Reconciliation of inconsistent data sources by correction for measurement error: the feasibility of parameter re-use

Paulina Pankowska[1], Bart Bakker[2,3], Daniel L. Oberski[4] & Dimitris Pavlopoulos[2]

Abstract: National Statistical Institutes (NSIs) often obtain information about a single variable from separate data sources. Administrative registers and surveys, in particular, often provide overlapping information on a range of phenomena of interest to official statistics. However, even though the two sources overlap, they both contain measurement error that prevents identical units from yielding identical values. Reconciling such separate data sources and providing accurate statistics, which is an important challenge for NSIs, is typically achieved through macro-integration. In this study we investigate the feasibility of an alternative method based on the application of previously obtained results from a recently introduced extension of the Hidden Markov Model (HMM) to newer data. The method allows a reconciliation of separate error-prone data sources without having to repeat the full HMM analysis, provided the estimated measurement error processes are stable over time. As we find that these processes are indeed stable over time, the proposed method can be used effectively for macro-integration, to reconciliate both first-order statistics – e.g. the size of temporary employment in the Netherlands – and second-order statistics – e.g. the amount of mobility from temporary to permanent employment.

Keywords: hidden Markov model, register data, survey data, data quality, labour market transitions, measurement error, administrative data

[1] Corresponding author. Department of Sociology, Faculty of Social Sciences, Vrije Universiteit Amsterdam, de Boelelaan 1105, 1081 HV Amsterdam. Tel.: +31 20 59 83178. E-mail: p.k.p.pankowska@vu.nl

[2] Vrije Universiteit Amsterdam

[3] Statistics Netherlands

[4] University of Utrecht

1. Introduction

National Statistical Institutes (NSIs) often obtain information about the same phenomenon from different data sources [1]. For example, the Dutch Labour Force Survey administered by Statistics Netherlands includes data that overlap to some extent with register data from the Dutch social security administration. The overlapping component of these datasets can be linked at the individual level. Such linked survey-register data often concern longitudinal measures of categorical variables such as employment, housing, and education, and are subject to editing procedures to improve data quality [2, pp. 10-13,3]. However, even then, identical units do not always yield identical values [4].

Two types of error may account for these discrepancies: measurement error and linkage error. However, linkage error, while an important error source in official statistics generally, is less of a concern for Statistics Netherlands due to the use of unique resident identifier numbers. We will therefore focus on the problem of measurement error[i].

Measurement error in surveys is a well-known and extensively studied phenomenon [5,6]. Measurement error in administrative registers, by contrast, has only recently attracted attention [2,7–10]. Such errors occur because registers result from data collection of public administration and are not originally intended for social-scientific research. When it occurs during data-entry, measurement errors in administrative registers mirror familiar survey response errors; however, errors unique to registers also occur, including administrative delay, definition error, and errors caused by administrative incentives [11–13].

Where measurement error is random or “classical”, the resulting data will not tend to bias “first-order” population estimates such as means, proportions, and totals [14]. However, “second-order” estimates, such as domain mean differences, hazard ratios, and transition rates over time, are well-known to be severely biased by random measurement error [15–18]. This bias may refer to an overestimation or an underestimation of these statistics.

For example, an important issue in labour market policy is the proportion of workers who change from employment with a flexible contract to employment with a permanent contract. If there is random measurement error in the type of employment contract, these transition rates are artificially (and severely) inflated as every misclassification in the contract type may lead to two errors in the measurement of transitions [19].

On the other hand, if errors are carried over between time points, the observed transitions rates are artificially dampened, as some real changes are not observed. Considering that the source and the type of the measurement error differs between data sources, the problem faced by NSIs is not only that different data sources yield different statistics, but also that measurement error may bias statistics in a different way in each of these sources.

There are several methods dealing with these differences in NSIs. Most commonly, the differences are ignored and only estimates from the source assumed to have the best quality are published. Another way is to assume that the quality of both sources is similar and take the mean of the estimates. However, a more advanced way of dealing with these differences is to apply macro-integration techniques. One of the usual integration strategies is that in the first step the stock data of two reference dates are integrated. In this step, the concepts, classifications and reference dates are harmonized, the data are completed by weighting or imputation if the data do not cover the entire target population, and, in order to minimize measurement error, the data are forced to meet identity relations defined beforehand. In the second step, data on the events between the two reference dates are made consistent by making use of the identity relations that the stock at reference date t plus all the changes add up to the stock at reference date $t + 1$. However, for the second step, only the source that is assumed to be of superior quality is used. In this second macro-integration step, one can try to preserve the original transitions in the (sub)populations as much as possible.

An alternative strategy to deal with this problem of inconsistency was recently introduced by Bakker [7] for continuous cross-sectional data, by Oberski et al. [8] for mixed type cross-sectional data, and by Pavlopoulos and Vermunt [20] for categorical longitudinal data. In this *latent variable modeling* approach, the reconciliation and measurement error problems are solved simultaneously by modeling the two sources as conditionally independent measures of an underlying true value. In the cross-sectional models, this true value is related to other, similar, true values. Since repeated observations of a single linked survey-register variable may be more common in practice, we focus on the case of longitudinal data. In these models, the true value is related to itself over time in an autoregressive process that yields an extended – multiple data source – version of the Hidden Markov Model popularized by Biemer [21,22], which in turn is a special case of the Latent Class Model [23,24]. Previous work done at Statistics Netherlands used such models to integrate data from Labour Force Survey and social security administration [20].

A problem with this procedure is that it is very time consuming and therefore expensive, since it requires the NSI to perform linkage between register and survey followed by re-estimation of the model for each new time period. This paper therefore considers the option of re-using existing parameter estimates from the above study in order to integrate data sources and correct statistics for measurement error. Re-use is potentially attractive because (1) it does not require re-estimation of the model, and (2) it can be applied not only to linked survey-register data, but also to each data source separately, forgoing the need for a time-intensive linkage exercise.

However, parameter re-use can only be applied to regular production at NSIs if the parameters of the model remain the same over time. If the parameter estimates do not exhibit stability over time, the corrections themselves will be biased. Therefore, an important question for the practical application of latent variable modeling at NSIs is whether there is indeed stability in the estimates when applying this procedure to real data. In this paper, we demonstrate how this question can be investigated using newly collected data on a topic studied previously and for earlier years by Pavlopoulos and Vermunt [20]. In other words, our analysis allows us to determine whether the aforementioned time- and cost-efficient methodology (based on using previously obtained parameters) can actually work in practice.

The next section describes the data used in the analysis; this is followed by a discussion of the empirical methodology, the results, and finally a brief conclusion.

2. Data

The dataset used for the analysis contains information from the Netherlands' Labour Force Survey that is conducted by Statistics Netherlands and the *'Polisadministratie'* (administrative data collected by the Employee Insurance Agency).

The Dutch Labour Force Survey (LFS) is a sample survey aimed at providing information about the relationship between individuals and the labour market. The target population consists of individuals aged 15 and older who reside in the Netherlands (excluding those in homes and institutions) and the information is collected at both the individual and household level.[5] Since the last quarter of 1999 the survey has been a rotating panel survey, consisting of five waves.

The Employment Register data (i.e. the *'Polisadministratie'* or ER) is an administrative dataset administered by the Dutch Employee Insurance Agency (EIA, or *UWV* in Dutch). The dataset contains monthly information on wages, benefits, and labour relations for all insured employees in the Netherlands. EIA uses the information collected to determine the level of benefits. The dataset combines information from various sources; the core of the information is delivered by the employers on their employees each month for tax purposes to the Dutch

5 http://www.cbs.nl/en-GB/menu/methoden/dataverzameling/dutch-labour-force-survey-characteristics.htm

Tax Authorities, information from temporary work agencies and the Population Register (PR, in Dutch: *Basis Registratie Personen*- BRP)[6] is also used.

The data from both sources were linked at the individual level to the population register (PR) of the Netherlands. Therefore, the target population is restricted to the registered population in the Netherlands. For the linkage of the LFS with the PR, the linkage key is the combination of birth date, gender, postal code and house number. The ER is linked to the PR based on the social security number (BSN)[7], birth date, gender, postal code and house number. After selection of the individuals aged 25-55, the linkage effectiveness of the combined sources is approximately 97%.

The sample used for the analysis consists of 8,886 LFS respondents aged between 25 and 55 who participated in the LFS for the first time in the first trimester of 2009. For each individual included in the sample, the dataset contains information for a period of 15 months with the variables coming from the ER data observed on a monthly basis (i.e. 15 observations) and those from the LFS observed every 3 months (i.e. 5 observations). The time period the data correspond to, January 2009 to May 2010, is illustrated in Fig. 1 (with the time period from January 2009 to March 2010 corresponding to those individuals first interviewed in January 2009; those from February 2009 to April 2010 to those firstly interviewed in February 2009; and those from March 2009 to May 2010 to those firstly interviewed in March 2009).

[Insert Fig. 1 here]

The panel dataset is unbalanced for the LFS as it suffers from attrition. More specifically, 8,708 individuals participated in the first round of the survey, 7,458 in the second, 6,856 in the third, 6,739 in the fourth and 6,560 in the fifth. For the non-survey months observations are assumed to be missing at random. While the ER officially cannot be subject to

[6] http://www.uwv.nl/overuwv/english/about-us-executive-board-organization/detail/organization/data-services

[7] A unique personal number allocated to everyone registered in the Netherland; https://www.government.nl/topics/identification-documents/contents/the-citizen-service-number)

drop-out as submission of reports is obligatory for all employers, 2,619 observations (out of a total of 133,290) are missing. We assume that these missing observations are also missing at random[8]

The key variable of interest in the analysis is the contract type held by the individual for his or her main job (any secondary jobs are ignored in the analysis). The contract type can take on three distinctive and mutually exclusive values: 'permanent contract' (i.e. a contract for an unlimited duration of time), 'temporary contract' (i.e. a fixed contract for a limited duration of time) and 'other' (which includes all other alternatives, i.e. self-employment, unemployment, unpaid employment and full-time education). While the last category is not of crucial importance for answering the research questions themselves (as the focus is on transition rates from temporary to permanent employment) it has to be included in the analysis to assure that the Markov assumption of mutual exclusivity and exhaustiveness of the latent classes is fulfilled.

The distributions of the contract types according to both data sources are displayed in table 1. As can be observed, the aforementioned distributions as indicated by the survey and register data accordingly differ to a larger extent for permanent and temporary contracts than other types of contracts.

[Insert Table 1 here]

In order to gain more insight into the extent of the aforementioned inconsistencies, the contract type according to both datasets has been cross- tabulated for the entire sample. The results, presented in table 2, show that while the discrepancies between the survey and register data concerning individuals who hold a permanent contract or occupy the state 'other' are relatively small, those regarding individuals employed on a temporary contract are highly substantial.

[8] Those are primarily observations of workers who have passed away or emigrated from the Netherlands and, thus, there is no reason to believe their missingness is related to the variable of interest

[Insert Table 2 here]

The disparities between the two datasets with regards to the contract type (and in particular to temporary contracts) outlined above have implications for the estimation of the transition rates between the different contract types. Namely, as depicted in table 3, the transition rate from temporary employment in month *t-3* to permanent employment in month *t* equals 5.8% according to the survey data while it amounts to 7.3% according to the register data.

[Insert Table 3 here]

3. Methods

3.1. Classification error model for survey and register

The methodology applied in this paper is based on the extended Hidden Markov Model used by Pavlopoulos and Vermunt [20]. The standard Hidden Markov models discussed by Biemer [22] assumes that an observed categorical variable Y_t is generated in the following way:

- At $t = 0$,
  - Sample a "true value" x_0 from the unknown distribution $p(X_0)$,
  - Sample the observed value y_0 from the unknown conditional distribution $p(Y_t \mid X_t)$. The off-diagonal entries in this unobserved cross-table are the misclassification rates and the diagonal entries the probability of a correct classification.
- At $t > 0$,
  - Sample a "true value" x_t from the unknown distribution $p(X_t \mid X_{t-1})$. The unobserved cross-table between X_t and X_{t-1} contains the unobserved

transition rates of substantive interest. In our example, the parameter $p(X_t \; = \; permanent \,|\, X_{t-1} \; = \; temporary)$ is specifically of interest,

- As before, sample the observed value y_t from the unknown conditional distribution $p(Y_t \,|\, X_t)$.

- Advance one step in discrete time by setting $t \leftarrow t + 1$ until the maximum number of observed time points $t = T$ is reached.

The unknown parameters of this model are those describing the initial state distribution $p(X_0)$, the misclassification rates $p(Y_t \,|\, X_t)$, and the AR(1) autoregressive transition rates $p(X_t \,|\, X_{t-1})$. These parameters are identifiable by assuming equal misclassification and transition rates over time, i.e. $p(Y_t \,|\, X_t) = p(Y_{t'} \,|\, X_{t'})$ and $p(X_t \,|\, X_{t-1}) = p(X_{t'} \,|\, X_{t'-1})$ for all $t \neq t'$. Since only the joint distribution of the observed variables $p(Y_{t0}, Y_{t1}, \dots, Y_t, Y_T)$ is observed and X_t is entirely missing, estimation of the unknown parameters often proceeds by marginal maximum likelihood, expectation-maximization, or Markov Chain Monte Carlo methods. In what follows we employ the Latent GOLD software, which uses a combination of expectation-maximization and marginal maximum likelihood estimation. It is also straightforward to implement covariates affecting the distribution of *X*; for the sake of clarity we have omitted these in the description but do include them in our extended model.

The standard Hidden Markov model has the substantial disadvantage that it makes the assumption of conditional independence of errors, sometimes also referred to as the "independent classification errors" or ICE assumption. In other words, it assumes that when y_t was generated, its probability of occurring only depended on x_t and nothing else. This precludes, for example, the possibility that any errors that occurred at the previous time point were copied over to the current time point, since that would make the observed value dependent on both the true value and the observed value at the previous time point, i.e. $p(Y_t \,|\, X_t) \neq p(Y_t \,|\, X_t, X_{t-1}, Y_{t-1})$. Since there are considerable indications that register errors are copied over time, the standard Hidden Markov model is inappropriate.

As mentioned before, in this paper, we follow Pavlopoulos and Vermunt (2015) in employing an extension to the standard HMM that allows for error-copying over time in the register. The parameters of this model are identified by linking the register to a survey measuring the same true value over time, in addition to assuming parameters are equal over time. A graphical illustration of the model for the first 4 months is given in Fig. 2

[Inset Fig. 2 here]

The extended HMM assumes that the observed register values $Y_t^{(r)}$ and survey values $Y_t^{(s)}$ were generated as follows:

- At $t = 0$,
  - Sample a "true value" x_0 from the unknown distribution $p(X_0)$,
  - Sample the observed register value $y^{(r)}_0$ from the unknown conditional distribution $p(Y^{(r)} \mid X)$,
  - Sample the observed survey value $y^{(s)}_0$ from the unknown conditional distribution $p(Y^{(s)} \mid X)$.
- At $t > 0$,
  - Sample a "true value" x_t from the unknown distribution $p(X_t \mid X_{t-1})$,
  - Sample the observed survey value $y_t^{(s)}$ from the unknown conditional distribution $p(Y^{(s)} \mid X)$,
  - If the register at the previous time point had an error and no change in true value occurred, i.e. if $x_{t-1} \neq y_{t-1}^{(r)}$, and $x_{t-1} = x_t$ ("*previous error and no change*"),
    - Sample the observed register value $y_t^{(r)}$ from the unknown distribution $p(Y_t^{(r)} \mid \text{previous error and no change})$. This distribution contains the probability of copying an error when no change occurred in the true value, $p(Y_t^{(r)} = y_{t-1} \mid \text{previous error and no change})$,

  - Else if $x_{t-1} = y_{t-1}$ or $x_{t-1} \neq x_t$ (there was no error, or true change occurred),
    - Sample the observed register value $y_t^{(r)}$ from the unknown conditional distribution $p(Y^{(r)} \mid X)$.
- Advance one step in discrete time by setting $t \leftarrow t + 1$ until the maximum number of observed time points $t = T$ is reached.

Again, covariates Z are easily included by extending $p(X \mid .)$ to $p(X \mid ., Z)$, where "." may indicate a set of random variables. In our model, this set of covariates always includes the timepoint to allow for variation over time in the transition probabilities. To control for unobserved heterogeneity in the transition probabilities, we further extend $p(X \mid ., Z)$ to $p(X \mid ., Z, k)$, where k denotes the latent class that the individual belongs to.

In addition to the output from the HMM, which also provides estimates of the transition rates and misclassification rates, the extended HMM also provides estimates of the error-copying rates. Moreover, the misclassification rates estimated for the register are conditional on no error having occurred previously. Since this cannot be known in practice, we will report both these estimates, and the overall error rates that average over previously occurring errors and correct reports.

The extended model allows for error-copying over time and therefore relaxes the ICE assumption. However, it does this by introducing the assumption that the survey and register values are conditionally independent, given the true value. In what follows we will evaluate the fit of these models before turning to interpretation.

4. Results

We first apply the extended HMM described above to the data from 2009; then, we repeat the analysis for the same cohort while fixing the measurement error specific parameters to those obtained by Pavlopoulos and Vermunt [20] when analysing data from 2007. The results of the two analyses are then compared to verify whether it is possible to correct for measurement error in data sources over the course of several years while only applying the full extended HMM analysis once at the initial stage.

4.1.Model fit

To assure that the model specification used by Pavlopoulos and Vermunt [20] fits more recent data equally well, we estimated a total of nine different specifications of the Hidden Latent Markov Model. Those specifications were also estimated by Pavlopoulos and Vermunt [20] to reach the final version of the model. The goodness-of-fit measures of those models are summarized in table 4. In more detail, the table includes the following information: the log-likelihood, the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC) values as well as the number of model parameters.

The first three models used *(A', A" and A)* assumed, respectively, that only the survey data, only the register data and both datasets are subject to independent classification errors (ICE). The fact that the last of the three models fits the data best provides support for the hypothesis that both data sources contain measurement error. As such, the subsequent six models are extensions of the model assuming the presence of classification errors in both the survey and register data.

The next three models estimated *(B', B" and B)* relax the ICE independence assumption of the measurement error for the survey, the register and both data sources, respectively. In the survey data this is related to the fact that the likelihood of making an error often varies according to age and proxy interview [26–28].

The serial correlation of the measurement error in the register data is likely to result from the fact that companies submit information – including the contract type - to the Employment Office once or twice a year. This is likely to result in errors being carried over until an actual change in the contract type occurs or until some form of data quality control takes place [12,13,29,30]. Therefore, the probabilities of having an error in the register data are modelled in such a way that they depend on the lagged observed and lagged true contract.

As can be seen from table 4, models B" and B, which relax the ICE assumption only for the register data and for both datasets respectively, perform somewhat better than model B', which assumes that only the survey errors do not satisfy the local independence assumption. This means that it is realistic to conclude that the error is indeed serially correlated in the register data but not in the survey data. Therefore, the final set of models (C', C'' and C) extends those two models by including covariates for the latent transition and for the latent initial state probabilities and thus assumes that those transitions and probabilities are heterogeneous.

In more detail, model C' can be seen as a restricted extension of model B'' as it assumes that the measurement errors are not locally independent for the register data and that the latent transitions depend on gender, age, education and country of origin. Model C'' can be seen as a full extension of B'' as it also assumes that ICE does not hold for the register data but, in addition to the latent transitions, it also assumes that the aforementioned covariates influence the initial state probabilities. Finally, model C can be seen as a full extension of model B as it assumes that ICE should be relaxed for both data sources and that the covariates influence both the latent transitions and initial state probabilities. The covariates are allowed to be time heterogeneous.

As can be seen in table 4 models C'' and C appear to fit the data best. However, as the differences in the AIC and BIC between the two models are rather minimal and model C'' is slightly less complex, it has been selected as our final model. The results from the comparison of the model fit statistics are similar to those of Pavlopoulos and Vermunt [20] where model C'' was also selected as the final model. This confirms that for a certain period of time the same model specification can be used to correct for measurement error.

[Insert Table 4 here]

4.2. The size of the measurement error

The size of measurement error in the survey and register data according to our analysis and that of Pavlopoulos and Vermunt [20] is depicted in tables 5 and 6 respectively. In order to estimate the error, the posterior probabilities of having a specific type of latent contract in each month have been used; those were estimated for all individuals included in the sample using the hidden Markov model.

In more detail, the tables report the classification error probabilities which in equation (1) are represented by the probabilities $P(Y_{i0}^{(r)} = y_0^{(r)} | X_{i0} = x_0)$ and $P\left(Y_{it}^{(r)} = y_t^{(r)} \middle| X_{it} = x_t, X_{i(t-1)} = x_{t-1}, Y_{i(t-1)}^{(r)} = y_t^{(r)}\right)$, for the survey and register data, respectively.

Overall, both analyses produce very similar results and point to the same trends with regards to the level of measurement error indicating that the error is stable for this period of time. In other words, the analyses show that overall all three contract types are measured very accurately by the survey. The overall size of measurement error in the register data, on the other hand, appears very high especially for individuals holding a temporary contract.

[Insert Table 5 here]

[Insert Table 6 here]

Given that the error probabilities in the register data are assumed to be serially correlated – by estimating an additional coefficient when a classification error was made in time point *t-1* and this error can be repeated in time *t*, we can extract more information on the structure of this error by studying more closely the conditional error probabilities. Fig. 3 reports our estimates of the conditional probabilities of the error in the register data in time t for all 9 combinations of latent and observed state in time *t-1*. The 3 diagonal matrices represent the cases where no error in the register data was made in time *t-1*, while off diagonal matrices represent the different cases of measurement error in *t-1*. Fig. 3 gives a completely different picture than table 6. The diagonal matrices – which are by construction identical - indicate that when no classification error is made in *t-1*, the probability of an error in *t* is rather minimal.

[Insert Fig. 3 here]

The conditional error probabilities obtained by Pavlopoulos and Vermunt [20] are presented in table 7. The results are almost the same as the ones reported by us and presented in Fig. 3.

[Insert Table 7 here]

The left-hand side of table 8 extracts from Fig. 3 the probabilities that an error repetition is possible. All the error probabilities when an error repetition is possible are extremely high. The relevant probabilities from Pavlopoulos and Vermunt [20]), as presented on the right hand side of the table, are very similar to our results.

The last remaining situation to examine is to study the probability for a different classification error in time *t* when a classification error is made in *t-1*. These probabilities are presented in table 9 for the three latent states and for both our own data and those of Pavlopoulos and Vermunt [20]. All these probabilities are rather small and similar to those where no error is made in t-1. The probabilities are also almost identical between the two studies.

[Insert Table 8 here]

Thus, the estimates of the conditional probabilities of the measurement error in the register data show a clear picture. The large size of the error that was illustrated in table 6 is only due to the error in the initial registration of the contract type in the register. Once a mistaken value for the contract type is entered, then this will be carried over almost for sure for many months. However, if a correct entry is made, then the probability of an error in the subsequent months is very small.

[Insert Table 9 here]

Overall we can conclude that the nature and size of the measurement error in both the survey and register data appear very similar in 2007 (analysed by Pavlopoulos and Vermunt [20]) and in 2009 (as shown by us). The stability of the measurement error for this period of time enables us to apply the aforementioned error correction method in which we fix the error parameters according to the results obtained by Pavlopoulos and Vermunt [20] in the analysis of our own data from 2009. This in turn allows us to correct for measurement error without having to undertake the full HMM analysis. The accuracy of this method when estimating first- and second- order statistics is explored below.

4.3.First-order statistics: the size of temporary employment

The latent distribution of the contract types, approximated according to our analysis and when substituting in the measurement error specific parameters from Pavlopoulos and Vermunt [20], is presented in table 10 and is contrasted with the observed distributions of the contract type according to the survey and register data, respectively. As in the case of the estimation of the average size of the measurement error, this has been carried out by using the average posterior probabilities of individuals holding a certain type of latent contract.

As can be seen from the table, the results of our own analysis are almost identical to those using the fixed error parameters. Furthermore, the latent probability of belonging to a certain state always lies between the observed probabilities coming from the two data sources. Specifically, the latent probability of having a temporary contract equals approx. 13% for both analyses and is higher than is reported by the survey data while lower than reported by the register data (11.1% and 15.1%, respectively). The latent probability of being employed with a permanent contract (approx. 61%), is lower than suggested by the survey data (65.3%) while higher than suggested by the register data (58.5%). Finally, the latent probability of belonging to the 'other' state equals approximately 26% and lies also in between the figures estimated using the survey and register data (23.7% and 26.4%, respectively).

This conveys good news for official statistics. In the presence of measurement error in our data, a macro-integration of two data sources – even by using a crude measure such as the average of the two observed probabilities - can produce reliable results for the size of temporary employment.

[Insert Table 10 here]

4.4. Second-order statistics: the transition probabilities

Besides providing a reliable estimate of the size of temporary employment, the challenge for official statistics is to present a correct estimate of mobility from temporary employment. The dominant argument in the policy debate is that although temporary employment is inferior to permanent employment, it provides an effective stepping stone to permanent employment. For this argument to be true, mobility rates from temporary to permanent employment should be high. Table 11 presents the average latent transition probabilities between the various states associated with the three contract types. These transition probabilities have been calculated using model C'' in such a way that they refer to a 3-month period and are an average of the twelve 3-month periods that are included in the dataset.

When looking at the estimates presented in table 11, it can be once more noted that the two analyses provide almost identical results. Furthermore, when analysing the transition rates in combination with those presented in table 3 - i.e. the observed transition rates based on the survey and register data – it can be inferred that the latent transition rate from temporary to permanent employment is much lower than those estimated using both the survey and register data. That is, while according to the survey and register data out of all temporary employees in time *t-3*, 5.8% and 7.3% respectively obtain a permanent contract in time *t*, our analysis suggests that this is only true for 1.6 - 1.7% of all temporary workers.

These findings suggest that a simple macro-integration of the two data sources, which typically aims at the reconciliation of the distribution of the variable of interest in the two data sources at a given point in time, cannot provide reliable estimates of second-order statistics, namely mobility from temporary to permanent employment. These transition rates are overestimated by both the survey and the register data although these datasets contain a very different size and structure of measurement error. Therefore, macro-integration would possibly not lead to transition rates lower than both sources.

[Insert Table 11 here]

5. Conclusions

National Statistical Institutes often have more than one indicator available for the same variable. The development of register data means that, increasingly, information on survey respondents can also be traced at the administrative level. This offers new opportunities to NSIs as they can corroborate findings from one data source using the other. However, these opportunities present new challenges that ought to be addressed. As all data sources contain some measurement error, discrepancies emerge between the data sources in the measurement of a single variable even for the same individuals. Measurement error leads to bias in the estimates for first order statistics (estimates on one reference date) and second order statistics (estimates of transition rates between two reference dates).

Besides ignoring the problem or taking averages, these discrepancies are usually resolved by NSIs with the use of macro-integration. After separate integration of the stock data of two reference dates by harmonization, completion, and by forcing the data to meet certain identity relationships, on an aggregate level, a large part of the measurement error has been removed. However, the aggregate transitions rates are only corrected by the application of one identity relationship that the stock at reference date t plus all the transitions add up to the stock at t+1. For this second step, most of the time only one source is used, the one that is assumed to be of superior quality. In practice this means that the first order statistics usually are close to the real values. However, the adjustment in the transitions is marginal when using only one identity relationship and only one source. Therefore, one can expect that the real transition rates, the second order statistics, differ more from the observed ones because not all measurement error could be removed.

In this paper, we study whether an alternative macro-integration method of the two datasets can produce more accurate results. In doing so we rely on the micro-integration approach undertaken by Pavlopoulos and Vermunt [20]. This approach requires re-linkage of data and re-estimation of the model at every time interval. For this reason, it is considered extremely time-consuming and expensive by Statistics Netherlands. We therefore investigate whether estimates obtained by using this approach are invariable to time and therefore can be re-used in later time points without the need to re-link the datasets and re-estimate the statistical model.

Our results indicate that the size of the error in the measurement of the employment contract in the LFS and the ER is indeed stable over time. Therefore, the HMM models provide a way to develop a powerful macro-integration method; as measurement error rates can be considered time constant, we can develop an error correction method, based on the use of fixed parameters from an initial HMM analysis, that can be easily included in the production of official statistics.

The method can be considered superior to traditional macro-integration approaches in particular for second-order statistics. That is, while the findings suggest that first-order statistics can be approximated using traditional macro-integration, this is probably not the case when estimating second-order statistics. In more detail, the size of temporary employment in the Netherlands always lies between the estimates from the two data sources. Therefore, traditional macro-integration can easily provide rather accurate estimates of these statistics. In contrast, findings on second-order statistics indicate a different picture; according to the survey and register data, 5.8% and 7.3% of workers with temporary contracts are employed with permanent contracts 3 months later. This indicates a substantial amount of mobility in the labour market. However, our HMM model suggests that this mobility is only 1.7%. Therefore, a static reconciliation approach, which is what typically traditional macro-integration methods do, is unlikely to provide an effective error correction of second order statistics.

However, a formal comparison of the outcomes of traditional macro-integration and macro-integration based on the HMM-model should confirm that the last method is superior to the first. Moreover, it would be interesting to test also a more advanced way of traditional macro-integration. Instead of doing the integration process in two steps (first, the integration of the stock data on the two reference dates and second, make the transitions consistent with the stock data) doing it in one go in such a way that the identity relationships on the two reference dates and the transitions are met. Further research will provide a formal comparison of our approach with traditional macro-integration approaches.

Nevertheless, before the macro-integration approach based on HMM-model enters the production of official statistics, further issues have to be addressed. Our analysis has ignored the effect of a possible linkage error between the two data sources. Although the probability of a linkage error is very small in our data, this error is particularly important when it is correlated with the outcome variable – here, the type of contract – and, thus, might bias the results. Moreover, the analysis did not fully take into account the overtime changes in the way LFS is conducted, such as the transition from dependent to independent interviewing or to a different mode of interviewing (i.e. face-to-face vs. phone or online surveys). Those aspects are likely to impact the error in the survey data and, therefore, further analysis which will investigate those aspects is required.

Acknowledgements

The first author acknowledges the contribution of Statistics Netherlands for financing her PhD project and for making the data available for this research. The authors also thank Harry Ganzeboom and all other members of the SILC research group of the Vrije Universiteit Amsterdam for reviewing the paper and providing valuable comments and constructive feedback.

References

[1] Delden A van, Pannekoek J, Banning R, Boer A de. Analysing correspondence between administrative and survey data. Stat J IAOS 2016;32:569–84. doi:10.3233/SJI-160972.

[2] Waal T de., Pannekoek J, Scholtus S. Handbook of statistical data editing and imputation. Wiley; 2011.

[3] Waal T de. Obtaining numerically consistent estimates from a mix of administrative data and surveys. Stat J IAOS 2016;32:231–43. doi:10.3233/SJI-150950.

[4] Guarnera U, Varriale R. Estimation from contaminated multi-source data based on latent class models. Stat J IAOS 2016;32:537–44. doi:10.3233/SJI-150951.

[5] Saris WE, Gallhofer IN. Design, Evaluation, and Analysis of Questionnaires for Survey Research. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2007. doi:10.1002/9780470165195.

[6] Alwin DF. Margins of Error. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2007. doi:10.1002/9780470146316.

[7] Bakker BFM. Estimating the validity of administrative variables. Stat Neerl 2012;66:8–17. doi:10.1111/j.1467-9574.2011.00504.x.

[8] Oberski DL, Kirchner A, Eckman S, Kreuter F. Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. J Am Stat Assoc 2017;forthcomin.

[9] Scholtus S, Bakker BFM, van Delden A. Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables. Den Haag/Heerlen: 2015.

[10] Oberski DL. Estimating error rates in an administrative register and survey questions using a latent class model. In: Biemer P, West B, Eckman S, Edwards B, Tucker C, editors. Total Surv. Error, New York: Wiley; 2015.

[11] Huynh M, Rupp K, Sears J. The assessment of Survey of Income and Program Participation (SIPP) benefit data using longitudinal administrative records. Washington, DC: 2001.

[12] Bakker BFM, Daas PJH. Methodological challenges of register-based research. Stat Neerl 2012;66:2–7. doi:10.1111/j.1467-9574.2011.00505.x.

[13] Zhang L-C. Topics of statistical theory for register-based statistics and data integration. Stat Neerl 2012;66:41–63. doi:10.1111/j.1467-9574.2011.00508.x.

[14] Bound J, Brown C, Mathiowetz N. Measurement Error in Survey Data. In: Heckman JJ, Leamer E, editors. Handb. Econom., vol. 5, Amsterdam: Elsevier; 2001, p. 3705–843.

[15] Fuller WA, editor. Measurement Error Models. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 1987. doi:10.1002/9780470316665.

[16] Carroll RJ, Midthune D, Freedman LS, Kipnis V. Seemingly Unrelated Measurement Error Models, with Application to Nutritional Epidemiology. Biometrics 2006;62:75–84. doi:10.1111/j.1541-0420.2005.00400.x.

[17] Bolck A, Croon M, Hagenaars JA. Estimating Latent Structure Models with Categorical Variables: One-Step Versus Three-Step Estimators. Polit Anal 2004;12:3–27. doi:10.1093/pan/mph001.

[18] Pavlopoulos D, Muffels R, Vermunt JK. How Real is Mobility between Low Pay, High Pay and Non-Employment. J R Stat Soc Ser A 2012;175:749–73. doi:10.1111/j.1467-985X.2011.01017.x.

[19] Hagenaars JA. Latent Variables in Log-Linear Models of Repeated Observations. In: von Eye A, Clogg CC, editors. Latent Var. Anal. Appl. Dev. Res., Thousand Oaks, CA: Sage Publications; 1994, p. 329–52.

[20] Pavlopoulos D, Vermunt JK. Measuring temporary employment. Do survey or

register data tell the truth? Surv Methodol 2015;41:197–214.

[21] Biemer P. An Analysis of Classification Error for the Revised Current Population Survey Employment Questions. Surv Methodol 2004;30:127–40.

[22] Biemer P. Latent Class Analysis of Survey Error. New Jersey: John Wiley & Sons; 2011.

[23] Vermunt JK. A General Latent Class Approach to Unobserved Heterogeneity in the Analysis of Event History Data. In: Hagenaars JA, McCutcheon AL, editors. Appl. Latent Cl. Anal., Cambridge: Cambridge University Press; 2002, p. 383–407. doi:10.1017/CBO9780511499531.

[24] van der Pol F, Langeheine R. Mixed Markov Latent Class Models. Sociol Methodol 1990;20:213–47.

[25] Manzoni A, Vermunt JK, Luijkx R, Muffels R. Memory Bias in Retrospectively Collected Employment Careers: A Model-Based Approach to Correct for Measurement Error. Sociol Methodol 2010;40:39–73.

[26] Bingley P, Martinello A. Measurement error in the Survey of Health, Ageing and Retirement in Europe: A Validation Study with Administrative Data for Education Level, Income and Employment. 2014.

[27] Bergin A. Job Changes and Wage Changes: Estimation with Measurement Error in a Binary Variable. Maynooth: 2013.

[28] King TR, Cook SL, Childs J. Interviewing Proxy versus Suspected Duplicate Respondents to Obtain Information Regarding Potential Duplicate Living Situations 2012.

[29] Groen JA. Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures 2012;28:173–98.

[30] Conrad FG, Rips LJ, Fricker SS. Seam Effects in Quantitative Responses 2009;25.

[31] Amuedo-Dorantes C. Work Transitions into and out of Involuntary Temporary Employment in a Segmented Market: Evidence from Spain. Ind Labor Relations Rev 2000;53:309–25.

[32] Guell M, Petrongolo B. How binding are legal limits? Transitions from temporary to permanent work in Spain. Labour Econ 2007;14:153–83.

Tables

Table 1- Distribution of contract types according to the survey and the register, January, February and March 2009

January '09

	Survey	Register
Permanent	0.637	0.586
Temporary	0.113	0.154
Other	0.250	0.261
Total	1.000	1.000
Cases	3,173	3,175

February '09

	Survey	Register
Permanent	0.627	0.576
Temporary	0.120	0.163
Other	0.254	0.262
Total	1.000	1.000
Cases	2,857	2,849

March '09

	Survey	Register
Permanent	0.642	0.596
Temporary	0.112	0.155
Other	0.247	0.249
Total	1.000	1.000
Cases	2,678	2,692

Table 2- Cross- tabulation of contract type according to survey and the register

	Survey Data				
Register Data	**Permanent**	**Temporary**	**Other**	**Total**	**Cases**
Permanent	0.934	0.052	0.015	1.000	21,840
Temporary	0.517	0.441	0.043	1.000	5,347
Other	0.060	0.059	0.881	1.000	8,411
Total	0.665	0.112	0.224	1.000	35,598
Cases	23,654	3,983	7,961	35,598	-

* The frequency distributions are calculated for all observations in the sample which are non-missing for both the LFS and ER.

Table 3- Observed 3-month transitions in LFS and PA

	Observed transitions from the survey data (LFS)		
		Contract in t	
Contract in t-3	**Permanent**	**Temporary**	**Other**
Permanent	0.983	0.006	0.011
Temporary	0.058	0.879	0.063
Other	0.016	0.037	0.947
Total	0.672	0.110	0.218
	Observed transitions from the register data (ER)		
		Contract in t	
Contract in t-3	**Permanent**	**Temporary**	**Other**
Permanent	0.976	0.012	0.012
Temporary	0.073	0.869	0.058
Other	0.019	0.043	0.938
Total	0.623	0.148	0.229

* For both tables, these are the transition rates over a 3-month period and for 34,387 cases of the pooled sample. These cases come from the LFS- respondents that appear at least twice in the sample and have an observation for both LFS and AP.

Table 4- Fit measures for nine models estimated with the linked LFS and ER data

	LL	BIC (LL)	AIC (LL)	Parameters	L²	df	p-value
A': ICE Error LFS	-35,983	72,365	72,053	44	32,458460	8,842	8.9e-2635
A'': ICE Error ER	-58,742744	11,78857887	11,7574	44	77,979	8,842	4.6e-10837
A: ICE Error LFS and ER	-35,852	72,159	71,805	50	32,198	8,836	2.2e-2595
B': non-ICE Error LFS	-35,717719	71,926928	71,544	54	55,691	8,832	2.1e-6647
B'': non-ICE Error ER	-30,875	62,313	61,873	62	54,435	8,824	1.3e-6421
B: non-ICE Error LFS and ER	-31,048050	62,697699	62,230	66	60,361363	8,820	3.3e-7512
C': non-ICE Error ER with covariates	-31,025027	62,942944	62,248	98	61,588590	8,788	1.3e-7753
C'': non-ICE Error ER with covariates also initial state	-30,647	62,240	61,503	104	60,831	8,782	2.6 e-7615
C: non-ICE Error LFS and ER with covariates also initial state	-30,634636	62,287289	61,493	112	60,806	8,774	5.6e-7614

Note: A'- ICE for the survey; A"- ICE for the register, A- ICE for both datasets, respectively.
B' – survey error depends on age and proxy interview; B"- register errors serially correlated; B- combines B' and B".
C' extend B" by introducing gender, age, education and country of origin as predictors for the transitions.
C' extend B" by introducing gender, age, education and country of origin as predictors for both the initial state and the transitions, respectively.
C extends B by introducing gender, age, education and country of origin as predictors for both the initial state and the transitions.

Table 5- The size of the measurement error in the survey data according to Model C''

	Own analysis			Pavlopoulos and Vermunt (2015)		
	Observed contract in t			Observed contract in t		
Latent contract in t	**Permanent**	**Temporary**	**Other**	**Permanent**	**Temporary**	**Other**
Permanent	0.996	0.003	0.002	0.998	0.001	0.002
Temporary	0.090	0.878	0.033	0.125	0.832	0.042
Other	0.011	0.006	0.984	0.004	0.005	0.991

Table 6- The size of the measurement error in the register data according to Model C''

	Own analysis			Pavlopoulos and Vermunt (2015)		
	Observed contract in t			Observed contract in t		
Latent contract in t	**Permanent**	**Temporary**	**Other**	**Permanent**	**Temporary**	**Other**
Permanent	0.877	0.106	0.017	0.888	0.081	0.031
Temporary	0.247	0.635	0.118	0.237	0.684	0.079
Other	0.033	0.013	0.954	0.032	0.017	0.951

Table 7- Conditional probabilities of measurement error in the register data in time t when no error has been made in t-1 according to the C'' model with fixed error parameters

	Latent contract in t		
Observed contract in t	**Permanent**	**Temporary**	**Other**
Permanent	0.986	0.045	0.005
Temporary	0.009	0.930	0.005
Other	0.004	0.025	0.990

Table 8- Conditional probabilities of repeating an error in time t that has been made in t-1

	Own analysis			Pavlopoulos and Vermunt (2015)		
	Observed contract in t			Observed contract in t		
Latent contract in t	**Permanent**	**Temporary**	**Other**	**Permanent**	**Temporary**	**Other**
Permanent		0.977	0.954		0.973	0.961
Temporary	0.970		0.921	0.968		0.896
Other	0.935	0.848		0.913	0.842	

Table 9- Conditional probabilities of making an error in time t that is different from the error made in t-1

Own data			Pavlopoulos and Vermunt (2015)		
Latent contract in t			Latent contract in t		
Permanent	**Temporary**	**Other**	**Permanent**	**Temporary**	**Other**
0.013	0.079	0.009	0.014	0.070	0.001

Note: the probabilities of the last row come from table 4.3 of Pavlopoulos and Vermunt

Table 10- The average size of temporary employment according to Model C''

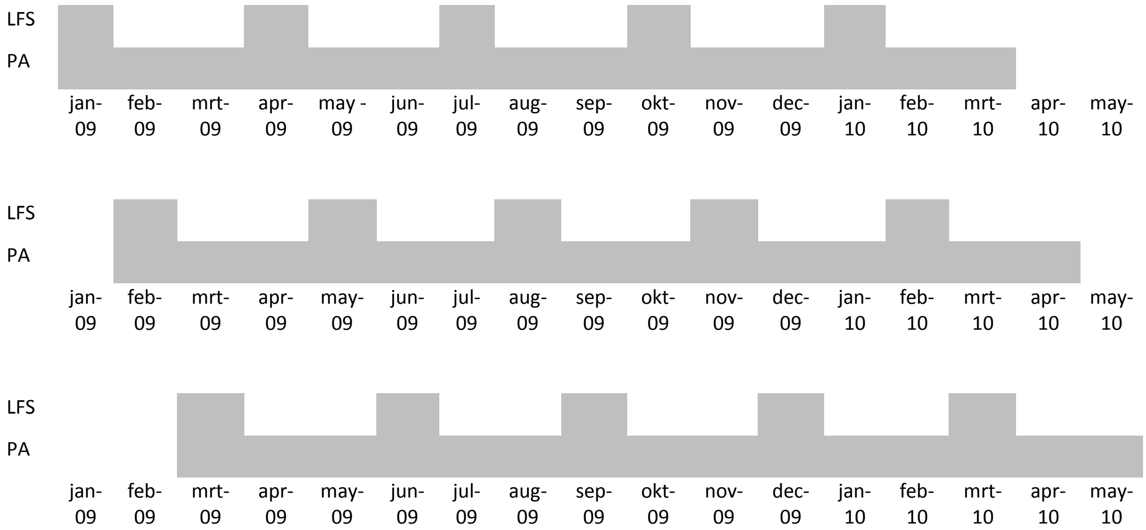
	Survey	Register	Latent- own analysis	Using fixed error parameters according to Pavlopoulos and Vermunt (2015)
Permanent	0.653	0.585	0.611	0.613
Temporary	0.110	0.151	0.128	0.131
Other	0.237	0.264	0.261	0.257
Cases	36,321	130,671	133,290	133,290

Table 11- Latent 3-months transitions according to model C''

	Own analysis			Using classification tables from Pavlopoulos and Vermunt (2015)		
Contract in t-3	**Permanent**	**Temporary**	**Other**	**Permanent**	**Temporary**	**Other**
Permanent	0.987	0.004	0.009	0.989	0.004	0.008
Temporary	0.017	0.929	0.054	0.016	0.928	0.056
Other	0.006	0.030	0.963	0.006	0.029	0.965
Total	0.610	0.128	0.263	0.610	0.132	0.258

Figures

Fig. 1 - An illustration of the sample



* The figure illustrates how the 3-monthly rotation panel of the LFS corresponds to monthly observations from the ER. A grey shaded cell indicates a valid observation.

Fig. 2- Path diagram for the hidden Markov model with two indicators, serially correlated and covariate dependent register errors and predictors for latent transitions and latent state probabilities

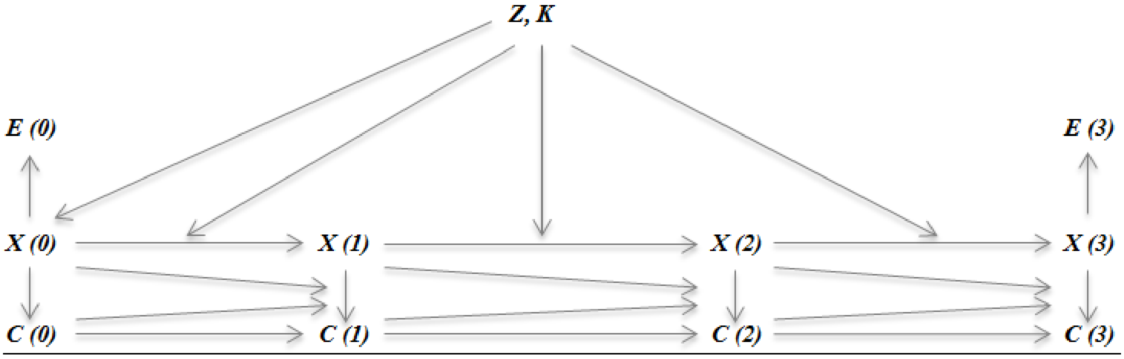
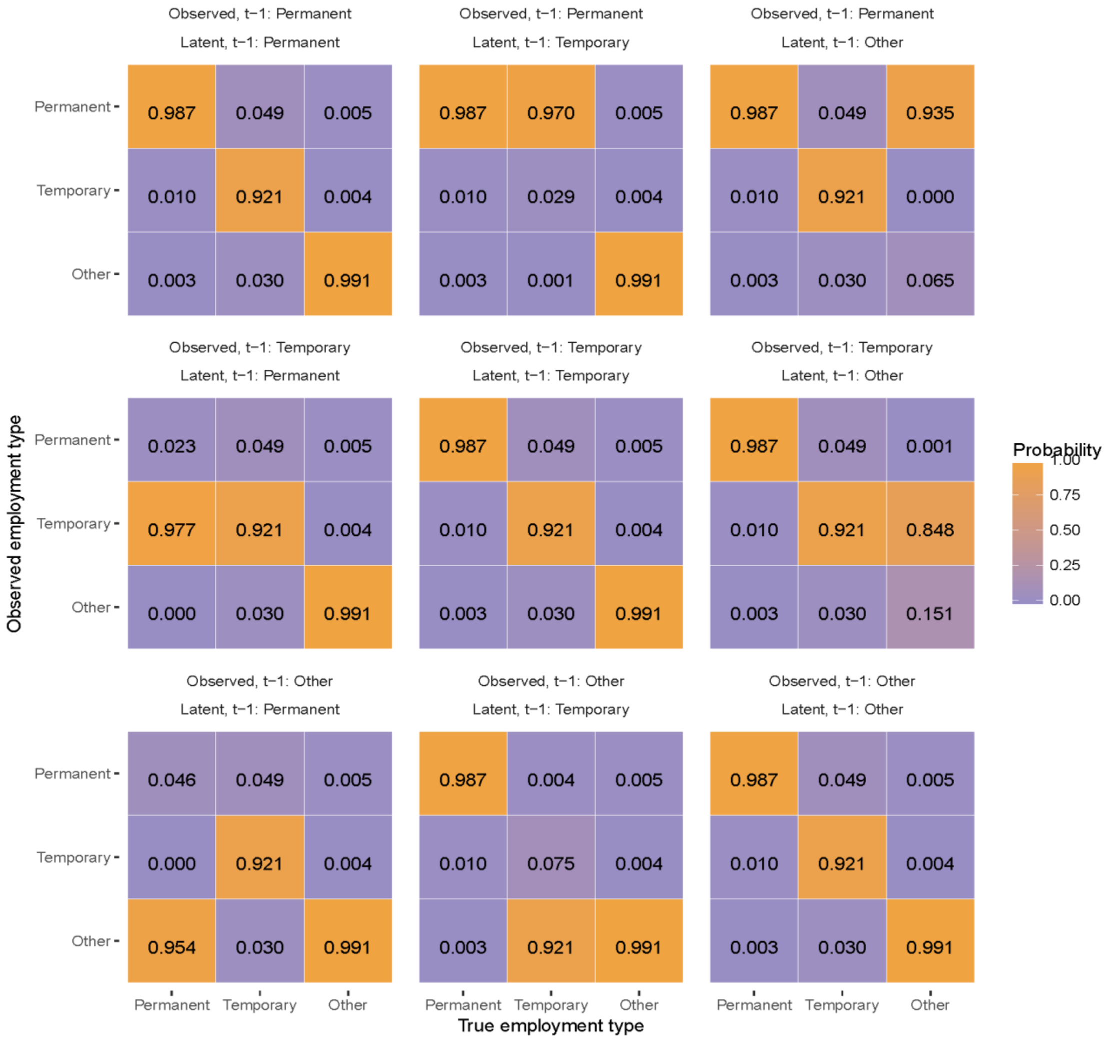


Fig. 3- Conditional probabilities of measurement error in register data according to Model C" (own analysis)



Note: use of average posterior probabilities