Statistical Journal of the IAOS 39 (2023) 137–154
DOI 10.3233/SJI-210886
IOS Press

Estimating a time series of temporary employment using a combination of survey and register data

Nino Mushkudiani* and Jeroen Pannekoek
Department of Methodology, Statistics Netherlands, The Hague, The Netherlands

Abstract. In this paper we investigate the application of macro-integration methods to combine two sources of labor force statistics: a survey and an administrative source. In particular, we aim to arrive at a single estimate of the time series of temporary employment that efficiently combines the information from both sources. By varying the specifications of the objective function and constraints, four different macro-integration models were defined. The most plausible results were of a model that treats neither of the sources as fixed and uses multiplicative adjustments. The results were compared with the previous research where a latent Markov model was used to estimate the same time series. This Markov model approach does not lead to very different estimates of the time-series of temporary (or permanent) employment contracts but results in smaller estimates of the proportion of "movers", persons that change contract status from temporary to permanent or the other way around. The model-based approach also provides estimates of the measurement errors in each of the sources. On the other hand, the macro-integration approach is less restrictive in the sense that it does not impose a Markov property of the integrated times series of proportions and it is more easy to implement.

Keywords: Macro-integration, labor force statistics, time series, multi-source statistics

1. Introduction

As the information era is booming, statistical agencies have to deal with an increasing number of all kinds of data sources, including administrative registers, censuses, big data and traditional sample surveys. Combining different sources to produce official statistics is essential in view of the necessity to meet the increasing demand for detailed information while facing decreasing response rates and a demand for reducing both response burden and survey costs [1–4]. "The 20th century witnessed the birth and maturing of sample surveys; the 21st century will be the age of data integration" [2].

*Corresponding author: Nino Mushkudiani, Department of Methodology, Statistics Netherlands, Henri Faasdreef 312, Postbus 24500, 2490 HA, The Hague, The Netherlands. E-mail: n.mushkudiani@cbs.nl.

Both survey non-response and the effort to reduce response burden lead to smaller sample sizes and, as a result, larger sampling errors. When population means or totals of some survey variables are available from another, reliable, source this information can be used to reduce sampling errors and non-response bias by using classical weighting and calibration methods. Other uses of multiple sources involve combining data sets that can (partially) be linked at unit-level. For instance, when an already existing administrative source can be linked to a survey, the linked data set becomes enriched with additional variables without increasing the response burden and with little additional costs.

Producing statistics from multiple sources has its own specific difficulties. [5] give an overview of the main problems that can occur when working with multi-source statistics and describe some methods that can be used to overcome these problems. A central issue is that the different sources often overlap in the sense that they have some variables in common, but the values of

these common variables differ. This can be because of (slight) differences in definition, in reference period or question wording, or due to measurement errors. It is one of the purposes of data integration to resolve these apparent inconsistencies, thereby enhancing the data quality. In practice this process of data integration can be divided into two phases: phase I deals with micro editing, in which some inconsistencies at micro-level can be resolved by, for instance, the harmonization of definitions and the identification of erroneous values, often by using simple knowledge rules and using information from combinations of variables. After these micro-level corrections, the attention shifts to estimates of the required output, based on all sources. In this second phase, modelling techniques can be used to combine the information from the different sources into single estimates of the output parameters of interest.

In this paper we want to combine the Labour Force Survey (LFS) data with the Employment Register (ER) data. Both data sources contain information on labour market variables and there are some inconsistencies in these variables between sources. Some studies have been conducted in the past that investigate possible reasons of the inconsistencies, see e.g. [6]. From these studies it was found that the differences in the figures are due to: slightly different definitions, measurement errors in certain groups due to e.g. misunderstanding of the questions in LFS or due to administrative delays in ER. But after correcting for these issues as much as possible, there still remain inconsistencies in the key figures from these data sources.

Other National Statistics Institutes (NSIs) also face similar problems and investigate integration methods for labour force statistics. [7] describe how the Basque Statistics Office integrated administrative information in the Labour Force Survey using a micro calibration method. This was done in order to improve the quality of survey estimations and gain coherency between internal and external results (registered employed and unemployed populations). One of the many methodological challenges is that survey respondents are confronted with the administrative information that is linked beforehand. The administrative data is used as auxiliary information also during the weighting and calibration procedure. Here it is assumed that errors in administrative data are small and can be neglected. Statistics Norway also investigated micro-integration techniques for combining employment data, see [3,8]. When differences are observed in the variable of interest between two data sources, additional information, for example on wages, is used to define the most probable value for the variable for each unit. These are simple knowledge based correction rules. These corrections are phase I in data integration process according to our definition above.

The goal of this paper is to investigate the possibility of obtaining reconciled figures of employment statistics using macro-integration methods when data editing of phase I was already carried out. Macro-integration techniques are commonly used in the compilation of the national accounts of a country, which involves combining a large number of estimates in a coherent system satisfying accounting rules, see e.g. [9]. In general macro-integration methods have as input a set of initial estimates that are inconsistent in the sense that they do not verify a set of known constraints (i.e. equality of estimates between sources) and have as output a new set of combined estimates that satisfy the constraints and is as close as possible, in the sense of a chosen distance function, to the initial estimates (see [10]). This generic macro-integration approach leads to different specific models for different choices of the constraints and of the distance function. In this paper we define four different models for our application, by varying the constraints and the distance function, and investigate the effects on the integrated output estimates.

For the same linked data of LFS and ER, we used in this paper, [11] proposed a different approach based on a latent Markov model for obtaining integrated results. The latent class models view the values obtained for the same variables in the different sources as fallible measurements of the same underlying true (or latent) variables that differ from the true values (and each other) due to (differences in) measurement errors. Estimates of proportions based on the estimated true values provide single estimates of the corresponding population values that are corrected for measurement errors. We compare the values of our estimates with those obtained with the latent Markov model.

This paper is organized as follows: In Section 2 we describe our data set. We also derive tables of proportions and transition proportions for the variable 'Contract type'; In Section 3 we define objective functions for macro-integration models: we introduce four different models, starting with the simplest additive model for fixed survey figures adjusting only register variables and making it more complex using multiplicative adjustments and adjusting both survey and register figures; In Section 4 we compare our results with the latent class model approach given in [11]; Finally in Section 5 we share our thoughts on the way the proposed macro-integration method could be imple-

mented in the production process. More efforts should be made before implementing the method propose here. We also discuss advantages and disadvantages of the macro-integration methods versus the latent class micro-integration method.

2. Data sources and targets of inference

For this research a tailored data set with linked individuals for several months was made available to us. Information on individuals that were in both the ER register and the LFS survey is linked.

We should mention that in practice it will not be easy to link these data within a reasonable time frame, in order to obtain up to date results. The ER register data has timeliness issues, it is a complex register including information from different institutions; not all information is available each month. On the other hand, the LFS is a quarterly rotating panel survey, with figures available for each month.

Both sources contain information on the employment status of each person, which is measured by the variable 'Contract type'. The targets of inference are the population proportions in each of the categories of 'Contract type' and, in particular, the development in time of these proportions.

2.1. The used data sets

The Labor Force Survey (LFS)[1] is collecting information about labour of households and individuals. For our study we consider only individuals and not households. The LFS is a rotating panel survey consisting of five waves. The rotating panel design entails that after a respondent joins the survey, he/she will receive follow-up questionnaires for four more times with three months intervals. Each month the sample is supplemented with new respondents to compensate for those that have finished five waves. As a result of this setup there are five waves in each month.

The ER is an administrative data set that combines information from different administrative sources, mainly from Tax authorities but also from the Centre for Work and Income (CWI) and the institute for employees insurances (in Dutch Uitkeringsinstituut Werknemers Verzekeringen (UWV)). The ER consists of administrative information on persons, households, jobs, benefits and pensions. It covers the entire Dutch population, including persons living abroad but working in the Netherlands or receiving a benefit or pension from a Dutch institution, see [12].

The LFS and the ER do not measure exactly the same and due to the different operational definitions some differences will occur when estimating population totals or proportions. Other than differences caused by definitions we will have differences in population coverage, measurement errors and maybe other large discrepancies. The population coverage is an important difference between these sources. For example, the LFS includes self-employed persons and excludes institutional residents and the ER includes all persons living abroad but working in the Netherlands and excludes self-employed persons. For the linked data coverage problems are completely avoided since it is only using the same set of persons from both sources. However, the discrepancies that we observe in the linked data may still be caused by different definitions and measurement errors. For example, if a person has 'a temporary contract' and there is an informal agreement that a contract will become permanent, then 'a permanent contract' will be filled in the LFS questionnaire. On the other hand, the ER variable 'contract type' could have errors due to delayed updates, if for e.g. a person had a temporary contract at time t and got a permanent contract at time $t+1$. It could take some time to change his/her status in the register data. This will certainly lead to some bias. Another difference in these sources is a definition of an unemployed person. In the LFS an unemployed person is a person that works less than 4 hours a week and is actively looking for a job. This could be different from an ER unemployed person. In the ER unemployed persons are persons that receive unemployment benefits.

The micro-data from the two sources are linked based on the social security number, birth date, gender, postal code and house number. In order to minimize differences due to definitions before linking data, the populations and the definition of permanent and temporary employment contracts were harmonized as much as possible in both sources. Other adjustments were also carried out, e.g. if an employee has more than one job in the ER register, only the one with the highest income was selected. The employees that are not Dutch residents were removed, these were mostly seasonal workers. The population of employees were limited to the employees with a job of more than 12 hours. After selection of the individuals aged 25–55, the linkage effectiveness of the combined sources was approximately 97%, see also [11].

[1] https://www.cbs.nl/en-gb/our-services/methods/surveys/korte-onderzoeksbeschrijvingen/dutch-labour-force-survey--lfs--.

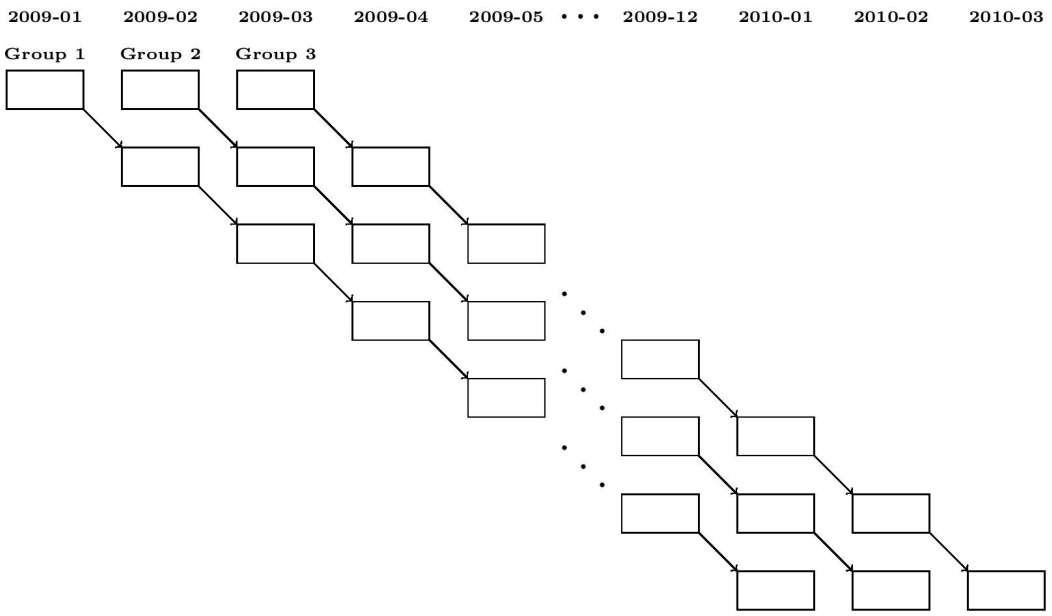

Fig. 1. The ER monthly data for three groups.

After linking the data on micro level there were still differences, that are due to differences in definitions, differences in time of measurement and measurement errors.

From the linked data set, we used the data of 8886 individuals of age 25–55, observed during 15 months, from January 2009 until March 2010. Our data include three cohorts or groups of the LFS. The first group starts in January 2009, the second in February 2009 and the third in March 2009. After the first measurement respondents are followed-up four times with three-month intervals. Assuming there is no attrition, we will have for each individual 5 quarterly observations of the employment status from the LFS. For the first group we have employment status for January 2009, April 2009, July 2009, October 2009 and January 2010. For the second group we have employment status for February 2009, May 2009, August 2009, November 2009 and February 2010 and for the third group for March 2009, June 2009, September 2009, December 2009 and March 2010. During these periods for each individual we have 13 monthly measurements of employment status from the ER. To each LFS cohort the monthly ER data are linked, see Fig. 1. This way we have observations for all 15 months.

The variable of interest: 'Contract type' is redefined having three categories: Permanent (1) including only the contracts for an unlimited duration of time; Temporary (2) including only the contracts for a limited duration of time; and Other (3). Here the category "Other" includes unemployment, self-employment and education, see also [11]. After linkage we have two variables 'Contract Type LFS' and 'Contract Type ER'. We have obtained the proportions of respondents in the three categories of both these variables. These proportions are presented in Tables 1 and 2. For the 'Contract Type ER' table we have monthly figures for each group. There are noticeable but small differences between groups. The proportions of temporary contracts are larger for group 2 than for the other two groups and group 3 has higher proportions of permanent contracts than group 1 and 2.

In Table 2 we have the proportions of the 5 quarterly measurements of 'Contract Type LFS' for each group. We see relatively big differences between the ER and LFS. In particular, the proportions of permanent contracts are much higher in the LFS than in the ER while the number of temporary contracts is much lower.

The Tables 1 and 2 show that the different sources lead to different proportions for the three contract types for time points where both sources apply.

2.2. Parameters of interest: distributions per time point and transitions

To describe the observed proportions of 'Contract type' for both sources and the integrated estimates of the corresponding population proportions, we introduce the following notation.

The measurements for a respondent i from group 1 at time point t will be denoted as follows:

Table 1
Proportions of contract type in the ER for the three groups

		Group 1			Group 2			Group 3		
N	Date	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.
1	2009–01	0.5858	0.1537	0.2605	–	–	–	–	–	–
2	2009–02	0.5901	0.1552	0.2547	0.5756	0.1629	0.2615	–	–	–
3	2009–03	0.5878	0.1562	0.2560	0.5752	0.1648	0.2600	0.5962	0.1545	0.2493
4	2009–04	0.5916	0.1508	0.2576	0.5741	0.1637	0.2621	0.6025	0.1512	0.2463
5	2009–05	0.5888	0.1549	0.2563	0.5727	0.1644	0.2628	0.6003	0.1501	0.2496
6	2009–06	0.5895	0.1517	0.2587	0.5712	0.1659	0.2629	0.5993	0.1498	0.2509
7	2009–07	0.5864	0.1473	0.2663	0.5683	0.1636	0.2682	0.5968	0.1468	0.2564
8	2009–08	0.5836	0.1489	0.2675	0.5697	0.1605	0.2698	0.5961	0.1472	0.2568
9	2009–09	0.5844	0.1474	0.2683	0.5673	0.1633	0.2694	0.5955	0.1472	0.2572
10	2009–10	0.5847	0.1461	0.2692	0.5703	0.1606	0.2691	0.5946	0.1457	0.2598
11	2009–11	0.5844	0.1492	0.2664	0.5676	0.1654	0.2669	0.5933	0.1439	0.2628
12	2009–12	0.5814	0.1479	0.2707	0.5693	0.1607	0.2700	0.5965	0.1402	0.2633
13	2010–01	0.5876	0.1415	0.2710	0.5763	0.1533	0.2704	0.5956	0.1414	0.2630
14	2010–02	–	–	–	0.5754	0.1483	0.2763	0.5967	0.1358	0.2675
15	2010–03	–	–	–	–	–	–	0.5981	0.1396	0.2624

Table 2
Proportions of contract type in the LFS for the three groups

		Group 1			Group 2			Group 3		
N	Date	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.
1	2009–01	0.6568	0.1117	0.2314	–	–	–	–	–	–
2	2009–02	–	–	–	0.6541	0.1159	0.2300	–	–	–
3	2009–03	–	–	–	–	–	–	0.6587	0.1121	0.2291
4	2009–04	0.6643	0.1050	0.2307	–	–	–	–	–	–
5	2009–05	–	–	–	0.6703	0.1059	0.2238	–	–	–
6	2009–06	–	–	–	–	–	–	0.6736	0.1034	0.2230
7	2009–07	0.6597	0.1070	0.2333	–	–	–	–	–	–
8	2009–08	–	–	–	0.6639	0.1066	0.2296	–	–	–
9	2009–09	–	–	–	–	–	–	0.6735	0.1030	0.2235
10	2009–10	0.6624	0.1039	0.2336	–	–	–	–	–	–
11	2009–11	–	–	–	0.6613	0.1168	0.2218	–	–	–
12	2009–12	–	–	–	–	–	–	0.6734	0.1044	0.2223
13	2010–01	0.6611	0.1008	0.2381	–	–	–	–	–	–
14	2010–02	–	–	–	0.6562	0.1148	0.2290	–	–	–
15	2010–03	–	–	–	–	–	–	0.6724	0.1009	0.2268

R_i^t : register value of 'Contract type' for individual i at time point t, for $t = 1, \ldots, 13$,

S_i^t : survey value of 'Contract type' for individual i at time point t, for $t \in \{1, 4, 7, 10, 13\}$,

with R_i^t and $S_i^t \in \{\text{Permanent}, \text{Temporary}, \text{Other}\}$

For group 2 we have register data R^t at the time points $t = 2, \ldots, 14$ and the survey data at the time points $t \in \{2, 5, 8, 11, 14\}$. Similarly for group 3 we have the register data R^t for $t = 3, \ldots, 15$ and the survey data S^t for $t \in \{3, 6, 9, 12, 15\}$. To keep it simple, from here on we will only deal with group 1. The results below hold similarly for groups 2 and 3.

For group 1 we denote the proportions for the 'Contract type' by:

$\mathbf{p}^{R.t} = (p_1^{R.t}, p_2^{R.t}, p_3^{R.t})$ for $t = 1, \ldots, 13$:

vectors with proportions from register data,

$\mathbf{p}^{S.t} = (p_1^{S.t}, p_2^{S.t}, p_3^{S.t})$ for $t \in \{1, 4, 7, 10, 13\}$:

vectors with proportions from survey data.

Here subscripts define the contract types: $1 =$ Permanent, $2 =$ Temporary and $3 =$ Other. The primary parameters of interest are the population distributions over the three contract types for each month. Let us denote these unknown population proportion by:

$$\boldsymbol{\pi}^t = (\pi_1^t, \pi_2^t, \pi_3^t) \text{ for } t = 1, \ldots, 13.$$

The observed information from one group to estimate these population proportions is depicted in Table 3. Here we see that for time points t and $t+3$ both sources provide information to estimate the vector with popula-

Table 3
Observed proportions from one group and population proportions for four consecutive time points

...	$\mathbf{p}^{S.t}$			$\rightarrow$			$\mathbf{p}^{S.t+3}$	...
	$\downarrow$						$\downarrow$	
...	$\boldsymbol{\pi}^{t}$	$\rightarrow$	$\boldsymbol{\pi}^{t+1}$	$\rightarrow$	$\boldsymbol{\pi}^{t+2}$	$\rightarrow$	$\boldsymbol{\pi}^{t+3}$	...
	$\uparrow$		$\uparrow$		$\uparrow$		$\uparrow$	
...	$\mathbf{p}^{R.t}$	$\rightarrow$	$\mathbf{p}^{R.t+1}$	$\rightarrow$	$\mathbf{p}^{R.t+2}$	$\rightarrow$	$\mathbf{p}^{R.t+3}$	...

tion proportions, whereas for the times in between, no direct information from the survey is available.

Besides the marginal distribution $\boldsymbol{\pi}^t$ over the three contract types for each month, the changes between these categories over time are also of interest. These can be described by the transition proportions between adjacent time points. That is, the proportions of individuals moving from one category at time point t to another category at time point $t+1$, for each of the 3×3 category combinations. To define these transition proportions, we first consider the matrix with bivariate proportions corresponding to two variables, 'Contract type' at time point t and 'Contract type' at time point $t+1$. The matrix with population proportions for the cross-classification of these two variables will be denoted by $\boldsymbol{\pi}^{t,t+1}$ with cell-proportions $\pi^{t,t+1}_{c_t,c_{t+1}}$ for $c_t, c_{t+1} \in \{\text{Permanent, Temporary, Other}\}$.

In general, we will denote the proportions of units for all combinations of any number of variables by repeated use of superscripts and subscripts, i.e. $\pi^{t,t+1,t+2,\ldots}_{c_t,c_{t+1},c_{t+2},\ldots}$, and the corresponding matrices will be denoted by $\boldsymbol{\pi}^{t,t+1,t+2,\ldots}$.

From each data source we can calculate the proportions of respondents for all possible combinations of 'Contract type' from month t till $t+3$ ($t \in \{1,4,7,10,13\}$). From the register data we observe also the intermediate months. These could be, for example, the proportion of persons that has 'contract types' "Permanent" $\rightarrow$ "Permanent" $\rightarrow$ "Permanent" $\rightarrow$ "Permanent" during the months 1 till 4, or any other sequence of categories. We define these proportions by $p^{R.t,t+1,t+2,t+3}_{c_1c_2c_3c_4}$ and the corresponding matrices by $\boldsymbol{p}^{R.t,t+1,t+2,t+3}$. Observe that:

$$p^{R.t,t+3}_{c_1c_4} = \sum_{c_2,c_3} p^{R.t,t+1,t+2,t+3}_{c_1c_2c_3c_4},$$

where $p^{R.t,t+3}_{c_1c_4}$ stands for the proportions of register units that have respectively 'Contract type' c_1 and c_4 in month t and $t+3$. For the survey data we observe $p^{S.t,t+3}_{c_1c_4}$ ($t \in \{1,4,7,10,13\}$) but do not have proportions for the intermediate months.

The observed values $p^{R.t,t+3}_{c_1c_4}$ and $p^{S.t,t+3}_{c_1c_4}$ are not consistent with each other. In the next section we define several reconciliation models that derive a single estimate $\pi^{t,t+3}_{c_tc_{t+3}}$ for these observed proportions.

3. Reconciliation methods

In this section several reconciliation methods will be presented that can be used to obtain univocal estimates of both marginal proportions and transitions. In general, reconciliation is applied when there is a need for finding single estimates of target parameters by combining separate estimates from different data sets. More generally, reconciliation (or macro-integration) is considered to be a technique for achieving consistency on an aggregated level of a large number of figures from different sources that are related with one another through a large number of constraints. Macro-integration techniques are based on constraint optimization methods that need to be able to handle large numbers of estimates and constraints, such as in applications to macro-economic accounts and supply and use tables (see [13,14]).

Our aim is to have a single estimate for the variable 'Contract type' that combines available information from the monthly and quarterly data of 'Contract type' from ER and LFS and transition probabilities.

We apply reconciliation methods that result in a single series of monthly figures of probabilities for the variable 'Contract type". The objective of this optimization is that the estimated monthly figures are as close as possible to the corresponding monthly values from ER and the aggregates of these figures into quarterly values are as close as possible to the original quarterly values from LFS.

3.1. Objective function

Consider the four-way probability table corresponding to the four time points t to $t+3$ as depicted in Table 3 with cell-probabilities $\pi^{t,t+1,t+2,t+3}_{c_tc_{t+1}c_{t+2}c_{t+3}}$. All parameters of interest can be obtained as univariate and bivariate marginal tables with probabilities obtained from this four-way table. A reconciliation strategy is now to estimate this table $\tilde{\boldsymbol{\pi}}^{t,t+1,t+2,t+3}$ using all the information from both sources, so that it is as close as possible to the register four-way proportions, $\mathbf{p}^{R.t,t+1,t+2,t+3}$ and has marginal probabilities $\tilde{\boldsymbol{\pi}}^{t,t+3}$ which are as close as possible to the corresponding survey proportions, $\mathbf{p}^{S.t,t+3}$. From this table we can also obtain all required univariate and bivariate marginal probabilities.

This reconciliation problem is defined as constrained minimization problem of an objective function, which can be formulated in general as:

$$\begin{aligned} D = {} & D^R(\mathbf{p}^{R.t,t+1,t+2,t+3}, \tilde{\boldsymbol{\pi}}^{t,t+1,t+2,t+3}) \\ & + D^S(\mathbf{p}^{S.t,t+3}, \tilde{\boldsymbol{\pi}}^{t,t+3}) \text{ for } t \in (1,4,7,10), \end{aligned} \tag{1}$$

with D^R a measure of the discrepancy between the reconciled proportions $\tilde{\boldsymbol{\pi}}^{t,t+1,t+2,t+3}$ and the observed register proportions $\mathbf{p}^{R.t,t+1,t+2,t+3}$ and D^S a measure of the discrepancy between the reconciled proportions $\tilde{\boldsymbol{\pi}}^{t,t+3}$ and the observed survey proportions $\mathbf{p}^{S.t,t+3}$. Note that in D we have separate components for the reconciled estimates of the four dimensional table $\tilde{\boldsymbol{\pi}}^{t,t+1,t+2,t+3}$ and its two-way margin $\tilde{\boldsymbol{\pi}}^{t,t+3}$. To enforce that this two-way table is indeed a margin of the four-way table, the following constraints must be satisfied:

$$\sum_{c_{t+1},c_{t+2}} \tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} = \tilde{\pi}^{t,t+3}_{c_t c_{t+3}} \quad \text{for } t \in (1,4,7,10).$$

Different reconciliation models can be defined by different specifications of the objective function and constraints. Below we define a number of such models.

Constraints

As mentioned above, since we have a linked data set on a micro level we mainly have to deal with measurement errors. These errors can occur in both ER and LFS variables for different reasons as discussed above.

For some of the reconciliation models we choose to adjust the register data and let the survey data remain unchanged. These models will have fixed survey-proportions and in Eq. (1) the component D^S vanishes. The constraints on the reconciled proportions $\pi^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}}$ for these models can be written as:

$$\sum_{c_{t+1},c_{t+2}} \tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} = p^{S.t,t+3}_{c_t c_{t+3}}, \quad t \in (1,4,7,10). \tag{2}$$

These constraints ensure that the bivariate marginal proportions for time points $t, t+3$ after reconciliation are equal to the observed survey-counterparts.

For models where both the register and the survey proportions can be adjusted, we have the following equality constraints:

$$\sum_{c_t c_{t+1} c_{t+2} c_{t+3}} \tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} = 1, \quad t \in (1,4,7,10), \tag{3a}$$

$$\sum_{c_{t+1},c_{t+2}} \tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} = \tilde{\pi}^{t,t+3}_{c_t c_{t+3}}, \quad t \in (1,4,7,10), \tag{3b}$$

$$\sum_{c_1} \tilde{\pi}^{t=1,t=4}_{c_1,c_4} = \sum_{c_7} \tilde{\pi}^{t=4,t=7}_{c_4 c_7}, \tag{3c}$$

$$\sum_{c_4} \tilde{\pi}^{t=4,t=7}_{c_4,c_7} = \sum_{c_{10}} \tilde{\pi}^{t=7,t=10}_{c_7 c_{10}}, \tag{3d}$$

$$\sum_{c_7} \tilde{\pi}^{t=7,t=10}_{c_7,c_{10}} = \sum_{c_{13}} \tilde{\pi}^{t=10,t=13}_{c_{10} c_{13}}. \tag{3e}$$

Constraint (3a) ensures that the values $\tilde{\pi}$ can be interpreted as probabilities.

Constraint (3b) ensures that the table $\tilde{\pi}^{t,t+3}_{c_t c_{t+3}}$ is a bivariate margin of the four-way table $\tilde{\boldsymbol{\pi}}^{t,t+1,t+2,t+3}$.

Constraints (3c)–(3e) enforce the equality of the common univariate margins of the bivariate tables.

3.2. Specific models

For the function D we will consider two alternatives. The first is the usual (weighted) sum of squares, which leads to additive adjustments and an explicit solution to the optimization problem and the second is the Kullback-Leibler divergence function which leads to multiplicative adjustments, see e.g. [15]. These adjustments cannot explicitly be calculated but a simple iterative algorithm can be used. The solutions of these optimization problems are derived in Appendix A. Together with the two choices of constraints, we arrive at the following four reconciliation models.

Model 1. Additive adjustment of register only

This model treats the survey estimates as fixed quantities and only adjusts the register proportions. The objective function for this model is the commonly used least-squares loss-function:

$$D^R_{LS} = \sum_{c_t c_{t+1} c_{t+2} c_{t+3}} \left(p^{R.t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} - \tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} \right)^2. \tag{4}$$

This objective function leads to additive adjustments. For the case of the fixed-sum constraints Eq. (2) these adjustments are very simple (see Appendix A): if a sum of register proportions is not equal to the corresponding survey proportion, a constant is added to all these register proportions such that the sum becomes equal the survey proportion:

$$\begin{aligned} \tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} &= p^{R.t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} \\ &\quad + \left(p^{S.t,t+3}_{c_t c_{t+3}} - p^{R.t,t+3}_{c_t c_{t+3}} \right) / 9 \\ &= p^{R.t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} + a^{t,t+3}_{c_t c_{t+3}}, \text{ for } t \in (1,4,7,10). \end{aligned} \tag{5}$$

Summation of Eq. (5) over c_{t+1} and c_{t+2} (9 components) shows that Eq. (2) is verified.

For the univariate reconciled proportions we have

Table 4
Reconciled proportions of contract types according to Model 1

	Group 1			Group 2			Group 3		
Month	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.
1	0.6568	0.1117	0.2314	–	–	–	–	–	–
2	0.5899	0.1553	0.2548	0.6541	0.1159	0.2300	–	–	–
3	0.5877	0.1562	0.2561	0.5752	0.1649	0.2598	0.6587	0.1121	0.2291
4	0.6643	0.1050	0.2307	0.5745	0.1639	0.2616	0.6032	0.1514	0.2454
5	0.5888	0.1549	0.2563	0.6703	0.1059	0.2238	0.6010	0.1502	0.2488
6	0.5894	0.1518	0.2588	0.5716	0.1660	0.2624	0.6736	0.1034	0.2230
7	0.6597	0.1070	0.2333	0.5695	0.1636	0.2670	0.5972	0.1469	0.2559
8	0.5838	0.1489	0.2673	0.6639	0.1066	0.2296	0.5961	0.1473	0.2566
9	0.5844	0.1474	0.2683	0.5673	0.1634	0.2692	0.6735	0.1030	0.2235
10	0.6624	0.1039	0.2336	0.5705	0.1606	0.2689	0.5950	0.1458	0.2592
11	0.5845	0.1494	0.2661	0.6613	0.1168	0.2218	0.5935	0.1439	0.2626
12	0.5817	0.1482	0.2702	0.5694	0.1609	0.2697	0.6734	0.1044	0.2223
13	0.6611	0.1008	0.2381	0.5764	0.1535	0.2701	0.5961	0.1415	0.2625
14	–	–	–	0.6562	0.1148	0.2290	0.5972	0.1359	0.2669
15	–	–	–	–	–	–	0.6724	0.1009	0.2268

$$\begin{aligned} \tilde{\pi}^{t}_{c_t} &= p^{S.t}_{c_t} \quad \text{for } t \in (1,4,7,10,13), \\ \tilde{\pi}^{t}_{c_t} &= p^{R.t}_{c_t} \quad \text{for } t \notin (1,4,7,10,13), \end{aligned} \tag{6}$$

where the first line is a consequence of Eq. (2) and the second line is true as summing Eq. (5) over c_t and c_{t+3} yields $p^{R.t+1,t+2}_{c_{t+1}c_{t+2}}$, because $\sum_{c_t c_{t+3}} p^{S.t,t+3}_{c_t c_{t+3}} = \sum_{c_t c_{t+3}} p^{R.t,t+3}_{c_t c_{t+3}} = 1$.

For the adjusted transition proportions for consecutive time points we can write,

$$\begin{aligned} \tilde{\pi}^{t,t+1}_{c_t c_{t+1}} &= p^{R.t,t+1}_{c_t c_{t+1}} + \left(p^{S.t}_{c_t} - p^{R.t}_{c_t}\right)/3 \\ \tilde{\pi}^{t+1,t+2}_{c_{t+1} c_{t+2}} &= p^{R.t+1,t+2}_{c_{t+1} c_{t+2}} \\ \tilde{\pi}^{t+2,t+3}_{c_{t+2} c_{t+3}} &= p^{R.t+2,t+3}_{c_{t+2} c_{t+3}} + \left(p^{S.t+3}_{c_{t+3}} - p^{R.t+3}_{c_{t+3}}\right)/3, \end{aligned} \tag{7}$$

with $p^{S.t}_{c_t}$ and $p^{S.t+3}_{c_{t+3}}$ the row- and column marginals of $p^{S.t,t+3}_{c_t c_{t+3}}$ and $p^{R.t}_{c_t}$ and $p^{R.t+3}_{c_{t+3}}$ the row- and column marginals of $p^{R.t,t+3}_{c_t c_{t+3}}$.

This shows that the univariate reconciled distributions of 'Contract type' for the time points t and $t+3$ are equal to those of the survey, which is an obvious consequence of the constraint Eq. (2), whereas for the intermediate time points the distributions are equal to the register ones. For the bivariate distribution or transition probabilities, we see that the reconciled transitions from t to $t+1$ and from $t+2$ to $t+3$ will differ from the register ones, whereas the intermediate transitions (from $t+1$ to $t+2$) will not change. The changes are such that to each row (or column) of the transition table a constant is added. This may be a counterintuitive result, but it follows directly from the cancellation of adjustments when summing Eq. (5) over c_t and c_{t+3}.

From the reconciled transition table we can easily obtain marginal proportions. These reconciled marginal time point distributions are presented in Table 4. These results confirm that for time points t and $t+3$ the proportions equal the ones obtained from the survey while for the intermediate time points the proportions are equal to the register proportions.

Note that this is rather a naive model. The reconciled series have large jumps and are therefore not realistic.

Model 2. Multiplicative adjustment of register only

Just as Model 1, this model treats the survey estimates as fixed quantities and only adjusts the register proportions. The objective function for this model is the Kullback-Leibler divergence or relative entropy. Although this objective function may not be the most common one, the resulting adjustment method is perhaps the most intuitive one. It leads to multiplicative adjustments that guarantee non-negative outcomes and are easy to understand and apply. The Kullback-Leibler divergence is given by:

$$\begin{aligned} D^{R}_{KL} = \sum_{c_t c_{t+1} c_{t+2} c_{t+3}} & \tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} \\ & \left(\log(\tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} / p^{R.t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}}) - 1\right) \end{aligned} \tag{8}$$

We show in Appendix A that the solution to minimizing D^{R}_{KL} subject to the constraints Eq. (2) can be obtained by multiplying the original register proportions by factors varying only over combinations of categories of t and $t+3$ but remaining constant over the combinations of categories of $t+1$ and $t+2$:

$$\begin{aligned} \tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} &= p^{R.t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} \\ &\quad \times p^{S.t,t+3}_{c_t c_{t+3}} / p^{R.t,t+3}_{c_t c_{t+3}} \\ &= p^{R.t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} \times f^{t,t+3}_{c_t c_{t+3}}, \quad t \in (1,4,7,10). \end{aligned} \tag{9}$$

Table 5
Reconciled proportions of contract types according to Model 2

	Group 1			Group 2			Group 3		
Month	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.
1	0.6568	0.1117	0.2314	–	–	–	–	–	–
2	0.6581	0.1120	0.2299	0.6541	0.1159	0.2300	–	–	–
3	0.6557	0.1117	0.2326	0.6539	0.1166	0.2296	0.6587	0.1121	0.2291
4	0.6643	0.1050	0.2307	0.6546	0.1146	0.2308	0.6672	0.1074	0.2254
5	0.6617	0.1107	0.2276	0.6703	0.1059	0.2238	0.6659	0.1066	0.2275
6	0.6628	0.1084	0.2287	0.6681	0.1076	0.2243	0.6736	0.1034	0.2230
7	0.6597	0.1070	0.2333	0.6650	0.1067	0.2282	0.6693	0.1030	0.2277
8	0.6575	0.1083	0.2343	0.6639	0.1066	0.2296	0.6661	0.1055	0.2284
9	0.6572	0.1071	0.2356	0.6630	0.1104	0.2266	0.6735	0.1030	0.2235
10	0.6624	0.1039	0.2336	0.6645	0.1105	0.2250	0.6703	0.1044	0.2253
11	0.6625	0.1049	0.2325	0.6613	0.1168	0.2218	0.6677	0.1046	0.2278
12	0.6594	0.1039	0.2367	0.6607	0.1153	0.2240	0.6734	0.1044	0.2223
13	0.6611	0.1008	0.2381	0.6592	0.1161	0.2247	0.6709	0.1040	0.2250
14	–	–	–	0.6562	0.1148	0.2290	0.6715	0.0995	0.2290
15	–	–	–	–	–	–	0.6724	0.1009	0.2268

Summation of Eq. (9) over c_{t+1} and c_{t+2} shows that Eq. (2) is verified.

For the adjusted univariate proportions we then have, for $t \in (1, 4, 7, 10)$:

$$\begin{aligned}
\tilde{\pi}^{t}_{c_t} &= p^{S.t}_{c_t},\\
\tilde{\pi}^{t+1}_{c_{t+1}} &= \sum_{c_t c_{t+3}} \Big(f^{t,t+3}_{c_t c_{t+3}} \times \sum_{c_{t+2}} p^{R.t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} \Big),\\
\tilde{\pi}^{t+2}_{c_{t+2}} &= \sum_{c_t c_{t+3}} \Big(f^{t,t+3}_{c_t c_{t+3}} \times \sum_{c_{t+1}} p^{R.t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} \Big),\\
\tilde{\pi}^{t+3}_{c_{t+3}} &= p^{S.t+3}_{c_{t+3}}.
\end{aligned} \tag{10}$$

Here we see that, contrary to the additive adjustments, none of the univariate distributions remain equal to the direct register estimates. The distributions for the intermediate time points $t+1$ and $t+2$ are influenced by the ratios $p^{S.t,t+3}_{c_t c_{t+3}} / p^{R.t,t+3}_{c_t c_{t+3}}$.

For the reconciled transition proportions for consecutive time points we can write:

$$\begin{aligned}
\tilde{\pi}^{t,t+1}_{c_t c_{t+1}} &= \sum_{c_{t+3}} \Big(f^{t,t+3}_{c_t c_{t+3}} \times \sum_{c_{t+2}} p^{R.t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} \Big),\\
\tilde{\pi}^{t+1,t+2}_{c_{t+1} c_{t+2}} &= \sum_{c_t c_{t+3}} \Big(f^{t,t+3}_{c_t c_{t+3}} \times p^{R.t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} \Big),\\
\tilde{\pi}^{t+2,t+3}_{c_{t+2} c_{t+3}} &= \sum_{c_t} \Big(f^{t,t+3}_{c_t c_{t+3}} \times \sum_{c_{t+1}} p^{R.t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} \Big),
\end{aligned} \tag{11}$$

where we see again that, contrary to the additive adjustments, there are now no transitions that remain equal to the observed register transitions: all transitions are affected by the differences between survey and register proportions as expressed by the factors $f^{t,t+3}_{c_t c_{t+3}}$.

The results for the marginal time point distributions for this adjustment model are in Table 5. These results show that the artificial 'jumps' in the distributions for the months where survey estimates are available, that were found for model 1, have now disappeared. Not surprisingly, we also see that the level of the intermediate proportions is now much closer to the ones obtained from the survey.

Model 3. Additive adjustment of both survey and register

For this model we consider the least-squares objective with components D^R_{LS} and D^S_{LS}:

$$\begin{aligned}
D_{LS} = & \sum_{c_t c_{t+1} c_{t+2} c_{t+3}} \\
& \Big(p^{R.t,t+1,t+2,t+3} - \tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} \Big)^2 \\
& + \sum_{c_t c_{t+3}} \Big(p^{S.t,t+3}_{c_t c_{t+3}} - \tilde{\pi}^{t,t+3}_{c_t c_{t+3}} \Big)^2 .
\end{aligned} \tag{12}$$

This function is minimised over $\tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}}$ and $\tilde{\pi}^{t,t+3}_{c_t c_{t+3}}$ subject to the constraints Eq. (3). The solution to this problem can still be obtained in closed form, but is slightly more involved and is given in Appendix A.

The results for the marginal time point distributions for this adjustment model are presented in Table 6. These results show that, similarly to the additive model 1, the proportions for the intermediate time points are equal the the register proportions but, contrary to model 1, the proportions for the time points where survey estimates are available are not equal to these survey estimates but are in-between the survey and register estimates.

Table 6
Reconciled proportions of contract types according to Model 3

	Group 1			Group 2			Group 3		
Month	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.
1	0.6497	0.1159	0.2343	–	–	–	–	–	–
2	0.5899	0.1553	0.2548	0.6463	0.1206	0.2330	–	–	–
3	0.5877	0.1562	0.2561	0.5752	0.1649	0.2598	0.6526	0.1164	0.2311
4	0.6542	0.1102	0.2356	0.5745	0.1639	0.2616	0.6032	0.1514	0.2454
5	0.5888	0.1549	0.2563	0.6537	0.1150	0.2313	0.6010	0.1502	0.2488
6	0.5894	0.1518	0.2588	0.5716	0.1660	0.2624	0.6629	0.1091	0.2279
7	0.6526	0.1106	0.2368	0.5695	0.1636	0.2670	0.5972	0.1469	0.2559
8	0.5838	0.1489	0.2673	0.6549	0.1115	0.2336	0.5961	0.1473	0.2566
9	0.5844	0.1474	0.2683	0.5673	0.1634	0.2692	0.6618	0.1089	0.2293
10	0.6522	0.1089	0.2389	0.5705	0.1606	0.2689	0.5950	0.1458	0.2592
11	0.5845	0.1494	0.2661	0.6529	0.1204	0.2267	0.5935	0.1439	0.2626
12	0.5817	0.1482	0.2702	0.5694	0.1609	0.2697	0.6633	0.1078	0.2289
13	0.6537	0.1048	0.2414	0.5764	0.1535	0.2701	0.5961	0.1415	0.2625
14	–	–	–	0.6481	0.1182	0.2337	0.5972	0.1359	0.2669
15	–	–	–	–	–	–	0.6649	0.1047	0.2303

Model 4. Multiplicative adjustment of both survey and register

For this model we consider a Kullback-Leibler objective that contains both the D^R_{KL} and D^S_{KL} components:

$$D_{KL} = \sum_{c_t c_{t+1} c_{t+2} c_{t+3}} \tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} \times \left(\log(\tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}} / p^{R.t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}}) - 1\right) + \sum_{c_t c_{t+3}} \tilde{\pi}^{t,t+3}_{c_t c_{t+3}} \left(\log(\tilde{\pi}^{t,t+3}_{c_t c_{t+3}} / p^{S.t,t+3}_{c_t c_{t+3}}) - 1\right). \quad (13)$$

Again this function is minimised over $\tilde{\pi}^{t,t+1,t+2,t+3}_{c_t c_{t+1} c_{t+2} c_{t+3}}$ and $\tilde{\pi}^{t,t+3}_{c_t c_{t+3}}$ and is subject to the constraints Eq. (3). The solution to this problem cannot be given in closed form, but a convenient iterative algorithm is given in Appendix A.

The results for the marginal time point distributions for this adjustment model are in Table 7. These results show that the artificial 'jumps' for the months where survey estimates are available, that were found for the additive models (for model 3 to a lesser extent than for model 1), have disappeared (just as for the multiplicative model 2). But, contrary to model 2, the level of all proportions has moved from what was found in the survey towards the level found in the register. For the category permanent contracts this means a decrease while for temporary contracts this is an increase.

We considered four different models. In two of these models we assumed that the quarterly survey figures were trustworthy and we could fix them, only adjusting the monthly register data. However, when we have no additional information about reliability of the two sources, nor whether one is more reliable than the other, we cannot fix the figures. In that case we observed that the optimisation model with a Kullback-Leiber objective function performs best.

4. Comparison with results from the measurement error model

In this section we describe results obtained by [11] using a latent class model on the same data set as used in this paper to estimate the true underlying variable 'Contract type'. The latent class models assume that the estimates from different sources differ from the true values (and each other) due to (differences in) measurement errors. The true variable is called a latent variable. Estimates of proportions based on the true values will provide single estimates of the corresponding population values that are corrected for measurement errors. In [11] hidden Markov models (HMM) have been proposed for analysing longitudinal categorical data. Similar data, but for an earlier period, between January 2007 and March 2009, was used by [16] to analyse the 'Contract type' variable applying a latent Markov model.

We compare the results presented in [11] with the results of the macro-integration approach presented above. For this comparison we shall first discuss the differences in combining the data from the different groups, then highlight some properties of the HMM that make this methodology different from our approach and finally compare the numerical results.

4.1. Differences in combining the groups

In our analyses we consider three different groups, that are defined by the starting point of the LFS panel.

Table 7
Adjusted monthly proportions of contract types in Model 4

	Group 1			Group 2			Group 3		
Month	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.
1	0.6247	0.1311	0.2441	–	–	–	–	–	–
2	0.6275	0.1319	0.2406	0.6213	0.1354	0.2433	–	–	–
3	0.6254	0.1321	0.2425	0.6207	0.1366	0.2427	0.6347	0.1296	0.2357
4	0.6286	0.1267	0.2447	0.6208	0.1350	0.2442	0.6421	0.1257	0.2322
5	0.6259	0.1316	0.2425	0.6205	0.1344	0.2451	0.6404	0.1247	0.2349
6	0.6269	0.1289	0.2442	0.6188	0.1360	0.2452	0.6400	0.1242	0.2358
7	0.6242	0.1254	0.2504	0.6163	0.1344	0.2493	0.6367	0.1227	0.2406
8	0.6220	0.1267	0.2512	0.6167	0.1322	0.2511	0.6345	0.1243	0.2412
9	0.6223	0.1254	0.2524	0.6153	0.1357	0.2491	0.6334	0.1248	0.2418
10	0.6223	0.1240	0.2537	0.6174	0.1347	0.2479	0.6312	0.1251	0.2437
11	0.6223	0.1258	0.2519	0.6156	0.1388	0.2457	0.6292	0.1245	0.2463
12	0.6192	0.1246	0.2561	0.6156	0.1359	0.2484	0.6306	0.1227	0.2467
13	0.6224	0.1202	0.2574	0.6171	0.1339	0.2490	0.6290	0.1228	0.2482
14	–	–	–	0.6146	0.1313	0.2541	0.6298	0.1177	0.2525
15	–	–	–	–	–	–	0.6307	0.1200	0.2493

Table 8
Proportions of 'Contract type' in the LFS for pooled data

Month	Permanent	Temporary	Other
1	0.6566	0.1131	0.2303
2	–	–	–
3	–	–	–
4	0.6692	0.1048	0.2261
5	–	–	–
6	–	–	–
7	0.6655	0.1056	0.2289
8	–	–	–
9	–	–	–
10	0.6656	0.1081	0.2262
11	–	–	–
12	–	–	–
13	0.6632	0.1052	0.2316

Table 9
Proportions of 'Contract type' in the ER for pooled data

Month	Permanent	Temporary	Other
1	0.5861	0.1571	0.2568
2	0.5892	0.1572	0.2536
3	0.5875	0.1569	0.2556
4	0.5880	0.1550	0.2570
5	0.5858	0.1561	0.2582
6	0.5850	0.1542	0.2608
7	0.5838	0.1517	0.2645
8	0.5819	0.1527	0.2654
9	0.5827	0.1506	0.2667
10	0.5830	0.1508	0.2662
11	0.5831	0.1507	0.2662
12	0.5847	0.1461	0.2691
13	0.5869	0.1433	0.2698

The groups entered the LFS in January, February and March of 2009. We applied reconciliation models separately to each group. In [11] these three groups were pooled together. Authors ignore time, by combining persons that start in January, February and March in one group. In order to compare our results with the results from [11] we combined the data in the same way. We call this data the pooled data. Tables 8 and 9 show the monthly distributions of pooled data for LFS and ER. In Table 1 we observe that the group 3 has slightly higher proportions of persons with permanent contract than the group 1 and group 2 has slightly lower proportions of persons with permanent contract than group 1. We observe a decrease in time in Temporary contracts in all three groups. In the pooled data, these differences are averaged for a permanent contract. Also in that case we are seeing a decrease in Temporary contracts.

Table 10
Path diagram for the latent Markov model with two (partially) observed indicators

$$\begin{array}{ccccccccc} \dots & S^t & & & & & & S^{t+3} & \dots \\ & \uparrow & & & & & & \uparrow & \\ \dots & X^t & \rightarrow & X^{t+1} & \rightarrow & X^{t+2} & \rightarrow & X^{t+3} & \dots \\ & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\ \dots & R^t & \rightarrow & R^{t+1} & \rightarrow & R^{t+2} & \rightarrow & R^{t+3} & \dots \end{array}$$

4.2. The latent Markov model approach versus macro-integration approach

The latent Markov model can be described globally with the help of the diagram below (see Table 10), which is taken (with some modifications) from [16]. It is also similar to the picture of the macro-integration approach in Table 3 in Section 2.2 but, contrary to that picture, it describes measurements and true values at the individual level rather than the observed proportions and unknown population proportions aggregated over units.

Here the S^t denote the quarterly survey measurements of 'Contract type' for any individual and the R^t denote the register measurements. The X^t denotes a persons true contract status which is an unobserved or latent variable. Associated with the latent variables are the contract status probability vectors $\boldsymbol{\pi}^t$, defined earlier. Each of the cells in this table represents a specific sequence of categories of X in time (a path) and the probability of a specific path is the cell probability, denoted by $\pi_{c_1 c_2 \ldots c_T}$. The arrows between the X^t indicate that the 'Contract type' at time point t depends on the 'Contract type' of the previous time point $t-1$, but given the 'Contract type' at $t-1$ it does not depend on any of the time points before $t-1$. This is the Markov property, which means that,

$$\Pi^{t|t-1,t-2,\ldots,1}_{c_t,c_{t-1},c_{t-2},\ldots,c_1} = \Pi^{t|t-1}_{c_t,c_{t-1}}. \tag{14}$$

Here $\Pi^{t|t-1}_{c_t,c_{t-1}}$ and $\Pi^{t|t-1,t-2,\ldots,1}_{c_t,c_{t-1},c_{t-2},\ldots,c_1}$ are conditional transition probabilities. $\Pi^{t|t-1}_{c_t,c_{t-1}}$ is a transition probability of the units having the value c_{t-1} at the time point $t-1$ and the value c_t at time point t. These probabilities sum to 1 for each category of c_t and are defined as follows:

$$\Pi^{t|t-1}_{c_t,c_{t-1}} = \pi^{t,t-1}_{c_t,c_{t-1}} / \pi^{t-1}_{c_{t-1}}. \tag{15}$$

Similarly, $\Pi^{t|t-1,t-2,\ldots,1}_{c_t,c_{t-1},c_{t-2},\ldots,c_1}$ denote the transition probability of units that have the values $c_{t-1}, c_{t-2}, \ldots, c_1$ at time points $t-1, t-2, \ldots, 1$ and c_t at time point t.

The consequence of the Markov property in Eq. (14) is that the probability of any possible sequence of categories of X in time (a path), is completely specified by the initial state probability and the transition probabilities, since:

$$\begin{aligned}\pi^{1,\ldots,T}_{c_1\ldots c_T} &= \pi^{1,\ldots,T-1}_{c_{1,\ldots,T-1}} \cdot \Pi^{T|1,\ldots,T-1}_{c_T,c_1,\ldots,c_{T-1}} \\ &= \pi^{1,\ldots,T-1}_{c_{1,\ldots,T-1}} \cdot \Pi^{T|T-1}_{c_T,c_{T-1}}\end{aligned}$$

and, by recursion on the first term on the right hand site,

$$\pi^{1,\ldots,T}_{c_1\ldots c_T} = \pi^1_{c_1} \Pi^{2|1}_{c_2,c_1} \Pi^{3|2}_{c_3,c_2} \ldots \Pi^{T|T-1}_{c_T,c_{T-1}}. \tag{16}$$

Now if we use Eq. (15) for the transition probabilities $\Pi^{2|1}_{c_2,c_1}, \ldots, \Pi^{T|T-1}_{c_T,c_{T-1}}$ we can rewrite Eq. (16) in terms of the bivariate probabilities $\pi^{2,1}_{c_2,c_1}/\pi^1_{c_1}, \ldots, \pi^{T,T-1}_{c_T,c_{T-1}}/\pi^{T-1}_{c_{T-1}}$ and so, by the Markov assumption, the cell probabilities are completely determined by the bivariate probabilities.

The model links the observed values from each of the sources to the true values by a matrix with misclassification probabilities that gives the probability of an observed category of R^t or S^t given X^t, the diagonal entries of these matrices are the probabilities of correct classification. These matrices, denoted by $p(R^t|X^t)$ and $p(S^t|X^t)$, together with the transition probabilities $\Pi^{t|t-1}_{c_t,c_{t-1}}$, are the parameters of the model and can be estimated provided that some restrictions are imposed. Usually it is assumed that the error matrices $p(R^t|X^t)$ and $p(S^t|X^t)$ only depend on X^t and not on other true or observed values, but [11] and [16] use an extensions of the model that allows for certain correlations between classification errors at successive time points.

In the model-based approach the observed values S^t and R^t are thought to be generated as a function of the true values X^t and added measurement errors, hence the direction of the arrows. The reconciliation approach is not based on a data generating model at the individual level, but begins with estimates at the macro-level and combines these different estimates into a single one. Thus, in the reconciliation approach the observed estimates are the inputs from which the reconciliation model can produce the single estimates whereas in the model based approach the single true values are thought to produce the different measurements. To reflect this difference in reasoning, the direction of the arrows is reversed for the reconciliation approach.

Despite these conceptual differences, both approaches will result in a single estimate of each of the parameters of interest: the univariate and bivariate proportions (or probabilities) of the 'Contract type' categories over time.

4.3. Numerical results

Since from the four macro-integration models defined in Section 3.2 the Model 4 performed best, we compare results only of this model. In order to compare results from the HMM and the Model 4, we run the Model 4 on the pooled data. Results of the reconciled time series for the pooled data are given in Tables 15 and 16.

In Table 10 in [11] the average size of 'Contract type' according to the LFS, ER and their optimal latent model, called C", are presented. Here we compare these results with the results we obtained from Model 4. In our model the average size of 'Contract type' is obtained in a different way. In [11] the model estimates new values on a micro level, whereas our models produce adjusted values per month for the total group, on a macro level. The average size of 'Contract type' for model 4 in Table 11 here is the average over 13 months for Model 4. Observe that the adjusted figures for Model 4 are closer to the average values obtained

Table 11
The average size of 'Contract type' according to model C" and model 4

	Survey	Register	Model C"	Model 4
Permanent	0.653	0.585	0.611	0.625
Temporary	0.110	0.151	0.128	0.129
Other	0.237	0.264	0.261	0.246

from these two sources. This is not surprising since in our optimization function we include all initial values as equally reliable. The latent class model gives slightly different results, with the largest deviation from Model 4 in the category "Other".

As mentioned in Section 2.2, interest is not only in the marginal distribution $\boldsymbol{\pi}^t$ over the three contract types for each month, the changes between these categories over time are also of interest. These can be described by the transition probabilities between adjacent time points. Estimates of these probabilities are shown in Table 12. This table compares conditional probabilities $P(c_t|c_{t-3})$ for observed survey data, observed register data, Model 4 results for pooled data and results from Table 11 in [11]. We again observe that the macro-integration model estimates are in between the observed values, but now this is not true for the estimates from Model C". The diagonal value estimates from model C" appear to be overall higher than the observed values in either LFS or ER while the transition probabilities are lower according to this model.

We anticipated that the macro-integration model estimates would lie between the observed values, since these estimates are required to be as close as possible to the observed values of both sources. The higher estimates on the diagonal in the HMM could be due to the Markov assumption. It seems that in HMM the transitions between different employment states are partly accounted for as measurement error (that exists in both data sources) resulting in lower off diagonal and higher diagonal estimates.

Table 12
Conditional probabilities of 'Contract type'

Observed transitions from LFS			
	Contract t		
Contract $t-3$	Permanent	Temporary	Other
Perm.	0.983	0.006	0.011
Temp	0.058	0.879	0.063
Oth.	0.016	0.037	0.947
Observed transitions from ER register data			
	Contract t		
Contract $t-3$	Permanent	Temporary	Other
Perm.	0.976	0.012	0.012
Temp	0.073	0.869	0.058
Oth.	0.019	0.043	0.938
Model 4 for pooled data			
	Contract t		
Contract $t-3$	Permanent	Temporary	Other
Perm.	0.979	0.008	0.012
Temp	0.062	0.874	0.064
Oth.	0.016	0.036	0.948
Model C" in [11], Table 11			
	Contract t		
Contract $t-3$	Permanent	Temporary	Other
Perm.	0.987	0.004	0.009
Temp	0.017	0.929	0.054
Oth.	0.006	0.030	0.963

5. Discussion

In this paper we applied macro integration methods to first-order (marginal probabilities) and second-order statistics (transition probabilities) in order to obtain consistent estimates of the variable 'Contract type' and also of the transition probabilities between different contract types.

Four different macro-integration models were defined in Section 3. We presented adjusted proportions of the variable 'Contract type' for these models, see Tables 4–7. From these results, it is easy to argue that model 4 shows the best results. First we observed that the multiplicative adjustments were superior to additive ones. The additive model adjusted only quarterly proportions, while the proportions of the intermediate months were not adjusted, see Tables 4 and 6, resulting in unrealistic monthly time series. Secondly, we assumed that both sources are equally reliable or equally wrong, since we do not have any evidence for one source to be more reliable than the other. When the adjustments to both sources are allowed we saw that a multiplicative adjustment model (Model 4) performed best.

Our macro-integration models resulted in estimated proportions for each group (survey cohort) separately. These estimates can easily be combined to form an overall estimate. From the proportions for the groups we first derive estimates for the totals for each group see Table 13. We then combine the totals of all three groups for each month and calculate the adjusted total proportions, see Table 14.

Some issues still remain unsolved. For example, by using linked data we avoided coverage problems. In practical applications, when we might not have linked data, coverage problems should be resolved before applying macro-integration methods. In addition, the discrepancies we have observed in linked data may still be

Table 13
Estimated monthly totals of contract types in Model 4

Group	1			2			3		
Contract type	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.	Perm.	Temp.	Oth.
Month 1	1983.5	416.4	775.2	–	–	–	–	–	–
2	1992.9	418.9	764.2	1770.1	385.9	693.1	–	–	–
3	1986.2	419.4	770.3	1766.6	388.7	690.8	1708.5	348.9	634.6
4	1996.4	402.3	777.3	1766.9	384.2	694.9	1728.5	338.3	625.2
5	1988.0	417.9	770.2	1765.9	382.6	697.4	1723.8	335.8	632.3
6	1991.7	409.4	776.0	1760.6	386.9	697.5	1721.6	334.1	634.3
7	1983.0	398.4	795.5	1755.8	382.8	710.4	1713.4	330.1	647.5
8	1976.2	402.6	798.1	1755.7	376.5	714.8	1707.4	334.6	649.1
9	1976.4	398.1	801.5	1751.6	386.3	709.1	1703.9	335.7	650.5
10	1976.5	393.7	805.8	1757.1	383.3	705.6	1698.6	336.7	655.7
11	1976.3	399.6	800.0	1752.5	395.0	699.5	1692.5	334.9	662.7
12	1967.2	396.0	813.8	1750.9	386.6	706.6	1695.7	329.9	663.5
13	1975.5	381.6	816.8	1755.0	380.8	708.2	1690.7	330.1	667.2
14	–	–	–	1748.5	373.5	723.0	1693.0	316.3	678.7
15	–	–	–	–	–	–	1694.7	322.6	669.7

Table 14
Estimates of combined monthly proportions of contract types

Contract type	Perm.	Temp.	Oth.
Month 1	0.6247	0.1311	0.2441
2	0.6246	0.1336	0.2419
3	0.6267	0.1328	0.2405
4	0.6302	0.1291	0.2407
5	0.6286	0.1304	0.2410
6	0.6283	0.1297	0.2419
7	0.6255	0.1275	0.2470
8	0.6241	0.1278	0.2481
9	0.6234	0.1286	0.2480
10	0.6235	0.1278	0.2487
11	0.6222	0.1296	0.2482
12	0.6216	0.1277	0.2507
13	0.6227	0.1255	0.2518
14	0.6220	0.1247	0.2533
15	0.6307	0.1200	0.2493

Table 15
Adjusted proportions of 'Contract type' in the Polis for combined groups for model 4

Month	Perm.	Temp.	Oth.
1	0.6267	0.1321	0.2412
2	0.6298	0.1316	0.2386
3	0.6285	0.1308	0.2407
4	0.6294	0.1284	0.2421
5	0.6269	0.1300	0.2431
6	0.6257	0.1293	0.2450
7	0.6245	0.1276	0.2479
8	0.6226	0.1292	0.2482
9	0.6227	0.1281	0.2492
10	0.6226	0.1284	0.2490
11	0.6218	0.1287	0.2495
12	0.6219	0.1255	0.2526
13	0.6225	0.1238	0.2537

caused by different definitions or measurement errors. These possible causes of discrepancies should be investigated thoroughly because macro-integration methods should be applied only after the data are cleaned from gross errors.

Table 16
Adjusted proportions of 'Contract type' in the LFS for combined groups for model 4

Month	Perm.	Temp.	Oth.
1	0.6267	0.1321	0.2412
2	–	–	–
3	–	–	–
4	0.6294	0.1284	0.2421
5	–	–	–
6	–	–	–
7	0.6245	0.1276	0.2479
8	–	–	–
9	–	–	–
10	0.6226	0.1284	0.2490
11	–	–	–
12	–	–	–
13	0.6225	0.1238	0.2537

In [11] another model-based approach, the latent class model with the Markov property is applied on the same data set to obtain the transition probabilities. In comparison to the macro-integration approach we proposed here, the individual level latent class model-based approach is more flexible because it allows the transition probabilities to vary across individuals according to covariates, which gives insight into the differences in transitions between subgroups and may also result, depending on the validity of the model assumptions, in more accurate estimates at the aggregate level. Moreover, contrary to the macro-integration approach, it also provides estimates of the misclassification probabilities. However this model-based approach has the problem of being quite demanding: a national statistical institute would have to run these computationally intensive models every time before publishing new figures. Al-

though [11] shows that measurement error probabilities – under certain conditions – can be carried over in time, a macro integration approach could still be much faster and less labour intensive. In addition, the macro-integration approach is less restrictive in the sense that it does not impose a Markov property of the integrated time series of proportions that it produces. Moreover, as the inputs for this method are population estimates, the method does not need sources that can be linked at the unit level. The method is also easy to extend and implement. In our model we did not include reliability weights for the data sources. If reliability information of some sort is available, reliability weights should be included in the model. One of the possibilities would be to use the estimates of the misclassification probabilities from the latent class Markov model to define the reliability weights. Another nice feature of our method is the simple algorithm that is easy to implement and runs fast. For the iterative algorithm described in our paper for Model 4 the elapsed system time in R was 2.5 seconds when we required the precision of 1e-09 for the constraints.

We compared the results of [11] with the results of Model 4 using the pooled data. We observed that the proportions estimates were not very different. However, the estimates of the transition rates did differ significantly. The most important difference between the findings of the two approaches is that the estimate of the transition rate from temporary to permanent employment obtained from the macro-integration method lies in between the observed transition rates of the LFS and the ER, while the hidden Markov Models find that a transition probability is much lower than both observed transition rates. The lower off diagonal transition rates estimates of the HMM could be explained by the fact that these transitions are to some extend interpreted as measurement errors.

Macro integration methods and hidden Markov models both can be used to produce consistent estimates. In this paper we specified several macro-integration models for longitudinal data on employment status. Comparing the numerical results with those of a previous application of the hidden Markov model on the same data gives some insights into the differences of both approaches. A thorough methodological study of the properties or assumptions of both approaches that explain these differences could be a subject of future research.

References

[1] Bakker BFM. Micro integration: State of the art. ESSnet on Data Integration Report of WP1. State of the art on statistical methodologies for data integration, Eurostat; 2011.

[2] Zhang LC. Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica. 2012; 66(1): 41-63.

[3] Fosen J, Zhang LC. Quality evaluation of employment status in register-based census. In: Int Statistical Inst: Proc 58th World Statistical Congress, 2011, Dublin (Session STS024); 2011. Available from: http://2011.isiproceedings.org/papers/650264.pdf.

[4] Guarnera U, Varriale R. Estimation from Contaminated Multi-Source Data Based on Latent Class Models. ISTAT; 2015. Available from https://ec.europa.eu/eurostat/cros/system/files/Guarnera-etal_abstract_NTTS_2015_Guarnera_Varriale.pdf.

[5] de Waal T, van Delden A, Scholtus S. Multi-source Statistics: Basic Situations and Methods. International Statistical Review. 2019; Accepted for publication.

[6] Bakker BFM, Daas P. Some Methodological Issues of Register Based Research. Statistica Neerlandica. 2012; 66: 2-7.

[7] Goni E, Serrano A, Aramendi J. Use of administrative data and data reconciliation in the Basque labour force survey. EUSTAT – Basque Statistics Office; 2019. Available from: https://ec.europa.eu/eurostat/cros/system/files/use-of-admin-data-basque-lfs.pdf.

[8] Fosen J. Register-based employment statistics. A case of microintegration. The micro-integration process and life-cycle from a quality perspective. Statistics Norway; 2011. Available from: https://ec.europa.eu/eurostat/cros/system/files/ESSnet%20DI%20Wp4.1%20Micro%20integration%20and%20life%20cycle.pdf.

[9] Dagum EB, Cholette PA. Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series. Lecture Notes in Statistics. New York: Springer-Verlag, 186; 2006.

[10] Mushkudiani N, Daalmans J, Bikker R. Solving large-data consistency problems at Statistics Netherlands using macro-integration techniques. Statistica Neerlandica. 2018; 72: 553-73.

[11] Pankowska PKP, Bakker BFM, Oberski DL, Pavlopoulos D. Reconciliation of inconsistent data sources by correction for measurement error: The feasibility of parameter re-use. Statistical Journal of the IAOS. 2018; 34(3): 317-29. doi: 10.3233/SJI-170368.

[12] Cobben F. Nonresponse in sample surveys: methods for analysis and adjustment. Statistics Netherlands. 2009.

[13] Bikker R, Daalmans J, Mushkudiani N. Benchmarking large accounting frameworks: a generalized multivariate model. Economic Systems Research. 2013.

[14] Daalmans J. Divide-and-Conquer solutions for estimating large consistent table sets. Statistical Journal of the IAOS. 2017.

[15] Pannekoek J, Zhang LC. Optimal adjustments for inconsistency in imputed data. Survey Methodology. 2015.

[16] Pavlopoulos D, Vermunt JK. Measuring temporary employment. Do survey or register data tell the truth? Survey Methodology. 2015; 41(1): 197-214.

[17] Censor Y, Zenios SA. Parallel Optimization. Theory, Algorithms, and Applications. New York: Oxford University Press; 1997.

[18] Boyd S, Vandenberghe L. Convex Optimization. New York: Cambridge University Press; 2004.

[19] Luenberger DG. Linear and Nonlinear Programming. Addison-Wesley, Reading; 1984.

Appendix A: Solutions to the optimisation problems

A.1. General solution

The objective and the constraints can be written more compactly in matrix notation, a formulation that allows for a concise description of the general optimisation problem, the structure of its solution and the algorithms involved. This formulation is also necessary to apply existing software for solving this optimisation problem.

First we collect all observed proportions, from both the register and the survey in one long vector $\mathbf{p}$ by the following steps. The observed register proportions in a four-way table, for a given t, can be written in vector form as an 81-vector given by (last index running fastest)

$$\vec{\mathbf{p}}^{R.t} = p^{R.t,t+1,t+2,t+3}_{1,1,1,1}, p^{R.t,t+1,t+2,t+3}_{1,1,1,2}, p^{R.t,t+1,t+2,t+3}_{1,1,1,3}, p^{R.t,t+1,t+2,t+3}_{1,1,2,1}, \ldots, p^{R.t,t+1,t+2,t+3}_{3,3,3,3}.$$

The observed register proportions for the four-way tables for all time points together are denoted by $\vec{\mathbf{p}}^R$, and obtained as the concatenation of the $\vec{\mathbf{p}}^{R.t}$ for $t \in (1, 4, 7, 10)$: $\vec{\mathbf{p}}^R = (\vec{\mathbf{p}}^{R.1}, \vec{\mathbf{p}}^{R.4}, \vec{\mathbf{p}}^{R.7}, \vec{\mathbf{p}}^{R.10})$.

Similarly, we can re-arrange the $\mathbf{p}^{S.t,t+3}$ in vector form as a concatenation of the vectors collecting the proportions in the four bivariate (3×3) tables $\vec{\mathbf{p}}^S = (\vec{\mathbf{p}}^{S.1}, \vec{\mathbf{p}}^{S.4}, \vec{\mathbf{p}}^{S.7}, \vec{\mathbf{p}}^{S.10})$. Note that the vectors $\vec{\mathbf{p}}^R$ and $\vec{\mathbf{p}}^S$ and the corresponding four-way and two-way tables only pertain to month 3 to 13, for the last two month there is only a single source (the register) and therefore no reconcilliation problem.

The complete vector of observed proportions, from both the register and the survey, can now be written as $\mathbf{p} = (\vec{\mathbf{p}}^R, \vec{\mathbf{p}}^S)$ with length $360 = 4 \times 81 + 4 \times 9$. The corresponding vectors with reconciled proportions will be denoted by $\tilde{\boldsymbol{\pi}}$,$\tilde{\vec{\boldsymbol{\pi}}}^R$ and $\tilde{\vec{\boldsymbol{\pi}}}^S$, respectively. The elements of $\mathbf{p}$ and $\tilde{\boldsymbol{\pi}}$ will be referred to as p_v and $\tilde{\pi}_v$, for $v = 1 \ldots 360$.

All constraints are linear equalities in sums of the components of $\tilde{\boldsymbol{\pi}}$ and can therefore be written in the form $\mathbf{a}^T\tilde{\boldsymbol{\pi}} = \mathbf{b}$. Here $\mathbf{a}$ is a vector defining the constraint. For the reconciliation models considered here, we distinguish two kinds of constraints, one that sets the reconciled bivariate proportions $\tilde{\boldsymbol{\pi}}^{t,t+3}$ equal to the corresponding survey proportions but allows the register proportions to be adjusted (*Fixed sums*) and one that allows both the survey and the register proportions to be adjusted in order to satisfy equality constraints (*Equalities*).

(i) *Fixed sums*. These are constraints that set sums of parameters equal to fixed constants (constraints Eqs (2) and (3a)). Then the vector $\mathbf{a}$ consists of 1's corresponding to the elements of $\tilde{\boldsymbol{\pi}}$ in the constraint and 0's for the other elements. The corresponding element of $\mathbf{b}$ contains the fixed constant.

(ii) *Equality*. These are constraints that define equalities between sums of parameters (constraints Eqs (3b)–(3e)). We move all elements from right side of equation to left side. Then the elements of $\mathbf{a}$ are 1's corresponding to the elements of $\tilde{\boldsymbol{\pi}}$ on the left side of the equality sign, -1's for the elements that were on the right side of the equality sign and 0's for all other elements of $\tilde{\boldsymbol{\pi}}$. The corresponding element of $\mathbf{b}$ is zero in this case.

To refer to all constraints simultaneously, we define the matrix $\mathbf{A}$ with rows $\mathbf{a}_k$ defining the K constraints.

The objective function can now be expressed as

$$D(\mathbf{p}, \tilde{\boldsymbol{\pi}}) = D^R(\vec{\mathbf{p}}^R, \tilde{\vec{\boldsymbol{\pi}}}^R) + D^S(\vec{\mathbf{p}}^S, \tilde{\vec{\boldsymbol{\pi}}}^S). \tag{17}$$

and the reconciled proportions are the solution to the optimisation problem

$$\begin{aligned} &\tilde{\boldsymbol{\pi}} = \underset{\tilde{\boldsymbol{\pi}}}{\operatorname{argmin}}\, D(\tilde{\boldsymbol{\pi}}, \mathbf{p}) \\ &\text{subject to } \mathbf{A}\tilde{\boldsymbol{\pi}} = \mathbf{b}, \end{aligned} \tag{18}$$

with the different reconciliation models defined by different specifications of constraints (rows of $\mathbf{a}$) and objective function D.

The optimisation problem associated with the reconciliation models described in Section 3 is a convex optimisation problem with linear constraints for which several algorithms can be used. For the models (1 and 3) that are based on the (weighted) least squares objective, analytic solutions exists. For models based on the Kullback-Leibler divergence D_{KL} no such closed form solutions exists in general, but for the separable model 2 with fixed constraints an analytic solution does exists. For model 4 an iterative algorithm that can be seen as a generalisation of iterative proportional fitting (IPF) is particularly convenient. These solutions will be described below.

In general, the solution to Eq. (18) can be obtained by using the Lagrangean function for this problem, which can be expressed as:

$$L(\tilde{\boldsymbol{\pi}}, \boldsymbol{\lambda}) = D(\tilde{\boldsymbol{\pi}}, \mathbf{p}) + \boldsymbol{\lambda}^T(\mathbf{A}\tilde{\boldsymbol{\pi}} - \mathbf{b}), \tag{19}$$

with $\boldsymbol{\lambda}$ the K-vector containing the Lagrange multipliers for the K-constraints. The constraint optimum of Eq. (18) is a stationary point of the Lagrangean Eq. (19), which can be found by equating the partial derivatives of Eq. (19) with respect to $\tilde{\boldsymbol{\pi}}$ and $\boldsymbol{\lambda}$ to zero (see, e.g. [18, Ch. 5], [19, Ch. 10]). Thus, we obtain the following equations:

$$\partial L(\tilde{\boldsymbol{\pi}}, \boldsymbol{\lambda})/\partial\tilde{\boldsymbol{\pi}} = \partial D(\tilde{\boldsymbol{\pi}}, \mathbf{p})/\partial\tilde{\boldsymbol{\pi}} + \mathbf{A}^T\boldsymbol{\lambda} = \mathbf{0} \tag{20a}$$

$$\partial L(\tilde{\boldsymbol{\pi}}, \boldsymbol{\lambda})/\partial\boldsymbol{\lambda} = \mathbf{A}\tilde{\boldsymbol{\pi}} - \mathbf{b} = \mathbf{0} \tag{20b}$$

A.2. Quadratic-loss, additive adjustments

If D is the quadratic-loss $\frac{1}{2}(\tilde{\boldsymbol{\pi}} - \mathbf{p})^T(\tilde{\boldsymbol{\pi}} - \mathbf{p})$, we have $\partial D(\tilde{\boldsymbol{\pi}}, \mathbf{p})/\partial\tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}} - \mathbf{p}$ so that Eq. (20a) becomes $\tilde{\boldsymbol{\pi}} = \mathbf{p} - \mathbf{A}^T\boldsymbol{\lambda}$ which by substitution in Eq. (20b) results in $\boldsymbol{\lambda} = (\mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{p} - \mathbf{b})$ and so the reconciled proportions are given by:

$$\tilde{\boldsymbol{\pi}} = \mathbf{p} + \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{b} - \mathbf{A}\mathbf{p}). \tag{21}$$

For the fixed-sum constraints we have $9 \times 4 = 36$ constraints defined by Eq. (2). These constraints are separable in the sense that they each pertain to a different part of $\tilde{\boldsymbol{\pi}}$, they have no $\tilde{\pi}$-parameters in common. The simultaneous equation Eq. (21) can then be written as a the set of separate equations:

$$\tilde{\boldsymbol{\pi}} = \mathbf{p} + \mathbf{a}_k\frac{1}{n_k}(b_k - \mathbf{a}_k^T\mathbf{p}), \tag{22}$$

where the $\mathbf{a}_k$ define the fixed-sum type constraints and $n_k = \mathbf{a}_k^T\mathbf{a}_k$ is the number of register values in constraint k. This equation means that reconciled proportions are obtained by adding the difference of the sum of the observed register values in constraint k and the corresponding survey proportion in b_k, divided by n_k, to each of these register proportions. This results in reconciled proportions that add-up to the corresponding survey proportion. For the equality constraints separability does not hold as can be seen from Eq. (3) and the solution for $\tilde{\boldsymbol{\pi}}$ must be obtained by the simultaneous Eq. (21).

A.3. KL-loss, multiplicative adjustments

If D is the KL-loss function, no explicit solution is available for the equality constraints but a convenient iterative algorithm will be presented below. For the case of fixed-sum constraints an explicit solution is available and presented before in Eq. (9). The iterative algorithm proceeds in a step-by-step manner as follows. It starts by minimising the objective subject to one of the constraints only. In the next step the resulting approximate solution is updated such that a next constraint is satisfied and the difference with the previous approximation is minimised. When all constraints are visited, the first iteration is completed and the next iteration starts that will again sequentially adjust the current approximation to satisfy each of the constraints. The minimisation carried out in each step solves the problem:

$$\begin{aligned} &\tilde{\boldsymbol{\pi}}^{r,k} = \underset{\tilde{\boldsymbol{\pi}}^{r,k}}{\operatorname{argmin}}\, D(\boldsymbol{\pi}^{r,k}, \tilde{\boldsymbol{\pi}}^{r,k-1}) \\ &\text{subject to } \mathbf{a}_k^T\tilde{\boldsymbol{\pi}}^{r,k} = b_k, \end{aligned} \tag{23}$$

with r indicating the iterations. This "succesive projection algorithm" is known to converge for convex objectives and linear (in)equality constraints, see e.g. [17].

To solve the minimisation problem Eq. (23) we set up the Lagrangean function for this problem, which can be expressed as

$$L(\tilde{\boldsymbol{\pi}}^{r,k}, \lambda_k^r) = D(\boldsymbol{\pi}^{r,k}, \tilde{\boldsymbol{\pi}}^{r,k-1}) + \lambda_k^r(\mathbf{a}_k^T\tilde{\boldsymbol{\pi}}^{r,k} - b_k), \tag{24}$$

with λ_k^r the Lagrange multiplier for constraint k in iteration r. Equating the partial derivatives of Eq. (24) with respect to $\tilde{\boldsymbol{\pi}}^{r,k}$ and λ_k^r to zero we obtain the following equations:

$$\partial L(\tilde{\boldsymbol{\pi}}^{r,k}, \lambda_k^r)/\partial \tilde{\boldsymbol{\pi}}^{r,k} = \partial D(\boldsymbol{\pi}^{r,k}, \tilde{\boldsymbol{\pi}}^{r,k-1})/\partial(\tilde{\boldsymbol{\pi}}^{r,k}) + \lambda_k^r \mathbf{a}_k = \mathbf{0} \tag{25a}$$

$$\partial L(\tilde{\boldsymbol{\pi}}^{r,k}\lambda_k^r)/\partial \lambda_k^r = \mathbf{a}_k^T \tilde{\boldsymbol{\pi}}^{r,k} - b_k = 0, \tag{25b}$$

If D is the KL-divergence, $\sum_v \tilde{\pi}_v^{r,k}(\log(\tilde{\pi}_v^{r,k}/\tilde{\pi}_v^{r,k-1}) - 1)$, we have from Eq. (25a):

$$\log \tilde{\pi}_v^{r,k} - \log \tilde{\pi}_v^{r,k-1} + \lambda_k^r a_{k,v} = 0$$

and so

$$\tilde{\pi}_v^{r,k} = \tilde{\pi}_v^{r,k-1} / \exp(a_{k,v}\lambda_k^r), \tag{26}$$

which by substitution in Eq. (25b) results in

$$\sum_v a_{k,v} \tilde{\pi}_v^{r,k-1} / \exp(a_{k,v}\lambda_k^r) - b_k = 0. \tag{27}$$

Since, as noted in Section A.1, the elements of $\mathbf{a}_k$ are 0, 1 or -1, we can re-express Eq. (27) as:

$$\frac{1}{\tau_k^r} \sum_{v \in k+} \tilde{\pi}_v^{r,k-1} - \tau_k^r \sum_{v \in k-} \tilde{\pi}_v^{r,k-1} - b_k = 0, \tag{28}$$

with $\tau_k^r = \exp(\lambda_k^r)$ and $k+$ and $k-$ the sets of indices v for which $a_{k,v} = 1$ and $a_{k,v} = -1$, respectively.

For the two types of constraints we are considering here, we now have:

i *Fixed sums*. In this case, with the $a_{k,v}$ equal to 1 or 0, Eq. (28) reduces to $\frac{1}{\tau_k^r} \sum_{v \in k+} \tilde{\pi}_v^{r,k-1} = b_k$, and so

$$\tau_k^r = \frac{\sum_{v \in k+} \tilde{\pi}_v^{r,k-1}}{b_k}$$

and from Eq. (26) we obtain for the update $\tilde{\pi}_v^{r,k}$

$$\tilde{\pi}_v^{r,k} = \tilde{\pi}_v^{r,k-1} \times \frac{b_k}{\sum_{v \in k+} \tilde{\pi}_v^{r,k-1}} \text{ for } v \in k+. \tag{29}$$

For a starting value $\tilde{\pi}_v^{r,k-1} = p_v$ we obtain

$$\tilde{\pi}_v^{k} = p_v \times \frac{b_k}{\sum_{v \in k+} p_v} \text{ for } v \in k+, \tag{30}$$

which gives the solution without further iterations because the fixed sum constraints are separable and hence the sets $k+$ for the different constraints are non-overlapping so that Eq. (30) only applies to different sets of reconciled proportions for each k. This result coincides with the re-scaling expressed by Eq. (9) in Section 3.

ii *Equalities*. In this case, with $b_k = 0$, we obtain from Eq. (28)

$$\tau_k^r = \left(\frac{\sum_{v \in k+} \tilde{\pi}_v^{r,k-1}}{\sum_{v \in k-} \tilde{\pi}_v^{r,k-1}} \right)^{\frac{1}{2}}$$

and from Eq. (26) we obtain for the update $\tilde{\pi}_v^{r,k}$:

$$\tilde{\pi}_v^{r,k} = \tilde{\pi}_v^{r,k-1} / \tau_k^r \text{ for } v \in k+, \tag{31a}$$

$$\tilde{\pi}_v^{r,k} = \tilde{\pi}_v^{r,k-1} \times \tau_k^r \text{ for } v \in k-. \tag{31b}$$

A.4. Algorithm

The algorithm to solve the reconciliation problem with D_{KL} can now proceed according to the following steps:

1. Create the constraint vectors $\mathbf{a}_k$, for $k \ldots K$. This also defines the sets $k-$ and $k+$ for each k.
2. For fixed-sum constraints set b_k equal to the fixed survey proportions.
3. Initialise the starting value $\tilde{\pi}_v^{1,0} = p_v$
4. For $k = 1$ to K use the update Eqs (30) or (31) to update $\tilde{\boldsymbol{\pi}}^{r,k}$ sequentially for each of the K constraints. This completes one iteration.
5. Repeat step 4 until some convergence criterion is met.