A critical evaluation of this assumption is beyond the scope of the present chapter; however, see EDIMBUS (2007, pp. 34–35). Here I simply note that, by improving the ability of automatic editing methods to mimic the results of manual editing, their usefulness in practice may be increased. In turn, this means that the share of automatic editing may be increased to improve the efficiency of the data editing process (Pannekoek et al., 2013).

To some extent, systematic differences between automatic and manual editing could be prevented by a clever choice of confidence weights. In general, however, the effects of a modification of the confidence weights on the results of automatic editing are difficult to predict. Moreover, if the editors apply a number of different complex adjustments, it might be impossible to model all of them under the Fellegi-Holt paradigm using a single set of confidence weights. Another option is to try to catch errors for which the Fellegi-Holt paradigm is known to provide an unsatisfactory solution at an earlier stage in the data editing process, i.e., during

deductive correction of systematic errors through automatic correction rules (De Waal et al., 2011; Scholtus, 2011a). This approach has practical limitations, however, because it may require a large collection of if-then rules, which would be difficult to design and maintain over time (Chen et al., 2003). Moreover, it is not self-evident that appropriate correction rules can be found for all errors that do not fit within the Fellegi-Holt paradigm.

In this chapter, a different approach is suggested. A new definition of the error localisation problem is proposed that allows for the possibility that errors affect more than one variable at a time. It is shown that this problem contains error localisation under the original Fellegi-Holt paradigm as a special case. Throughout this chapter, I restrict attention to numerical data and linear edits; a possible extension to categorical and mixed data will be discussed briefly in Section 5.8.

The remainder of this chapter is organised as follows. Section 5.2 briefly reviews relevant previous work done in this area. In Section 5.3, the concept of an edit operation is introduced and illustrated. The new error localisation problem is formulated in terms of these edit operations in Section 5.4. Section 5.5 generalises an existing method for identifying solutions to the Fellegi-Holt-based error localisation problem, and this result is used in Section 5.6 to outline a possible algorithm for solving the new problem. A small simulation study is discussed in Section 5.7. Finally, some conclusions and questions for further research follow in Section 5.8.

5.2 Background and related work

Let $\mathbf{x} = (x_1, \ldots, x_p)' \in \mathbb{R}^p$ be a record of p numerical variables. Suppose that this record has to satisfy k edit rules, in the form of the following system of linear (in)equalities:

$$\mathbf{A}\mathbf{x} + \mathbf{b} \odot \mathbf{0}, \tag{5.1}$$

where $\mathbf{A} = (a_{rj})$ is a $k \times p$-matrix of coefficients and $\mathbf{b} = (b_1, \ldots, b_k)'$ is a vector of constants. Here and elsewhere, $\mathbf{0}$ represents a vector of zeros of appropriate length; similarly, $\odot$ represents a symbolic vector of operators from the set $\{\geq, \leq, =\}$.

For a given record $\mathbf{x}$ that does not satisfy all edits in (5.1), the Fellegi-Holt-based error localisation problem amounts to finding the minimum of

$$\sum_{j=1}^{p} w_j \delta_j, \tag{5.2}$$

with $w_j > 0$ the confidence weight of variable x_j and $\delta_j \in \{0, 1\}$, under the

condition that the original record can be made consistent with the edits by imputing only those x_j with $\delta_j = 1$ (De Waal et al., 2011, p. 66).

Fellegi and Holt (1976) also proposed a method for solving the above error localisation problem, based on the generation of a sufficient set of so-called *implied edits* (see below). Unfortunately, the number of implied edits needed by this method is often extremely large in practice. Over the past decades, various dedicated algorithms for the error localisation problem have been developed by, among others, Schaffer (1987), Garfinkel et al. (1988), Kovar and Whitridge (1990), Ragsdale and McKeown (1996), De Waal (2003c), De Waal and Quere (2003), Riera-Ledesma and Salazar-González (2003, 2007), Bruni (2004), and De Jonge and Van der Loo (2014). Early algorithms mostly focused on strengthening the original method of Fellegi and Holt (1976) by reducing the number of required implied edits. More recent algorithms rely on the fact that the error localisation problem can be written as a mixed-integer programming problem, which makes it possible to apply standard optimisation techniques. See also De Waal and Coutinho (2005) or De Waal et al. (2011) for an overview and comparison of various error localisation algorithms.

Implied edits are constraints that follow logically from the original edits (5.1). In the present context (numerical data, linear edits), all relevant implied edits may be generated by a technique called *Fourier-Motzkin elimination* [FM elimination; cf. Williams (1986)]. FM elimination transforms a system of linear constraints having p variables into a system of implied linear constraints having at most $p - 1$ variables; thus, at least one of the original variables is eliminated. For mathematical details, see Appendix 5.A.

FM elimination has the following fundamental property: the system of implied constraints is satisfied by the values of the non-eliminated variables if, and only if, there exists a value for the eliminated variable that, together with the other values, satisfies the original system of constraints. In error localisation under the Fellegi-Holt paradigm, by repeatedly applying this fundamental property, one may verify whether any particular combination of variables can be imputed to obtain a consistent record, given the original values of the other variables. A clear illustration of this use of FM elimination is provided by the error localisation algorithm of De Waal and Quere (2003).

To conclude this section, it is interesting to look briefly at the statistical interpretation of the error localisation problem. In fact, in motivating their paradigm for automatic error localisation, Fellegi and Holt (1976) did not provide any formal statistical argument. Their reasoning was more intuitive:

> "The data in each record should be made to satisfy all edits by changing the

> fewest possible items of data (fields). This we believe to be in agreement with the idea of keeping the maximum amount of original data unchanged, subject to the constraints of the edits, and so manufacturing as little data as possible. At the same time, if errors are comparatively rare, it seems more likely that we will identify the truly erroneous fields." (Fellegi and Holt, 1976, p. 18)

A statistical argument for minimising the weighted number of imputed variables was provided by Liepins (1980) and Liepins et al. (1982), elaborating on earlier results of Naus et al. (1972). Suppose that errors occur according to a stochastic process, with each variable x_j being observed in error with a probability p_j that does not depend on its true value and with errors being independent across variables. Suppose furthermore that the confidence weights are defined as follows:

$$w_j = -\log\left(\frac{p_j}{1-p_j}\right). \tag{5.3}$$

Then it can be shown that minimising expression (5.2) is approximately equivalent to maximising the likelihood of the unobserved error-free record. Note that these authors tacitly assume that an error always affects one variable at a time.

Alternative error localisation procedures that are based more directly on statistical models have been proposed by, e.g., Little and Smith (1987) and Ghosh-Dastidar and Schafer (2006). These procedures use outlier detection techniques and require an explicit model for the true data. Unfortunately, they cannot handle edit rules such as (5.1) in a straightforward manner.

5.3 Edit operations

Continuing with the notation from Section 5.2, I define an *edit operation* g to be an affine function of the general form

$$g(\mathbf{x}) = \mathbf{T}\mathbf{x} + \mathbf{S}\boldsymbol{\alpha} + \mathbf{c}, \tag{5.4}$$

where $\mathbf{T}$ and $\mathbf{S}$ are known coefficient matrices of dimensions $p \times p$ and $p \times m$, respectively, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ is a vector of free parameters that may occur in g, and $\mathbf{c}$ is a p-vector of known constants. In the special case that g does not involve any free parameters ($m = 0$), the second term in (5.4) vanishes. Sometimes, it may be useful to impose one or several linear constraints on the free parameters in g:

$$\mathbf{R}\boldsymbol{\alpha} + \mathbf{d} \odot \mathbf{0}, \tag{5.5}$$

with $\mathbf{R}$ a known matrix, and $\mathbf{d}$ a known vector of constants. (Note: Matrix-vector notation will be used throughout this chapter because it leads to a concise description of results; however, using matrices to represent edits and edit operations is probably not the most efficient way to implement these results on a computer.)

As a first example, consider the operation that replaces one of the original values in $\mathbf{x}$ by an arbitrary new value (imputation). I will call this an *FH operation*, in view of its central role in automatic editing based on the Fellegi-Holt paradigm. Let $\mathbf{I}$ denote the $p \times p$ identity matrix and $\mathbf{e}_i$ the i^{th} standard basis vector in $\mathbb{R}^p$. The FH operation that imputes the variable x_j is given by (5.4) with $\mathbf{T} = \mathbf{I} - \mathbf{e}_j\mathbf{e}_j'$, $\mathbf{S} = \mathbf{e}_j$, and $\mathbf{c} = \mathbf{0}$. This yields: $g(\mathbf{x}) = \mathbf{x} + \mathbf{e}_j(\alpha - x_j) = (x_1, \ldots, x_{j-1}, \alpha, x_{j+1}, \ldots, x_p)'$, with $\alpha \in \mathbb{R}$ a free parameter that represents the imputed value. It should be noted that for a record of p variables, p distinct FH operations can be defined.

To further illustrate the concept of an edit operation, some other examples will now be given. For notational convenience, I restrict attention to the case $p = 3$.

- An edit operation that changes the sign of one of the variables:

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

- An edit operation that interchanges the values of two adjacent items:

$$g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_1 \\ x_3 \end{pmatrix}.$$

- An edit operation that transfers an amount between two items, where the amount transferred may equal at most K units in either direction:

$$\begin{aligned} g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \alpha + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} x_1 + \alpha \\ x_2 \\ x_3 - \alpha \end{pmatrix}, \end{aligned}$$

  with the constraint that $-K \leq \alpha \leq K$.

- An edit operation that imputes two variables simultaneously using a fixed ratio:

$$\begin{aligned} g\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ x_3 \end{pmatrix}, \end{aligned}$$

  with the constraint that $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)'$ satisfies $10\alpha_1 - \alpha_2 = 0$.

Intuitively, an edit operation is supposed to "reverse the effects" of a particular type of error that may have occurred in the observed data. That is to say, if the error associated with edit operation g actually occurred in the observed record $\mathbf{x}$, then $g(\mathbf{x})$ is the record that would have been observed if that error had not occurred. Somewhat more formally, it is assumed here that errors occurring in the data can be modelled by a stochastic "error generating process" $\mathcal{E}$, and that each edit operation acts as a "corrector" for one particular error that can occur under $\mathcal{E}$ (see Remark 4 in the next section).

If the edit operation g contains free parameters, the record $g(\mathbf{x})$ might not be determined uniquely even when the restrictions (5.1) and (5.5) are taken into account. In that case, one has to "impute" values for the free parameters that occur in an edit operation, which in turn means that some of the variables in $\mathbf{x}$ are imputed via the affine transformation given by (5.4). As in traditional Fellegi-Holt-based editing, finding appropriate "imputations" for the free parameters will not be considered part of the error localisation problem here. On the other hand, if g does not contain any free parameters, the imputed values in $g(\mathbf{x})$ follow directly from the edit operation itself and the distinction between error localisation and imputation is blurred.

In any particular application, only a small subset of potential edit operations of the form (5.4) would have a substantively meaningful interpretation, in the sense that the associated types of errors are known to occur. In what follows, I assume that a finite set of specific edit operations of the form (5.4) has been identified as relevant for a particular application. This will be called the set of *admissible edit operations* for that application. Some suggestions on how to construct this set will be given in Section 5.8.

5.4 A generalised error localisation problem

Let $\mathcal{G}$ be a finite set of admissible edit operations for a given application of automatic editing. Informally, I propose to generalise the error localisation problem of Fellegi and Holt (1976) by replacing "the smallest subset of variables that can be imputed to make the record consistent" with "the shortest sequence of admissible edit operations that can be applied to make the record consistent". To give a formal definition of this generalised error localisation problem, some new notation and concepts need to be introduced.

Consider a sequence of points $\mathbf{x} = \mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t = \mathbf{y}$ in $\mathbb{R}^p$. A *path* from $\mathbf{x}$ to $\mathbf{y}$ is defined as a sequence of *distinct* edit operations $g_1, \ldots, g_t \in \mathcal{G}$ such that $\mathbf{x}_n = g_n(\mathbf{x}_{n-1})$ for all $n \in \{1, \ldots, t\}$. (Note: In the case that g_n contains free

parameters, one should interpret this equality as "there exist feasible parameter values such that g_n maps $\mathbf{x}_{n-1}$ to $\mathbf{x}_n$".) A path is denoted by $P = [g_1, \ldots, g_t]$. The set of all possible paths from $\mathbf{x}$ to $\mathbf{y}$ is denoted by $\mathcal{P}(\mathbf{x}, \mathbf{y})$. This set may be empty. Later, I will use $\mathcal{P}(\mathbf{x}; G)$ to denote, for a given subset $G \subseteq \mathcal{G}$, the set of all paths starting in $\mathbf{x}$ that consist of the edit operations in G in some order (without specifying the free parameters); if G contains t elements, $\mathcal{P}(\mathbf{x}; G)$ contains $t!$ paths.

To each edit operation $g \in \mathcal{G}$, one can associate a weight $w_g > 0$ that expresses the costs of applying edit operation g. In particular, the weight of an FH operation is to be chosen equal to the confidence weight of the variable that it imputes. Now the *length* of a path $P = [g_1, \ldots, g_t]$ can be defined as the sum of the weights of its constituent edit operations: $\ell(P) = \sum_{n=1}^{t} w_{g_n}$, where, by convention, the empty path has length zero. The *distance* from $\mathbf{x}$ to $\mathbf{y}$ is defined as the length of the shortest path that connects $\mathbf{x}$ to $\mathbf{y}$:

$$d(\mathbf{x}, \mathbf{y}) = \begin{cases} \min \{\ell(P) \mid P \in \mathcal{P}(\mathbf{x}, \mathbf{y})\} & \text{if } \mathcal{P}(\mathbf{x}, \mathbf{y}) \neq \emptyset, \\ \infty & \text{otherwise.} \end{cases}$$

In general, $d(\mathbf{x}, \mathbf{y})$ satisfies the standard axioms of a metric *except* that it need not be symmetric in $\mathbf{x}$ and $\mathbf{y}$; it is a so-called *quasimetric* (Scholtus, 2014a). Accordingly, $d(\mathbf{x}, \mathbf{y})$ represents "the distance from $\mathbf{x}$ to $\mathbf{y}$" rather than "the distance between $\mathbf{x}$ and $\mathbf{y}$".

The distance from $\mathbf{x}$ to any closed, non-empty subset $D \subseteq \mathbb{R}^p$ is defined as the distance to the nearest $\mathbf{y} \in D$: $d(\mathbf{x}, D) = \min \{d(\mathbf{x}, \mathbf{y}) \mid \mathbf{y} \in D\}$. For the purpose of error localisation, the closed, non-empty subset of $\mathbb{R}^p$ that is of particular interest is the set D_0 of all points that satisfy (5.1).

I can now formulate the generalised error localisation problem.

Problem 5.1 *Consider a given set of consistent records D_0, a given set of admissible edit operations $\mathcal{G}$, and a given record $\mathbf{x}$. If $d(\mathbf{x}, D_0) = \infty$, then the error localisation problem for $\mathbf{x}$ is infeasible. Otherwise, any shortest path leading to a record $\mathbf{y} \in D_0$ such that $d(\mathbf{x}, \mathbf{y}) < \infty$ is called a* feasible solution *to the error localisation problem for $\mathbf{x}$. A feasible solution is called* optimal *if it leads to a record $\mathbf{x}^* \in D_0$ such that*

$$d(\mathbf{x}, \mathbf{x}^*) = d(\mathbf{x}, D_0). \tag{5.6}$$

Formally, then, the generalised error localisation problem consists of finding an optimal path of edit operations.

Remark 1. In general, there may be infinitely many records $\mathbf{x}^*$ in D_0 that satisfy (5.6) and can be reached by the same path of edit operations. To solve the error

localisation problem, it is sufficient to find an optimal path. Constructing an associated record $\mathbf{x}^* \in D_0$ may then be regarded as a generalisation of the consistent imputation problem; cf. the discussion on imputation at the end of Section 5.3. □

Remark 2. Problem 5.1 is infeasible for records that cannot be mapped onto D_0 by any combination of distinct edit operations in $\mathcal{G}$. To avoid this situation, $\mathcal{G}$ should be chosen sufficiently large so that $d(\mathbf{x}, D_0) < \infty$ for all $\mathbf{x} \in \mathbb{R}^p$. In what follows, I tacitly assume that $\mathcal{G}$ has this property. An easy way – not necessarily the only way – to achieve this is by letting $\mathcal{G}$ contain at least all FH operations. That this is sufficient follows from the fact that any two points in $\mathbb{R}^p$ are connected by a path that concatenates the FH operations associated with the coordinates on which they differ. □

Remark 3. It is not difficult to see that Problem 5.1 reduces to the original error localisation problem of Fellegi and Holt (1976) in the special case that $\mathcal{G}$ contains only the FH operations. □

Remark 4. As with the original Fellegi-Holt-based error localisation problem, it can be shown that, under certain assumptions, minimising $d(\mathbf{x}, \mathbf{y})$ over all $\mathbf{y} \in D_0$ for a given observed record $\mathbf{x}$ is approximately equivalent to maximising the likelihood of the associated unobserved error-free record. The argument closely follows that of Kruskal (1983, pp. 38–39) for the so-called Levenshtein distance in the context of approximate string matching. This requires first of all that the edits (5.1) be hard edits, i.e., failed only by erroneous values. In addition, it must be assumed that the stochastic "error generating process" $\mathcal{E}$ introduced in Section 5.3 has the following properties:

- There exists a one-to-one correspondence between the set of errors that can occur under $\mathcal{E}$ and the set of admissible edit operations $\mathcal{G}$ that correct them.
- The errors in $\mathcal{E}$ occur independently of each other.
- The error corresponding to operation g occurs with known probability p_g.

Finally, analogous to (5.3), the weights w_g should be chosen according to

$$w_g = -\log\left(\frac{p_g}{1 - p_g}\right). \tag{5.7}$$

Under these assumptions, Scholtus (2014a) adapted the argument of Kruskal (1983) to show that the optimal solution to error localisation problem (5.6) can be justified as an approximate maximum likelihood estimator. [Note: The derivation in Scholtus (2014a) assumed in addition that all $p_g \ll 1$, in which case $w_g \approx -\log p_g$. This assumption is unnecessary; cf. Liepins (1980).] □

5.5 Implied edits for general edit operations

In this section, a result will be derived that establishes whether a given path of edit operations of the form (5.4) can be used to make a given record consistent with a given system of edit rules (i.e., is a feasible solution to the error localisation problem). This result uses the FM elimination technique discussed in Section 5.2.

Let $\mathbf{x}$ be a given record and let $\mathbf{y}_t$ be any record that can be obtained by applying, in sequence, the edit operations $g_1, \ldots, g_t$ to $\mathbf{x}$:

$$\mathbf{y}_t = g_t \circ g_{t-1} \circ \cdots \circ g_1(\mathbf{x}). \tag{5.8}$$

Write $g_n(\mathbf{x}) = \mathbf{T}_n\mathbf{x} + \mathbf{S}_n\boldsymbol{\alpha}_n + \mathbf{c}_n$, for $n \in \{1, \ldots, t\}$. From (5.8) it follows by induction that

$$\begin{aligned}\mathbf{y}_1 &= \mathbf{T}_1\mathbf{x} + \mathbf{S}_1\boldsymbol{\alpha}_1 + \mathbf{c}_1,\\ \mathbf{y}_2 &= \mathbf{T}_2\mathbf{T}_1\mathbf{x} + \mathbf{S}_2\boldsymbol{\alpha}_2 + \mathbf{c}_2 + \mathbf{T}_2\left(\mathbf{S}_1\boldsymbol{\alpha}_1 + \mathbf{c}_1\right),\end{aligned}$$

and, in general,

$$\mathbf{y}_t = \mathbf{T}_t \cdots \mathbf{T}_1\mathbf{x} + \mathbf{S}_t\boldsymbol{\alpha}_t + \mathbf{c}_t + \sum_{n=2}^{t} \mathbf{T}_t \cdots \mathbf{T}_n\left(\mathbf{S}_{n-1}\boldsymbol{\alpha}_{n-1} + \mathbf{c}_{n-1}\right), \tag{5.9}$$

where the sum over n is defined to be zero when $t = 1$. Moreover, all terms involving $\mathbf{S}_n\boldsymbol{\alpha}_n$ vanish in these expressions when g_n does not contain any free parameters.

The path of edit operations $P = [g_1, \ldots, g_t]$ can be applied to $\mathbf{x}$ to obtain a record that is consistent with the edits (5.1) if, and only if, there exists a $\mathbf{y}_t$ of the form (5.9) that satisfies $\mathbf{A}\mathbf{y}_t + \mathbf{b} \odot \mathbf{0}$ and all relevant additional restrictions of the form (5.5) on $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$. Using (5.9), $\mathbf{A}\mathbf{y}_t + \mathbf{b} \odot \mathbf{0}$ can be written as:

$$\left(\mathbf{A}\mathbf{T}_t \cdots \mathbf{T}_1\right)\mathbf{x} + \left(\mathbf{A}\mathbf{S}_t\right)\boldsymbol{\alpha}_t + \sum_{n=2}^{t}\left(\mathbf{A}\mathbf{T}_t \cdots \mathbf{T}_n\mathbf{S}_{n-1}\right)\boldsymbol{\alpha}_{n-1} + \mathbf{b}_t \odot \mathbf{0}, \tag{5.10}$$

with $\mathbf{b}_t = \mathbf{b} + \mathbf{A}\mathbf{c}_t + \sum_{n=2}^{t} \mathbf{A}\mathbf{T}_t \cdots \mathbf{T}_n\mathbf{c}_{n-1}$ a vector of constants.

Interestingly, (5.10) and the possible additional restrictions of the form (5.5) constitute a linear system of the form (5.1) on the extended record $(\mathbf{x}', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_t')'$. Therefore, FM elimination may be used to remove all free parameters from this system. This yields a system of implied restrictions for $\mathbf{x}$. Moreover, a repeated application of the fundamental property of FM elimination establishes that $\mathbf{x}$ satisfies this system of implied edits if, and only if, there exist parameter values for $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$ that, together with $\mathbf{x}$, satisfy (5.10) and (5.5). Hence, it follows that a

path of edit operations $P = [g_1, \ldots, g_t]$ can lead to a consistent record for $\mathbf{x}$ if, and only if, $\mathbf{x}$ satisfies the system of implied edits obtained by eliminating $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$ from (5.10) and (if relevant) additional restrictions of the form (5.5).

Example. Consider the following edits in x_1 and x_2:

$$x_1 \geq 0, \tag{5.11}$$
$$x_2 \geq 0, \tag{5.12}$$
$$x_1 + x_2 \leq 5. \tag{5.13}$$

Let g be the edit operation that transfers an amount of at most four units between x_1 and x_2, in either direction: $g((x_1, x_2)') = (x_1 + \alpha, x_2 - \alpha)'$ with $-4 \leq \alpha \leq 4$. For this single edit operation, the system of transformed edits (5.10) is:

$$x_1 + \alpha \geq 0, \tag{5.14}$$
$$x_2 - \alpha \geq 0, \tag{5.15}$$
$$x_1 + x_2 \leq 5. \tag{5.16}$$

I also add the following restrictions of the form (5.5) on α:

$$\alpha \geq -4, \tag{5.17}$$
$$\alpha \leq 4. \tag{5.18}$$

This yields five linear constraints (5.14)–(5.18) on x_1, x_2, and α, from which α may be removed by FM elimination to obtain:

$$x_1 \geq -4, \tag{5.19}$$
$$x_2 \geq -4, \tag{5.20}$$
$$x_1 + x_2 \geq 0, \tag{5.21}$$
$$x_1 + x_2 \leq 5. \tag{5.22}$$

According to the theory, any record $(x_1, x_2)'$ that satisfies (5.19)–(5.22) can be made consistent with the original edits (5.11)–(5.13) by transferring an amount of α units (with $-4 \leq \alpha \leq 4$) between x_1 and x_2. The example record $(x_1, x_2)' = (-2, 3)'$ is inconsistent with the original edits (5.11)–(5.13) but satisfies (5.19)–(5.22). This implies that the record can be made consistent with the original edits by applying g. It is easy to see that this is true; any choice $2 \leq \alpha \leq 3$ will do. □

It is interesting to note that, for the special case that P consists of the single FH operation that imputes x_j, the transformed system of edits (5.10) is obtained by

replacing every occurrence of x_j in the original edits by an unrestricted parameter α. Eliminating α from (5.10) is equivalent in this case to eliminating x_j directly from the original edits. In this sense, the above result generalises the fundamental property of FM elimination for FH operations to all edit operations of the form (5.4).

In general, the set of records defined by expression (5.9) depends on the way the edit operations are ordered. Thus, two paths consisting of the same set of edit operations in a different order need not yield the same solution to the error localisation problem. In this respect, general edit operations differ from FH operations (Scholtus, 2014a).

5.6 An error localisation algorithm

In this section, I propose a relatively simple algorithm to solve the error localisation problem of Section 5.4, using the theoretical result from the previous section.

In practical applications of error localisation in official statistics, it is not unusual to have records of over 100 variables. To obtain a problem that is computationally feasible, existing applications of automatic editing based on the Fellegi-Holt paradigm usually specify an upper bound M on the number of variables that may be imputed in a single record (e.g., $M = 12$ or $M = 15$). De Waal and Coutinho (2005) argued that the introduction of such an upper bound is reasonable because a record that requires more than, say, fifteen imputations should be considered unfit for automatic editing anyway. Following this tradition, one can also introduce an upper bound R on the number of distinct edit operations that may be applied to a single record. Even with this additional restriction, the search space of potential solutions to Problem 5.1 will usually be too large in practice to find the optimal solution by an exhaustive search.

Figure 5.1 summarises the proposed error localisation algorithm. Its basic set-up was inspired by the *apriori algorithm* of Agrawal and Srikant (1994) for data mining. Upon completion, the algorithm returns a set $\mathcal{L}$ containing all paths of admissible edit operations that correspond to an optimal solution to Problem 5.1, as well as the optimal path length W. [Note: An error localisation problem may have multiple optimal solutions, and it may be beneficial to find all of them (Giles, 1988; De Waal et al., 2011, pp. 66–67).]

After initialisation in step 0, the algorithm cycles through steps 1, 2, and 3 at most R times. In step 1 of the algorithm, the search space is limited by using the following fact: if G has a proper subset $H \subset G$ for which $\mathcal{P}(\mathbf{x}; H)$ contains a path that leads to a consistent record, then $\mathcal{P}(\mathbf{x}; G)$ can contain only suboptimal

Figure 5.1: An algorithm that finds all optimal paths of edit operations for Problem 5.1.

Step 0. Let $\mathbf{x}$ be a given record and $\mathcal{G}$ a given set of admissible edit operations. Initialise: $\mathcal{L} := \emptyset$; $\mathcal{B}_0 := \{\emptyset\}$; $W := \infty$; and $t := 1$.

Step 1. Determine all subsets $G \subseteq \mathcal{G}$ of cardinality t that satisfy these conditions:

1. Every subset of $t-1$ elements in G is part of $\mathcal{B}_{t-1}$.
2. It holds that $\sum_{g \in G} w_g \leq W$.

Step 2. For each G found in step 1, construct $\mathcal{P}(\mathbf{x}; G)$ and, for each path $P \in \mathcal{P}(\mathbf{x}; G)$, evaluate whether it can lead to a consistent record. If so, then:

- if $\ell(P) < W$, define $\mathcal{L} := \{P\}$ and $W := \ell(P)$;
- if $\ell(P) = W$, define $\mathcal{L} := \mathcal{L} \cup \{P\}$.

If *none* of the paths $P \in \mathcal{P}(\mathbf{x}; G)$ leads to a consistent record, add G to $\mathcal{B}_t$.

Step 3. If $t < R$ and $\mathcal{B}_t \neq \emptyset$, define $t := t + 1$ and return to step 1.

solutions. Thus, any set G that has such a subset may be ignored by the algorithm. Similarly, G may also be ignored whenever the total weight of the edit operations in G exceeds the path length of the best feasible solution found so far.

During the t^{th} iteration, the number of subsets G encountered in step 1 of the algorithm equals $\binom{|\mathcal{G}|}{t}$. For each of these subsets, the conditions in step 1 have to be checked. If a subset passes these checks, in step 2 all $t!$ paths in $\mathcal{P}(\mathbf{x}; G)$ are evaluated using the theory of Section 5.5. The idea behind the apriori algorithm is that, as t becomes larger, the majority of subsets will not pass the checks in the first step, so that the total amount of computational work remains limited. In the context of data mining, this desirable behaviour has indeed been observed in practice. Whether it also occurs in the context of error localisation remains to be seen.

One possible improvement to the algorithm can be made by observing that the order in which edit operations are applied does not matter in all cases. Sometimes two paths in $\mathcal{P}(\mathbf{x}; G)$ are *equivalent* in the sense that any record that can be reached from $\mathbf{x}$ by the first path can also be reached by the second path, and vice versa. This property defines an equivalence relation on $\mathcal{P}(\mathbf{x}; G)$. Let $\tilde{\mathcal{P}}(\mathbf{x}; G)$ be

a set that contains one representative from each equivalence class of $\mathcal{P}(\mathbf{x}; G)$ under this relation. Clearly, the algorithm in Figure 5.1 remains correct if in step 2 the search is limited to $\tilde{\mathcal{P}}(\mathbf{x}; G)$ instead of $\mathcal{P}(\mathbf{x}; G)$. Scholtus (2014a) provides a simple method for constructing $\tilde{\mathcal{P}}(\mathbf{x}; G)$ from $\mathcal{P}(\mathbf{x}; G)$.

A detailed example illustrating the above algorithm is given in Appendix 5.B.

5.7 Simulation study

To test the potential usefulness of the new error localisation approach, I conducted a small simulation study, using the R environment for statistical computing (R Development Core Team, 2017). A prototype implementation was created in R of the algorithm in Figure 5.1. This prototype made liberal use of the existing functionality for Fellegi-Holt-based automatic editing available in the `editrules` package (Van der Loo and De Jonge, 2012; De Jonge and Van der Loo, 2014). The program was not optimised for computational efficiency, but it turned out to work sufficiently fast for the relatively small error localisation problems encountered in this simulation study. (Note: The R code used in this study is available from the author upon request.)

The simulation study involved records of five numerical variables that should satisfy the following nine linear edit rules:

$$\begin{aligned} x_1 + x_2 &= x_3, \\ x_3 - x_4 &= x_5, \\ x_j &\geq 0, \qquad j \in \{1, 2, 3, 4\}, \\ x_1 &\geq x_2, \\ x_5 &\geq -0.1x_3, \\ x_5 &\leq 0.5x_3. \end{aligned}$$

Edits of this form might typically be encountered for SBS, as part of a much larger set of edit rules (Scholtus, 2014a).

I created a random error-free data set of $2,000$ records by drawing from a multivariate normal distribution (using the `mvtnorm` package) with the following parameters:

$$\boldsymbol{\mu} = \begin{pmatrix} 500 \\ 250 \\ 750 \\ 600 \\ 150 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 10,000 & -1,250 & 8,750 & 7,500 & 1,250 \\ -1,250 & 5,000 & 3,750 & 4,000 & -250 \\ 8,750 & 3,750 & 12,500 & 11,500 & 1,000 \\ 7,500 & 4,000 & 11,500 & 11,750 & -250 \\ 1,250 & -250 & 1,000 & -250 & 1,250 \end{pmatrix}.$$

Only records that satisfied all of the above edits were added to the data set. Note that $\boldsymbol{\Sigma}$ is a singular covariance matrix that incorporates the two equality edits. Technically, the resulting data follow a so-called truncated multivariate singular normal distribution; see De Waal et al. (2011, pp. 318ff) or Tempelman (2007).

Table 5.1: Admissible edit operations for the simulation study.

name	operation	associated type of error	p_g	w_g
FH1	impute x_1	erroneous value of x_1	0.10	2.20
FH2	impute x_2	erroneous value of x_2	0.08	2.44
FH3	impute x_3	erroneous value of x_3	0.06	2.75
FH4	impute x_4	erroneous value of x_4	0.04	3.18
FH5	impute x_5	erroneous value of x_5	0.02	3.89
IC34	interchange x_3 and x_4	true values of x_3 and x_4 interchanged	0.07	2.59
TF21	transfer an amount from x_2 to x_1	part of the true value of x_1 reported as part of x_2	0.09	2.31
CS4	change the sign of x_4	sign error in x_4	0.11	2.09
CS5	change the sign of x_5	sign error in x_5	0.13	1.90

Table 5.1 lists the nine admissible edit operations that were considered in this study. Note that the first five lines contain the FH operations for this data set. As indicated in the table, each edit operation has an associated type of error. A synthetic data set to be edited was created by randomly adding errors of these types to the above-mentioned error-free data set. The probability of each type of error is listed in the fourth column of Table 5.1. The associated "ideal" weight according to (5.7) is shown in the last column.

To limit the amount of computational work, I only considered records that required three edit operations or less. Records without errors were also removed. This left $1,025$ records to be edited, each containing one, two, or three of the errors listed in Table 5.1.

Several error localisation approaches were applied to this data set. First of all, I tested error localisation according to the Fellegi-Holt paradigm (i.e., using only the edit operations FH1–FH5) and according to the new paradigm (i.e., using all edit operations in Table 5.1). Both approaches were tested once using the "ideal"

weights listed in Table 5.1 and once with all weights equal to 1 ("no weights"). The latter case simulates a situation where the relevant edit operations would be known, but not their respective frequencies. Finally, to test the robustness of the new error localisation approach to a lack of information about relevant edit operations, I also applied this approach with one of the non-FH operations in Table 5.1 missing from the set of admissible edit operations.

The quality of error localisation was evaluated in two ways. Firstly, I evaluated how well the optimal paths of edit operations found by the algorithm matched the true distribution of errors, using the following contingency table for all $1,025 \times 9 = 9,225$ combinations of records and edit operations:

	edit operation was suggested	edit operation was not suggested
associated error occurred	TP	FN
associated error did not occur	FP	TN

From this table, I computed indicators that measure the proportion of false negatives, false positives, and overall wrong decisions, respectively:

$$\alpha = \frac{FN}{TP + FN}; \qquad \beta = \frac{FP}{FP + TN}; \qquad \delta = \frac{FN + FP}{TP + FN + FP + TN}.$$

Similar indicators are discussed by De Waal et al. (2011, pp. 410–411). I also computed $\bar{\rho} = 1 - \rho$, with ρ the fraction of records in the data set for which the error localisation algorithm found exactly the right solution. A good error localisation algorithm should have low scores on all four indicators.

It should be noted that the above quality indicators put the original Fellegi-Holt approach at a disadvantage, as this approach does not use all the edit operations listed in Table 5.1. Therefore, I also calculated a second set of quality indicators α, β, δ, and $\bar{\rho}$ that look at erroneous values rather than edit operations. In this case, α measures the proportion of values in the data set that were affected by errors but left unchanged by the optimal solution of the error localisation problem, and similarly for the other measures.

Table 5.2 displays the results of the simulation study for both sets of quality indicators. In both cases, a considerable improvement in the quality of the error localisation results is seen for the approach that used all edit operations, compared to the approach that used only FH operations. In addition, leaving one relevant edit operation out of the set of admissible edit operations had a negative effect on the quality of error localisation. In some cases this effect was quite large – particularly in terms of edit operations used –, but the results of the new error localisation approach still remained substantially better than those of the Fellegi-Holt approach.

Table 5.2: Quality of error localisation in terms of edit operations used and identified erroneous values; computing time required.

	quality indicators (edit operations)				quality indicators (erroneous values)				
approach	α	β	δ	$\bar{\rho}$	α	β	δ	$\bar{\rho}$	time*
Fellegi-Holt (weights)	74%	12%	23%	80%	19%	10%	13%	32%	46
Fellegi-Holt (no weights)	70%	12%	21%	74%	13%	8%	9%	24%	33
all operations (weights)	14%	3%	5%	24%	10%	5%	7%	17%	98
except IC34	29%	5%	9%	35%	15%	9%	11%	29%	113
except TF21	34%	5%	10%	37%	10%	5%	7%	18%	80
except CS4	28%	6%	9%	39%	10%	5%	7%	17%	80
except CS5	35%	7%	10%	47%	11%	6%	7%	18%	82
all operations (no weights)	27%	5%	8%	36%	6%	4%	5%	13%	99

* Total computing time (in seconds) on a laptop PC with a 2.5 GHz CPU under Windows 7.

Contrary to expectation, not using different confidence weights actually improved the quality of the error localisation results somewhat for this data set under the Fellegi-Holt approach (both sets of indicators) and to some extent also under the new approach (only the second set of indicators). Finally, it is seen that using all edit operations led to an increase in computing time compared to using only FH operations, but this increase was not dramatic.

5.8 Conclusion

In this chapter, a new formulation was proposed of the error localisation problem in automatic editing. It was suggested to find the (weighted) minimal number of edit operations needed to make an observed record consistent with the edits. The new error localisation problem can be seen as a generalisation of the problem proposed in a seminal paper by Fellegi and Holt (1976), because the operation that imputes a new value for one variable at a time is an important special case of an edit operation.

The main focus here has been on developing the mathematical theory behind the new error localisation problem. It turns out that FM elimination, a technique that has been used in the past to solve the Fellegi-Holt-based error localisation problem, can be applied also in the context of the new problem (Section 5.5). Nevertheless, the task of solving the new error localisation problem is challenging from a computational point of view, at least for the numbers of variables, edits, and edit operations that would be encountered in practical applications at statistical institutes. A possible error localisation algorithm was outlined in Section 5.6. More efficient algorithms probably could and should be developed. Similarly to FM elimination, it may be possible to adapt other ideas that have been used to solve the Fellegi-Holt-based problem to the generalised problem considered here.

The discussion in this chapter was restricted to numerical data and linear edits. The original Fellegi-Holt paradigm has been applied also to categorical and mixed data. Several authors, including Bruni (2004) and De Jonge and Van der Loo (2014), have shown that a large class of edits for mixed data can be re-formulated in terms of numerical data and linear edits, with the additional restriction that some of the variables have to be integer-valued. In principle, this means that the results in this chapter could be applied also to mixed data. To accommodate the fact that some variables are integer-valued, Pugh (1992)'s extension of FM elimination to integers could be used; see also De Waal et al. (2011) for a discussion of this extended elimination technique in the context of Fellegi-Holt-based error localisation. It remains to be seen whether this approach is computationally feasible.

Remark 4 in Section 5.4 hinted at an analogy between error localisation in sta-

tistical microdata and the field of approximate string matching. In approximate string matching, text strings are compared under the assumption that they may have been partially corrupted (Navarro, 2001). Various distance functions have been proposed for this task. The Hamming distance, which counts the number of positions on which two strings differ, may be seen as an analogue of the Fellegi-Holt-based target function (5.2). The generalised error localisation problem defined in this chapter has its counterpart in the use of the Levenshtein distance or "edit distance" for approximate string matching. It may be interesting to explore this analogy further. In particular, efficient algorithms have been developed for computing edit distances between strings; it might be possible to apply some of the underlying ideas also to the generalised error localisation problem.

The new error localisation algorithm was applied successfully to a small synthetic data set (Section 5.7). Overall, the results of this simulation study suggest that the new error localisation approach has the potential to achieve a substantial improvement of the quality of automatic editing compared to the approach that is currently used in practice. However, this does require that sufficient information be available to identify all – or at least most – of the relevant edit operations in a particular application. Possible gains in the quality of error localisation also have to be weighed in practice against the higher computational demands of the generalised error localisation problem.

An obvious candidate for applying the new methodology in practice would be the SBS. However, more research is needed before this method could be applied during regular production. To apply the method in a particular context, it is necessary first to specify the relevant edit operations. Ideally, each edit operation should correspond to a combination of amendments to the data that human editors consider to be a correction for one particular error. In addition, a suitable set of weights w_g has to be determined for these edit operations. This would require information about the relative frequencies of the most common types of amendments made during manual editing. Both aspects could be investigated based on historical data before and after manual editing, editing instructions and other documentation used by the editors, and interviews with editors and/or supervisors of editing.

On a more fundamental level, a question of demarcation arises between deductive correction methods and automatic editing under the new error localisation problem. In principle, many known types of error could be resolved either by automatic correction rules or by error localisation using edit operations. Each approach has its own advantages and disadvantages (Scholtus, 2014a). It is likely that some compromise will produce the best results, with some errors handled deductively and others by edit operations. However, it is not obvious how best to make this

division in practice.

Ultimately, the aim of the new methodology proposed in this chapter is to improve the usefulness of automatic editing in practice. So far, the results are promising.

Appendix 5.A Fourier-Motzkin elimination

Consider a system of linear constraints (5.1) and let x_f be the variable to be eliminated. First, suppose that x_f is involved only in inequalities. For ease of exposition, suppose that the edits are normalised so that all inequalities use the $\geq$ operator. The FM elimination method considers all pairs (r, s) of inequalities in which the coefficients of x_f have opposite signs; that is, $a_{rf}a_{sf} < 0$. Suppose without loss of generality that $a_{rf} < 0$ and $a_{sf} > 0$. From the original pair of edits, the following implied constraint is derived:

$$\sum_{j=1}^{p} a_j^* x_j + b^* \geq 0, \tag{5.23}$$

with $a_j^* = a_{sf}a_{rj} - a_{rf}a_{sj}$ and $b^* = a_{sf}b_r - a_{rf}b_s$. Note that $a_f^* = 0$, so x_f is not involved in (5.23). An inequality of the form (5.23) is derived from each of the above-mentioned pairs (r, s). The full implied system of constraints obtained by FM elimination now consists of these derived constraints, together with all original constraints that do not involve x_f.

If there are linear equalities that involve x_f, the above technique could be applied after replacing each linear equality by two equivalent linear inequalities. De Waal and Quere (2003) suggested a more efficient alternative for this case. Suppose that the r^{th} constraint in (5.1) is an equality that involves x_f. This constraint can be rewritten as

$$x_f = \frac{-1}{a_{rf}} \left(b_r + \sum_{j \neq f} a_{rj} x_j \right). \tag{5.24}$$

By substituting the expression on the right-hand-side of (5.24) for x_f in all other constraints, one again obtains an implied system of constraints that does not involve x_f and that can be rewritten in the form (5.1).

For a proof that FM elimination has the fundamental property mentioned in Section 5.2, see, e.g., De Waal et al. (2011, pp. 69–70).

Appendix 5.B A small example

As an illustration of the algorithm of Section 5.6, consider the following small-scale example. Suppose that the following linear edits are defined:

$$x_1 + x_3 = 19, \tag{5.25}$$
$$x_1 \geq 4, \tag{5.26}$$
$$x_1 \leq 7, \tag{5.27}$$
$$x_3 - x_1 \geq 5, \tag{5.28}$$
$$x_3 - x_1 \leq 10, \tag{5.29}$$
$$x_3 \geq 0. \tag{5.30}$$

Note: The two numerical variables are denoted by x_1 and x_3 here to be consistent with the notation in Scholtus (2014a), where a more elaborate version of this example is described that also includes the variable x_2.

The record $(x_1, x_3) = (10, -3)$ requires editing as it fails edits (5.25), (5.27), (5.28), and (5.30). Suppose that the following edit operations of the form (5.4) are admissible:

- the FH operation FH1 that imputes variable x_1;
- the FH operation FH3 that imputes variable x_3;
- an edit operation CS1 that changes the sign of x_1;
- an edit operation TF13 that transfers an amount of at most $K = 15$ units between x_1 and x_3 (in either direction).

Representations of these edit operations in matrix-vector notation can be derived from the examples given in Section 5.3. Suppose in addition that the weights of the admissible edit operations are chosen as follows:

edit operation	FH1	FH3	CS1	TF13
weight	1	3	0.5	1

The algorithm in Figure 5.1 can be applied to find the optimal solution to Problem 5.1 for this record. In step 0 of the algorithm, $\mathcal{G}$ is defined as the set that contains FH1, FH3, CS1, and TF13, $\mathcal{L}$ is defined to be the empty set, and W is initialised at ∞.

In the first iteration ($t = 1$), the algorithm considers subsets of one edit operation from $\mathcal{G}$. There are four such subsets and they all satisfy the two conditions of step 1. (Note that the first condition is irrelevant when $t = 1$.) For each of the

paths [FH1], [FH3], [CS1], and [TF13], the theory of Section 5.5 can be used to check whether it leads to a consistent record with respect to (5.25)–(5.30) when applied to the original record $(x_1, x_3) = (10, -3)$.

As mentioned at the end of Section 5.5, for paths that contain only FH operations, the procedure defined in that section is equivalent to applying FM elimination directly to the variables associated with these operations. Thus, to check whether the path [FH1] is a feasible solution, one should remove x_1 from the edits (5.25)–(5.30) by FM elimination. After some simplification, this yields the following implied constraints for x_3:

$$x_3 \geq 12, \tag{5.31}$$

$$2x_3 \leq 29. \tag{5.32}$$

Upon substituting the original value $x_3 = -3$, it is seen that (5.31) is failed. Hence, the path [FH1] is not a feasible solution to the error localisation problem. Similarly, by eliminating x_3 from the original edits (5.25)–(5.30) and substituting $x_1 = 10$, it is found that the path [FH3] is not a feasible solution either.

For the path [CS1], the theory from Section 5.5 could be applied. However, edit operation CS1 (which just changes the sign of x_1) is clearly weaker than FH1 (which can impute any value for x_1). Since it was already seen that [FH1] is not a feasible solution to the error localisation problem, it is clear that [CS1] cannot lead to a consistent record in this example.

For the path [TF13], applying expression (5.10) (with $t = 1$) to the edits (5.25)–(5.30) yields a system that includes one free parameter:

$$\begin{aligned} x_1 + x_3 &= 19, \\ x_1 + \alpha &\geq 4, \\ x_1 + \alpha &\leq 7, \\ x_3 - x_1 - 2\alpha &\geq 5, \\ x_3 - x_1 - 2\alpha &\leq 10, \\ x_3 - \alpha &\geq 0, \\ \alpha &\geq -15, \\ \alpha &\leq 15. \end{aligned}$$

Note that the last two inequalities are constraints of the form (5.5) that follow from the definition of TF13. By eliminating α from this system, an implied system of edits is found for x_1 and x_3. It turns out that the original values $x_1 = 10$ and $x_3 = -3$ do not satisfy this implied system, so the single edit operation TF13 does

not lead to a consistent record either. This conclusion also follows directly from the first constraint above, since $x_1 + x_3 \neq 19$ independently of α.

In summary, applying single edit operations from $\mathcal{G}$ does not yield a feasible solution in this example. At the end of the first iteration, it holds that $\mathcal{B}_1 = \{\{\text{FH1}\}, \{\text{FH3}\}, \{\text{CS1}\}, \{\text{TF13}\}\}$, $\mathcal{L} = \emptyset$, and $W = \infty$.

In the next iteration, $t = 2$. There are $\binom{4}{2} = 6$ distinct subsets of two edit operations in this example:

$$\{\text{FH1}, \text{FH3}\}, \{\text{FH1}, \text{CS1}\}, \{\text{FH1}, \text{TF13}\},$$
$$\{\text{FH3}, \text{CS1}\}, \{\text{FH3}, \text{TF13}\}, \{\text{CS1}, \text{TF13}\}.$$

Since no feasible solutions were found in the first iteration, all of these subsets satisfy the first condition of step 1. They also satisfy the second condition. For each of these subsets G, the set $\mathcal{P}((10, -3)'; G)$ contains two ordered paths. However, it can be shown that for the first five subsets listed above, the two ordered paths are equivalent as defined at the end of Section 5.6, so they do not have to be considered separately; see Scholtus (2014a) for more details. The only pair of non-equivalent paths is found to be $[\text{CS1}, \text{TF13}]$ and $[\text{TF13}, \text{CS1}]$. Thus, in total, seven paths (out of a potential twelve) need to be evaluated in this iteration. For the sake of brevity, I discuss only two of these evaluations in detail.

For the path $[\text{FH1}, \text{FH3}]$, which uses only FH operations, the checking procedure is equivalent to eliminating x_1 and x_3 from the original edits (5.25)–(5.30). Eliminating x_1 yields (5.31) and (5.32), as seen above. Eliminating x_3 from this pair of implied edits yields the trivially true statement $0 \geq -5$. The fact that this statement is true implies, by the fundamental property of FM elimination, that there exists a value for x_3 that satisfies the edits (5.31) and (5.32). (It is not difficult to see that this is indeed the case.) By another application of the fundamental property, this in turn implies that there exist values for x_1 and x_3 that satisfy the edits (5.25)–(5.30). Hence, it follows that imputing x_1 and x_3 (i.e., the path $[\text{FH1}, \text{FH3}]$) is a feasible solution to the error localisation problem here. The path length associated with this solution is 4. If the original Fellegi-Holt paradigm were used, this would be the optimal solution to the error localisaton problem.

Now consider the path $[\text{FH1}, \text{TF13}]$. To see whether this path leads to a feasible solution, I apply the transformation (5.10) (with $t = 2$) to the edits (5.25)–(5.30), with $\mathbf{T}_1$, $\mathbf{S}_1$ and $\mathbf{c}_1$ coming from the definition of FH1, and $\mathbf{T}_2$, $\mathbf{S}_2$ and $\mathbf{c}_2$ coming from the definition of TF13. This yields the following system of constraints, with the parameter α_1 introduced by FH1 and the parameter α_2 introduced by TF13.

(Again, the last two restrictions are added from the definition of TF13.)

$$x_3 + \alpha_1 = 19, \tag{5.33}$$
$$\alpha_1 + \alpha_2 \geq 4, \tag{5.34}$$
$$\alpha_1 + \alpha_2 \leq 7, \tag{5.35}$$
$$x_3 - \alpha_1 - 2\alpha_2 \geq 5, \tag{5.36}$$
$$x_3 - \alpha_1 - 2\alpha_2 \leq 10, \tag{5.37}$$
$$x_3 - \alpha_2 \geq 0, \tag{5.38}$$
$$\alpha_2 \geq -15, \tag{5.39}$$
$$\alpha_2 \leq 15. \tag{5.40}$$

The two parameters have to be eliminated from (5.33)–(5.40). Elimination of α_1 yields the following non-redundant edits:

$$\begin{aligned} x_3 - \alpha_2 &\geq 12, \\ 2x_3 - 2\alpha_2 &\leq 29, \\ \alpha_2 &\geq -15, \\ \alpha_2 &\leq 15. \end{aligned}$$

Elimination of α_2 from this system yields, upon simplification:

$$\begin{aligned} x_3 &\geq -3, \\ 2x_3 &\leq 59. \end{aligned}$$

It is seen that the original value $x_3 = -3$ satisfies this system of implied edits. Therefore, I conclude that a consistent record can be obtained by applying the edit operations FH1 and TF13 to $(x_1, x_3) = (10, -3)$. The associated path length is 2, which is an improvement compared to the solution found previously.

The other five paths mentioned above may be handled similarly. The remaining paths yield just one additional feasible solution, given by the path $[\text{FH3}, \text{TF13}]$. This solution has a path length of 4. At the end of the second iteration, it holds that: $\mathcal{B}_2 = \{\{\text{FH1}, \text{CS1}\}, \{\text{FH3}, \text{CS1}\}, \{\text{CS1}, \text{TF13}\}\}$, $\mathcal{L} = \{[\text{FH1}, \text{TF13}]\}$, and $W = 2$.

In the next iteration ($t = 3$), the algorithm is stopped, because none of the subsets of three edit operations from $\mathcal{G}$ satisfies the first criterion of step 1. Thus, the optimal solution returned by the algorithm is: "impute x_1 and transfer an amount between x_1 and x_3". The corresponding distance $d((10, -3)', D_0)$ equals 2. In addition, two suboptimal feasible solutions were found by the algorithm, both with

path length 4: "impute x_1 and impute x_3" and "impute x_3 and transfer an amount between x_1 and x_3". Note that, for all three solutions, the order in which the two edit operations are applied does not matter.

It is interesting to work out which consistent records can be reached from the original record by these feasible paths of edit operations. I start with the (suboptimal) solution that uses both FH operations, because this is the easiest case. Denote the imputed values under this solution by x_1^* and x_3^* and suppose that $x_1^* = 7 - \beta$ for some, as yet unspecified, parameter β. By equation (5.25), $x_3^* = 12 + \beta$. Furthermore, by (5.31) and (5.32), it has to hold that $0 \leq \beta \leq 5/2$. In summary, the potential consistent records that can be reached by imputing both x_1 and x_3 in this example are $(x_1^*, x_3^*) = (7 - \beta, 12 + \beta)$, for $0 \leq \beta \leq 5/2$.

For the optimal solution "impute x_1 and transfer an amount between x_1 and x_3", the amended record has the form $(x_1^*, x_3^*) = (\alpha_1 + \alpha_2, -3 - \alpha_2)$, where the parameters (α_1, α_2) have to satisfy the system of restrictions found by substituting $x_3 = -3$ in (5.33)–(5.40). From (5.33), it follows immediately that $\alpha_1 = 22$. The remaining restrictions for α_2 are satisfied only when $\alpha_2 = -15$. Thus, for this solution, the edits determine unique feasible values for the parameters. The unique amended record is found to be $(x_1^*, x_3^*) = (7, 12)$. This is a special case of the solution found previously, with $\beta = 0$. Hence, the optimal solution "impute x_1 and transfer an amount between x_1 and x_3" is found to be more restrictive than "impute x_1 and impute x_3". In a similar way, it can be shown that the other feasible (but suboptimal) solution found above – "impute x_3 and transfer an amount between x_1 and x_3" – is *not* more restrictive than "impute x_1 and impute x_3"; i.e., every record of the above form with $0 \leq \beta \leq 5/2$ can be reached using these edit operations.

Figure 5.2 summarises the results for this example in graphical form. The boundary of the region defined by each edit from (5.25)–(5.30) is plotted as a solid line in the (x_1, x_3) plane. The feasible region defined jointly by these edits is shown as the bold line segment AB, with $A = (9/2, 29/2)$ and $B = (7, 12)$; note that AB contains precisely all points of the form $(x_1, x_3) = (7 - \beta, 12 + \beta)$ with $0 \leq \beta \leq 5/2$. The original record $(x_1, x_3) = (10, -3)$ is plotted as point C. This point does not lie on the line segment AB, since it corresponds to an inconsistent record.

As was derived algebraically above, the Fellegi-Holt-based solution "impute x_1 and impute x_3" can be used to reach any point E on the line segment AB (in fact, any point in $\mathbb{R}^2$) from C, by varying the imputed values. One potential path, shown in Figure 5.2, consists of the line segment CD (i.e., an imputation for x_3) followed by DE (i.e., an imputation for x_1). The suboptimal solution "impute x_3 and transfer an amount between x_1 and x_3" also reaches any point on AB; a

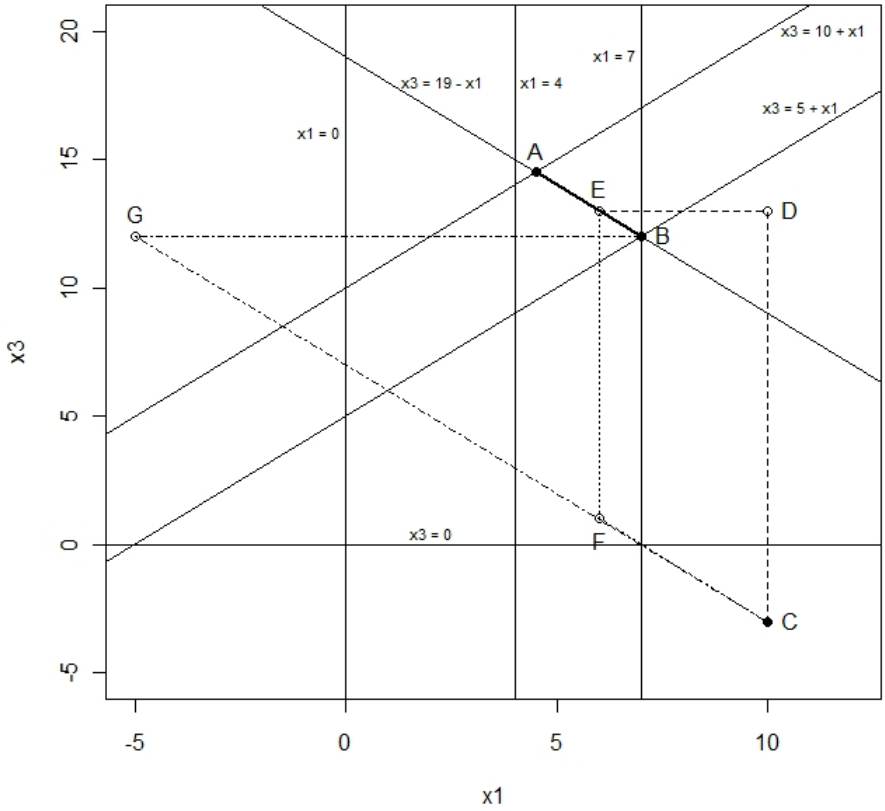


Figure 5.2: Illustration of three feasible solutions in the (x_1, x_3) plane.

potential path is shown as CF (i.e., a transferred amount from x_1 to x_3) followed by FE (an imputation for x_3). The *optimal* solution "impute x_1 and transfer an amount between x_1 and x_3" reaches only the point B. The corresponding path is displayed as CG (a transferred amount from x_1 to x_3; note that the maximal allowed amount $K = 15$ is needed here) followed by GB (an imputation for x_1).

In terms of distances, it holds that $d(C, B) = 2$ and $d(C, E) = 4$ for all points $E \neq B$ on AB. Apparently, for the weights chosen in this example, it is considered better to adjust C towards B than towards any other point on AB.

Chapter 6

Estimating the Validity and Bias of Administrative and Survey Variables

This chapter was co-authored by Bart F. M. Bakker (Statistics Netherlands, VU University) and Arnout van Delden (Statistics Netherlands). Author contributions: all authors contributed ideas; Van Delden and Scholtus set up the data for the application; Scholtus developed the mathematical part of the work, carried out the analysis and wrote the report; Bakker and Van Delden edited the report. An earlier version of this chapter was published by Statistics Netherlands as Scholtus et al. (2015). The contents of this chapter have been submitted for publication.

6.1 Introduction

In recent years, the use of administrative data has grown in official statistics as well as in academic research (Bethlehem, 2008; Bakker and Daas, 2012). Governmental organisations, such as tax authorities, social security offices, and municipalities, collect data on a large number of social and economic phenomena as part of their regular activities. In many countries, national statistical institutes (NSIs) and other producers of official statistics have access to these administrative data sources. Many NSIs are looking at ways to use administrative data to reduce – and ideally replace – their own data collection by means of questionnaires. Reasons for this include tighter budgets and a decreasing willingness of persons and businesses to participate in surveys. Administrative data also offer possibilities for more detailed statistical analyses than surveys based on relatively small samples.

When examining the suitability of a given administrative source for statistical purposes, several questions need to be addressed (Bakker, 2011b; Zhang, 2012). In this chapter, we will focus on issues related to the quality of measurement. In

general, all data sources may contain errors. In the case of administrative data, a particular source of error arises from potential differences between the variable that is measured for administrative purposes and the variable that is needed for statistical purposes.

To give an example, European NSIs are supposed to publish short-term statistics on Turnover as defined in the short-term statistics regulation (European Commission, 2006). The tax authorities in the Netherlands also collect information on Turnover from businesses to levy value-added tax (VAT). Conceptually, these two Turnover variables are not the same for all businesses; for instance, some economic activities are included in 'statistical' Turnover but exempt from taxes. (We will return to this example in the application below.) It is, therefore, important to assess the measurement quality of administrative variables for statistical use (Bakker and Daas, 2012; Groen, 2012).

In the context of questionnaire design, there is a well-established tradition of using linear structural equation models (SEMs) to assess the measurement quality of survey variables; key references include Andrews (1984); Saris and Andrews (1991); Scherpenzeel and Saris (1997); Saris and Gallhofer (2007); Alwin (2007). The models used in this approach can be seen as an extension of the classical test theory from psychology as set out by Lord and Novick (1968) and Jöreskog (1971). Each observed variable is modelled as an imperfect measure of an underlying latent (unobserved) variable. To quantify the measurement quality of an observed variable, one can estimate its *validity* which, under the simplest model, is defined as its standardised factor loading on the underlying latent variable (see Section 6.2). These models are usually identified by taking repeated measurements on each target variable, which requires a carefully-planned research design. It should also be noted that SEMs require variables that are measured on an interval scale or higher. For nominal and ordinal variables, latent class models are more appropriate (Biemer, 2011).

Applying the same modelling approach to administrative data is not straightforward. As administrative data are collected by an external party, it is usually not possible to conduct methodological experiments. Bakker (2012) suggested that repeated measurements may be obtained by linking an administrative data set to data from an independent sample survey. This is useful in particular for examining whether questions in an existing survey can be replaced by corresponding administrative variables, at least in terms of validity. Similarly, Pavlopoulos and Vermunt (2015) and Oberski (2017) have used latent class models to compare the amount of classification error in categorical administrative and survey variables. An important advantage of approaches that use latent variables is that they do not

assume that either the administrative or the survey data are error-free. In fact, it is not necessary to know in advance which source provides the measurement with the highest validity: this is estimated from the data.

While validity captures the correlation of an observed variable to the underlying concept, producers of official statistics are often interested in population means or totals. Therefore, in addition to the validity, it may be important to know whether any substantial measurement bias occurs in the levels of individual variables (so-called *intercept bias*). The main objective of the present chapter is to extend the approach using SEMs to also assess the bias of an administrative variable. To illustrate, we describe an application at Statistics Netherlands to assess the validity and intercept bias of VAT Turnover for short-term statistics.

The remainder of this chapter is organised as follows. Section 6.2 describes the proposed methodology for estimating the validity and intercept bias of observed variables. The above-mentioned application to VAT Turnover is discussed in Section 6.3. Section 6.4 closes the chapter with a discussion of the possibilities and limitations of the proposed method. While the main focus of this chapter is on evaluating the measurement quality of administrative data, some potential other applications are also outlined in Section 6.4.

6.2 Methodology

6.2.1 Assessing validity and intercept bias using SEMs

Let $y_1, \ldots, y_p$ denote a set of observed variables that may be affected by measurement errors, and let $\eta_1, \ldots, \eta_m$ denote the underlying variables of interest that are error-free and not observed directly. The relationship between each observed and unobserved variable, as well as the relations that exist among the unobserved variables, may be described by an SEM.

For our purposes here, an SEM may be defined as a system of linear regression equations:

$$\eta_j = \alpha_j + \sum_{j' \neq j} \beta_{jj'} \eta_{j'} + \zeta_j, \quad (j = 1, \ldots, m), \tag{6.1}$$

$$y_k = \tau_k + \lambda_k \eta_{j_k} + \epsilon_k, \quad (k = 1, \ldots, p). \tag{6.2}$$

Equations of the form (6.1) are *structural equations* relating the unobserved variables to each other: the coefficient $\beta_{jj'}$ represents a direct effect of $\eta_{j'}$ on η_j, ζ_j represents a zero-mean disturbance term, and α_j represents a structural intercept. Equations of the form (6.2) are *measurement equations* relating an observed y_k to an unobserved η_{j_k} in terms of a factor loading λ_k, a measurement intercept τ_k, and

a zero-mean measurement error ϵ_k that is uncorrelated with η_{j_k}. Observed variables act as indicators for latent variables. Note that latent variables often have more than one indicator. By contrast, we restrict attention in this chapter to SEMs in which each observed variable loads on exactly one latent variable. More general SEMs that do not have this restriction are discussed, e.g., by Bollen (1989).

The SEM given by (6.1)–(6.2) contains the following parameters: α_j, $\beta_{jj'}$, τ_k, λ_k, $\psi_{jj'} = \text{cov}(\zeta_j, \zeta_{j'})$, and $\theta_{kl} = \text{cov}(\epsilon_k, \epsilon_l)$. It is standard practice to restrict some of these to zero a priori, based on substantive considerations. Provided that the model is identified, the unknown parameters can be estimated from the observed variance-covariance matrix and the observed vector of means of $y_1, \ldots, y_p$; see Section 6.2.2. The absolute value of the standardised factor loading

$$|\lambda_k^s| \equiv |\lambda_k| \frac{\text{sd}(\eta_{j_k})}{\text{sd}(y_k)} = \sqrt{1 - \frac{\text{var}(\epsilon_k)}{\text{var}(y_k)}} \tag{6.3}$$

may be used as a measure of the validity[1] of y_k (Bakker, 2012). The intercept bias of y_k may be evaluated in terms of the parameters τ_k and λ_k. Having estimated the model, we can derive formulae to correct each observed variable to the scale of the corresponding error-free variable; see Section 6.2.4 for more details.

By linking administrative data to survey data, one will usually obtain at most two indicators per latent variable (Scholtus and Bakker, 2013a). The smallest SEM that is then identified has $m = 2$ correlated latent variables with two indicators each. If covariates are available that are considered to be measured (essentially) without error, these can also be included in the model to obtain identification. In addition, identification of any SEM with latent variables requires that each latent variable be given a scale and, if the model contains intercept terms, that the origins of these scales be fixed as well. When one is interested only in the validity, identification may be achieved by standardising each latent variable to have mean 0 and variance 1. However, this is not an option if the intercept bias is to be evaluated. In fact, none of the standard SEM identification procedures [see Little et al. (2006) for an overview] is then suitable because, as argued by Bielby (1986a), these procedures define an 'arbitrary' metric for the latent variables.

A procedure for achieving model identification in a 'non-arbitrary' way was suggested by Sobel and Arminger (1986) and discussed in the present context by Scholtus (2014b). The basic idea is to collect additional 'gold standard' data on each latent variable for a random subsample of the original data set. Many of

[1] In the terminology of Saris and Andrews (1991), $|\lambda_k^s|$ measures the *indicator validity* of y_k. This is actually the product of its 'pure' validity and its reliability as defined by Saris and Andrews (1991). Biemer (2011) uses the terms *empirical* and *theoretical validity* instead of indicator validity and 'pure' validity, respectively. This point is taken up in Section 6.4.

the variables encountered in official statistics are factual (e.g., Age, Educational attainment, Turnover, Number of employees), so that it is theoretically possible to obtain the true score for each unit. In practice, it is usually prohibitively expensive or otherwise inconvenient to do so for the entire population or even for a sizeable sample. But it may often be feasible to obtain 'gold standard' data for a small subsample of units. Provided that this *audit sample* is obtained by randomised selection from the original data set, we can use it to assign a 'non-arbitrary' metric to the latent variables, thereby identifying the SEM. We can still use the entire data set to estimate the model parameters in terms of this metric.

Figure 6.1 shows an example of a path diagram of an SEM that is identified in this way, having $m = 3$ latent variables with two indicators each (outside the audit sample). The task of estimating this model can be cast as a missing-data problem that may be solved by fitting a two-group SEM; see Section 6.2.3. Results on simulated data in Scholtus (2014b) suggested that a relatively small audit sample of 50 units may often be sufficient.

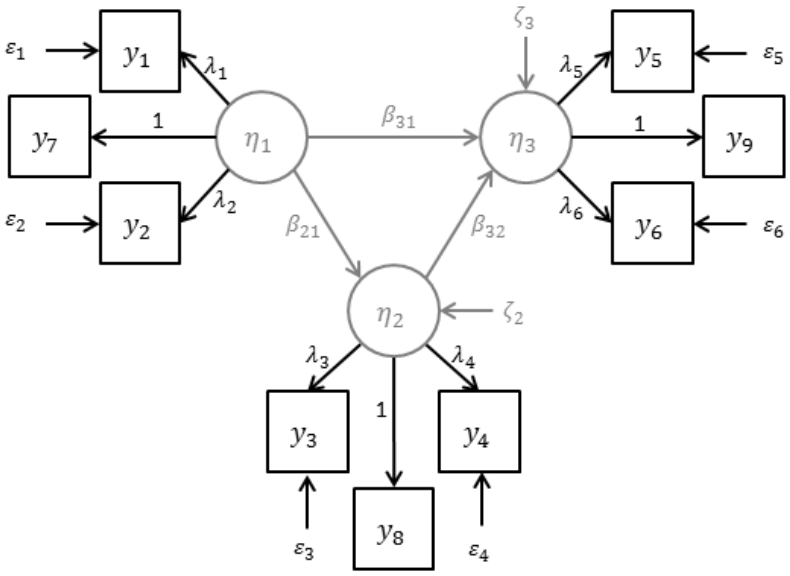
In practice, the 'gold standard' data could be obtained by some form of re-editing by subject-matter experts, as was done in a different context by Nordbotten (1955). In Figure 6.1 and throughout this chapter, it is assumed that the audit data do not contain any measurement errors: in (6.2) for these variables, $\tau = 0$, $\lambda = 1$, and $\text{var}(\epsilon) = 0$. In fact, the model can be identified by the audit sample also when $\text{var}(\epsilon) \neq 0$ but the other two assumptions do hold. In that case, the 'gold standard' data are supposed to contain only measurement errors that do not affect the scale of measurement. While this assumption is theoretically weaker than the assumption of no errors, it is not necessarily more plausible in practice. When the 'gold standard' data are obtained by re-editing, it actually seems less plausible from a practical point of view.

6.2.2 Estimating an SEM

Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and $\boldsymbol{\Sigma} = (\sigma_{kl})$ denote, respectively, the population mean vector and population variance-covariance matrix of the observed variables in the SEM. That is, $\mu_k = E(y_k)$ and $\sigma_{kl} = \text{cov}(y_k, y_l)$. Under the model given by (6.1)–(6.2), these moments are expressible in terms of the unknown model parameters: $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\vartheta})$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\vartheta})$, with $\boldsymbol{\vartheta}$ a vector containing all distinct parameters (Bollen, 1989).

For a given sample of size n, let $\bar{\mathbf{y}}$ and $\mathbf{S}$ denote the empirical means and covariances of $y_1, \ldots, y_p$. A conventional way to estimate $\boldsymbol{\vartheta}$ is by minimising a certain distance function F_{ML} between $(\bar{\mathbf{y}}, \mathbf{S})$ and $(\boldsymbol{\mu}(\boldsymbol{\vartheta}), \boldsymbol{\Sigma}(\boldsymbol{\vartheta}))$, which leads to maximum likelihood (ML) estimation if the sample consists of independent, identi-

(group 1: audited units)



(group 2: other units)

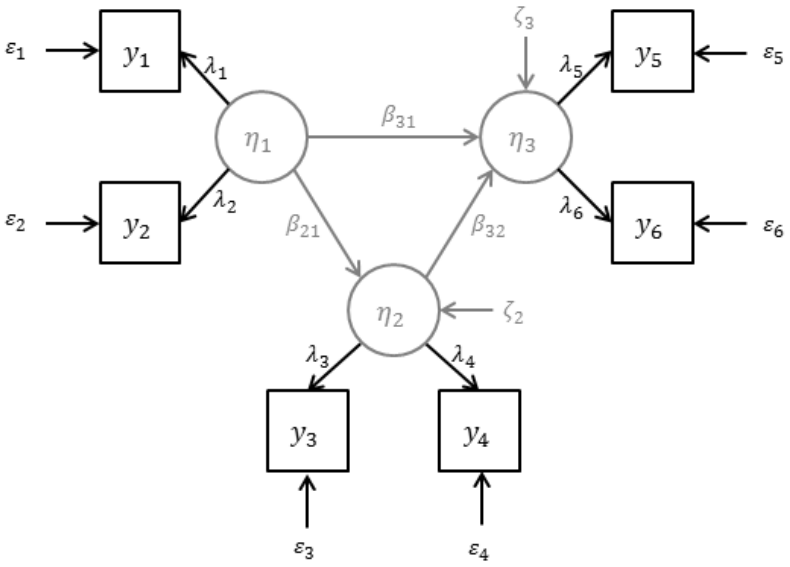


Figure 6.1: Example of a two-group SEM identified by means of an audit sample. The model for the first group contains additional error-free variables that are observed only in the audit sample. The structural part of the model is shown in grey.

cally distributed (i.i.d.) observations from a multivariate normal distribution. This method of estimation also produces asymptotic standard errors for the estimated parameters, as well as a test statistic that can be used as a measure of overall fit: under the above assumption of normality, $X^2_{\mathrm{ML}} = (n-1)F_{\mathrm{ML}}$ should follow a chi-square distribution with known degrees of freedom if the model holds. More details are given in Appendix 6.A.1 and in Bollen (1989).

In many practical applications, including the VAT application to be discussed in Section 6.3, the assumption of having i.i.d. observations from a normal distribution is not satisfied. Firstly, the data may come from a different (unknown) distribution. In this situation, it can be shown that minimising F_{ML} still produces a consistent point estimator for $\boldsymbol{\vartheta}$ under mild conditions, but the estimated standard errors are typically incorrect and the above test statistic need not follow a chi-square distribution. It is known how to obtain asymptotically correct standard errors (Satorra, 1992; Muthén and Satorra, 1995); see also Appendix 6.A.1. A correction to the chi-square test statistic was proposed by Satorra and Bentler (1986, 1994). The resulting corrected statistic is denoted by X^2_{SB} here. The terms Robust Maximum Likelihood and Pseudo Maximum Likelihood (PML) are used to refer to this estimation strategy when the data are not normally distributed.

Secondly, the above assumption is violated when the sample is obtained by some complex survey design, possibly involving:

- without-replacement sampling from a finite population;
- stratification;
- clustering; and/or
- multi-stage selection.

In this case, to obtain a consistent point estimator, one should use design-consistent estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in place of $\bar{\mathbf{y}}$ and $\mathbf{S}$. After this adjustment, essentially the same results apply as in the i.i.d. case with non-normal data (Muthén and Satorra, 1995). Thus, the same PML approach may be used to obtain corrected standard errors and test statistics. Some more details are given in Appendix 6.A.1. In the application to be discussed below, we used PML estimation to account for both phenomena: non-normal data and finite-population sampling.

Many software packages are available for estimating SEMs, including LISREL, EQS, and Mplus. For the analyses in this chapter, we made use of two packages from the R environment for statistical computing (R Development Core Team, 2017): the package `lavaan` (Rosseel, 2012) which contains the basic

functionality for estimating a variety of latent variable models and the package `lavaan.survey` (Oberski, 2014) which implements the PML approach for SEM estimation with complex samples.

6.2.3 Incorporating the audit sample

In theory, the estimation of an SEM that is identified by means of an audit sample is straightforward. Consider the example in Figure 6.1. We set up a two-group SEM, where the first group contains the n_1 observations from the audit sample and its model is defined in terms of the observed variables $y_1, \ldots, y_9$, while the second group contains the $n - n_1$ remaining observations and its model is defined in terms of $y_1, \ldots, y_6$ alone. In the first group, the model is identified by assuming that $y_7 = \eta_1$, $y_8 = \eta_2$, and $y_9 = \eta_3$. The model for the second group is identified by restricting all parameters of the overlapping part of the model to be equal in both groups; this makes sense if the audit sample is a random subsample of the original data.[2]

In practice, some complications arise because all standard software packages that can estimate multi-group SEMs require that the same set of observed variables be used in each group. Thus, in the example of Figure 6.1 we need to account for the missing data for y_7, y_8, y_9 in the second group.

Allison (1987) proposed a general-purpose method for estimating SEMs with missing data, which provides ML estimates provided that the data are i.i.d. multivariate normal and the missing values are Missing At Random (MAR) in the terminology of Little and Rubin (2002). In the context considered here, the data are missing by design (i.e., the design of the audit sample) and we can ensure that the MAR condition is satisfied. Baumgartner and Steenkamp (1998) described an extension of Allison's method that is usable in combination with the PML approach of Section 6.2.2, so that the condition of multivariate normality can be dropped. This method involves imputing random values from a normal distribution for the missing variables in the second group, in such a way that the observed means of these variables are identically zero, the observed variances are identically one, and the observed covariances with all other variables are zero. In the model for the second group, the measurement equations $y_7 = \epsilon_7$, $y_8 = \epsilon_8$, and $y_9 = \epsilon_9$ are in-

[2] To identify the model, it is in fact sufficient to restrict only the parameters of the *measurement* model to be invariant across groups; the *structural* parameters could be estimated freely. In practice, the model fit will improve if these parameters are left free across groups. However, this may be seen as overfitting if the audit sample is truly a random subsample of the original data, as all differences between the groups should then be due to sampling fluctuations. Having a separate structural model in each group would also complicate the derivation of a correction formula for the observed variables (see Section 6.2.4). We therefore restrict all overlapping parameters to be invariant.

cluded. In addition, we fix $\theta_{77} = \theta_{88} = \theta_{99} = 1$ for this group. Basically, this ensures that the observed means and covariances for y_7, y_8, y_9 in the second group are exactly reproduced while the estimation of the rest of the model is not affected by the imputed values. Because some of the observed moments are now fixed by design, some care must be taken in defining the correct degrees of freedom for the model. See Appendix 6.A.2 for more details.

A more generally applicable way to deal with missing values in an SEM is by multiple imputation (Oberski, 2014). We did not explore this option here.

6.2.4 Deriving a correction formula

Having estimated the SEM given by (6.1)–(6.2), we obtain for each observed variable y_k an estimate of the validity $|\hat{\lambda}^s_k|$ from (6.3) and an estimated regression line for y_k given η_{j_k}:

$$\hat{y}_k = \hat{\tau}_k + \hat{\lambda}_k \eta_{j_k}. \tag{6.4}$$

For notational simplicity, we drop the indices k and j_k in the remainder of this section. In broad terms, we can distinguish between three cases:

(a) the validity of y is high and $(\hat{\tau}, \hat{\lambda}) \approx (0, 1)$;

(b) the validity of y is high but $(\hat{\tau}, \hat{\lambda})$ differs significantly from $(0, 1)$;

(c) the validity of y is low.

With case (a), the observed values are strongly correlated to the true values and there is no indication of measurement bias. Observed variables that fall under case (c) apparently contain too much measurement error to be of use. In the remainder of this section, we will focus on case (b), where there is a strong correlation between the observed and true values but the observed values are systematically too high or too low. Suppose we would like to correct this measurement bias. Formula (6.4), which predicts the value of y for a given value of η, cannot be used directly for this purpose. Rather, we need a formula that predicts η, given the observed values.

From the literature, it is known how to predict the true scores of the latent variables in an SEM from the observed ones; see, e.g., formula (6) in Meijer et al. (2012). This predictor is unbiased but it involves a linear combination of (in general) all the observed variables from the original model. In the present context, these variables have most likely been obtained specifically for a methodological evaluation study, e.g., by linking data from different sources, and they will typically not all be available during regular statistical production. Consider the extreme case

that only y itself is available. By using the linear regression model $\eta = a + by + \omega$ and expression (6.2), it may be derived that the best predictor (in a least-squares sense) of η for an arbitrary given value $y = y_0$ is:

$$\eta_0 = \mu_\eta + \lambda \frac{\sigma_\eta^2}{\sigma_y^2}(y_0 - \mu_y); \tag{6.5}$$

see Scholtus et al. (2015) for details. The unknown parameters in (6.5) can all be expressed as simple functions of $\boldsymbol{\vartheta}$ (Bollen, 1989). Thus, having estimated the original SEM, we can use the following formula for predicting η given $y = y_0$:

$$\hat{\eta}_0 = \hat{\mu}_\eta + \hat{\lambda} \frac{\hat{\sigma}_\eta^2}{\hat{\sigma}_y^2}(y_0 - \hat{\mu}_y), \tag{6.6}$$

with $\hat{\mu}_\eta = \mu_\eta(\hat{\boldsymbol{\vartheta}})$, $\hat{\sigma}_\eta^2 = \sigma_\eta^2(\hat{\boldsymbol{\vartheta}})$, etc. Furthermore, since $a = \mu_\eta - \lambda(\sigma_\eta^2/\sigma_y^2)\mu_y$ and $b = \lambda(\sigma_\eta^2/\sigma_y^2)$ are differentiable functions of $\boldsymbol{\vartheta}$, approximate standard errors for the estimated intercept and slope in (6.6) can be obtained by linearisation; `lavaan` and most other modern SEM software packages provide this option.

A similar formula to (6.6) can be derived for predicting η from any given subset of the observed variables in the original SEM, by considering the multiple regression of η on those variables. Such a formula may be useful in practice if several (but not necessarily all) observed variables from the SEM are available during regular production, for instance because they come from the same data source.

Two further remarks are in order. Firstly, it should be noted that solving (6.4) for η directly yields $\tilde{\eta}_0 = (y_0 - \hat{\tau})/\hat{\lambda}$, which is *not* equivalent to (6.6). This approach is invalid in general because it ignores the fact that ϵ and y are correlated. However, $\hat{\eta}_0$ does converge to $\tilde{\eta}_0$ as the validity of y approaches 1 (Scholtus et al., 2015).

Secondly, in the context of a repeated survey (where the same set of statistics is produced on a regular basis), it is desirable to use the same instance of formula (6.6) to correct observations on y for measurement error in multiple survey rounds, without the need to re-estimate the correction every time. Clearly, this requires that the measurement model remains stable over time. In fact, the parameters a and b also depend on μ_η and σ_η^2 and could therefore change as the structural part of the model evolves over time, even when the measurement model remains stable. However, it can be shown that this effect is negligible in practice provided that the validity of y is high enough and the structural parameters evolve gradually over time; see Scholtus et al. (2015). Of course, the measurement model cannot be expected to remain stable indefinitely. Therefore, it will be necessary to update formula (6.6) by conducting a new audit sample at regular intervals and/or whenever

changes are made to the data collection process that may affect the measurement parameters of y. In the case of administrative data, an NSI should monitor actively whether such changes are being made by the administrative authority.

6.3 Application: Using VAT Turnover for the Netherlands' quarterly short-term statistics

6.3.1 Introduction

From 2011 onwards, Statistics Netherlands has been publishing quarterly short-term statistics (STS) on Turnover that are based on a combination of VAT data for small to medium-sized businesses and a census survey for the largest and/or most complex units (Van Delden and De Wolf, 2013). The VAT data are obtained from tax declarations submitted to the tax authorities. The primary output of STS consists of estimated growth rates of Turnover for different sectors of the economy (classified by NACE code). Levels of total Turnover by sector are also estimated and used to calibrate the Netherlands' structural business statistics (SBS) and to weight the contribution of each sector to the Netherlands' national accounts. Given the use of those level estimates, it is vital that they do not suffer from intercept bias. The relation between VAT Turnover and STS Turnover is known to vary by type of economic activity, depending on the specific tax regulations that apply (Van Delden et al., 2016). Hence, direct use of the VAT data may give a distorted view of the contribution of each sector to the economy of the Netherlands.

Van Delden et al. (2016) previously analysed the measurement quality of VAT data by a direct linear regression of Turnover as measured in the SBS survey[3] on VAT Turnover. This analysis was done separately for each NACE group (i.e., a stratum of units with the same NACE code). The results of these analyses were used, in combination with qualitative knowledge on tax regulations, to decide for each NACE group whether:

(a) VAT data could be used as a direct replacement of survey data;

(b) VAT data could be used after applying a linear correction to VAT Turnover; or

(c) VAT data could not be used.

The correction formulae for NACE groups in class (b) followed directly from the linear regression model [similar to formula (6.6) in this chapter, but with SBS

[3]The definitions of SBS and STS Turnover are identical in nearly all cases.

Turnover taking the role of true Turnover]. Class (c) also included NACE groups for which the analysis was inconclusive, e.g., because the results did not agree with expectations based on the tax regulations. In fact, a drawback of linear regression is that measurement errors in the SBS and VAT data are not explicitly taken into account. It is well known that estimates of regression parameters may be biased in the presence of measurement errors (Bound et al., 2001). Van Delden et al. (2016) did use a robust regression to avoid bias due to incidental (large) errors, but this does not provide protection against the effects of structural measurement errors. Therefore, we decided to do an alternative analysis using an SEM to account for potential measurement errors.

As mentioned above, the analyses in this application were done in R. The R code can be obtained from the first author upon request.

6.3.2 Data

In this chapter, we focus on the results for four NACE groups within the sector "Trade" listed in Table 6.1. A similar analysis was done for four NACE groups in another sector but the results are omitted here to save space; see Scholtus et al. (2015). For all of these NACE groups, the VAT data are currently not used to produce STS. In addition to Turnover, we included the following concepts in the SEM: Number of employees, Costs of purchases, and Total operating costs. All data referred to the year 2012.

Table 6.1: Overview of NACE groups considered in this application

NACE	description
45112	Sale and repair of passenger cars and light motor vehicles
45190	Sale and repair of trucks, trailers, and caravans
45200	Specialised repair of motor vehicles
45400	Sale and repair of motorcycles and related parts

For all concepts, one indicator is available from the SBS sample survey data. As a second indicator for the Number of employees, we used the value listed in the General Business Register (GBR) which is the population frame of business units maintained by Statistics Netherlands. Additional indicators for the other three variables were obtained from the Profit Declaration Register (PDR). This is an administrative data set provided to Statistics Netherlands by the tax authorities. Businesses are obliged to provide information to the PDR annually, but delayed reporting is accepted by the tax authorities up to several years after the reference period. In this study, we used the PDR data that were available by October 2014. Finally, VAT data on Total turnover were included. Businesses usually declare

VAT on a monthly or quarterly basis. In this study, we used the derived annual VAT Turnover.

Table 6.2 lists the population size in each NACE group as well as the number of units for which data were available. Businesses from the group of very large and/or complex units were excluded from this analysis, as Statistics Netherlands is not planning to use administrative data to replace the STS survey for this group. The SBS data set has survey weights to account for the sampling design and non-response. The SBS uses simple random sampling stratified by NACE group and size class (based on number of employees in the GBR). Correction for non-response is based on a weighting model involving NACE group, size class and legal form.

Table 6.2: Number of units in each NACE group. All figures refer to 2012 and, apart from the first line, to the population with large and/or complex units excluded.

NACE group	45112	45190	45200	45400
population (total)	18,680	1,790	6,054	1,763
population (w/o complex units)	18,556	1,739	6,018	1,759
SBS net sample	934	180	281	76
SBS net sample linked to admin. data	819	170	238	60
net audit sample	44	47	44	43

We could not link all units from the SBS data set to the two administrative data sets used here (PDR and VAT), mainly because not all fiscal units from the administrative data could be linked unambiguously to an SBS unit. In addition, some units had missing data in the PDR or VAT data sets (unit non-response). This explains the loss of units between the third and fourth line in Table 6.2. To account for potential selectivity introduced in this step, we recalibrated the survey weights within each NACE group, using a simplified version of the standard SBS weighting model. Since SBS Turnover was available for all units in the third row of Table 6.2, we could check whether the loss of records that were not linked to administrative data yielded selection bias that was not corrected by reweighting, at least for our target variable Turnover. We found no indication of such selection bias (Scholtus et al., 2015).

The necessary 'gold standard' data for evaluating intercept bias were obtained by having two senior subject-matter experts re-edit the SBS data for an audit sample of 50 units in each NACE group. The audit sample was stratified by a coarsened version of size class, reduced to just two strata, with 25 units taken in each stratum. The net audit sample was slightly smaller (see Table 6.2), mainly because we had selected the audit sample before linking the SBS data to administrative data. In addition, a few audited units turned out to be inactive or misclassified by type of

economic activity, which means that they were not part of the target population.

All variables considered here have skew distributions. For instance, most of the Turnover in each NACE group is concentrated among a few largest units. In theory, the PML estimation method should account for the fact that the data are not normally distributed. We also considered possible transformations to obtain data that were closer to being normal, or to account for heteroskedastic measurement errors. In some cases, this led to a slightly improved model fit (not shown here). On the other hand, these transformations made the interpretation of the measurement model in terms of the original variables less intuitive. We therefore decided to work with the untransformed data in this application, since we could find a model that fitted these data reasonably well (see below). In what follows, all financial variables are measured in millions of Euros.

Preliminary analyses revealed that some correlations between the original SBS data and the audited data were extremely high. This could be explained by the fact that relatively few values were changed during the audit, combined with the skewness of the data. These correlations close to 1 led to some computational problems, with covariance matrices close to being singular, so that `lavaan` could not estimate the parameters of the SEM. To avoid these problems, we decided to only include SBS Turnover in the model and exclude the other SBS variables.

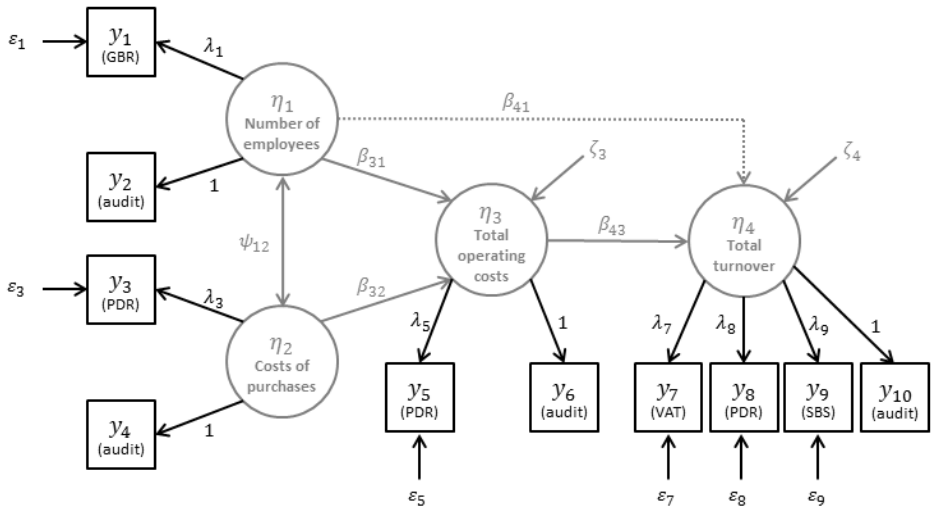


Figure 6.2: Path diagram of the basic model used in this application (intercepts not shown). For the group of non-audited units, remove variables y_2, y_4, y_6, and y_{10}.

The path diagram of the basic SEM used here is shown in Figure 6.2. For the structural part of the model, we used a nearly-saturated recursive model. The direction of the arrows was prompted by accounting rules that underlie these con-

ceptual variables: Costs of purchases (η_2) is a component of Total operating costs (η_3), which in turn contributes to the Total turnover (η_4); in addition, Number of employees (η_1) is closely related to Staff costs which is another component of η_3. The structural model is not fully saturated: we excluded the direct effect of η_2 on η_4 because there is no substantive reason why Costs of purchases would have an additional effect on Total turnover besides its contribution to Total operating costs. The direct effect of Number of employees on Total turnover was included in the initial model but fixed at zero for those NACE groups where it was found to be insignificant.

6.3.3 Results

Table 6.3 shows the following fit measures for the final chosen model in each NACE group: the robust chi-square test statistic X^{2*}_{SB} and robust versions of the CFI, TLI, and RMSEA; see Appendices 6.A.2 and 6.A.3 for precise definitions of these measures. For NACE groups 45112 and 45200, all measures indicate an excellent fit. For NACE group 45190, the robust chi-square statistic is somewhat high compared to the degrees of freedom and the other fit measures mostly indicate a reasonable fit. Finally, for NACE group 45400, the overall fit is rather poor. Note that the sample size in this last group is small, both compared to the other NACE groups and compared to minimal sample sizes that are recommended in the SEM literature [see, e.g., Boomsma (1982)].

Table 6.3: Fit measures for the final model

NACE group	45112	45190	45200	45400
X^{2*}_{SB}	57.5	91.0	41.0	172.9
df^*	62	61	61	62
p value	0.637	0.008	0.977	0.000
$\mathrm{CFI}^*_{\mathrm{SB}}$	1.000	0.966	1.000	0.925
$\mathrm{TLI}^*_{\mathrm{SB}}$	1.001	0.966	1.034	0.927
$\mathrm{RMSEA}^*_{\mathrm{SB}}$	0.000	0.077	0.000	0.255

We checked the residuals of the fitted models. In cases where the overall model fit was not very good, some large residuals did occur for the exogenous variables Number of employees and/or Costs of purchases, but never for Turnover. Thus, to the extent that the model may be misspecified, we assumed that these misspecifications were related only to other variables than Turnover. Moreover, results on simulated data in Scholtus and Bakker (2013a) suggest that, for the type of SEM considered here, the effects of local model misspecifications are not propagated to other parts of the model. Hence, for the purpose of making valid inferences

about the measurement quality of Turnover, we considered the fitted models to be adequate.

Table 6.4 displays the estimated factor loadings, measurement intercepts, and validities of Turnover as observed in VAT, PDR, and SBS. For completeness, the full set of parameter estimates is given in Appendix 6.B. Recall that the validity $|\lambda^s|$ is defined by (6.3). It is seen that the validity of VAT Turnover was high in all NACE groups considered here. On the other hand, the unstandardised λ indicates that the observed VAT Turnover was often systematically too high or too low compared to the true Turnover. For the intercept τ, no significant deviations from 0 were found in these NACE groups.

Table 6.4: Parameter estimates for Turnover (with standard errors)

	45112		45190		45200		45400	
parameter	estim.	s.e.	estim.	s.e.	estim.	s.e.	estim.	s.e.
λ (VAT)	0.79	0.01	0.89	0.02	1.29	0.19	0.80	0.04
τ (VAT)	−0.04	0.04	−0.00	0.05	−0.04	0.08	0.01	0.03
λ^s (VAT)	0.98		0.97		0.99		0.97	
λ (PDR)	1.02	0.01	0.95	0.02	1.23	0.20	0.99	0.03
τ (PDR)	0.00	0.05	0.06	0.03	−0.02	0.08	0.00	0.01
λ^s (PDR)	1.00		0.98		0.99		1.00	
λ (SBS)	1.01	0.01	1.01	0.00	1.21	0.19	0.98	0.02
τ (SBS)	−0.01	0.05	−0.00	0.00	−0.04	0.08	0.00	0.01
λ^s (SBS)	0.99		1.00		0.98		1.00	

Interestingly, the overall measurement quality of PDR Turnover was better than that of VAT Turnover. Unfortunately, the PDR data cannot be used directly for STS, because they are available only on an annual basis and because they suffer from administrative delay as mentioned above.

For all NACE groups, a correction formula for VAT Turnover was derived as described in Section 6.2.4. The results are shown in Table 6.5. Thus, for instance, to correct VAT Turnover to the scale of true Turnover in NACE group 45112, the following formula was obtained:

$$\widehat{\text{Turnover}} = 0.11 + 1.22 \times \text{VAT Turnover}.$$

Analogous correction formulae could be derived, if necessary, for the other observed Turnover variables (SBS and PDR).

Scholtus et al. (2015) compared the results of the SEM method in this application to those that would be obtained by the robust regression method of Van Delden et al. (2016). In some NACE groups, the two methods yielded similar correction formulae for VAT Turnover, while in other groups some large and significant differences occurred, in particular for the slope parameter. In all cases, the direction

Table 6.5: Intercept and slope of a correction formula for VAT Turnover (with standard errors)

	45112		45190		45200		45400	
parameter	estim.	s.e.	estim.	s.e.	estim.	s.e.	estim.	s.e.
a (VAT)	0.11	0.05	0.07	0.06	0.04	0.06	0.01	0.03
b (VAT)	1.22	0.02	1.07	0.04	0.76	0.12	1.18	0.06

of the difference was as expected from the estimated measurement parameters of SBS Turnover under the SEM model (Table 6.4). That is to say, for these NACE groups the assumption of the robust regression method that SBS Turnover is a good proxy for true Turnover was not satisfied.

6.3.4 Effect on publication figures

To conclude this section, we illustrate the effect of applying the correction formulae from Table 6.5 in practice. Recall that the Netherlands' short-term statistics on Turnover are based on a census survey for the largest or most complex units, and VAT data for the rest of the population. Estimates of Turnover levels (by NACE group) are obtained by summing the observed values for all units (within a NACE group) in both data sources, and estimated annual growth rates can be obtained as a ratio of estimated Turnover levels for two years.[4] Since no form of sampling is involved, these estimates are affected only by non-sampling errors, such as measurement errors.

Table 6.6: Effect of correction on estimated Turnover levels (2012 and 2013) and annual growth rate (2012/2013) for the target population

	45112		45190		45200		45400	
	2012	2013	2012	2013	2012	2013	2012	2013
Turnover	($\times 10^9$ Euro)							
VAT	25.2	24.0	5.1	4.9	2.7	2.6	1.6	1.5
VAT adjusted	29.1	27.7	5.2	5.1	2.2	2.1	1.7	1.6
rel. difference	$+15\%$	$+16\%$	$+3\%$	$+3\%$	-19%	-19%	$+7\%$	$+7\%$
growth rate								
VAT		0.951		0.971		0.979		0.916
VAT adjusted		0.953		0.971		0.984		0.916

Table 6.6 compares population estimates that would be obtained by using the VAT data directly and by applying the correction formulae estimated above; we look at annual Turnover levels for 2012 and 2013 and the corresponding annual

[4] The computation of growth rates in the actual statistical process is more complicated to account for population dynamics, but we ignore this aspect here.

growth rate. It is seen that correcting for measurement errors in the VAT data often has a substantial effect on the estimates of Turnover levels for these NACE groups. In other words, the uncorrected estimates seriously under- or overestimate the contributions of many NACE groups to the economy of the Netherlands. The effect on growth rates is much smaller and usually negligable, because the bias in the numerator and denominator mostly cancels out.

6.4 Conclusions and discussion

6.4.1 Discussion of results

In this chapter, we explored the possibility of using structural equation modelling to assess the measurement quality of administrative variables for statistical use. We specifically looked at validity and intercept bias. Estimating the intercept bias of an observed variable in a meaningful way requires the collection of additional 'gold standard' data for a random subsample of the original data. To illustrate the method, we applied it to assess the suitability of VAT data on Turnover for the Netherlands' quarterly STS and their derived annual values.

As the method is relatively expensive and complex, it might be useful in practice to apply a staged approach. Begin by making some preliminary comparisons between the administrative data and data from other sources (e.g., survey data), for instance by visual inspection of scatter plots or by robust linear regression, as was done by Van Delden et al. (2016). This preliminary analysis may already be conclusive in two possible ways: either by revealing that the administrative data are only weakly correlated to the other data (in which case the data are probably not useful), or by revealing that the two data sources contain nearly identical values (in which case the data may be considered to have high validity, provided that the errors in the two sources are independent). In all other cases, it seems premature to draw conclusions about the validity of the administrative data at this stage. Moreover, nothing can be concluded at this stage about the presence of systematic bias in either data source.

For the second stage, if the preliminary analysis is inconclusive, one may proceed with the estimation of an SEM to evaluate the validity. For this, the collection of additional audit data is not required. If the validity turns out to be low, the administrative data should probably not be used.

If the validity is high and one is also interested in the bias, then one may proceed to the final stage. For this stage, an audit sample is conducted. An extended SEM can then be used to evaluate the intercept bias and, if necessary, estimate a formula for correcting the bias. Results on simulated data in Scholtus (2014b) sug-

gest that this method provides valid results even for small audit samples and that the precision of the estimated SEM parameters increases slowly with the size of the audit sample. Hence, from a cost-benefit point of view it may be reasonable to keep the audit sample small in practice. On the other hand, the audit samples in these simulations were selected by simple random sampling and it may in fact be possible to obtain significant improvements in precision by optimising the design of the audit sample. This could be an interesting topic for future research. Within the method as discussed here, any form of probability sampling can be used to select the audit sample so long as the design is known.

In traditional applications of SEMs to survey data, model identification is often achieved by asking multiple variants of the same question, either within the same interview or in a follow-up interview (Saris and Gallhofer, 2007). For panel surveys, an alternative is the so-called simplex design which involves asking the same question to the same respondents at ($\geq$ 3) different time points (Alwin, 2007). With administrative data sources, asking follow-up questions is almost never possible. In addition, while many longitudinal administrative data sources are available, the recorded values often remain unchanged until an event occurs that triggers an alteration (Bakker, 2011a). This implies that measurement errors in a single administrative source at different time points are often strongly correlated. A more generally applicable way to achieve model identification with administrative data may be to link them to survey data, as we did in this chapter. This approach does require that the data sources can be perfectly linked (no linkage errors). In practice, there may be records that cannot be linked. In that case, one should check whether the linked data are sufficiently representative of the population, and possibly weigh the data to improve this.

6.4.2 Assumptions and limitations

A strong assumption of the method is that it is possible to obtain 'gold standard' versions of the target variables, at least for a small subsample of units. In practice, applications where absolute levels are of concern are likely to arise only for 'factual' variables. For such variables, an objective true value can be determined in principle, although the measurement procedure that is required to obtain this value may be difficult, expensive, or otherwise inconvenient to implement in practice. Clearly, the outcome of the method relies on the quality of the audit data. In our application, the audit data were obtained through re-editing by subject-matter experts. An important, albeit difficult, question is whether it is realistic to consider these data as a 'gold standard'. As a topic for future research, it may be interesting to investigate the re-editing process in more detail and to find out how confident the

experts are about their decisions. It is conceivable that the quality of the audit data actually differs by sub-population, e.g., because less information is available on smaller units. It may also be interesting to investigate by simulation to what extent the estimate of validity and the correction formula for intercept bias are robust to minor violations of the assumption that the audit data do not contain measurement error.

A limitation of the application in Section 6.3 is that the model was fitted to data of only one year, so we could not test whether the estimated measurement parameters change over time. It would be good to repeat the analysis on data from a different year. Note that this would also require a new audit sample.

In the method as described here, we did not introduce any prior assumptions about the relative measurement quality of each observed variable (apart from the audit data). In this respect, the comparison between the validities of the variables in this application was completely data-driven. If a researcher does have prior knowledge about the relative merits of each data source, this could be incorporated in the model by means of (in)equality constraints on parameters (Rosseel, 2012). Alternatively, it may then be natural to use a Bayesian SEM (Palomo et al., 2007).

In the type of model that was used here, measurement errors are considered to follow a continuous distribution. In practice, measurements on the same theoretical variable in a survey and an administrative source are sometimes found to be exactly equal for a substantial subset of the units. This is often explained by assuming that measurement errors in survey and administrative data are 'intermittent', i.e., there is a non-zero probability of observing the true value. Guarnera and Varriale (2016) consider a latent class model for measurement errors in numerical variables which explicitly takes this property into account. An alternative interpretation of the above phenomenon is that measurement errors in different sources are correlated because the measurement procedures cannot be considered independent. For instance, it might happen that some units simply report the same value of Turnover in the survey that they provided previously to the tax authorities, without going back to their original records. Correlated measurement errors can be taken into account in the SEM framework, provided that sufficient other indicators of the latent variables are available.

As remarked in footnote 1, the SEM in this chapter yields estimates of the so-called indicator validity or empirical validity of the observed variables. Estimating the theoretical validity by factoring out the reliability component requires a more complex SEM, the so-called *multitrait-multimethod model*. This approach has been applied successfully in survey questionnaire design (Scherpenzeel and Saris, 1997), but it is not readily applicable to administrative data.

6.4.3 Potential applications

In the context of the application in Section 6.3, estimating the validity and intercept bias was useful to help deciding whether a specific administrative source could replace an existing sample survey, possibly after a model-based correction. Another, similar type of application might involve comparing several potential (administrative) sources for the same target variable and choosing the best one. This could be relevant for instance for NSIs that are moving towards a population census based on register data (Berka et al., 2012). Of course, the decision to use or not to use an administrative data source for statistics should be based on other criteria as well, besides the measurement quality. See, e.g., Daas et al. (2011) for a comprehensive overview of relevant quality indicators for administrative data. In addition, the outcome of a model-based analysis should always be compared with expectations based on other, qualitative knowledge about an administrative data source. For statistics that are already based on administrative data or mixed sources, the method described in this chapter could be useful to quantify the influence of measurement errors on published statistical results.

The multi-group SEM with an audit sample as used in this chapter can be applied to answer other research questions too. One interesting application in official statistics might be to compare the effects of automatic editing and manual editing on administrative or survey data (Scholtus et al., 2015). In some applications, a model-based bias correction might replace part of the manual editing to yield a more efficient statistical process. This alternative seems interesting in particular for processing large administrative data sets, where even modern selective editing methods may be too resource-demanding.

Finally, the use of an audit sample to identify an SEM may also be relevant in some applications outside official statistics. SEMs are frequently used as an analysis technique in sociology, political science, and other social sciences, as well as in econometrics. Researchers in these areas are seldom interested in the true metrics of latent variables, and intercept bias is not usually a direct concern. However, this type of study often involves a comparison between groups (e.g., across countries, across subpopulations, or across time) and in that case different amounts of intercept bias or unequal factor loadings between groups can lead to invalid conclusions (Bielby, 1986a). Using an audit sample for model identification (when possible) may reduce this risk (Sobel and Arminger, 1986; Scholtus et al., 2015).

Appendix 6.A Additional methodology and results

In this appendix, a more detailed and technical description is given of the methodology from Section 6.2. Section 6.A.1 reviews some general results on SEM estimation. Section 6.A.2 considers adjustments regarding missing data that are needed for the application in this chapter. Section 6.A.3 provides formulae for additional fit measures used in this application.

6.A.1 PML estimation for SEMs

We begin by reviewing some of the theory behind SEM estimation, starting with ML estimation for i.i.d. normal data and moving on to PML estimation for non-normal data and complex survey data. A more comprehensive discussion of these topics, as well as other estimation methods, can be found in Muthén and Satorra (1995) or Oberski (2014) (for a single group) and Satorra (2002) (for multiple groups).

We consider a multiple-group SEM, from which the single-group model follows as a special case. Suppose there are G groups with n_g sampled units in group g, and the samples are independent between groups. The total sample size is $n = \sum_{g=1}^{G} n_g$. Let $\mathbf{y}_{gi} = (y_{gi1}, \ldots, y_{gip})'$ denote the vector of observed variables for unit i in group g. In contrast with the notation of Section 6.2, we use a matrix of uncentered cross-product moments to summarise the observed data in each group: $\mathbf{S}_g^0 = (1/n_g) \sum_{i=1}^{n_g} \mathbf{y}_{gi}\mathbf{y}_{gi}'$, where it is assumed that a constant 1 is included as one of the observed variables. In addition, let $\mathbf{s}_g^0 = \text{vech}(\mathbf{S}_g^0)$, where vech(.) denotes the operator that vectorises a symmetric matrix by stacking the non-redundant elements column-wise (Harville, 1997). The population equivalents of $\mathbf{S}_g^0$ and $\mathbf{s}_g^0$ (i.e., the matrix and vector to which these quantities converge as $n_g \to \infty$) are denoted by $\boldsymbol{\Sigma}_g^0$ and $\boldsymbol{\sigma}_g^0$, respectively. We also define $\mathbf{s}^0 = ((\mathbf{s}_1^0)', \ldots, (\mathbf{s}_G^0)')'$ and $\boldsymbol{\sigma}^0 = ((\boldsymbol{\sigma}_1^0)', \ldots, (\boldsymbol{\sigma}_G^0)')'$.

For G groups, the distance function F_{ML} that was mentioned in Section 6.2 is given by:

$$F_{\text{ML}}(\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \frac{n_g}{n} \left\{ \log |\boldsymbol{\Sigma}_g^0(\boldsymbol{\vartheta})| + \text{tr}(\mathbf{S}_g^0 \boldsymbol{\Sigma}_g^0(\boldsymbol{\vartheta})^{-1}) - \log |\mathbf{S}_g^0| - p \right\}, \qquad (6.7)$$

where tr(.) denotes the trace of a matrix. Let $\hat{\boldsymbol{\vartheta}}$ be the estimator that is obtained by minimising (6.7). In addition, define $\hat{\mathbf{V}}_{\text{ML}}$ as a block-diagonal matrix with $(n_g/n)2^{-1}\mathbf{D}' \left\{ (\mathbf{S}_g^0)^{-1} \otimes (\mathbf{S}_g^0)^{-1} \right\} \mathbf{D}$ as blocks along the main diagonal, with $\mathbf{D}$ the so-called duplication matrix (Harville, 1997); let $\mathbf{V}_{\text{ML}}$ denote the population equivalent of $\hat{\mathbf{V}}_{\text{ML}}$. If the data are i.i.d. multivariate normal, then it can be shown

that $\mathbf{V}_{\mathrm{ML}}$ is identical to $\boldsymbol{\Gamma}^{-1}$, where $\boldsymbol{\Gamma}$ denotes the asymptotic variance-covariance matrix of $\sqrt{n}\mathbf{s}^0$. Under this assumption of normality, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\vartheta}}$ is given by

$$\mathrm{avar}(\hat{\boldsymbol{\vartheta}}) = \frac{1}{n}(\boldsymbol{\Delta}'\mathbf{V}_{\mathrm{ML}}\boldsymbol{\Delta})^{-1}, \tag{6.8}$$

with $\boldsymbol{\Delta} = \partial\boldsymbol{\sigma}^0(\boldsymbol{\vartheta})/\partial\boldsymbol{\vartheta}'$. Furthermore, under the hypothesis that the model holds, the test statistic $X^2_{\mathrm{ML}} = (n-1)F_{\mathrm{ML}}$ is then asymptotically distributed as a chi-square variate with degrees of freedom equal to $df = \mathrm{rank}(\boldsymbol{\Delta}'_\perp\boldsymbol{\Gamma}\boldsymbol{\Delta}_\perp)$. Here, $\boldsymbol{\Delta}_\perp$ denotes an orthogonal complement to the matrix $\boldsymbol{\Delta}$ (Harville, 1997; Satorra, 2002).

When the data are not normally distributed (but the i.i.d. assumption does hold), minimising (6.7) still provides consistent point estimates under rather general conditions (Bollen, 1989). Asymptotic standard errors based on (6.8) may be too small in this case. The correct expression for the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\vartheta}}$ is now:

$$\mathrm{avar}(\hat{\boldsymbol{\vartheta}}) = \frac{1}{n}(\boldsymbol{\Delta}'\mathbf{V}_{\mathrm{ML}}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\mathbf{V}_{\mathrm{ML}}\boldsymbol{\Gamma}\mathbf{V}_{\mathrm{ML}}\boldsymbol{\Delta}(\boldsymbol{\Delta}'\mathbf{V}_{\mathrm{ML}}\boldsymbol{\Delta})^{-1}, \tag{6.9}$$

which reduces to (6.8) when $\mathbf{V}_{\mathrm{ML}} = \boldsymbol{\Gamma}^{-1}$. The asymptotic distribution of X^2_{ML} also need not be chi-square in this case. Satorra and Bentler (1994) proposed a relatively simple adjustment to X^2_{ML}. Define

$$\hat{\mathbf{U}} = \hat{\mathbf{V}}_{\mathrm{ML}} - \hat{\mathbf{V}}_{\mathrm{ML}}\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\Delta}}'\hat{\mathbf{V}}_{\mathrm{ML}}\hat{\boldsymbol{\Delta}})^{-1}\hat{\boldsymbol{\Delta}}'\hat{\mathbf{V}}_{\mathrm{ML}}$$

and

$$\hat{c}_{\mathrm{SB}} = \mathrm{tr}(\hat{\mathbf{U}}\hat{\boldsymbol{\Gamma}})/df. \tag{6.10}$$

In (6.10), $\hat{\boldsymbol{\Delta}}$ is obtained by evaluating $\boldsymbol{\Delta}$ at $\boldsymbol{\vartheta} = \hat{\boldsymbol{\vartheta}}$ and $\hat{\boldsymbol{\Gamma}}$ is an appropriate estimate of $\boldsymbol{\Gamma}$ (see below). The Satorra-Bentler-corrected test statistic is $X^2_{\mathrm{SB}} = X^2_{\mathrm{ML}}/\hat{c}_{\mathrm{SB}}$, with the chi-square distribution with df degrees of freedom as its reference distribution (if the model holds).

An estimate of $\boldsymbol{\Gamma}$ may be obtained as follows. Define $\mathbf{d}^0_{gi} = \mathrm{vech}(\mathbf{y}_{gi}\mathbf{y}'_{gi})$, so that $\mathbf{s}^0_g = (1/n_g)\sum_{i=1}^{n_g}\mathbf{d}^0_{gi}$. Since this re-defines $\mathbf{s}^0_g$ as a sample mean, it can be shown that an appropriate estimator for $\mathrm{avar}(\sqrt{n_g}\mathbf{s}^0_g)$ is given by

$$\hat{\boldsymbol{\Gamma}}_g = \frac{1}{n_g - 1}\sum_{i=1}^{n_g}(\mathbf{d}^0_{gi} - \mathbf{s}^0_g)(\mathbf{d}^0_{gi} - \mathbf{s}^0_g)'.$$

Hence, $\boldsymbol{\Gamma} = \mathrm{avar}(\sqrt{n}\mathbf{s}^0)$ may be estimated by

$$\hat{\boldsymbol{\Gamma}} = \begin{bmatrix} \frac{n}{n_1}\hat{\boldsymbol{\Gamma}}_1 & & & \\ & \frac{n}{n_2}\hat{\boldsymbol{\Gamma}}_2 & & \\ & & \ddots & \\ & & & \frac{n}{n_G}\hat{\boldsymbol{\Gamma}}_G \end{bmatrix}. \tag{6.11}$$

For complex survey designs, one should first of all replace $\mathbf{s}^0$ by a design-consistent estimator of $\boldsymbol{\sigma}^0$. Muthén and Satorra (1995) and Oberski (2014) consider the general case of a survey design that involves stratification, multi-stage selection and clustering. Essentially, in this case we can write

$$\mathbf{s}_g^0 = (1/N_g) \sum_{i=1}^{n_g} w_{gi} \mathbf{d}_{gi}^0$$

for some weights w_{gi} that depend on the survey design, with $N_g = \sum_{i=1}^{n_g} w_{gi}$. To apply the PML approach, we can still use (6.9), (6.10), and (6.11), provided that each $\hat{\boldsymbol{\Gamma}}_g$ is replaced by a variance estimator that takes the sample design for group g into account. The R package `lavaan.survey` implements this by referring to the variance estimation functionality of the `survey` package (Lumley, 2004).

It should be noted that expression (6.11) is based on the assumption that the samples are independent between groups. For survey designs that involve without-replacement sampling, this will be false in general unless the survey happens to be stratified by the variable that defines the groups. In particular, this assumption was violated in the application of Section 6.3; note that, conditionally on the total sample, units that are not selected in the first group (the audit sample) automatically belong to the second group. To obtain correct standard errors and fit measures for applications where the samples are not independent across groups, it would make sense from a design-based point of view to include the off-diagonal blocks $\frac{n}{\sqrt{n_g n_h}} \hat{\boldsymbol{\Gamma}}_{gh}$ in (6.11), where $\hat{\boldsymbol{\Gamma}}_{gh}$ denotes an estimate of $\text{acov}(\sqrt{n_g}\mathbf{s}_g^0, \sqrt{n_h}\mathbf{s}_h^0)$ ($g \neq h$). As far as we are aware, this problem has not been treated in the SEM literature. Papadopoulos and Amemiya (2005) considered correlation between groups in the case where the groups represent waves of a longitudinal study and the same respondents are observed multiple times, but they did not take other aspects of finite-population sampling into account. Deng and Yuan (2015) proposed a more general solution, but still focussed on correlations due to multiple observations on the same set of respondents. The general case of between-group dependencies due to a finite-population sampling design remains open.

An approximate adjustment to $\hat{\boldsymbol{\Gamma}}$ to account for inter-group dependency in the application of Section 6.3 was derived in Scholtus et al. (2015). As discussed there, the effect on the fit measures and estimated standard errors was very small in this application. For ease of exposition, we ignored the adjustment in this chapter.

6.A.2 Missing data

The use of an audit sample leads naturally to a two-group SEM with some of the variables missing by design in the second group. As described in Section 6.2.3,

Baumgartner and Steenkamp (1998) suggested that these missing values can be accounted for by imputing random, normally-distributed values with mean zero and variance one, such that the imputed variables are uncorrelated to all other variables in the second group. (That is, they are both uncorrelated amongst themselves and uncorrelated to the observed variables.) In case a complex survey design is used, the design-consistent estimates of the mean, variance and covariances should equal 0, 1 and 0, respectively.

As described in Section 6.2.3, the measurement equations for the missing variables in the second group are then chosen in such a way that the means, variances and covariances involving these variables are reproduced exactly by the SEM, while the estimation of the rest of the model is not affected by these variables. The sample moments involving the missing variables have thus been fixed so that they do not contribute to F_{ML} (or any other fitting function). The degrees of freedom of the model should be corrected to take this into account. Let q denote the number of missing variables in the second group and let df the denote the uncorrected degrees of freedom of the model, computed as if the imputed values were ordinary observed values. Since we have fixed q means and

$$p + (p-1) + \cdots + (p-q+1) = pq - \frac{q(q-1)}{2}$$

distinct covariances, the correct degrees of freedom should be:

$$df^* = df - q\left(p - \frac{q-3}{2}\right). \tag{6.12}$$

Baumgartner and Steenkamp (1998) applied the above approach only in the context of standard ML estimation. For PML estimation, we have to make an additional adjustment to $\hat{\mathbf{\Gamma}}_2 = \widehat{\mathrm{avar}}(\sqrt{n_2}\mathbf{s}_2^0)$ (or, more generally, to $\hat{\mathbf{\Gamma}}_g$ for each group g in which missing variables have been imputed in this way). Since the observed means and covariances involving the imputed variables are fixed, all elements of the corresponding rows and columns of $\hat{\mathbf{\Gamma}}_2$ should be set to zero. In particular, the Satorra-Bentler correction factor (6.10) is replaced in this context by $\hat{c}^*_{\mathrm{SB}} = \mathrm{tr}(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}}^*)/df^*$, where df^* is given by (6.12) and $\hat{\mathbf{\Gamma}}^*$ is obtained by making the above-mentioned adjustment to $\hat{\mathbf{\Gamma}}$ from (6.11). The overall fit of the model can now be tested by comparing $X^{2*}_{\mathrm{SB}} = X^2_{\mathrm{ML}}/\hat{c}^*_{\mathrm{SB}}$ to a chi-square distribution with df^* degrees of freedom.

6.A.3 Other fit measures

In Section 6.3, several other measures were used in addition to X^{2*}_{SB} to evaluate the model fit. For the sake of completeness, we provide expressions for the robust

(PML) versions of these fit measures, with adjustments to account for the imputed values in the second group (see Section 6.A.2). The following formulae are based on the default implementation in `lavaan`.

- Comparative Fit Index (CFI):

$$\text{CFI}^*_{\text{SB}} = 1 - \frac{\max\left\{X^{2*}_{\text{SB}} - df^*, 0\right\}}{\max\left\{X^{2*}_{\text{SB}} - df^*, X^{2*}_{\text{SB},0} - df^*_0, 0\right\}}. \tag{6.13}$$

- Tucker-Lewis Index (TLI):

$$\text{TLI}^*_{\text{SB}} = \frac{(X^{2*}_{\text{SB},0}/df^*_0) - (X^{2*}_{\text{SB}}/df^*)}{(X^{2*}_{\text{SB},0}/df^*_0) - 1}. \tag{6.14}$$

- Root Mean Square Error of Approximation (RMSEA):

$$\text{RMSEA}^*_{\text{SB}} = \sqrt{G \max\left\{n^{-1}(X^{2*}_{\text{SB}} - df^*), 0\right\} / df^*}. \tag{6.15}$$

Note: The CFI and TLI compare the fit of the model to that of a so-called baseline model. In the application of Section 6.3, we used the default baseline model selected by `lavaan`: this is the independence model with no restrictions across groups and with each observed variable modelled as $y_k = \tau_k + \epsilon_k$, with $\text{cov}(\epsilon_k, \epsilon_l) = 0$ for all $k \neq l$. In expressions (6.13) and (6.14), $X^{2*}_{\text{SB},0}$ and df^*_0 refer to this baseline model. These adjusted quantities can be obtained from their unadjusted versions $X^2_{\text{SB},0}$ and df_0 analogously to Section 6.A.2, with one subtle difference in the definition of df^*_0. Under the baseline model, the intercepts and error variances of the q imputed variables in the second group are not fixed (as in our original model) but estimated. This means that our adjustment to df_0 needs to account for $2q$ degrees of freedom less than before. Hence, the correction formula for the degrees of freedom of the baseline model becomes:

$$df^*_0 = df_0 - q\left(p - \frac{q-3}{2}\right) + 2q = df_0 - q\left(p - \frac{q+1}{2}\right). \tag{6.16}$$

Appendix 6.B Parameter estimates

For each NACE group of the application in Section 6.3, the full set of parameter estimates for the final model is listed in Table 6.7.

Table 6.7: Parameter estimates for the final model; parameter names and indices refer to Figure 6.2.

	45112		45190		45200		45400	
parameter	estim.	s.e.	estim.	s.e.	estim.	s.e.	estim.	s.e.
λ_1	0.83	0.02	0.90	0.05	0.74	0.07	0.81	0.08
λ_2	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–
λ_3	1.03	0.01	0.95	0.03	1.30	0.32	0.98	0.01
λ_4	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–
λ_5	1.03	0.01	0.97	0.02	1.22	0.23	1.01	0.01
λ_6	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–
λ_7	0.79	0.01	0.89	0.02	1.29	0.19	0.80	0.04
λ_8	1.02	0.01	0.95	0.02	1.23	0.20	0.99	0.03
λ_9	1.01	0.01	1.01	0.00	1.21	0.19	0.98	0.02
λ_{10}	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–	$1/0^a$	–
θ_{11}	1.24	0.57	9.57	2.32	2.77	1.33	1.02	0.42
θ_{22}	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–
θ_{33}	0.04	0.01	0.36	0.13	0.05	0.05	0.00	0.00
θ_{44}	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–
θ_{55}	0.03	0.02	0.41	0.18	0.05	0.04	0.00	0.00
θ_{66}	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–
θ_{77}	1.00	0.20	0.57	0.19	0.04	0.01	0.04	0.01
θ_{88}	0.06	0.02	0.41	0.19	0.05	0.03	0.01	0.00
θ_{99}	0.87	0.21	0.00	0.00	0.06	0.03	0.00	0.00
$\theta_{10,10}$	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–	$0/1^a$	–
τ_1	1.04	0.17	0.81	0.37	1.17	0.30	1.04	0.15
τ_2	0^a	–	0^a	–	0^a	–	0^a	–
τ_3	−0.01	0.04	0.03	0.03	−0.00	0.06	0.00	0.00
τ_4	0^a	–	0^a	–	0^a	–	0^a	–
τ_5	−0.01	0.04	0.03	0.04	−0.01	0.08	−0.00	0.01
τ_6	0^a	–	0^a	–	0^a	–	0^a	–
τ_7	−0.04	0.04	−0.00	0.05	−0.04	0.08	0.01	0.03
τ_8	0.00	0.05	0.06	0.03	−0.02	0.08	0.00	0.01
τ_9	−0.01	0.05	−0.00	0.00	−0.04	0.08	0.00	0.01
τ_{10}	0^a	–	0^a	–	0^a	–	0^a	–
β_{31}	0.05	0.00	0.04	0.01	0.04	0.01	0.04	0.01
β_{32}	1.03	0.00	1.14	0.03	1.19	0.11	1.16	0.04
β_{41}	0^a	–	0.01	0.00	0.01	0.00	0^a	–
β_{43}	1.02	0.00	1.00	0.01	0.96	0.04	1.02	0.02
ψ_{11}	157	0.62	96.8	0.65	67.4	0.07	10.5	2.06
ψ_{22}	28.6	1.75	8.29	1.52	0.36	0.18	0.56	0.10
ψ_{12}	59.7	1.33	25.2	2.49	3.59	0.87	2.11	0.31
ψ_{33}	0.01	0.00	0.05	0.01	0.02	0.02	0.00	0.00
ψ_{44}	0.01	0.00	0.02	0.01	0.00	0.00	0.00	0.00
α_1	3.44	0.21	4.06	0.40	3.02	0.36	1.31	0.22
α_2	1.21	0.06	0.94	0.08	0.15	0.02	0.29	0.06
α_3	0.02	0.01	0.02	0.03	0.02	0.02	0.00	0.01
α_4	0.03	0.01	−0.01	0.02	0.02	0.01	0.01	0.01

[a] parameter fixed a priori, value may depend on group as indicated

Chapter 7

Estimating the Quality of Business Survey Data before and after Automatic Editing

This chapter was co-authored by Bart F. M. Bakker (Statistics Netherlands, VU University) and Sam Robinson (Leiden University). Author contributions: all authors contributed ideas; Robinson contributed to the mathematical work and implemented the EM algorithm for Model 2; Scholtus developed the rest of the mathematical work, carried out the analysis and wrote the report; Bakker edited the report. A condensed version of this chapter was presented as a conference paper at the 2017 UN/ECE Work Session on Statistical Data Editing as Scholtus et al. (2017).

7.1 Introduction

Statistical results can be affected by measurement errors in the underlying data. National statistical institutes (NSIs) and other producers of official statistics therefore edit their data for errors as part of the process of generating statistical output (De Waal et al., 2011). Editing can be done manually or automatically. Statistics Netherlands uses both automatic and manual editing in the production of economic statistics. Automatic editing methods are more efficient than manual editing – in terms of both costs and time – and yield results that are reproducible (Pannekoek et al., 2013). On the other hand, it is generally held that the measurement quality of automatically edited data is lower than that of manually edited data (EDIMBUS, 2007), although little quantitative evidence exists either for or against this belief.

In this chapter, we propose to evaluate the measurement quality of automatically edited survey data in an objective way, by modelling the residual measurement errors after editing. We will compare the quality of an observed variable before and after automatic editing, in terms of validity (correlation of the observed variable to the true variable of interest) and intercept bias (systematic deviation be-

tween the observed variable and the variable of interest). To obtain an identified model, auxiliary variables are included by linking the survey data before and after automatic editing to data from administrative sources.

Two different measurement error models will be considered. The first model is a structural equation model, in which the true values of several conceptual variables are represented by latent (unobserved) variables. Each latent variable is measured by one or more observed variables. Measurement errors in these observed variables are represented by error terms in a linear regression equation. Survey variables before and after automatic editing are included as observed variables in this model, along with several administrative variables. In addition, identification of all relevant model parameters requires that a small audit sample is included; for the units in this audit sample, it is assumed that error-free versions of the variables of interest have been obtained by an additional manual editing effort. We apply this model to survey data of the Netherlands' Structural Business Statistics (SBS) linked to administrative data, building on a previous application of structural equation modelling in Scholtus et al. (2015) (see also Chapter 6 of this thesis). The previous application focussed on estimating the measurement quality of the administrative variables, whereas here we focus on the survey variables.

The structural equation model assumes that – apart from the audit sample – all units have errors on all observed variables. We also apply a different latent variable model that assumes that errors occur according to a so-called "intermittent" mechanism. This means that each observation has a certain non-zero probability of being error-free. The latter assumption may be more appropriate for the data at hand and is in line with much of the existing literature on data editing [see, e.g., Di Zio and Guarnera (2013) and Chapter 2 of this thesis]. The observations that do contain errors are again modelled using linear regression techniques. A practical advantage of this second model is that no additional audit data are needed to identify all model parameters.

The remainder of this chapter is organised as follows. We begin by briefly describing the data editing process for the Netherlands' SBS and the data that will be used here in Section 7.2. An introduction to the two models and a summary of their results are given in Section 7.3 (Model 1) and Section 7.4 (Model 2). Possible implications and limitations of these results are discussed in Section 7.5. Finally, some concluding remarks follow in Section 7.6.

7.2 Application

7.2.1 Automatic editing in the Netherlands' SBS

The SBS aim to provide an overview of employment and the financial structure (costs and revenues) of different sectors of the economy. Data are collected in a sample survey of businesses. The sample is stratified by type of economic activity and size class. Businesses are classified by main economic activity according to the so-called NACE classification. We use the term "NACE group" to refer to a stratum of units with similar economic activities for which separate SBS estimates are published. The SBS questionnaire is tailored separately to each NACE group. On average, the SBS questionnaire produces a data set of about 100 different variables.

Figure 7.1: Overview of the process for automatic editing in the Netherlands' SBS.

Figure 7.1 gives an overview of the automatic editing process for the Netherlands' SBS. Each box corresponds to a version of the data and each arrow between boxes corresponds to a process step during which changes can be made to the data.

The automatic process steps are, in chronological order:

- *Input processing (IP)*: Technical checks on the initial, unedited data and correction of uniform thousand-errors.
- *Deductive processing 1 (DP1)*: Deterministic IF-THEN-rules to resolve common errors with a known cause.
- *Error localisation and imputation (EL&I)*: Automatic error localisation followed by imputation of missing and discarded values.
- *Deductive processing 2 (DP2)*: Deterministic IF-THEN-rules to resolve inconsistencies not handled during EL&I (e.g., consistency between financial variables and stock variables).

Of these process steps, the EL&I step is the most complex from a methodological point of view. In this step, the data are made consistent with a given set of restrictions (so-called edit rules) by replacing observed values with new values if necessary. The selection of values to change is based on the paradigm of Fellegi and Holt (1976), which aims to minimise the number of changed values given that

the resulting record has to satisfy all restrictions. This leads to a mathematical optimisation problem which can be solved automatically (De Waal et al., 2011).

The methodology of the two deductive processing steps DP1 and DP2 is less complex. These steps consist of applying a number of deterministic rules that can make changes to the data. An example of a rule that is used during process step DP1 is:

$$\begin{aligned} &\text{IF Depreciations } < 0 \\ &\text{THEN Depreciations } := -\text{ Depreciations.} \end{aligned}$$

According to this rule, if any negative values are encountered for the variable Depreciations, these have to be replaced by their absolute values. There is in fact an edit rule (restriction) in the SBS which states that the value of Depreciations must be non-negative.

We will not describe the process steps of Figure 7.1 in more detail here. A general overview of methodology for automatic data editing can be found elsewhere, e.g., in De Waal et al. (2011) or Pannekoek et al. (2013). A detailed description of the data editing process of the Netherlands' SBS is provided by De Jong (2002) and Hoogland and Smit (2008).

A feature of the above automatic editing process is that, during each process step, the vast majority of observed values are not changed. Thus, most of the observed values in the unedited data (first box in Figure 7.1) are equal to the corresponding values in the edited data (final box in Figure 7.1). This happens because the editing methods used for the Netherlands' SBS all assume, either explicitly or implicitly, that most of the observed values are correct to begin with. For instance, the Fellegi-Holt paradigm that is used during the EL&I step is based on the assumption that errors are rare, and that a record should therefore be made consistent with the edit rules by changing the least possible number of values.

For this study, we want to compare the measurement quality of variables in the input data (second box in Figure 7.1) and edited data (right-most box). This will give an impression of the overall effect of automatic editing on data quality. We take the second box as a starting point rather than the first box, because some of the technical checks carried out during the IP step are required to know whether the data are accessible at all. In fact, for questionnaires that are submitted on paper – which is still done by a minority of responding units – the data are digitised as part of the IP step, so no unedited data are available in digital form for these units.

In the actual production process of the Netherlands' SBS, only a subset of the data after input processing is treated by the remaining automatic process steps in Figure 7.1. The other records are edited manually instead. A selection procedure

is applied to the input data to assign records either to automatic or manual editing (Hoogland, 2006). For the present study, we created a version of the data in which as many records as possible were edited automatically, regardless of the selection that was made during actual production. By focussing on this data set, we can evaluate the "pure" effect of automatic editing on the measurement quality of SBS data, rather than the combined effect of measurement and selection.

In practice, not all records can be edited automatically. During the IP step, records can be rejected (discarded from further automatic processing) if certain key variables such as total turnover are missing. During the EL&I step, a small number of records for which no solution to the error localisation problem can be found are also rejected. In the actual production process, these records would then be treated manually instead. For the purpose of this study, they are treated as non-response. Fortunately, this concerns only a handful of records.

7.2.2 Data

For this application, we used SBS data of reference year 2012 for four different NACE groups within the economic sector "Trade". These NACE groups are listed in Table 7.1. The SBS data were linked to administrative data from three different sources. Firstly, we used the General Business Register (GBR) which is maintained by Statistics Netherlands as a population frame of businesses in the Netherlands. We also used two data sets collected by the Netherlands' tax authorities: Value-Added Tax declarations (VAT) on turnover and the Profit Declaration Register (PDR) which contains many administrative variables that are similar to SBS variables. Finally, for a small random subsample of units the SBS data were re-edited by subject-matter experts with the aim of recovering the true values for all variables (audit data). Some additional information about these different data sources can be found in Scholtus et al. (2015) or Chapter 6 of this thesis.

Table 7.1: Overview of NACE groups considered in this application

NACE	description
45112	Sale and repair of passenger cars and light motor vehicles
45190	Sale and repair of trucks, trailers, and caravans
45200	Specialised repair of motor vehicles
45400	Sale and repair of motorcycles and related parts

Table 7.2 lists the number of available records in each NACE group. The editing process for very large and/or complex units differs from that of the other units (in particular, they are never edited automatically), so these were not included in the present study (second line in Table 7.2).

Table 7.2: Number of units in each NACE group. All figures refer to 2012 and, apart from the first line, to the population with large and/or complex units excluded.

NACE group	45112	45190	45200	45400
population (total)	18,680	1,790	6,054	1,763
population (w/o complex units)	18,556	1,739	6,018	1,759
SBS net sample, edited	914	165	269	74
SBS net sample, edited and linked to admin. data	810	158	231	58
net audit sample	43	45	43	43

The GBR and SBS data could be linked directly by business identification number. The other administrative sources contain information for fiscal units rather than statistical units. The relationship between fiscal and statistical units is known in the GBR, but not all fiscal units can be linked to a single statistical unit. Therefore, and also due to missing data in the administrative sources, it was not possible to link all units in the SBS data to administrative data (fourth line in Table 7.2). Scholtus et al. (2015) investigated whether the linked data might suffer from selection bias but found no indication that such a bias occurred.

It is worth noting that the application in this chapter used mostly the same data that were used in the application of Chapter 6 of this thesis. The only difference is that the present application used versions of the SBS data before and after automatic editing, whereas the previous application used SBS data after (automatic or manual) editing during regular production. In particular, differences between the numbers of units in Table 7.2 and those in Table 6.2 are due to SBS records that could not be edited automatically. Although the main aim of the present study is to evaluate the measurement quality of SBS data before and after automatic editing, this application also allows us to assess the robustness of the previously estimated measurement parameters for the administrative variables. Ideally, the estimated validity of administrative variables should not be influenced by the choice of survey data included in the model.

7.3 Model 1: A structural equation model

7.3.1 Methodology

The first model we considered is an extension of the structural equation model used in Chapter 6. In general, a linear structural equation model (SEM) consists of two types of regression equations. Firstly, a number of latent variables $\eta_1, \ldots, \eta_m$ are introduced and linear regression models are defined to describe the relations between these latent variables. Secondly, each of the observed variables $y_1, \ldots, y_p$

in the data set is related to (at least) one of the latent variables by a linear regression model. In our application, the latent variables represent the true variables of interest and the observed variables are error-prone measurements of these variables. In this special case, each observed variable y_k is related to exactly one latent variable (say, η_{j_k}) and the SEM is given by the following regression equations (where i denotes a unit):

$$\eta_{ji} = \alpha_j + \sum_{j' \neq j} \beta_{jj'} \eta_{j'i} + \zeta_{ji}, \quad (j = 1, \ldots, m), \tag{7.1}$$

$$y_{ki} = \tau_k + \lambda_k \eta_{j_k i} + \epsilon_{ki}, \quad (k = 1, \ldots, p). \tag{7.2}$$

Here, ζ_j and ϵ_k denote zero-mean disturbance terms, with $\text{cov}(\zeta_j, \zeta_{j'}) = \psi_{jj'}$ and $\text{cov}(\epsilon_k, \epsilon_{k'}) = \theta_{kk'}$. See, e.g., Bollen (1989) for a general introduction to SEMs.

For our purposes here, we are mainly interested in the parameters of the measurement model (7.2). The intercept τ_k and factor loading λ_k describe the effect of systematic measurement errors in y_k. To the extent that τ_k deviates from 0 and λ_k deviates from 1, the observed variable y_k is biased with respect to the true value η_{j_k}. The absolute value of the standardised version of λ_k (say, λ_k^s) can be used to quantify the so-called indicator validity coefficient of y_k as a measure of η_{j_k}:

$$\text{IVC}(y_k) = |\lambda_k^s| = |\lambda_k| \sqrt{\frac{\text{var}(\eta_{j_k})}{\text{var}(y_k)}} = \sqrt{1 - \frac{\text{var}(\epsilon_k)}{\text{var}(y_k)}}. \tag{7.3}$$

The indicator validity coefficient lies between 0 and 1. Values close to 1 indicate a strong linear relationship between the observed value of y_k and the true value of η_{j_k}. The term "indicator validity" is due to Saris and Andrews (1991); see also Chapter 2.

For this study, we focussed on four key SBS variables: Number of employees, Costs of purchases, Total operating costs, and Total turnover. We only discuss results related to Total turnover here, as this was the main variable of interest both in this application and the previous application in Chapter 6.

A path diagram of the SEM that was used in this application is shown in Figure 7.2. This path diagram is identical to Figure 6.2 in Chapter 6, except that the observed variable SBS Total turnover (y_9 in that model) is now split into two versions, corresponding to input data (y_{9I}) and edited data (y_{9E}). Since any measurement errors that occur in the input data and are not resolved during automatic editing will also be present in the edited data, it is likely that the error terms ϵ_{9I} and ϵ_{9E} are correlated. The covariance between these errors was therefore added as a model parameter ($\theta_{9I,9E}$). All other disturbance terms were assumed to be mutually uncorrelated.

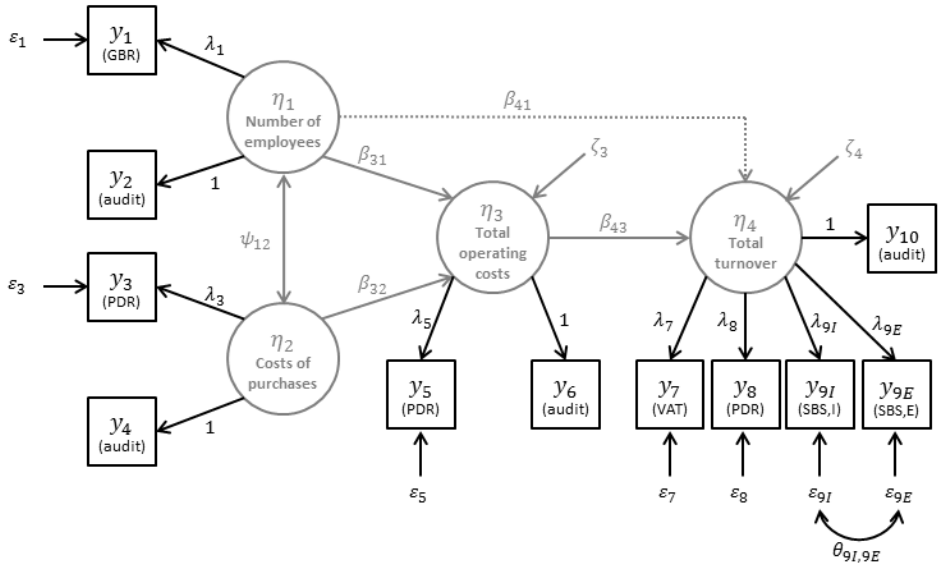


Figure 7.2: Path diagram of the basic SEM used in this application (intercepts not shown). For the group of non-audited units, remove variables y_2, y_4, y_6, and y_{10}.

The error-free observed variables y_2, y_4, y_6 and y_{10} in the model of Figure 7.2 were available only for the units in the audit sample. Since the audit sample was selected by subsampling from the original SBS data, the missing values on these variables were known to be missing at random in the terminology of Little and Rubin (2002). We refer to Scholtus et al. (2015) or Chapter 6 of this thesis for a description of the way these missing data were taken into account in the estimation of the SEM.

SEMs are often estimated by maximum likelihood, under the assumptions that the observed data can be seen as (1) independent draws (2) from a multivariate normal distribution. In the present application, neither of these assumptions holds. We used pseudo maximum likelihood (PML) to account for non-normality of the data and for the finite-population sampling design of the SBS survey and audit sample; see, e.g., Muthén and Satorra (1995). The models were estimated in the R environment for statistical computing (R Development Core Team, 2017), using the functionality in the packages `lavaan` (Rosseel, 2012) and `lavaan.survey` (Oberski, 2014). See Scholtus et al. (2015) or Chapter 6 for more details about the estimation procedure.

Due to the fact that many observed values in the SBS input and edited data were equal, as mentioned above, y_{9I} and y_{9E} were highly correlated (and usually also highly correlated to the audit variable y_{10}). For one NACE group (45190), y_{9I} and y_{9E} were perfectly correlated and it was therefore not possible to fit a model

that involved both variables simultaneously. In this case, we attempted instead to fit a model that involved only one of y_{9I} and y_{9E}.

7.3.2 Results

Fit measures and parameter estimates related to Total turnover for the best fitting models are shown in Table 7.3 and Table 7.4. See Chapter 6 for definitions of these fit measures.

Table 7.3: Fit measures for the final model

NACE group	45112	45190	45200	45400
X^{2*}_{SB}	67.7	150.4	88.2	230.2
df^*	78	61[a]	78	78
p value	0.790	0.000	0.201	0.000
$\mathrm{CFI}^*_{\mathrm{SB}}$	1.000	0.885	0.984	0.915
$\mathrm{TLI}^*_{\mathrm{SB}}$	1.003	0.886	0.985	0.917
$\mathrm{RMSEA}^*_{\mathrm{SB}}$	0.000	0.138	0.034	0.266

[a] No differences between y_{9I} and y_{9E} in observed data (correlation is 1). To avoid numerical issues, a model was fitted with y_{9I} excluded from the data.

Table 7.4: Estimated indicator validity and measurement parameters (intercept and slope) for Turnover under Model 1 (with standard errors)

	45112		45190		45200		45400	
parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
τ (VAT)	−0.01	0.05	0.03	0.07	0.01	0.04	0.01	0.03
λ (VAT)	0.79	0.01	0.88	0.05	1.08	0.16	0.80	0.05
IVC (VAT)	0.95		0.98		0.97		0.97	
τ (PDR)	0.00	0.05	0.08	0.05	0.02	0.05	0.00	0.01
λ (PDR)	1.02	0.01	0.96	0.03	1.05	0.19	0.99	0.03
IVC (PDR)	1.00		1.00		0.97		1.00	
τ (SBS,I)	0.00	0.05	−[a]	−	−0.00	0.00	0.00	0.01
λ (SBS,I)	1.00	0.01	−	−	1.00	0.00	0.98	0.02
IVC (SBS,I)	0.98		−		1.00		1.00	
τ (SBS,E)	0.02	0.05	0.03	0.06	−0.00	0.00	0.01	0.01
λ (SBS,E)	1.00	0.01	1.04	0.04	1.00	0.00	0.98	0.02
IVC (SBS,E)	0.98		0.97		1.00		1.00	

[a] See footnote for Table 7.3.

The main results are:

- For the administrative versions of Turnover (VAT and PDR), in most NACE groups the parameter estimates were very similar to those in Chapter 6. In particular, it is seen that the PDR data provide a better measurement of Turnover than the VAT data for these NACE groups, both in terms of IVC and bias. VAT Turnover underestimates the true Turnover ($\lambda < 1$) for all NACE groups except 45200 where $\lambda > 1$.

- In all NACE groups the estimated IVC and bias of SBS input data and SBS edited data were virtually equal. By extension, this result also holds for NACE group 45190 where the two SBS variables were perfectly correlated.

- In all NACE groups where both y_{9I} and y_{9E} were included in the model, the estimated correlation between ϵ_{9I} and ϵ_{9E} differed significantly from 0. In fact, it was estimated to be larger than 0.90 in all of these groups but one.

The first result is positive, because it shows that the estimated measurement properties of the administrative data do not depend significantly on the choice of survey data to include in the model. The latter two results are natural, given that the observed values of the two SBS turnover variables were equal for most units. However, this also undermines the assumption that an SEM can be used to describe these data: under this type of model, the event that two observed variables are equal should occur with probability zero. In the next section, we will discuss a different model that allows for the possibility that two observed variables are exactly equal in some records.

7.4 Model 2: A finite mixture model

7.4.1 Methodology

Guarnera and Varriale (2015, 2016) considered a measurement error model for data from multiple sources which takes into account the possibility that a latent target variable is sometimes measured without error. In contrast to the model of Section 7.3, we now focus on a single variable of interest.

Let η_i denote the true value of a variable of interest (say, Total turnover) for unit i. For each observed variable y_k that measures η, a 0-1-indicator z_k is introduced such that $y_{ki} = \eta_i$ if $z_{ki} = 0$. For units with $z_{ki} = 1$, y_{ki} contains a measurement error which is assumed to have a similar structure as under Model 1:

$$y_{ki} = \begin{cases} \eta_i & \text{if } z_{ki} = 0, \\ \tau_k + \lambda_k \eta_i + e_{ki} & \text{if } z_{ki} = 1, \end{cases} \tag{7.4}$$

where τ_k and λ_k are constants and e_{ki} follows a normal distribution with mean zero and variance σ_k^2. In a single formula, this model for y_k can be expressed as follows:

$$y_{ki} = (1 - z_{ki})\,\eta_i + z_{ki}\,(\tau_k + \lambda_k \eta_i + e_{ki})\,. \tag{7.5}$$

It is assumed that, for each unit, all z_k and all e_k are independent across different observed variables. This is similar to assuming that the measurement errors in an SEM are uncorrelated. The probability of observing an error on y_k is represented by the parameter $\pi_k = P(z_k = 1) = E(z_k)$.

In addition to the above measurement model, we also use an ordinary linear regression model to describe the variation in the true values η_i across units as a function of covariates $\mathbf{x}$:

$$\eta_i = \boldsymbol{\beta}' \mathbf{x}_i + u_i, \tag{7.6}$$

where $\boldsymbol{\beta}$ denotes a vector of regression coefficients and it is assumed that u_i is normally distributed with mean zero and variance σ^2.

The parameters of the model given by (7.5) and (7.6) again provide several interesting indicators for the measurement quality of each observed variable. Firstly, we can look at the error probability π_k: a value of π_k closer to 1 indicates that more errors occur for variable y_k. Secondly, the intercept τ_k and slope λ_k in (7.5) provide information about the amount of systematic bias in y_k, conditional on the event that an error occurs. Finally, to quantify the effect of the random measurement errors e_k on y_k, we can again use the indicator validity coefficient. For the observations y_{ki} that contain errors, the indicator validity coefficient is given by the standardised slope parameter $|\lambda_k^s|$, analogously to (7.3) for Model 1, except that the standardisation should now be based only on the part of the data for which y_k contains errors:

$$\text{IVC}(y_k | z_k = 1) = |\lambda_k^s| = |\lambda_k| \sqrt{\frac{\text{var}(\eta | z_k = 1)}{\text{var}(y_k | z_k = 1)}} = \sqrt{1 - \frac{\sigma_k^2}{\lambda_k^2 \sigma_\eta^2 + \sigma_k^2}}. \tag{7.7}$$

Here, σ_η^2 denotes the total variance of η which, under model (7.6), is given by $\sigma_\eta^2 = \boldsymbol{\beta}' \boldsymbol{\Sigma}_{xx} \boldsymbol{\beta} + \sigma^2$, where $\boldsymbol{\Sigma}_{xx}$ denotes the variance-covariance matrix of $\mathbf{x}$. (Note that, in the presence of covariates, σ^2 represents the unexplained variance in η.) Furthermore, the error-free y_{ki} can be seen as observations with a validity coefficient of 1. Hence, a natural definition of the unconditional indicator validity coefficient of y_k under Model 2 is:

$$\begin{aligned} \text{IVC}(y_k) &= \pi_k \times \text{IVC}(y_k | z_k = 1) + (1 - \pi_k) \times 1 \\ &= 1 - \pi_k \left(1 - \sqrt{1 - \frac{\sigma_k^2}{\lambda_k^2 \sigma_\eta^2 + \sigma_k^2}} \right). \end{aligned} \tag{7.8}$$

This unconditional indicator validity coefficient has the same interpretation as the IVC defined under Model 1.

Under this model, there is a non-zero probability (namely $1 - \pi_k$) that an observed value y_{ki} is equal to the true value η_i and therefore error-free. The event that two different observed values y_{ki} and y_{li} for a given unit i are identical occurs with probability $(1-\pi_k)(1-\pi_l)$, which is also non-zero. By contrast, these events have probability zero under the SEM of Section 7.3. Thus, when the observed data contain a non-negligible number of records with $y_{ki} = y_{li}$, Model 2 may be more appropriate than Model 1.

In the previous paragraph, we used the fact that the event $y_{ki} = \eta_i$ occurs with probability zero if y_{ki} contains an error. This follows from the above assumption that the measurement errors are normally distributed, since the probability of drawing any specific value from such a distribution equals zero. In fact, the same property holds for any random variable with a continuous distribution, so we do not need the assumption of normality here. By extension, since we also assumed that errors in different observed variables are independent, it follows that if we observe $y_{ki} = y_{li}$, it must hold that $y_{ki} = y_{li} = \eta_i$. That is to say, under the assumptions of this model, if a record contains the same value for two (or more) observed variables, then that value must also be equal to the corresponding true value. Thus, not only can the observed values be error-free under this model, in the presence of multiple observed values it is possible to *recognise* some of these error-free values from the observed data themselves. Of course, all of this need not be true if the model does not hold for the data at hand. In particular, the assumption that errors in different variables are independent is a strong assumption that may not always be satisfied in practice.

The above model formulation is more general than that of Guarnera and Varriale (2015, 2016) who assumed that $\tau_k = 0$ and $\lambda_k = 1$ for all observed variables [in which case formula (7.5) can be simplified to $y_{ki} = \eta_i + z_{ki}e_{ki}$]. In principle, the model could be extended to multiple target variables (Guarnera and Varriale, 2016) but an estimation procedure for such an extended model has not yet been developed. Note that it is possible to introduce covariates $\mathbf{x}$ to predict the true value of the target variable, but potential errors in these covariates are not taken into account. Another important limitation of this model is that it relies heavily on the assumption that measurement errors in different observed variables are independent. Thus, correlated measurement errors cannot be taken into account under the model as formulated here.

As Model 2 is less well-known, we will provide some more details about the estimation procedure than we did for Model 1. Technically, model (7.5)–(7.6) is

an example of a so-called finite mixture model. For n observations on K observed variables $y_1, \ldots, y_K$ with error patterns specified by $z_1, \ldots, z_K$, the complete-data loglikelihood function is as follows:

$$\begin{aligned}\ell_c(\boldsymbol{\theta}) = C &+ \sum_{k=1}^{K} \left[n \log(1-\pi_k) + n_k \log\left(\frac{\pi_k}{1-\pi_k}\right)\right] \\ &- \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\eta_i - \boldsymbol{\beta}'\mathbf{x}_i)^2 \\ &- \sum_{k=1}^{K}\left[\frac{n_k}{2}\log\sigma_k^2 + \frac{1}{2\sigma_k^2}\sum_{i=1}^{n} z_{ki}(y_{ki} - \tau_k - \lambda_k\eta_i)^2\right]. \end{aligned} \tag{7.9}$$

In this expression, $\boldsymbol{\theta}$ denotes a vector containing all distinct parameters of the model, $n_k = \sum_{i=1}^{n} z_{ki}$ denotes the number of observations with an error on y_k, and C denotes a constant term that does not depend on any unknown parameters. See, e.g., McLachlan and Peel (2000, p. 48) for the complete-data loglikelihood of a general finite mixture model.

In practice, not all η_i and $z_{1i}, \ldots, z_{Ki}$ are observed – although some of them are. As noted above, the true value η_i (and hence all z_{ki}) can be inferred from the observed data when $y_{ki} = y_{li}$ for (at least) two different observed values. To give an example: in the presence of $K = 3$ observed variables there are $2^3 = 8$ possible error patterns. We can derive η_i and the error pattern for all observations with

$$(z_{1i}, z_{2i}, z_{3i}) \in \{(0,0,0),\ (1,0,0),\ (0,1,0),\ (0,0,1)\}, \tag{7.10}$$

as for these observations at least two of the observed values are identical. For the remaining observations, we can infer that (z_{1i}, z_{2i}, z_{3i}) must have one of the four remaining patterns – that is, at least two of the three observed values must be erroneous. However, we cannot derive the exact error pattern and we cannot obtain the value of η_i for these observations.

The model parameters can still be estimated from (7.9) by using maximum likelihood estimation for incomplete data. Guarnera and Varriale (2016) worked out an EM algorithm (Expectation – Maximisation) for the model with $K = 3$ observed variables, under the restriction that $\tau_k = 0$ and $\lambda_k = 1$. Robinson (2016) gives a detailed description of this algorithm, including an extension to estimate τ_k and λ_k. For the present study, we wrote an implementation of this algorithm in R. We refer to Little and Rubin (2002) for an introduction to EM algorithms in general. Below we also report asymptotic standard errors for the estimated parameters; details on the derivation of these standard errors are given in Appendix 7.A.

In our application, there are four different sources of Turnover: VAT, PDR, SBS-input and SBS-edited. However, measurement errors in SBS-input and SBS-edited are highly correlated, as discussed above. We therefore estimated separate models with either SBS-input or SBS-edited included besides VAT and PDR, so $K = 3$. In addition, for a subset of units we have an audit variable for which it is assumed that $y_{ki} = \eta_i$ with certainty. In the context of this model, this simply means that there may be some additional units for which η_i and (z_{1i}, z_{2i}, z_{3i}) can be obtained from the observed data. It is interesting to note that for Model 2 – in contrast to Model 1 – the intercept and slope parameters τ_k and λ_k are identified even without audit data, due to the presence of observations where the true value η_i can be inferred from the data.

The loglikelihood (7.9) is correct under the assumptions that (1) the residuals u_i in (7.6) are $N(0, \sigma^2)$ distributed and (2) the measurement errors e_{ki} in (7.5) are $N(0, \sigma_k^2)$ distributed. For our data on Turnover, neither of these assumptions holds. In contrast to Model 1, this non-normality may be more problematic for Model 2, as we do not have a robust estimation procedure for this model.

By examining the cases with error patterns (7.10), for which η_i is known, it was found that the above assumptions were more reasonable after applying a logarithmic transformation to the data. We will therefore present results for Model 2 with all variables measured on a log scale. To be precise, for error-prone observations ($z_{ki} = 1$), the measurement model (7.4) was replaced by

$$\log(y_{ki} + 0.5) = \tau_k + \lambda_k \log(\eta_i + 0.5) + e_{ki}. \tag{7.11}$$

We used $\log(y + 0.5)$ rather than $\log y$ to be able to handle cases with $y = 0$. A similar transformation was applied to (7.6). Note that model (7.11) is equivalent to

$$y_{ki} + 0.5 = \exp(\tau_k)(\eta_i + 0.5)^{\lambda_k} \exp(e_{ki}). \tag{7.12}$$

Thus, for the variables on the original scale the error structure is now multiplicative rather than additive.

It is clear that the parameters τ_k and λ_k in (7.12) have a very different interpretation in comparison to model (7.4), except for the trivial special case where $\tau_k = 0$ and $\lambda_k = 1$. Unfortunately, this means that the estimated parameters for Model 2 with this transformation are not directly comparable to those obtained for Model 1 in Table 7.4. In order to have sets of parameter estimates that could be compared between the two models, we therefore estimated the SEM of Figure 7.2 again with the same logarithmic transformation (7.11) applied to our data. The resulting estimated measurement parameters for Turnover are shown in Table 7.5. The estimated intercept and slope parameters for Model 2 that will be reported

below should therefore be compared to those in Table 7.5 for Model 1. As noted above, there is no straightforward way to relate the results in Table 7.5 to those in Table 7.4.

Table 7.5: Estimated indicator validity and measurement parameters (intercept and slope) for Turnover on a log scale under Model 1 (with standard errors)

	45112		45190		45200		45400	
parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
τ (VAT)	0.08	0.15	0.54	0.16	0.13	0.16	−0.34	0.33
λ (VAT)	0.93	0.03	0.89	0.03	0.98	0.03	1.01	0.07
IVC (VAT)	0.93		0.95		0.98		0.94	
τ (PDR)	−0.12	0.09	0.40	0.14	0.25	0.16	−0.32	0.12
λ (PDR)	1.02	0.02	0.94	0.02	0.96	0.03	1.06	0.03
IVC (PDR)	0.98		0.93		0.98		1.00	
τ (SBS,I)	0.09	0.17	$-^a$	–	0.15	0.26	−0.89	0.43
λ (SBS,I)	0.97	0.03	–	–	0.94	0.05	1.15	0.09
IVC (SBS,I)	0.91		–		0.83		0.96	
τ (SBS,E)	−0.09	0.19	−0.02	0.01	0.17	0.19	0.17	0.29
λ (SBS,E)	1.00	0.03	1.00	0.00	0.95	0.04	0.96	0.06
IVC (SBS,E)	0.92		1.00		0.92		0.96	

[a] See footnote for Table 7.3.

7.4.2 Results

Table 7.6 shows a selection of estimated parameters for the model with SBS input data; Table 7.7 shows the corresponding results for the model with SBS edited data. In all models, we used the observed (edited) Number of employees and Total operating costs in SBS and a constant as covariates $\mathbf{x}$ to predict the true value of Turnover [cf. (7.6)]. In these tables, we have included both the conditional IVC (7.7) for error-prone observations (rows labelled "cIVC") and the unconditional IVC (7.8) for all observations (rows labelled "IVC"). Note that the unconditional IVC is always closer to 1 than the conditional IVC, due to the contribution of observations that are error-free.

The main results that can be seen in these tables are:

- The estimated parameters for the administrative sources do not differ significantly between both tables – i.e., the choice between using input or edited SBS data does not affect the parameter estimates for the other observed variables.

- According to the model, the VAT variable has relatively large error proba-

Table 7.6: Estimated indicator validity and measurement parameters (error probability, intercept and slope) for Turnover on a log scale under Model 2 (with standard errors); model with SBS input data for Turnover

	45112		45190		45200		45400	
parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
π (VAT)	0.91	0.01	0.82	0.03	0.76	0.03	0.85	0.05
τ (VAT)	−0.26	0.06	0.67	0.22	0.07	0.10	−0.19	0.26
λ (VAT)	1.00	0.01	0.90	0.03	0.99	0.02	0.99	0.04
cIVC (VAT)	0.98		0.96		0.98		0.97	
IVC (VAT)	0.98		0.96		0.98		0.98	
π (PDR)	0.64	0.02	0.34	0.05	0.55	0.03	0.68	0.06
τ (PDR)	0.04	0.11	0.85	0.55	0.26	0.13	−0.07	0.06
λ (PDR)	0.99	0.01	0.89	0.08	0.97	0.02	1.01	0.01
cIVC (PDR)	0.96		0.84		0.98		1.00	
IVC (PDR)	0.97		0.95		0.99		1.00	
π (SBS,I)	0.10	0.01	0.40	0.05	0.13	0.02	0.08	0.04
τ (SBS,I)	0.68	0.68	−0.25	0.10	1.21	1.80	−1.87	0.89
λ (SBS,I)	0.80	0.11	1.04	0.01	0.40	0.31	1.20	0.17
cIVC (SBS,I)	0.70		1.00		0.26		0.92	
IVC (SBS,I)	0.97		1.00		0.91		0.99	

Table 7.7: Estimated indicator validity and measurement parameters (error probability, intercept and slope) for Turnover on a log scale under Model 2 (with standard errors); model with SBS edited data for Turnover

	45112		45190		45200		45400	
parameter	estimate	s.e.	estimate	s.e.	estimate	s.e.	estimate	s.e.
π (VAT)	0.91	0.01	0.82	0.03	0.76	0.03	0.85	0.05
τ (VAT)	−0.21	0.06	0.67	0.22	0.08	0.10	−0.20	0.26
λ (VAT)	0.99	0.01	0.90	0.03	0.99	0.02	0.99	0.04
cIVC (VAT)	0.98		0.96		0.98		0.97	
IVC (VAT)	0.98		0.96		0.98		0.98	
π (PDR)	0.64	0.02	0.34	0.05	0.54	0.03	0.68	0.06
τ (PDR)	0.11	0.12	0.85	0.55	0.29	0.13	−0.08	0.06
λ (PDR)	0.98	0.01	0.89	0.08	0.96	0.02	1.01	0.01
cIVC (PDR)	0.95		0.84		0.97		1.00	
IVC (PDR)	0.97		0.95		0.99		1.00	
π (SBS,E)	0.11	0.02	0.40	0.05	0.14	0.03	0.10	0.04
τ (SBS,E)	−0.30	0.49	−0.25	0.10	0.72	0.90	0.08	0.84
λ (SBS,E)	0.98	0.07	1.04	0.01	0.70	0.16	0.94	0.17
cIVC (SBS,E)	0.83		1.00		0.68		0.88	
IVC (SBS,E)	0.98		1.00		0.96		0.99	

bilities in all NACE groups ($\hat{\pi}$ lies between 0.76 and 0.91), whereas the corresponding probabilities for PDR and SBS are smaller. In all NACE groups except 45190, SBS has the smallest error probabilities, both before and after automatic editing.

- Just as for Model 1, we find that automatic editing does not have a large impact on the overall IVC of SBS Turnover. The error probabilities before and after editing also do not differ significantly, and in some NACE groups the probability of observing an error after editing is actually slightly larger. On the other hand, we can see that after editing the intercept and slope parameters τ and λ for the SBS variable are closer to 0 and 1, respectively, in all NACE groups except 45190, where no change occurs. In NACE groups 45112 and 45200, it is also seen that the conditional IVC for the error-prone observations is improved by the editing procedure; in NACE group 45400 editing actually slightly reduces the IVC.

- Comparing the results of Models 1 and 2, it is seen that the λ values in Table 7.5 are often closer to 1 than the corresponding values in Table 7.6 and Table 7.7, in particular for the SBS variables. This is as expected, since Model 1 estimates a single regression line for both the error-free and error-prone observations, whereas τ and λ in Model 2 refer only to the error-prone observations. Most of the exceptions to this rule occur for VAT, where error-free observations in fact appear to be rare. It is also seen that the unconditional IVC values are nearly always closer to 1 under Model 2 than under Model 1, both before and after editing.

Overall, we conclude that the effect of automatic editing on the measurement quality of SBS Turnover for these data is slightly larger than was suggested by the results of Model 1, but is still fairly small. In particular, the error probabilities in Table 7.6 and Table 7.7 suggest that the automatic editing process has successfully corrected only a small subset of the errors in SBS Turnover that were actually present in these data.

7.5 Discussion

The results of both Model 1 and Model 2 for the data in our application suggest that, according to our measurement quality indicators, the effect of automatic editing on SBS Turnover is very limited. Looking at the data, this is not unexpected, because in fact only a small number of Turnover values in our data set were changed during the automatic editing process of Figure 7.1. This naturally leads to the question

whether automatic editing has any added value for the Netherlands' SBS. We believe that it does, but that this added value is mainly related to other quality criteria that are not considered by the above measurement error models.

In particular, an important aim of automatic editing is to obtain a data set which is consistent with respect to a pre-defined set of edit rules. The edit rules define univariate and multivariate restrictions that would be expected to hold if the data were error-free. Two examples of edit rules for SBS are:

$$\text{Turnover} \geq 0;$$
$$\text{Turnover} - \text{Total operating costs} = \text{Profit}.$$

It should be noted that it is possible that an SBS observation satisfies all edit rules while still containing one or more errors. On the other hand, an observation that does not satisfy all edit rules certainly contains errors, but these errors might be very small and therefore hardly affect the validity or bias as defined under Model 1 or 2. For instance, it is not uncommon for SBS data to contain so-called rounding errors [see Scholtus (2011a) or Chapter 3 of this thesis].

Regarding the importance for NSIs of obtaining data that are consistent with edit rules, Pannekoek and De Waal (2005) noted the following:

> "Statistically speaking there is indeed hardly any reason to let a data set satisfy all edits, other than the *hope* that enforcing internal consistency will result in data of higher statistical quality. NSIs, however, have the responsibility to supply data for many different academic and nonacademic users in society. For the majority of these users, inconsistent data are incomprehensible. They may reject the data as being an invalid source of information or make adjustments themselves. This hampers the unifying role of an NSI in providing data that are undisputed by different parties (...)."

Thus, even when automatic editing does not significantly improve the measurement quality of a data set (in terms of validity and bias), it can still be useful for NSIs as a relatively cheap way of obtaining consistent data.

On the other hand, the results of Model 2 do suggest that the automatic editing process has successfully corrected only a small subset of the errors in SBS Turnover that were actually present in these data. While the validity of the variables after editing is quite close to 1, the remaining errors do appear to cause a noticeable bias in some of the NACE groups. It may therefore be useful to apply a measurement error model such as Model 2 during regular production to estimate the effects on statistical output of errors that remain in the data after automatic editing. If these effects are significant, improved output may be estimated by applying a correction for measurement errors. This correction requires a prediction of the true value η_i, given one or more of the observed values y_{ki} and the estimated model

parameters. For Model 1, a correction procedure was outlined in Chapter 6. For Model 2, predicted values for η based on all observed data are actually computed as part of the E step of the EM algorithm [see Robinson (2016)], so the extension to a correction procedure is straightforward from a theoretical point of view.

Alternatively, Model 2 could be used as part of a selective editing procedure during regular production to identify records that are likely to contain errors (either before or after automatic editing) for manual follow-up. Having estimated the model, one can compute the posterior probability that y_{ki} contains an error, taking into account all observed data and the estimated model parameters; see Guarnera and Varriale (2016). Note that for observations such as those in (7.10), these posterior probabilities are equal to 0 or 1. In combination with a measure of the expected error size (which could be derived from the predicted value of η_i under the model), these posterior probabilities can provide a basis for selecting the observations that are likely to contain the most important errors. See Di Zio and Guarnera (2013) for a detailed discussion of the use of "intermittent-error" models for selective data editing.

As it stands, Model 2 has some important limitations that were highlighted by the above application. Firstly, the assumption that the true values and measurement errors are normally distributed may often be violated in practice. It is not known to what extent the maximum likelihood estimates for this model are robust to non-normality. In principle, other versions of the model could be developed for different distributions, but this has not been done yet. In fact, for many variables that occur in business statistics (such as Turnover) a log-normal distribution is reasonable, in which case the current model can be applied to the data after a logarithmic transformation. However, this does complicate the interpretation of the model parameters τ_k and λ_k, as can be seen in (7.11) and (7.12). An interesting alternative solution to handle non-normal data that is sometimes used in finite mixture models is to model these non-normal distributions themselves as mixtures of two or more normal distributions, which leads to a "mixture of mixture models" (McLachlan and Peel, 2000; Di Zio et al., 2007). It remains to be seen whether such an approach would work in our situation (e.g., identifiability might be a problem).

The model also assumes that measurement errors in different observed variables are independent. In principle, this assumption could be relaxed, but this would make the estimation procedure more complicated. Furthermore, for a given number of observed variables K, only a limited number of dependencies can be added before the model becomes under-identified.

In this application, we focussed on a single variable of interest (Turnover). Since automatic editing – in particular: error localisation based on the Fellegi-Holt

paradigm – is a multivariate procedure, it would actually be more interesting to model several target variables simultaneously. A relatively straightforward extension of Model 2 could be made if errors for different variables of interest are independent, but this assumption may often be unreasonable in practice. Without such an assumption, the model quickly becomes very complex as more latent variables are added (Guarnera and Varriale, 2016).

Finally, the maximum likelihood estimation procedure used here assumes that the data consist of independent, identically distributed observations. It would be good to extend the estimation procedure to take the effects of finite-population sampling into account, as survey observations are hardly ever independent in practice. In particular, this is likely to affect the standard errors of the estimated parameters.

7.6 Conclusion

In this chapter, we have applied two different latent variable models to estimate and compare the measurement quality of survey data on Turnover from the Netherlands' SBS before and after automatic editing. Model 1 is a structural equation model, while Model 2 is a finite mixture model which takes into account the possibility that some observed values are error-free. In principle, the latter type of model seems to be more suitable for SBS data. However, as indicated in Section 7.5, Model 2 currently has some important limitations. We maintain that it is useful to develop this model further to address these limitations. Several potential practical applications of Model 2 in the context of automatic and selective editing were discussed in Section 7.5.

In our application, we found that automatic editing methods had a minor effect on the validity and bias of business survey data. Overall, the measurement quality of the edited data was, at best, only marginally better than that of the input data. Of course, these results are based on a single data set for a small number of NACE groups. Also, the target variable Turnover is usually reported with relatively high accuracy in the SBS survey. Thus, our findings may not extend to all applications of automatic editing in business statistics.

Nevertheless, we can tentatively conclude that the main merit of automatic editing may be that it provides consistent data at low costs, but that it often does not significantly improve the measurement quality of individual variables in terms of validity or bias. This suggests that it would be good to develop measurement error models that can be used to estimate the effects of residual errors in edited data during regular production, which could then be used to correct statistics for measurement error, or to select observations for further manual editing. Model

2 could be used as a starting point for the development of such a model for the Netherlands' SBS and for other applications in business statistics.

Appendix 7.A Asymptotic standard errors for Model 2

Asymptotic standard errors and confidence intervals for the estimated parameters of Model 2 can be obtained from the following general result [see, e.g., Van der Vaart (1998)]: under regularity conditions and assuming that the model is correct, as $n \to \infty$ the complete-data maximum-likelihood estimates $\hat{\boldsymbol{\theta}}$ based on ℓ_c in (7.9) converge to a joint normal distribution with the true parameter values $\boldsymbol{\theta}$ as mean and a variance-covariance matrix given by

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathbf{I}_{\text{obs}}^{-1}(\hat{\boldsymbol{\theta}}) = -\left(\frac{\partial^2 \ell_c}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)^{-1},$$

where $\mathbf{I}_{\text{obs}}(\hat{\boldsymbol{\theta}})$ denotes the *observed information matrix*, i.e., minus one times the Hessian matrix of second-order partial derivatives of $\ell_c(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. In particular, asymptotic standard errors for the parameter estimates in $\hat{\boldsymbol{\theta}}$ are given by the square roots of the diagonal elements of $\mathbf{V}(\hat{\boldsymbol{\theta}})$.

With incomplete data, these standard errors will be inflated. Little and Rubin (2002) describe a two-step procedure to obtain correct standard errors in this situation. Firstly, $\mathbf{I}_{\text{obs}}(\hat{\boldsymbol{\theta}})$ is replaced by its conditional expectation given the observed data and the estimated parameters, the so-called *complete information matrix*. Let $\mathbf{V}_{\text{com}}(\hat{\boldsymbol{\theta}})$ denote the variance matrix that is obtained by inverting this complete information matrix. Secondly, according to formula (9.7) in Little and Rubin (2002), the variance matrix of the incomplete-data maximum likelihood estimates is:

$$\mathbf{V}_{\text{obs}}(\hat{\boldsymbol{\theta}}) = \mathbf{V}_{\text{com}}(\hat{\boldsymbol{\theta}})(\mathbf{I} - \mathbf{DM})^{-1}.$$

Here, $\mathbf{I}$ denotes an identity matrix and $\mathbf{DM}$ is a matrix that describes the loss of information due to having incomplete data. The latter matrix is directly related to the rate of convergence of the EM algorithm. A numerical approximation to $\mathbf{DM}$ can be obtained by applying a so-called Supplemented EM algorithm; see Little and Rubin (2002, pp. 191–196) for details.

For model (7.5) with $K = 3$ observed variables, it can be derived from (7.9) that $\mathbf{V}(\hat{\boldsymbol{\theta}})$ based on complete data is a block-diagonal matrix of the following form:

$$\begin{pmatrix} \mathbf{V}(\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3) & & & & \\ & \mathbf{V}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) & & 0 & \\ & & \mathbf{V}(\hat{\tau}_1, \hat{\lambda}_1, \hat{\sigma}_1^2) & & \\ & 0 & & \mathbf{V}(\hat{\tau}_2, \hat{\lambda}_2, \hat{\sigma}_2^2) & \\ & & & & \mathbf{V}(\hat{\tau}_3, \hat{\lambda}_3, \hat{\sigma}_3^2) \end{pmatrix},$$

where

$$\mathbf{V}(\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3) = \begin{pmatrix} \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n} & 0 & 0 \\ 0 & \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n} & 0 \\ 0 & 0 & \frac{\hat{\pi}_3(1-\hat{\pi}_3)}{n} \end{pmatrix}$$

and

$$\mathbf{V}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \begin{pmatrix} \hat{\sigma}^2 \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'\right)^{-1} & \mathbf{0} \\ \mathbf{0}' & \frac{2\hat{\sigma}^4}{n} \end{pmatrix}$$

and

$$\mathbf{V}(\hat{\tau}_k, \hat{\lambda}_k, \hat{\sigma}_k^2) = \begin{pmatrix} \frac{\hat{\sigma}_k^2}{n_k \Delta_k} \sum_{i=1}^n z_{ki}\eta_i^2 & -\frac{\hat{\sigma}_k^2}{n_k \Delta_k} \sum_{i=1}^n z_{ki}\eta_i & 0 \\ -\frac{\hat{\sigma}_k^2}{n_k \Delta_k} \sum_{i=1}^n z_{ki}\eta_i & \frac{\hat{\sigma}_k^2}{\Delta_k} & 0 \\ 0 & 0 & \frac{2\hat{\sigma}_k^4}{n_k} \end{pmatrix},$$

with $\Delta_k = \sum_{i=1}^n z_{ki}\eta_i^2 - \frac{1}{n_k}\left(\sum_{i=1}^n z_{ki}\eta_i\right)^2$. The standard errors reported in Section 7.4.2 were obtained from these expressions by applying Little and Rubin's two-step procedure to correct for missing information.

Note: For the π and σ^2 parameters, standard errors based on the asymptotic normal distribution may be misleading in practice unless the sample size is very large. Little and Rubin (2002) note that a better approach is to first form confidence intervals for the transformed parameters $\log \pi/(1-\pi)$ and $\log \sigma^2$ – for which the normal approximation works better in practice – and then transform the end points of these intervals back to obtain confidence intervals for the original parameters. We also used this approach. This made little difference for the π parameters in our application, as none of the point estimates for these parameters were close to 0 or 1. The alternative approach was more useful for some of the σ^2 parameters, but these are not discussed separately in this chapter. For simplicity, we therefore report standard errors based on the direct approach.

Chapter 8

Summary, Conclusions and Discussion

8.1 Summary and conclusions

In this thesis, five chapters of original research have been presented. Chapters 3–5 focussed on new methods for automatic editing. Chapters 6 and 7 focussed on applications of measurement error models.

In Chapter 3, we looked at deductive correction methods for systematic errors. Correcting systematic errors in a separate step at the beginning of a data editing process can improve the efficiency of data editing as well as the quality of the edited data. This is true because, if a systematic error can be corrected accurately by a deductive rule, it does not have to be treated later on by a human editor or a more complex algorithm for automatic error localisation. This means that editors and more complex algorithms can focus their attention on cases with more complicated error structures, where their contribution is more likely to be worthwhile.

With the above aims in mind of improving efficiency and quality, we have developed two new deductive methods for correcting two errors that are known to occur in data of the so-called Structural Business Statistics (SBS) at Statistics Netherlands: sign errors and rounding errors. Sign errors occur for variables in a particular subsection of the questionnaire (the so-called profit-and-loss account), while rounding errors can occur throughout the data. Both methods require an algorithm that is more complex than a simple if-then rule, but they are still relatively easy and cheap to implement. Theoretical properties of the algorithms were investigated. By way of illustration, both algorithms were applied to real data from the Netherlands' SBS of 2007. For these data, we found that the deductive method for sign errors reduced the number of records with inconsistent profit-and-loss accounts by about twenty per cent. We also found that, of all records that contained

inconsistencies with respect to the edit rules, about one in five contained at least one rounding error. Moreover, by resolving these rounding errors, the number of violated balance edit rules could be reduced by about thirteen per cent. These results show that these deductive methods can achieve a substantial reduction of the amount of editing that remains to be done by editors or complex error localisation algorithms.

Chapters 4 and 5 focussed on error localisation for random errors. Two generalisations of the Fellegi-Holt paradigm were proposed that aim to improve the quality of automatically-edited data. Both generalisations address a different limitation of the Fellegi-Holt paradigm.

The starting point for Chapter 4 was the idea that some of the systematic differences that have been found between manual and automatic editing may be explained by the fact that human editors make use of soft edits as well as hard edits, whereas the Fellegi-Holt paradigm for automatic editing assumes that only hard edit rules occur. Under the Fellegi-Holt paradigm, existing soft edits have to be either ignored or treated as hard edits during automatic error localisation. We proposed a new formulation of the error localisation problem that can distinguish between hard and soft edit rules. The new approach involves solving a minimisation problem that is a generalisation of the problem of Fellegi and Holt, with an extra term that measures the extent to which soft edit rules are violated. The new problem can be solved by an extension of the existing error localisation algorithm of De Waal and Quere (2003). A simulation study was conducted with synthetic data. For these data, it was found that the new error localisation approach achieved better results than the Fellegi-Holt paradigm, both in terms of false positives (correct values that were identified as erroneous by the algorithm) and false negatives (erroneous values that were identified as correct).

The Fellegi-Holt paradigm and the underlying model based on Naus et al. (1972) tacitly assume that errors independently affect one variable at a time. By contrast, human editors often make adjustments to the data that involve more than one variable at a time. It is in fact likely that respondents often commit errors that simultaneously affect several variables. In Chapter 5 we therefore introduced a generalised error localisation problem in which the assumption is relaxed that errors affect one variable at a time. This problem is based on a new minimisation criterion which involves the number of required edit operations rather than the number of changed values. Here, each edit operation is a well-defined elementary adjustment that can be made to a record to correct one particular error, which might involve changing the values of one, two, or more variables simultaneously. We suggested to choose these edit operations such that they mimic as closely as possible

the manual corrections made by editors. The Fellegi-Holt-based error localisation problem is in fact a special case of the new problem, obtained by restricting the set of admissible edit operations to one particular class (i.e., operations that impute a new value for a single variable).

An algorithm was developed for solving the new error localisation problem. This algorithm was used in a simulation study with synthetic data to compare the new approach to Fellegi and Holt's original error localisation problem. The results of this study indicated that the new method can be used to achieve a significant improvement of the quality of automatically-edited data (again in terms of both false negatives and false positives). This does require that all (or nearly all) appropriate edit operations are included. Finding the appropriate edit operations for a given application is not trivial; we provided some suggestions on how this might be done in practice.

Turning to measurement error models, in Chapter 6 we used a structural equation model (SEM) to estimate the quality of administrative and survey data for official statistics. It was shown how both the indicator validity and intercept bias of administrative and survey variables can be estimated in this way. In particular, the indicator validity can be used as a measure to decide whether the administrative concept is sufficiently related to the true variable of interest to be of use. In cases where the validity is high but significant intercept bias occurs, a correction formula can be derived from the SEM by predicting the true value of the variable of interest from the observed value. To identify the model, we took a random subsample of our original observations and attempted to measure the true values for these units (an audit sample). The inclusion of an audit sample was necessary for the estimation of the true intercept bias and true correction formulas for the observed variables, but not for the estimation of indicator validity.

The methodology was applied to real data at Statistics Netherlands to estimate the validity and intercept bias of value-added tax (VAT) turnover for short-term statistics (monthly or quarterly statistics on the development of the economy). SEMs were fitted to linked data from three administrative sources (VAT, the Profit Declaration Register and the General Business Register) and one survey (SBS). Additional data for an audit sample were obtained by re-editing the survey data. It was found that the target variable turnover was measured with indicator validity close to 1 in all data sources. However, often the VAT data did suffer from substantial intercept bias. For cases where intercept bias occurred, a correction formula was derived from the SEM. We simulated an application of the estimated correction formulas from the SEM to publication figures for the short-term statistics. As expected, it was found that the correction hardly affected the estimated annual growth

rates but it did have a substantial effect on the estimated annual turnover levels.

Finally, in Chapter 7 we used measurement error modelling to gain insight into the quality of edited data. The indicator validity and bias of observed variables in a data set of the Netherlands' SBS before and after automatic editing were evaluated and compared. We analysed the data using two different models: an SEM and a contamination model. The latter model seemed more appropriate for the data at hand, but its current formulation does have some limitations that require further development. In our application, the effect of automatic editing on data quality in terms of validity and bias turned out to be very limited. In particular, the models suggested that the data after editing still contained a substantial amount of measurement error.

8.2 Potential applications

The results in this thesis may be applied to improve the production process for official statistics in several ways. In this section, we will mention some potential applications and discuss one of them in more detail.

The new editing methods introduced in Chapters 3–5 have been developed with the aim of improving the effectiveness of automatic editing. The underlying idea is that by improving the quality of automatically-edited data, it will be possible to change the balance between manual and automatic work in an editing process in favour of automatic editing (Pannekoek et al., 2013). By reducing the amount of manual editing, the efficiency, timeliness and reproducibility of editing processes can be increased.

Although these methods were not developed with the editing of administrative data in mind per se, they could be very useful in that context. As noted in Section 2.2.2, the size of most administrative data sets means that they cannot be processed effectively using traditional selective editing strategies. A feasible editing process for administrative data should therefore consist mostly of methods that can be automated, i.e., deductive correction rules, automatic localisation of random errors, and imputation. The main tasks of subject-matter specialists would then be to set and, if necessary, adjust the parameters of the automatic editing procedures, and to check the plausibility of the outcome of the automated process. For the latter task, selective or macro-editing methods could be used to detect and correct influential errors that may have slipped through. For economic statistics, the specialists may additionally check the data of a few very large businesses on a regular basis. In the past, editing processes of this form have been advocated and to some extent realised for survey data (Granquist and Kovar, 1997; Pannekoek et al., 2013).

Probably, they are even more relevant for the editing of administrative data.

The measurement error model discussed in Chapter 6 can be used to estimate the indicator validity of variables in linked administrative and survey data. In the presence of additional "gold standard" data for a random subsample of units in the linked data set, the method also provides an estimate of the intercept bias of each observed variable. Once the model parameters have been estimated, they can be used for various purposes.

Firstly, if different sources (administrative data sets or surveys) are available for the same variable of interest, these can be compared in terms of indicator validity and intercept bias. This situation can arise when a statistical office wants to produce new statistical output or change the input data of an existing statistical process. In particular, it is relevant for countries that are considering to switch from a survey-based to a (partly) register-based population census (Berka et al., 2012). All else being equal, the source with the validity that is closest to 1 and/or the least amount of intercept bias is to be preferred. A more quantitative approach can be obtained by using the outcome of a measurement error model to compare different proposed estimators in terms of mean squared error (MSE). This makes it possible to compare, for instance, a variable with a validity of 0.70 in an administrative source that covers almost the entire population (large measurement error, virtually no sampling error) to a variable with a validity of 0.99 in a small sample survey (large sampling error, virtually no measurement error).

Secondly, an estimated error model can be used to obtain a model-based adjustment to statistical results, to correct for the effects of measurement errors in the data. The attenuation formula for correlation coefficients that was discussed in Section 1.4 provides a well-known example. In Section 6.3.4, we applied a correction formula derived from an SEM to short-term statistics (turnover levels and growth rates), to correct for systematic bias in the observed VAT turnover values. Alternatively, predicted true values could be derived directly from the model and used as imputations instead of the originally observed values.

A model-based correction may reduce the MSE of the statistical output, but only if the model is appropriate for the application at hand and if the parameters of the model can be estimated with sufficient accuracy. This depends on the type of application. For example, when a model is used to impute predicted true values, the relation of the imputed variable to other variables may be distorted if these variables have not been included in the model (either as covariates or as other target variables). A similar problem is known to occur in the context of mass imputation (Kooiman, 1998; De Waal, 2015).

A third possible application of measurement error models was illustrated in

Chapter 7: evaluating the effect of an editing process on data quality. This information can be useful during the design of a new editing process: to compare the results of different editing approaches and to help decide how much effort should be put into each process step. It is also useful once an editing process has been set up, to check whether it is working as expected. In this case, a contamination model is more appropriate than an SEM, as it is more in line with the "intermittent-error" assumption of most data editing methods. It should be noted that an editing process may be important for other aspects of data quality that are not captured by a measurement error model, such as the fact that the edited data are consistent with a set of edit rules (see Section 7.5).

In applications like the one in Chapter 7, the model provides an estimate of the amount of measurement error that remains in the data after editing. In principle, it is therefore possible to correct statistical output for this residual measurement error, with the above-mentioned caveat that the model should be appropriate for the data and estimated with sufficient accuracy. When these conditions are satisfied, this approach could be used to either improve the quality of statistics or to reduce the costs of a statistical process while retaining the same quality. We will now discuss this in some more detail.

Currently at Statistics Netherlands, the editing process for survey data in most economic statistics consists of the following main steps (in order):

1. deductive editing of systematic errors (e.g., unit of measurement errors);
2. selective manual editing of a cut-off sample of records that are likely to contain the most important errors;
3. automatic editing of the remaining records;
4. macro-editing, and, if necessary, manual follow-up of remaining influential errors.

The methods used in these steps were reviewed in Section 2.2. The third step (automatic editing) is sometimes omitted; in that case the non-selected records are not edited.

An editing process of this form has two drawbacks. Firstly, as noted above, the process is not suitable for administrative data sets that contain a substantial number of influential errors, because it would be too costly and/or time-consuming to edit these errors manually. Secondly, the effect on statistical output of errors that remain in the data at the end of the editing process cannot be evaluated. Without further assumptions that cannot be tested in practice, it is not possible to infer the

amount of measurement error in the non-selected part of the data from the part that has been edited, because the selection was made by cut-off sampling.

A potential alternative editing process that does not have these drawbacks could be set up as follows:

1. deductive editing of systematic errors;
2. automatic editing of all[1] records;
3. macro-editing to detect a few remaining influential errors and selective manual editing of a (small) probability sample of records;
4. estimating a measurement error model for the edited data and using the estimated model parameters to evaluate the quality of intended statistical output.

In step 3, probabilistic selective editing is applied as proposed by Ilves and Laitila (2009); see also Section 2.2.1. In this case, the main goal of manual editing is not to correct all influential errors, but rather to obtain a second, improved measurement of the variables in the data set that will help us to estimate the amount of measurement error that remains in the data after automatic editing. In other words, probabilistic selective editing is used to obtain an audit sample.

In step 4, a measurement error model is fitted to the automatically-edited data, with the manually-edited variables included for the cases where they are available. To obtain an identified model, the edited data should be linked to auxiliary data from an independent administrative source. If an SEM like the one in Chapter 6 is used in this step, then the inclusion of an audit sample of manually-edited "gold standard" data is required for model identification. If a contamination model is used then, as noted in Chapter 7, an audit sample is not strictly necessary. However, it could still be useful to include these data to improve the accuracy of parameter estimates and to provide an opportunity for testing some of the model assumptions.

Having obtained an estimate of the amount of residual measurement error after automatic editing (in terms of error probability, intercept bias and indicator validity), we can then evaluate the effect of these remaining errors on intended statistical output. If the effect is large, we may go back and perform some additional macro-editing, or we may apply a model-based adjustment to the output to correct for measurement error.

This alternative editing process has not yet been tested. A question that remains to be investigated is whether sufficiently accurate estimates can be obtained of

[1] In practice, it is likely that a few records cannot be edited automatically due to computational problems. These records could be added to the set that is edited manually, or they could even be treated as non-response.

the parameters of the measurement error model, based on a realistic number of manually-edited cases. This is particularly important if a model-based correction formula is to be used for the statistical output. Furthermore, this approach seems to be feasible mainly in cases where the statistical output consists of one or a few target variables (e.g., the short-term statistics on turnover). It is not immediately clear how it could be applied, for instance, to the structural business statistics which contain dozens of variables.

8.3 Discussion

Over the past years, administrative data have increasingly been used in official statistics and academic research to replace traditional surveys. This trend is likely to continue in the future. It is important to realise that administrative data nearly always suffer from measurement errors. In this sense, they are just like surveys. In fact, the problem of measurement errors may be even more important for administrative data than for surveys because administrative concepts can differ from statistical concepts.

The methods that have been discussed and developed in this thesis can be applied to address the problem of measurement errors in both administrative data and survey data. We have looked both at methods that try to correct individual errors (*editing*) and methods that try to model the overall effect of measurement errors (*estimation*). We will now summarise the main contributions of this thesis in relation to the three points that were mentioned in Section 1.1. We will also point out some topics for future research.

1. Increasing the usefulness of methods for automatic editing
For the editing approach, we have worked on automatic editing methods. Currently, methods for automatic editing of random errors are used only to a limited extent at Statistics Netherlands – and even less at most other national statistical institutes – because the underlying assumptions of these methods are rather restrictive. In particular, the Fellegi-Holt paradigm for finding random errors is based on an implicit measurement error model that assumes that errors affect one variable at a time, that they are independent across variables and that the probability of observing a variable in error does not depend on the underlying true value (see Section 2.4). Furthermore, all subject-matter information that is relevant for finding these errors is supposed to be available in the form of hard edit rules. These assumptions are rarely, if ever, satisfied in practice.

The new methods for automatic error localisation that were developed in Chap-

ters 4 and 5 are more flexible than existing methods. The method of Chapter 4 can handle additional subject-matter information in the form of soft edit rules. The method of Chapter 5 can handle a more general class of errors than the original Fellegi-Holt paradigm that includes errors that affect multiple variables simultaneously. This added flexibility can be exploited to bring the results of automatic editing closer to those of manual editing. This means that NSIs could use these methods to reduce the amount of manual work in data editing processes and increase the amount of automatic editing. This will increase the efficiency and timeliness of statistical production processes. Furthermore, the attention of subject-matter specialists could then be focussed more on editing the most difficult cases, which will improve the quality of the edited data and hence the quality of statistical output.

The added flexibility of these new methods for automatic editing also means that they are more complicated to set up in practice. For instance, the formulation of the generalised error localisation problem in Chapter 5 involves a set of "admissible edit operations" which has to be chosen for each application, and a set of weights that have to be assigned to these operations. To realise the potential of these new editing methods, more research is therefore needed to test these methods on realistic data and to develop simple recommendations for their use in practice.

Some of the work in Chapters 4 and 5 was based on existing ideas. The formulation of the error localisation problem that was proposed in Chapter 4 makes use of a well-known technique in mathematical optimisation (accounting for soft restrictions by adding a term to the target function). The generalised error localisation problem in Chapter 5 has a strong similarity to the so-called Levenshtein distance that is used for approximate string matching. As far as we are aware, our application of these ideas to statistical data editing is new. Moreover, the context of error localisation in statistical data has some specific features so that we had to develop new algorithms for solving these error localisation problems. For instance, the requirement in Chapter 5 that the edited record must satisfy a set of hard edit rules has no counterpart in standard applications of the Levenshtein distance. Nevertheless, an interesting topic for future research could be to improve the efficiency of the algorithms that we have proposed here by adapting results from the fields of mathematical optimisation with soft restrictions (Chapter 4) and distance computation in approximate string matching (Chapter 5).

2. Constructing a measurement error model that is useful for official statistics
For the estimation approach, we noted in Chapter 1 that applications in official statistics often involve target parameters such as population means and totals of "factual" variables, for which true values rather than true scores (as defined in Sec-

tion 1.2.3) are of interest. For these applications, it is important to estimate the relation between the scale of an observed variable and the true scale of the underlying (latent) target variable. By contrast, most social-science applications deal with bivariate and multivariate statistics (e.g., correlations) about "non-factual" phenomena, for which it is sufficient to model the latent variables on a standardised scale or on an arbitrarily chosen reference scale. Because of this gap, some useful modelling techniques that have been developed in the social and behavioural sciences are currently not applied in official statistics as often as they could be. This holds in particular for the use of SEMs to account for measurement errors in observed variables.

In Chapter 6, we used an SEM to estimate the effects of measurement errors in linked administrative and survey data on several factual variables. To identify the true scales of the latent variables in this SEM, we introduced the assumption that "gold standard" observations could be obtained for a random subsample of the original data set (an audit sample). This assumption was suggested previously by Sobel and Arminger (1986). For the application in Chapter 6, we generalised the method of Sobel and Arminger to allow for complex sample designs and non-normality of the data by using Pseudo Maximum Likelihood (PML). Although the PML method for estimating SEMs under complex sampling and non-normal data is well established, its application in combination with audit data that are missing by design required some non-trivial adjustments to the method of Sobel and Arminger (1986). As far as we are aware, the combined method as described in Chapter 6 is new. This method allows for a wider use of SEMs in official statistics. Such models can be useful in official statistics to assess the suitability of administrative data and other new data sources, as illustrated by our application in Chapter 6. Several other potential uses of measurement error models in official statistics were described in Section 8.2, which shows that they can be used to evaluate the accuracy of statistics and to obtain statistical output of higher quality.

The SEM of Chapter 6 is an instance of a congeneric measures design. As was shown in Section 2.3, this type of model can be used to estimate the indicator validity and – provided that an audit sample is included – the intercept bias of observed variables, but not their true-score validity or reliability. If the latter parameters are of interest, more complicated designs such as the multitrait-multimethod (MTMM) design should be used. These designs were originally developed for survey data. In Section 2.3.4, we discussed several possible approaches to identify and estimate an MTMM design for administrative data. For future research, it would be interesting to test these approaches in practice.

3. Using a measurement error model to evaluate the effects of automatic editing
The traditional approach to evaluate the quality of automatic editing works by comparing automatically-edited data to manually-edited data under the assumption that the latter data are error-free. This assumption is not realistic. In Chapter 7, we have shown that the effects of automatic editing on the amount of measurement error in survey data can be evaluated without making this assumption, again by modelling the errors in a linked data set of administrative and survey variables. The results of this application suggested that a contamination model may be more appropriate than an SEM for modelling measurement errors in most of the sources that were examined in this thesis (both survey and administrative data), as these all appear to contain a non-negligible fraction of observations without errors. An additional advantage of the contamination model is that this model can be identified without an audit sample even when true values are of interest.

The contamination model that was estimated in Chapter 7 is a generalisation of the model of Guarnera and Varriale (2016). Our generalisation accounts for a possible bias due to a difference in scale between each observed variable and the underlying target variable. In the experience of Statistics Netherlands so far, such differences in scale often occur for administrative variables. Before this model could be applied in practice at Statistics Netherlands, more work is needed. Firstly, the model should be extended to more than one target variable. Secondly, the sensitivity of the model should be investigated to violations of the assumption that the errors in different observed variables are independent, and possible extensions of the model should be developed that avoid this assumption. Thirdly, an estimation procedure should be developed for this model that is robust to non-normal data and/or data that arise from complex sampling designs. For this, it may be possible to develop a variation of the PML method that is used for SEMs. In fact, the ultimate outcome of the above refinements may be a model that is a "contaminated SEM", i.e., a structural equation model that includes a distinction between observations with and without errors on the observed variables.

As the potential applications discussed in Section 8.2 show, both the editing and estimation approaches can be useful for official statistics. In fact, we have argued in Section 2.5 that an approach that combines the two approaches (as suggested in Section 8.2) may lead to a production process for official statistics that is more efficient and timely than the current practice and also produces better statistical output. Moreover, such a combined approach may be the only effective way to handle measurement errors in large administrative data sets. It is clear that more research is needed before this combined approach could be applied in practice. We hope that the contents of this thesis will provide inspiration for this research.

Bibliography

Agrawal, R. and R. Srikant (1994). Fast Algorithms for Mining Association Rules. Technical report, IBM Almaden Research Center, San Jose, California.

Al-Hamad, A., D. Lewis, and P. L. N. Silva (2008). Assessing the Performance of the Thousand Pounds Automatic Editing Procedure at the Office for National Statistics and the Need for an Alternative Approach. Working Paper No. 21, UN/ECE Work Session on Statistical Data Editing, Vienna.

Allison, P. D. (1987). Estimation of Linear Models with Incomplete Data. *Sociological Methodology 17*, 71–103.

Alwin, D. F. (2007). *Margins of Errors*. New York: John Wiley & Sons.

Andrews, F. M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. *Public Opinion Quarterly 48*, 409–442.

Arbués, I., P. Revilla, and D. Salgado (2013). An Optimization Approach to Selective Editing. *Journal of Official Statistics 29*, 489–510.

Bakker, B. F. M. (2011a). Micro-Integration. Method Series, Statistics Netherlands, The Hague.

Bakker, B. F. M. (2011b). Micro-Integration: State of the Art. In *ESSnet on Data Integration, Report on WP1*, pp. 77–107.

Bakker, B. F. M. (2012). Estimating the Validity of Administrative Variables. *Statistica Neerlandica 66*, 8–17.

Bakker, B. F. M. and P. J. H. Daas (2012). Methodological Challenges of Register-Based Research. *Statistica Neerlandica 66*, 2–7.

Bakker, B. F. M. and L. Kuijvenhoven (2010). Registers en Sociaalwetenschappelijk Onderzoek: een Geslaagde Combinatie? In Bakker and Kuijvenhoven (Eds.), *Registers in Sociaalwetenschappelijk Onderzoek – Mogelijkheden en Valkuilen*, pp. 7–14. Statistics Netherlands/Vrije Universiteit. In Dutch.

Banff Support Team (2003). Functional Description of the Banff System for Edit and Imputation. Technical report, Statistics Canada.

Bankier, M. and S. Crowe (2009). Enhancements to the 2011 Canadian Census E&I System. Working Paper No. 15, UN/ECE Work Session on Statistical Data Editing, Neuchâtel.

Bankier, M., M. Lachance, and P. Poirier (2000). 2001 Canadian Census Minimum Change Donor Imputation Methodology. Working Paper No. 17, UN/ECE Work Session on Statistical Data Editing, Cardiff.

Bassi, F., J. A. Hagenaars, M. A. Croon, and J. K. Vermunt (2000). Estimating True Changes when Categorical Panel Data are Affected by Uncorrelated and Correlated Classification Errors. *Sociological Methods and Research 29*, 230–268.

Baumgartner, H. and J.-B. E. M. Steenkamp (1998). Multi-Group Latent Variable Models for Varying Numbers of Items and Factors with Cross-National and Longitudinal Applications. *Marketing Letters 9*, 21–35.

Bavdaž, M. (2010). Sources of Measurement Error in Business Surveys. *Journal of Official Statistics 26*, 25–42.

Berka, C., S. Humer, M. Moser, M. Lenk, H. Rechta, and E. Schwerer (2012). Combination of Evidence from Multiple Administrative Data Sources: Quality Assessment of the Austrian Register-Based Census 2011. *Statistica Neerlandica 66*, 18–33.

Bethlehem, J. (2008). Surveys without Questions. In De Leeuw, Hox, and Dillman (Eds.), *International Handbook of Survey Methodology*, pp. 500–511. New York: Psychology Press.

Bethlehem, J. (2009). *Applied Survey Methods: A Statistical Perspective*. Hoboken, New Jersey: John Wiley & Sons.

Bethlehem, J., F. Cobben, and B. Schouten (2011). *Handbook of Nonresponse in Household Surveys*. Hoboken, New Jersey: John Wiley & Sons.

Bielby, W. T. (1986a). Arbitrary Metrics in Multiple-Indicator Models of Latent Variables. *Sociological Methods and Research 15*, 3–23.

Bielby, W. T. (1986b). Arbitrary Normalizations; Comments on Issues Raised by Sobel, Arminger, and Henry. *Sociological Methods and Research 15*, 62–63.

Biemer, P. (2004). Modeling Measurement Error to Identify Flawed Questions. In Presser, Rothgeb, Couper, Lessler, Martin, Martin, and Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires*, pp. 225–246. Hoboken, New Jersey: John Wiley & Sons.

Biemer, P. and S. L. Stokes (1991). Approaches to the Modeling of Measurement Error. In Biemer, Groves, Lyberg, Mathiowetz, and Sudman (Eds.), *Measurement Errors in Surveys*, pp. 487–516. New York: John Wiley & Sons.

Biemer, P. P. (2011). *Latent Class Analysis of Survey Error*. Hoboken, New Jersey: John Wiley & Sons.

Biemer, P. P. and L. E. Lyberg (2003). *Introduction to Survey Quality*. Hoboken, New Jersey: John Wiley & Sons.

Bikker, R. (2003). Evaluatie Automatisch versus Handmatig Gaafmaken van Productiestatistieken 2000 Handel & Transport – Aanvullende Verklaringen. Internal report 1900-03-TMO (in Dutch), Statistics Netherlands, Voorburg.

Boeschoten, L., D. Oberski, and T. De Waal (2016). Estimating Classification Error under Edit Restrictions in Combined Survey-Register Data. Discussion Paper 2016-12, Statistics Netherlands, The Hague.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Boomsma, A. (1982). The Robustness of LISREL against Small Sample Sizes in Factor Analysis Models. In Jöreskog and Wold (Eds.), *Systems under Indirect Observation*, Volume I, pp. 149–173. Amsterdam: North-Holland Publishing Company.

Borsboom, D., G. J. Mellenbergh, and J. Van Heerden (2003). The Theoretical Status of Latent Variables. *Psychological Review 110*, 203–219.

Boskovitz, A., R. Goré, and P. Wong (2005). Data Editing and Logic. Working Paper No. 33, UN/ECE Work Session on Statistical Data Editing, Ottawa.

Bound, J., C. Brown, and N. Mathiowetz (2001). Measurement Error in Survey Data. In Heckman and Leamer (Eds.), *Handbook of Econometrics*, Volume 5, pp. 3705–3843. Amsterdam: Elsevier.

Brancato, G., S. Macchia, M. Murgia, M. Signore, G. Simeoni, K. Blanke, T. Körner, A. Nimmergut, P. Lima, R. Paulino, and J. H. P. Hoffmeyer-Zlotnik

(2006). *Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System*. Luxembourg: Eurostat.

Bruni, R. (2004). Discrete Models for Data Imputation. *Discrete Applied Mathematics 144*, 59–69.

Bruni, R. (2005). Error Correction for Massive Datasets. *Optimization Methods and Software 20*, 297–316.

Campbell, D. T. and D. W. Fiske (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin 56*, 81–105.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective* (Second ed.). Boca Raton: Chapman & Hall/CRC.

Casado Valero, C., F. Del Castillo Cuervo-Arango, J. Mateo Ayerra, and A. De Santos Ballesteros (1996). Quantitative Data Editing: Quadratic Programming Method. Presented at the COMPSTAT 1996 Conference, Barcelona.

Černikov, S. N. (1963). The Solution of Linear Programming Problem by Elimination of Unknowns. In *Soviet Mathematics Doklady 2*.

Chen, B., Y. Thibaudeau, and W. E. Winkler (2003). A Comparison Study of ACS If-Then-Else, NIM, DISCRETE Edit and Imputation Systems using ACS Data. Working Paper No. 7, UN/ECE Work Session on Statistical Data Editing, Madrid.

Cochran, W. G. (1977). *Sampling Techniques* (Third ed.). New York: John Wiley & Sons.

Daas, P., S. Ossen, M. Tennekes, L.-C. Zhang, C. Hendriks, K. Foldal Haugen, A. Bernardi, F. Cerroni, T. Laitila, A. Wallgren, and B. Wallgren (2011). List of Quality Groups and Indicators Identified for Administrative Data Sources. BLUE-ETS Project, Deliverable 4.1. Available at: http://www.blue-ets.istat.it/.

De Jong, A. (2002). Uni-Edit: Standarized Processing of Structural Business Statistics in The Netherlands. Working Paper No. 27, UN/ECE Work Session on Statistical Data Editing, Helsinki.

De Jonge, E. and M. Van der Loo (2011). Manipulation of Linear Edits and Error Localization with the Editrules Package. Discussion Paper 201120, Statistics Netherlands, The Hague.

De Jonge, E. and M. Van der Loo (2014). Error Localization as a Mixed Integer Problem with the Editrules Package. Discussion Paper 2014-07, Statistics Netherlands, The Hague.

De Leeuw, E. D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics 21*, 233–255.

De Leeuw, E. D., J. J. Hox, and D. A. Dillman (2008). *International Handbook of Survey Methodology*. New York: Psychology Press.

De Waal, T. (2002). Algorithms for Automatic Error Localisation and Modification. Paper prepared for the DATACLEAN 2002 Conference, Jyväskylä.

De Waal, T. (2003a). A Simple Branching Scheme for Solving the Error Localisation Problem. Discussion Paper 03010, Statistics Netherlands, Voorburg.

De Waal, T. (2003b). Processing of Erroneous and Unsafe Data. PhD Thesis, Erasmus University, Rotterdam.

De Waal, T. (2003c). Solving the Error Localization Problem by Means of Vertex Generation. *Survey Methodology 29*, 71–79.

De Waal, T. (2005). SLICE 1.5: A Software Framework for Automatic Edit and Imputation. Working Paper No. 39, UN/ECE Work Session on Statistical Data Editing, Ottawa.

De Waal, T. (2015). General Approaches for Consistent Estimation based on Administrative Data and Surveys. Discussion Paper 2015-11, Statistics Netherlands, The Hague.

De Waal, T. and W. Coutinho (2005). Automatic Editing for Business Surveys: An Assessment for Selected Algorithms. *International Statistical Review 73*, 73–102.

De Waal, T., J. Pannekoek, and S. Scholtus (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, New Jersey: John Wiley & Sons.

De Waal, T. and R. Quere (2003). A Fast and Simple Algorithm for Automatic Editing of Mixed Data. *Journal of Official Statistics 19*, 383–402.

De Waal, T. and S. Scholtus (2011). Methods for Automatic Statistical Data Editing. Paper presented at the 2011 KSS International Conference on Statistics and Probability, Busan.

Deng, L. and K.-H. Yuan (2015). Multiple-Group Analysis for Structural Equation Modeling with Dependent Samples. *Structural Equation Modeling 22*, 552–567.

Di Zio, M. and U. Guarnera (2013). A Contamination Model for Selective Editing. *Journal of Official Statistics 29*, 539–555.

Di Zio, M. and U. Guarnera (2014). Theme: Selective Editing. In *MEMOBUST Handbook on Methodology for Modern Business Statistics*. Luxembourg: Eurostat.

Di Zio, M., U. Guarnera, and O. Luzi (2005a). Editing Systematic Unity Measure Errors through Mixture Modelling. *Survey Methodology 31*, 53–63.

Di Zio, M., U. Guarnera, and O. Luzi (2005b). Improving the Effectiveness of a Probabilistic Editing Strategy for Business Data. Report, ISTAT, Rome.

Di Zio, M., U. Guarnera, and R. Rocci (2007). A Mixture of Mixture Models for a Classification Problem: The Unity Measure Error. *Computational Statistics & Data Analysis 51*, 2573–2585.

Di Zio, M. and O. Luzi (2014). Theme: Editing Administrative Data. In *MEMOBUST Handbook on Methodology for Modern Business Statistics*. Luxembourg: Eurostat.

Durbin, J. (1954). Errors in Variables. *Review of the International Statistical Institute 22*, 23–32.

EDIMBUS (2007). *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. Eurostat manual prepared by ISTAT, Statistics Netherlands, and SFSO.

European Commission (2006). Commission Regulation (EC) No 1503/2006 of 28 September 2006 implementing and amending Council Regulation (EC) No 1165/98 concerning short-term statistics as regards definitions of variables, list of variables and frequency of data compilation. Published in the Official Journal of the European Union L281, 12 October 2006, pp. 15–30.

Federal Committee on Statistical Methodology (1990). Data Editing in Federal Statistical Agencies. Statistical Policy Working Paper 18, U.S. Office of Management and Budget, Washington, DC.

Fellegi, I. P. and D. Holt (1976). A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association 71*, 17–35.

Fraleigh, J. B. and R. A. Beauregard (1995). *Linear Algebra* (Third ed.). Reading, MA: Addison-Wesley.

Freund, R. J. and H. O. Hartley (1967). A Procedure for Automatic Data Editing. *Journal of the American Statistical Association 62*, 341–352.

Fuller, W. A. (1987). *Measurement Error Models*. New York: John Wiley & Sons.

Garfinkel, R. S., A. S. Kunnathur, and G. E. Liepins (1986). Optimal Imputation of Erroneous Data: Categorical Data, General Edits. *Operations Research 34*, 744–751.

Garfinkel, R. S., A. S. Kunnathur, and G. E. Liepins (1988). Error Localization for Erroneous Data: Continuous Data, Linear Constraints. *SIAM Journal on Scientific and Statistical Computing 9*, 922–931.

Ghosh-Dastidar, B. and J. L. Schafer (2006). Outlier Detection and Editing Procedures for Continuous Multivariate Data. *Journal of Official Statistics 22*, 487–506.

Giesen, D. (2007). Does Mode Matter? First Results of the Comparison of the Response Burden and Data Quality of a Paper Business Survey and an Electronic Business Survey. Paper presented at QUEST 2007, 24-26 April, Ottawa.

Giles, P. (1988). A Model for Generalized Edit and Imputation of Survey Data. *The Canadian Journal of Statistics 16*, 57–73.

Golub, G. H. and C. F. Van Loan (1996). *Matrix Computations* (Third ed.). Baltimore: The Johns Hopkins University Press.

Granquist, L. (1990). A Review of Some Macro-Editing Methods for Rationalizing the Editing Process. In *Proceedings of the Statistics Canada Symposium*, pp. 225–234.

Granquist, L. (1995). Improving the Traditional Editing Process. In Cox, Binder, Chinnappa, Christianson, Colledge, and Kott (Eds.), *Business Survey Methods*, pp. 385–401. John Wiley & Sons.

Granquist, L. (1997). The New View on Editing. *International Statistical Review 65*, 381–387.

Granquist, L. and J. Kovar (1997). Editing of Survey Data: How Much is Enough? In Lyberg, Biemer, Collins, De Leeuw, Dippo, Schwartz, and Trewin (Eds.), *Survey Measurement and Process Quality*, pp. 415–435. John Wiley & Sons.

Groen, J. A. (2012). Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures. *Journal of Official Statistics 28*, 173–198.

Groves, R. M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons.

Groves, R. M., F. J. Fowler, Jr., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau (2009). *Survey Methodology* (Second ed.). New York: John Wiley & Sons.

Guarnera, U. and R. Varriale (2015). Estimation and Editing for Data from Different Sources. An Approach Based on Latent Class Model. Working Paper No. 32, UN/ECE Work Session on Statistical Data Editing, Budapest.

Guarnera, U. and R. Varriale (2016). Estimation from Contaminated Multi-Source Data Based on Latent Class Models. *Statistical Journal of the IAOS 32*, 537–544.

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. New York: Springer-Verlag.

Hedlin, D. (2003). Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. *Journal of Official Statistics 19*, 177–199.

Heise, D. (1969). Separating Reliability and Stability in Test-Retest Correlation. *American Sociological Review 34*, 93–101.

Heise, D. R. and G. W. Bohrnstedt (1970). Validity, Invalidity, and Reliability. In Borgatta and Bohrnstedt (Eds.), *Sociological Methodology*, pp. 104–129. San Francisco: Jossey Bass.

Hidiroglou, M. A. and J.-M. Berthelot (1986). Statistical Editing and Imputation for Periodic Business Surveys. *Survey Methodology 12*, 73–83.

Hoogland, J. (2006). Selective Editing using Plausibility Indicators and SLICE. In *Statistical Data Editing, Volume No. 3, Impact on Data Quality*, pp. 106–130. New York and Geneva: United Nations.

Hoogland, J. and R. Smit (2008). Selective Automatic Editing of Mixed Mode Questionnaires for Structural Business Statistics. Working Paper No. 2, UN/ECE Work Session on Statistical Data Editing, Vienna.

Ilves, M. and T. Laitila (2009). Probability-Sampling Approach to Editing. *Austrian Journal of Statistics 38*, 171–182.

Jöreskog, K. G. (1971). Statistical Analysis of Sets of Congeneric Tests. *Psychometrika 36*, 109–133.

Kenny, D. A. (1976). An Empirical Application of Confirmatory Factor Analysis to the Multitrait-Multimethod Matrix. *Journal of Experimental Social Psychology 12*, 247–252.

Knottnerus, P. (2003). *Sample Survey Theory: Some Pythagorean Perspectives*. New York: Springer-Verlag.

Kohler, D. A. (1967). Projections of Convex Polyhedral Sets. Operational Research Center Report ORC 67-29, University of California, Berkeley.

Költringer, R. (1990). Analysis of Multitrait Multimethod Matrices. In Saris and Van Meurs (Eds.), *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait-Multimethod Studies*, pp. 81–92. Amsterdam: North-Holland.

Kooiman, P. (1998). Massa-imputatie: waarom niet!? Report 8792-98-RSM (in Dutch), Statistics Netherlands, Voorburg.

Kovar, J. and P. Whitridge (1990). Generalized Edit and Imputation System; Overview and Applications. *Revista Brasileira de Estadistica 51*, 85–100.

Kruskal, J. B. (1983). An Overview of Sequence Comparison. In Sankoff and Kruskal (Eds.), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, pp. 1–44. Addison-Wesley.

Langeheine, R. (1994). Latent Variable Markov Models. In Von Eye and Clogg (Eds.), *Latent Variables Analysis: Applications to Developmental Research*, pp. 373–395. Thousand Oaks: SAGE Publications.

Lawrence, D. and R. McKenzie (2000). The General Application of Significance Editing. *Journal of Official Statistics 16*, 243–253.

Liepins, G. E. (1980). A Rigorous, Systematic Approach to Automatic Data Editing and its Statistical Basis. Report ORNL/TM-7126, Oak Ridge National Laboratory.

Liepins, G. E., R. S. Garfinkel, and A. S. Kunnathur (1982). Error Localization for Erroneous Data: A Survey. *TIMS/Studies in the Management Sciences 19*, 205–219.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (Second ed.). New York: John Wiley & Sons.

Little, R. J. A. and P. J. Smith (1987). Editing and Imputation of Quantitative Survey Data. *Journal of the American Statistical Association 82*, 58–68.

Little, T. D., D. W. Slegers, and N. A. Card (2006). A Non-Arbitrary Method of Identifying and Scaling Latent Variables in SEM and MACS Models. *Structural Equation Modeling 13*, 59–72.

Lord, F. M. and M. R. Novick (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software 9*, 1–19.

Magnus, J. R., J. W. Van Tongeren, and A. F. De Vos (2000). National Accounts Estimation using Indicator Ratios. *Review of Income and Wealth 46*, 329–350.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: John Wiley & Sons.

Meijer, E., S. Rohwedder, and T. Wansbeek (2012). Measurement Error in Earnings Data: Using a Mixture Model Approach to Combine Survey and Register Data. *Journal of Business & Economic Statistics 30*, 191–201.

Mirsky, L. (1971). *Transversal Theory*. New York: Academic Press, Inc.

Mittag, N. (2013). A Method of Correcting for Misreporting Applied to the Food Stamp Program. Discussion Paper CES 13-28, U.S. Census Bureau, Center for Economic Studies, Washington, DC.

Muthén, B. O. and A. Satorra (1995). Complex Sample Data in Structural Equation Modeling. *Sociological Methodology 25*, 267–316.

Naus, J. I., T. G. Johnson, and R. Montalvo (1972). A Probabilistic Model for Identifying Errors in Data Editing. *Journal of the American Statistical Association 67*, 943–950.

Navarro, G. (2001). A Guided Tour to Approximate String Matching. *ACM Computing Surveys 33*, 31–88.

Nemhauser, G. L. and L. A. Wolsey (1988). *Integer and Combinatorial Optimization*. New York: John Wiley & Sons.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection (with discussion). *Journal of the Royal Statistical Society 97*, 558–625.

Norberg, A. (2016). SELEKT – A Generic Tool for Selective Editing. *Journal of Official Statistics 32*, 209–229.

Nordbotten, S. (1955). Measuring the Error of Editing the Questionnaires in a Census. *Journal of the American Statistical Association 50*, 364–369.

Nordbotten, S. (1963). Automatic Editing of Individual Statistical Observations. In *Conference of European Statisticians – Statistical Standards and Studies No. 2*. New York: United Nations.

Oberski, D. L. (2014). lavaan.survey: An R Package for Complex Survey Analysis of Structural Equation Models. *Journal of Statistical Software 57*, 1–27.

Oberski, D. L. (2017). Estimating Error Rates in an Administrative Register and Survey Questions Using a Latent Class Model. In Biemer, De Leeuw, Eckman, Edwards, Kreuter, Lyberg, Tucker, and West (Eds.), *Total Survey Error in Practice*. New York: John Wiley & Sons.

Oberski, D. L., A. Kirchner, S. Eckman, and F. Kreuter (2017). Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models. *Journal of the American Statistical Association*. Accepted for publication.

Palomo, J., D. B. Dunson, and K. Bollen (2007). Bayesian Structural Equation Modeling. In Lee (Ed.), *Handbook of Latent Variable and Related Models*, pp. 163–188. Amsterdam: Elsevier.

Pannekoek, J. and T. De Waal (2005). Automatic Edit and Imputation for Business Surveys: The Dutch Contribution to the EUREDIT Project. *Journal of Official Statistics 21*, 257–286.

Pannekoek, J., S. Scholtus, and M. Van der Loo (2013). Automated and Manual Data Editing: A View on Process Design and Methodology. *Journal of Official Statistics 29*, 511–537.

Papadopoulos, S. and Y. Amemiya (2005). Correlated Samples with Fixed and Nonnormal Latent Variables. *The Annals of Statistics 33*, 2732–2757.

Pavlopoulos, D. and J. K. Vermunt (2015). Measuring Temporary Employment. Do Survey or Register Data Tell the Truth? *Survey Methodology 41*, 197–214.

Presser, S., J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer (2004). *Methods for Testing and Evaluating Survey Questionnaires*. New York: John Wiley & Sons.

Pugh, W. (1992). The Omega Test: A Fast and Practical Integer Programming Algorithm for Data Dependence Analysis. *Communications of the ACM 35*, 102–114.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: http://www.R-project.org/.

Ragsdale, C. T. and P. G. McKeown (1996). On Solving the Continuous Data Editing Problem. *Computers & Operations Research 23*, 263–273.

Revilla, M. and W. E. Saris (2013). The Split-Ballot Multitrait-Multimethod Approach: Implementation and Problems. *Structural Equation Modeling 20*, 27–46.

Riera-Ledesma, J. and J. J. Salazar-González (2003). New Algorithms for the Editing and Imputation Problem. Working Paper No. 5, UN/ECE Work Session on Statistical Data Editing, Madrid.

Riera-Ledesma, J. and J. J. Salazar-González (2007). A Branch-and-Cut Algorithm for the Continuous Error Localization Problem in Data Cleaning. *Computers & Operations Research 34*, 2790–2804.

Robinson, S. P. (2016). Modelling Measurement Errors in Linked Administrative and Survey Data. Master Thesis, Leiden University.

Rodgers, W. L. (1989). Reliability and Validity in Measures of Subjective Well-Being. Paper presented at the International Conference on Social Reporting, Berlin.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software 48*, 1–36.

Salazar-González, J. J., P. Lowthian, C. Young, G. Merola, S. Bond, and D. Brown (2004). Getting the Best Results in Controlled Rounding with the Least Effort. In Domingo-Ferrer and Torra (Eds.), *Privacy in Statistical Databases*, pp. 58–72. Berlin: Springer-Verlag.

Sande, G. (1978). An Algorithm for the Fields to Impute Problems of Numerical and Coded Data. Technical report, Statistics Canada.

Saris, W. E. (1990a). Models for Evaluation of Measurement Instruments. In Saris and Van Meurs (Eds.), *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait-Multimethod Studies*, pp. 52–80. Amsterdam: North-Holland.

Saris, W. E. (1990b). The Choice of a Research Design for MTMM Studies. In Saris and Van Meurs (Eds.), *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait-Multimethod Studies*, pp. 160–167. Amsterdam: North-Holland.

Saris, W. E. and F. M. Andrews (1991). Evaluation of Measurement Instruments Using a Structural Modeling Approach. In Biemer, Groves, Lyberg, Mathiowetz, and Sudman (Eds.), *Measurement Errors in Surveys*, pp. 575–597. New York: John Wiley & Sons.

Saris, W. E. and I. N. Gallhofer (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York: John Wiley & Sons.

Saris, W. E. and A. Münnich (1995). *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments*. Budapest: Eötvös University Press.

Saris, W. E., D. Oberski, M. Revilla, D. Zavala, L. Lilleoja, I. Gallhofer, and T. Gruner (2011). The Development of the Program SQP 2.0 for the Prediction of the Quality of Survey Questions. RECSM Working Paper Number 24, Universitat Pompeu Fabra, Barcelona.

Saris, W. E., A. Satorra, and G. Coenders (2004). A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design. *Sociological Methodology 34*, 311–347.

Sarle, W. S. (1997). Measurement Theory: Frequently Asked Questions. Report (version 3), SAS Institute Inc., Cary, NC.

Särndal, C.-E., B. Swensson, and J. H. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Satorra, A. (1992). Asymptotic Robust Inferences in the Analysis of Mean and Covariance Structures. *Sociological Methodology 22*, 249–278.

Satorra, A. (2002). Asymptotic Robustness in Multiple Group Linear-Latent Variable Models. *Econometric Theory 18*, 297–312.

Satorra, A. and P. M. Bentler (1986). Some Robustness Issues of Goodness of Fit Statistics in Covariance Structure Analysis. In *ASA 1986 Proceedings of the Business and Economic Statistics Section*, pp. 549–554.

Satorra, A. and P. M. Bentler (1994). Corrections to Test Statistics and Standard Errors in Covariance Structure Analysis. In Von Eye and Clogg (Eds.), *Latent Variables Analysis: Applications to Developmental Research*, pp. 399–419. Thousand Oaks: SAGE Publications.

Schaffer, J. (1987). Procedure for Solving the Data-Editing Problem with Both Continuous and Discrete Data Types. *Naval Research Logistics 34*, 879–890.

Scherpenzeel, A. C. (1995). A Question of Quality: Evaluating Survey Questions by Multitrait-Multimethod Studies. PhD Thesis, University of Amsterdam.

Scherpenzeel, A. C. and W. E. Saris (1997). The Validity and Reliability of Survey Questions: A Meta-Analysis of MTMM Studies. *Sociological Methods and Research 25*, 341–383.

Scholtus, S. (2008). Algorithms for Correcting Some Obvious Inconsistencies and Rounding Errors in Business Survey Data. Discussion Paper 08015, Statistics Netherlands, The Hague.

Scholtus, S. (2009). Automatic Correction of Simple Typing Errors in Numerical Data with Balance Edits. Discussion Paper 09046, Statistics Netherlands, The Hague.

Scholtus, S. (2011a). Algorithms for Correcting Sign Errors and Rounding Errors in Business Survey Data. *Journal of Official Statistics 27*, 467–490.

Scholtus, S. (2011b). Automatic Editing with Soft Edits. Discussion Paper 201130, Statistics Netherlands, The Hague.

Scholtus, S. (2013). Automatic Editing with Hard and Soft Edits. *Survey Methodology 39*, 59–89.

Scholtus, S. (2014a). Error Localisation using General Edit Operations. Discussion Paper 2014-14, Statistics Netherlands, The Hague.

Scholtus, S. (2014b). Explicit and Implicit Calibration of Covariance and Mean Structures. Discussion Paper 2014-09, Statistics Netherlands, The Hague.

Scholtus, S. (2014c). Method: Manual Editing. In *MEMOBUST Handbook on Methodology for Modern Business Statistics*. Luxembourg: Eurostat.

Scholtus, S. (2015). New Results on Automatic Editing using Hard and Soft Edit Rules. Working Paper No. 35, UN/ECE Work Session on Statistical Data Editing, Budapest.

Scholtus, S. (2016). A Generalized Fellegi-Holt Paradigm for Automatic Error Localization. *Survey Methodology 42*, 1–18.

Scholtus, S. and B. F. M. Bakker (2013a). Estimating the Validity of Administrative and Survey Variables by means of Structural Equation Models. Paper presented at the conference New Techniques and Technologies for Statistics 2013, Brussels.

Scholtus, S. and B. F. M. Bakker (2013b). Estimating the Validity of Administrative and Survey Variables through Structural Equation Modeling: A Simulation Study on Robustness. Discussion Paper 201302, Statistics Netherlands, The Hague.

Scholtus, S., B. F. M. Bakker, and S. P. Robinson (2017). Evaluating the Quality of Business Survey Data before and after Automatic Editing. Working Paper No. 6, UN/ECE Work Session on Statistical Data Editing, The Hague.

Scholtus, S., B. F. M. Bakker, and A. Van Delden (2015). Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables. Discussion Paper 2015-17, Statistics Netherlands, The Hague.

Scholtus, S. and S. Göksen (2012). Automatic Editing with Hard and Soft Edits - Some First Experiences. Discussion Paper 201225, Statistics Netherlands, The Hague.

Schrijver, A. (1986). *Theory of Linear and Integer Programming*. New York: John Wiley & Sons.

Schulte-Nordholt, E., M. Hartgers, and R. Gircour (2004). *The Dutch Virtual Census of 2001. Analysis and Methodology*. Voorburg/Heerlen: Statistics Netherlands.

Schulte-Nordholt, E., J. Van Zeijl, and L. Hoeksma (2014). *Dutch Census 2011. Analysis and Methodology*. The Hague/Heerlen: Statistics Netherlands.

Skinner, C. J., D. Holt, and T. M. F. Smith (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons.

Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling*. Boca Raton: Chapman & Hall/CRC.

Snijkers, G., G. Buiten, and R. Van de Boom (2016). Issues in Automated Data Collection for Financial Information in The Netherlands. Paper presented at the ICES-V Conference, Geneva.

Sobel, M. E. and G. Arminger (1986). Platonic and Operational True Scores in Covariance Structure Analysis. *Sociological Methods and Research 15*, 44–58.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science 103*, 677–680.

Stevens, S. S. (1959). Measurement. In Churchman (Ed.), *Measurement: Definitions and Theories*, pp. 18–36. New York: John Wiley & Sons. Reprinted in: Maranell (1974), Scaling: A Sourcebook for Behavioral Scientists, pp. 22–41, Aldine Publishing Company, Chicago.

Stoer, J. and R. Bulirsch (2002). *Introduction to Numerical Analysis* (Third ed.). New York: Springer-Verlag.

Stoop, I. A. L. (2005). The Hunt for the Last Respondent. PhD Thesis, Universiteit Utrecht, Social and Cultural Planning Office, The Hague.

Stuart, W. J. (1966). Computer Editing of Survey Data. Five Years of Experience in BLS Manpower Surveys. *Journal of the American Statistical Association 61*, 375–383.

Tempelman, D. C. G. (2007). Imputation of Restricted Data. PhD Thesis, University of Groningen.

Todaro, T. A. (1999). Overview and Evaluation of the AGGIES Automated Edit and Imputation System. Working Paper No. 19, UN/ECE Work Session on Statistical Data Editing, Rome.

Tourangeau, R., L. J. Rips, and K. Rasinski (2000). *The Psychology of Survey Response*. New York: Cambridge University Press.

UN/ECE (2000). Glossary of Terms on Statistical Data Editing. Report, United Nations, Geneva.

UN/ECE (2014). *Practices of UNECE Countries in the 2010 Round of Censuses*. New York: United Nations.

Van Delden, A., R. Banning, A. De Boer, and J. Pannekoek (2016). Analysing Correspondence between Administrative and Survey Data. *Statistical Journal of the IAOS 32*, 569–584.

Van Delden, A. and P.-P. De Wolf (2013). A Production System for Quarterly Turnover Levels and Growth Rates Based on VAT Data. Paper presented at the conference New Techniques and Technologies for Statistics 2013, Brussels.

Van Delden, A., S. Scholtus, P.-P. De Wolf, and J. Pannekoek (2014). Methods to Assess the Quality of Mixed-Source Estimates. Report, Statistics Netherlands, The Hague.

Van der Loo, M. and E. De Jonge (2011). Manipulation of Categorical Data Edits and Error Localization with the Editrules Package. Discussion Paper 201129, Statistics Netherlands, The Hague.

Van der Loo, M. and E. De Jonge (2012). Automatic Data Editing with Open Source R. Working Paper No. 33, UN/ECE Work Session on Statistical Data Editing, Oslo.

Van der Loo, M., E. De Jonge, and S. Scholtus (2011). Correction of Rounding, Typing, and Sign Errors with the Deducorrect Package. Discussion Paper 201119, Statistics Netherlands, The Hague.

Van der Pijll, E. and J. Hoogland (2003). Evaluatie Automatisch versus Handmatig Gaafmaken van Productiestatistieken 2000 Handel & Transport. Internal report 286-03-TMO (in Dutch), Statistics Netherlands, Voorburg.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Van Meurs, A. and W. E. Saris (1990). Memory Effects in MTMM Studies. In Saris and Van Meurs (Eds.), *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait-Multimethod Studies*, pp. 134–146. Amsterdam: North-Holland.

Wallgren, A. and B. Wallgren (2014). *Register-Based Statistics – Statistical Methods for Administrative Data* (Second ed.). Chichester: John Wiley & Sons.

Williams, H. P. (1986). Fourier's Method of Linear Programming and its Dual. *The American Mathematical Monthly 93*, 681–695.

Winkler, W. E. (1995). Editing Discrete Data. Working Paper, UN/ECE Work Session on Statistical Data Editing, Athens.

Winkler, W. E. and L. R. Draper (1997). The SPEER Edit System. In *Statistical Data Editing, Volume No. 2, Methods and Techniques*, pp. 51–55. New York and Geneva: United Nations.

Winkler, W. E. and T. F. Petkunas (1997). The DISCRETE Edit System. In *Statistical Data Editing, Volume No. 2, Methods and Techniques*, pp. 55–62. New York and Geneva: United Nations.

Wothke, W. and M. W. Browne (1990). The Direct Product Model for the MTMM Matrix Parameterized as a Second Order Factor Analysis Model. *Psychometrika 55*, 255–262.

Zhang, L.-C. (2012). Topics of Statistical Theory for Register-Based Statistics and Data Integration. *Statistica Neerlandica 66*, 41–63.

Zhang, L.-C. (2014). Data Integration. *The Survey Statistician 70*, 15–24.

Summary

Data that are collected for the production of official statistics or, more generally, for statistical analyses nearly always contain measurement errors. National statistical institutes, other statistical agencies and academic researchers have therefore developed methods to handle error-prone data. Two broad classes of approaches can be distinguished: editing methods that aim to identify and correct individual errors in the data and estimation methods that try to correct for measurement errors at the analysis stage, without adjusting the data themselves. The aim of this thesis was to contribute to the development of both approaches for dealing with measurement errors, with a particular focus on their extension and application to large data sets from administrative sources.

In particular, the following points have been addressed in this thesis. Firstly, current methods for automatic data editing – based on the seminal work of Fellegi and Holt (1976) – have limited practical applicability because they are based on rather restrictive assumptions. In this thesis, two new methods for automatic editing have been developed that relax some of these assumptions. Secondly, we have discussed the estimation of measurement error models with latent variables in an official-statistics context. Here, often univariate descriptive statistics such as population totals and means are of interest. It was demonstrated how latent-variable models could be used to assess the suitability of new data sources for official statistics, to gain better insight into the accuracy of statistics and to improve the quality of statistical output. Thirdly, an application at Statistics Netherlands was described in which a measurement error model was used to compare the quality of data before and after an automatic editing procedure.

Having given this brief overview, we now provide a more detailed summary of the contents of this thesis. In total, five chapters of original research have been presented in this thesis. Chapters 3–5 focussed on new methods for automatic editing. Chapters 6 and 7 focussed on applications of measurement error models.

In Chapter 3, we looked at deductive correction methods for systematic errors. Correcting systematic errors in a separate step at the beginning of a data editing

process can improve the efficiency of data editing as well as the quality of the edited data. This is true because, if a systematic error can be corrected accurately by a deductive rule, it does not have to be treated later on by a human editor or a more complex algorithm for automatic error localisation. This means that editors and more complex algorithms can focus their attention on cases with more complicated error structures, where their contribution is more likely to be worthwhile.

With the above aims in mind of improving efficiency and quality, we have developed two new deductive methods for correcting two errors that are known to occur in data of the so-called Structural Business Statistics (SBS) at Statistics Netherlands: sign errors and rounding errors. Sign errors occur for variables in a particular subsection of the questionnaire (the so-called profit-and-loss account), while rounding errors can occur throughout the data. Both methods require an algorithm that is more complex than a simple if-then rule, but they are still relatively easy and cheap to implement. Theoretical properties of the algorithms were investigated. By way of illustration, both algorithms were applied to real data from the Netherlands' SBS of 2007. For these data, we found that the deductive method for sign errors reduced the number of records with inconsistent profit-and-loss accounts by about twenty per cent. We also found that, of all records that contained inconsistencies with respect to the edit rules, about one in five contained at least one rounding error. Moreover, by resolving these rounding errors, the number of violated balance edit rules could be reduced by about thirteen per cent. These results show that these deductive methods can achieve a substantial reduction of the amount of editing that remains to be done by editors or complex error localisation algorithms.

Chapters 4 and 5 focussed on error localisation for random errors. Two generalisations of the Fellegi-Holt paradigm were proposed that aim to improve the quality of automatically-edited data. Both generalisations address a different limitation of the Fellegi-Holt paradigm.

The starting point for Chapter 4 was the idea that some of the systematic differences that have been found between manual and automatic editing may be explained by the fact that human editors make use of soft edits as well as hard edits, whereas the Fellegi-Holt paradigm for automatic editing assumes that only hard edit rules occur. Under the Fellegi-Holt paradigm, existing soft edits have to be either ignored or treated as hard edits during automatic error localisation. We proposed a new formulation of the error localisation problem that can distinguish between hard and soft edit rules. The new approach involves solving a minimisation problem that is a generalisation of the problem of Fellegi and Holt, with an extra term that measures the extent to which soft edit rules are violated. The new

problem can be solved by an extension of the existing error localisation algorithm of De Waal and Quere (2003). A simulation study was conducted with synthetic data. For these data, it was found that the new error localisation approach achieved better results than the Fellegi-Holt paradigm, both in terms of false positives (correct values that were identified as erroneous by the algorithm) and false negatives (erroneous values that were identified as correct).

The Fellegi-Holt paradigm and the underlying model based on Naus et al. (1972) tacitly assume that errors independently affect one variable at a time. By contrast, human editors often make adjustments to the data that involve more than one variable at a time. It is in fact likely that respondents often commit errors that simultaneously affect several variables. In Chapter 5 we therefore introduced a generalised error localisation problem in which the assumption is relaxed that errors affect one variable at a time. This problem is based on a new minimisation criterion which involves the number of required edit operations rather than the number of changed values. Here, each edit operation is a well-defined elementary adjustment that can be made to a record to correct one particular error, which might involve changing the values of one, two, or more variables simultaneously. We suggested to choose these edit operations such that they mimic as closely as possible the manual corrections made by editors. The Fellegi-Holt-based error localisation problem is in fact a special case of the new problem, obtained by restricting the set of admissible edit operations to one particular class (i.e., operations that impute a new value for a single variable).

An algorithm was developed for solving the new error localisation problem. This algorithm was used in a simulation study with synthetic data to compare the new approach to Fellegi and Holt's original error localisation problem. The results of this study indicated that the new method can be used to achieve a significant improvement of the quality of automatically-edited data (again in terms of both false negatives and false positives). This does require that all (or nearly all) appropriate edit operations are included. Finding the appropriate edit operations for a given application is not trivial; we provided some suggestions on how this might be done in practice.

Turning to measurement error models, in Chapter 6 we used a structural equation model (SEM) to estimate the quality of administrative and survey data for official statistics. It was shown how both the indicator validity and intercept bias of administrative and survey variables can be estimated in this way. In particular, the indicator validity can be used as a measure to decide whether the administrative concept is sufficiently related to the true variable of interest to be of use. In cases where the validity is high but significant intercept bias occurs, a correction

formula can be derived from the SEM by predicting the true value of the variable of interest from the observed value. To identify the model, we took a random subsample of our original observations and attempted to measure the true values for these units (an audit sample). The inclusion of an audit sample was necessary for the estimation of the true intercept bias and true correction formulas for the observed variables, but not for the estimation of indicator validity.

The methodology was applied to real data at Statistics Netherlands to estimate the validity and intercept bias of value-added tax (VAT) turnover for short-term statistics (monthly or quarterly statistics on the development of the economy). SEMs were fitted to linked data from three administrative sources (VAT, the Profit Declaration Register and the General Business Register) and one survey (SBS). Additional data for an audit sample were obtained by re-editing the survey data. It was found that the target variable turnover was measured with indicator validity close to 1 in all data sources. However, often the VAT data did suffer from substantial intercept bias. For cases where intercept bias occurred, a correction formula was derived from the SEM. We simulated an application of the estimated correction formulas from the SEM to publication figures for the short-term statistics. As expected, it was found that the correction hardly affected the estimated annual growth rates but it did have a substantial effect on the estimated annual turnover levels.

Finally, in Chapter 7 we used measurement error modelling to gain insight into the quality of edited data. The indicator validity and bias of observed variables in a data set of the Netherlands' SBS before and after automatic editing were evaluated and compared. We analysed the data using two different models: an SEM and a contamination model. The latter model seemed more appropriate for the data at hand, but its current formulation does have some limitations that require further development. In our application, the effect of automatic editing on data quality in terms of validity and bias turned out to be very limited. In particular, the models suggested that the data after editing still contained a substantial amount of measurement error.

Acknowledgements

— Mag ik u hartelijk bedanken voor uw aandacht?
— Gaat uw gang.
— Ik wil u hartelijk bedanken voor uw aandacht.
— Gaat uw gang.
— Hartelijk bedankt voor uw aandacht.
— Geen dank.
— En ik had het nog zo gevraagd!
— Tja, is niks meer aan te doen.
— O, nou, u wordt bedankt.
— Ja, dat idee had ik al.
Herman Finkers – *Het meisje van de slijterij* (1987)

First and foremost, I would like to thank prof. dr. Bart Bakker for instigating this thesis project and for providing excellent support and supervision throughout. In a very literal sense, this thesis would not have existed without him.

I would also like to thank prof. dr. Harry Ganzeboom and prof. dr. Cees Elzinga for acting as co-supervisors. They have both, each in their own way, provided me with much-needed advice. In particular, I have benefited a lot from their comments during the final stages of writing this thesis.

I am very grateful to prof. dr. Ineke Maas, prof. dr. Thomas Laitila, prof. dr. Joop Hox, prof. dr. Ton de Waal, and dr. Daniel Oberski for acting as members of the reading committee. Unfortunately, due to a clash in schedules, professor Hox was not able to attend the PhD defence ceremony. Many thanks to prof. dr. Li-Chun Zhang for agreeing to step in as a replacement on short notice.

Many colleagues and former colleagues at Statistics Netherlands have contributed, directly or indirectly, to the development of this thesis. I would like to thank my co-authors for Chapters 6 and 7: Bart Bakker, Arnout van Delden, and Sam Robinson. Guus van de Burgt, Danny van Elswijk, Joost van der Leeden, Roy Manglie, Roos Smit, and Mathieu van der Vijgh provided access to data and metadata from production processes at Statistics Netherlands. (In the case of Roy

and Mathieu, this involved actually creating the data that we needed for the audit sample in Chapters 6 and 7.)

I would like to thank all members past and present of the Department of Methodology and Process Development at Statistics Netherlands for providing a work environment that is supportive and stimulating. In particular, I am grateful to Jacco Daalmans, Arnout van Delden, Bram Duyx, Jeffrey Hoogland, Edwin de Jonge, Léander Kuijvenhoven, Mark van der Loo, Jeroen Pannekoek, Barry Schouten, Marc Smeets, Caren Tempelman, Léon Willenborg, and Peter-Paul de Wolf for collaborating with me on related projects, for commenting on early drafts, and in general for many interesting discussions. Jeroen and Arnout in particular have helped to shape the research behind this thesis, in their roles as co-ordinators of the research programs at Statistics Netherlands on data throughput and on the use of administrative and combined data, respectively, for most of the time that I have worked on this thesis. I have really enjoyed our collaborations over the years. A special mention is also due to Edwin and Mark for their work on developing R packages for data editing, such as `editrules`, which proved invaluable for obtaining the results in Chapters 4, 5, and 7.

Sevinç Göksen, Sam Robinson, and Judith ter Schure have done Master's thesis projects at Statistics Netherlands under my supervision on topics closely related to this thesis: automatic editing with hard and soft edit rules (Sevinç) and finite mixture models for estimating measurement errors (Sam and Judith). I want to thank them for a pleasant collaboration and for sharpening my understanding of these topics.

Two people at Statistics Netherlands have probably taught me more about statistics – both in theory and in practice – than anyone else: Abby Israëls and Paul Knottnerus. I feel very privileged to have met and to have had the opportunity to work with you both.

Finally, I would like to thank my parents and my brother for putting up with me for the last 34.5 years.

Sander Scholtus
December 2017