**Integration of inconsistent data sources using Hidden Markov Models**

**Keywords:** Measurement Error; Hidden Markov Models (HMMs); data integration

1. Introduction

National Statistical Institutes (NSI’s) often obtain information on the same phenomena from different data sources (Bakker, 2011; Van Delden et al., 2016). Even though those sources are in most cases subjected to editing which aim at ensuring that they measure exactly the same variable and cover the same target population (De Waal et al. 2011; De Waal, 2016), identical units do not always yield identical values (Guarnera & Variale, 2016). Such mismatch is the result of measurement error in the data sources that are involved and is likely to lead to the unwanted publication of differing statistics.

NSIs apply several methods to account for the inconsistencies in data sources caused by measurement error. Most commonly, differences are ignored and only estimates from the source assumed to have the best quality are published. Alternatively, it is assumed that the quality of both sources is similar and so the mean of the estimates from both sources is used. A more advanced approach relies on applying different macro-integration techniques.

An alternative, and arguably superior, strategy which resolves the problem of inconsistencies in categorical data and has been increasingly applied across various disciplines is the one based on latent class modelling (LCM). LCM is a methodology in which the problems of data reconciliation and measurement error are solved simultaneously by linking two or more sources and modeling them as conditionally independent measures of an underlying true value. A specific group of latent class models applied to longitudinal data specifically are Hidden Markov Models (HMMs).

While LCM in general and HMMs in particular serve as an attractive solution to the problems of discrepancy across data sources, several issues need to be considered before they can be utilized in the production of official statistics. First, the procedures involved in applying and estimating HMMs are very complicated, time-consuming and expensive and, therefore, cannot be applied regularly. Thus, is it desirable to re-use HMMs estimates from previous time points with more recent data. This procedure may produce accurate estimates only if measurement error is time invariant.

Second, HMMs are based on the local independence assumption.[[1]](#footnote-1) Although this assumption is crucial for model identification, it is rather restrictive and unrealistic as it implies that the measurement error is random. To relax this assumption while maintaining model identifiability it is necessary to use multiple indicators per time point and hence multiple sources. However, the use of more than one data source requires record linkage which might result in linkage error – a new potential source of bias.

In our research, we test the feasibility of using HMMs as a way to reconciliate different sources which measure the same phenomenon. In doing so we apply an extended, two-indicator HMM to Dutch data on transitions from temporary to permanent employment coming from the Labour Force Survey (LFS) and the Employment Register (ER).

In addition, we also show that HMMs can be used to evaluate how effective various data collections techniques are in producing accurate statistics. We illustrate this application of HMMs by examining how the use of different modes of interviewing (dependent vs. independent) affects the level measurement error in the LFS.

1. Methodology

Hidden Markov Models (HMMs) are a group of latent class models increasingly used to estimate and correct for measurement error in longitudinal categorical data (Biemer, 2011). The basic HMM assumes that, at each time point t, the observed answer $Y\_{t}$ is generated independently, with a certain probability, from the true, but unobserved, value $X\_{t}$ (the local independence assumption); and that, $X\_{t}$ follows a first order autoregressive process, where each value carries over partly to the next time point (the Markov assumption). Thus, the model can be written as follows:

$$P\left(Y\right)=\sum\_{x=1}^{k}P\left(X\_{0}\right)\prod\_{t=0}^{T}P\left(X\_{\left(t-1\right)}\right)\prod\_{t=0}^{T}P\left(X\_{t}\right)$$

Where $P\left(X\_{0}\right)$ and $P\left(X\_{\left(t-1\right)}\right)$ denote the true/ latent initial state and transition probabilities and $P\left(X\_{t}\right)$ denote the observed, emission probabilities.

As mentioned above, while HMMs are an attractive methodology to correct for measurement error, they also relies on the rather unrealistic local independence assumption. When conditional independence is violated, a single measure at each time point is no longer sufficient, as a HMM with a single measure that relaxes this assumption is not identifiable. However, a model with local dependence can be identified by linking additional data sources, in which the same phenomenon is measured at least twice at each time point. Such a multiple-indicator HMM can, identifiably, account for local dependence among the errors thanks to the presence of an additional indicator.

Therefore, in our analysis, we make use of an extended HMM with two observed indicators - $C$ and $E$ - which denote the employment contract type according to the register and the survey data. This model assumes that the error in the register data is serially correlated. The model assumes further that the latent transition probabilities as well as the initial state probabilities, depend on observed and unobserved individual characteristics. Figure 1 depicts the path for the first 4 months.



Figure 1- Path diagram for the extended Hidden Markov Model with two observed indicators

1. Results
	1. *The feasibility of parameter re-use*

The main challenge associated with the application of (extended) Hidden Markov Models in official statistics production is their complicated nature. Namely, utilizing HMMs in this domain is very time consuming and therefore expensive, as it requires NSIs to perform record linkage followed by model re-estimation for each new time period. Therefore, in our research we first studied the plausibility of re-using existing error parameter estimates from Pavlopoulos and Vermunt (2015) in order to estimate the true contract type with more recent data. This is a potentially attractive solution as (1) it does not require re-estimating 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.

To study the feasibility of parameter re-use, we applied the aforementioned extended HMM on linked LFS and ER data from 2009; then, we repeated the analysis for the same sample while fixing the measurement error specific parameters to those obtained by Pavlopoulos and Vermunt (2015) when analyzing the data 2007 from the same data sources; finally we compared the results of the two analyses.

Our analysis showed that the nature and size of the measurement error in both the survey and register data were very similar in 2007 and in 2009 which suggested it is possible to correct for measurement error in 2009 without having to undertake the full HMM analysis. The results presented in Table 1 confirmed this intuition; namely, as can be observed, the latent distributions of the contract types and the 3-monthly transition rates from temporary to permanent employment obtained using the ‘full’ analysis are almost identical to those obtained fixing the error parameters to those provided by Pavlopoulos and Vermunt (2015).

**Table 1 - Latent distribution of contract type and transitions from temporary to permanent contracts**

|  |  |  |
| --- | --- | --- |
|  | Full analysis | Re-using error parameters |
| Permanent | 0.611 | 0.613 |
| Temporary | 0.128 | 0.131 |
| Other | 0.261 | 0.257 |
| Temp to perm transition rate | 0.017 | 0.016 |

* 1. *Sensitivity of the extended HMM to linkage error*

The second challenge related to the application of HMMs is to investigate their sensitivity to linkage error. As data linkage is necessary to apply HMM’s that relax the local independence assumption, linkage error is potentially a serious threat to these models. Therefore, in our analysis we investigated the sensitivity of the two-indicator HMM to false-positive and false- negative linkage errors. False negatives occur if records of the same person are not linked. False positives occur if records of two different persons are linked.

We carried out a simulation study in which we used the linked 2009 LFS and ER data (which we assumed to be perfectly linked) and simulated various levels and types of linkage error onto this dataset. We then estimated the transition rates from temporary to permanent employment for the datasets with the simulated error using a restricted version of the extended HMM described above (which did not contain covariates or autocorrelated error in the register data); we compared the results to those obtained using the original dataset (without simulated linkage error) to approximate the bias.

The results are summarized in figure 3 and show that, in most cases, the biasing effects of both false-negative and false-positive linkage errors are negligible. The bias is substantial only when the exclusion or mislinkage probabilities depend on a covariate strongly correlated with the model outcomes (i.e. transitioning from a temporary to a permanent contract according to the register data), and the overall level of the error is high (i.e. 20%). Thus, it can be concluded that the model estimates of the extended HMM are only sensitive to linkage error in rather extreme cases.

**Figure 3 - Relative bias by level of false-negative (left) and false-positive (right) linkage error**



1. Conclusions

Measurement error is a serious threat in the production of official statistics and Hidden Markov Models present an attractive opportunity of mitigating this threat. The increasing availability of linked data from difference sources offers the opportunity of making these models more realistic as we can relax the unattractive assumption of local independence. However, several problems remain to be solved before these models can be applied in the production of official statistics. In this paper, we discuss two of these problems: the feasibility of parameter re-use and the sensitivity of HMM’s to linkage error.

The first problem is that HMMs are complicated and time consuming to apply in official statistics. However, as we showed in our analysis, if the error parameters are time invariant, parameter estimates from previous time points can be re-used on more recent data without having to link datasets again and apply the HMM.

The second problem is that the application of the extended HMM often requires record linkage, which can lead to linkage error – a new source of bias. In our simulation study, however, we showed that the sensitivity of the method to linkage error is low. Only scenarios with very high levels of linkage error and where the probability of exclusion or mislinkage is highly correlated with model estimates lead to substantial bias. Such extreme scenarios, though, are unlikely to occur in practice.

Finally, in our research we have also examined how the transition to a different mode of interviewing affected measurement error in the LFS[[2]](#footnote-2). More specifically we looked at how moving from proactive dependent interviewing to fully independent interviewing, when asking about the contract type of individuals who had previously indicated to hold a temporary contract, affected the misclassification rate of contracts. Our results showed that the overall level of measurement error was unaffected by this change suggesting that measurement error is rather robust to at least certain changes or alterations in the data collection processes.

References

Bakker, B. F. (2012). Estimating the validity of administrative variables. *Statistica Neerlandica*, *66*(1), 8-17.

Biemer, P. P. (2011). *Latent class analysis of survey error* (Vol. 571). John Wiley & Sons.

Delden, A. V., Pannekoek, J., Banning, R., & Boer, A. D. (2016). Analysing correspondence between administrative and survey data. *Statistical Journal of the IAOS*, *32*(4), 569-584.

Guarnera, U., & Varriale, R. (2016). Estimation from contaminated multi-source data based on latent class models. *Statistical Journal of the IAOS*, *32*(4), 537-544.

Pavlopoulos, D., & Vermunt, J. K. (2015). Measuring temporary employment. Do survey or register data tell the truth? *Survey Methodology*, 41, 197–214.

Waal, T. D. (2016). Obtaining numerically consistent estimates from a mix of administrative data and surveys. *Statistical Journal of the IAOS*, *32*(2), 231-243.

De Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation* (Vol. 563). John Wiley & Sons.

1. According to which the latent state at time t only depends on the latent state at time t-1 and conditional on that it is independent of everything else. [↑](#footnote-ref-1)
2. This analysis has not been mentioned in the results section due to space limitations but will be discussed in the presentation (and the paper following) in greater detail. [↑](#footnote-ref-2)