TI 2020-078/III
Tinbergen Institute Discussion Paper

Dynamic Factor Models with Clustered Loadings: Forecasting Education Flows using Unemployment Data

Revision: January 2021

Francisco Blasques[1,2]
Meindert Heres Hoogerkamp[3]
Siem Jan Koopman[1,2]
Ilka van de Werve[1,2]

[1] Vrije Universiteit Amsterdam
[2] Tinbergen Institute
[3] Dutch Ministry of Education, Culture and Science

Dynamic Factor Models with Clustered Loadings: Forecasting Education Flows using Unemployment Data

Francisco Blasques[1,2,*], Meindert Heres Hoogerkamp[1,3,4],
Siem Jan Koopman[1,2] and Ilka van de Werve[1,2,5]

Abstract

We propose a dynamic factor model which we use to analyze the relationship between education participation and national unemployment, as well as to forecast the number of students across the many different types of education. By clustering the factor loadings associated with the dynamic macroeconomic factor, we can measure to what extent the different types of education exhibit similarities in their relationship with macroeconomic cycles. To utilize the feature that unemployment data is available for a longer time period than our detailed education panel data, we propose a two-step procedure. First, we consider a score-driven model which filters the conditional expectation of the unemployment rate. Second, we consider a multivariate model in which we regress the number of students on the dynamic macroeconomic factor, and we further apply the k-means method to estimate the clustered loading matrix. In a Monte Carlo study, we analyze the performance of the proposed procedure in its ability to accurately capture clusters and preserve or enhance forecasting accuracy. For a high-dimensional, nation-wide data set from the Netherlands, we empirically investigate the impact of the rate of unemployment on choices in education over time. Our analysis confirms that the number of students in part-time education covaries more strongly with unemployment than those in full-time education.

Keywords: Dynamic Factor Models, Cluster Analysis, Forecasting, Education, Unemployment

JEL codes: C38, C53, I25

[1]Vrije Universiteit Amsterdam [2]Tinbergen Institute [3]Dutch Ministry of Education, Culture and Science [4]Dienst Uitvoering Onderwijs (DUO) [5]Netherlands Institute for the Study of Crime and Law Enforcement (NSCR)

[*]F. Blasques is thankful to the Dutch Science Foundation (NWO) for financial support (VI.Vidi.195.099). We would like to thank the handling Editors, the Associate Editor and the two referees for their careful reading of earlier versions of our paper and their many insightful comments.

1 Introduction

Quality education is one of the Sustainable Development Goals of the United Nations. For example, the Dutch government has allocated roughly 11% of its total expenditures in the national budget of 2019 towards education[1,2]. It illustrates the importance of education to our society. To secure a reliable budgetary policy, the Dutch government forecasts the numbers of students in each type of education on a nation-wide level. Although education systems are complex, dynamic and evolving, accurate forecasts are of key importance for the overall operational and financial planning, but also for providing insights into what drives participation in education. For the purpose of fiscal policy, it is valuable to understand the interaction between education participation and macroeconomic circumstances.

Although Spijkerman (2006) did not find a strong relation between macroeconomic indicators and the total number of students, macroeconomic circumstances do seem to affect the demand for certain types of education. In particular, the share of part-time education appears to be inversely related with unemployment rates, especially in vocational education. This analysis is relevant because educational institutions receive less funding for part-time students compared to full-time students. However, whether or not distinct groups react differently to macroeconomic circumstances has not been studied in full. The more recent availability of low-level data allows us to revisit this research question. Our data set for vocational and higher education is high-dimensional on the cross-section but low-dimensional on the time series. The models for such panel data sets typically strike a balance between interpretability and performance. However, to policy makers in government and educational institutions, both interpretability and performance are of interest.

In this paper we develop a dynamic factor model where unemployment rates and education participation are modeled simultaneously. The model consists of two components. First, we focus on the time series dimension and model the historical unemployment rates through a score-driven local level model as proposed by Creal, Koopman, and Lucas (2013). By using all available observations of the unemployment rate, we get more reliable results when filtering the factor than when we would use the few observations that correspond to the same years as that the education data is also available. Second, in the cross-section dimension of our methodology, we anticipate that

[1] `https://www.rijksfinancien.nl/visuele-begroting/2018/owb/u` (visual in Dutch), last accessed 2020-04-30.
[2] `https://www.rijksoverheid.nl/onderwerpen/financiering-onderwijs/overheidsfinanciering-onderwijs` (in Dutch), last accessed 2020-04-30.

many education flows respond in a similar way to changes in the unemployment rate. We take the extracted dynamic economic factor as given and model the education data set effectively as a multivariate regression model. We also propose to cluster their dependence on the dynamic economic factor through the parameters of the loading matrix. It imposes a structure on the model that benefits interpretation. Moreover, since we represent many education series by a couple of cluster centroids, it is also more efficient with respect to forecasting education participation.

In accordance with these two components of our dynamic factor model, we propose a two-step estimation procedure. First, in the score-driven local level model for the unemployment rates, we estimate the static parameters by maximum likelihood and extract the dynamic economic factor. This step is important to filter out the noise and preserve the signal in economic data such as the unemployment rate. Second, given the extracted dynamic economic factor, we estimate the multivariate regression model of the education data by the method of least squares. To gain insights into what types of education respond similarly to changes in unemployment rates, we perform a cluster analysis. By using the k-means method, we are able to represent all loading matrix elements by a few cluster centroids. This ability implies that cluster analysis can support the testing of joint significance for a group of variables without pre-imposition of group compositions. At the same time, we avoid the possible insignificance of individual variables, given that the education data for each variable is limited as the time series dimension is small. Once the clusters are identified, we can provide accurate forecasts for all series in the panel.

Dynamic factor models are well-suited to extract common factors from large data sets, see Stock and Watson (2002), Bai and Ng (2002) and Jungbacker and Koopman (2015) amongst others. It has become more prevalent to estimate the parameters in dynamic factor models using a step-wise approach. For example, Doz, Giannone, and Reichlin (2011) first proxy the factors by principal components to estimate the static parameters and then use Kalman filter techniques for signal extraction and forecasting. Bräuning and Koopman (2014) take a slightly different approach, they also first use a principal component analysis to obtain a dimension reduction, but then model all relevant variables jointly in a state space framework such that parameter estimation, signal extraction and forecasting are done by Kalman filter methods. The approaches in both papers are parameter-driven in which the stochastic processes of the factors have their own sources of error.

Our procedure differs in adopting an observation-driven approach: we allow the factors to evolve as dynamic processes which are formulated as functions of past data. In particular, we adopt the approach taken by Creal et al. (2013) and Harvey (2013) where the dynamic specification is based on autoregressive processes with the innovations defined as score functions with respect to the predictive likelihood function. We first model the unemployment rate data as a score-driven model to filter the dynamic factor, which we then consider as given in the second step of our proposed estimation procedure to estimate the model for the education data. Unrestricted parameter estimates of the loading matrix will then be clustered to cluster types of education according to their dependence on the unemployment rate. Just as Stock and Watson (2008) do, we use the k-means algorithm in the cluster analysis, although their approach differs in that they base it on the residuals of the dynamic factor model. In our case, we represent the large vector of unrestricted loading estimates by a much smaller vector of cluster centroids. The introduction of clustering within a dynamic factor model has more recently also been explored by Hallin and Liška (2011), Barnichon and Mesters (2018) and Alonso, Galeano, and Peña (2020). Similarly, Ando and Bai (2016) incorporate the k-means procedure within the dynamic factor model and they also adopt an estimation procedure that consists of several steps.

In an extensive simulation study we assess the overall performane of our proposed estimation procedure and the accuracy of the forecasts. We discuss the most important insights from this study. First of all, the estimation of the parameters in our dynamic factor model show very small biases while the clustered pattern in the loading matrix is correctly identified. Secondly, we can report high levels of forecast accuracy. We also find that the forecasts based on the clusters perform are at least as good as the unrestricted forecasts. Indeed, we present improvements in forecast accuracy when a clear clustered structure is present in the data. Even for cases where clusters are less present in the data, we do not lose much on forecasting accuracy while we gain much on interpretation.

The remainder of this paper is organized as follows. We describe the education data and unemployment rate data in Section 2. In Section 3 we introduce the dynamic factor model and we discuss the features of the model, the two-step estimation procedure and the clustering method for forecasting. Section 4 describes the design and results of our Monte Carlo study. We apply the methodology to our education participation data in Section 5 and show that the proposed methodology can capture important clusters in the flows across the Dutch education system. Section 6 concludes and reviews the most important findings of our study.

2 Data

In 2019, 3.7 million students are registered at a Dutch educational institution: around 1.5 million (40.2%) of which in primary education, 960 thousand (25.6%) in secundary education, 500 thousand (13.4%) in vocational education, 460 thousand (12.4%) in higher vocational education (university of applied sciences), and 300 thousand (8.2%) in university. For this study, we use data from DUO, the executive agency for the Dutch Ministry of Education, Culture and Science. Student registrations are observed on October 1st each year (reference date). The primary use of this data set is the Dutch student forecasts (Ministry of Education, Culture and Science, 2020) that feed the governmental education budget. Everyone in the Dutch population not in education, is labeled 'outside education'. Using information from two consecutive reference dates, flows through the educational system are constructed. The educational data set lists the number of people in each flow from one state to another, by age and sex, in a given year. Between these dates, people might obtain a diploma. Those people are said to transition from an origin state to a diploma state, and then from this diploma state to a destination state in or outside of education. Each state (one of 820 education types, 173 diploma types or no education) has up to 5 descriptive labels: a sector (such as vocational education, university), type (such as on-the-job training, full-time), level (such as bachelor, master), direction (such as health care, economics), and grade. Not all flows between the states are viable (for example, one cannot move from university to primary education) or have not been observed. 326,000 unique transitions from origin to destination have been observed since 2005. Data from 2006 to 2019 is used.

Flows at the lowest level are filtered, aggregated and transformed to form the data set for this study. We are interested in the short-term relation between unemployment rates and inflow into first year of vocational and higher education. We include all inflow from no education and diploma categories. The origin states are aggregated by sector, category, and type. Furthermore, to reduce noise, some ages (those that contain fewer first-year students) are combined as follows: <17, 23-25, 26-30, 31-40, >40. In 2019, each age bin contains a total of between 6,900 and 60,600 students. This leaves $N = 1,155$ time series. This is a short, wide panel (large N, and small T), implying that many series have to be forecasted while limited data on the time dimension is available.

The average level of the 1,155 time series is 240 persons per year, with a minimum of 14.9 (22 year old males moving from no education to fulltime school-based level 2 vocational education direction technics/science) and a maximum of 5,658 (<17 year old females moving from prepratory vocational education diploma to fulltime school-based level 4 vocational education direction healthcare). 50%

count	1155
mean	239.6
std	487.3
min	14.9
25%	38.3
50%	82.8
75%	206.8
max	5658.0

Table 1: Descriptive statistics of average level of 1,155 time series.

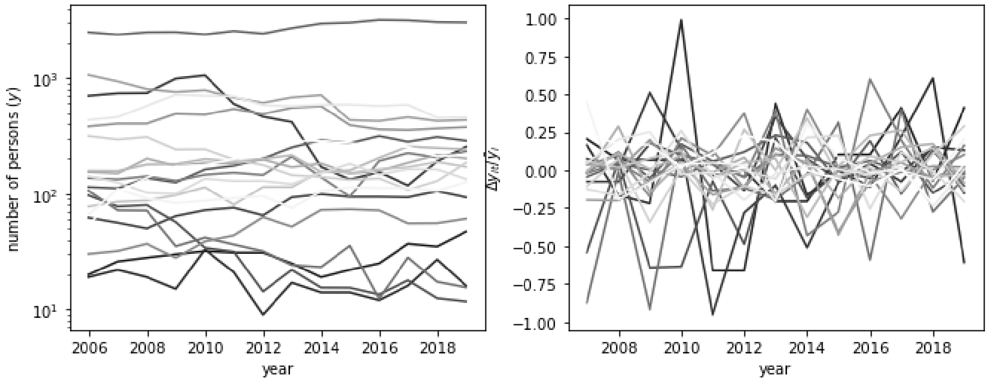


Figure 1: Sample of 20 time series. Left: levels (log-scaled vertical axis), right transformed variable.

of the time series contain on average between 38.3 and 206.8 persons per year (see Table 1). To illustrate, a random sample of 20 time series is plotted in Figure 1. The left panel presents the actual time series (on a log-scale) which may not be all stationary, some may be subject to trend behaviour. Transitions with a higher number of students on average have a higher variance. To normalize, each cross-sectional unit is divided by its average level: $\tilde{y}_{it} = \Delta y_{it}/\bar{y}_i$, with y_{it} the number of people in transition i at time t. We assume that the transformed variables are stationary. Using the KPSS test we find that only the null hypothesis of mean stationarity can be rejected ($p < .05$) for only 15 (1.2%) time series. These are still included, since one expects some rejections under the null when many tests are conducted. Most series show positive autocorrelation at lag the first lags (Figure 2). This fits the autoregressive model that will be developed in section 3.

We enrich the data by including historical macro-level data on the Dutch economy. Yearly unemployment rates (operationalized as unemployed labor force divided by total labor force, ages 15-74) are obtained from 1970 to 2019 (Bureau for Economic Policy Analysis, 2019). In our modeling, we will make use of the fact that the macroeconomic data is available for a much longer period than the education data. Figure 3 provides the instigation for this particular research. Spijkerman

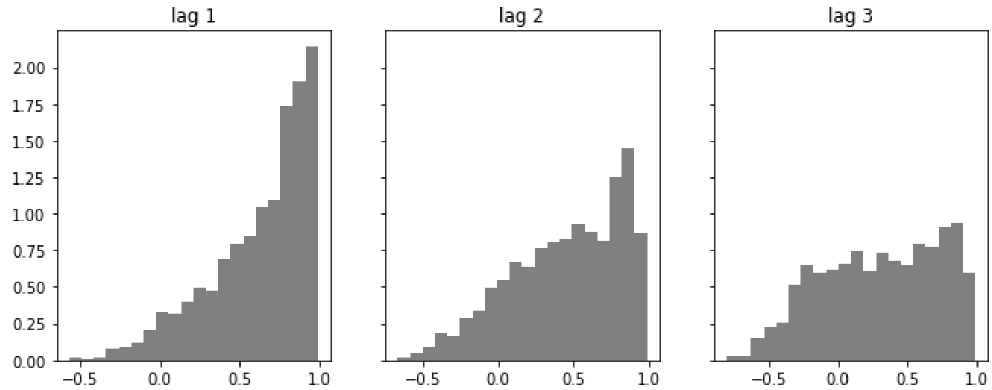
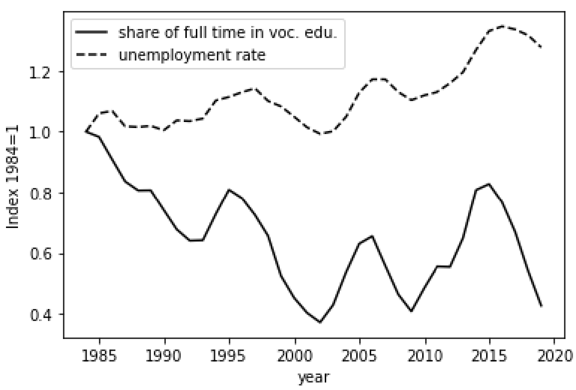
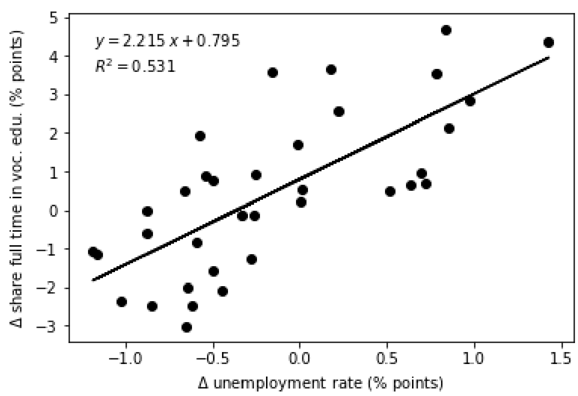


Figure 2: Histogram of autocorrelations for all time series at first three lags. At lag 1, most time series show positive autocorrelation. At higher lags, the estmates spread out more evenly and the mean decreases.

(2006) did not find a strong correlation between unemployment and nationwide educational enrolment, but he did show that the share of full-time education covaries positively with unemployment, especially in vocational studies. The left panel shows that peaks and valleys in the share of full time in vocational studies indeed correspond with the labor market cycle. The right-side panel suggests a linear relationship in first differences.



(a) Share of full-time in vocational education and unemployment rate (indexed at 1984=1)



(b) Increase in unemployment (x-axis) vs. share of full time in vocational education (y-axis)

Figure 3

In accordance with the scatterplot in Figure 3 the first differences of both datasets are calculated. This also prevents regressing possibly (co)integrated time series. Our macro-economic time series has dimension $T_x = 49$, and the education panel has dimension $T_y = 13$, where the last T_y observations of both series refer to the same years.

3 Methodology for Modeling and Forecasting

Our education panel data set consists of many possible education flows which can be selected on the basis of gender and age groups. We want to understand to which extent a macroeconomic variable (the unemployment rate) can help us to forecast the number of students for each category, for a number of years ahead, despite that we only have data available for the last thirteen years. The forecasting method is oftentimes regarded as more convincing when the model preserves a level of interpretability for policy purposes. We therefore develop a dynamic factor model where types of education are clustered based on their dependence on changes in the unemployment rate.

3.1 The Dynamic Factor Modeling Framework

Let y_t be the N-dimensional series of education flows in year t. The basic dynamic factor model is given by

$$y_t = \Lambda f_t + \varepsilon_t, \qquad t = 1, \ldots, T_y, \tag{1}$$

where Λ is a $N \times \ell$ loading matrix, $\ell \times 1$ vector f_t an unobserved factor and the sequence $\varepsilon_1, \ldots, \varepsilon_{T_y}$ is independent and identically Gaussian distributed with mean zero and variance matrix $\sigma_\varepsilon^2 I_N$. Since the unobserved factor in our model is a proxy for the macroeconomic circumstances (measured by the unemployment rates), we have $\ell = 1$, implying that the loading matrix Λ is effectively a vector and f_t a scalar. A vector of unit-specific intercepts $\mu = (\mu_1, \ldots, \mu_N)'$ can be added to the model, we then have $y_t = \mu + \Lambda f_t + \varepsilon_t$, but to facilitate the clustered forecasting method in our analysis, we assume that the data is demeaned and we have $\mu = 0$ without loss of generality.

To emphasize that the education and macroeconomic data have different time series dimensions, we introduce the following notation. The index t and time series dimension T_y correspond to the education data denoted by y_t, while index $\bar{t}$ and time series dimension T_x correspond to the unemployment rate data denoted by $x_{\bar{t}}$. More specifically, the macroeconomic data $x_{\bar{t}}$ is available at $\bar{t} \in \{1970, \ldots, 2019\}$ (so $T_x = 49$ years) while the education data y_t is available at $t \in \{2006, \ldots, 2019\}$ (so $T_y = 13$ years). The macroeconomic data is available over a longer time-span so that only its last T_y time periods coincide with the availability of the education data.

We let $x_{\bar{t}}$ be the growth in the unemployment rate in year $\bar{t}$; this time series of yearly observations has dimension T_x. The location model is given by

$$x_{\bar{t}} = f_{\bar{t}} + \xi_{\bar{t}}, \qquad \bar{t} = 1, \dots, T_x, \tag{2}$$

where the signal $f_{\bar{t}}$ can be regarded as the time-varying mean of the observed time series $x_{\bar{t}}$, and where $\xi_1, \dots, \xi_{T_x}$ is assumed to be an independent and identically distributed Gaussian sequence with mean zero and variance σ_ξ^2.

In our analysis, we take $x_{\bar{t}}$ as the yearly unemployment rate, which we consider to be a proxy of general macroeconomic circumstances. However, we note that we could generalize the model by allowing for multiple economic indicators (multivariate $x_{\bar{t}}$) and then still easily filter $\ell \geq 1$ factors f_t from it. This imposes extra structure on the model that requires extra assumptions to identify f_t. We leave this extension for future research.

Equation (2) enables us to estimate the signal $f_{\bar{t}}$ in our modeling framework. We use the extracted signal for the estimation of loading parameters in Λ of equation (1) and for the forecasting of time series variables in y_t of equation (1). For the filtering of the factor, we adopt the score-driven model as introduced by Creal et al. (2013) and Harvey (2013). In an observation-driven approach, we introduce

$$\mathcal{X}_{\bar{t}-1} = \{x_1, \dots, x_{\bar{t}-1}\} = \{\{x_1, \dots, x_{\bar{t}-1}\}, \{f_1, \dots, f_{\bar{t}-1}\}\},$$

so that the information set at time $\bar{t}$ is generated by $\{f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}\}$. Since the location model for $x_{\bar{t}}$ in equation (2) is linear Gaussian, we have $x_{\bar{t}} \sim p(x_{\bar{t}}|f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}; \theta)$, where $p(\cdot\,|f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}; \theta)$ is the univariate Gaussian density with mean $f_{\bar{t}}$, variance σ_ξ^2, and parameter vector θ. The unknown variance σ_ξ^2 is placed in the parameter vector θ, together with the unknown coefficients that we introduce below.

We follow Creal et al. (2013) in their formulation of the filtering or updating equations for $f_{\bar{t}}$; these are referred to as the generalized autoregressive score (GAS) model and are given by

$$\begin{aligned} f_{\bar{t}+1} &= \omega + \sum_{i=1}^{p} \phi_i f_{\bar{t}+1-i} + \sum_{j=1}^{q} \alpha_j s_{\bar{t}+1-j}, \\ s_{\bar{t}} &= S_{\bar{t}} \cdot \nabla_{\bar{t}}, \qquad S_{\bar{t}} = S(\bar{t}, f_{\bar{t}}, \mathcal{X}_{\bar{t}}), \qquad \nabla_{\bar{t}} = \frac{\partial \log p(x_{\bar{t}}|f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}; \theta)}{\partial f_{\bar{t}}}, \end{aligned} \tag{3}$$

where ω is the intercept, $\phi_1, \dots, \phi_p$ and $\alpha_1, \dots, \alpha_q$ are the weight coefficients for the updating mechanism of $f_{\bar{t}+1}$, and $s_{\bar{t}}$ is the scaled score with local score function $\nabla_{\bar{t}}$ and scaling term $S_{\bar{t}}$. Typically, we base the scaling on a variance measure of the score function. We refer to this score-driven model by GAS(p,q), where the integers $p \geq 0$ and $q \geq 0$ can be chosen on the basis of fit, residual diagnostics and forecast performance considerations. In many cases, it is sufficient to take $p = q = 1$ and we have the GAS(1,1) model.

Given that $p(\cdot|f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}; \theta)$ is the univariate Gaussian density with the time-varying mean (or location) $f_{\bar{t}}$, we have that

$$\begin{aligned}
\log p(x_{\bar{t}}|f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}; \theta) &= -\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\log \sigma_\xi^2 - \tfrac{1}{2}(x_{\bar{t}} - f_{\bar{t}})^2/\sigma_\xi^2, \\
\nabla_{\bar{t}} &= (x_{\bar{t}} - f_{\bar{t}})/\sigma_\xi^2, \\
S_{\bar{t}} = \mathcal{I}_{\bar{t}}^{-1} &= \left(-\frac{\partial^2 \log p(x_{\bar{t}}|f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}; \theta)}{\partial f_{\bar{t}}^2}\right)^{-1} = \sigma_\xi^2, \\
s_{\bar{t}} &= \sigma_\xi^2 \cdot (x_{\bar{t}} - f_{\bar{t}})/\sigma_\xi^2 = x_{\bar{t}} - f_{\bar{t}},
\end{aligned}$$

with parameter vector $\theta = \left(\omega\,,\, \phi_1\,, \ldots,\, \phi_p\,,\, \alpha_1\,, \ldots, \alpha_q\,,\, \sigma_\xi^2\right)'$. By placing the above elements in the filtering equation (3), it becomes apparent that the factor f_t depends on previous years of unemployment. It is further implied by the observation equation (1) that education participation is allowed to depend on past unemployment rates through the factor f_t. An alternative approach is to model these dynamics explicitly by allowing lags of the factor in the observation equation, but we leave this extension for future research.

Our dynamic factor modeling framework is represented by the equations (1), (2) and (3). It facilitates the linkage of the two available data sets (the education panel data y_t and the unemployment rate time series $x_{\bar{t}}$) and it provides feasible methods for parameter estimation and forecasting.

3.2 The Two-Step Estimation Procedure

We propose a two-step estimation procedure for our dynamic factor modeling framework given by equations (1), (2) and (3). In the first step, we focus on the time series component and use the unemployment rate time series to estimate the score-driven model of equations (2) and (3) using the method of maximum likelihood. The static parameters in $\theta = (\omega\,,\, \phi_1, \ldots, \phi_p\,,\, \alpha_1\,, \ldots,\, \alpha_q\,,\, \sigma_\xi^2)'$ are estimated via the maximization of the loglikelihood function as given by

$$\hat{\theta} = \arg\max_{\theta} T_x^{-1} \sum_{\bar{t}=1}^{T_x} \left(-\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\log \sigma_\xi^2 - \tfrac{1}{2}(x_{\bar{t}} - f_{\bar{t}})^2/\sigma_\xi^2\right),$$

where $f_{\bar{1}}$ is initialized by setting it equal to the unconditional mean $\omega/(1 - \sum_{i=1}^{p} \phi_i)$, and where $f_{\bar{t}}$ is obtained from the GAS filter (3), for a given θ and $\bar{t} = 1, \ldots, T_x$. When the maximum likelihood estimate $\widehat{\theta}$ is obtained, we denote the factors obtained from the GAS filter (3) with $\theta = \widehat{\theta}$ by $\hat{f}_{\bar{t}}$, with $\bar{t}$ as before. We consider $\hat{f}_{\bar{t}}$ as a proxy of macroeconomic circumstances.

In the second step we consider the education data, focus on the cross-section component, and view equation (1) as a multivariate regression model. We recall that the time series dimension of the unemployment data $x_{\bar{t}}$ is longer than of the education data y_t, so we replace the factors f_t by those estimated in the first step and only keep the last T_y periods such that $\hat{f}_t \equiv \hat{f}_{\bar{t}}$. The loadings in equation (1) are estimated by the method of least squares. In this way we obtain an estimate of the variance σ_ε^2 and the unrestricted estimate of the loading matrix as denoted by $\tilde{\Lambda} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N)'$. Next we carry out a cluster analysis on the estimated column representing the loading matrix.

Cluster analysis is often used in statistics and machine learning to partition many individuals, cities or products into groups. As part of the second step in our estimation procedure, we are proposing to cluster the elements of the loading matrix Λ such that it does not consist of N different elements, but of $K << N$ cluster centroids instead. It implies that we can analyze the similarities and differences between many types of education in their dependence on the macroeconomic data as measured by the dynamic factor. In addition, it also implies that the number of unique forecasts decreases considerably.

There are several possibilities to proceed with a cluster analysis. We have opted for the k-means algorithm because it provides clear interpretation and it is computationally simple. The clustering method limits the number of forecasts that needs to be produced, because we let a cluster of similar education flows be represented by one centroid. In case of the k-means algorithm, we cluster the N different and unrestricted values of column $\tilde{\Lambda}$ into K cluster centroids for $\hat{\Lambda}$ by conducting the following steps:

1. Initialize the cluster centroids randomly from the data range: draw centroids $\delta_1, \ldots, \delta_K$ from $U(\tilde{\Lambda}_{min}, \tilde{\Lambda}_{max})$, where $U(\cdot)$ is the uniform distribution

2. Obtain distances between data points and cluster centroids and add cluster labels to the data points: label $c^{(i)} = \arg\min_k ||\tilde{\lambda}_i - \delta_k||^2, \quad \forall i$

3. For each cluster, assign new centroids $\delta_k = \sum_{i=1}^{N} \mathbb{1}\{c^{(i)} = k\}\tilde{\lambda}_i \ / \ \sum_{i=1}^{N} \mathbb{1}\{c^{(i)} = k\}, \quad \forall k$

4. Verify whether the cluster centroids changed

5. Repeat steps 2-4 until convergence

6. Replicate steps 1-5 for $R = 15$ different random seeds

7. Keep clustered loading matrix $\hat{\Lambda}$ with shortest total distance to the cluster centroids

In our empirical study, the number of clusters is not known beforehand. We therefore run this algorithm for several values of K and proceed heuristically by using the simple elbow method. The elbow method is a graphical representation of plotting the performance against several cluster sizes, where performance is usually measured as a distance metric. The typical pattern is a sharp decreasing distance for cluster sizes $1, \ldots, K$, with K small. For cluster sizes $K+1, K+2, \ldots$, the returns diminish. One then selects the kink or elbow in the graph as optimal cluster size, because decreasing it would lead to a big performance loss while increasing it would lead to a negligible improvement.

The resulting small vector of cluster centroids facilitates interpretation for policy makers. It gives insights on which education flows covary similarly with changes in the unemployment rate. We want to emphasize that we do not attach any causal interpretation to it because the regressor might be endogenous because of reverse causality. If we assume a simple labor supply model, then people spend time in either education or employment. So, it might well be that more education participation leads to higher unemployment rates. This is the reverse relationship as in our model where we try to explain education by unemployment, so we emphasize that we measure associations in our current study.

To show the properties of the estimators, we need to take into account that we follow a two-step estimation approach. In the first step, we estimate the factor f_t by maximum likelihood, using a standard score-driven filter for the conditional expectation. Consequently, the asymptotic properties follow the usual results covered in the score-driven literature; see, for example, Blasques, Koopman, and Lucas (2014a) and Blasques, Gorgi, Koopman, and Wintenberger (2018). In the second step, we take the filtered f_t as given and attempt to estimate the unrestricted loadings in a simple linear regression model where the filtered f_t is already observed. The asymptotics of the second step estimator then follow standard regression conditions. The estimated unrestricted loadings are then clustered using the k-means algorithm. In Appendix A, we discuss the asymptotic properties in more detail. The finite-sample performance of the k-means clustering is analyzed by the Monte Carlo study below. We leave the theoretical characterization of the estimators in our two-step method for future research.

3.3 The Clustered Forecasting Method

After the two-step estimation procedure, we use the estimated static parameters and filtered time-varying factor of the score-driven model to forecast future values of the factor in a recursive manner from equation (3). Next, together with the estimated clustered loading matrix, we forecast future education participation. Since the clustered loading matrix $\hat{\Lambda}$ consists of K distinct values only, we just need to forecast $K << N$ series. This leads to computation time savings because we have thousands of education flows for each combination of age and gender.

4 Monte Carlo Study

We carry out a Monte Carlo study to verify the performance of our estimation and measurement methodology. For this study, we consider the dynamic factor model as given by the equations (1), (2) and (3) with the GAS updating orders $p = 1$ and $q = 1$, that is, the GAS(1,1) specification. The dimensions in our simulation design are motivated by the empirical problem at hand. Hence we set the cross-section dimension of y_t to be relatively large and the time series dimension to be relatively small. In particular, we have $N = 500$ and $T_y \in \{10, 20\}$. We set the time series dimension of $x_{\bar{t}}$ to be moderate, with $T_x \in \{50, 100\}$. The last T_y time units of $x_{\bar{t}}$ are equal to the T_y time units of y_t. To verify the forecasting performance, we set the forecast horizon to be $F = 3$ periods ahead. Moreover, we take the static parameters in the score-driven model as $\sigma_\xi = 1, \omega = 0.3, \phi = 0.95$ and $\alpha = 0.1$, making the unconditional mean of the process for the stationary factors equal to $\omega/(1 - \phi) = 6$. The five equally-sized clusters of the loading matrix have centroids 4, 11, 19, 23 and 35. This implies that the observations roughly vary between $4 \times 6 = 24$ and $35 \times 6 = 210$. Finally, we set the error variance to follow from $\sigma_\varepsilon = 20$. The reported Monte Carlo results in our study are based on $M = 1,000$ simulations, for the different model specifications and data dimensions.

The specific choices of the static parameters and cluster centroids are for illustrative purposes. We do need to assume that there is some underlying form of clustering in the data present and we can visualize that by taking a non-zero unconditional mean of the factors and distinct choices of the centroids. However, the behavior over time can vary between the series within a cluster. To visualize that, we present a couple of example time series with varying variance σ_ε^2 in Figure 4 for $T_x = 100, T_y = 10$, two clusters with loadings 23 and 35 and remaining settings as above.

In these plots we see the trade-off when clustering is beneficial and when not. For a very small variance, such as the $\sigma_\varepsilon^2 = 10$ in the top-left plot, there is no real need for imposing the cluster

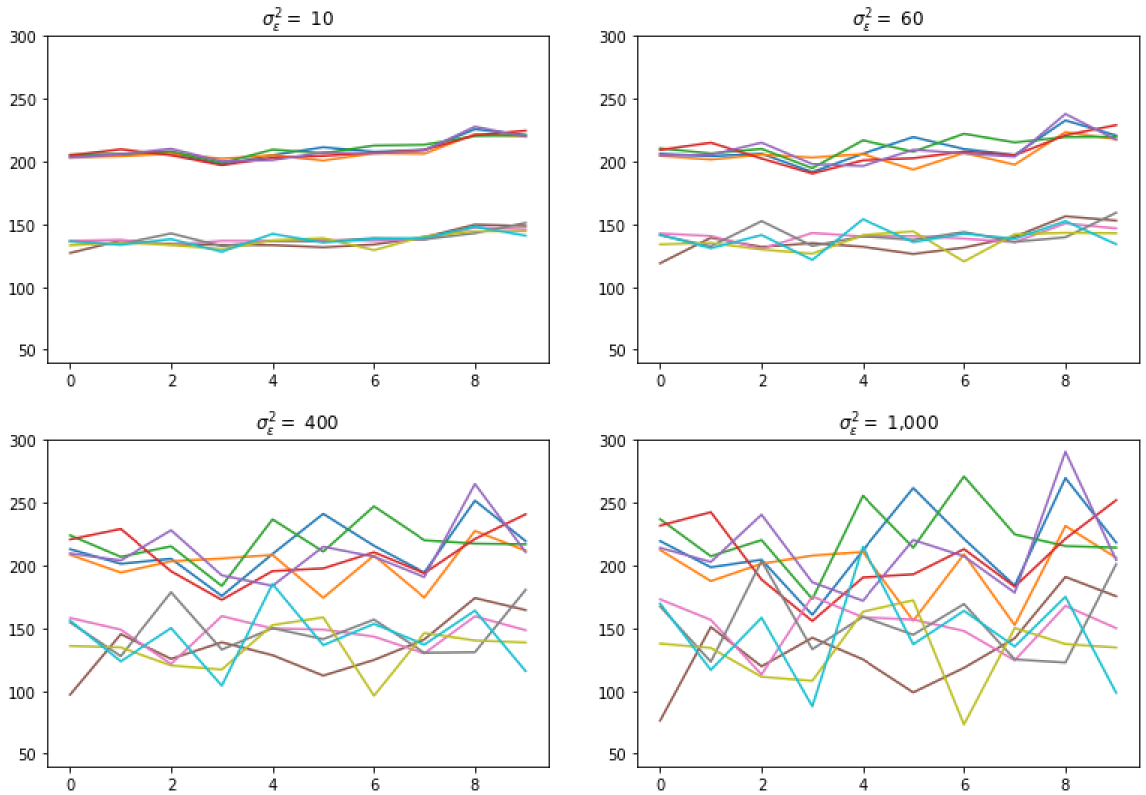


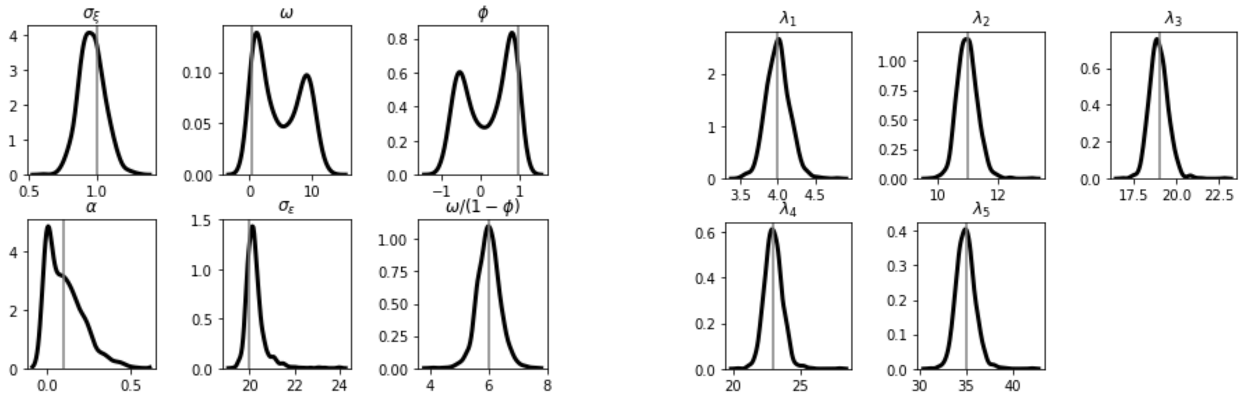
Figure 4: Example time series y_t plotted against time. The same five time series from two clusters with loadings 23 and 35 are given in all plots. Only the variance σ_ε^2 differs over the plots, the other simulation settings are kept fixed ($T_x = 100, T_y = 10, N = 500, \sigma_\xi = 1, \omega = 0.3, \phi = 0.95, \alpha = 0.1$).

analysis. By basically scanning over such data plots, one gains already the knowledge on patterns in the data. But also practically, as there is basically no variation over time, the unrestricted estimate will be as good as the clustered one. In the other extreme, where the variance is very large as in the bottom-right plot with $\sigma_\varepsilon^2 = 1,000$, clustering is also not useful because there are no clusters to really distinguish. The large variation over time will make getting an unrestricted estimate already challenging and the cluster classification is therefore also difficult. More reasonable values in-between, such as the plots with $\sigma_\varepsilon^2 \in \{60, 400\}$, show where extracting the clustered pattern can be of added value. With some variation over time, the unrestricted estimates may be a bit too far off in either direction for each of the series within a cluster, however, this is offset by using clustered estimates. In such cases, one might think that the unrestricted estimates are very different, but the clustered estimate clearly shows that their behavior is actually similar. In the discussion of the full simulation study next, we will continue with $\sigma_\varepsilon^2 = 400$.

4.1 Parameter Estimation Results

The performance of the two-step estimation procedure will be visualized by densities of the estimated parameters and cluster centroids. We will judge the clustering classification by confusion matrices[3]. Furthermore, we give in-sample statistics to compare the fit of the unrestricted and clustered models.

Figures 5 and 6 give the density plots of the estimated static parameters and cluster centroids. For both figures we take $T_y = 10$ fixed and first consider $T_x = 50$ and then $T_x = 100$. We also obtained these results for $T_y = 20$ with $T_x \in \{50, 100\}$, but since the results are very similar, we refer to the Appendix for these figures. In each figure, the plotted densities of the estimated static parameters are given in the first set of results and the plotted densities of the cluster centroids in the loading matrix in the second set of results, all based on $M = 1,000$ simulations.



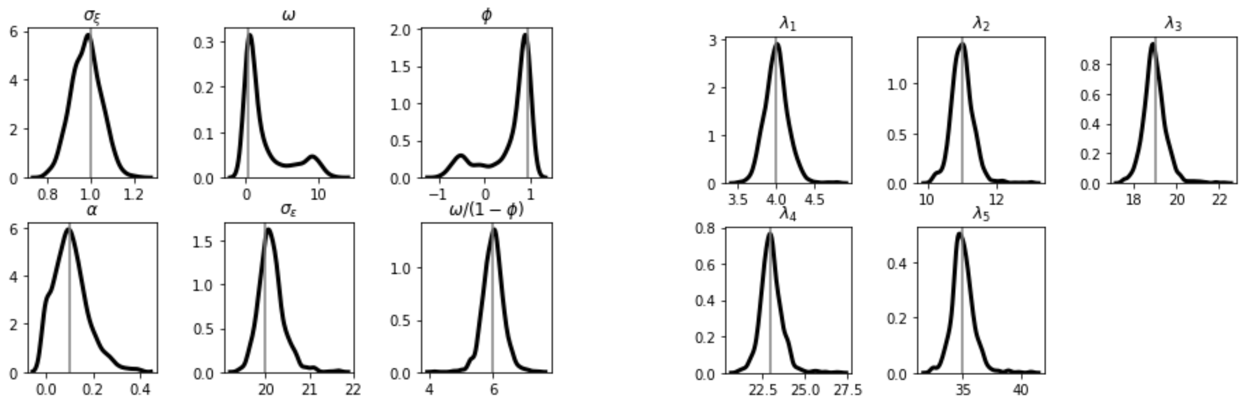
(a) Densities of estimated static parameters. (b) Densities of estimated cluster centroids.

Figure 5: Parameter estimation results of $M = 1,000$ simulations for $N = 500, T_y = 10, T_x = 50$. Vertical lines represent true values.

If we focus on the first step of our proposed estimation procedure, then we consider the static parameter vector $(\sigma_\xi, \omega, \phi, \alpha)'$. Only the time series dimension T_x is of importance for these parameters of the score-driven model. Comparing the corresponding densities in Figures 5 and 6 (or similarly comparing Figures B.1 and B.2 in the Appendix, since the varying time series dimension T_y is not relevant in this step) clearly shows that the parameter estimates become much more precise as the time series dimension T_x increases. We obtain more improvements when we take T_x even larger, but such cases do not match our empirical study, so we do not consider this further.

[3] A confusion matrix gives insight on the correct assignment to the clusters instead of the specific values of the centroids. It gives the counts of correct and incorrect assignments to the clusters with smallest, second-to-smallest, . . . , largest centroid. Perfect classification would give a diagonal matrix.

It is apparent from the density plots for $T_x = 50$, but also for $T_x = 100$ to some extent, that the densities of ω and ϕ appear to be bimodal. Since the density plot of the unconditional mean $\omega/(1-\phi)$ is unimodal around the true value, it suggests that in small samples it is challenging to empirically separate the two parameters ω and ϕ. When the time series dimension increases, our results show that the bimodality vanishes and we get the more unimodal results as expected. In both cases, the filtered factors are well estimated. Hence, there are no further consequences for our parameters of interest such as the cluster centroids. All estimated cluster centroids are centered around their true values with just small deviations.



(a) Densities of estimated static parameters.

(b) Densities of estimated cluster centroids.

Figure 6: Parameter estimation results of $M = 1,000$ simulations for $N = 500, T_y = 10, T_x = 100$. Vertical lines represent true values.

Besides the estimated values of the cluster centroids, the assignment to the correct cluster is also of importance. For this purpose we present confusion matrices in Table 2. In the columns we vary time series dimension $T_y \in \{10, 20\}$, while in the rows $T_x \in \{50, 100\}$ varies. First, we recall at this point that the results of the first step of the estimation procedure are taken as given and the time series dimension of the second step, denoted by T_y, is now of interest. For the case of $T_y = 10$, the confusion matrices are already very satisfactory, with lower boundaries of 96.9% being correctly assigned; for the case of $T_y = 20$, the lower boundaries are even as high as 99.6%. This confirms our earlier finding that even though it might be empirically challenging to separate some of the score-driven model parameters, it does not lead to any problem in identifying the clusters and cluster centroids in the second step. Hence we can correctly identify the structure in the data.

The estimated cluster centroids and confusion matrices are rather precise. However, our key interest is the comparison between the unrestricted model and the clustered model. For that matter, we present in-sample statistics in Table 3 for time series dimension $T_y = 10$. We again found very

	$\mathbf{T_y = 10, T_x = 50}$					$\mathbf{T_y = 20, T_x = 50}$				
	E1	**E2**	**E3**	**E4**	**E5**	**E1**	**E2**	**E3**	**E4**	**E5**
C1	99.960	0.040	0	0	0	100	0	0	0	0
C2	0.057	99.931	0.012	0	0	0.001	99.999	0	0	0
C3	0	0.009	96.938	3.053	0	0	0	99.574	0.426	0
C4	0	0	2.954	97.046	0	0	0	0.378	99.622	0
C5	0	0	0	0	100	0	0	0	0	100

	$\mathbf{T_y = 10, T_x = 100}$					$\mathbf{T_y = 20, T_x = 100}$				
	E1	**E2**	**E3**	**E4**	**E5**	**E1**	**E2**	**E3**	**E4**	**E5**
C1	99.952	0.048	0	0	0	100	0	0	0	0
C2	0.043	99.945	0.012	0	0	0	100	0	0	0
C3	0	0.010	96.925	3.065	0	0	0	99.602	0.398	0
C4	0	0	2.996	97.004	0	0	0	0.383	99.617	0
C5	0	0	0	0	100	0	0	0	0	100

Table 2: Confusion matrices of estimated cluster centroids. All results are based on $M = 1,000$ simulations with $N = 500$. The top panels have $T_x = 50$ while $T_y \in \{10, 20\}$ varies and the bottom panels have $T_x = 100$ fixed while $T_y \in \{10, 20\}$ varies. In each panel, the row labels indicate the true clusters and the columns labels to the assigned clusters in estimation. Frequencies are given (here also equal to percentages), perfect classification would be $100I_5$. From smallest to largest, the true values are $4, 11, 19, 23$ and 35.

similar results for $T_y = 20$, so we refer to Table B.1 in the Appendix for the corresponding tables. In the simulation study, we used five equally-sized clusters with centroids $4, 11, 19, 23$ and 35 as loading matrix. We report for each of the clusters, and the full sample, the mean squared error and mean absolute error of the unrestricted estimates and estimated cluster centroids compared to the true values and the difference between the two estimates. Overall, the estimated cluster centroids always outperform the unrestricted estimates. Furthermore, we also report the contribution of each cluster to the log likelihood. We can then compare the models by the AIC and, since the clustered model is a restricted version of the unrestricted model, the LR-test. In all cases, the LR-test provides evidence that there is no significant difference between the log likelihood values. However, taking into account that the clustered model has much less parameters than the unrestricted model, the AIC shows that the clustered model should be preferred. For each combination of time series dimensions, the differences between the clusters are small because the clusters are here chosen to be the same in size and without deviation around the cluster centroid, but in empirical studies such statistics will give insight on the homogeneity of the different clusters.

$\mathbf{T_y = 10, T_x = 50}$

	MSE			MAE			Unrestricted		Clustered		
	Unr.	**Cl.**	**Diff.**	**Unr.**	**Cl.**	**Diff.**	**LL**	**AIC**	**LL**	**AIC**	**LR**
C1	1.134	0.047	1.106	0.849	0.127	0.839	-4,324	8,850	-4,375	8,754	102
C2	1.234	0.150	1.117	0.884	0.267	0.842	-4,325	8,852	-4,376	8,756	102
C3	1.432	0.827	1.015	0.948	0.538	0.817	-4,332	8,866	-4,378	8,760	92
C4	1.575	0.929	1.008	0.992	0.623	0.813	-4,335	8,872	-4,381	8,766	92
C5	2.168	1.053	1.115	1.157	0.786	0.839	-4,350	8,902	-4,399	8,802	98
Full	1.508	0.601	1.072	0.966	0.468	0.830	-21,884	44,770	-22,130	44,272	492

$\mathbf{T_y = 10, T_x = 100}$

	MSE			MAE			Unrestricted		Clustered		
	Unr.	**Cl.**	**Diff.**	**Unr.**	**Cl.**	**Diff.**	**LL**	**AIC**	**LL**	**AIC**	**LR**
C1	1.137	0.047	1.111	0.849	0.121	0.840	-4,323	8,848	-4,374	8,752	102
C2	1.216	0.134	1.109	0.876	0.243	0.839	-4,323	8,848	-4,374	8,752	102
C3	1.408	0.802	1.010	0.937	0.508	0.814	-4,326	8,854	-4,373	8,750	94
C4	1.519	0.888	1.004	0.971	0.575	0.811	-4,328	8,858	-4,375	8,754	94
C5	2.054	0.950	1.104	1.114	0.705	0.837	-4,338	8,878	-4,388	8,780	100
Full	1.467	0.564	1.067	0.949	0.430	0.828	-21,856	44,714	-22,105	44,222	498

Table 3: Model fit for unrestricted and clustered model. All results are based on $M = 1,000$ simulations with $N = 500$ and $T_y = 10$, with $T_x = 50$ in the top panel and $T_x = 100$ in the bottom panel. Each row in a panel represents a cluster and the last row is the full sample. In the columns are loss functions given of the unrestricted loadings compared to the true ones ("Unr."), the clustered centroids compared to the true ones ("Cl.") and the unrestricted loadings minus the clustered centroids ("Diff."). The first three columns have the MSE as loss function and the last three columns the MAE. For both models, the log likelihood and AIC are given and they are compared via the LR-statistic in the last column. For the latter, the critical values are 123 for each cluster (100-1 degrees of freedom) and 548 overall (500-5 degrees of freedom), at the 5% significance level.

4.2 Forecasting Results

By enforcing the clustered structure in our modeling framework, we do not want to compromise on the forecasting performance. To judge the accuracy of clustered forecasting, we compute loss functions of forecasting with and without clustering. We then consider the fraction of the two loss functions and prefer clustered forecasting if

$$\text{Relative Accuracy} = M^{-1} \sum_{m=1}^{M} \left(\frac{LF_m^{clustered}}{LF_m^{unrestricted}} \right) < 1,$$

where the numerator reflects clustered forecasting (with $\hat{\Lambda}$ after running k-means) and the denominator reflects unrestricted forecasting (with $\tilde{\Lambda}$ directly after least squares) for loss function $LF \in \{MSE, MAE\}$. We use $M = 1,000$ simulations to obtain the Relative Accuracy.

Table 4 reports these average fractions; in the columns we vary time series dimension $T_y \in \{10, 20\}$, while in the rows we vary $T_x \in \{50, 100\}$. Furthermore, row f in any cell of Table 4 reports the performance for forecasting $f \in \{1, 2, 3\}$ steps ahead.

We saw before that the time series dimension T_x of the first step in our proposed estimation procedure does not have a big impact on the estimation results of the second step, this is also confirmed by the forecasting results. An improvement of more than 4% in the mean squared error is obtained if $T_y = 20$. This becomes 7% if T_y is only half of it. This reveals the strength of our proposed procedure: for data sets where forecasting is of interest but the time series dimension is small although the cross-section dimension is large, it is beneficial to exploit the clustered nature of the data. As the time series dimension T_y increases, clustered forecasting goes to unrestricted forecasting because the unrestricted estimates of the loadings become less biased in the second step. For small time series dimension T_y, the clustering averages out these biases such that the forecasting performance improves.

		$\mathbf{T_y = 10}$		$\mathbf{T_y = 20}$	
		MSE-ratio	*MAE-ratio*	*MSE-ratio*	*MAE-ratio*
$\mathbf{T_x = 50}$	$f = 1$	0.930	0.963	0.957	0.978
	$f = 2$	0.929	0.963	0.957	0.978
	$f = 3$	0.929	0.963	0.958	0.978
$\mathbf{T_x = 100}$	$f = 1$	0.928	0.962	0.956	0.978
	$f = 2$	0.928	0.962	0.956	0.977
	$f = 3$	0.930	0.963	0.958	0.978

Table 4: Forecasting performance of clustered forecasting versus unrestricted forecasting. All results are based on $M = 1,000$ simulations with $N = 500$. The top panel has $T_x = 50$ while $T_y \in \{10, 20\}$ varies and the bottom panel has $T_x = 100$ fixed while $T_y \in \{10, 20\}$ varies. Clustered forecasting is preferred if $M^{-1} \sum_{m=1}^{M} \left(\frac{LF^{clustered}}{LF^{unrestricted}} \right)_m < 1$, where the numerator reflects clustered forecasting (with $\hat{\Lambda}$ after running k-means) and the denominator reflects unrestricted forecasting (with $\tilde{\Lambda}$ directly after least squares) for loss function $LF \in \{MSE, MAE\}$. The first row in each cell represents one-step ahead forecasting, the second row two steps ahead and the last row three steps ahead.

4.3 Sensitivity Analysis

In the Monte Carlo study we have assumed the number of clusters as known while this is not the case in an empirical study. To learn about the sensitivity of assuming the number of clusters as known, we present results for cases where we deviate from the true number of clusters. We remain simulating data from our model with five clusters, but now we also let the k-means algorithm use another number of clusters. We present summary statistics of the model performance in Table 5 for $N = 500, T_x = 50, T_y = 10$ and static parameters as before. The table reports the average cluster centroid values, its average MSE, the log likelihood of the clusterd model and the MSE-ratio of forecasting $f = 1$ step ahead, all based on $M = 1,000$ simulations. The true number of clusters is five and we show results based on three to seven clusters.

K	**Average clustered centroids**							**MSE**	**LL**	**f = 1**
3	7.495	21.001	35.000					6.905	-23,273	1.437
4	4.002	11.010	21.012	35.000				2.032	-22,470	1.045
5	*4.002*	*10.998*	*18.974*	*23.032*	*35.000*			*0.601*	*-22,130*	*0.930*
6	3.860	9.948	16.479	20.642	24.612	35.111		0.797	-22,091	0.946
7	3.657	8.315	13.110	18.928	21.760	27.276	35.323	0.960	-22,055	0.958

Table 5: Estimation and forecasting results of using $3, \ldots, 7$ clusters in the k-means algorithm, while the data is simulated using five equally-sized clusters with loadings 4, 11, 19, 23 and 35. All results are based on $M = 1,000$ simulations with $N = 500, T_x = 50, T_y = 10$. The first column gives the number of clusters used in the algorithm, where the row of 5 clusters in italic indicates the true number of clusters. The average estimated cluster centroids are given in the second column and the average MSE in the third column. The log likelihood (LL) is given in the fourth column and the Relative Accuracy (the MSE-ratio) for forecasting one-step ahead is given in the last column.

The simulations are done with cluster centroids 4, 11, 19, 23 and 35, and with equally-sized clusters, such that the average value of the loadings is 18.4. When using less clusters than five, we find that the average means are on the higher side, implying a high MSE and low log likelihood values. When using three clusters instead of five, the forecasting performance decreases considerably (MSE-ratio of 1.437). However, when using four instead of five clusters, the performance of clustered forecasting is as good as for unrestricted forecasting (MSE-ratio ≈ 1). It implies that the accuracy loss is small when imposing more structure while computational efficiency is higher and interpretation remains. The overall performance is higher when using more clusters than five rather than using less clusters than five, both for in-sample (lower MSE of the clustered centroids) and out-of-sample (MSE-ratios < 1) criteria. Given that more clusters can embed a structure with less clusters, we can expect that the latter case shows better results. Of course, using the true number of five clusters remains optimal and using too many clusters will lead to overfitting.

Finally, in our main Monte Carlo study, the time series dimensions T_x and T_y and the cross-section dimension N were chosen in line with the empirical study. However, the static parameters and loadings were chosen under the assumption that the data consists of clearly identifiable clusters. In practice, this might be less the case. For example, in our empirical study, we will find $\omega \approx 0, \phi \approx 0.35, \alpha \approx 0.9$ and the cluster centroids are smaller and also zero or negative. To verify whether such other settings alter the performance of our proposed methodology, we redo the analysis with parameters $(\sigma_\xi, \omega, \phi, \alpha, \sigma_\varepsilon)' = (1, 0, 0.35, 0.9, 1)'$ and with cluster centroids $-4, -2, 0, 3$ and 5. We present these results in the Appendix; they do not lead to very different conclusions regarding the performance of our proposed methodology. Figures B.3 to B.6 show the densities of the estimated parameters and Table B.2 summarizes the forecasting performance.

For this other set of parameter values and cluster centroids, the estimation results are even more precise than those for the original set, even for the smaller time series dimensions. The forecasting results are somewhat less strong (MSE-ratios $\rightarrow 1$), but with clustered forecasting we still don't lose on accuracy compared to unrestricted forecasting. This would mean that unrestricted and clustered forecasting should be equally preferred if one is only interested in forecasting. However, in the policy relevant context of our empirical study, we are especially interested in the interpretation. By the structure that we put on the model, we gain a lot on interpretation while we don't pay for forecasting performance. This in combination with the decent improvements in forecasting performance that we found in the main simulation study above gives enough evidence that our methodology can also be generalized and applied in different contexts in future research. All Monte Carlo study results together give sufficient evidence that we can rely on the results from our empirical study.

5 Empirical Study

In a study on education enrollment in the Netherlands, Spijkerman (2006) found that educational choices are related to unemployment rates, in particular in part-time and on-the-job education. In the literature, proposed causal relations usually concern demand for education vis-a-vis supply of labor. Economists typically assume substitution: people allocate their time towards education and the labor market. In this line of thought, enrollment is understood as an investment decision (Clark, 2011, pp. 524-525). Labor market characteristics such as vacancies, unemployment rates, and wages, influence the choice made at any given moment. For example, higher demand for

	GAS(1,1) w/o ω	GAS(1,1)	GAS(2,1) w/o ω	GAS(1,2) w/o ω
$\hat{\sigma}_\xi^2$	0.005	0.005	0.005	0.005
$\hat{\omega}$	-	-0.000	-	-
$\hat{\phi}_1$	0.365	0.365	0.355	0.329
$\hat{\phi}_2$	-	-	-0.117	-
$\hat{\alpha}_1$	0.933	0.937	0.956	0.936
$\hat{\alpha}_2$	-	-	-	0.039
logL	-181.90	-181.06	-181.35	-181.01
AIC	369.80	370.12	370.70	370.02

Table 6: Parameter estimates and metrics for several GAS specifications.

labor increases the opportunity costs of education, decreasing its relative preference. Conversely, when confronted with a weak labor market, young students are more likely to remain in education; see Lamb, Walstab, Teese, Vickers, and Rumberger (2004), Clark (2011, p. 523). For post-initial education, a tight labor market might induce to re-educate, looking for better job prospects or to protect their position (Groenez, Desmedt, & Nicaise, 2007). There are effects on the supply-side of education as well, especially in on-the-job learning. When demand for labor is high, employers are more likely to provide apprenticeships opportunities. This relationship is assumed to have multiple causes: apprenticeships can be a substitute for hard to find workers (especially for middle-skill vacancies), or they are a way to attract talent. This is the reasoning behind the correction for unemployment vocational education in the Dutch student forecasts (Ministry of Education, Culture and Science, 2020).

There is limited recognition of issues related to the non-stationarity in the regressed time series (participation in education and economic indicators) in this literature; see, for example, Lamb et al. (2004, pp. 126-132). As a result, regressions might be spurious. It is further suggested that causal relations are based on cross-country comparative analysis in which participation choices of individuals cannot be distinguished from the varying institutional landscapes; see, for example, (Groenez et al., 2007, p. 2). We use data from the Netherlands only, and increase N by lowering the level of analysis. In particular, we study transitions from one type of education to another. The change in unemployment rate is regressed on the changes in transitions into first grade studies in Dutch vocational and higher education. Figure 3 suggests a linear relation in differenced unemployment rates and differenced share of full-time students in vocational education. We will not test causal claims, for which a structural causal model should be developed.

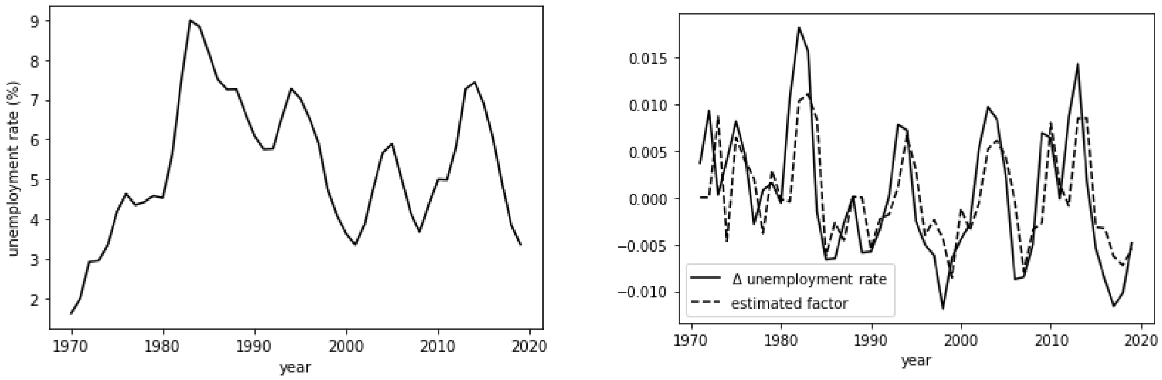


Figure 7: Unemployment rate and estimated factor. Left: unemployment rate at from 1970 to 2019. Right: differenced unemployment rate and estimated factor $\hat{f}_t$ from 1971 to 2019 ($t = 1, 2, \ldots, T_x$, $T_x = 49$).

Transitions with a higher number of students on average have a higher variance. To normalize, each cross-sectional unit is divided by its average level: $\tilde{y}_{it} = \Delta y_{it}/\bar{y}_i$, with y_{it} the number of people in transition i at time $t = 1, ..., T_y$. Several specifications of the GAS model have been tested; see Table 6. Based on the AIC, for our yearly observed macroeconomic time series, the GAS(1,1) without intercept seems to be most suitable. The filter is initialized using the unconditional mean (0). The right panel in Figure 7 presents the differenced unemployment rate (solid) and the estimated common factor (dashed) over time.

With $\hat{\phi}_1 = 0.365$, extrapolating from this filter results in rapid regression towards the zero mean. Comparing clustered and unclustered loadings will not be very meaningful when the forecasted factor is close to zero. Moreover, the model with one factor is sensitive to variance in the forecast of x_t. Since we have a short panel and have not much space to vary T_x when forecasting many steps ahead, the out-of-sample tests are somewhat less reliable[4]. Instead, we will rely on in-sample metrics to compare the clustered and unrestricted models.

The unrestricted model reveals that the unemployment rate explains about 7.7% of the total variation in the education flows of y_t. This low number is to be expected since unemployment is a relevant factor for only a part of the transitions into education. The R^2 of the restricted model is naturally lower, but it converges to this rate of 7.7% as the number of clusters increases. Similarly, the MSE- and MAE-ratios in the right panel converge to 1, see Figure 8. We aim for a relatively small number of centroids since the predictive performance worsens when increasing the number

[4]In an exploration of forecasting performance on random subsets of the data, we found the clustered model did not underperform for the unrestricted one. MSE-ratios center closely around 1.0. More research is needed to draw conclusions.

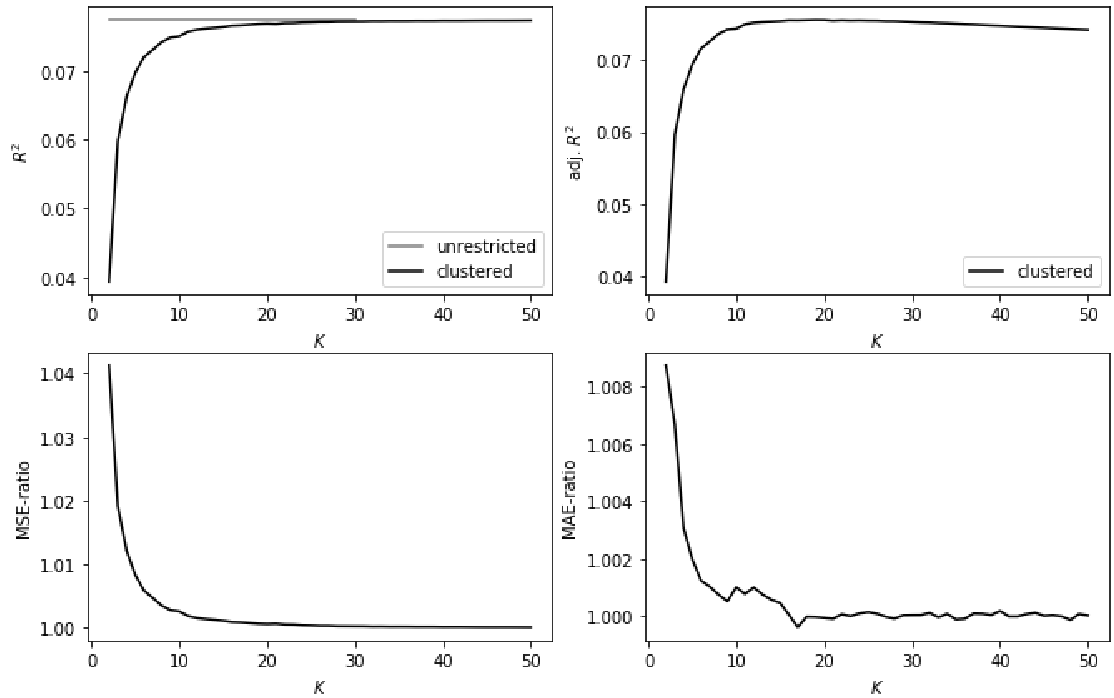


Figure 8: Various metrics (R^2, $\bar{R}^2$, MSE-ratio and MAE-ratio) for $K = 2, \cdots, 50$ (with $R = 1{,}000$ repetitions using different centroid seeds)

of clusters. To find the optimal number of clusters, the adjusted R^2 ($\bar{R}^2$) can be used which has its maximum value at $K = 19$. The MSE-ratio is 1.001, meaning that the predictive performance of the clustered model is almost on par with the unrestricted one. Thus, clustering the loading matrix is an effective way to reduce the number of parameters in the model. We can refer to this procedure as the "elbow method", see section 3.2. Alternatively, other criteria to select the optimal number of clusters can be adopted such as the information criterion AIC.

Figure 9 presents the estimated loading coefficients. It shows a density of unrestricted loadings and a step-wise cumulative distribution of clustered loadings. The estimated unrestricted loadings have sample average -2.67 (std.dev. 13.95), the clustered loadings have average -2.68 (std. dev. 13.89). The smallest and largest loadings are -60.58 and 47.32, respectively, meaning that for these transitions 1% change in the unemployment factor corresponds to a ¿40% change (relative to a transition's sample average) in student counts. The unrestricted loading coefficients do not reveal obvious clusters (multimodal distribution). This is possibly due to high variances of the estimates (small sample) and to model misspecification. We have considered a simple linear model with constant loadings. In reality, the student's decision depends on more than only the unemployment rate. This model therefore is at present not suitable for structural or causal interpretation.

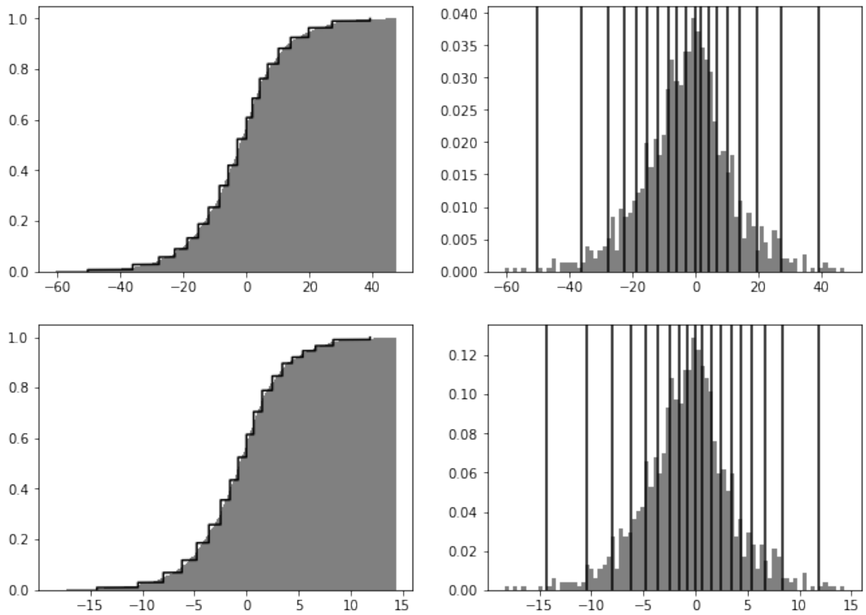


Figure 9: Distribution of loadings. Up with OLS, down with ridge regression ($\alpha = 0.01$). Left: cumulative distribution. The stepwise line indicates the position of clusters and distribution of clustered loadings. Right: density. The vertical lines indicate estimated cluster positions.

The main advantage of this model is that decreasing the number of estimated coefficients reduces model complexity, without losing notable explanatory performance. Clustering might be helpful when model simplicity is considered attractive. In the context of the governmental student forecasts, model interpretability is highly valued (Ministry of Education, Culture and Science, 2018, p. 67). Since the governmental student forecasts feed the education budget, and thus the allocation of scarce, public resources, transparency and explainability are key. We argue that a clustered model is easier to convey, especially if the clusters themselves are meaningfully presented. Both the linear model and the k-means clustering are transparent and widely known. Litle explanatory performance is traded in for gains in model simplicity. Restricting the model does not lead to notable performance loss.

Figure 10 presents the R^2 per cluster. As one would expect, the common factor is most relevant for clusters with larger loadings. In one cluster, 32.1% of variance is explained by the filtered unemployment factor. In the left tail we find transitions into vocational education, which seem to be negatively related to unemployment rates. It supports the hypotheses that participation in on-the-job learning depends on the availability of apprenticeships position and/or that low unemployment induces people to learn market-oriented skills. Similarly, the right tail of the distribution contains many transitions into school-based vocational education, moving pro-cyclical with unemployment.

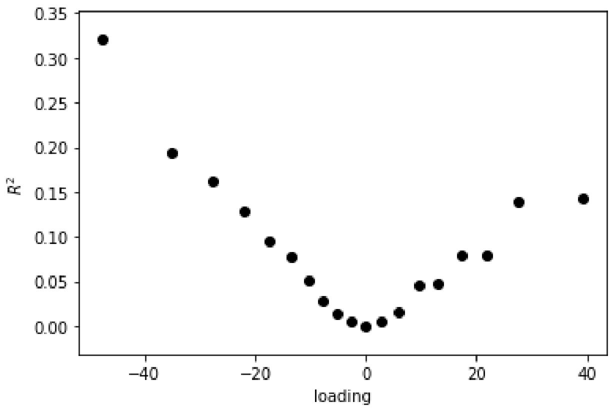


Figure 10: R^2 per cluster

Across the panel, we find that part-time/on-the-job education tends to covary negatively with unemployment; see Table C.1. Similarly, the results indicate that unemployment rates affects people aged 31 and above most; see Figure 11. Also, those not in education are less likely to study with increasing unemployment, whereas transitions from diploma origins do not show a strong covariance. Thus, the results do not suggest that a weak labor market induces people to reorient. Instead, the results favor the notion that people are more likely to participate in post-initial education when having better job prospects. Alternatively, post-initial education depends on the availability of apprenticeship positions. Although clustering strongly reduces the number of parameters, Figure 11 indicates that the structure of the loading matrix stays intact. Loadings also vary to some extent across directions in education, but these results do not provide enough basis to make conclusions.

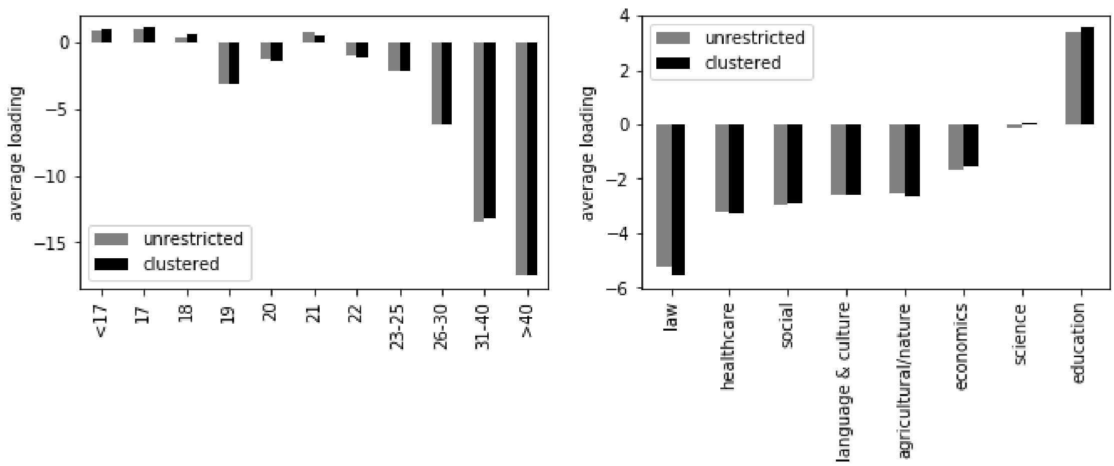


Figure 11: Average loadings in unrestricted and clustered model (weighted by average level of the transition $\bar{y}$) for age (groups) and directions

A drawback of the linear regression model on the short panel is that estimated parameter coefficients are largely affected by random noise. One way to reduce model variance is to include a penalty in the objective criterion. For an illustration, see the right panel of Figure 9 for the distribution of estimated loadings using ridge regression with $\alpha = 0.001$). Model variance might also be reduced by using the descriptive labels of the educational time series. The categories could be included in a k-means type algorithm; see Huang (1998) for an extension to cluster in a setting with mixed categorical and numerical data.

To complete the empirical analysis, we present impulse response functions (IRFs) as produced by a unit shock in the differenced unemployment rate x_t at time $t = 0$ ($\xi_0 = 1$). Figure 12 plots these IRFs which show the dynamic impact of the unemployment shock in education flows. Through the updating equation, the common factor f_t responds with 1 step delay. Additionally, the education flows predicted by the linear regression model covary synchronously with f_t. Some of the clusters covary positively and some negatively. The right panel shows the effects on the differenced transitions into education for the 19 clusters.

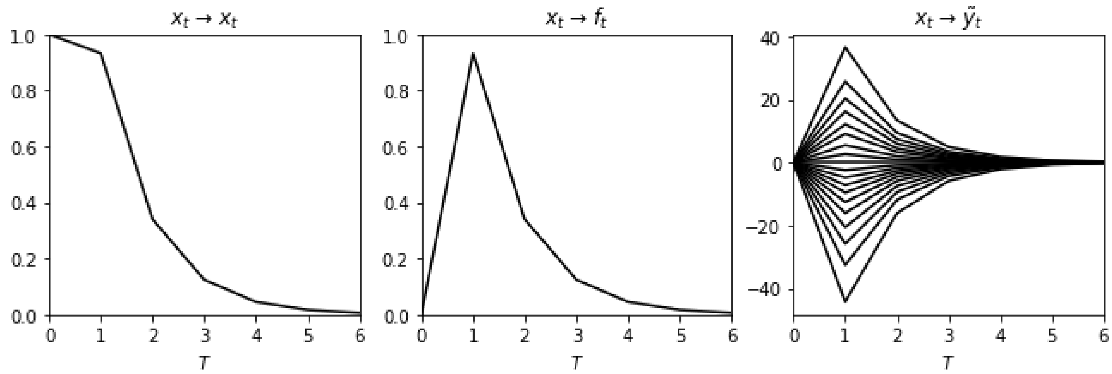


Figure 12: Impulse response functions for unit shock in differenced unemployment rate x_t on (from left to right) x_t, common factor f_t and differenced transitions into education $\tilde{y}_t$

6 Conclusion

In this paper we have introduced a novel dynamic factor model that is capable of forecasting the number of students across the many different types of education. The model can also be used to analyze the relationship between education participation and relevant macroeconomic variables such as the unemployment rate. We further have proposed an econometric treatment for this flexible modeling framework. The empirical analysis is carried out for a large dataset for the educational system in the Netherlands. In this study we have found that, overall, changes in the unemployment rate account for approximately 7.7% of the changes in the flows across the educational system. Given that the panel data dimension is huge, we have allowed for clustering in the factor loadings that are associated with the dynamic macroeconomic factor. As a result we have been able to measure the extent in which the different types of education exhibit similarities in their relationship with macroeconomic cycles. In the empirical study we have highlighted the practical feasibility and good forecasting performance of our modeling framework. In future research, we plan to generalize the methodology further and verify its theoretical properties.

References

Alonso, A. M., Galeano, P., & Peña, D. (2020). A robust procedure to build dynamic factor models with cluster structure. *Journal of Econometrics*, *216*(1), 35-52.

Ando, T., & Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, *31*(1), 163-191.

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.

Barnichon, R., & Mesters, G. (2018). On the demographic adjustment of unemployment. *Review of Economic Statistics*, *100*(2), 219–231.

Blasques, F. (2019). *Advanced econometric methods: Estimation and inference for nonlinear dynamic models*. A. Publications.

Blasques, F., Gorgi, P., Koopman, S. J., & Wintenberger, O. (2018). Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electronic Journal of Statistics*, *12*(1), 1019–1052.

Blasques, F., Koopman, S. J., & Lucas, A. (2014a). Maximum likelihood estimation for generalized autoregressive score models. *Tinbergen Institute Discussion Paper* (TI 2014-029/III).

Blasques, F., Koopman, S. J., & Lucas, A. (2014b). Stationarity and ergodicity of univariate generalized autoregressive score processes. *Electronic Journal of Statistics*, *8*(1), 1088–1112.

Bougerol, P. (1993). Kalman filtering with random coefficients and contractions. *SIAM Journal on Control and Optimization*, *31*(4), 942-–959.

Bräuning, F., & Koopman, S. J. (2014). Forecasting macroeconomic variables using collapsed dynamic factor analysis. *International Journal of Forecasting*, *30*(3), 572-584.

Bureau for Economic Policy Analysis. (2019, November). *MLT-raming november 2019, cijfers.* Retrieved 2020-05-21, from https://www.cpb.nl/middellangetermijnverkenning-2022-2025#docid-160027

Clark, D. (2011). Do recessions keep students in school? The impact of youth unemployment on enrolment in post-compulsory education in England. *Economica*, *78*(311), 523–545. (Publisher: Wiley Online Library)

Creal, D., Koopman, S. J., & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, *28*(5), 777-795.

Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics*, *164*(1), 188-205.

Groenez, S., Desmedt, E., & Nicaise, I. (2007). Participation in lifelong learning in the eu-15: the role of macro-level determinants. In *Paper for the ecer conference.*

Hallin, M., & Liška, R. (2011). Dynamic factors in the presence of blocks. *Journal of Econometrics*, *163*, 29–41.

Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: with applications to financial and economic time series* (Vol. 52). Cambridge University Press.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, *2*(3), 283–304.

Jungbacker, B., & Koopman, S. J. (2015). Likelihood-based dynamic factor analysis for measurement and forecasting. *The Econometrics Journal*, *18*(2), C1–C21.

Lamb, S., Walstab, A., Teese, R., Vickers, M., & Rumberger, R. (2004). Staying on at school: Improving student retention in australia. *Brisbane: Queensland Department of Education and the Arts*.

Ministry of Education, Culture and Science. (2018, September). *Referentieraming 2018.* Retrieved from https://www.rijksoverheid.nl/documenten/rapporten/2018/09/18/referentieraming-ocw-2018

Ministry of Education, Culture and Science. (2020, June). *Referentieramingen 2020.*

Pötscher, B. M., & Prucha, I. (1997). *Dynamic nonlinear econometric models: Asymptotic theory.* Springer-Verlag Berlin.

Spijkerman, M. (2006, October). *De invloed van conjunctuureffecten op onderwijsdeelname.* SEOR, Erasmus Universiteit Rotterdam.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*(460), 1167–1179.

Stock, J. H., & Watson, M. W. (2008). The evolution of national and regional factors in u.s. housing construction. In T. Bollerslev, J. Russell, & M. Watson (Eds.), *Volatility and time series econometrics: Essays in honor of Robert F. Engle.* Oxford University Press.

Straumann, D., & Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *Annals of Statistics*, *34*(5), 2449–2495.

APPENDICES

A Assumptions for Asymptotic Properties

In this Appendix we provide and discuss the assumptions that are needed to have consistency of the maximum likelihood estimates in our two-step method for the factor model. In the first step, the consistency of the MLE $\hat{\boldsymbol{\theta}}_{T_x}$ for the parameters $\boldsymbol{\theta}$ of the score filter of f_t relies on the following assumptions:

(i.a) the sequence $\{x_{\bar{t}}\}_{\bar{t}\in\mathbb{Z}}$ is strictly stationary and ergodic, with one bounded moment $\mathbb{E}|x_{\bar{t}}| < \infty$;

(ii) the filter is invertible (ensured by $|\beta - \alpha| < 1$);

(iii) when $\boldsymbol{\theta}_0$ is the identifiably unique maximizer of the limit log likelihood function.

If the score filter is assumed to be correctly specified, then condition (i.a) above is ensured as long as $|\beta| < 1$ and the innovations $\xi_{\bar{t}}$ are iid with a bounded moment; see Blasques et al. (2014a) or Blasques, Koopman, and Lucas (2014b) for stationarity and moment conditions for score models, which build on Pötscher and Prucha (1997) and Straumann and Mikosch (2006), among others. The invertibility condition in (ii) follows from the contraction theory in Bougerol (1993) and Straumann and Mikosch (2006), and is discussed in more detail in Blasques et al. (2014a) and Blasques (2019). Additionally, condition (iii) is ensured under appropriate regularity conditions stated in Blasques et al. (2014a) and Blasques (2019), among others. The same references provide the proofs of consistency for both correctly specified and misspecified score models.

In the second step, the consistency of the MLE for the regression coefficients Λ can be ensured using standard regularity conditions for consistency of MLE in linear regressions. Since the MLE estimator $\hat{\Lambda}_{T_y}$ depends on the MLE for the score parameters $\hat{\boldsymbol{\theta}}_{T_x}$, the consistency of $\hat{\Lambda}_{T_y}$ is obtained in the context of a plug-in estimator. $\hat{\Lambda}_{T_y}(\hat{\boldsymbol{\theta}}_{T_x})$. Naturally, if $T_x \to \infty$ and $T_y \to \infty$ sequentially, then the proof is trivial since $\hat{\Lambda}_{T_y}(\boldsymbol{\theta}_0)$ no longer relies on $\hat{\boldsymbol{\theta}}_{T_x}$, in the second step. If T_x and T_y are taken to infinity jointly, then the consistency of the plug-in estimator. $\hat{\Lambda}_{T_y}(\hat{\boldsymbol{\theta}}_{T_x})$ follows naturally by the consistency of $\hat{\boldsymbol{\theta}}_{T_x}$ and the continuity of $\hat{\Lambda}_{T_y}(\cdot)$ on $\hat{\boldsymbol{\theta}}_{T_x}$, which holds naturally as $\hat{\Lambda}_{T_y}(\cdot)$ is continuous on the filter $\hat{f}_t$, and the filter $\hat{f}_t$ is continuous on $\boldsymbol{\theta}$. In the present context, we can thus ensure consistency of $\hat{\Lambda}_{T_y}$ as long as:

(iv) the sequence $\{y_t\}_{t\in\mathbb{Z}}$ is strictly stationary and ergodic, with two bounded moments $\mathbb{E}|y_t|^2 < \infty$

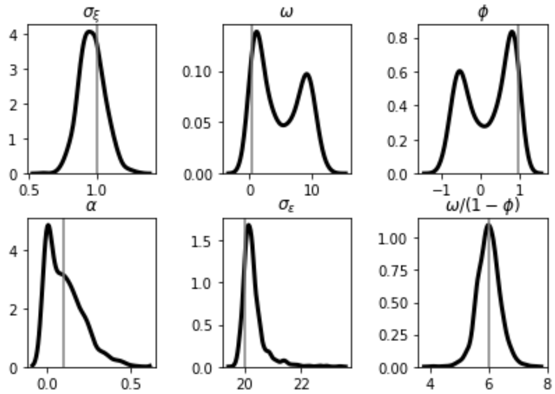
(v) the filtered f_t is asymptotically stationary with two bounded moments.

Condition (v) is easily ensured by the filter invertibility condition (ii) and as long as two bounded moments are assumed in condition (i.a) rather than just one. Hence, in addition we assume:

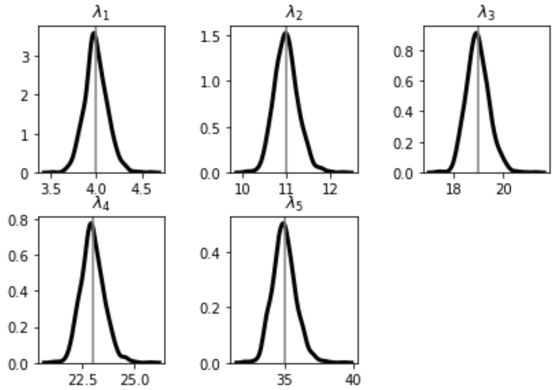
(i.b) the sequence $\{x_{\bar{t}}\}_{\bar{t}\in\mathbb{Z}}$ is strictly stationary and ergodic, with one bounded moment $\mathbb{E}|x_{\bar{t}}|^2 < \infty$;

If the filter is assumed to be correctly specified, condition (i.b) is ensured as long as $|\beta| < 1$ and the innovations $\xi_{\bar{t}}$ are iid with two bounded moments. If the regression model is also assumed correct, then condition (iv) is ensured as long as the innovations ε_t are iid with two bounded moments and independent of f_t. Similar results can be found in Blasques (2019).

B Additional Monte Carlo Results

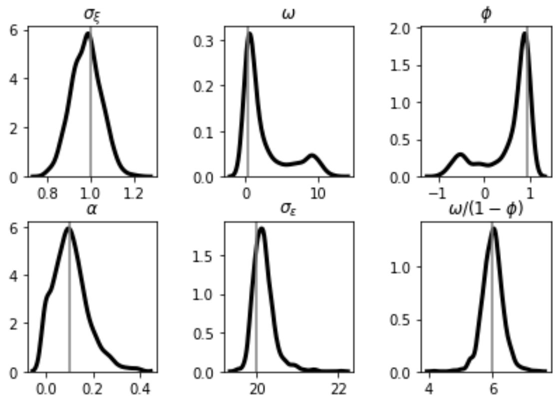


(a) Densities of estimated static parameters.

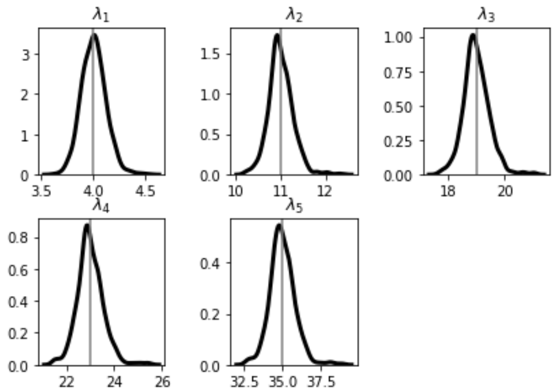


(b) Densities of estimated cluster centroids.

Figure B.1: Parameter estimation results of $M = 1,000$ simulations for $N = 500, T_y = 20, T_x = 50$. Vertical lines represent true values.



(a) Densities of estimated static parameters.



(b) Densities of estimated cluster centroids.

Figure B.2: Parameter estimation results of $M = 1,000$ simulations for $N = 500, T_y = 20, T_x = 100$. Vertical lines represent true values.

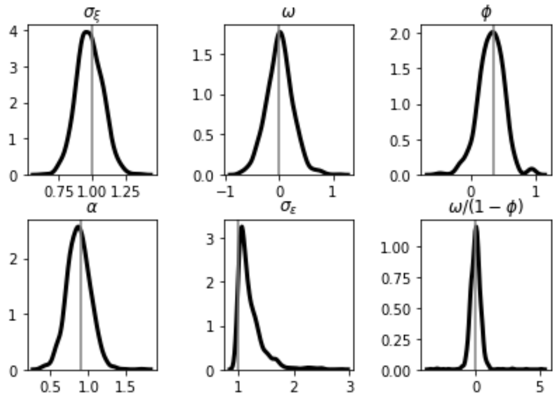
$\mathbf{T_y = 20, T_x = 50}$

	MSE			MAE			Unrestricted		Clustered		
	Unr.	**Cl.**	**Diff.**	**Unr.**	**Cl.**	**Diff.**	**LL**	**AIC**	**LL**	**AIC**	**LR**
C1	0.570	0.016	0.555	0.602	0.097	0.594	-8,694	17,590	-8,745	17,494	102
C2	0.629	0.076	0.554	0.631	0.213	0.593	-8,702	17,606	-8,751	17,506	98
C3	0.769	0.284	0.548	0.697	0.375	0.592	-8,717	17,636	-8,765	17,534	96
C4	0.865	0.372	0.544	0.738	0.448	0.590	-8,727	17,656	-8,775	17,554	96
C5	1.277	0.722	0.555	0.891	0.657	0.593	-8,767	17,736	-8,814	17,632	94
Full	0.822	0.294	0.551	0.712	0.358	0.592	-44,048	89,098	-44,293	88,598	490

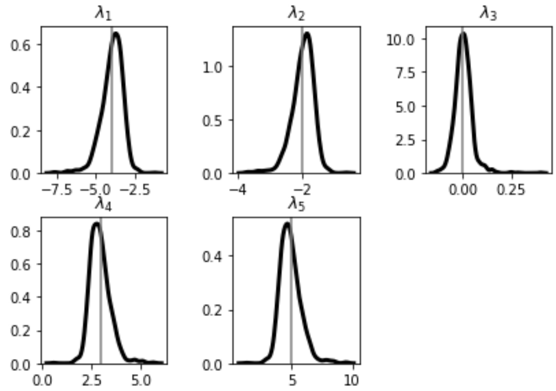
$\mathbf{T_y = 20, T_x = 100}$

	MSE			MAE			Unrestricted		Clustered		
	Unr.	**Cl.**	**Diff.**	**Unr.**	**Cl.**	**Diff.**	**LL**	**AIC**	**LL**	**AIC**	**LR**
C1	0.571	0.014	0.557	0.603	0.093	0.595	-8,694	17,590	-8,745	17,494	102
C2	0.630	0.075	0.555	0.631	0.207	0.593	-8,697	17,596	-8,747	17,498	100
C3	0.765	0.276	0.547	0.691	0.361	0.590	-8,706	17,614	-8,755	17,514	98
C4	0.851	0.358	0.544	0.727	0.426	0.589	-8,709	17,620	-8,758	17,520	98
C5	1.240	0.686	0.554	0.869	0.623	0.592	-8,735	17,672	-8,784	17,572	98
Full	0.812	0.282	0.552	0.704	0.342	0.592	-43,982	88,966	-44,230	88,472	496

Table B.1: Model fit for unrestricted and clustered model. All results are based on $M = 1,000$ simulations with $N = 500$ and $T_y = 20$, with $T_x = 50$ in the top panel and $T_x = 100$ in the bottom panel. Each row in a panel represents a cluster and the last row is the full sample. In the columns are loss functions given of the unrestricted loadings compared to the true ones ("Unr."), the clustered centroids compared to the true ones ("Cl.") and the unrestricted loadings minus the clustered centroids ("Diff."). The first three columns have the MSE as loss function and the last three columns the MAE. For both models, the log likelihood and AIC are given and they are compared via the LR-statistic in the last column. For the latter, the critical values are 123 for each cluster (100-1 degrees of freedom) and 548 overall (500-5 degrees of freedom), at the 5% significance level.

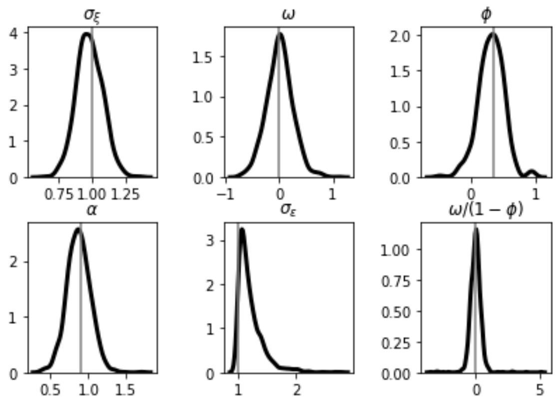


(a) Densities of estimated static parameters.

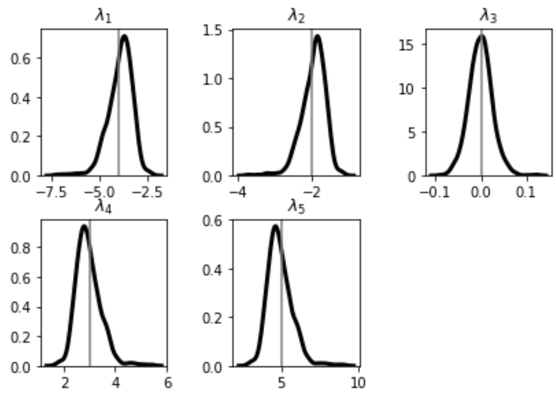


(b) Densities of estimated cluster centroids.

Figure B.3: Parameter estimation results of $M = 1,000$ simulations for $N = 500, T_y = 10, T_x = 50$ in sensitivity analysis of simulation study when parameter values and cluster centroids are more in line with empirical study. Vertical lines represent true values.

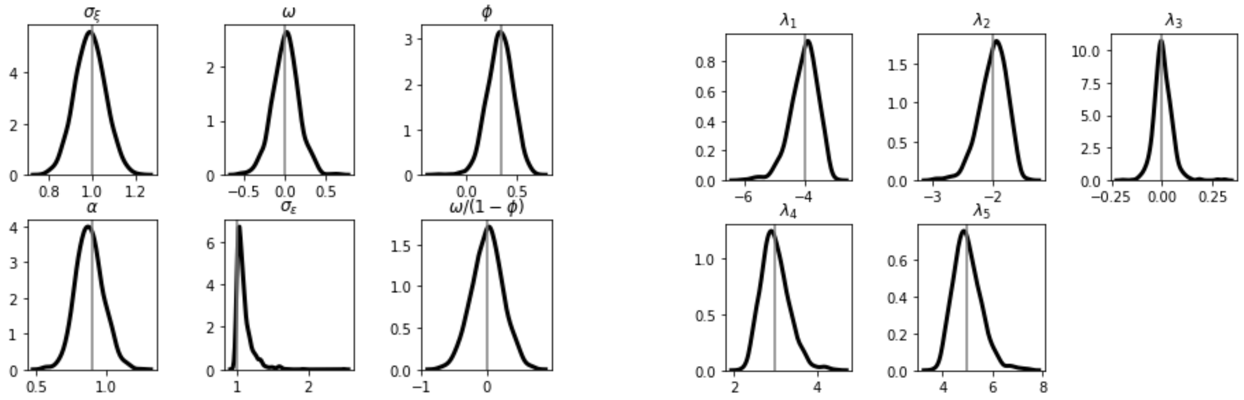


(a) Densities of estimated static parameters.



(b) Densities of estimated cluster centroids.

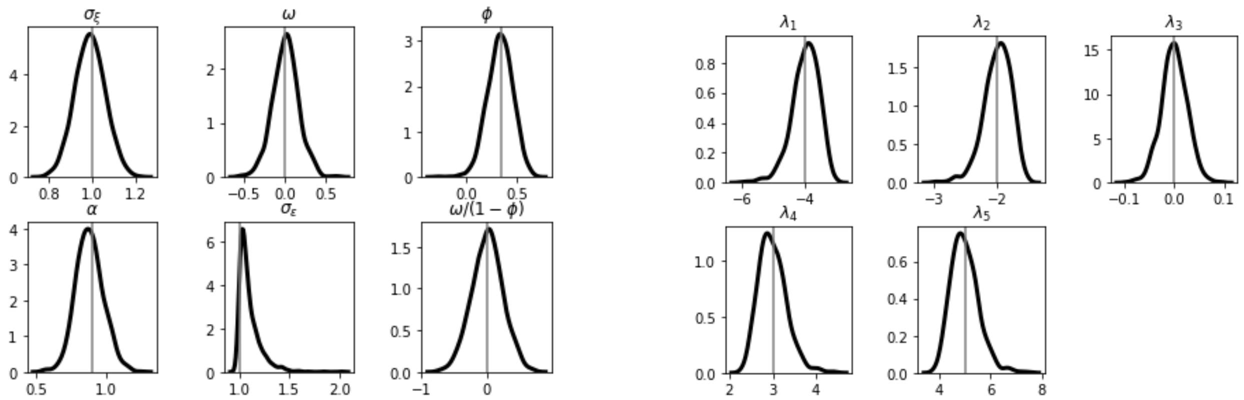
Figure B.4: Parameter estimation results of $M = 1,000$ simulations for $N = 500, T_y = 20, T_x = 50$ in sensitivity analysis of simulation study when parameter values and cluster centroids are more in line with empirical study. Vertical lines represent true values.



(a) Densities of estimated static parameters.

(b) Densities of estimated cluster centroids.

Figure B.5: Parameter estimation results of $M = 1,000$ simulations for $N = 500, T_y = 10, T_x = 100$ in sensitivity analysis of simulation study when parameter values and cluster centroids are more in line with empirical study. Vertical lines represent true values.



(a) Densities of estimated static parameters.

(b) Densities of estimated cluster centroids.

Figure B.6: Parameter estimation results of $M = 1,000$ simulations for $N = 500, T_y = 20, T_x = 100$ in sensitivity analysis of simulation study when parameter values and cluster centroids are more in line with empirical study. Vertical lines represent true values.

		$\mathbf{T_y = 10}$		$\mathbf{T_y = 20}$	
		MSE-ratio	*MAE-ratio*	*MSE-ratio*	*MAE-ratio*
	$f = 1$	0.981	0.991	0.989	0.995
$\mathbf{T_x = 50}$	$f = 2$	0.994	0.997	0.996	0.998
	$f = 3$	0.996	0.998	0.998	0.999
	$f = 1$	0.984	0.992	0.991	0.996
$\mathbf{T_x = 100}$	$f = 2$	0.995	0.998	0.997	0.999
	$f = 3$	0.998	0.999	0.999	0.999

Table B.2: Forecasting performance of clustered forecasting versus unrestricted forecasting in sensitivity analysis of simulation study when parameter values and cluster centroids are more in line with empirical study. All results are based on $M = 1,000$ simulations with $N = 500$. The top panel has $T_x = 50$ while $T_y \in \{10, 20\}$ varies and the bottom panel has $T_x = 100$ fixed while $T_y \in \{10, 20\}$ varies. Clustered forecasting is preferred if $M^{-1} \sum_{m=1}^{M} \left(\frac{LF^{clustered}}{LF^{unrestricted}} \right)_m < 1$, where the numerator reflects clustered forecasting (with $\hat{\Lambda}$ after running k-means) and the denominator reflects unrestricted forecasting (with $\tilde{\Lambda}$ directly after least squares) for loss function $LF \in \{MSE, MAE\}$. The first row in each cell represents one-step ahead forecasting, the second row two steps ahead and the last row three steps ahead.

C Additional Empirical Results

		-47.5	-35.0	-27.8	-22.1	-17.3	-13.5	-10.3	-7.7	-5.2
hbo	fulltime	1	2	7	8	17	18	13	33	30
	parttime	4	2	3	8	14	4	9	9	8
mbo	school-based	0	2	1	4	1	6	5	4	11
	on-the-job	6	13	17	18	16	14	11	12	8
wo	fulltime	1	1	6	9	18	31	25	29	30
	parttime	0	0	0	3	2	4	2	2	1

		-2.6	0.0	2.8	5.8	9.7	13.0	17.3	21.8	27.5	39.2
hbo	fulltime	32	33	40	36	20	10	9	3	2	0
	parttime	10	2	5	4	3	3	0	0	2	2
mbo	school-based	13	26	23	27	31	28	18	15	19	8
	on-the-job	11	9	5	15	3	4	6	5	4	0
wo	fulltime	37	47	36	18	10	3	3	1	1	0
	parttime	1	4	1	3	0	0	0	0	0	1

Table C.1: Clusters (centroid position in column names) with composition of education types. Translations: *mbo*: vocational education, *hbo*: hbo, *wo*: university