Correcting for Linkage Errors in Contingency Tables: New Methods to Improve the Correction Approach

Eikenhout, Sjarai

Correcting for Linkage Errors in Contingency Tables: New Methods to Improve the Correction Approach

Sjarai Sancha Eikenhout

Thesis advisors:
Prof.dr. T. de Waal, Statistics Netherlands
Dr. S. Scholtus, Statistics Netherlands
Prof.dr. F.P. Pijpers, Statistics Netherlands
Prof.dr. M.J. de Rooij, Institute of Psychology, Leiden University

Defended on 10 July 2023

MASTER THESIS

STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Abstract

Record linkage aims to bring records together from two or more files that belong to the same statistical entity. Linkage errors can occur during this process. Ignoring these linkage errors can lead to biased inference. There is a growing emphasis on accounting for linkage errors in the statistical analysis of categorical data and contingency tables.

In this thesis, we developed three new approaches for compensating for linkage errors in contingency tables. The first approach, the regularised estimator, uses ideas from the application of regularisation of ill-conditioned matrices. Two other approaches use probabilities to compute the expected contingency table given the observed contingency table and to weight three existing correction methods with their estimated mean square error. The new approaches were tested together with two existing estimators $\mathbf{Q}$ and $\mathbf{Q}^{-1}$ by means of a simulation study.

For dependent contingency tables, we propose to use the expected value approach with a prior distribution that uses information about the observed values of the contingency table. Moreover, we propose to use the existing $\mathbf{Q}$ approach for independent contingency tables. The regularised estimator seems to have a lot of potential for both dependent and independent tables, but improvement is still needed.

Keywords: contingency table, exchangeable linkage error model, linkage error correction, probabilistic record linkage

Contents

1 Introduction

With the increasing use and availability of routinely collected data, linking data from multiple sources is becoming more useful. Record linkage is a solution to the problem of recognizing records in two or more files that represent identical persons, objects, or events and aims to bring those records together (Fellegi & Sunter, 1969). To illustrate the concept of record linkage, a simple scenario is shown in Figure 1. Assume there are two separate files, with four records in each file. Information about an individual's highest education qualification is stored in file 1 and information about the individual's occupation is in file 2. By linking these two files together, the linked file on the right is obtained. The unlinked files (file 1 and file 2) provide information on the distribution of *education qualification* and *occupation* separately, e.g. the percentage of individuals with a PhD degree or the percentage of individuals that are methodologists by profession. The linked file additionally provides information on the joint distribution of *education qualification* and *occupation*, e.g. the percentage of methodologists with a PhD degree.

Figure 1
Example of the Concept of Record Linkage

File 1

	First name	Last name	Education qualification
1.	Liam	Smith	BSc
2.	Olivia	Johnson	MSc
3.	Noah	Williams	MSc
4.	Luna	Brown	PhD

File 2

	First name	Last name	Occupation
1.	Liam	Smith	Business analist
2.	Luna	Brown	Methodologist
3.	Olivia	Johnson	Methodologist
4.	Noah	Williams	Data scientist

Linked file (File 1 + File 2)

	First name	Last name	Education qualification	Occupation
1.	Liam	Smith	BSc	Business analist
2.	Olivia	Johnson	MSc	Methodologist
3.	Noah	Williams	MSc	Data scientist
4.	Luna	Brown	PhD	Methodologist

Note. File 1 contains the first name, last name, and highest educational qualification of four individuals. File 2 contains the first name, last name, and occupation of the same four individuals. By using record linkage, records from these two files can be brought together into one file. The linked file that is obtained after linking files 1 and 2 is shown on the right.

Record linkage is also used for the deduplication of individual records within a single database and case re-identification in capture-recapture studies (Sayers et al., 2016). However, this thesis focuses on linking data from multiple sources together.

Record linkage is used in all kinds of scientific areas, e.g. health, epidemiology, demography, and sociology. However, it is not limited to these scientific areas. National Statistical Institutes (NSIs) increasingly rely on linking surveys to administrative registers to provide more accurate

measurements and to reduce respondent burden. Two examples of the use of record linkage in official statistics are given by Chambers (2009). The first example is the development of a linked longitudinal employer-employee dataset based on administrative data at Statistics New Zealand. This dataset allows the analysis of job and worker flows, multiple job holdings, and business demography. The second example is the development of an integrated longitudinal dataset by the Office for National Statistics in the United Kingdom to improve migration and population statistics. Records from administrative, health register, school enrollment, and university student data are linked with incoming passenger survey and labour force survey data, which is used for the analysis of the migrant experience in the UK.

During the record linkage process, linkage errors can occur. Linkage errors are the errors caused by incorrectly linking different population units as well as by not linking the same population units; incorrect links and missed links, respectively. These linkage errors are a particular type of measurement error, which can lead to biased inference if no appropriate steps are taken to control and/or adjust this bias. Typically, the errors in this type of record matching are ignored. Hence, this leads to bias and additional variability in standard statistical estimation techniques (Chambers, 2009).

NSIs aim to publish high-quality and accurate descriptive statistics of the population. Consequently, the estimation of contingency tables is important as they are used to inform policies by government agencies and other stakeholders. In earlier research on correcting for linkage errors, the focus mainly is on compensating for linkage errors in regression models, given data from a probabilistically linked file (see also Section 2.1). There has been comparatively little research carried out on compensating for linkage errors in contingency tables. However, due to the increasing use of administrative and new forms of data in statistical systems, more research is needed to understand the impact of linkage errors on categorical data and to continue developing record linkage approaches. In particular, there is a growing emphasis on accounting for linkage errors in the statistical analysis of contingency tables (Chipperfield & Chambers, 2015; Scholtus et al., 2022).

To emphasise this, a real example from Chipperfield and Chambers (2015) is shown. In this example, a census and a database are linked; where the records in the database are a subset of the census. The census contains economic and social information from over 20 million people living in Australia. The database contains information about 1,315,000 immigrants who were granted visas to live permanently in Australia. The 'true' contingency table for *level of qualification* (a categorical variable in the census) by *visa class* (a categorical variable in the database) is shown in Table 1, where it is assumed that all the links are correct. Here, the links are made by using the variables *name* and *address*. The frequency counts are expressed as proportions of the marginal counts by *visa class*.

Table 1
Real Example of Accounting for Linkage Errors in the Statistical Analysis of Contingency Tables: The True Contingency Table of Level of Qualification and Visa Class

	Level of Qualification		
Visa Class	1	2	3
1	0.273	0.642	0.083
2	0.385	0.391	0.222

Note. *Level of qualification* is a categorical variable in the census data and *visa class* is a categorical variable in the database. The frequency counts are expressed as proportions of the marginal counts by *visa class*.

Now, we will look at the results when the linkage is not perfect. The difference with Table 1 is that now the links are made by using all variables, except *name* and *address*. The naïve contingency table that is obtained for *level of qualification* by *visa class* when the data is linked without any compensation for linkage errors is shown in Table 2.

Table 2
Real Example of Accounting for Linkage Errors in the Statistical Analysis of Contingency Tables: The Contingency Table Using a Naïve Estimator (Not Correcting for Linkage Errors)

	Level of Qualification		
Visa Class	1	2	3
1	0.220	0.750	0.029
2	0.315	0.641	0.043

Note. *Level of qualification* is a categorical variable in the census data and *visa class* is a categorical variable in the database. The frequency counts are expressed as proportions of the marginal counts by *visa class*.

It is clearly visible that the naïve contingency table differs from the true contingency table. If the naïve contingency table would be used instead of the (unknown) true contingency table, this may lead to biased inference.

In Scholtus et al. (2022) the effect of linkage errors on contingency tables is researched in its purest form, assuming there are no other errors besides linkage errors. That paper is based on a basic problem, namely the estimation of contingency tables where the two target variables are from different linked datasets. Two fundamental correction methods; an unbiased correction and a biased correction, are investigated and compared to a naïve approach where linkage error is not compensated for. Results showed that the unbiased correction approach performs rather poorly compared to the biased and naïve approach. In practice, the variance of the unbiased estimator is often so large that the naïve and biased estimators produce a smaller total mean square error. Further research is needed to construct an estimator correcting for linkage errors in contingency tables that performs consistently better than the naïve estimator, the biased estimator, and the unbiased estimator. Preferably, this estimator is (approximately) unbiased and has a lower variance.

The aim of this current project is to develop, implement and test new methods for the correction of linkage errors in contingency tables. A simulation study will be performed where the quality of the new estimators will be measured by computing the estimation errors, bias, variance, and mean square error. The first new estimator, presented in Pijpers (2021), uses notions from the application of regularisation of ill-conditioned matrices. Furthermore, two new estimators will be developed and evaluated based on other ideas for a possible improved estimator presented in Scholtus and De Waal (2020-2022). Moreover, in Scholtus et al. (2022) conditions are derived that describe which correction method is best to use in a given situation. However, these conditions cannot be applied directly as they are based on the true contingency table, which is not known in reality. This thesis aims to find a way to determine which estimator to use in a specific situation.

The remainder of this thesis is structured as follows. Section 2 provides background information about the main topics of this project. Section 3 describes the used methods. The correction methods are tested with a simulation study in Section 4. Lastly, Section 5 discusses all the results of this study and gives further recommendations.

2 Background information

This section provides a brief introduction to the main topics of this project; namely, record linkage and linkage errors. As this thesis builds on Scholtus et al. (2022), a repetition of the notation and terminology used in that paper is given next. Lastly, the exchangeable linkage error model is explained.

2.1 Record linkage

As mentioned in the Introduction, record linkage is the process of bringing information from two or more distinct sources together. In general, there are two broad types of record linkage, namely deterministic and probabilistic record linkage.

Deterministic record linkage is the process of linking information by a unique shared key or by identifying variables using a fixed linking ruleset. Deterministic linkage results in two mutually exclusive categories of linked and nonlinked records, where the nonlinked records are only present in one of the files. Unfortunately, deterministic linkage does not reflect whether the linking fields partially or fully agree, e.g. if a single character is different due to a spelling mistake but the remaining characters are all identical, the two records will not be linked while they probably should be linked. This can be solved by cleaning the data first to reduce heterogeneity or by applying some form of probabilistic linkage.

Probabilistic record linkage attempts to link pieces of information together using multiple, possibly non-unique keys (Sayers et al., 2016). First, all matching fields have to be uniquely identifying in both files. Next, a join between the two files is performed containing all possible links. Each possible link is given a score based on the likelihood that the records belong to the same unit. Optimisation algorithms are used to select which pairs are declared as links. Nowadays, probabilistic methods for record linkage are well established (Chipperfield & Chambers, 2015).

When using probabilistic linkage, a researcher is in a more informed position than when using deterministic linkage (Sayers et al., 2016). An advantage of probabilistic linkage is its capability of finding the optimal balance between keeping sufficient discriminative power and allowing disagreements to overcome registration errors. Deterministic linkage does not have the capability of finding this optimal balance and therefore has to be based on a priori knowledge (Tromp et al., 2011).

2.2 Linkage errors

Ideally, the linkage between two datasets will be perfect, i.e. all records belonging to the same unit are linked and no other records are linked. However, this does not happen in many situations, especially when records may contain incorrect values or missing values (Chipperfield et al., 2011).

Two types of errors can occur in record linkage: the failure to link two records that belong to the same entity (missed links) and the linking of two records that belong to different entities (incorrect links). Missed links occur when there is disagreement on linking variables while the records belong to the same entity, a problem that can be caused by data entry errors. Incorrect links occur when two different entities share the same value on several linking variables by coincidence (Tromp et al., 2011). Incorrectly linked pairs may potentially incur bias in statistical analyses, and missed links impact coverage and potential bias if those missed links differ in their characteristics from the found links (Scholtus et al., 2022). Typically, the linkage errors are ignored, and thus bias and additional variability are introduced into statistical estimation techniques. This poses a significant barrier to policy-relevant research using probabilistically linked data (Chambers, 2009).

Statistical methods for linking datasets are now well established. Previous research in this area is mainly focused on the confidentiality issues that arise as a consequence of linkage (Chambers, 2009) and regression model-based linkage correction (Scholtus et al., 2022). However, as the focus of NSIs emerges on administrative-based censuses, it is important to understand the impact of linkage errors on categorical data and census tables. NSIs aim to publish high-quality and accurate descriptive statistics of the population. The estimation of such contingency tables is therefore important, as they are mainly used to inform policies. Meanwhile, the statistical analyses of such tables, e.g. for tests for independence, are important for research purposes (Scholtus et al., 2022).

2.3 Notation and terminology

In Scholtus et al. (2022) the focus is on incorrectly probabilistically linked pairs which result from records in two datasets being linked incorrectly due to errors, missing values, or changes over time in the variables that are used in the linking procedure. The missed links are not included. It is assumed that both datasets contain the same units. The aim is to link all of the units, corresponding to a one-to-one linkage. The scenario of having a one-to-one linkage is relevant at NSIs, where there is a move towards administrative-based censuses which link multiple administrative data sources, e.g. administrative data from patients registers, income tax, and social security authorities. There is an expectation of a one-to-one linkage in this scenario (Scholtus et al., 2022). Consequently, the focus of this thesis is also on one-to-one incorrectly probabilistically linked pairs from datasets that contain the same units.

Assume two data files A and B of n records ($n \geq 2$) for observed units in a population. File A contains variables (x, y) and file B contains variables (x, z). To estimate the contingency table of the categorical variables y and z, the two datasets are linked first on the common variable(s) x. For the purpose of constructing the contingency table of y and z, variables y and z can be re-coded into dummy variables as a binary $n \times J$ matrix $\mathbf{Y}$ and $n \times K$ matrix $\mathbf{Z}$ respectively, where J is the number of categories of variable y and K the number of categories of variable z.

The elements of $\mathbf{Y}$ are equal to one ($y_{ij} = 1$) if the unit in record i of the first dataset belongs to category j of variable y, and equal to zero ($y_{ij} = 0$) otherwise. Similarly, the elements of $\mathbf{Z}$ are equal to one ($z_{ik} = 1$) if the unit in record i of the second dataset belongs to category k of variable z, and equal to zero ($z_{ik} = 0$) otherwise. As the focus is on the effect of linkage errors on estimators, $\mathbf{Y}$ and $\mathbf{Z}$ are treated as fixed. After linking the two data files, the $J \times K$ target contingency table is given by $\mathbf{T} = \mathbf{Y}^\mathsf{T}\mathbf{Z}$ (where $^\mathsf{T}$ denotes taking the transpose of a matrix) with typical element $t_{jk} = \sum_{i=1}^{n} y_{ij} z_{ik}$. The marginal counts of the contingency table of y and z can be obtained from the separate data files. Therefore, they are unaffected by linkage errors. The marginal count for category j of variable y is denoted by $r_j = \mathbf{y}_j^\mathsf{T}\mathbf{u} = \mathbf{u}^\mathsf{T}\mathbf{y}_j$, where $\mathbf{y}_j$ denotes a column of $\mathbf{Y}$ and $\mathbf{u}$ denotes the n vector of ones. Similarly, the marginal count for category k of variable z is denoted by $s_k = \mathbf{z}_k^\mathsf{T}\mathbf{u} = \mathbf{u}^\mathsf{T}\mathbf{z}_k$, where $\mathbf{z}_k$ denotes a column of $\mathbf{Z}$.

As both datasets contain the same entities, random linkage errors can be represented by a random permutation of the order of some units in the second dataset. As a result, $\mathbf{Z}^* = \mathbf{C}\mathbf{Z}$ is observed instead of $\mathbf{Z}$, where $\mathbf{C}$ is a stochastic permutation matrix of order n due to random linkage errors. This means that each row and column of $\mathbf{C}$ contains exactly one element equal to one and all other elements equal to zero. The target contingency table is observed as $\widehat{\mathbf{T}}^* = \mathbf{Y}^\mathsf{T}\mathbf{Z}^* = \mathbf{Y}^\mathsf{T}\mathbf{C}\mathbf{Z}$ with typical element $\widehat{t}^*_{jk} = \sum_{i=1}^{n} y_{ij}\ (\sum_{l=1}^{n} c_{il} z_{lk})$. See Example 1 for an illustration of the notation.

Example 1. Take $n = 10$ and assume that variables y and z both have 2 categories. Suppose that the two dummy variables are:
$\mathbf{Y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}^\mathsf{T}$ and $\mathbf{Z} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}^\mathsf{T}$.
The true contingency table of y and z is: $\mathbf{T} = \mathbf{Y}^\mathsf{T}\mathbf{Z} = \begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix}$. Suppose that linkage errors occur according to permutation matrix $\mathbf{C}$ where the 4^{th} and 10^{th} unit in the second dataset are permuted. The other eight units are linked correctly. Hence,
$\mathbf{Z}^* = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}^\mathsf{T}$. The observed contingency table is then given by:
$\widehat{\mathbf{T}}^* = \mathbf{Y}^\mathsf{T}\mathbf{Z}^* = \begin{pmatrix} 3 & 4 \\ 3 & 0 \end{pmatrix} \neq \mathbf{Y}^\mathsf{T}\mathbf{Z}$.

2.4 Exchangeable linkage error model

In Chambers (2009), a methodological framework is proposed that can be used to provide appropriate adjustments to standard statistical analysis methods to ensure that they remain unbiased when used with probabilistically linked data. This framework is based on modelling the relationship between the probabilistically linked data and the true data that would be obtained if perfect

linkage (linkage without errors) were possible. The proposed model is known as *the exchangeable linkage error* model and is widely used. It was originally suggested by Neter et al. (1965) in a groundbreaking paper. Despite the simplicity of the model, it provides important insights into the properties of various compensation approaches for linkage errors. These insights are also useful for more complicated linkage error models (Scholtus et al., 2022).

The framework in Chambers (2009) assumes the existence of a population of n units. Two registers are given: one with a scalar random variable and one with a vector random variable. It is assumed that both registers refer to the same population and do not contain any duplicates, hence both registers consist of n records. Probabilistic linkage is used to link the records. The linkage is complete and one-to-one. In practice, probabilistic linkage is often carried out within blocks. The linked records are then partitioned into W distinct blocks where each block contains m_w linked records, so $n = \sum_w m_w$. There is no possibility that linked records in different blocks contain data for the same population unit. Blocking allows one to consider more general linkage error models for different sub-groups of the datasets and makes record linkage more manageable in practice (Scholtus et al., 2022).

In Scholtus et al. (2022), the focus is on compensating for linkage errors when analysing a two-way contingency table, where one variable is from one file and the second variable is from another file. Therefore, it is more natural to take two scalar random variables in this context. Moreover, the focus is on one block. For all these reasons, we assume the existence of a population of n units, two registers A and B with a scalar random variable, and one block in what follows.

Let i index the records in the linked dataset. In total, there are n linked pairs (y_i, z_i^*), where y_i denotes the value of y on register A and z_i^* the linked value of z on register B. Let $\mathbf{z}^*$ denote the n vector defined by the linked values in z_i^*, $\mathbf{y}$ the n vector defined by the values y_i, and $\mathbf{z}$ the unknown n vector indexed as in register B that corresponds to the true values of y on register A associated with $\mathbf{y}$. As mentioned previously, randomness is modelled in the outcome of the linkage process by $\mathbf{z}^* = \mathbf{C}\mathbf{z}$, where $\mathbf{C} = [c_{ij}]$ is an unknown random permutation matrix of order n. Inference based on linked data involves assumptions about the distribution of $\mathbf{C}$, which is assumed to be independent of $\mathbf{z}$ given $\mathbf{y}$. Let

$$\mathbb{E}\left(\mathbf{C}|\mathbf{y}\right) = \mathbf{Q}. \tag{1}$$

Assume that the probability of correct linkage is the same for all records. Moreover, assume that it is equally likely that any two records in register A that are not linked to a specific record in register B could in fact be the correct link for this record. These assumptions can be characterised by the exchangeable linkage error model:

$$\Pr(\text{correct linkage}) = \Pr(c_{ii} = 1) = q, \tag{2}$$

and for $i \neq j$:

$$\Pr(\text{incorrect linkage}) = \Pr(c_{ij} = 1) = \delta, \tag{3}$$

where Pr reflects the stochastic process by which $\mathbf{C}$ is generated. The model specified by (2) and (3) represents a simple way of characterising the behaviour of a probabilistic linkage process. Given that (2) and (3) hold, it then follows that (1) is of the form:

$$\begin{aligned} \mathbf{Q} &= (q - \delta)\mathbf{I} + \delta\mathbf{u}\mathbf{u}^\mathsf{T} \\ &= q\mathbf{I} + \delta(\mathbf{u}\mathbf{u}^\mathsf{T} - \mathbf{I}), \end{aligned} \tag{4}$$

where $\mathbf{I}$ denotes the identity matrix of order n. Since $\mathbf{u}^\mathsf{T}\mathbf{C} = \mathbf{u}^\mathsf{T}$ and $\mathbf{C}\mathbf{u} = \mathbf{u}$, it follows that $\mathbf{u}^\mathsf{T}\mathbf{Q} = \mathbf{u}^\mathsf{T}$ and $\mathbf{Q}\mathbf{u} = \mathbf{u}$. This means that (4) implies:

$$q + (n - 1)\delta = 1. \tag{5}$$

So, if a value for q (the probability of a correct link) is specified, it follows from (5) that $\delta = \frac{1-q}{n-1}$ (the probability that the unit in register A should be linked to a specific other unit in register B). This is useful, as the estimation of q only requires information on whether a defined link is correct or incorrect and not the identity of the correct link. The model in (4), or equivalently (2) and (3) in combination with (5), is the essence of the previously mentioned *exchangeable linkage error model*.

The following weak technical assumption is made:

$$\frac{1}{n} < q < 1. \tag{6}$$

The left inequality implies that the linking process is at least better than linking the records completely at random and the right inequality rules out the trivial case where the linkage process is deterministic and perfect ($q = 1$). The inverse of $\mathbf{Q}$ is:

$$\mathbf{Q}^{-1} = \frac{n - 1}{nq - 1}\mathbf{I} - \frac{1 - q}{nq - 1}\mathbf{u}\mathbf{u}^\mathsf{T}. \tag{7}$$

The left inequality assumption in (6) is necessary to ensure that the inverse of $\mathbf{Q}$ exists.

The assumptions of the exchangeable linkage error model are followed in Scholtus et al. (2022) to develop a theoretical framework. As mentioned, the focus is on one block with a single value of q and it is assumed that the linkage error rate is known. Therefore, the focus in this thesis is also on one block and the linkage error rates are also assumed to be known.

3 Methods

In this section, the used methods in this study are covered. First, the existing correction methods are repeated. Next, a regularised estimator is presented. Lastly, new correction methods are presented based on conditional probabilities.

3.1 Existing estimators

In Scholtus et al. (2022), several different estimators are considered that have been proposed to improve on the naïve estimator. The main formulas of these estimators will now be repeated. The full derivations of these formulas can be found in Scholtus et al. (2022).

The naïve estimator is given by:

$$\widehat{\mathbf{T}}^{*} = \mathbf{Y}^{\mathsf{T}}\mathbf{Z}^{*}, \tag{8}$$

where $\mathbf{Z}^{*} = \mathbf{C}\mathbf{Z}$. It is easy to see that the naïve estimator is biased for the true contingency table, namely: $\mathbb{E}(\widehat{\mathbf{T}}^{*}) = \mathbb{E}(\mathbf{Y}^{\mathsf{T}}\mathbf{C}\mathbf{Z}) = \mathbf{Y}^{\mathsf{T}}\mathbb{E}(\mathbf{C})\mathbf{Z} \equiv \mathbf{Y}^{\mathsf{T}}\mathbf{Q}\mathbf{Z}$, which in general is not equal to $\mathbf{T} = \mathbf{Y}^{\mathsf{T}}\mathbf{Z}$. The following estimators are considered to correct for linkage errors:

$$\widehat{\mathbf{T}}^{Q} = (\mathbf{Q}\mathbf{Y})^{\mathsf{T}}\mathbf{Z}^{*} = \mathbf{Y}^{\mathsf{T}}\mathbf{Q}^{\mathsf{T}}\mathbf{C}\mathbf{Z}, \tag{9}$$

$$\widehat{\mathbf{T}}^{BC} = \mathbf{Y}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{Z}^{*} = \mathbf{Y}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{C}\mathbf{Z}. \tag{10}$$

The idea of the $\mathbf{Q}$ approach in (9) is that $(\mathbf{C}\mathbf{Y})^{\mathsf{T}}\mathbf{Z}^{*} = (\mathbf{C}\mathbf{Y})^{\mathsf{T}}\mathbf{C}\mathbf{Z} = \mathbf{Y}^{\mathsf{T}}\mathbf{Z} = \mathbf{T}$. The second equality follows from the fact that $\mathbf{C}$ is a permutation matrix, so $\mathbf{C}^{\mathsf{T}}\mathbf{C} = \mathbf{I}$. By replacing the unknown matrix $\mathbf{C}$ by its expectation $\mathbf{Q}$, the estimator $\widehat{\mathbf{T}}^{Q}$ is obtained. Using the expectation $\mathbf{Q}$ instead of the unknown matrix $\mathbf{C}$ is the defining property of this approach. The idea of the $\mathbf{Q}^{-1}$ approach in (10) is that $\mathbf{Y}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{Z}^{*}$ is an unbiased estimator for the true contingency table of y and z, namely: $\mathbb{E}(\mathbf{Y}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{Z}^{*}) = \mathbb{E}(\mathbf{Y}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{C}\mathbf{Z}) = \mathbf{Y}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbb{E}(\mathbf{C})\mathbf{Z} = \mathbf{Y}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Z} = \mathbf{Y}^{\mathsf{T}}\mathbf{Z} = \mathbf{T}$.

The errors of the single entries are examined in estimated contingency tables under the exchangeable linkage error model. Consider a single entry of the true contingency table $\mathbf{T}$: $t_{jk} = (\mathbf{Y}^{\mathsf{T}}\mathbf{Z})_{jk} = \mathbf{y}_j^{\mathsf{T}}\mathbf{z}_k$. The corresponding entries of the estimated tables $\widehat{\mathbf{T}}^{*}$, $\widehat{\mathbf{T}}^{Q}$, and $\widehat{\mathbf{T}}^{BC}$ are:

$$\begin{aligned}
\widehat{t}^{*}_{jk} &= (\mathbf{Y}^{\mathsf{T}}\mathbf{C}\mathbf{Z})_{jk} = \mathbf{y}_j^{\mathsf{T}}\mathbf{C}\mathbf{z}_k, \\
\widehat{t}^{Q}_{jk} &= (\mathbf{Y}^{\mathsf{T}}\mathbf{Q}^{\mathsf{T}}\mathbf{C}\mathbf{Z})_{jk} = \mathbf{y}_j^{\mathsf{T}}\mathbf{Q}^{\mathsf{T}}\mathbf{C}\mathbf{z}_k, \\
\widehat{t}^{BC}_{jk} &= (\mathbf{Y}^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{C}\mathbf{Z})_{jk} = \mathbf{y}_j^{\mathsf{T}}\mathbf{Q}^{-1}\mathbf{C}\mathbf{z}_k.
\end{aligned}$$

Denote the errors per element as $e^*_{jk} = \widehat{t}^*_{jk} - t_{jk}$, $e^Q_{jk} = \widehat{t}^Q_{jk} - t_{jk}$, and $e^{BC}_{jk} = \widehat{t}^{BC}_{jk} - t_{jk}$:

$$
\begin{aligned}
e^*_{jk} &= \mathbf{y}_j^\mathsf{T}(\mathbf{C} - \mathbf{I})\mathbf{z}_k, \\
e^Q_{jk} &= \mathbf{y}_j^\mathsf{T}(\mathbf{Q}^\mathsf{T}\mathbf{C} - \mathbf{I})\mathbf{z}_k = \frac{nq-1}{n-1}\mathbf{y}_j^\mathsf{T}(\mathbf{C} - \mathbf{I})\mathbf{z}_k + \frac{1-q}{n-1}\mathbf{y}_j^\mathsf{T}(\mathbf{u}\mathbf{u}^\mathsf{T} - n\mathbf{I})\mathbf{z}_k, \\
e^{BC}_{jk} &= \mathbf{y}_j^\mathsf{T}(\mathbf{Q}^{-1}\mathbf{C} - \mathbf{I})\mathbf{z}_k = \frac{n-1}{nq-1}\mathbf{y}_j^\mathsf{T}(\mathbf{C} - \mathbf{I})\mathbf{z}_k - \frac{1-q}{nq-1}\mathbf{y}_j^\mathsf{T}(\mathbf{u}\mathbf{u}^\mathsf{T} - n\mathbf{I})\mathbf{z}_k.
\end{aligned}
$$

For the derivations of the expressions e^Q_{jk} and e^{BC}_{jk} formulas (4) and (7) are used.

It follows that the expressions for the bias are:

$$
\begin{aligned}
B(\widehat{t}^*_{jk}) &= -\delta(nt_{jk} - r_j s_k), && (11)\\
B(\widehat{t}^Q_{jk}) &= -(1 + \frac{nq-1}{n-1})\delta(nt_{jk} - r_j s_k), && (12)\\
B(\widehat{t}^{BC}_{jk}) &= 0. && (13)
\end{aligned}
$$

Furthermore, expressions for the variance are:

$$
\begin{aligned}
\mathrm{Var}(\widehat{t}^*_{jk}) &= \mathrm{Var}(\mathbf{y}_j^\mathsf{T}\mathbf{C}\mathbf{z}_k), \\
\mathrm{Var}(\widehat{t}^Q_{jk}) &= \left(\frac{nq-1}{n-1}\right)^2 \mathrm{Var}(\mathbf{y}_j^\mathsf{T}\mathbf{C}\mathbf{z}_k), \\
\mathrm{Var}(\widehat{t}^{BC}_{jk}) &= \left(\frac{n-1}{nq-1}\right)^2 \mathrm{Var}(\mathbf{y}_j^\mathsf{T}\mathbf{C}\mathbf{z}_k).
\end{aligned}
$$

Note that the variance of $\widehat{t}^{BC}_{jk}$ can diverge. The variance of the estimator plus the square of the corresponding bias gives the mean square error for each of the three estimators. In Scholtus (2020), the following extended formula for the variance $\mathrm{Var}(\widehat{t}^*_{jk})$ is given for $n \geq 3$, assuming that

linkage errors can be described by the simple simulation method with 2-cycles (see Appendix B):

$$\begin{aligned}
\text{Var}\left(\widehat{t}^{*}_{jk}\right) = {} & q(1-q)t_{jk} + \frac{1-q}{n-1}\left(1-\frac{1-q}{n-1}\right)(r_j s_k - t_{jk}) \\
& - q\frac{1-q}{n-1}2t_{jk}(r_j + s_k - 2) \\
& + \left\{\frac{2q}{n-1} + \left(1-\frac{2}{n-1}\right)\left(\frac{nq-1}{n-1}\right)^{2-\frac{2}{n}} - q^2 - \left(\frac{1-q}{n-1}\right)^2\right\} t_{jk}(t_{jk}-1) \\
& + \frac{1}{n-2}\left\{\frac{n-3}{n-1}\left[q^2 - \left(\frac{nq-1}{n-1}\right)^{2-\frac{2}{n}}\right] + \left(\frac{1-q}{n-1}\right)^2\right\} \times \\
& \quad 2t_{jk}[(r_j-1)(s_k-1) - (t_{jk}-1)] \\
& - \left(\frac{1-q}{n-1}\right)^2 (r_j s_k - 2t_{jk})(r_j + s_k - 2) \\
& + \frac{1}{n-2}\left\{\frac{1}{n-1}\left[\left(\frac{nq-1}{n-1}\right)^{2-\frac{2}{n}} - q^2\right] + \left(\frac{1-q}{n-1}\right)^2\right\} \times \\
& \quad \{r_j s_k (r_j-1)(s_k-1) - 2t_{jk}\left[2(r_j-1)(s_k-1) - (t_{jk}-1)\right]\}. \qquad (14)
\end{aligned}$$

3.2 Regularised estimator

In Pijpers (2021), a new estimator is proposed. In this section, the main formulas of this estimator will be given. The full derivations of these formulas are given in Pijpers (2021). The proposed estimator uses notions from the application of regularisation of ill-conditioned matrices. A square matrix is ill-conditioned if it is invertible (non-singular) but can become non-invertible (singular) if some of its entries are changed slightly. The matrix $\mathbf{Q}$ is generally not ill-conditioned except when q approaches the lower bound $\frac{1}{n}$, which is exactly where the variance of the unbiased estimator $\widehat{\mathbf{T}}^{BC}$ in (10) diverges. Two different variants of the regularised estimator are proposed. The following estimator is the first variant proposed to correct for linkage errors:

$$\widehat{\mathbf{T}}^{reg} = \mathbf{Y}^{\mathsf{T}}[\nu\mathbf{Q}^{\mathsf{T}}\mathbf{Q} + (1-\nu)\mathbf{I}]^{-1}\mathbf{Q}^{\mathsf{T}}\mathbf{Z}^{*} = \mathbf{Y}^{\mathsf{T}}[\nu\mathbf{Q}^{\mathsf{T}}\mathbf{Q} + (1-\nu)\mathbf{I}]^{-1}\mathbf{Q}^{\mathsf{T}}\mathbf{C}\mathbf{Z}, \qquad (15)$$

where $0 \leq \nu \leq 1$. The parameter ν is to be chosen such that it balances minimizing the bias and the variance. As ν approaches zero the estimator becomes identical to $\widehat{\mathbf{T}}^{Q}$ and as ν approaches 1 the estimator becomes identical to $\widehat{\mathbf{T}}^{BC}$. The entries of the estimated table $\widehat{\mathbf{T}}^{reg}$ are:

$$\widehat{t}^{reg}_{jk} = \left(\mathbf{Y}^{\mathsf{T}}[\nu\mathbf{Q}^{\mathsf{T}}\mathbf{Q} + (1-\nu)\mathbf{I}]^{-1}\mathbf{Q}^{\mathsf{T}}\mathbf{C}\mathbf{Z}\right)_{jk} = \mathbf{y}_j^{\mathsf{T}}[\nu\mathbf{Q}^{\mathsf{T}}\mathbf{Q} + (1-\nu)\mathbf{I}]^{-1}\mathbf{Q}^{\mathsf{T}}\mathbf{C}\mathbf{z}_k.$$

The expression for the error per element, $e_{jk}^{reg} = \widehat{t}_{jk}^{reg} - t_{jk}$, becomes:

$$\begin{aligned} e_{jk}^{reg} &= \mathbf{y}_j^\mathsf{T}([\nu\mathbf{Q}^\mathsf{T}\mathbf{Q} + (1-\nu)\mathbf{I}]^{-1}\mathbf{Q}^\mathsf{T}\mathbf{C} - \mathbf{I})\mathbf{z}_k \\ &= \frac{nq-1}{n-1-n\nu\lambda(1-q)}\mathbf{y}_j^\mathsf{T}(\mathbf{C}-\mathbf{I})\mathbf{z}_k + \frac{(1-q)(1-\nu\lambda)}{n-1-n\nu\lambda(1-q)}\mathbf{y}_j^\mathsf{T}(\mathbf{u}\mathbf{u}^\mathsf{T} - n\mathbf{I})\mathbf{z}_k, \end{aligned}$$

where $\lambda = \frac{(n-1)^2-(nq-1)^2}{n(n-1)(1-q)}$. The expression for the bias of this family of estimators is:

$$B(\widehat{\mathbf{T}}^{reg}) = \frac{1}{n}\frac{[(n-1)^2-(nq-1)^2](1-\nu)}{(n-1)^2(1-\nu)+\nu(nq-1)^2}\mathbf{Y}^\mathsf{T}[\mathbf{u}\mathbf{u}^\mathsf{T} - n\mathbf{I}]\mathbf{Z},$$

and the expression for the variance is:

$$\mathrm{Var}(\widehat{t}_{jk}^{reg}) = \left(\frac{(nq-1)(n-1)}{(n-1)^2(1-\nu)+\nu(nq-1)^2}\right)^2 \mathrm{Var}(\mathbf{y}_j^\mathsf{T}\mathbf{C}\mathbf{z}_k).$$

One possible approach to choosing a value for ν is to determine whether there is a value for ν for which the mean squared error has a minimum. This yields:

$$\nu_{opt} = 1 - \frac{n^2(n-1)^2}{(n-1)^2-(nq-1)^2}\frac{\mathrm{Var}(\mathbf{y}_j^\mathsf{T}\mathbf{C}\mathbf{z}_k)}{(\mathbf{y}_j^\mathsf{T}[\mathbf{u}\mathbf{u}^\mathsf{T} - n\mathbf{I}]\mathbf{z}_k)^2}. \tag{16}$$

There are two problems with the expression for ν_{opt}. The first problem is that for values of $q \uparrow 1$ or when $(\mathbf{y}_j^\mathsf{T}[\mathbf{u}\mathbf{u}^\mathsf{T} - n\mathbf{I}]\mathbf{z}_k)^2 << \mathrm{Var}(\mathbf{y}_j^\mathsf{T}\mathbf{C}\mathbf{z}_k)$ the value of ν will approach $-\infty$. This would violate the design that $0 \leq \nu \leq 1$. It is inconvenient to have to determine whether the ν_{opt} is in the correct range before applying the estimator. The second problem is that the second fraction of the expression depends on the true contingency table $\mathbf{T}$, which is not known in practice. Therefore, this term needs to be estimated first.

A pragmatic choice proposed by Pijpers (2021) is to set ν to:

$$\nu_{prag} = \frac{(nq-1)^2}{(n-1)^2+(nq-1)^2}. \tag{17}$$

With the pragmatic value for ν, the bias and variance become:

$$\begin{aligned} B(\widehat{\mathbf{T}}^{reg,prag}) &= \frac{1}{n}\frac{[(n-1)^2-(nq-1)^2](n-1)^2}{(n-1)^4+(nq-1)^4}\mathbf{Y}^\mathsf{T}[\mathbf{u}\mathbf{u}^\mathsf{T}-n\mathbf{I}]\mathbf{Z} \\ &= \frac{(n-1)^4}{(n-1)^4+(nq-1)^4}B(\widehat{\mathbf{T}}^{Q}) \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(\widehat{t}_{jk}^{reg,prag}) &= \left(\frac{[(n-1)^2+(nq-1)^2](n-1)(nq-1)}{(n-1)^4+(nq-1)^4}\right)^2 \mathrm{Var}(\mathbf{y}_j^\mathsf{T}\mathbf{C}\mathbf{z}_k) \\ &= \left(\frac{[(n-1)^2+(nq-1)^2](n-1)^2}{(n-1)^4+(nq-1)^4}\right)^2 \mathrm{Var}(\widehat{t}_{jk}^{Q}). \end{aligned}$$

The second variant of the estimator that is proposed in Pijpers (2021), is not included in this project as the variance is slightly higher than for the naïve estimator, which is higher than for the first variant of the regularised estimator (see Appendix A).

3.3 New estimators using probabilities

As mentioned before, the true contingency table is unknown in practice. In what follows, new correction methods will be constructed by using the probabilities of the true t_{jk} given the observed $\widehat{t}^*_{jk}$, i.e. $\Pr(t|\widehat{t}^*)$ where t and $\widehat{t}^*$ denote the true and observed value for some cell, respectively. For notational convenience, we will often drop subscripts in this section. Multiple steps are necessary to compute these probabilities. First, the probabilities $\Pr(\widehat{t}^*|t)$ are computed, after which this conditional probability is reversed using Bayes' rule. Based on these probabilities, two different correction methods are constructed.

3.3.1 Prior probabilities

To derive the reverse probability $\Pr(t|\widehat{t}^*)$, Bayes' rule (Hoff, 2009) will be used. A prior distribution of the true value t_{jk} is needed to be able to use Bayes' rule to compute this probability. Three different prior distributions are constructed for t_{jk}. Note that the prior distribution has to be computed for each individual cell of the contingency table and that different values for t (i.e. potential true values) are considered. The values for t run over all integers from $L = \max\{r_j + s_k - n, 0\}$ up to and including $U = \min\{r_j, s_k\}$. L and U are the so-called Fréchet bounds and hold true for any cell in any contingency table (Nelsen, 1987). Due to computational difficulties, which will be clarified in Section 4.1, the possible integers for t are noted as $L \leq L' \leq t \leq U' \leq U$.

The first prior distribution is given by:

$$\Pr_0(t) = \frac{1}{U' - L' + 1} \text{ for } L' \leq t \leq U'. \tag{18}$$

With this prior, each value of t that will be considered has the same probability of occurring. As this prior distribution just assigns the same probability to each possible value for t, it is hardly informative.

The second prior distribution, proposed by Scholtus and De Waal (2020-2022), is given by:

$$\Pr_0(t) = C \frac{\binom{r_j}{t}\binom{s_k}{t} t! \frac{(n-r_j)!(n-s_k)!}{(n-r_j-s_k+t)!}}{n!} \text{ for } L' \leq t \leq U', \tag{19}$$

where C is a constant to ensure that the probabilities sum up to 1. It is assumed that this prior probability is the probability of obtaining t_{jk} in cell (j,k) when links between the n units in dataset A and the n units in dataset B are made randomly, with r_j units with category j in dataset A and s_k units with category k in dataset B. $\binom{r_j}{t}\binom{s_k}{t} t!$ is the number of ways t can be selected from the r_j units with category j in dataset A and from the s_k units with category k in dataset B, and subsequently, link these sets of units to each other randomly. $\frac{(n-s_k)!}{(n-r_j-s_k+t)!}$ is the number of ways in which $r_j - t$ units with category j in dataset A can be linked to $n - s_k$ units in dataset B not having category k. This way the value of t is not affected. The remaining $n - r_j$ units in dataset A can be linked in every way to $n - r_j$ units in dataset B without affecting the value of t, which can be done in $(n - r_j)!$ ways (Scholtus & De Waal, 2020-2022). This prior distribution is more informative in comparison to the first prior distribution in (18), as it makes use of the marginals r_j and s_k.

The third and last prior distribution is given by:

$$\Pr_0(t) = C \binom{U'}{t} p^t (1-p)^{U'-t} \text{ for } L' \leq t \leq U', \tag{20}$$

where C is a constant to ensure that the probabilities sum up to 1. A truncated binomial distribution is assumed for t. Since the observed values $\widehat{t^*}$ are known, we assume that $\mathbb{E}_{\Pr_0}(t) = \widehat{t^*}$, where $\mathbb{E}_{\Pr_0}$ denotes the expectation with respect to the truncated binomial model. To find a value for p, we assume that $\mathbb{E}_{\Pr_0}(t)$ can be closely approximated by using the corresponding binomial model instead of the truncated binomial model. Then, we can say: $\mathbb{E}_{\Pr_0}(t) = U' \times p$. Subsequently, p can be estimated by $\widehat{p} = \frac{\widehat{t^*}}{U'}$. Thus, this prior distribution can be estimated by:

$$\widehat{\Pr_0}(t) = C \binom{U'}{t} \widehat{p}^t (1-\widehat{p})^{U'-t} \text{ for } L' \leq t \leq U'. \tag{21}$$

As this prior distribution makes use of the observed values $\widehat{t}^*$, it is again more informative in comparison to the first prior distribution in (18).

In practice, the informativeness of the prior distribution for the true value t can be both an advantage and a disadvantage. If the value of $\widehat{t}^*$ is close to the true value t_{jk}, it is useful to make use of an informative prior distribution, e.g. (19) and (21). However, if the value of $\widehat{t}^*$ deviates more from t_{jk}, it might be better to use a less informative prior, e.g. (18).

3.3.2 Computing the conditional probabilities

To compute the conditional probabilities $\Pr(\widehat{t}^*|t)$, an iterative method is proposed in Scholtus and De Waal (2020-2022) to approximate these probabilities for the observed value $\widehat{t}^*$ given the true value t:

$$\Pr(\widehat{t}^*|t) = \sum_{d=0}^{\infty} p(d, \widehat{t}^*) e^{-\lambda(n,q)} \frac{\lambda^d(n,q)}{d!}. \tag{22}$$

Here, $p(d, \widehat{t}^*)$ is the probability of observing $\widehat{t}^*$ links after having drawn exactly d 2-cycles. A 2-cycle (ij) represents an incorrect link between the records of unit i in file A and unit j in file B, and vice versa (see Appendix B). This probability is conditional on the true number of links t, so it actually is $p(d, \widehat{t}^*|t)$. In practice, only the first few terms of the sum in (22) have to be computed, as the sum can be truncated once the contributions of terms with larger values of d become negligible.

To be able to compute the probability $p(d, \widehat{t}^*)$, three different probabilities have to be computed first. The sum of these three probabilities is equal to 1 (Scholtus & De Waal, 2020-2022). The first probability is $\Pr(+|\widehat{t}^*)$. This is the probability of drawing a 2-cycle that leads to an extra linked pair of units with category j in dataset A and category k in dataset B, which leads to an increase of 1 in cell (j,k) when there are already $\widehat{t}^*$ linked pairs. In other words, it is the probability of swapping one of the $(s_k - \widehat{t}^*)$ units from dataset B with category k that is not linked to a unit from dataset A with category j yet, with one of the $(r_j - \widehat{t}^*)$ units from dataset B with another category than k that is linked to a unit from dataset A. That is:

$$\Pr(+|\widehat{t}^*) = \frac{(r_j - \widehat{t}^*)(s_k - \widehat{t}^*)}{\binom{n}{2}} = 2\frac{(r_j - \widehat{t}^*)(s_k - \widehat{t}^*)}{n(n-1)}, \tag{23}$$

since there are $(r_j - \widehat{t}^*)(s_k - \widehat{t}^*)$ 2-cycles that create a new linked pair with categories j and k and $\binom{n}{2}$ 2-cycles in total.

Next, we have the probability $\Pr(-|\widehat{t}^*)$. This is the probability of drawing a 2-cycle that leads to one linked pair of units with category j in dataset A and category k in dataset B less, which leads to a decrease of 1 in cell (j,k) when there are already $\widehat{t}^*$ linked pairs. In other words, this is the probability of swapping one of the $\widehat{t}^*$ units from dataset B with category k that is linked to a unit from dataset A with category j, with one of the $(n - r_j - s_k + \widehat{t}^*)$ units from dataset

B with another category than k that are not linked to a unit from dataset A with category j. This probability can be computed by:

$$\Pr(-|\widehat{t^*}) = \frac{\widehat{t^*}(n - r_j - s_k + \widehat{t^*})}{\binom{n}{2}} = 2\frac{\widehat{t^*}(n - r_j - s_k + \widehat{t^*})}{n(n-1)}, \tag{24}$$

since there are $\widehat{t^*}(n - r_j - s_k + \widehat{t^*})$ 2-cycles that have a linked pair less with categories i and j and $\binom{n}{2}$ 2-cycles in total.

Lastly, we have the probability of drawing a 2-cycle that has no effect on the number of linked pairs of units with category j in dataset A and category k in dataset B when we already have $\widehat{t^*}$ linked pairs, $\Pr(= |\widehat{t^*})$. Since $\Pr(+|\widehat{t^*}) + \Pr(-|\widehat{t^*}) + \Pr(= |\widehat{t^*}) = 1$, we have:

$$\begin{aligned}\Pr(= |\widehat{t^*}) &= 1 - \frac{(r_j - \widehat{t^*})(s_k - \widehat{t^*}) + \widehat{t^*}(n - r_j - s_k + \widehat{t^*})}{\binom{n}{2}} \\ &= 1 - 2\frac{r_j s_k + n\widehat{t^*} - 2(r_j + s_k)\widehat{t^*} + 2(\widehat{t^*})^2}{n(n-1)}.\end{aligned} \tag{25}$$

With these three probabilities, we can compute:

$$p(d, \widehat{t^*}) = \Pr(+|\widehat{t^*} - 1)p(d-1, \widehat{t^*} - 1) + \Pr(= |\widehat{t^*})p(d-1, \widehat{t^*}) + \Pr(-|\widehat{t^*} + 1)p(d-1, \widehat{t^*} + 1), \tag{26}$$

with boundary conditions $p(0, \widehat{t^*}) = 1$ if $\widehat{t^*} = t$ and $p(0, \widehat{t^*}) = 0$ otherwise. By computing $\Pr(+|\widehat{t^*}), \Pr(-|\widehat{t^*})$ and $\Pr(= |\widehat{t^*})$ for all feasible values of $\widehat{t^*}$, and substituting those probabilities in (26), $p(d, \widehat{t^*})$ can be computed for all $\widehat{t^*}$.

3.3.3 Applying Bayes' Rule

Now that we have the probabilities $\Pr(t = t_{jk}|\widehat{t^*} = \widehat{t^*_{jk}})$, we can apply Bayes' rule to reverse this conditional probability to find $\Pr(t = t_{jk}|\widehat{t^*} = \widehat{t^*_{jk}})$:

$$\Pr(t|\widehat{t^*}) = \frac{\Pr(\widehat{t^*}|t)\Pr_0(t)}{\sum_{z=L'}^{U'} \Pr(\widehat{t^*}|t = z)\Pr_0(z)}. \tag{27}$$

These probabilities describe the likelihood that t_{jk} is the true value for a certain cell (j, k), given the observed value, the assumed model for linkage errors, and the prior distribution (Scholtus & De Waal, 2020-2022). With these probabilities, new correction methods can be constructed. Unlike the previous existing and regularised estimators in Sections 3.1 and 3.2, we do not have explicit formulas for the bias and variance of the new estimators that exploit (27). The bias and

variance of the following estimators could be estimated using the bootstrap method, as performed in Chipperfield and Chambers (2015). However, this is beyond the scope of this project.

3.3.4 Expected value of the contingency table

The first correction method that is constructed, uses the probabilities $\Pr(t = t_{jk}|\widehat{t}^* = \widehat{t}^*_{jk})$ to directly compute the expected value $\mathbb{E}(\mathbf{T}|\widehat{\mathbf{T}}^*)$. The expected value of true links in cell (j,k) given the observed links $\widehat{t}^*_{jk}$ is given by:

$$t^{(\mathbb{E})}_{jk} = \mathbb{E}(t_{jk}|\widehat{t}^*_{jk}) = \sum_{t=L'}^{U'} t \cdot \Pr(t|\widehat{t}^*). \tag{28}$$

Denote the error per element as $e^{(\mathbb{E})}_{jk} = t^{(\mathbb{E})}_{jk} - t_{jk}$. The expected contingency table can also be used in combination with the regularised estimator from Section 3.2, by using it in the computations of ν_{opt} in (16).

3.3.5 Weighted correction method by using MSEs

The second correction method uses the estimated mean square errors (MSE) to weight the three existing correction methods from Scholtus et al. (2022) (see Section 3.1). That is,

$$\widehat{t}^W_{jk} = \frac{\widehat{MSE}(\widehat{t}^Q_{jk})^{-1}\widehat{t}^Q_{jk} + \widehat{MSE}(\widehat{t}^*_{jk})^{-1}\widehat{t}^*_{jk} + \widehat{MSE}(\widehat{t}^{BC}_{jk})^{-1}\widehat{t}^{BC}_{jk}}{\widehat{MSE}(\widehat{t}^Q_{jk})^{-1} + \widehat{MSE}(\widehat{t}^*_{jk})^{-1} + \widehat{MSE}(\widehat{t}^{BC}_{jk})^{-1}}, \tag{29}$$

where

$$\widehat{MSE}(\widehat{t}^Q_{jk}) = \widehat{B}(\widehat{t}^Q_{jk})^2 + \widehat{\mathrm{Var}}(\widehat{t}^Q_{jk}) = \left(-(1 + \tfrac{nq-1}{n-1})\delta(n\widehat{t}^{(\mathbb{E})}_{jk} - r_j s_k)\right)^2 + \tfrac{(nq-1)^2}{(n-1)^2}\widehat{\mathrm{Var}}(\widehat{t}^*_{jk}),$$

$$\widehat{MSE}(\widehat{t}^{BC}_{jk}) = \widehat{B}(\widehat{t}^{BC}_{jk})^2 + \widehat{\mathrm{Var}}(\widehat{t}^{BC}_{jk}) = 0 + \tfrac{(n-1)^2}{(nq-1)^2}\widehat{\mathrm{Var}}(\widehat{t}^*_{jk}),$$

$$\widehat{MSE}(\widehat{t}^*_{jk}) = \widehat{B}(\widehat{t}^*_{jk})^2 + \widehat{\mathrm{Var}}(\widehat{t}^*_{jk}) = \left(-\delta(n\widehat{t}^{(\mathbb{E})}_{jk} - r_j s_k)\right)^2 + \widehat{\mathrm{Var}}(\widehat{t}^*_{jk}).$$

The estimated bias for the naïve correction method is computed using the expressions for the bias in (11), (12), and (13); where t_{jk} is replaced by $t^{(\mathbb{E})}_{jk}$. The estimated variance of the naïve correction method is computed using formula (14), where again t_{jk} is replaced by $t^{(\mathbb{E})}_{jk}$. This estimator weights for each individual cell of the contingency table which of the three correction methods from Section 3.1 (the naïve, $\mathbf{Q}$, or $\mathbf{Q}^{-1}$ approach) performs best and weights them based on their estimated MSE.

4 Simulation study

In this section, all the correction methods are tested by means of a simulation study. The settings for the simulation study are presented first. Thereafter, the results of the simulation study are presented.

4.1 Simulation design

The simulation study uses an extended version of the design that was used in Scholtus et al. (2022). It is based on a dataset of $n = 300$ records. As mentioned in Section 2.4, we assume we are dealing with one block of data and a single error rate according to the exchangeable linkage error model. Nine different error matrices $\mathbf{Q}$ will be considered, with error rates $q = 0.9, 0.8, \ldots, 0.1$ on the diagonal. All off-diagonal elements of the $\mathbf{Q}$ matrices are equal to $\delta = \frac{1-q}{n-1} = \frac{1-q}{299}$.

First, four continuous variables are generated: $x \sim N(20, 10^2)$, $z \sim N(20, 15^2)$, $y = 20 + 2x + 1z + \epsilon$, where $\epsilon \sim N(0, 4^2)$, and $w \sim N(20, 10^2)$. Then the variables y, z, and w are discretised into five categories each and converted to vectors of dummy variables $\mathbf{Y}_g$, $\mathbf{Z}_g$, and $\mathbf{W}_g$, respectively. Next, the dataset is split into two files. File A contains the vector $\mathbf{Y}_g$ and matching variables. File B contains the vectors $\mathbf{Z}_g$ and $\mathbf{W}_g$ and matching variables. Two true contingency tables are obtained, namely $\mathbf{Y}_g^\mathsf{T}\mathbf{Z}_g$ with dependent attributes (i.e. the target variables in file A and B are dependent on each other), and $\mathbf{Y}_g^\mathsf{T}\mathbf{W}_g$ with independent attributes (i.e. the target variables in file A and B are independent of each other).

For each error matrix $\mathbf{Q}$, 10,000 permutation matrices are generated. The approach that was used to generate the permutation matrices is explained in Appendix B. For each permutation matrix, a row of file A is linked with the corresponding row of file B according to the permutation matrix. This way, we simulate the situation where we observe $\mathbf{Z}_g^* = \mathbf{C}\mathbf{Z}_g$ and $\mathbf{W}_g^* = \mathbf{C}\mathbf{W}_g$ instead of the true (and unknown) $\mathbf{Z}_g$ and $\mathbf{W}_g$, respectively. Next, on each of the 10,000 linked files, 5×5 corrected contingency tables of $\mathbf{Y}^\mathsf{T}\mathbf{Z}_g$ and $\mathbf{Y}_g^\mathsf{T}\mathbf{W}_g$ are produced after using the following correction methods:

1. the naïve approach (8): $\mathbf{Y}_g^\mathsf{T}\mathbf{Z}_g^*$ and $\mathbf{Y}_g^\mathsf{T}\mathbf{W}_g^*$, without any correction.
2. the $\mathbf{Q}$ approach (9): $(\mathbf{Q}\mathbf{Y}_g)^\mathsf{T}\mathbf{Z}_g^*$ and $(\mathbf{Q}\mathbf{Y}_g)^\mathsf{T}\mathbf{W}_g^*$.
3. the $\mathbf{Q}^{-1}$ approach (10): $\mathbf{Y}_g^\mathsf{T}\mathbf{Q}^{-1}\mathbf{Z}_g^*$ and $\mathbf{Y}_g^\mathsf{T}\mathbf{Q}^{-1}\mathbf{W}_g^*$. For this estimator, estimated cell values can become negative. These negative values are set equal to zero.
4. the expected value approach (28): $\mathbb{E}(\mathbf{Y}_g^\mathsf{T}\mathbf{Z}_g)$ and $\mathbb{E}(\mathbf{Y}_g^\mathsf{T}\mathbf{W}_g)$, by using the first (18), second (19), or third (21) prior distribution.
5. the regularised approach (15): $\mathbf{Y}_g^\mathsf{T}[\nu\mathbf{Q}^\mathsf{T}\mathbf{Q}+(1-\nu)\mathbf{I}]^{-1}\mathbf{Q}^\mathsf{T}\mathbf{Z}_g^*$ and $\mathbf{Y}_g^\mathsf{T}[\nu\mathbf{Q}^\mathsf{T}\mathbf{Q}+(1-\nu)\mathbf{I}]^{-1}\mathbf{Q}^\mathsf{T}\mathbf{W}_g^*$, where different values for ν are used, namely:
   (a) ν_{prag} (17)

(b) ν_{opt} (16) computed with $\mathbf{Y}_g^\top \mathbf{Z}_g$ and $\mathbf{Y}_g^\top \mathbf{W}_g$

(c) ν_{opt} computed with $\mathbf{Y}_g^\top \mathbf{Z}_g^*$ and $\mathbf{Y}_g^\top \mathbf{W}_g^*$

(d) ν_{opt} computed with the three different variants of $\mathbb{E}(\mathbf{Y}_g^\top \mathbf{Z}_g)$ and $\mathbb{E}(\mathbf{Y}_g^\top \mathbf{W}_g)$

6. the weighted MSE approach (29), computed with the three different variants of $\mathbb{E}(\mathbf{Y}_g^\top \mathbf{Z}_g)$ and $\mathbb{E}(\mathbf{Y}_g^\top \mathbf{W}_g)$

In total, we obtain 15 corrected contingency tables for $\mathbf{Y}_g^\top \mathbf{Z}_g$ with dependent attributes and 15 corrected contingency tables for $\mathbf{Y}_g^\top \mathbf{W}_g$ with independent attributes, for each of the 10,000 permutation matrices. Moreover, the whole simulation is repeated 10 times, with different true contingency tables $\mathbf{Y}^\top \mathbf{Z}_g$ and $\mathbf{Y}^\top \mathbf{W}_g$ by setting a different random seed. The average Cramer's V (Sakoda, 1977) of the ten tables with dependent attributes is 0.3167 with a minimum of 0.2482 and a maximum of 0.3534. For the ten tables with independent attributes, the average Cramer's V is equal to 0.1145 with a minimum of 0.0901 and a maximum of 0.1590. Note that the regularised approach where ν_{opt} is computed with $\mathbf{Y}_g^\top \mathbf{Z}_g$ and $\mathbf{Y}_g^\top \mathbf{W}_g$ cannot be used in practice, as the true contingency tables are unknown. However, these are included in the simulation study to be able to compare the performance of the regularised approach where ν_{opt} is computed with the true contingency tables with the performances of the other regularised approaches where ν_{opt} is estimated by using the observed and expected contingency tables.

All the calculations of the above-mentioned simulation are performed in RStudio, version 2022.2.1.461 (RStudio Team, 2022). The code that was used for the simulation study is stored in a repository. The link to this repository can be found in Appendix C. For computational reasons, some adjustments are made to the theoretical calculations covered in Section 3. In Section 3.3.1, prior probabilities are computed for each individual contingency table cell, for all values of t from L up to and including U. During the simulation study, it appeared that the number of possible values for t can be too large to be able to compute the probabilities that are needed for the new estimators. Therefore, the potential true values of t are considered from $L' = \max(L, \widehat{t}^* - 19) \geq L$ up to and including $U' = \min(U, \widehat{t}^* + 19) \leq U$. Moreover, a different approach is used to calculate the probabilities from Sections 3.3.2 and 3.3.3. The numerical procedure to obtain $p(d, \widehat{t}^*)$ and the reversed probabilities $\Pr(t|\widehat{t}^*)$ are described in matrix-vector notation and programmed using this approach. The full explanation of this approach can be found in Appendix D.

4.2 Simulation results

In this section, we will look at the results of the simulation study. First, we will extract two contingency tables $\mathbf{Y}_g^\top \mathbf{Z}_g$ with dependent attributes and $\mathbf{Y}_g^\top \mathbf{W}_g$ with independent attributes that were used in the simulation study. We also look at four corresponding naïve contingency tables. After that, we will inspect the general results of the simulation study by looking at the average percentages where the naïve approach outperforms the alternative approach and the average total relative differences from the naïve empirical RMSE. Finally, we will go back to the

two examples. Based on the general results, one correction approach that performs well and one that performs worse are applied to these tables.

4.2.1 Example tables from the simulation study

To give a better idea of the simulation study that was performed, two examples of the contingency tables are extracted and shown in detail. In Table 3, one of the ten true contingency tables $\mathbf{Y}_g^\top \mathbf{Z}_g$ with dependent attributes is shown. In Table 4 one of the ten true contingency tables $\mathbf{Y}_g^\top \mathbf{W}_g$ with independent attributes is shown.

Table 3
One of the Ten Contingency Tables $\mathbf{Y}_g^\top \mathbf{Z}_g$ with Dependent Attributes that Was Used in the Simulation Study

	C_1	C_2	C_3	C_4	C_5
R_1	4	3	1	0	0
R_2	4	17	17	9	0
R_3	1	18	59	27	4
R_4	0	8	42	36	15
R_5	0	1	9	16	9

Table 4
One of the Ten Contingency Tables $\mathbf{Y}_g^\top \mathbf{W}_g$ with Independent Attributes that Was Used in the Simulation Study

	C_1	C_2	C_3	C_4	C_5
R_1	0	2	3	3	0
R_2	2	10	23	10	2
R_3	0	30	56	19	4
R_4	1	20	45	23	12
R_5	1	8	15	8	3

For this example, we will only look at two error matrices $\mathbf{Q}$ with $q = 0.8$ and $q = 0.2$ in detail. Note that for the error matrix with $q = 0.8$, 80% of the links are correct and 20% of the links are incorrect in expectation. Similarly, 20% of the links are correct and 80% of the links are incorrect in expectation for the error matrix with $q = 0.2$. Two examples of the observed naïve contingency table $\mathbf{Y}_g^\top \mathbf{Z}_g^*$ with dependent attributes are given in Table 5 and Table 6 with a probability of a correct link of $q = 0.8$ and $q = 0.2$, respectively. When we compare these naïve contingency tables to the true contingency table in Table 3, we see that the values of Table 5 are closer to the true values compared to the values of Table 6, which is as expected.

Table 5
One of the 10,000 Naïve Contingency Tables $\mathbf{Y}_g^\top \mathbf{Z}_g^*$ with Dependent Attributes where the Probability of a Correct Link (q) Is 0.8

	C_1	C_2	C_3	C_4	C_5
R_1	4	2	1	1	0
R_2	4	15	18	9	1
R_3	1	17	56	29	6
R_4	0	10	43	34	14
R_5	0	3	10	15	7

Table 6
One of the 10,000 Naïve Contingency Tables $\mathbf{Y}_g^\top \mathbf{Z}_g^*$ with Dependent Attributes where the Probability of a Correct Link (q) Is 0.2

	C_1	C_2	C_3	C_4	C_5
R_1	1	0	4	3	0
R_2	2	8	19	11	7
R_3	3	17	46	31	12
R_4	2	16	42	35	6
R_5	1	6	17	8	3

Two examples of the observed naïve contingency tables $\mathbf{Y}_g^\top \mathbf{W}_g^*$ with independent attributes are given in Table 7 and Table 8 with a probability of a correct link of $q = 0.8$ and $q = 0.2$,

respectively. Comparing these naïve contingency tables to the true contingency table in Table 4, we see that the values of Table 7 are closer to the true values compared to the values of Table 8, which again is as expected.

Table 7
One of the 10,000 Naïve Contingency Tables $\mathbf{Y}_g^\mathsf{T}\mathbf{W}_g^*$ *with Independent Attributes where the Probability of a Correct Link (q) Is 0.8*

	C_1	C_2	C_3	C_4	C_5
R_1	0	2	3	3	0
R_2	3	9	23	11	1
R_3	0	28	57	20	4
R_4	1	25	40	22	13
R_5	0	6	19	7	3

Table 8
One of the 10,000 Naïve Contingency Tables $\mathbf{Y}_g^\mathsf{T}\mathbf{W}_g^*$ *with Independent Attributes where the Probability of a Correct Link (q) Is 0.2*

	C_1	C_2	C_3	C_4	C_5
R_1	0	1	4	3	0
R_2	2	10	14	16	5
R_3	1	31	56	17	4
R_4	1	20	47	22	11
R_5	0	8	21	5	1

If we look at the values of the naïve tables with dependent attributes in Table 5 and 6, we see that the values are very different for $q = 0.8$ and $q = 0.2$. For the naïve tables with independent attributes in Table 7 and 8, the values are less different for $q = 0.8$ and $q = 0.2$. Moreover, values of the naïve contingency tables with independent attributes differ less from the true contingency table in Table 4 than the tables with dependent attributes differ from the true contingency table in Table 3. An explanation for this is that linkage errors, provided they comply with the exchangeable linkage error assumption, cause more damage to dependent tables (i.e. tables where the target variables are dependent on each other) than to independent tables (i.e. tables where the target variables are independent of each other). Incorrectly linking records that do not belong to the same entity disturbs the relationships between $\mathbf{Y}$ and $\mathbf{Z}$ by attenuating the association towards zero. As a result, the association between $\mathbf{Y}$ and $\mathbf{Z}$ in the naïve table is underestimated when the target variables are dependent. On the other hand, if there is no dependency between the target variables, the true association is already zero. The expected association then remains equal to zero after incorrectly linking records. In Zhang and Tuoto (2021), these findings were also shown in the context of regression models and using non-informative linkage errors instead of exchangeable linkage errors.

4.2.2 General results of the simulation study

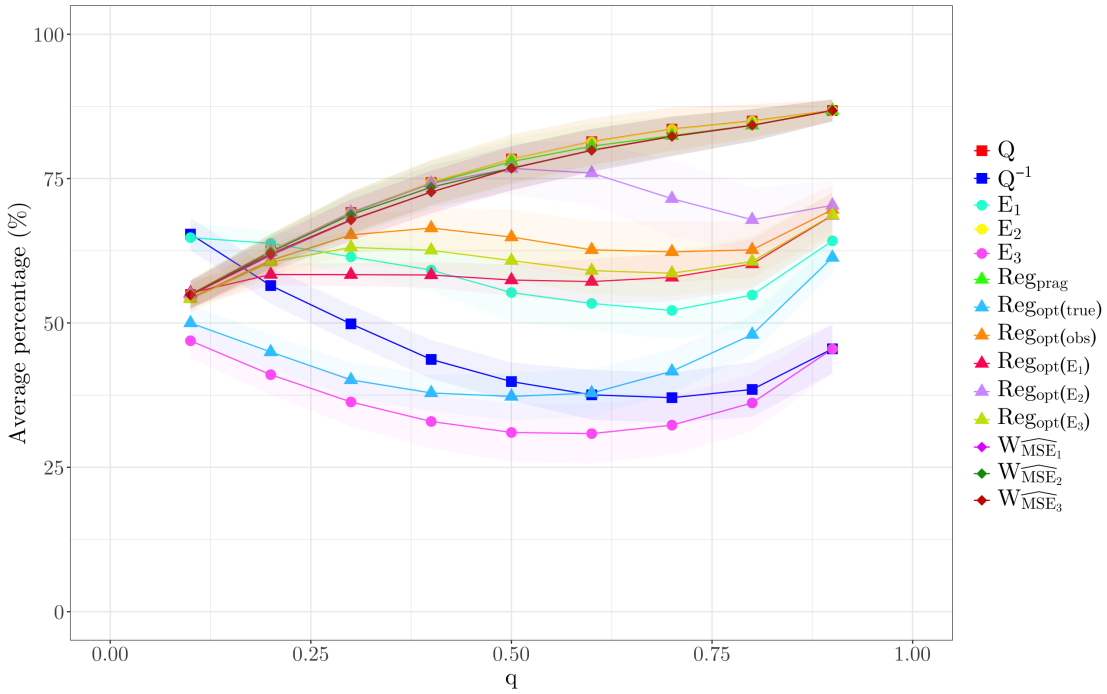
In what follows, we will look at the overall results of the simulation study. In the figures, short names are used to indicate the alternative approaches. A list of these short names and the corresponding approaches can be found in Table E.10 in Appendix E.

We begin by assessing whether the naïve approach is expected to perform better than the alternative approaches for each of the 25 individual cells of the tables among all the 10,000 tables produced from the linked files. In other words, this is the percentage of times that $|e_{jk}^{\text{alternative}}| > |e_{jk}^*|$. Then, the average of all 25 individual cells is computed over the tables.

Thereafter, we again take the average of these 10,000 percentages. This is repeated for the different values of q. A low average percentage would indicate that the corresponding approach is expected to perform better than the naïve approach, which is desirable. The average percentages over the 10 contingency tables with dependent attributes with corresponding standard deviation are shown in Figure 2. The full results can be found in Table E.11 in Appendix E.

Figure 2

Average Percentages where the Naïve Approach Outperforms the Alternative Approach over the 10 Generated Contingency Tables with Dependent Attributes for Different Error Rates q with Corresponding Standard Deviation



Note. The values of q are on the x-axis and the average percentage is on the y-axis. Each alternative approach has its own colour, which can be found in the legend. Moreover, each group of estimators has its own shape to indicate the data points. The existing estimators from Section 3.1 are indicated by a square, the expected values from Section 3.3.4 by a dot, the regularised estimators from Section 3.2 by a triangle, and the weighted MSE estimators from Section 3.3.5 by a diamond. The percentages are the average over the ten contingency tables with dependent attributes. The ribbon around the graph in the same colour as the line and data points corresponding to the approach indicates one standard deviation above and below the average percentage. For an overview of the short names used for the alternative approaches, see Table E.10 in Appendix E.

In the plot, we see that the $\mathbf{Q}^{-1}$ approach has the highest average percentage for $q = 0.1$. However, as q increases, the average percentage that the naïve approach will outperform the $\mathbf{Q}^{-1}$ approach decreases and is below most of the alternative approaches. For values $q > 0.3$, the average percentage is below 50%. The expected value with the first (uninformative) prior distribution also has a high average percentage for $q = 0.1$, which slightly decreases as q increases.

For $q > 0.7$ the average percentage increases again. For this approach, the average percentage being outperformed by the naïve approach is always above 50%.

The $\mathbf{Q}$ approach, the expected value approach with the second prior, the regularised approach with the pragmatic choice for ν and the three variants of the weighted MSE approach perform the worst for values of q between 0.5 and 0.9. As the value of q increases, the average percentages of the naïve approach outperforming these six alternative approaches also increases. Overall, the average percentages remain above 50% for these six alternative approaches for all considered values of q.

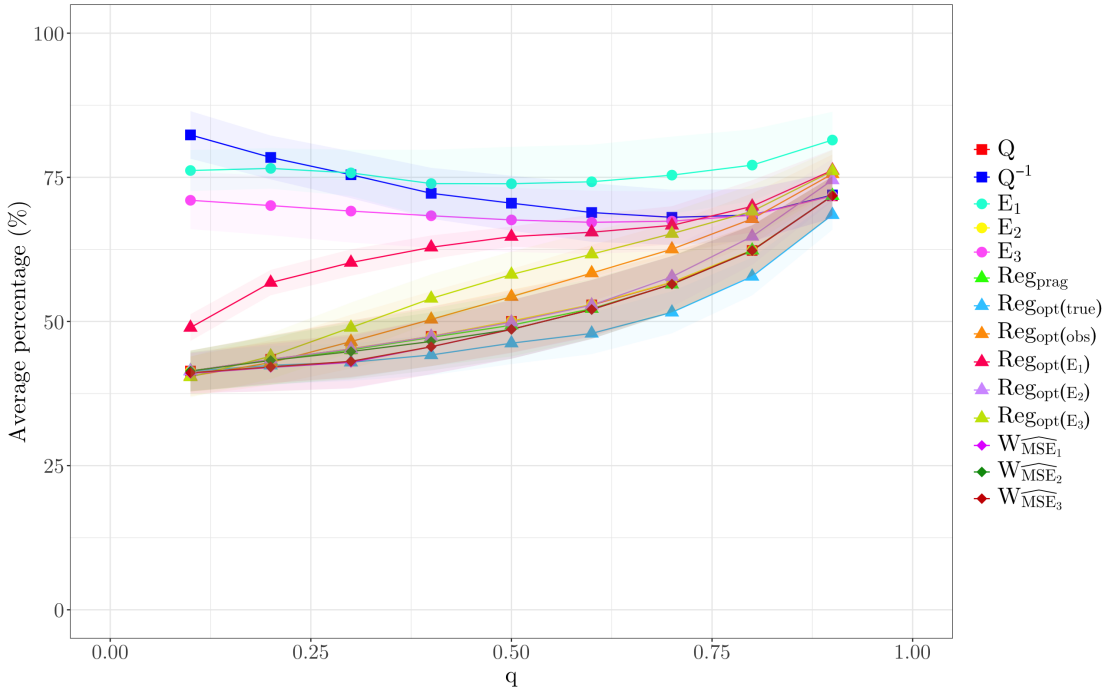
The regularised approaches where the optimal value of ν is estimated with the observed and the three expected contingency tables, perform all similarly. Between the values $q = 0.2$ and $q = 0.8$, the average percentages slightly increase for all types of the regularised estimator, where the regularised estimator with ν estimated with the second prior rises the most and the regularised estimator with ν estimated with the first prior rises the least. When we compare these four regularised estimators with the regularised approach that uses the true contingency table to compute the optimal ν, it can be concluded that the average percentage is higher if we do not use the true contingency table for all values of q. If we knew the true contingency table in practice, the regularised estimator with optimal ν would perform better. From these results, it appears that the expected value with the first prior estimates ν_{opt} best for all values of q. However, using an estimated value of ν_{opt}, the average percentages results in considerably higher percentages than using the true value of ν_{opt}.

The expected value approach with the third prior distribution has the lowest average percentage for all values of q that were considered. The average percentage is below 50% for all values of q. Hence, no matter what value of q, the expected value approach with the third prior is likely better than the naïve approach when the table has dependent attributes. The $\mathbf{Q}^{-1}$ approach is also likely to perform better than the naïve approach for values of $q \geq 0.4$. For smaller values of q, the average percentages of the naïve approach outperforming the alternative approaches are all close to each other. For larger values of q, the average percentages are more separated compared to the smaller values of q.

The average percentages over the 10 contingency tables with independent attributes with corresponding standard deviation are shown in Figure 3. The full results can be found in Table E.13 in Appendix E. Again, the lower the average percentage, the better the alternative approach performs in comparison to the naïve approach.

Figure 3

Average Percentages where the Naïve Approach Outperforms the Alternative Approach over the 10 Generated Contingency Tables with Independent Attributes for Different Error Rates q with Corresponding Standard Deviation



Note. The values of q are on the x-axis and the average percentage is on the y-axis. Each alternative approach has its own colour, which can be found in the legend. Moreover, each group of estimators has its own shape to indicate the data points. The existing estimators from Section 3.1 are indicated by a square, the expected values from Section 3.3.4 a dot, the regularised estimators from Section 3.2 by a triangle, and the weighted MSE estimators from Section 3.3.5 by a diamond. The percentages are the average over the ten contingency tables with independent attributes. The ribbon around the graph in the same colour as the line and data points corresponding to the approach indicates one standard deviation above and below the average percentage. For an overview of the short names used for the alternative approaches, see Table E.10 in Appendix E.

We see that the $\mathbf{Q}^{-1}$ approach has the highest average percentage for smaller values of q and slightly decreases as q increases. The average percentages of the expected value approaches computed with the first and third prior both remain around the same value for all considered values of q. The average percentages for these two approaches are above 63% for all values of q. The expected value approach with the second prior performs better. For values of $q \leq 0.4$ the average percentage of the naïve approach outperforming this alternative approach is below 50%. As the value of q increases, the average percentage also increases. For values of $q > 0.4$, the average percentage is above 50%.

When we look at the regularised estimators, we see that the regularised estimator with ν_{opt} estimated with the expected value with the first prior has the highest average percentage of all

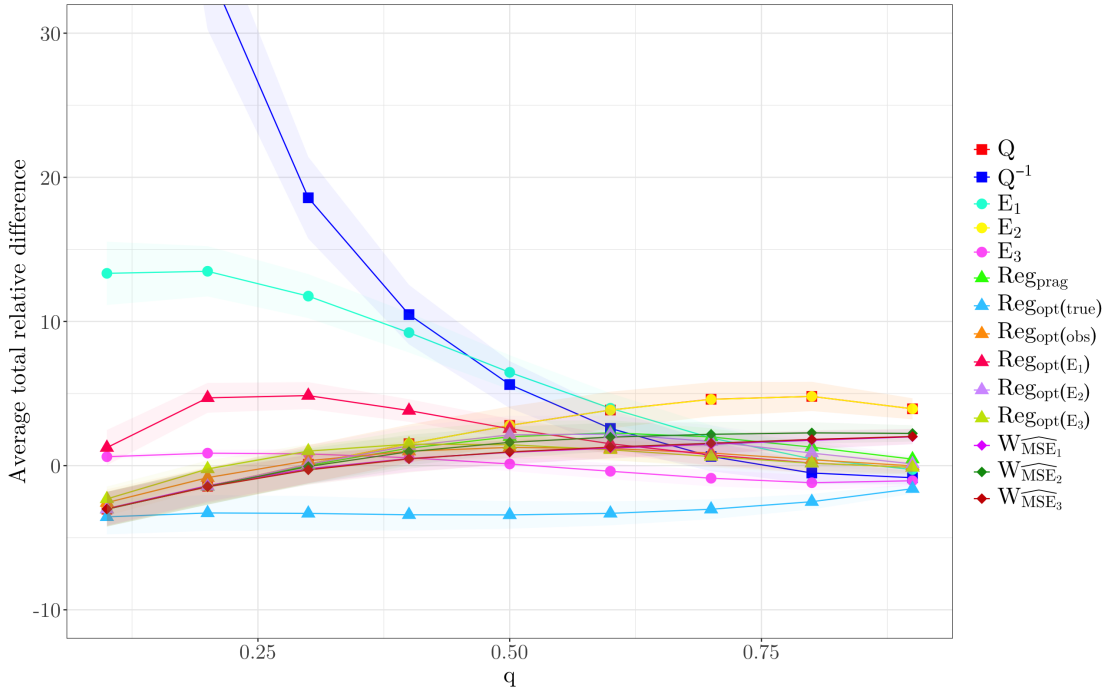
variants of the regularised estimator for all values of q. All the regularised estimators where ν_{opt} is estimated are outperformed by the naïve approach for $q > 0.4$. The average percentages are below 50% for all the regularised estimators with estimated ν_{opt} when $q \leq 0.3$, except when using the first variant of the expected value. For almost all values of q, the regularised estimator with the pragmatic choice for ν has a lower average percentage compared to the other regularised estimators that use an estimated value of ν_{opt}. Hence, the optimal value of ν appears not to be optimal for tables with independent attributes when ν_{opt} has to be estimated. The regularised approach that uses the true contingency table to compute ν_{opt} has the lowest average percentage of all alternative approaches for almost all values of q. So, just like for the contingency tables with dependent attributes, the regularised estimator with optimal ν would perform better if we knew the true contingency table in practice.

The $\mathbf{Q}$ approach and the three weighted MSE approaches all perform quite similarly and have the lowest average percentages of all the alternative approaches that can be used in practice. Overall, it seems that most approaches perform better for smaller values of q. The higher the probability of a correct link, the higher the average percentage where the naïve approach outperforms the alternative approach. For larger values of q, the average percentages of the naïve approach outperforming the alternative approaches are all close to each other. For smaller values of q, the average percentages are more separated compared to the larger values of q. This is the opposite of what we saw for tables with dependent attributes in Figure 2.

Next, the relative difference from the naïve approach for the alternative approach is investigated. That is: $(RMSE^{\text{alternative}} - RMSE^{\text{naïve}})/RMSE^{\text{naïve}}$, where RMSE is the empirical root mean square error. The relative difference is computed for each individual cell of the contingency table, after which the relative differences of all 25 cells are added up for each table. Using the relative difference makes it easy to see whether the alternative approaches improve upon the naïve approach or not. Negative values mean that the alternative approach performs better than the naïve approach, and positive values mean that the naïve approach performs better than the alternative approach. In Figure 4, the average total relative differences over the 10 contingency tables with dependent attributes with corresponding standard deviations are given for all considered values of q. Note that the y-axis is truncated to be able to see all the graphs with negative values more clearly. Because of the high average total relative differences for the $\mathbf{Q}^{-1}$ approach for small values of q, the lines of the other alternative approaches became hard to recognize. The average total relative difference of the $\mathbf{Q}^{-1}$ approach keeps increasing as q becomes smaller, up until an average total relative difference of approximately 80. The full plot and a table with the corresponding results can be found in Figure E.6 and Table E.12 in Appendix E, respectively.

Figure 4

The Average Total Relative Differences from the Naïve Empirical RMSE over the 10 Generated Contingency Tables with Dependent Attributes for Different Error Rates q with Corresponding Standard Deviation



Note. The values of q are on the x-axis and the average total relative difference is on the y-axis. Each alternative approach has its own colour, which can be found in the legend. Moreover, each group of estimators has its own shape to indicate the data points. The existing estimators from Section 3.1 are indicated by a square, the expected values from Section 3.3.4 by a dot, the regularised estimators from Section 3.2 by a triangle, and the weighted MSE estimators from Section 3.3.5 a diamond. The total relative differences are the average over the ten contingency tables with dependent attributes. The ribbon around the graph in the same colour as the line and data points corresponding to the approach indicates one standard deviation above and below the average total relative difference. For clarity, the y-axis is truncated from -10 to 30. The graph of the $\mathbf{Q}^{-1}$ approach keeps increasing until approximately 80 for $q < 0.2$. For an overview of the short names used for the alternative approaches, see Table E.10 in Appendix E.

In the plot, we see more positive values than negative values for the average total relative difference. The average total relative difference for the $\mathbf{Q}$ approach is negative for $q = 0.1$ and $q = 0.2$ and increases and becomes positive for larger values of q. The $\mathbf{Q}^{-1}$ approach has very high average total relative differences for small values of q which decrease as q increases. For $q = 0.8$ and $q = 0.9$, the average total relative difference is negative.

When we look at the expected value approaches with the different prior distributions, we see that when using the first prior the average total relative difference is barely negative for $q = 0.9$ and becomes positive for all smaller values of q. When we use the second prior the average total relative difference is negative for $q \leq 0.2$ and positive for larger values of q. Using the third prior

results in negative average total relative differences for values of q between 0.6 and 0.9. When the probability of a correct link becomes smaller than 0.6, the average total relative difference becomes positive.

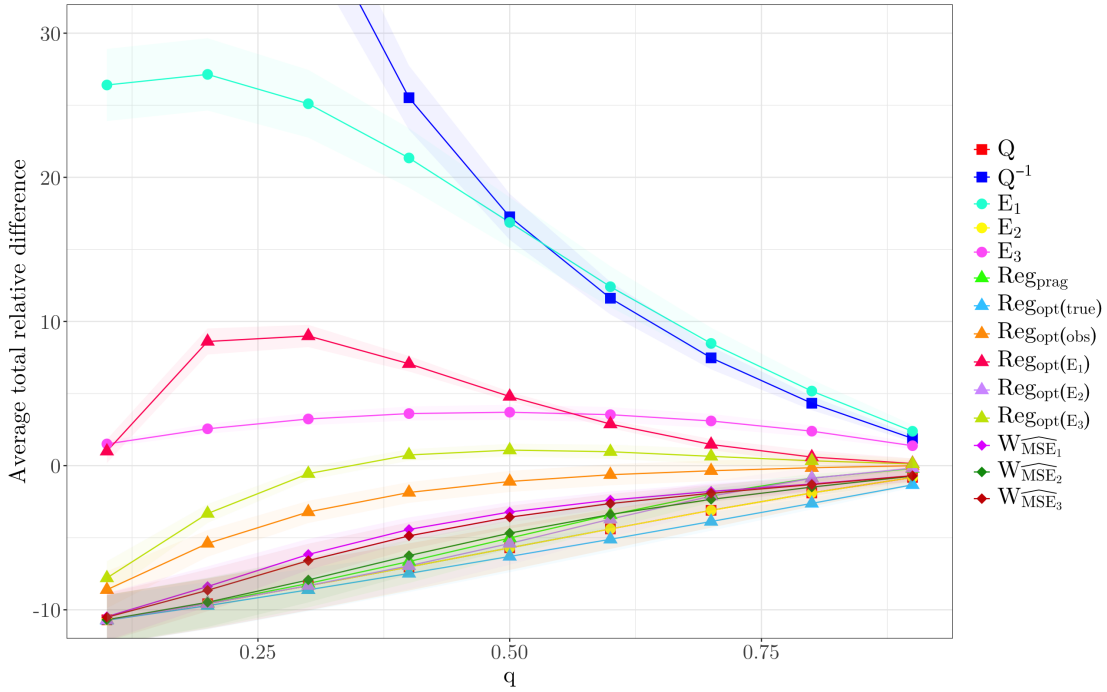
The regularised estimator where the optimal value of ν is computed with the true contingency table performs best of all estimators. Unfortunately, this estimator cannot be used in practice, as mentioned previously. The other regularised estimators do not perform as well. Only for $q \leq 0.2$ the regularised estimators with ν_{prag} and ν_{opt} estimated with the observed contingency table and the expected values with priors two and three have a negative average total relative difference.

The three versions of the weighted estimators all perform quite similarly. The average total relative differences are positive for values of q between 0.4 and 0.9. For $q \leq 0.3$, the average total relative difference becomes negative for all three weighted estimators. Using the third variant of the expected value in the computations seems to give the best results, but the performances of the other variants are close.

In Figure 5, the average total relative differences over the 10 contingency tables with independent attributes with corresponding standard deviations are given for all considered values of q. Note that again the y-axis is truncated to be able to see all the graphs with negative values more clearly. This is because of the same reason as in the previous figure, namely the high average total relative differences for the $\mathbf{Q}^{-1}$ approach for small values of q. The average total relative difference of the $\mathbf{Q}^{-1}$ approach keeps increasing as q becomes smaller, up until an average total relative difference of approximately 140. The full plot and a table with the corresponding results can be found in Figure E.7 and Table E.14 in Appendix E.

Figure 5

The Average Total Relative Differences from the Naïve Empirical RMSE over the 10 Generated Contingency Tables with Independent Attributes for Different Error Rates q with Corresponding Standard Deviation



Note. The values of q are on the x-axis and the average total relative difference is on the y-axis. Each alternative approach has its own colour, which can be found in the legend. Moreover, each group of estimators has its own shape to indicate the data points. The existing estimators from Section 3.1 are indicated by a square, the expected values from Section 3.3.4 by a dot, the regularised estimators from Section 3.2 by a triangle, and the weighted MSE estimator from Section 3.3.5 by a diamond. The total relative differences are the average over the ten contingency tables with independent attributes. The ribbon around the graph in the same colour as the line and data points corresponding to the approach indicates one standard deviation above and below the average total relative difference. For clarity, the y-axis is truncated from -10 to 30. The graph of the $\mathbf{Q}^{-1}$ approach keeps increasing until approximately 140 for $q < 0.4$. For an overview of the short names used for the alternative approaches, see Table E.10 in Appendix E.

In this plot, we have more negative values for the average total relative difference in comparison to the plot for tables with dependent attributes in Figure 4. The $\mathbf{Q}$ approach is clearly outperforming the $\mathbf{Q}^{-1}$ approach for all values of q. The smaller the value of q, the higher the average total relative difference from the naïve empirical RMSE for the $\mathbf{Q}^{-1}$ approach. It is the other way around for the $\mathbf{Q}$ approach: the smaller q, the lower the average relative difference. Moreover, the average total relative difference is negative for all values of q for the $\mathbf{Q}$ approach.

The expected value approach using the second prior gives the same results as the $\mathbf{Q}$ approach and thus performs well. The expected value approach using the first prior does not perform well. The smaller the value of q, the higher the average total relative difference. The performance of

the third variant of the expected value approach is between the other two. It only has positive average total relative differences for all values of q.

Again, we see that the regularised estimator with ν_{opt} computed with the true contingency table performs best of all the variants of the regularised estimators and of all the alternative approaches in general. For all regularised estimators that can be used in practice, the average total relative difference becomes more negative as the probability of a correct link becomes smaller, except when using the first variant of the expected value. The regularised estimator with the pragmatic choice of ν also performs well, better than most of the regularised estimators where an estimated value of ν_{opt} is used. Hence, again the optimal value of ν does not appear to be optimal for tables with independent attributes when it has to be estimated.

The weighted approaches all have negative average total relative differences, which become more negative as the value of q becomes smaller. The three variants perform similarly, but the weighted estimator using the second expected value seems to perform best.

4.2.3 Example tables revisited

Now that we have investigated the average percentages where the naïve approach outperforms the alternative approaches and the average total relative difference from the naïve empirical RMSE, we go back to the contingency tables that we looked at in detail in Tables 3 and 4. We will now show the application of one correction approach that performs well and one correction approach that performs worse for $\mathbf{Y}_g^\intercal \mathbf{Z}_g$ with $q = 0.8$. The application of one correction approach that performs well and one correction approach that performs worse for $\mathbf{Y}_g^\intercal \mathbf{Z}_g$ with $q = 0.2$ and for $\mathbf{Y}_g^\intercal \mathbf{W}_g$ with $q = 0.8$ and $q = 0.2$ can be found in Appendix F.

In the results, we saw that for tables with dependent attributes and larger values of q the expected value approach with the third prior distribution performed well and the $\mathbf{Q}$ approach performed poorly. The naïve contingency table $\mathbf{Y}_g^\intercal \mathbf{Z}_g^*$ with $q = 0.8$ that we saw in Table 5 after correction with these two correction methods are shown in Table 9.

Table 9

Two Examples of the Corrected Contingency Tables with Dependent Attributes where the Probability of a Correct Link (q) Was 0.8 by Using the Expected Value Approach with the Third Prior Distribution and the $\mathbf{Q}$ *Approach*

	C_1	C_2	C_3	C_4	C_5
R_1	4.63	2.13	0.72	0.82	0.00
R_2	4.44	16.30	17.69	8.31	0.66
R_3	0.73	17.00	57.44	28.60	5.37
R_4	0.00	9.16	43.04	34.70	14.76
R_5	0.00	2.63	9.22	15.80	7.66

	C_1	C_2	C_3	C_4	C_5
R_1	3.25	1.85	1.48	1.27	0.15
R_2	3.48	13.47	18.41	9.96	1.68
R_3	1.46	17.02	54.09	29.60	6.84
R_4	0.61	11.17	43.02	33.12	13.08
R_5	0.21	3.50	10.99	14.05	6.25

Note. The corrected contingency table on the left is corrected using the expected value approach that uses the third prior distribution. The corrected contingency table on the right is corrected using the $\mathbf{Q}$ approach.

The corrected table by using the expected value approach with the third prior distribution in Table 9 indeed performs very well, as the values of the corrected table are quite close to the real values. The values of the corrected contingency table with the worse correction method, the **Q** approach, differ more from the true values. If the values are all rounded to whole numbers, the corrected contingency table using the expected value approach with the third prior distribution results in 9 correct values compared to the true contingency table in Table 3. When we look at the times that $|e_{jk}^{(\mathbb{E}_3)}| > |e_{jk}^*|$, we find that this holds for 6 out of 25 cells (i.e. 24%). The total relative difference from the naïve empirical RMSE is approximately -0.281. If we do the same for the corrected contingency table using the **Q** approach, we find 4 correct values compared to the true contingency table in Table 3. It holds for 24 out of 25 cells (i.e. 96%) that $|e_{jk}^{\mathbf{Q}}| > |e_{jk}^*|$. The total relative difference from the naïve empirical RMSE is approximately 0.642.

5 Discussion

In this final section, the findings of this project are discussed and limitations and suggestions for future research are given.

Linkage errors can occur during linking data from multiple sources together. There is a growing emphasis on accounting for linkage errors in the statistical analysis of categorical data and contingency tables. This is relevant for NSIs, as their published statistics are used by government agencies and stakeholders. If the linkage errors are ignored, this can lead to biased inference. In Scholtus et al. (2022) two general approaches for compensating for linkage errors (the biased $\mathbf{Q}$ approach and the unbiased $\mathbf{Q}^{-1}$ approach) were presented and tested for dependent tables, i.e. tables where the target variables in two data files are dependent on each other, and independent tables, i.e. tables where the target variables in two data files are independent of each other. Results showed that the unbiased correction approach performed rather poorly compared to the biased and naïve approach where linkage errors were not compensated for.

The aim of this project was to develop, implement and test new methods for the correction of linkage errors in contingency tables that perform consistently better than the naïve estimator and the two previously proposed estimators presented in Scholtus et al. (2022). Moreover, the aim was to ascertain which correction method is best for a given situation by comparing their performances on simulated data. Three new approaches for compensating for linkage error when calculating and analysing two-way contingency tables for categorical data were presented: the regularised approach, the expected value approach, and the weighted MSE approach. The existing and the new approaches (and their different variants) were tested in an extended simulation study. The correction methods were both tested on dependent tables and independent tables. Moreover, the performances of the correction approaches were tested with different matrices of linkage error probabilities $\mathbf{Q}$, with 9 different values of q (the probability of a correct link). For each of the alternative correction methods, the errors and the root mean square error were compared to the naïve approach.

Determining the best correction approach for the cell values of a contingency table depends on if the table has dependent or independent attributes, and on how many linkage errors might be present, i.e. the probability of a correct link (q). For dependent tables, the naïve approach performs just as well or even better than most of the alternative approaches. The $\mathbf{Q}^{-1}$ approach seemed to perform well when we looked at the average percentage where this approach is outperformed by the naïve approach. If q is larger than 0.4, i.e. 40% or more of the links are correct, the average percentage is below 50%. However, when we looked at the average total relative difference for the $\mathbf{Q}^{-1}$ approach, we found high positive values for all values of $q < 0.8$. This indicates that this approach does not perform well when less than 80% of the links are correct. The expected value approach with the third prior distribution also performed well according to the average percentage of being outperformed by the naïve approach. For all values of q, the

average percentage was below 50%. Hence, regardless of the probability of a correct link q, the expected value approach with the third prior distribution (see Section 3.3.1) is likely to perform better than the naïve approach. This is confirmed for values of $q > 0.5$ by the average total relative difference. For $q > 0.5$ the average total difference of this approach from the naïve empirical RMSE is negative, which indicates that it performs better than the alternative approach. For values $q \leq 0.5$, the average total relative difference is between 0 and 1. For these reasons, we recommend using the expected value approach with the third prior distribution for dependent tables. However, the performance may be worse if the probability of a correct link is less than 0.5. If the probability of a correct link is larger than 0.8, the $\mathbf{Q}^{-1}$ approach may also be a good alternative that requires less computational work.

For independent tables, we found that the $\mathbf{Q}$ approach, the expected value approach with the second prior distribution (see Section 3.3.1), and the three weighted MSE approaches all had the lowest average percentages where the naïve approach outperforms the alternative approach for all values of q. When we looked at the average total relative difference from the naïve empirical RMSE, we found negative values for all of these five alternative approaches for all values of q. The smaller the probability of a correct link, the better the correction methods perform. Moreover, the regularised estimator with the second variant of the expected value and the regularised estimator with the pragmatic choice for ν also have negative values for all q for the average total relative difference from the naïve empirical RMSE. However, the average percentages of these two approaches are always slightly higher than the previous five highlighted alternative approaches. For these reasons, we recommend using the $\mathbf{Q}$ approach. For very small values of q (i.e. $q \leq 0.2$), the weighted MSE approach with the second variant of the expected value may be a good alternative.

In Scholtus et al. (2022), it was recommended to use the $\mathbf{Q}$ approach for independent tables. This is in line with what is found in this thesis. In addition, the expected value approach with the second prior distribution performs the same as the $\mathbf{Q}$ approach. As the $\mathbf{Q}$ approach is easier to calculate than the expected value approach, it is recommended to use the $\mathbf{Q}$ approach in practice. When the probability of a correct link is very small, the weighted MSE approach with the second variant of the expected value is a good approach to use for independent tables as well. In both the previous research and this thesis, it was found that the $\mathbf{Q}^{-1}$ approach performs badly for independent tables. As only the values $q = 0.8$ and $q = 0.9$ were considered in Scholtus et al. (2022), it was shown in this thesis that the $\mathbf{Q}^{-1}$ approach also performs poorly for independent tables when the error matrix has less than 0.8 on the diagonal. For dependent tables, it was recommended in Scholtus et al. (2022) to use the naïve approach when the probability of a correct link is larger than 0.8. In this thesis, the expected value approach with the third prior distribution is recommended to use for tables with dependent attributes. Moreover, it is found in this thesis that the $\mathbf{Q}^{-1}$ approach performs well for tables with $q > 0.8$. This is in contrast to the previous findings, where it was suggested to use the naïve approach when the probability of a correct link is larger than or equal to 0.8.

We are aware that this research may have some limitations as well. First of all, three different prior distributions for t_{jk} were used to apply Bayes' rule to compute the probability of obtaining t_{jk} in cell (j,k) given the observed $\widehat{t}^*_{jk}$. Using the third prior distribution to compute the probabilities and thereafter correct with the expected value approach gave positive results for tables with dependent attributes. For tables with independent attributes, using the second prior distribution to compute the probabilities and subsequently correct with the expected value approach gave reasonable results. In future research, new priors can be constructed and used in the computations, which may lead to better results.

Secondly, if the regularised estimator with ν_{opt} is used to correct for linkage errors, the marginals of the corrected contingency table generally do not sum up to the fixed marginal totals of the true contingency table. This also holds for the expected value approach and the weighted correction approach using mean square errors. To obtain tables where the row and column totals are preserved, an additional iterative proportional fitting (IPF) can be performed (Deming & Stephan, 1940). This is not implemented in this project, but it might positively influence the correction approaches.

Thirdly, a point of improvement that can be made in future research, is the estimation of the value of ν_{opt} for the regularised approach. It was found that the regularised estimator with ν_{opt} computed with the true contingency table performed well for both tables with dependent attributes as with independent attributes. In practice, the true contingency table is unknown and ν_{opt} has to be estimated. In this project, ν_{opt} was estimated by using the naïve contingency table and by using the three variants of the expected value tables. The performance of the regularised approach with estimated ν_{opt} was worse than when ν_{opt} was computed with the true contingency table. Therefore, if the estimation of ν_{opt} can be optimised, this may lead to a better performance which can also be used in practice. We already tried to optimise the value of optimal ν by iteratively computing ν_{opt} until convergence. Unfortunately, this did not lead to better results.

The iterative method that we just mentioned might also be interesting to use in future research for any of the other methods that require an estimate for the real t_{jk} as input, i.e. all the new estimators. In this case, one would compute the estimator by using the naïve observed value as input for the first round $t = 0$. Then for each round $t > 0$, the estimator can be computed by using the estimator of round $t - 1$ as input until the estimator no longer changes. Possibly, this can lead to better estimators.

Moreover, for the existing approaches $\mathbf{Q}$ and $\mathbf{Q}^{-1}$ and the regularised approach, explicit formulas were given for the bias and variance. For the new estimators that use the probabilities of the true t_{jk} given the observed $\widehat{t}^*_{jk}$, no explicit expressions were derived in this thesis. The bias and variance of these estimators can be estimated using the bootstrap method, as performed in Chipperfield and Chambers (2015). This was out of the scope of this project but it can be performed in future research. Lastly, the correction methods are now only tested by means of a simulation study. In further research, it might be interesting to also apply and test the correction

methods on real probabilistically linked data.

In summary, we have found a new successful correction approach for dependent contingency tables, namely the expected value approach with the third prior distribution that uses the observed values of the contingency table. Moreover, it is confirmed that the **Q** approach is best to use for independent contingency tables, also when the probability of a correct link is smaller than 0.8. The regularised estimator could eventually become the most recommendable correction method to use for tables with both dependent and independent attributes, but more research is needed to get there.

References

Chambers, R. (2009). Regression analysis of probability-linked data. *Official Statistics Research Series*, *4*.

Chipperfield, J. O., Bishop, G., Campbell, P. D., et al. (2011). Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data. *Survey Methodology*, *37*(1), 13–24.

Chipperfield, J. O., & Chambers, R. L. (2015). Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. *Journal of Official Statistics*, *31*(3), 397–414.

Cox, L. H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, *82*(398), 520–524.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*(4), 427–444.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods* (Vol. 580). Springer.

Nelsen, R. B. (1987). Discrete bivariate distributions with given marginals and correlation: Discrete bivariate distributions. *Communications in Statistics-Simulation and Computation*, *16*(1), 199–208.

Neter, J., Maynes, E. S., & Ramanathan, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, *60*(312), 1005–1027.

Pijpers, F. P. (2021). *A new estimator for linkage error correction.* [Unpublished manuscript].

RStudio Team. (2022). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Sakoda, J. M. (1977). Measures of association for multivariate contingency tables. *Proceedings of the Social Statistics Section of the American Statistical Association*, 777–780.

Sayers, A., Ben-Shlomo, Y., Blom, A. W., & Steele, F. (2016). Probabilistic record linkage. *International journal of epidemiology*, *45*(3), 954–964.

Scholtus, S. (2020). *A simple probability distribution for exchangeable linkage errors.* [Unpublished manuscript].

Scholtus, S. (2023). *Some further notes on probabilities for contingency tables with linkage errors.* [Unpublished manuscript].

Scholtus, S., & De Waal, T. (2020-2022). *Correcting for linkage errors in contingency tables: Using probabilities to improve the correction approach.* [Unpublished manuscript].

Scholtus, S., Shlomo, N., & De Waal, T. (2022). Correcting for linkage errors in contingency tables—a cautionary tale. *Journal of Statistical Planning and Inference*, *218*, 122–137.

Tromp, M., Ravelli, A. C., Bonsel, G. J., Hasman, A., & Reitsma, J. B. (2011). Results from simulated data sets: Probabilistic record linkage outperforms deterministic record linkage. *Journal of clinical epidemiology*, *64*(5), 565–572.

Willenborg, L., & De Waal, T. (2012). *Elements of statistical disclosure control* (Vol. 155). Springer Science & Business Media.

Zhang, L.-C., & Tuoto, T. (2021). Linkage-data linear regression. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *184*(2), 522–547.

Appendices

A Second variant of the regularised estimator

In this appendix, the second variant of the regularised estimator is given as an addition to Section 3.2. As mentioned there, the second variant of this estimator is not as attractive as the first variant. For this reason, it was decided to leave the second variant aside in this study. The main formulas of the second variant of the regularised estimator will be given (Pijpers, 2021).

The first variant of the proposed regularised estimator in (15) is constructed in such a way that even if $\mathbf{Q}$ were not symmetric, the matrix inverse needed for the estimator is guaranteed to exist for all values of $0 \leq \nu \leq 1$. For the $\mathbf{Q}$ that is actually considered, i.e. (4) which is symmetric and positive definite as long as $\frac{1}{n} < q < 1$, a second alternative estimator can be considered:

$$\begin{aligned}
\widehat{\mathbf{T}}^{reg,2} &= \mathbf{Y}^{\mathsf{T}}[\mu\mathbf{Q} + (1-\mu)\mathbf{I}]^{-1}\mathbf{Z}^{*} \\
&= \mathbf{Y}^{\mathsf{T}}[\mu\mathbf{Q} + (1-\mu)\mathbf{I}]^{-1}\mathbf{C}\mathbf{Z} \\
&= \mathbf{Y}^{\mathsf{T}}\left[\frac{n-1}{\mu(nq-1)+(1-\mu)(n-1)}\mathbf{I} - \frac{\mu(1-q)}{\mu(nq-1)+(1-\mu)(n-1)}\mathbf{u}\mathbf{u}^{\mathsf{T}}\right]\mathbf{C}\mathbf{Z}, \qquad \text{(A.30)}
\end{aligned}$$

where $0 \leq \mu \leq 1$. The expression for the bias of this estimator is:

$$\begin{aligned}
B(\widehat{\mathbf{T}}^{reg,2}) &= \mathbf{Y}^{\mathsf{T}}\left[\frac{n-1}{\mu n(q-1)+(n-1)}\mathbf{Q} - \mathbf{I} - \frac{\mu(1-q)}{\mu n(q-1)+(n-1)}\mathbf{u}\mathbf{u}^{\mathsf{T}}\right]\mathbf{Z} \\
&= \mathbf{Y}^{\mathsf{T}}\left[\frac{n(1-\mu)(q-1)}{\mu n(q-1)+(n-1)}\mathbf{I} + \frac{(1-\mu)(1-q)}{\mu n(q-1)+(n-1)}\mathbf{u}\mathbf{u}^{\mathsf{T}}\right]\mathbf{Z} \\
&= \frac{(1-\mu)(1-q)}{\mu n(q-1)+(n-1)}\mathbf{Y}^{\mathsf{T}}\left[\mathbf{u}\mathbf{u}^{\mathsf{T}} - n\mathbf{I}\right]\mathbf{Z} \\
&= \frac{(1-\mu)(n-1)^2}{(nq+n-2)(\mu n(q-1)+(n-1))}B(\widehat{\mathbf{T}}_Q). \qquad \text{(A.31)}
\end{aligned}$$

and the expression for the variance is:

$$\begin{aligned}
\mathrm{Var}(\widehat{t}_{jk}^{reg,2}) &= \left(\frac{n-1}{\mu n(q-1)+(n-1)}\right)^2 \mathrm{Var}(\mathbf{y}_j^{\mathsf{T}}\mathbf{C}\mathbf{z}_k) \\
&= \frac{(n-1)^4}{(nq-1)^2[\mu n(q-1)+(n-1)]^2}\mathrm{Var}(\widehat{t}_{jk}^{Q}). \qquad \text{(A.32)}
\end{aligned}$$

Even in the limit $q \downarrow \frac{1}{n}$, the variance remains finite. In general, smaller values of μ decrease the bias but increase the variance. As done for the ν with the first variant of the regularised

estimator, an optimal value for μ is determined:

$$\mu_{opt} = 1 - \frac{(n-1)^2}{(nq-1)(1-q)^2} \frac{\text{Var}(\mathbf{y}_j^\intercal \mathbf{C} \mathbf{z}_k)}{\left(\mathbf{y}_j^\intercal [\mathbf{u}\mathbf{u}^\intercal - n\mathbf{I}] \mathbf{z}_k\right)^2}. \tag{A.33}$$

As for ν_{opt}, the same two problems occur for μ_{opt}. For $q \uparrow 1$ or if $(\mathbf{y}_j^\intercal [\mathbf{u}\mathbf{u}^\intercal - n\mathbf{I}] \mathbf{z}_k)^2 << \text{Var}(\mathbf{y}_j^\intercal \mathbf{C} \mathbf{z}_k)$ the value of μ can go to -∞, which would violate $0 \leq \mu \leq 1$, and the second fraction of the expression depends on $\mathbf{T}$, which needs to be estimated first.

A pragmatic choice is to set μ to:

$$\mu_{prag} = \frac{nq-1}{nq+n-2}. \tag{A.34}$$

Using the pragmatic choice for μ results in a lower bias for $\widehat{\mathbf{T}}^{reg,2}$ compared to the biases of $\widehat{\mathbf{T}}^{*}$ and $\widehat{\mathbf{T}}^{Q}$. However, this is not the case when comparing the bias of $\widehat{\mathbf{T}}^{reg,2}$ with μ_{prag} with the bias of $\widehat{\mathbf{T}}^{reg,1}$ with ν_{prag}. The variance of $\widehat{\mathbf{T}}^{reg,2}$ is slightly higher than for $\widehat{\mathbf{T}}^{*}$ which is considerably higher than for $\widehat{\mathbf{T}}^{reg,1}$. Hence, the second variant is not as attractive as the first variant.

B Generating permutation matrices

In this appendix, a relatively simple approach by Scholtus (2020) is proposed to generate permutation matrices $\mathbf{C}$. This approach is necessary for the computations of the variance of $\widehat{t}^*_{jk}$ in (14) and the conditional probabilities $\Pr(\widehat{t}^*|t)$ for the new methods that were constructed in Section 3.3. Lastly, the approach is used in the simulation study in Section 4. The full derivations and proof of this approach can be found in Scholtus (2020).

When two datasets that are to be linked one-to-one contain the same n entities, random linkage errors can be presented by a random permutation of order n. For each of the error matrices $\mathbf{Q}$, permutation matrices $\mathbf{C}$ have to be generated by using a stochastic procedure. Generating these random matrices $\mathbf{C}$ is non-trivial when $n > 2$ as the permutation matrices, with exactly one entry of one in each row and column and zero otherwise, have to be drawn from an appropriate distribution such that the expectation of the generated permutation matrices is equal to the error matrix $\mathbf{Q}$ under consideration (see (1)). When $n = 1$, there are no linkage errors.

In Scholtus et al. (2022), appropriate permutation matrices $\mathbf{C}$ with $\mathbf{Q}$ as expectation were generated by applying Cox's algorithm (Cox, 1987) which is presented in Willenborg and De Waal (2012). This procedure control-rounds the $\mathbf{Q}$ matrix to base 1 according to the probabilities provided in this matrix. It ensures that there is exactly one entry equal to 1 in each column and row and that the other entries are equal to zero. Therefore, the expectation of these generated permutation matrices is equal to the corresponding $\mathbf{Q}$ of the exchangeable linkage error model (see (1)). However, this method becomes relatively time-consuming when n is large.

In Scholtus (2020), a relatively simple probability distribution for $\mathbf{C}$ that satisfies (2) and (3) is proposed to efficiently generate permutation matrices. Assume that $n \geq 3$. The set of permutation matrices $\mathbf{C}$ of order n can be mapped one-to-one to the permutation group S_n. There are $n!$ permutations in S_n. To describe these permutations the cycle notation can be used. Let $T_n = \{(12), (13), \ldots, (1n), (23), \ldots, ((n-1)n)\}$ denote the set of all distinct 2-cycles of n elements. The 2-cycle (ij) represents an incorrect link between the records of unit i in file A and unit j in file B, and vice versa. The number of elements in T_n is $\binom{n}{2} = \frac{n(n-1)}{2}$.

Every possible linked dataset can be obtained by taking combinations of pairwise linkage errors of the form (ij). Consider the following two-step procedure for drawing a permutation from S_n (i.e., a permutation matrix $\mathbf{C}$):

1. Draw a random number d from a probability distribution on $\{0, 1, 2, \ldots\}$
2. If $d = 0$, take the identity permutation. Otherwise, draw d random 2-cycles from T_n and take their composition.

This procedure always yields a valid permutation from S_n and every permutation S_n can be constructed this way. Note that a pairwise linkage error can also correct for other pairwise

linkage errors, e.g. when (ij) is used immediately after (ij) itself. The second pairwise linkage error then corrects the first one. To complete the procedure, a probability distribution has to be specified in each step. The following specification is proposed, which is in line with the exchangeable linkage error model when n is not small:

- In the first step, d is drawn from a Poisson distribution with parameter λ. That is to say, $\Pr(d = t) = e^{-\lambda}\frac{\lambda^t}{t!}$ for $t \in \{0, 1, 2, \dots\}$.
- In the second step, 2-cycles are drawn from T_n with equal probability and with replacement, i.e. for each draw all $\binom{n}{2}$ 2-cycles have the same selection probability $\frac{1}{\binom{n}{2}} = \frac{2}{n(n-1)}$.

We speak of an appropriate value of λ when this generation procedure results in permutation matrices with the correct expectation, namely $\mathbf{Q}$. The following formula can be used to find an appropriate value of λ:

$$\lambda(n, q) = \frac{n-1}{2}\{\log(n-1) - \log(nq-1)\}, \tag{B.35}$$

where log denotes the natural logarithm. The proof that (B.35) indeed results in appropriate values of λ can be found in Scholtus (2020). Note that the weak assumption $\frac{1}{n} < q$ from (6) is necessary to ensure that this solution exists. Under this assumption, $\lambda(n, q)$ decreases monotonically to 0 as q increases to 1 for fixed n.

C Online repository

In addition to the simulation study in Section 4, this appendix provides a link to an online repository, where the files with the R-code can be downloaded. These files are ready to use and reproduce the study. Moreover, a short description is given for each file.

The files containing the R-code that was used for the simulation study is available at the following online repository: https://github.com/sjarai/master_thesis

The following files are in the repository:

- READ.ME: Explanatory file that gives a short description of the project and the files in the repository. It is recommended to read this file before use.
- Functions.R: R-file with all the functions that are needed in the simulation study.
- Simulation_study.R: R-file with parallelised code that performs the simulation study that was described in Section 4.1.
- Results.R: R-file with code that gives the average results and the plots from Section 4.2.2.

D Alternative approach to calculate probabilities

In this appendix, the alternative approach to compute the probabilities $\Pr(\widehat{t^*}|t)$ (Scholtus, 2023) is given as an addition to Section 3.3 and Section 4. This approach is used in the R-code to perform the simulation study, as this is a more convenient approach to program.

First, the numerical procedure of computing the probabilities $p(d, \widehat{t^*})$, covered in Section 3.3.2, is rewritten in matrix-vector notation. Let $\mathbf{A}$ be a transition matrix built up from the probabilities $\Pr(+|\widehat{t^*})$, $\Pr(-|\widehat{t^*})$, and $\Pr(-|\widehat{t^*})$ (i.e., (23), (24), (25, respectively):

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_{L'}^{\mathsf{T}} \\ \vdots \\ \mathbf{a}_{y}^{\mathsf{T}} \\ \vdots \\ \mathbf{a}_{U'}^{\mathsf{T}} \end{pmatrix}, \mathbf{a}_{y}^{\mathsf{T}} = (a_{yz}), a_{yz} = \begin{cases} \Pr(+|y-1), & \text{if } z = y-1 \text{ and } y > L' \\ \Pr(-|y+1), & \text{if } z = y+1 \text{ and } y < U' \\ \Pr(=|y), & \text{if } z = y \\ 0, & \text{otherwise.} \end{cases} \tag{D.36}$$

Note that matrix $\mathbf{A}$ is a band matrix and that it does not depend on the assumed true value t. Let $\mathbf{p}_t(d)$ be a vector containing all possible probabilities $p(d, \widehat{t^*})$, i.e. for all $L \leq L' \leq \widehat{t^*} \leq U' \leq U$, under the assumption that t is the true value. We then have:

$$\begin{aligned} \mathbf{p}_t(0) &= \mathbf{e}_t, \\ \mathbf{p}_t(d) &= \mathbf{A} \cdot \mathbf{p}_t(d-1) = \mathbf{A}^d \mathbf{e}_t, \end{aligned} \tag{D.37}$$

where $\mathbf{e}_t$ is a standard basis vector with 1 in position t and 0 otherwise.

Next, the computation of the reversed probabilities using Bayes' rule, covered in Section 3.3.3, is rewritten in matrix-vector notation. Let $\mathbf{P} = (p_{\widehat{t^*}t})$ denote the matrix of all conditional probabilities $\Pr(\widehat{t^*}|t)$, with $p_{\widehat{t^*}t} = \Pr(\widehat{t^*}|t)$ and let $\tilde{\mathbf{P}} = (p_{t\widehat{t^*}})$, with $p_{t\widehat{t^*}} = \Pr(t|\widehat{t^*})$. Let π_{0t} be the chosen prior distribution for the true cell t. We collect the prior probabilities into a vector: $\boldsymbol{\pi}_0 = (\pi_{0t})$, where t runs over all integers from L' up to and including U'. Then, applying Bayes' rule to obtain $\tilde{\mathbf{P}}$ from $\mathbf{P}$ and $\boldsymbol{\pi}_0$ can be represented in two steps:

$$\begin{aligned} \mathbf{P}_{joint} &= \mathbf{P} \cdot \mathrm{diag}(\boldsymbol{\pi}_0), \\ \tilde{\mathbf{P}} &= \mathrm{diag}(\mathbf{P}_{joint} \cdot \mathbf{u})^{-1} \cdot \mathbf{P}_{joint}, \end{aligned} \tag{D.38}$$

where $\mathbf{P}_{joint}$ is a matrix of all joint probabilities $\Pr(t, \widehat{t^*}) = \Pr(\widehat{t^*}|t)\Pr_0(t)$, $\mathbf{u}$ denotes a vector of ones, $\mathrm{diag}(\boldsymbol{\pi}_0)$ denotes a diagonal matrix with the prior probabilities on the diagonal corresponding to each value of t. Matrix $\mathbf{P}$ has an interesting property that the numerical procedure

in (22) and (D.37) can be computed in one step by:

$$\mathbf{P} = \mathbf{V} e^{\mathbf{D} - \lambda(n,q)\mathbf{I}} \mathbf{V}^{-1}, \tag{D.39}$$

where $\mathbf{B} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$ denotes the eigenvalue decomposition of matrix $\mathbf{B} = \lambda(n,q)\mathbf{A}$ and $\mathbf{D}$ a diagonal matrix of eigenvalues. The full proof of this property is given in Scholtus (2023).

E Other results simulation study

In addition to the results of the simulation study in Section 4.2.2, this appendix provides an overview of the short names used in the results and tables with the exact results corresponding to the line graphs with the average percentages of the naïve approach outperforming the alternative approach and the average total relative differences from the naïve empirical RMSE. Moreover, full plots are given for the latter, as they were shown with a truncated y-axis.

In the figures in Section 4.2.2 and in the tables and figures in the remainder of this appendix, short names are used for the different alternative approaches. For clarity, an overview is given with all the alternative approaches with the corresponding short name in Table E.10.

Table E.10
Overview of the Short Names Used in the Figures and Tables in This Thesis for Each Alternative Correction Approach

Short name	Corresponding alternative approach
$\mathbf{Q}$	The $\mathbf{Q}$ approach
$\mathbf{Q}^{-1}$	The $\mathbf{Q}^{-1}$ approach
$\mathbb{E}_1$	The expected value approach using the first prior distribution
$\mathbb{E}_2$	The expected value approach using the second prior distribution
$\mathbb{E}_3$	The expected value approach using the third prior distribution
Reg_{prag}	The regularised approach with the pragmatic value for ν
$\text{Reg}_{opt(true)}$	The regularised approach with the optimal value for ν which is computed using the true contingency table
$\text{Reg}_{opt(obs)}$	The regularised approach with the optimal value for ν which is estimated using the naïve contingency table
$\text{Reg}_{opt(\mathbb{E}_1)}$	The regularised approach with the optimal value for ν which is estimated using the first variant of the expected value
$\text{Reg}_{opt(\mathbb{E}_2)}$	The regularised approach with the optimal value for ν which is estimated using the second variant of the expected value
$\text{Reg}_{opt(\mathbb{E}_3)}$	The regularised approach with the optimal value for ν which is estimated using the third variant of the expected value
$W_{\widehat{MSE}_1}$	The weighted MSE approach using the first variant of the expected value
$W_{\widehat{MSE}_2}$	The weighted MSE approach using the second variant of the expected value
$W_{\widehat{MSE}_3}$	The weighted MSE approach using the third variant of the expected value

Note. In the legends of the figures, the text format differs as the short names are displayed without the bold and italic letters and $\mathbb{E}$ is displayed as a normal E.

In Table E.11, the average percentages of the naïve approach outperforming the alternative approach over the 10 contingency tables with dependent attributes with corresponding standard deviation are given. These results were also graphically shown in Figure 2. In Table E.12, the average total relative differences from the naïve empirical RMSE over the 10 contingency tables with dependent attributes with corresponding standard deviation are given. These results were also graphically shown in Figure 4.

Table E.11

The Average Percentages where the Naïve Approach Outperforms the Alternative Approach over the 10 Generated Contingency Tables with Dependent Attributes for Different Error Rates q

	Dependent attributes		
	$q = 0.1$	$q = 0.2$	$q = 0.3$
$\mathbf{Q}$	54.93 (2.45)	62.46 (2.96)	69.14 (3.57)
$\mathbf{Q}^{-1}$	65.39 (2.78)	56.47 (3.14)	49.84 (3.24)
$\mathbb{E}_1$	64.78 (2.41)	63.78 (2.26)	61.45 (3.21)
$\mathbb{E}_2$	54.97 (2.64)	62.55 (3.19)	69.14 (3.64)
$\mathbb{E}_3$	46.91 (2.97)	41.04 (3.49)	36.31 (4.20)
Reg_{prag}	54.93 (2.53)	62.34 (2.99)	68.99 (3.64)
$\text{Reg}_{opt(true)}$	50.00 (3.04)	44.97 (2.99)	40.14 (3.00)
$\text{Reg}_{opt(obs)}$	54.33 (2.39)	60.84 (2.72)	65.25 (3.24)
$\text{Reg}_{opt(\mathbb{E}_1)}$	55.19 (2.31)	58.39 (1.87)	58.36 (1.99)
$\text{Reg}_{opt(\mathbb{E}_2)}$	54.93 (2.45)	62.46 (2.96)	69.13 (3.57)
$\text{Reg}_{opt(\mathbb{E}_3)}$	54.16 (2.40)	60.54 (2.70)	63.08 (2.86)
$W_{\widehat{MSE}_1}$	54.79 (2.43)	61.70 (2.99)	67.79 (3.58)
$W_{\widehat{MSE}_2}$	54.93 (2.44)	62.28 (2.97)	68.77 (3.61)
$W_{\widehat{MSE}_3}$	54.82 (2.44)	61.97 (3.00)	67.86 (3.57)
	$q = 0.4$	$q = 0.5$	$q = 0.6$
$\mathbf{Q}$	74.33 (3.90)	78.39 (4.21)	81.40 (4.06)
$\mathbf{Q}^{-1}$	43.68 (3.36)	39.85 (3.30)	37.55 (4.33)
$\mathbb{E}_1$	59.15 (3.71)	55.26 (4.97)	53.38 (4.72)
$\mathbb{E}_2$	74.38 (4.03)	78.39 (4.19)	81.35 (4.02)
$\mathbb{E}_3$	32.93 (4.65)	31.03 (5.01)	30.83 (5.15)
Reg_{prag}	74.14 (3.95)	77.93 (4.02)	80.61 (3.92)
$\text{Reg}_{opt(true)}$	37.88 (3.10)	37.27 (4.13)	37.86 (4.08)
$\text{Reg}_{opt(obs)}$	66.44 (3.44)	64.87 (4.77)	62.66 (4.95)
$\text{Reg}_{opt(\mathbb{E}_1)}$	58.31 (2.21)	57.44 (2.73)	57.17 (3.91)
$\text{Reg}_{opt(\mathbb{E}_2)}$	74.14 (3.87)	76.78 (4.52)	75.96 (5.50)
$\text{Reg}_{opt(\mathbb{E}_3)}$	62.59 (3.53)	60.80 (3.39)	59.06 (4.10)
$W_{\widehat{MSE}_1}$	72.67 (3.68)	76.77 (3.82)	79.89 (3.72)
$W_{\widehat{MSE}_2}$	73.50 (3.77)	76.81 (3.83)	79.89 (3.72)
$W_{\widehat{MSE}_3}$	72.68 (3.68)	76.78 (3.82)	79.89 (3.72)
	$q = 0.7$	$q = 0.8$	$q = 0.9$
$\mathbf{Q}$	83.57 (3.76)	84.97 (2.91)	86.79 (1.90)
$\mathbf{Q}^{-1}$	37.06 (4.32)	38.48 (4.75)	45.52 (4.24)
$\mathbb{E}_1$	52.16 (4.93)	54.84 (5.59)	64.22 (3.92)
$\mathbb{E}_2$	83.53 (3.67)	84.93 (2.81)	86.77 (1.83)
$\mathbb{E}_3$	32.30 (5.07)	36.15 (4.86)	45.49 (4.18)
Reg_{prag}	82.44 (3.38)	84.23 (2.80)	86.79 (1.90)
$\text{Reg}_{opt(true)}$	41.64 (3.72)	47.99 (3.40)	61.33 (2.84)
$\text{Reg}_{opt(obs)}$	62.31 (5.26)	62.67 (4.81)	69.62 (4.18)
$\text{Reg}_{opt(\mathbb{E}_1)}$	57.90 (4.15)	60.20 (4.30)	68.64 (3.91)
$\text{Reg}_{opt(\mathbb{E}_2)}$	71.52 (6.58)	67.87 (5.51)	70.38 (4.23)
$\text{Reg}_{opt(\mathbb{E}_3)}$	58.61 (3.96)	60.64 (4.35)	68.71 (3.80)
$W_{\widehat{MSE}_1}$	82.30 (3.41)	84.23 (2.81)	86.79 (1.90)
$W_{\widehat{MSE}_2}$	82.31 (3.41)	84.23 (2.81)	86.79 (1.90)
$W_{\widehat{MSE}_3}$	82.30 (3.41)	84.23 (2.81)	86.79 (1.90)

Note. The rows contain the 14 considered alternative approaches. The columns contain the different values of q. Due to the table's width, the columns are split into three parts. The percentages where the naïve approach outperforms the alternative approach are the average over the ten contingency tables with dependent attributes. The corresponding standard deviation is given behind the average percentages in parentheses. For an overview of the short names used for the alternative approaches, see Table E.10.

Table E.12

The Average Total Relative Differences from the Naïve Empirical RMSE over the 10 Generated Contingency Tables with Dependent Attributes for Different Error Rates q with Corresponding Standard Deviation

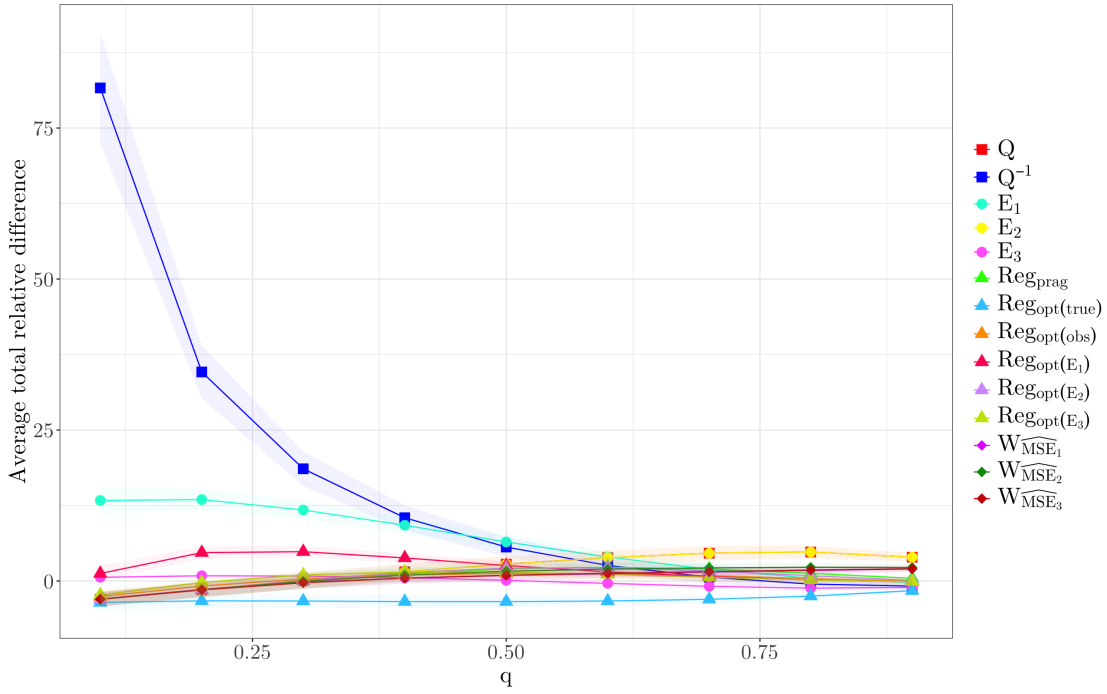
Dependent attributes			
	$q = 0.1$	$q = 0.2$	$q = 0.3$
$\mathbf{Q}$	-3.02 (1.21)	-1.43 (1.27)	0.10 (1.32)
$\mathbf{Q}^{-1}$	81.64 (9.07)	34.61 (4.36)	18.58 (2.83)
$\mathbb{E}_1$	13.34 (2.19)	13.48 (1.74)	11.76 (1.52)
$\mathbb{E}_2$	-3.02 (1.21)	-1.43 (1.27)	0.10 (1.32)
$\mathbb{E}_3$	0.62 (0.13)	0.87 (0.23)	0.82 (0.33)
Reg_{prag}	-3.02 (1.21)	-1.45 (1.26)	0.01 (1.28)
$\text{Reg}_{opt(true)}$	-3.55 (1.21)	-3.28 (1.24)	-3.31 (1.19)
$\text{Reg}_{opt(obs)}$	-2.56 (0.92)	-0.83 (0.74)	0.34 (0.67)
$\text{Reg}_{opt(\mathbb{E}_1)}$	1.25 (1.22)	4.70 (1.03)	4.86 (0.93)
$\text{Reg}_{opt(\mathbb{E}_2)}$	-3.02 (1.21)	-1.43 (1.27)	0.09 (1.32)
$\text{Reg}_{opt(\mathbb{E}_3)}$	-2.31 (0.85)	-0.23 (0.59)	1.01 (0.47)
$W_{\widehat{MSE}_1}$	-2.98 (1.17)	-1.39 (1.10)	-0.22 (1.01)
$W_{\widehat{MSE}_2}$	-3.02 (1.21)	-1.45 (1.25)	-0.06 (1.24)
$W_{\widehat{MSE}_3}$	-3.01 (1.17)	-1.47 (1.12)	-0.29 (1.03)
	$q = 0.4$	$q = 0.5$	$q = 0.6$
$\mathbf{Q}$	1.53 (1.33)	2.79 (1.31)	3.86 (1.27)
$\mathbf{Q}^{-1}$	10.48 (2.06)	5.63 (1.63)	2.58 (1.30)
$\mathbb{E}_1$	9.23 (1.38)	6.47 (1.18)	3.98 (1.04)
$\mathbb{E}_2$	1.52 (1.33)	2.79 (1.31)	3.85 (1.28)
$\mathbb{E}_3$	0.55 (0.41)	0.12 (0.49)	-0.39 (0.55)
Reg_{prag}	1.21 (1.22)	2.01 (1.09)	2.28 (0.88)
$\text{Reg}_{opt(true)}$	-3.41 (1.09)	-3.42 (0.95)	-3.31 (0.82)
$\text{Reg}_{opt(obs)}$	1.00 (0.69)	1.27 (0.76)	1.19 (0.83)
$\text{Reg}_{opt(\mathbb{E}_1)}$	3.82 (0.78)	2.56 (0.69)	1.52 (0.70)
$\text{Reg}_{opt(\mathbb{E}_2)}$	1.38 (1.33)	2.14 (1.32)	2.19 (1.25)
$\text{Reg}_{opt(\mathbb{E}_3)}$	1.49 (0.51)	1.45 (0.60)	1.10 (0.71)
$W_{\widehat{MSE}_1}$	0.49 (0.91)	0.92 (0.82)	1.22 (0.73)
$W_{\widehat{MSE}_2}$	0.97 (1.17)	1.62 (1.09)	1.98 (0.98)
$W_{\widehat{MSE}_3}$	0.47 (0.93)	0.95 (0.84)	1.28 (0.76)
	$q = 0.7$	$q = 0.8$	$q = 0.9$
$\mathbf{Q}$	4.60 (1.17)	4.80 (1.01)	3.94 (0.73)
$\mathbf{Q}^{-1}$	0.64 (1.02)	-0.51 (0.77)	-0.84 (0.47)
$\mathbb{E}_1$	1.92 (0.88)	0.47 (0.75)	-0.27 (0.48)
$\mathbb{E}_2$	4.60 (1.19)	4.80 (1.01)	3.94 (0.73)
$\mathbb{E}_3$	-0.87 (0.56)	-1.18 (0.52)	-1.05 (0.38)
Reg_{prag}	1.99 (0.60)	1.28 (0.33)	0.45 (0.10)
$\text{Reg}_{opt(true)}$	-3.03 (0.68)	-2.50 (0.53)	-1.60 (0.35)
$\text{Reg}_{opt(obs)}$	0.88 (0.84)	0.41 (0.77)	-0.01 (0.52)
$\text{Reg}_{opt(\mathbb{E}_1)}$	0.74 (0.73)	0.18 (0.70)	-0.12 (0.50)
$\text{Reg}_{opt(\mathbb{E}_2)}$	1.68 (1.12)	0.87 (0.92)	0.12 (0.57)
$\text{Reg}_{opt(\mathbb{E}_3)}$	0.63 (0.74)	0.17 (0.69)	-0.12 (0.50)
$W_{\widehat{MSE}_1}$	1.48 (0.67)	1.76 (0.62)	2.00 (0.52)
$W_{\widehat{MSE}_2}$	2.17 (0.86)	2.27 (0.74)	2.24 (0.56)
$W_{\widehat{MSE}_3}$	1.55 (0.69)	1.82 (0.63)	2.02 (0.53)

Note. The rows contain the 14 considered alternative approaches. The columns contain the different values of q. Due to the table's width, the columns are split into three parts. The total relative differences from the naïve empirical RMSE are the average over the ten contingency tables with dependent attributes. The corresponding standard deviation is given behind the average total relative differences in parentheses. For an overview of the short names used for the alternative approaches, see Table E.10.

The corresponding full plot (i.e. the plot we saw in Figure 4 with full axes) is shown in Figure E.6.

Figure E.6

Full plot of the Average Total Relative Differences from the Naïve Empirical RMSE over the 10 Generated Contingency Tables with Dependent Attributes for Different Error Rates q with Corresponding Standard Deviation



Note. The values of q are on the x-axis and the average total relative difference is on the y-axis. Each alternative approach has its own colour, which can be found in the legend. Moreover, each group of estimators has its own shape to indicate the data points. The existing estimators from Section 3.1 are indicated by a square, the expected values from Section 3.3.4 by a dot, the regularised estimators from Section 3.2 by a triangle, and the weighted MSE estimators from Section 3.3.5 by a diamond. The total relative differences are the average over the ten contingency tables with dependent attributes. The ribbon around the graph in the same colour as the line and data points corresponding to the approach indicates one standard deviation above and below the average total relative difference. For an overview of the short names used for the alternative approaches, see Table E.10.

In Table E.13, the average percentages of the naïve approach outperforming the alternative approach over the 10 contingency tables with independent attributes with corresponding standard deviation are given. These results were also graphically shown in Figure 3. In Table E.14, the average total relative differences from the naïve empirical RMSE over the 10 contingency tables with independent attributes with corresponding standard deviation are given. These results were also graphically shown in Figure 5.

Table E.13

The Average Percentages where the Naïve Approach Outperforms the Alternative Approach over the 10 Generated Contingency Tables with Independent Attributes for Different Error Rates q

Independent attributes			
	$q = 0.1$	$q = 0.2$	$q = 0.3$
$\mathbf{Q}$	41.41 (3.55)	43.44 (4.11)	45.24 (4.92)
$\mathbf{Q}^{-1}$	82.36 (4.12)	78.45 (3.80)	75.45 (4.06)
$\mathbb{E}_1$	76.18 (3.55)	76.55 (3.51)	75.77 (4.01)
$\mathbb{E}_2$	41.41 (3.55)	43.36 (4.19)	45.29 (4.86)
$\mathbb{E}_3$	71.02 (4.98)	70.10 (5.14)	69.15 (5.45)
Reg_{prag}	41.44 (3.53)	43.26 (4.26)	45.07 (4.88)
$\text{Reg}_{opt(true)}$	41.38 (3.36)	42.40 (3.33)	42.92 (3.20)
$\text{Reg}_{opt(obs)}$	40.51 (3.44)	42.87 (3.88)	46.51 (4.65)
$\text{Reg}_{opt(\mathbb{E}_1)}$	48.94 (2.31)	56.75 (2.19)	60.22 (2.29)
$\text{Reg}_{opt(\mathbb{E}_2)}$	41.41 (3.55)	43.44 (4.11)	45.24 (4.92)
$\text{Reg}_{opt(\mathbb{E}_3)}$	40.37 (3.52)	43.98 (3.93)	48.95 (4.32)
$W_{\widehat{MSE}_1}$	41.00 (3.49)	42.03 (4.06)	42.92 (4.57)
$W_{\widehat{MSE}_2}$	41.40 (3.55)	43.25 (4.17)	44.79 (4.84)
$W_{\widehat{MSE}_3}$	41.11 (3.53)	42.13 (4.26)	43.10 (4.64)
	$q = 0.4$	$q = 0.5$	$q = 0.6$
$\mathbf{Q}$	47.41 (5.15)	50.02 (5.46)	52.87 (5.44)
$\mathbf{Q}^{-1}$	72.23 (4.41)	70.52 (4.76)	68.90 (5.08)
$\mathbb{E}_1$	73.92 (5.85)	73.89 (6.38)	74.23 (6.42)
$\mathbb{E}_2$	47.43 (5.14)	50.08 (5.43)	52.95 (5.38)
$\mathbb{E}_3$	68.34 (5.53)	67.61 (5.55)	67.20 (5.46)
Reg_{prag}	47.25 (4.99)	49.35 (5.41)	52.21 (5.19)
$\text{Reg}_{opt(true)}$	44.19 (3.50)	46.24 (3.61)	47.93 (3.58)
$\text{Reg}_{opt(obs)}$	50.36 (4.86)	54.31 (4.77)	58.40 (4.85)
$\text{Reg}_{opt(\mathbb{E}_1)}$	62.87 (2.04)	64.72 (1.73)	65.48 (3.09)
$\text{Reg}_{opt(\mathbb{E}_2)}$	47.39 (5.14)	49.87 (5.26)	52.83 (5.50)
$\text{Reg}_{opt(\mathbb{E}_3)}$	53.99 (4.17)	58.15 (3.97)	61.67 (3.99)
$W_{\widehat{MSE}_1}$	45.62 (4.76)	48.64 (5.09)	52.08 (5.11)
$W_{\widehat{MSE}_2}$	46.53 (5.01)	48.63 (5.10)	52.08 (5.11)
$W_{\widehat{MSE}_3}$	45.62 (4.76)	48.64 (5.10)	52.08 (5.11)
	$q = 0.7$	$q = 0.8$	$q = 0.9$
$\mathbf{Q}$	56.78 (5.16)	62.32 (4.42)	71.80 (3.35)
$\mathbf{Q}^{-1}$	68.07 (4.78)	68.46 (4.50)	71.95 (4.02)
$\mathbb{E}_1$	75.38 (6.67)	77.12 (6.19)	81.47 (4.89)
$\mathbb{E}_2$	56.81 (5.18)	62.37 (4.40)	71.78 (3.37)
$\mathbb{E}_3$	67.40 (5.13)	68.43 (4.61)	71.78 (3.88)
Reg_{prag}	56.44 (4.95)	62.34 (4.35)	71.82 (3.34)
$\text{Reg}_{opt(true)}$	51.59 (3.63)	57.80 (3.26)	68.50 (2.57)
$\text{Reg}_{opt(obs)}$	62.54 (4.59)	67.80 (4.39)	75.59 (3.35)
$\text{Reg}_{opt(\mathbb{E}_1)}$	66.67 (3.31)	69.97 (4.53)	76.23 (3.58)
$\text{Reg}_{opt(\mathbb{E}_2)}$	57.75 (5.16)	64.74 (4.89)	74.53 (3.63)
$\text{Reg}_{opt(\mathbb{E}_3)}$	65.24 (3.82)	69.11 (4.54)	76.09 (3.63)
$W_{\widehat{MSE}_1}$	56.48 (4.86)	62.29 (4.37)	71.80 (3.35)
$W_{\widehat{MSE}_2}$	56.48 (4.86)	62.29 (4.37)	71.80 (3.35)
$W_{\widehat{MSE}_3}$	56.48 (4.86)	62.30 (4.37)	71.80 (3.35)

Note. The rows contain the 14 considered alternative approaches. The columns contain the different values of q. Due to the table's width, the columns are split into three parts. The percentages where the naïve approach outperforms the alternative approach are the average over the ten contingency tables with independent attributes. The corresponding standard deviation is given behind the average percentages in parentheses. For an overview of the short names used for the alternative approaches, see Table E.10.

Table E.14
The Average Total Relative Differences from the Naïve Empirical RMSE over the 10 Generated Contingency Tables with Independent Attributes for Different Error Rates q with Corresponding Standard Deviation

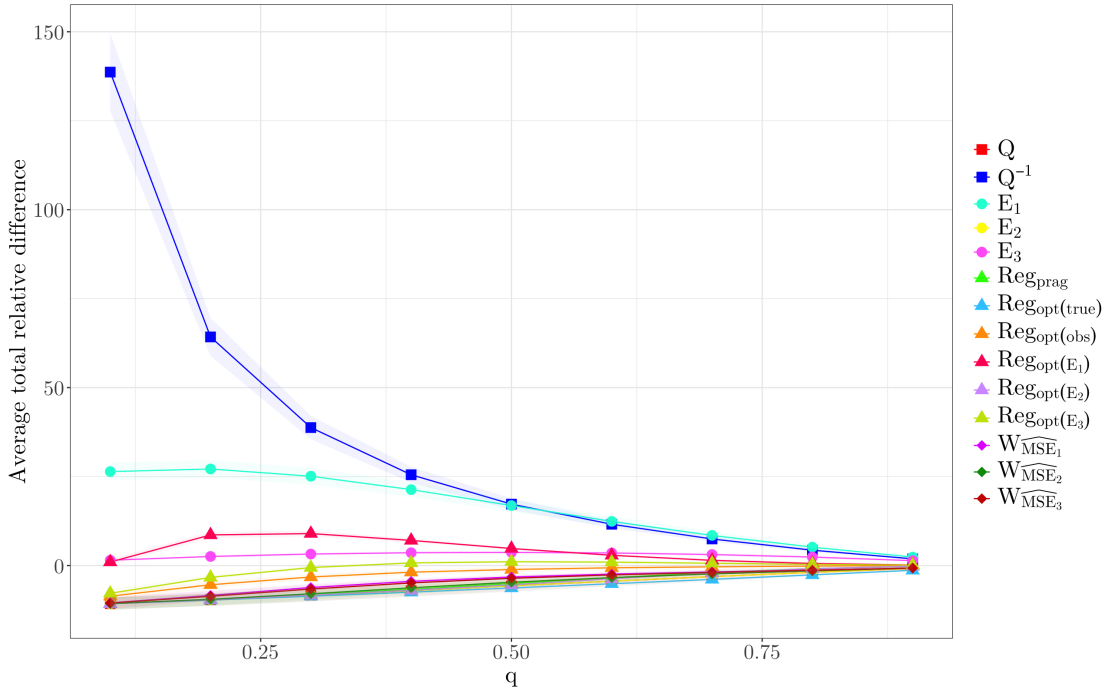
Independent attributes			
	$q = 0.1$	$q = 0.2$	$q = 0.3$
$\mathbf{Q}$	-10.69 (1.73)	-9.58 (1.74)	-8.33 (1.71)
$\mathbf{Q}^{-1}$	138.67 (10.94)	64.25 (5.32)	38.78 (3.26)
$\mathbb{E}_1$	26.41 (2.50)	27.14 (2.50)	25.10 (2.35)
$\mathbb{E}_2$	-10.69 (1.73)	-9.58 (1.74)	-8.33 (1.71)
$\mathbb{E}_3$	1.50 (0.21)	2.56 (0.34)	3.24 (0.42)
Reg_{prag}	-10.69 (1.73)	-9.54 (1.73)	-8.17 (1.66)
$\text{Reg}_{opt(true)}$	-10.73 (1.70)	-9.71 (1.64)	-8.61 (1.49)
$\text{Reg}_{opt(obs)}$	-8.61 (1.27)	-5.40 (0.92)	-3.21 (0.77)
$\text{Reg}_{opt(\mathbb{E}_1)}$	1.00 (0.82)	8.61 (0.90)	9.00 (0.77)
$\text{Reg}_{opt(\mathbb{E}_2)}$	-10.69 (1.73)	-9.58 (1.74)	-8.32 (1.71)
$\text{Reg}_{opt(\mathbb{E}_3)}$	-7.79 (1.15)	-3.32 (0.65)	-0.56 (0.44)
$W_{\widehat{MSE}_1}$	-10.47 (1.66)	-8.38 (1.39)	-6.16 (1.07)
$W_{\widehat{MSE}_2}$	-10.69 (1.73)	-9.48 (1.70)	-7.93 (1.55)
$W_{\widehat{MSE}_3}$	-10.52 (1.67)	-8.65 (1.45)	-6.59 (1.17)
	$q = 0.4$	$q = 0.5$	$q = 0.6$
$\mathbf{Q}$	-7.03 (1.63)	-5.70 (1.53)	-4.39 (1.39)
$\mathbf{Q}^{-1}$	25.52 (2.22)	17.26 (1.59)	11.61 (1.13)
$\mathbb{E}_1$	21.34 (2.06)	16.87 (1.74)	12.41 (1.41)
$\mathbb{E}_2$	-7.03 (1.63)	-5.70 (1.53)	-4.39 (1.39)
$\mathbb{E}_3$	3.61 (0.48)	3.71 (0.52)	3.54 (0.52)
Reg_{prag}	-6.65 (1.49)	-5.02 (1.25)	-3.42 (0.93)
$\text{Reg}_{opt(true)}$	-7.48 (1.29)	-6.30 (1.09)	-5.11 (0.86)
$\text{Reg}_{opt(obs)}$	-1.86 (0.72)	-1.10 (0.73)	-0.63 (0.73)
$\text{Reg}_{opt(\mathbb{E}_1)}$	7.07 (0.54)	4.80 (0.33)	2.88 (0.28)
$\text{Reg}_{opt(\mathbb{E}_2)}$	-6.95 (1.62)	-5.40 (1.47)	-3.71 (1.27)
$\text{Reg}_{opt(\mathbb{E}_3)}$	0.74 (0.40)	1.08 (0.43)	0.97 (0.49)
$W_{\widehat{MSE}_1}$	-4.43 (0.83)	-3.21 (0.69)	-2.39 (0.59)
$W_{\widehat{MSE}_2}$	-6.24 (1.33)	-4.68 (1.11)	-2.34 (0.76)
$W_{\widehat{MSE}_3}$	-4.87 (0.94)	-3.57 (0.78)	-2.63 (0.67)
	$q = 0.7$	$q = 0.8$	$q = 0.9$
$\mathbf{Q}$	-3.10 (1.20)	-1.89 (0.94)	-0.82 (0.58)
$\mathbf{Q}^{-1}$	7.48 (0.79)	4.33 (0.48)	1.89 (0.24)
$\mathbb{E}_1$	8.48 (1.09)	5.17 (0.79)	2.38 (0.44)
$\mathbb{E}_2$	-3.10 (1.20)	-1.89 (0.94)	-0.82 (0.58)
$\mathbb{E}_3$	3.10 (0.49)	2.39 (0.39)	1.38 (0.23)
Reg_{prag}	-1.97 (0.59)	-0.86 (0.28)	-0.20 (0.08)
$\text{Reg}_{opt(true)}$	-3.88 (0.63)	-2.62 (0.41)	-1.33 (0.20)
$\text{Reg}_{opt(obs)}$	-0.35 (0.71)	-0.15 (0.62)	-0.01 (0.41)
$\text{Reg}_{opt(\mathbb{E}_1)}$	1.47 (0.38)	0.59 (0.45)	0.15 (0.36)
$\text{Reg}_{opt(\mathbb{E}_2)}$	-2.12 (1.03)	-0.88 (0.75)	-0.17 (0.45)
$\text{Reg}_{opt(\mathbb{E}_3)}$	0.65 (0.53)	0.34 (0.52)	0.12 (0.38)
$W_{\widehat{MSE}_1}$	-1.79 (0.54)	-1.27 (0.50)	-0.69 (0.41)
$W_{\widehat{MSE}_2}$	-2.34 (0.76)	-1.49 (0.62)	-0.73 (0.44)
$W_{\widehat{MSE}_3}$	-1.91 (0.59)	-1.31 (0.52)	-0.70 (0.41)

Note. The rows contain the 14 considered alternative approaches. The columns contain the different values of q. Due to the table's width, the columns are split into three parts. The average total relative differences from the naïve empirical RMSE are the average over the ten contingency tables with independent attributes. The corresponding standard deviation is given behind the average total relative differences in parentheses. For an overview of the short names used for the alternative approaches, see Table E.10.

The corresponding full plot (i.e. the plot we saw in Figure 5 with full axes) is shown in Figure E.7.

Figure E.7

Full plot of the Average Total Relative Differences from the Naïve Empirical RMSE over the 10 Generated Contingency Tables with Independent Attributes for Different Error Rates q with Corresponding Standard Deviation



Note. The values of q are on the x-axis and the average total relative difference is on the y-axis. Each alternative approach has its own colour, which can be found in the legend. Moreover, each group of estimators has its own shape to indicate the data points. The existing estimators from Section 3.1 are indicated by a square, the expected values from Section 3.3.4 by a dot, the regularised estimators from Section 3.2 by a triangle, and the weighted MSE estimators from Section 3.3.5 by a diamond. The total relative differences are the average over the ten contingency tables with independent attributes. The ribbon around the graph in the same colour as the line and data points corresponding to the approach indicates one standard deviation above and below the average total relative difference. For an overview of the short names used for the alternative approaches, see Table E.10.

F Additional results revisiting example tables

In Section 4.2.3, we revisited the contingency tables with dependent attributes and $q = 0.8$ with two correction approaches applied, one that performs well and one that performs worse (see Table 9). In this appendix, the application of one correction approach that performs well and one correction approach that performs worse for the table with dependent attributes $\mathbf{Y}_g^\intercal \mathbf{Z}_g$ with $q = 0.2$ and for the tables with independent attributes $\mathbf{Y}_g^\intercal \mathbf{W}_g$ with $q = 0.8$ and $q = 0.2$ are given.

For tables with dependent attributes and smaller values of q, we saw that the regularised estimator with ν_{prag} performs well and the expected value approach with the first prior performs worse. The naïve contingency table $\mathbf{Y}_g^\intercal \mathbf{Z}_g^*$ with $q = 0.2$ that we saw in Table 6 after correction with these two correction methods are shown in Table F.15.

Table F.15

Two Examples of the Corrected Contingency Tables with Dependent Attributes where the Probability of a Correct Link (q) Was 0.2 by Using the Regularised Approach with ν_{prag} and the Expected Value Approach with the First Prior Distribution

	C_1	C_2	C_3	C_4	C_5
R_1	0.40	1.00	3.53	2.48	0.59
R_2	1.53	7.49	19.84	13.22	4.92
R_3	3.21	17.06	46.40	31.77	10.55
R_4	2.82	15.86	42.87	30.73	8.73
R_5	1.04	5.59	15.36	9.80	3.21

	C_1	C_2	C_3	C_4	C_5
R_1	4.21	2.41	4.31	4.25	2.49
R_2	4.53	12.59	17.28	10.27	16.09
R_3	4.13	17.58	45.47	29.84	16.10
R_4	3.52	17.31	40.81	41.37	6.73
R_5	3.73	12.53	20.97	9.51	8.13

Note. The corrected contingency table on the left is corrected using the regularised approach with the pragmatic choice for ν. The corrected contingency table on the right is corrected using the expected value approach with the first prior distribution.

The values of the corrected table by using the regularised approach with ν_{prag} in Table F.15 seem to differ quite a lot from the true contingency table in Table 3. The values of the corrected contingency table with the worst correction method, the expected value approach with the first prior, also differ a lot from the true values. If the values are all rounded to whole numbers, the corrected contingency table using the regularised approach with ν_{prag} results in 0 correct values compared to the true contingency table in Table 3. It holds for 12 out of 25 cells (i.e. 48%) that $|e_{jk}^{reg,prag}| > |e_{jk}^*|$. The total relative difference from the naïve empirical RMSE is approximately -0.006. For the corrected contingency table using the expected value approach with the first prior distribution, we find 3 correct values compared to the true contingency table in Table 3. For this correction method, we find that it holds for 14 out of 25 cells (i.e. 56%) that $|e_{jk}^{(\mathbb{E}_1)}| > |e_{jk}^*|$. The total relative difference from the naïve empirical RMSE is approximately 0.161. Overall, it is clearly visible that the values of the corrected contingency tables under the $q = 0.8$ error matrices are closer to the true values than the values of the corrected contingency tables under

the $q = 0.2$ error matrices.

For tables with independent attributes and larger values of q we found that the weighted MSE approach that uses the second variant of the expected value performed well and the expected value approach using the third prior distribution performed worse. The naïve contingency table $\mathbf{Y}_g^\intercal \mathbf{W}_g^*$ with q = 0.8 that we saw in Table 7 after correction with these two correction methods are shown in Table F.16. These corrected contingency tables can then be compared to the true contingency table $\mathbf{Y}_g^\intercal \mathbf{W}_g$ in Table 4.

Table F.16

Two Examples of the Corrected Contingency Tables with Independent Attributes where the Probability of a Correct Link (q) Was 0.8 by Using the Weighted MSE Approach Using the Second Variant of the Expected Value and the Expected Value Approach Using the Third Prior Distribution

	C_1	C_2	C_3	C_4	C_5
R_1	0.02	1.97	3.15	2.76	0.11
R_2	2.61	9.33	22.85	10.79	1.37
R_3	0.27	27.62	56.55	20.41	4.46
R_4	1.07	24.74	40.43	21.85	12.48
R_5	0.09	6.35	18.62	7.07	2.89

	C_1	C_2	C_3	C_4	C_5
R_1	0.00	2.02	2.87	3.22	0.00
R_2	3.31	8.71	23.13	11.19	0.73
R_3	0.00	28.42	57.85	19.59	3.47
R_4	0.95	25.25	38.92	22.15	13.93
R_5	0.00	5.65	19.41	6.93	3.09

Note. The corrected contingency table on the left is corrected using the weighted MSE approach that uses the second variant of the expected value. The corrected contingency table on the right is corrected using the expected value approach with the third prior distribution.

The values of the corrected contingency table with the better correction method, the weighted MSE approach with the second variant of the expected value, are close to the true values in Table 4. The values of the corrected contingency table using the expected value approach with the third prior distribution differ more from the true values. Rounding the corrected values to whole numbers, the corrected contingency table using the weighted MSE approach results in 11 correct values compared to the true contingency table in Table 4. For 13 out of 25 cells (i.e. 52%) it holds that $|e_{jk}^{W_{\widehat{MSE}_2}}| > |e_{jk}^*|$. The total relative difference from the naïve empirical RMSE is approximately -0.047. If we do the same for the corrected contingency table using the expected value approach with the third prior distribution, we find 9 correct values compared to the true contingency table in Table 4. It holds for 18 of the 25 cells (i.e. 72%) that $|e_{jk}^{(\mathbb{E}_3)}| > |e_{jk}^*|$. The total relative difference from the naïve empirical RMSE is approximately 0.185.

For tables with independent attributes and smaller values of q, the regularised estimator where ν_{opt} is estimated with the first variant of the expected value performs well and the $\mathbf{Q}^{-1}$ approach performs worse. The naïve contingency table $\mathbf{Y}_g^\intercal \mathbf{W}_g^*$ with $q = 0.2$ that we saw in Table 8 after correction with these two correction methods are shown in Table F.17. These corrected contingency tables can be compared to the true contingency table $\mathbf{Y}_g^\intercal \mathbf{W}_g$ in Table 4.

Table F.17

Two Examples of the Corrected Contingency Tables with Independent Attributes where the Probability of a Correct Link (q) Was 0.2 by Using the Regularised Approach where ν_{opt} Is Computed with the First Variant of the Expected Value and $\mathbf{Q}^{-1}$ Approach

	C_1	C_2	C_3	C_4	C_5
R_1	0.09	1.70	3.83	1.94	0.45
R_2	0.90	10.78	20.62	11.08	3.63
R_3	1.36	26.53	52.46	21.73	6.91
R_4	1.28	22.86	47.65	21.37	7.85
R_5	0.37	8.13	17.44	6.89	2.16

	C_1	C_2	C_3	C_4	C_5
R_1	0.00	0.00	4.87	8.37	0.00
R_2	7.59	6.07	0.00	40.94	11.96
R_3	0.00	53.64	73.93	0.00	0.00
R_4	0.00	5.49	43.72	25.21	26.99
R_5	0.00	7.32	39.03	0.00	0.00

Note. The corrected contingency table on the left is corrected using the regularised approach using ν_{opt} that is estimated with the first variant of the expected value. The corrected contingency table on the right is corrected using the $\mathbf{Q}^{-1}$ approach.

Some of the values of the corrected contingency table where the regularised approach is used with ν_{opt} estimated by using the first variant of the expected value, are close to the true values in Table 4, while others differ more. The values of the corrected contingency table using the $\mathbf{Q}^{-1}$ approach clearly differ even more from the true values. Rounding the corrected values to whole numbers, the corrected contingency table using the regularised value approach with ν_{opt} results in 5 correct values compared to the true contingency table in Table 4. For 16 out of 25 cells (i.e. 64%) it holds that $|e_{jk}^{reg_{opt(\mathbb{E}_1)}}| > |e_{jk}^{*}|$. The total relative difference from the naïve empirical RMSE is approximately 0.011. For the corrected contingency table using the $\mathbf{Q}^{-1}$ approach, we find 3 correct values compared to the true contingency table in Table 4. For this correction method, we find that it holds for 20 out of 25 cells (i.e. 80%) that $|e_{jk}^{\mathbf{Q}^{-1}}| > |e_{jk}^{*}|$. The total relative difference from the naïve empirical RMSE is approximately 4.448. For the independent tables, we also find that the values of the corrected contingency tables under the $q = 0.8$ error matrices are closer to the true values than the values of the corrected contingency tables under the $q = 0.2$ error matrices.