Catalogue no. 11-522-X
ISSN 1709-8211

**Proceedings of Statistics Canada Symposium 2022:
Data Disaggregation: building a more representative data portrait of society**

Modelling intra-annual measurement in linked administrative and survey data

by Jetske Marcelis, Arnout van Delden, Sander Scholtus and Femke Kessels

Release date: March 25, 2024

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2022
Data Disaggregation: Building a more representative data portrait of society

Modelling intra-annual measurement in linked administrative and survey data

Jetske Marcelis, Arnout van Delden, Sander Scholtus and Femke Kessels[1]

Abstract

At Statistics Netherlands (SN) for some economic sectors two partly-independent intra-annual turnover index series are available: a monthly series based on survey data and a quarterly series based on value added tax data for the smaller units and re-used survey data for the other units. SN aims to benchmark the monthly turnover index series to the quarterly census data on a quarterly basis. This cannot currently be done because the tax data has a different quarterly pattern: the turnover is relatively large in the fourth quarter of the year and smaller in the first quarter. With the current study we aim to describe this deviating quarterly pattern at micro level. In the past we developed a mixture model using absolute turnover levels that could explain part of the quarterly patterns. Because the absolute turnover levels differ between the two series, in the current study we use a model based on relative quarterly turnover levels within a year.

Key Words: Mixture models; Measurement errors; Reporting errors; Tax data; Seasonal patterns.

1. Introduction

Like other countries, Statistics Netherlands is using value added tax (VAT) data to estimate intra-annual turnover levels and changes. In many of those cases the statistical agency first used surveys to estimate turnover levels and/or changes and then starts to investigate whether VAT can be used instead. When one compares turnover values derived from VAT with those obtained by a survey, measurement errors can be found in both source types. Examples of studies on measurement errors in VAT data are Țîru et al. (2019) and Lewis and Woods (2013).

In the current paper we deal with a specific form of measurement error that can occur in VAT data, namely the occurrence of so-called quarterly effects: relatively large turnover values in the fourth quarter of the year and relatively small values in the first quarter. As far as we know these patterns occur because businesses close their booking year and come up with corrections to make sure that the yearly amount of declared VAT is correct. This phenomenon is likely to occur in other countries also. We could study these seasonal patterns since, for some economic sectors in the Netherlands, two partly-independent intra-annual turnover index series are available: a monthly and a quarterly series. The monthly index series is based on a sample survey and is used to produce output for the short-term statistics (STS). The quarterly index series consists of turnover derived from Value Added Tax (VAT) data for the smaller and simple enterprises and re-use of the monthly survey data (aggregated to a quarter) for the more complex enterprises. The two sources enclosed in the quarterly series cover nearly all units in the target population, and are therefore also referred to as census data. From the quarterly census data yearly turnover totals are derived which are used to calibrate the outcomes of the structural business statistics (SBS). In turn, the SBS is input into the National Accounts. The reason for this calibration step is that the census turnover data are considered to be of higher quality since they are not prone to sampling error. The National Accounts are published in different releases: late releases are based on the SBS, while early releases are based on the STS. As a result, differences between the monthly survey series and the quarterly census series may lead to differences between early and late(r) National Account releases.

At Statistics Netherlands we would like to benchmark the monthly turnover index series to the quarterly census data on a quarterly basis. That would not only lead to consistency between the quarterly figures of the survey and the census data but it would also help to reduce the National Accounts adjustments because inaccuracies in the survey series are

[1] All Authors, Statistics Netherlands, P.O. Box 24500 2490 HA The Hague, The Netherlands (j.marcelis@cbs.nl, a.vandelden@cbs.nl (corresponding author), s.scholtus@cbs.nl, flm.kessels@cbs.nl).

then corrected on a quarterly basis and do not accumulate. The survey series is prone to variance especially in industries with a limited sample size. Furthermore, a cut-off sample is used in which the smaller enterprises are not observed, which could lead to a (small) bias.

When applying this benchmark, the growth rates of the monthly series of the final releases are adjusted under the restriction that the adjusted indices per quarter are identical to the quarterly indices from the census data. In the past, we have compared different ways to perform this benchmarking, such as ratio adjustment and a Denton method (Daalmans, 2018). Back in 2016, CBS benchmarked the monthly series of 2015 with the quarterly series of that year using a Denton method (see Bikker et al., 2013; Denton, 1971). This led to some unexpected results since, for the majority of economic sectors, the year-on-year (yoy) growth rates of quarterly turnover from the sample survey data were adjusted downwards in the first quarter of the year and upwards in the fourth quarter of the year (see Van Delden and Scholtus, 2017). For example, for the economic sector Retail trade, the adjustments on yoy growth rates of quarterly turnover in the four quarters were –0.5, +0.5, +0,2, +1.0. Further study showed that this effect was partly due to the use of the Denton method, and partly because of the quarterly effects in the VAT data.

Van Delden and Scholtus (2017) made a first analysis of the quarterly turnover differences between the two sources in 2014 and 2015. They analysed observed survey and VAT data that were linked at enterprise level, for all enterprises for which observed data from both sources were available. They analysed the data by using a linear regression model for the quarterly data with a slope that could vary for each of the quarters and a regression method that can cope with outliers. To handle outliers they tested a robust linear regression model as well as a two-group mixture model similar to Di Zio and Guarnera (2013). In this two-group model, one group had a moderate residual variance and a second group had a large residual variance. The seasonal effects that were found with that two-group model were relatively small and not entirely consistent over the different economic sectors. As a next step, Van Delden et al. (2020) developed an extended mixture model in which they aimed to explain the observed quarterly effect by using different groups that could vary in size of quarterly effect and in the size and structure of the (co)variances. Van Delden et al. (2020) compared a six-group model with the previous two-group model for the economic sector Job placement and found that they could already explain more of the quarterly effects with the six-group model.

Results on other economic sectors (Retail trade, Manufacturing, Construction) showed that this extended mixture model could still only explain part of the seasonal effects. One of the problems we were facing is that the total yearly turnover levels as reported by VAT were larger than those reported in the sample survey (see **Error! Reference source not found.** below). As far as we know the turnover at enterprise level based on VAT declarations are sometimes a bit too large because it is not corrected for internal deliveries among legal units within an enterprise. Furthermore, the turnover reported by the survey is sometimes too small, for instance when enterprises only reported turnover of their main activities in the survey and not their secondary activities. The extended mixture model sometimes picked up groups in the population whose VAT turnover levels were larger than their survey turnover, whereas we were seeking to explain the quarterly seasonal effects. In the current paper we therefore present a new mixture model, in which we use the relative turnover values within a year. That way, we eliminate the effects of the yearly turnover level differences. The aim of the current paper is to determine to what extent the new model is able to describe the deviating quarterly patterns. The ultimate aim is to correct the VAT data for these quarterly effects.

The remainder of this paper is organised as follows. We start with section 2 that describes the mixture model based on the relative turnover. Next, section 3 we describe the empirical data for which we try to explain the quarterly effects. Section 4 describes some tests on the new mixture model and in section 5 we apply the model to Job placement data. Finally, section 6 discusses the results.

2. A mixture model for relative turnover differences

In this section, we introduce a mixture model for detecting enterprises with deviating quarterly turnover patterns between the survey and VAT data. For a given year, let $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$ and $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})'$ denote the observed vectors of four quarterly turnover values of enterprise i, in the survey and VAT data respectively. We divide each vector by its total (annual turnover) to obtain vectors of relative quarterly turnover values, $\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})'$ and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})'$, with $\sum_{k=1}^{4} y_{ik} = \sum_{k=1}^{4} x_{ik} = 1$. For simplicity, we assume that all $y_{ik} \geq 0$ and $x_{ik} \geq 0$ and develop the mixture model only for enterprises that satisfy this assumption. Both $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ are examples of so-called *compositional data*; see, e.g., Aitchison (1986).

The $(p-1)$-dimensional Dirichlet distribution provides a flexible way to model compositional data, i.e., vectors $\boldsymbol{a} = (a_1, \dots, a_p)'$ with all $a_k \geq 0$ and $\sum_{k=1}^{p} a_k = 1$. The probability density function of this distribution is given by:

$$f_{Dir}(\boldsymbol{a};\boldsymbol{\beta}) = \frac{\Gamma\left(\sum_{k=1}^{p} \beta_k\right)}{\prod_{k=1}^{p} \Gamma(\beta_k)} \prod_{k=1}^{p} a_k^{\beta_k - 1}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ denotes a vector of positive parameters and $\Gamma(u) = \int_0^{\infty} v^{u-1} e^{-v} \mathrm{d}v$ is the gamma function. Let $\beta_{tot} = \sum_{k=1}^{p} \beta_k$ and $\beta_k^* = \beta_k / \beta_{tot}$. The marginal first and second central moments of the Dirichlet distribution in (1) are given by $E(a_k) = \beta_k^*$ and $\mathrm{var}(a_k) = \beta_k^*(1-\beta_k^*)/(\beta_{tot}+1)$. It is seen that, by increasing (decreasing) all β_k by the same factor, we can obtain a distribution with the same central point but smaller (larger) variances.

To model the differences between $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ using Dirichlet distributions, we define a transformed difference vector:

$$\boldsymbol{d}_i = \frac{1}{4}\iota_4 - \frac{\boldsymbol{y}_i - \boldsymbol{x}_i}{4}, \tag{2}$$

where $\iota_4 = (1,1,1,1)'$. The elements of the vector $\boldsymbol{d}_i$ satisfy $\sum_{k=1}^{4} d_{ik} = 1$ with $0 \leq d_{ik} \leq 1/2$. In the absence of systematic differences between the distributions of $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$, we would expect that $E(d_{ik}) = 1/4$ for all k.

As in Van Delden et al. (2020), we allow for multiple subpopulations (groups) of enterprises, each with a different relation between survey and VAT turnover values. Our proposed model for $\boldsymbol{d}_i$ is a mixture of Dirichlet distributions:

$$f(\boldsymbol{d}_i) = \prod_{g=1}^{G} \left\{ \alpha_g \cdot f_{Dir}\left(\boldsymbol{d}_i; \kappa_g \left(\frac{1}{4}\iota_4 + \boldsymbol{\delta}_g \right) \right) \right\}^{z_{gi}}. \tag{3}$$

Here, G denotes the number of groups and $z_{gi} \in \{0,1\}$ is an indicator denoting whether unit i belongs to group g (with $\sum_{g=1}^{G} z_{gi} = 1$). Within each group, a Dirichlet distribution of the form (1) is assumed with a parameter vector of the form $\boldsymbol{\beta}_g = \kappa_g(\iota_4/4 + \boldsymbol{\delta}_g)$. The scalar parameter κ_g determines the amount of variance within group g, while the parameters $\boldsymbol{\delta}_g = (\delta_{g1}, \delta_{g2}, \delta_{g3}, \delta_{g4})'$ describe potential systematic quarterly differences between survey and VAT data. We use the natural restriction that $\sum_{k=1}^{4} \delta_{gk} = 0$. In some groups, we may add the restriction $\boldsymbol{\delta}_g = \boldsymbol{0}$, indicating that no systematic differences occur for enterprises in that group. Finally, the model parameters $\boldsymbol{\alpha} = (\alpha_1, \dots \alpha_G)'$ denote the relative sizes of the different groups in the population, with $\sum_{g=1}^{G} \alpha_g = 1$. The full set of model parameters is denoted by $\boldsymbol{\theta}$ and contains $\boldsymbol{\alpha}$, $\kappa_1, \dots, \kappa_G$, and all $\boldsymbol{\delta}_g$ that have not been restricted to $\boldsymbol{0}$.

In practice, the group indicators z_{gi} are not observed. To estimate a mixture model of the form (3), we can use an Expectation (Conditional) Maximisation (E(C)M) algorithm (McLachlan & Peel, 2000). In this algorithm, two steps are repeated until convergence:

- E step: Given the current parameter estimates $\boldsymbol{\theta}$, evaluate $\tau_{gi} = E(z_{gi} | \boldsymbol{d}_i, \boldsymbol{\theta}) = P(z_{gi} = 1 | \boldsymbol{d}_i, \boldsymbol{\theta})$.
- M step: Update the parameter estimates by maximizing the log likelihood function based on (3), with all unknown z_{gi} replaced by their expected value τ_{gi}.

To simplify the computations in the M step, the parameters $\kappa_1, \dots, \kappa_G$ and $\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_G$ are updated separately, conditional on the current values of the other parameters, which makes this an ECM algorithm rather than an EM algorithm. The algorithm requires starting values for $\boldsymbol{\theta}$ and may converge to a suboptimal solution depending on these starting values. To ensure that the optimal solution is found, multiple sets of starting values should be tried and only the solution with the best value of the log likelihood function is retained.

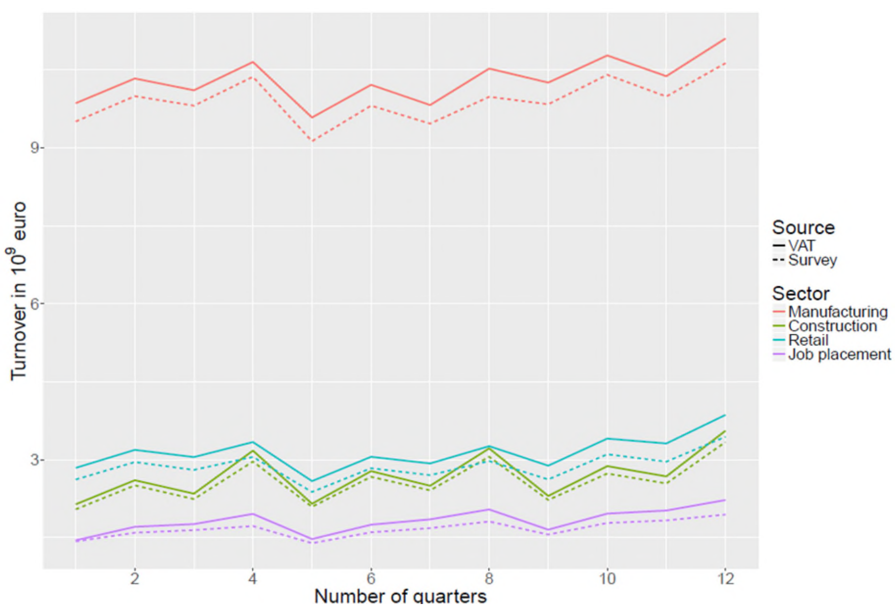
To find the best specification of model (3) for the data at hand – which includes selecting the number of groups G and deciding which of these groups contain systematic differences represented by $\boldsymbol{\delta}_g$ – we can fit multiple models and compare their AIC and BIC based on the log likelihood and the number of model parameters. A third possible fit measure is the ICL-BIC, which extends the BIC to also take into account how well the estimated model is able to assign units to a single group based on τ_{gi}. For more details on these fit measures, see McLachlan & Peel (2000).

3. Empirical data

In the current paper we use data of four economic sectors (Manufacturing, Construction, Retail trade and Job placement) and three years (2014–2016) to study differences in seasonal patterns between survey and VAT data. Manufacturing, Construction, and Retail trade have monthly survey data. STS output for Job placement was based on quarterly survey data during 2014-2016, but in more recent years the STS output is based on the census data and the survey series has ended. We included Job placement data because this sector has clear seasonal effects (see Van Delden and Scholtus, 2019), so these data were very suitable to develop the model. Therefore, most of the results in section 0 refer to Job Placement. The data that we used have been described before in Van Delden and Scholtus (2017) and in Van Delden et al. (2020), so here we describe them only very briefly. Each of the four economic sectors is further divided into a number of industries. We estimate the seasonal patterns per economic sector rather than per industry, because differences among the seasonal patterns of the industries were found to be too subtle to estimate.

For enterprises that responded to the sample survey we linked the VAT turnover values. We included only enterprises that responded to the sample survey for all four quarters of the year, and for which the VAT turnover was available for all four quarters of the year. For units where $Y_{ik}/X_{ik} \geq 100$ or $Y_{ik}/X_{ik} \leq 0.01$ we assumed that there was a large error in either Y_{ik}or X_{ik}, those units were therefore omitted. Furthermore, industries within the economic sectors for which the turnover level or change estimates based on VAT were considered unreliable because of differences in definition between VAT and sample survey turnover (Van Delden et al., 2016), were omitted. Post-stratum weights (w_{ki}=$w_{k\ell}$) for the enterprises i in quarter q of stratum ℓ were computed as the ratio of the population size ($N_{k\ell}$) to the size of the included units ($n_{k\ell}$). As a stratum we used a combination of industry and one-digit size class. Note that all presented results on sample and VAT data in the present study refer to the small and simple units (both included units and the corresponding populations) since only for those units we have both sources.

Figure 3-1
Estimated total turnover of the small and simple units for VAT and survey that report both to the survey and the VAT data. Quarters are numbered from the first quarter of 2014 onwards.



We estimated the quarterly turnover totals levels as $\hat{Y}_k = \sum_i w_{ki} y_{ki}$ for the survey turnover and $\hat{X}_k = \sum_i w_{ki} x_{ki}$ for the VAT turnover (**Figure 3-1**). For all quarters and economic sectors, the estimated turnover totals were larger for the VAT data than for the survey data. Moreover, the difference between VAT and survey turnover is often larger in the fourth quarter of the year and smaller in the first quarter of the year. That can most easily be seen for the Job placement sector (purple continuous line).

4. Testing the model

Using simulated data we tested to what extent the mixture model from Section 2 can be estimated reliably. These simulated data are generated from a mixture of Dirichlet distributions. We based the parameter values of this distribution on the parameter estimates of a mixture model applied to the available data sets (i.e., the data sets of four economic sectors in 2014, 2015, and 2016) to get realistic simulated data. Thus, firstly, we applied a three-group model to these data to determine common values for the sample sizes and free parameters of the mixture model. We made two different sets of parameter values to explore data with a relatively strong quarterly effect (based on Job placement) and data with a relatively weak quarterly effect (based on Manufacturing). These sets are displayed in **Table 4-1**. For the $\boldsymbol{\delta}_g$-values, the values of the third group are displayed as this is the only group which has a systematic effect (i.e., $\boldsymbol{\delta}_1 = \boldsymbol{\delta}_2 = \mathbf{0}$ in these simulations). Moreover, we did all simulations using the sample size of the original data sets on which the parameter sets are based (see column 'size' in **Table 4-1**) and using the smallest sample size of the available data sets (750 units) in order to investigate whether the simulations show similar results for a smaller sample size. We did three simulation studies (see sections 4.1 - 4.3). For each simulation study, we generated data from a mixture of Dirichlet distributions with these parameter sets 100 times.

Table 4-1
Parameter values based on the data of Manufacturing (set 1) and Job placement (set 2) which will be used to form predefined Dirichlet distributions to generate the simulated data from

	Size	$\boldsymbol{\alpha}_g$			$\boldsymbol{\kappa}_g \cdot \mathbf{10^3}$			$\boldsymbol{\delta}_{3k} \cdot \mathbf{10^{-4}}$			
		α_1	α_2	α_3	κ_1	κ_2	κ_3	δ_{31}	δ_{32}	δ_{33}	δ_{34}
Set 1	1100	0.25	0.15	0.60	0.15	700	5	-20	0	0	20
Set 2	2250	0.25	0.20	0.55	0.5	2000	20	-3.5	-0.25	0	3.75

4.1 Finding parameter estimates

In the first simulation study, we investigated how close the estimated parameter values were to the true ones given the correct number of groups. We explored four different scenarios which form a 2 × 2 design where the settings were equal group sizes ($\alpha_1, \alpha_2, \alpha_3$ = 1/3) versus unequal group sizes ($\alpha_1, \alpha_2, \alpha_3$ according to **Table 4-1**) and good versus poor starting values. Overall, the ECM algorithm performed well: it correctly estimated the true parameter values given reasonable starting values. There was only a slight bias in the estimation of some $\boldsymbol{\delta}_g$-values in the data of parameter set 2. For the simulations with unequal group sizes and poor starting values (scenario 4), the ECM algorithm could end up in local maximum in case of the data based on parameter set 2. This led to bias and variance in the parameter estimates. The variability in estimates of the α_g- and κ_g-values was small in scenarios 1, 2, and 3. However, standard errors of the $\boldsymbol{\delta}_g$-values were relatively large in all scenarios. Further inspection of the results showed that the variability and slight bias in estimates for $\boldsymbol{\delta}_g$ in scenarios 1, 2, and 3 are mainly due to the fact that the values for $\boldsymbol{\delta}_g$ in the samples differ from those in the population, indicating a sampling effect. In particular, the accuracy of the estimates did not improve much if the true group indicators z_{gi} were used. For the simulations with the smaller sample size (750 units), the parameter estimates resembled the true parameter values equally well, but the variance increased as compared to the simulations with the original sample size.

4.2 Finding the number of groups

In the second simulation study, we explored whether the ECM algorithm was able to recover the true number of groups given reasonable starting values. We ran different models in which we varied the number of groups from two to seven. Seven was the maximum number of groups identified in the earlier approach by Van Delden et al. (2020). To compare the performance of the different models, we computed the AIC, BIC, and the ICL-BIC. Next, we determined whether the model with the correct number of groups had the best fit measures in most of the simulations. For both parameter sets and both sample sizes, the results indicated that the ECM algorithm was able to recover the true number of groups in nearly all or even all simulations depending on the fit measure.

4.3 Test the effect of starting values

In the third simulation study, we further analysed the effect of different starting values on the performance of the mixture model. We found that for simulations with reasonable starting values, the algorithm correctly estimated the parameters, while for simulations with poor starting values, the ECM algorithm sometimes ended up in local maxima. In case of a local maximum the corresponding fit measures were less optimal.. The algorithm was most sensitive to the starting values for κ_g. In the remainder (section 5) we therefore used multiple sets of starting values, we especially used sufficiently varying starting values for κ_g, and we selected the final model based on the best fit measures.

5. Apply the model to Job placement

In order to evaluate how well the algorithm could detect seasonal patterns, we applied it to the sector Job placement for the years 2014–2016. As mentioned above, we used Job placement because we have found that it has clear seasonal patterns (Van Delden & Scholtus, 2019). We applied mixture models of the form (3) with two to six groups. The models with two to four groups had one group with a systematic effect, while the models with five or six groups had two groups with a systematic effect. In all data sets, the two- and three-group models showed results that were inferior to the results of models with four to six groups. So, we will only discuss the four-, five- and six-group models.

5.1 Basic models

The groups in the four-, five-, and six-group models are specified in **Table 5.1-1**. For the four-group model, there were three groups that varied in their degree of measurement error and one group with quarterly effects (systematic differences between the quarterly distribution of VAT and survey data). In the five-group model, the four-group model was extended with an extra group with quarterly effects. The six-group model has also an extra group without a quarterly effect. Moreover, there were small differences in size of variance between the groups of the three models.

Table 5.1-1
A specification of the four-, five-, and six- models on the Job placement data 2014, 2015, and 2016. This specification is in line with both the starting values and the final parameter estimates. The number of plus signs indicates the relative size of the variance (+++++ very large , + very small).

	Four-group model		Five-group model		Six-group model	
	Variance	Systematic effect	Variance	Systematic effect	Variance	Systematic effect
Group 1	+++++	No	+++++	No	+++++	No
Group 2	+	No	+	No	++	No
Group 3	+++	No	++++	No	++++	No
Group 4	+++	Yes	+++	Yes	+++	Yes
Group 5			++	Yes	+	No
Group 6					++++	Yes

5.2 Results

With respect to the optimal number of groups, all three years gave the same results: the six-group model was preferred according to AIC and BIC, while the five-group model was preferred according to the ICL-BIC.

We used the following procedure to analyse how well the model could detect the seasonal patterns: we determined the ratio between the total absolute survey (Y_k) and the VAT (X_k) turnover per quarter of the year, without and with adjustment for the quarterly effects detected by the model. We used the estimated $\boldsymbol{\delta}$ –values of the groups with a quarterly effect to derive adjusted quarterly VAT turnover values from the original ones X_{ki} and their corresponding adjusted totals, denoted by $\tilde{X}_k$. For totals, the turnovers are weighted using the post-stratum weights (w_{ki}).

In **Figure 5.2-1**, the Y_k/X_k ratios are displayed for the six-groups model, for the three years separately. These ratios were determined for both groups with systematic differences, for all groups (indicated by a colour), and for adjusted and unadjusted VAT turnover (indicated by solid and dashed lines, respectively). In all cases, the adjustment of X_{ki} is only applied to the group(s) with the systematic differences. In these plots, the Y_k/X_k ratios with unadjusted VAT turnover determined for all groups showed two types of differences between the two time series. Firstly, the Y_k/X_k ratios were not equal to one, but smaller than one. This indicates that $X_k > Y_k$, that is, the VAT turnover totals were larger than the survey turnover totals which corresponds to a level difference. Secondly, the Y_k/X_k ratios deviated from a flat line: the ratios varied per quarter which corresponds to quarterly differences. So, for our application, we aimed to achieve that the Y_k/X_k after adjustment would be more horizontal than the unadjusted ratios. **Figure 5.2-1** shows that the adjustment of VAT turnover indeed resulted in smaller quarterly effects. In other words, the adjustment resulted in a more stable pattern of the Y_k/X_k ratios within the year, which means that the current mixture model explained part of the seasonal effects. Especially in 2014, we found that the second group (the green line) had a clear seasonal effect, which was well adjusted by the model.

Figure 5.2-1
The Y_k/X_k ratios based on the six-group model for Job placement in 2014–2016. The ratios are determined for both groups with systematic differences and for all groups, and for adjusted and unadjusted VAT turnover.

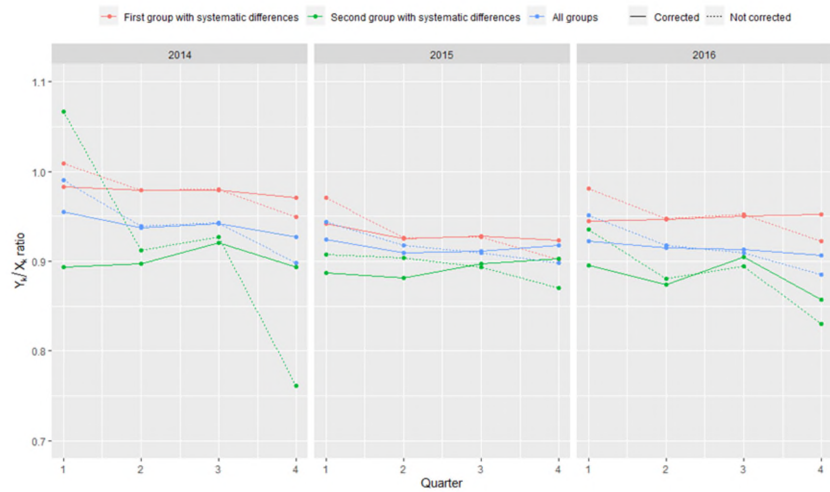
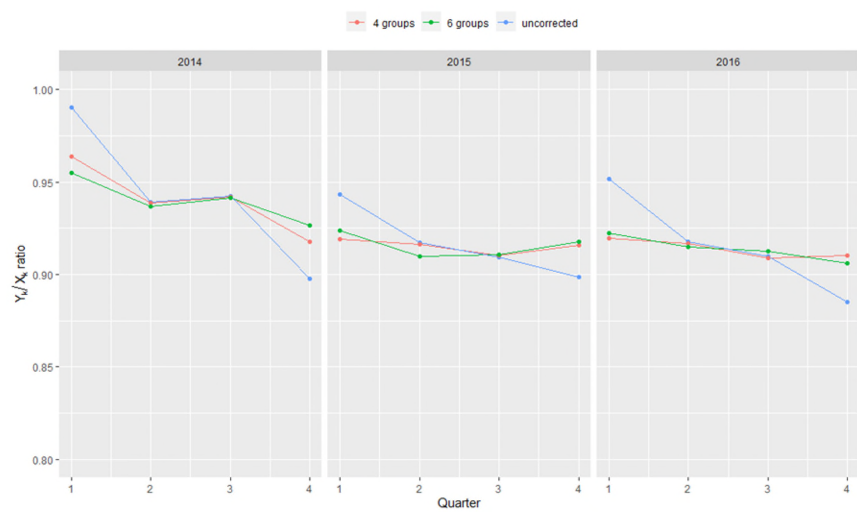


Figure 5.2-2
The Y_k/X_k ratios with adjusted VAT turnover for Job placement in 2014-2016 for the four- and six-group model. Moreover, the Y_k/X_k ratios with unadjusted VAT turnover are included.



Next, we compared the Y_k/X_k ratios of the four-, five-, and six-group models. In **Figure 5.2-2**, we only displayed the ratios of the four- and six-group models as these models could explain more of the seasonal effects than the five-group model. Here, the ratios were determined for all groups. For both models, the adjustment of X_{ki} led to lines with less deviations between quarters than the line of the unadjusted VAT turnover. Thus, both models explained part of the seasonal effects. Especially in 2014, the six-group model outperformed the four-group model.

6. Conclusions and future work

We developed a mixture model using the relative quarterly turnover distribution within a year as input to explain (and correct) seasonal differences between the VAT and the survey turnover distribution. We also applied an ECM algorithm to estimate this model. Based on simulation studies we conclude that the ECM algorithm performs well: under the conditions tested and when given reasonable starting values, it recovers the true parameter values and the correct number of groups. When the algorithm is initiated with poor starting values, it can end up in local maxima in which case the fit measures are less optimal. It is therefore important to initiate the algorithm with multiple starting values, especially for the parameters κ_g. We applied the model to real data of the economic sector Job placement and found that the model could indeed explain a considerable part of the seasonal effects. Although the model we presented was tailored to our specific application, we believe that the approach can also be of use for other situations where one uses sources with both random and structural intra-annual measurement errors.

There are a number of next steps before we can use the model in a statistical production process. First of all, we want to apply the model to other economic sectors and to more recent years. Second, we want to investigate whether the probability of a unit to belong to a certain group (i.e., the values predicted by the current ECM algorithm) can be predicted by a new model, using available register data, such as the relative VAT distribution. We could then use this new model to predict the group membership for *all* units of the population and subsequently compute corrected VAT turnover values at micro-level. These could be used to derive adjusted turnover indices for census data after which we could compare benchmarking the monthly series to the adjusted versus the original quarterly census indices. Finally, we are interested to find out which of the two mixture models, the one based on relative or on absolute turnover values is most useful in describing the seasonal patterns.

References

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, London, UK: Chapman & Hall.

Bikker, R.P., J. Daalmans, and N. Mushkudiani (2013), “Benchmarking Large Accounting Frameworks: a Generalised Multivariate Model”, *Economic Systems Research*, 25, pp. 390-408.

Daalmans, J. (2018), “Kwartaalinpassing KICR DRT: pro rata versus Denton”, internal report (in Dutch), The Hague: Statistics Netherlands.

Van Delden, A., and S. Scholtus (2017), “Correspondence between survey and administrative data on quarterly turnover”, discussion paper 2017-3, The Hague: Statistics Netherlands. Available at https://www.cbs.nl/en-gb/background/2017/07/correspondence-between-survey-and-admin-data-on-quarterly-turnover.

Van Delden, A., and S. Scholtus (2019), “Analysing response differences between sample survey and VAT turnover”, discussion paper, The Hague: Statistics Netherlands. Available at https://www.cbs.nl/en-gb/background/2019/22/analysing-response-differences-in-vat.

Van Delden, A., J. Pannekoek, R. Banning, and A. de Boer (2016), “Analysing correspondence between administrative and survey data”, *Statistical Journal of the IAOS*, 32, pp. 569-584.

Van Delden, A., S. Scholtus, and N. Ostlund (2020), "Modelling measurement errors to enable consistency between monthly and quarterly turnover growth rates", *Proceedings of the 2018 International Methodology Symposium*, *Statistics Canada*. Available at https://www.statcan.gc.ca/eng/conferences/symposium2018/program.

Denton, F.T. (1971), "Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization", *Journal of the American Statistical Association*, 66, pp. 99-102.

Di Zio, M., and U. Guarnera (2013), "A Contamination Model for Selective Editing", *Journal of Official Statistics*, 29, pp. 539-555.

Lewis, D. and Woods, J. (2013), Issues to consider when turning to the use of administrative data: the UK experience. Paper presented at the New Techniques and Technologies for Statistics (NTTS) conference 2013. Available at https://ec.europa.eu/eurostat/cros/system/files/NTTS2013fullPaper_142.pdf

McLachlan, G.J., and D. Peel (2000), *Finite Mixture Models*, New York: John Wiley & Sons.

Țîru, A.M., Pop, I. and Oancea, B. (2019), Error detection and data imputation methods for administrative data sources used in business statistics. Paper presented at the ESS workshop on administrative data for business, agriculture and fisheries statistics in Bucharest. Romania, 17-18 October 2019.