Received: 1 February 2019 | Accepted: 15 September 2020

DOI: 10.1111/rssa.12626

ORIGINAL ARTICLE

A dynamic factor model approach to incorporate Big Data in state space models for official statistics

Caterina Schiavoni[1,2] | **Franz Palm**[2] | **Stephan Smeekes**[2] | **Jan van den Brakel**[1,2]

[1]Statistics Netherlands, Heerlen, The Netherlands

[2]Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands

Correspondence
Caterina Schiavoni, Department of Quantitative Economics, Maastricht University, The Netherlands.
Email: c.schiavoni@maastrichtuniversity.nl

Abstract

In this paper we consider estimation of unobserved components in state space models using a dynamic factor approach to incorporate auxiliary information from high-dimensional data sources. We apply the methodology to unemployment estimation as done by Statistics Netherlands, who uses a multivariate state space model to produce monthly figures for unemployment using series observed with the labour force survey (LFS). We extend the model by including auxiliary series of Google Trends about job-search and economic uncertainty, and claimant counts, partially observed at higher frequencies. Our factor model allows for nowcasting the variable of interest, providing reliable unemployment estimates in real-time before LFS data become available.

KEYWORDS

factor models, Google trends, high-dimensional data analysis, nowcasting, state space, unemployment

1 | INTRODUCTION

There is an increasing interest among national statistical institutes (NSIs) to use data that are generated as a by-product of processes not directly related to statistical production purposes in the production of official statistics. Such data sources are sometimes referred to as 'Big Data'; examples are time and location of network activity available from mobile phone companies, social media messages from Twitter and

Facebook, sensor data, and internet search behaviour from Google Trends. A common problem with this type of data sources is that they are likely selective with respect to an intended target population. If such data sources are directly used to produce statistical information, then the potential selection bias of these data sources must be accounted for, which is often a hard task since Big Data sources are often noisy and generally contain no auxiliary variables, which are required for bias correction. These problems can be circumvented using them as covariates in model-based inference procedures to make precise detailed and timely survey estimates, since they come at a high frequency and are, therefore, very timely. These techniques are known in the literature as small area estimation and nowcasting (Rao & Molina, 2015).

Official statistics are generally based on repeated samples. Therefore, multivariate time series models are potentially fruitful to improve the precision and timeliness of domain estimates with survey data obtained in preceding the reference periods and other domains. The predictive power of these models can be further improved by incorporating auxiliary series that are related with the target series observed with a repeated survey.

In this paper, we investigate how auxiliary series derived from Big Data sources and registers can be combined with time series observed with repeated samples in high-dimensional multivariate structural time series (STS) models. We consider Google Trends and claimant counts as auxiliary series for monthly unemployment estimates observed with a continuously conducted sample survey. Big Data sources have the problem that they are noisy and potentially (partly) irrelevant, and, as such, care must be taken when using them for the production of official statistics. We show that, using a dynamic factor model in state space form, relevant information can be extracted from such auxiliary high-dimensional data sources, while guarding against the inclusion of irrelevant data.

Statistical information about a country's labour force is generally obtained from labour force surveys, since the required information is not available from registrations or other administrative data sources. The Dutch LFS is based on a rotating panel design, where monthly household samples are observed five times with quarterly intervals. These figures are, however, considered too volatile to produce sufficiently reliable monthly estimates for the employed and the unemployed labour force at monthly frequency. For this reason Statistics Netherlands estimates monthly unemployment figures, together with its change, as unobserved components in a state space model where the observed series come from the monthly Dutch LFS, using a model originally proposed by Pfeffermann (1991). This method improves the precision of the monthly estimates for unemployment with sample information from previous periods and can, therefore, be seen as a form of small area estimation. In addition it accounts for rotation group bias (Bailar, 1975), serial correlation due to partial sample overlap and discontinuities due to several major survey redesigns (van den Brakel & Krieg, 2015).

Time series estimates for the unemployment can be further improved by including related auxiliary series. The purpose is twofold. First, auxiliary series can further improve the precision of the time series predictions. In this regard, Harvey and Chung (2000) propose a bivariate state space model to combine a univariate series of the monthly unemployed labour force derived from the UK LFS, with the univariate auxiliary series of claimant counts. The latter series represents the number of people claiming unemployment benefits. It is an administrative source, which is not available for every country, and, as for the Netherlands, it can be affected by the same publication delay of the labour force series. Second, auxiliary series derived from Big Data sources like Google Trends are generally available at a higher frequency than the monthly series of the LFS. Combining both series in a time series model allows to make early predictions for the survey outcomes in real-time at the moment that the outcomes for the auxiliary series are available, but the survey data not yet, which is in the literature known as *nowcasting*, in other words, 'forecasting the present'.

In this paper, we extend the state space model used by Statistics Netherlands in order to combine the survey data with the claimant counts and the high-dimensional auxiliary series of Google Trends

about job-search and economic uncertainty, as they could yield more information than a univariate one, which is not affected by publication lags and that can eventually be observed at a higher frequency than the labour force series. This paper contributes to the existing literature by proposing a method to include a high-dimensional auxiliary series in a state space model in order to improve the (real-time) estimation of unobserved components. The model accounts for the rotating panel design underlying the sample survey series, combines series observed at different frequencies and deals with missing observations at the end of the sample due to publication delays. It handles the curse of dimensionality that arises from including a large number of series related to the unobserved components, by extracting their common factors.

Besides claimant counts, the majority of the information related to unemployment is nowadays available on the internet; from job advertisements to resumé's templates and websites of recruitment agencies. We, therefore, follow the idea originating in Choi and Varian (2009), Askitas and Zimmermann (2009) and Suhoy (2009) of using terms related to job and economic uncertainty, searched on Google in the Netherlands. Since 2004, these time series are freely downloadable in real-time from the Google Trends tool, at a monthly or higher frequency. As from the onset it is unclear which search terms are relevant and if so, to which extent, care must be taken not to model spurious relationships with regards to the labour force series of interest, which could have a detrimental effect on the estimation of unemployment, such as happened for the widely publicised case of Google Flu Trends (Lazer et al., 2014). We avoid this problem by employing a targeting technique which allows us to carry out a first selection of those Google Trends that are related to the Dutch unemployment. This is achieved with the elastic net (Hastie & Zou, 2005), which is commonly used in statistics and econometrics as a variable selection technique in high-dimensional time series models (Bai & Ng, 2008; Hastie et al., 2015).

Our method allows to exploit the high-frequency and/or real-time information of the auxiliary series, and to use it in order to nowcast the unemployment, before the publication of labour force data. As the number of search terms related to unemployment can easily become large, we employ the two-step estimator of Doz et al. (2011), which combines factor models with the Kalman filter, to deal both with the high-dimensionality of the auxiliary series and with the estimation of the state space model. The above-mentioned estimator is generally used to improve the nowcast of variables that are observed such as GDP (see Giannone et al., 2008; Hindrayanto et al., 2016 for applications to the United States and the euro area), which is not the case for the unemployment. Nonetheless, D'Amuri and Marcucci (2017), Naccarato et al. (2018) and Maas (2019) are all recent studies that use Google Trends to nowcast and forecast the unemployment, by treating the latter as *known* dependent variable in time series models where the Google searches are part of the explanatory variables. To the best of our knowledge, our paper is the first one to use Google Trends in order to nowcast the latent, *unobserved*, structural components of unemployment, that is, trend and seasonal, in a state space model where the observed series are the survey measures.

We evaluate the performance of our proposed method via Monte Carlo simulations and find that our method can yield large improvements in terms of Mean Squared Forecast Error (MSFE) of the unobserved components' nowcasts. We then assess whether the accuracy of the unemployment's estimation and nowcast improves with our high-dimensional state space model, respectively, from in-sample and out-of-sample results. The latter consists of a recursive nowcast. We do not venture into forecasting exercises as Google Trends are considered to be more helpful in predicting the present rather than the future of economic activities (Choi & Varian, 2012). We conclude that Google Trends can significantly improve the fit of the model, although the magnitude of these improvements is sensitive to aspects of the data and the model specification, such as the frequency of observation of the Google Trends, the number of Google Trends' factors included in the model, and the level of

estimation accuracy provided by the first step of the two-step estimation procedure. We finally note that the Google Trends in our analysis have a purely predictive role. We extract the common information contained in these series, by estimating their common factors, and use it in order to (possibly) improve the in-sample estimation and especially the nowcast of the Dutch unemployment. We do not attach any causal relationship between the individual search terms and the latter series.

The remainder of the paper is organised as follows. Section 2 discusses the data used in the empirical analysis. Section 3.1 describes the state space model that is currently used by Statistics Netherlands to estimate the unemployment. Section 3.2 focuses on our proposed method to include a high-dimensional auxiliary series in the aforementioned model. Sections 4 and 5 report, respectively, the simulation and empirical results for our method. Section 6 concludes.

2 | DATA

The Dutch LFS is conducted as follows. Each month a stratified two-stage cluster design of addresses is selected. Strata are formed by geographical regions. Municipalities are considered as primary sampling units and addresses as secondary sampling units. All households residing on an address are included in the sample with a maximum of three (in the Netherlands there is generally one household per address). All household members with age of 16 or older are interviewed. Since October 1999, the LFS has been conducted as a rotating panel design. Each month a new sample, drawn according to the above-mentioned design, enters the panel and is interviewed five times at quarterly intervals. The sample that is interviewed for the j^{th} time is called the j^{th} wave of the panel, $j = 1, \ldots, 5$. After the fifth interview, the sample of households leaves the panel. This rotation design implies that in each month five independent samples are observed. The generalised regression (GREG, i.e. design-based) estimator (Särndal et al., 1992) is used to obtain five independent direct estimates for the unemployed labour force, which is defined as a population total. This generates over time a five-dimensional time series of the unemployed labour force. Table 1 provides a visualisation for the rotation panel design of the Dutch LFS.

Rotating panel designs generally suffer from Rotation Group Bias (RGB), which refers to the phenomena that there are systematic differences among the observations in the subsequent waves (Bailar, 1975). In the Dutch LFS the estimates for the unemployment based on the first wave are indeed systematically larger compared to the estimates based on the follow-up waves (van den Brakel & Krieg, 2015). This is the net results of different factors:

1. Selective nonresponse among the subsequent waves, that is, panel attrition.
2. Systematic differences due to different data collection models that are applied to the waves. Until 2010 data collection in the first wave was based on face-to-face interviewing. Between 2010 and 2012 data collection in the first wave was based on telephone interviewing for households for which a telephone number of a landline telephone connection was available and face-to-face interviewing for the remaining households. After 2012 data collection in the first wave was based on a sequential mixed-mode design that starts with Web interviewing with a follow-up using telephone interviewing and face-to-face interviewing. Data collection in the follow-up waves is based on telephone interviewing only.
3. Differences in wording and questionnaire design used in the waves. In the first wave a block of questions is used to verify the status of the respondent on the labour force market. In the follow-up waves the questionnaire focuses on differences that occurred compared to the previous interview, instead of repeating the battery of questions.
4. Panel conditioning effects, that is, systematic changes in the behaviour of the respondents. For example, questions about activities to find a job in the first wave might increase the search activities

TABLE 1 An 18-months visualisation for the rotation panel design of the Dutch LFS. Each letter represents a sample. The lower-case letters denote samples that entered the panel before the start of this 18-months period, whereas the capital letters indicate samples that entered the panel from the first month of this period onwards. Each column shows the samples that are interviewed in the corresponding month. Every month a new sample enters the panel and is interviewed five times at a quarterly frequency. After the fifth interview, the sample of households leaves the panel

	month t							
	quarter							
	month							
	A	B C	D E F	G H I	J K L	M N O	P Q R	} wave 1
wave j	x	y z	A B C	D E F	G H I	J K L	M N O	} wave 2
	u	v w	x y z	A B C	D E F	G H I	J K L	} wave 3
	r	s t	u v w	x y z	A B C	D E F	G H I	} wave 4
	o	p q	r s t	u v w	x y z	A B C	D E F	} wave 5

of the unemployed respondents in the panel. Respondents might also systematically adjust their answers in the follow-up waves, since they learn how to keep the routing through the questionnaire as short as possible.

The Dutch labour force is subject to a 1-month publication delay, which means that the sample estimates for month t become available in month $t + 1$. In order to have more timely and precise estimates of the unemployment, we extend the model by including, respectively, auxiliary series of weekly/monthly Google Trends about job-search and economic uncertainty, and monthly claimant counts, in the Netherlands. Claimant counts are the number of registered people that receive unemployment benefits. The claimant counts for month t become available in month $t + 1$.

Google Trends are indexes of search activity. Each index measures the fraction of queries that include the term in question in the chosen geography at a particular time, relative to the total number of queries at that time. The maximum value of the index is set to be 100. According to the length of the selected period, the data can be downloaded at either monthly, weekly or higher frequencies. The series are standardised according to the chosen period and their values can, therefore, vary according to the period's length (Stephens-Davidowitz & Varian, 2015). We use weekly and monthly Google Trends for each search term. Google Trends are available in real-time (i.e. they are available in period t for period t, independently on whether the period is a week or a month).

The list of Google search terms used in the empirical analysis of this paper, together with their translation/explanation, is reported in Tables S1 and S2. A first set of terms was chosen by thinking of queries that could be made by unemployed people in the Netherlands. The rest of the terms has been chosen using the Google Correlate tool and selecting the queries that are highly correlated to each term of the initial set, and that have a meaningful relation to unemployment and, more generally, economic uncertainty. Later in the paper we mention that we need non-stationary (e.g. persistent) Google Trends for our model. Correlations between non-stationary series can be spurious and in this respect Google Correlate is not an ideal tool in order to choose search terms. In Section 5 we explain how to circumvent this problem (i.e. by first 'targeting' the Google Trends).

Figure 1 displays the time series of the five waves of the unemployed labour force, together with the claimant counts and an example of job-related Google query. They all seem to be following the same trend, which already shows the potential of using this auxiliary information in estimating the unemployment.

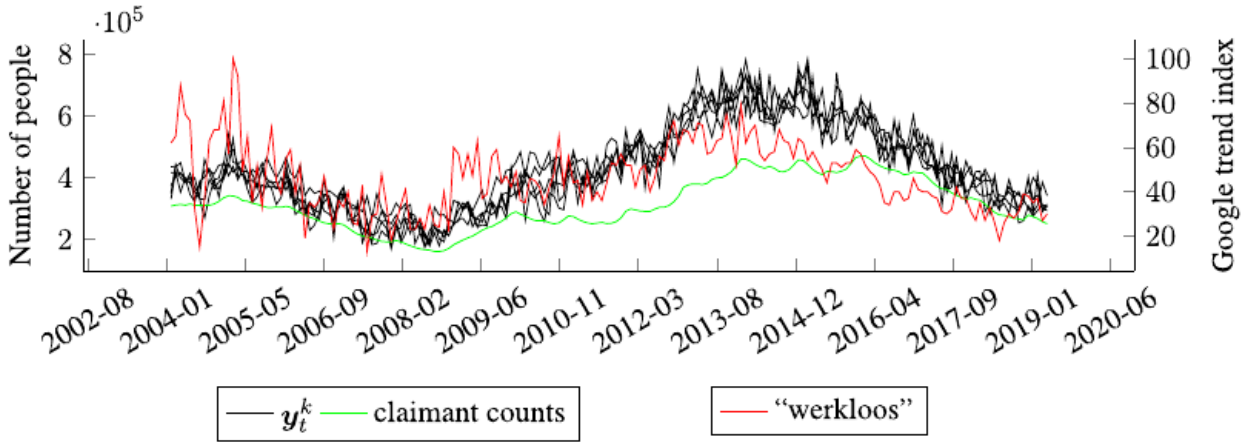

FIGURE 1 Monthly time series of the five waves of the Dutch unemployed labour force ($\boldsymbol{y}_t^k$), the claimant counts, and the Google search term 'werkloos', which means 'unemployed', in the Netherlands. The period starts in January 2004 and ends in May 2019

3 | THE DUTCH LABOUR FORCE MODEL AND EXTENSIONS

We first describe the model in use at Statistics Netherlands in Section 3.1. Next we explain how high-dimensional auxiliary series can be added to this model in Section 3.2.

3.1 | The Dutch labour force model

The monthly sample size of the Dutch LFS is too small to produce sufficiently precise estimates directly. In the past, rolling quarterly figures were published on a monthly frequency. This has the obvious drawback that published figures are unnecessarily delayed since the reference period is the mid-month of the rolling quarter. Also, real monthly seasonal effects are smoothed over the rolling quarter. Another problem that arose after the change from a cross-sectional survey to a rotating panel design in 2000, was that the effects of RGB became visible in the labour force figures. Both problems are solved with a STS model that is used by Statistics Netherlands, since 2010, for the production of monthly statistics about the Dutch labour force (van den Brakel & Krieg, 2015). In a STS model, an observed series is decomposed in several unobserved components, such as a trend, a seasonal component, one or more cycles with a period longer than 1 year, regression components and a white noise component. After writing an STS model in state space form, the Kalman filter can be applied in order to estimate the unobserved components. See Durbin and Koopman (2012) for an introduction to STS modelling.

Let $y_{j,t}^k$ denote the GREG estimate for the unemployment in month t based on the sample observed in wave j. Now $\boldsymbol{y}_t^k = (y_{1,t}^k, \ldots, y_{5,t}^k)'$ denotes the vector with the five GREG estimates for the unemployment in month t. The $y_{j,t}^k$ are treated as five distinct time series in a five-dimensional time series model in order to account for the RGB. The superscript $k > 1$ indicates that the vector is observed at the low frequency. We need this notation (see e.g. Bańbura et al., 2013) to distinguish between series observed at different frequencies, because later on we will make use of Google Trends which are available on a weekly basis. If $\boldsymbol{y}_t^k$ is observed at the monthly frequency, as in the case of the unemployed labour force, then $k = 4, 5$ if the high frequency series is observed at the weekly frequency, since a month can have either 4 or 5 weeks.

The unemployment is estimated, with the Kalman filter, as a state variable in a state space model where $\boldsymbol{y}_t^k$ represents the observed series. The measurement equation takes the form (Pfeffermann, 1991; van den Brakel & Krieg, 2009):

$$\boldsymbol{y}_t^k = \iota_5 \theta_t^{k,y} + \boldsymbol{\lambda}_t^k + \boldsymbol{e}_t^k, \tag{1}$$

where ι_5 is a five-dimensional vector of ones and $\theta_t^{k,y}$, that is, the unemployment, is the common population parameter among the five-dimensional waves of the unemployed labour force. It is composed of the level of a trend (L_t) and a seasonal component (S_t):

$$\theta_t^{k,y} = L_t^{k,y} + S_t^{k,y}.$$

The transition equations for the level (L_t) and the slope (R_t) of the trend are, respectively:

$$\begin{aligned} L_t^{k,y} &= L_{t-1}^{k,y} + R_{t-1}^{k,y}, \\ R_t^{k,y} &= R_{t-1}^{k,y} + \eta_{R,t}^{k,y}, \eta_{R,t}^{k,y} \sim N(0, \sigma_{R,y}^2), \end{aligned}$$

which characterise a smooth trend model. This implies that the level of the trend is integrated of order 2, denoted as $I(2)$, which means that the series of the level is stationary (i.e. mean-reverting) after taking two times successive differences. The slope of the trend, $R_t^{k,y}$, is a first-order integrated series, denoted as $I(1)$. This state variable represents the change in the level of the trend, $L_t^{k,y}$ and not in the unemployment, $\theta_t^{k,y}$, directly. Nevertheless, since the$I(2)$ property of the unemployment is generated by its trend and not by its seasonal component, the change in $\theta_t^{k,y}$ will also mainly be captured by $R_t^{k,y}$ and we can, therefore, consider the latter as a proxy for the change in unemployment. The model originally contained an innovation term for the population parameter $\theta_t^{k,y}$. However, the maximum likelihood estimate for its variance tended to be zero and Bollineni-Balabay et al. (2017) showed via simulations that it is better to not include this term in the model.

The trigonometric stochastic seasonal component allows for the seasonality to vary over time and it is modelled as in Durbin and Koopman (2012) Chapter 3):

$$\begin{aligned} S_t^{k,y} &= \sum_{l=1}^{6} S_{l,t}^{k,y}, \\ \begin{pmatrix} S_{l,t}^{k,y} \\ S_{l,t}^{*k,y} \end{pmatrix} &= \begin{bmatrix} \cos(h_l) & \sin(h_l) \\ -\sin(h_l) & \cos(h_l) \end{bmatrix} \begin{pmatrix} S_{l,t-1}^{k,y} \\ S_{l,t-1}^{*k,y} \end{pmatrix} + \begin{pmatrix} \eta_{\omega,l,t}^{k,y} \\ \eta_{\omega,l,t}^{*k,y} \end{pmatrix}, \begin{pmatrix} \eta_{\omega,l,t}^{k,y} \\ \eta_{\omega,l,t}^{*k,y} \end{pmatrix} \sim N(\mathbf{0}, \sigma_{\omega,y}^2 \boldsymbol{I}_2), \end{aligned}$$

where $h_l = \frac{\pi l}{6}$, for $l = 1, \ldots, 6$ and $\boldsymbol{I}_2$ is a 2×2 identity matrix.

The second component in Equation (1), $\boldsymbol{\lambda}_t^k = (\lambda_{1,t}^k, \ldots, \lambda_{5,t}^k)'$, accounts for the RGB. Based on the factors that contribute to the RGB, as mentioned in Section 2, the response observed in the first wave is assumed to be the most reliable one and not to be affected by the RGB (van den Brakel & Krieg, 2009). Therefore, it is assumed that $\lambda_{1,t}^k = 0$. The remaining four components in $\boldsymbol{\lambda}_t^k$ are random walks that capture time-dependent differences between the follow-up waves with respect to the first wave:

$$\begin{aligned} \lambda_{1,t}^k &= 0, \\ \lambda_{j,t}^k &= \lambda_{j,t-1}^k + \eta_{\lambda,j,t}^k, \eta_{\lambda,j,t}^k \sim N(0, \sigma_\lambda^2), j = 2, \ldots, 5. \end{aligned}$$

As a result the Kalman filter estimates for $\theta_t^{k,y}$ in Equation (1) are benchmarked to the level of the GREG series of the first wave.

The third component in Equation (1), $\boldsymbol{e}_t^k = (e_{1,t}^k, \ldots, e_{5,t}^k)^t$, models the autocorrelation among the survey errors ($e_{j,t}^k$) in the follow-up waves due to the sample overlap of the rotating panel design. In

order to account for this autocorrelation, the survey errors are treated as state variables, which follow the transition equation below.

$$
\begin{aligned}
& e^k_{j,t} = c_{j,t}\tilde{e}^k_{j,t},\ c_{j,t} = \sqrt{\widehat{\text{var}}(y^k_{j,t})}, \qquad j = 1, \ldots, 5, \\
& \tilde{e}^k_{1,t} \sim N(0, \sigma^2_{v_1}), \\
& \tilde{e}^k_{j,t} = \delta \tilde{e}^k_{j-1,t-3} + v^k_{j,t},\ v^k_{j,t} \sim N(0, \sigma^2_{v_j}), \qquad j = 2, \ldots, 5, |\delta| < 1, \\
& \text{var}(\tilde{e}^k_{j,t}) = \sigma^2_{v_j}/(1-\delta^2), \qquad j = 2, \ldots, 5,
\end{aligned}
\tag{2}
$$

with $\widehat{\text{var}}\ (y^k_{j,t})$ being the design variance of the GREG estimates $y^k_{j,t}$. The scaled sampling errors, $\tilde{e}^k_{j,t}$, for $j = 1, \ldots, 5$, account for the serial autocorrelation induced by the sampling overlap of the rotating panel. Samples in the first wave are observed for the first time and, therefore, its survey errors are not autocorrelated with survey errors of previous periods. The survey errors of the second to fifth wave are correlated with the survey errors of the previous wave 3 months before. Based on the approach proposed by Pfeffermann et al. (1998), van den Brakel and Krieg (2009) motivate that these survey errors should be modelled as an AR(3) process, without including the first two lags. Moreover, the survey errors of all waves are assumed to be proportional to the standard error of the GREG estimates. In this way the model accounts for heterogeneity in the variances of the survey errors, which are caused by changing sample sizes over time. As a result the maximum likelihood estimates of the variances of the scaled sampling errors, $\sigma^2_{v_j}$, will have values approximately equal to one.

The STS model (1) as well as the models proposed in the following sections are fitted with the Kalman filter after putting the model in this state space form. We use an exact initialisation for the initial values of the state variables of the sampling error and a diffuse initialisation for the other state variables. It is common to call *hyperparameters* the parameters that define the stochastic properties of the measurement equation and the transition equation of the state space model. These are the parameters that are assumed to be known in the Kalman filter (Durbin & Koopman, 2012, Chapter 2). In our case the hyperparameters are δ and all the parameters that enter the covariance matrices of the innovations. These hyperparameters are estimated by maximum likelihood using the Broyden–Fletcher–Goldfarh–Shanno (BFGS) optimisation algorithm. The additional uncertainty of using maximum likelihood estimates for the hyperparameters in the Kalman filter is ignored in the standard errors of the filtered state variables. Since the observed time series contains 185 monthly periods, this additional uncertainty can be ignored. See also Bollineni-Balabay et al. (2017) for details. Both the simulation and estimation results in Sections 4 and 5 are obtained using the statistical software R.

Assuming normality of the innovations is common in state space models because the hyperparameters of the model are estimated by maximising a Gaussian log-likelihood which is evaluated by the Kalman filter. Moreover, under normality, the Kalman filter yields the minimum variance unbiased estimator of the state variables. Nonetheless, as long as the state space model is linear, if the true distribution of the error terms is non-Gaussian, then the Kalman filter still provides the minimum variance *linear* unbiased estimator of the state variables (Durbin & Koopman, 2012, Chapter 4). In this case we can further rely on quasi-maximum likelihood (QML) theory in order to perform inference based on the QML estimates of the hyperparameters. This means that the hyperparameters can still be consistently estimated by maximising the Gaussian log-likelihood (or in general, as Gourieroux et al. (1984) argue, a density function that belongs to the family of linear exponential distributions), but we shall use, if needed, the appropriate expression for the covariance matrix of the QML estimators, which should capture the additional uncertainty caused by the model's misspecification (Hamilton, 1994, Chapter 13). In Section 4 of the supplementary material we conduct a Monte Carlo simulations study and find that deviations from normality are not of concern for the performance our method.

This time series model addresses and solves the mentioned problems with small sample sizes and RGB. Every month a filtered estimate for the trend ($L_t^{k,y}$) and the population parameter, which is defined as the filtered trend plus the filtered seasonal effect ($\theta_t^{k,y} = L_t^{k,y} + S_t^{k,y}$), are published in month $t + 1$. The time series model uses sample information from previous months in order to obtain more stable estimates. The estimates account for RGB by benchmarking the estimates for $L_t^{k,y}$ and $\theta_t^{k,y}$ to the level of the first wave, which makes them comparable with the outcomes obtained under the cross-sectional design before 2000.

We now introduce some further notation to distinguish between in-sample estimates and out-of-sample forecasts. In the case of in-sample estimates, $\widehat{\theta}_{t|\Omega_t}^{k,y}$ denotes the filtered estimate of the population parameter $\theta_t^{k,y}$, assuming that all data for time t is released and available at time t. We, therefore, condition on the information set Ω_t which does not contain any missing data at time t. In the case of out-of-sample forecasts, we condition on the data set Ω_t^- that is actually available in real-time at time t. For instance, $\boldsymbol{y}_t^k$ only gets published during month $t + 1$, and is, therefore, not available yet at time t, and not part of Ω_t^-. Thus $\widehat{\theta}_{t|\Omega_t^-}^{k,y}$ is the filtered forecast for $\theta_t^{k,y}$, based on the information that is available at time t. Under model (1), which does not contain auxiliary information other than the labour force series, $\widehat{\theta}_{t|\Omega_t^-}^{k,y}$ is in fact the one-step-ahead prediction $\widehat{\theta}_{t|\Omega_{t-1}}^{k,y}$, since $\boldsymbol{y}_t^k$ is not available yet in month t, but $\boldsymbol{y}_{t-1}^k$ is; therefore, $\Omega_t^- = \Omega_{t-1} = \{\boldsymbol{y}_{t-1}^k, \boldsymbol{y}_{t-2}^k, \ldots\}$.

3.2 | Including high-dimensional auxiliary series

To improve precision and timeliness of the monthly unemployment figures, we extend the labour force model by including auxiliary series of weekly/monthly Google Trends about job-search and economic uncertainty, and monthly claimant counts, in the Netherlands. Since the claimant counts for month t become available in month $t + 1$, it is anticipated that this auxiliary series is particularly useful to further improve the precision of the trend and population parameter estimates after finalising the data collection for reference month t. The Google Trends come at a higher frequency already during the reference month t. It is, therefore, anticipated that these auxiliary series can be used to make first provisional estimates for the trend and the population parameter of the LFS during month t, when the sample estimates $\boldsymbol{y}_t^k$ are not available, but the Google Trends become available on weekly basis.

Weekly and monthly Google Trends are throughout the paper denoted by $\boldsymbol{x}_t^{GT}$ and $\boldsymbol{x}_t^{k,GT}$, respectively. We denote the dimension of the vector $\boldsymbol{x}_t^{GT}$ by n, which can be large. In addition, we can expect the Google Trends to be very noisy, such that the signal about unemployment contained in them is weak. We, therefore, need to address the high-dimensionality of these auxiliary series, in order to make the dimension of our state space model manageable for estimation and extract the relevant information from these series. For this purpose we employ a factor model which achieves both by retaining the information of these time series in a few common factors. Moreover, when dealing with mixed frequency variables and with publication delays, we can encounter 'jagged edge' datasets, which have missing values at the end of the sample period. The Kalman filter computes a prediction for the unobserved components in presence of missing observations for the respective observable variables. The two-step estimator by Doz et al. (2011) combines factor models with the Kalman filter and hence addresses both of these issues. In the remainder of this section we explain in detail how this estimator can be employed to nowcast the lower-frequency unobserved components of the labour force model using information from higher-frequency or real-time auxiliary series. The idea is to first use principal component analysis (PCA) to reduce the dimensionality of the Google Trends into a few common factors and then, to re-estimate the factors with the Kalman filter, together with the state variables of the labour force series.

We consider the following state space representation of the dynamic factor model for the Google Trends, with respective measurement and transition equations, as we would like to link it to the state space model used to estimate the unemployment (1):

$$\begin{aligned} \boldsymbol{x}_t^{GT} &= \Lambda \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_t \sim N(\boldsymbol{0}, \Psi) \\ \boldsymbol{f}_t &= \boldsymbol{f}_{t-1} + \boldsymbol{u}_t, \boldsymbol{u}_t \sim N(\boldsymbol{0}, \boldsymbol{I}_r), \end{aligned} \tag{3}$$

where $\boldsymbol{x}_t^{GT}$ is a $n \times 1$ vector of observed series, $\boldsymbol{f}_t$ is a $r \times 1$ vector of latent factors with $r \ll n$, $\boldsymbol{\Lambda}$ is a $n \times r$ matrix of factor loadings, $\boldsymbol{\varepsilon}_t$ is the $n \times 1$ vector of idiosyncratic components and $\boldsymbol{\Psi}$ its $n \times n$ diagonal variance matrix; $\boldsymbol{u}_t$ is the $r \times 1$ vector of factors' innovations and $\boldsymbol{I}_r$ is a $r \times r$ identity matrix (which corresponds to the identification conditions implied by estimation of the model using PCA since the factors are only identified up to rotation). Notice that the dynamic equation for $\boldsymbol{f}_t$ implies that we are making the assumption that $\boldsymbol{x}_t^{GT}$ is $I(1)$ of dimension n and $\boldsymbol{f}_t$ is $I(1)$ of dimension r. Later in this section the need of this assumption will become clearer; the intuition behind it is that the factors and the change in unemployment, $\boldsymbol{R}_t^{k,y}$, must be of the same order of integration.

Among others, Bai (2004) proves the consistency of the estimator of $I(1)$ factors by PCA, under the assumptions of limited time and cross-sectional dependence and stationarity of the idiosyncratic components, $\boldsymbol{\varepsilon}_t$, and non-trivial contributions of the factors to the variance of $\boldsymbol{x}_t$ (for the exact formulation we refer to Assumptions A-D in Bai (2004)). We assume absence of cointegration among the factors. We further assume normality of the innovations for the same reasons outlined in Section 3.1. The consistency of the two-step estimator has been originally proven in the stationary framework by Doz et al. (2011) and extended to the nonstationary case by Barigozzi and Luciani (2017).

In the first step, the factors ($\boldsymbol{f}_t$), the factor loadings ($\boldsymbol{\Lambda}$) and the covariance matrix of the idiosyncratic components ($\boldsymbol{\Psi}$) in model (3) are estimated by PCA as in Bai (2004). The matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ are then replaced, in model (3), by their estimates $\widehat{\Lambda}$ and $\widehat{\Psi} = \text{diag}(\widehat{\psi}_{11}, \ldots, \widehat{\psi}_{nn})$ obtained in this first step. These estimates are kept as fixed in the second step, because their high-dimensionality and associated curse of dimensionality complicates re-estimation by maximum likelihood. Moreover, restricting the covariance matrix of the idiosyncratic components $\boldsymbol{\Psi}$ as being diagonal is standard in the literature. The specification of the dynamic factor model with spherical idiosyncratic components is often called 'approximate' dynamic factor model. Doz et al. (2011) and Barigozzi and Luciani (2017) mention that misspecifications of this model arising from time or cross-sectional dependence of the idiosyncratic components, do not affect the consistency of the two-step estimator of the unobserved common factors, if n is large.

In order to make use of the auxiliary series to nowcast the unemployment, we stack together the measurement equations for $\boldsymbol{y}_t^k$ and $\boldsymbol{x}_t^{k,GT}$, respectively, (1) and the first equation of (3) with $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ replaced, respectively, by $\widehat{\Lambda}$ and $\widehat{\Psi}$ and express them at the lowest frequency (in our case the monthly observation's frequency of $\boldsymbol{y}_t^k$). The transition equations for the RGB and survey error component in combination with the rotation scheme applied in the Dutch LFS hamper a formulation of the model on the high frequency. This means that $\boldsymbol{x}_t^{GT}$ needs to be first temporally aggregated from the high to the low frequency (either before or after the first step which estimates $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$). Since $\boldsymbol{x}_t^{GT}$ are the $I(1)$ weekly Google Trends, which are flow variables as they measure the number of queries made during each week, they are aggregated according to the following rule (Bańbura et al., 2013):

$$\boldsymbol{x}_{j,t}^{k,GT} = \sum_{i=1}^{j} \boldsymbol{x}_{t-k+i}^{GT}, \quad j = 1, \ldots, k, \quad t = k, 2k, \ldots. \tag{4}$$

The aggregated $\boldsymbol{x}_{j,t}^{k,GT}$ are then rescaled in order to be bounded again between 0 and 100. The subscript j allows for real-time updating of the aggregated Google Trends in week j when new data become available.

As such, this index indicates that we aggregate weeks 1 up to j. When $j = k$ we are at the end of the month and we simply write $\boldsymbol{x}_t^{k,GT}$ to indicate the end-of-month aggregate value.

In order to get the final model, we also include a measurement equation for the univariate auxiliary series of the claimant counts, assuming that its state vector, $\theta_t^{k,CC}$, has the same composition of our population parameter $\theta_t^{k,y}$ (i.e. composed of a smooth trend and a seasonal component):

$$\begin{pmatrix} \boldsymbol{y}_t^k \\ x_t^{k,CC} \\ \boldsymbol{x}_t^{k,GT} \end{pmatrix} = \begin{pmatrix} \iota_5 \theta_t^{k,y} \\ \theta_t^{k,CC} \\ \hat{\Lambda} \boldsymbol{f}_t^k \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t^k \\ 0 \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{e}_t^k \\ \varepsilon_t^{k,CC} \\ \boldsymbol{\varepsilon}_t^{k,GT} \end{pmatrix}, \begin{pmatrix} \varepsilon_t^{k,CC} \\ \boldsymbol{\varepsilon}_t^{k,GT} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{\varepsilon,CC} & \mathbf{0} \\ \mathbf{0} & \hat{\Psi} \end{bmatrix} \right), \tag{5}$$

$$\begin{pmatrix} \theta_t^{k,y} \\ \theta_t^{k,CC} \end{pmatrix} = \begin{pmatrix} L_t^{k,y} \\ L_t^{k,CC} \end{pmatrix} + \begin{pmatrix} S_t^{k,y} \\ S_t^{k,CC} \end{pmatrix}, \tag{6}$$

$$\begin{pmatrix} L_t^{k,y} \\ L_t^{k,CC} \end{pmatrix} = \begin{pmatrix} L_{t-1}^{k,y} \\ L_{t-1}^{k,CC} \end{pmatrix} + \begin{pmatrix} R_{t-1}^{k,y} \\ R_{t-1}^{k,CC} \end{pmatrix}, \tag{7}$$

$$\begin{pmatrix} R_t^{k,y} \\ R_t^{k,CC} \\ \boldsymbol{f}_t^k \end{pmatrix} = \begin{pmatrix} R_{t-1}^{k,y} \\ R_{t-1}^{k,CC} \\ \boldsymbol{f}_{t-1}^k \end{pmatrix} + \begin{pmatrix} \eta_{R,t}^{k,y} \\ \eta_{R,t}^{k,CC} \\ \boldsymbol{u}_t^k \end{pmatrix}, \tag{8}$$

$$\operatorname{cov} \begin{pmatrix} \eta_{R,t}^{k,y} \\ \eta_{R,t}^{k,CC} \\ \boldsymbol{u}_t^k \end{pmatrix} = \begin{bmatrix} \sigma^2_{R,y} & \rho_{CC}\sigma_{R,y}\sigma_{R,CC} & \rho_{1,GT}\sigma_{R,y} & \cdots & \rho_{r,GT}\sigma_{R,y} \\ \rho_{CC}\,\sigma_{R,y}\sigma_{R,CC} & \sigma^2_{R,CC} & 0 & \cdots & 0 \\ \rho_{1,GT}\sigma_{R,y} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{r,GT}\sigma_{R,y} & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{9}$$

The last equality allows the innovations of the trends' slopes, $R_t^{k,y}$ and $R_t^{k,CC}$, and of the factors of the Google Trends, to be correlated. Harvey and Chung (2000) show that there can be potential gains in precision, in terms of mean squared error (MSE) of the Kalman filter estimators of $\theta_t^{k,y}$, $L_t^{k,y}$ and $R_t^{k,y}$, if the correlation parameters $|\rho|$s are large. Specifically, if $|\rho_{CC}| = 1$, then $\boldsymbol{y}_t^k$ and $x_t^{k,CC}$ have a common slope. This means that $\boldsymbol{y}_t^k$ and $x_t^{k,CC}$ are both $I(2)$, but there is a linear combination of their first differences which is stationary. Likewise, if $|\rho_{m,GT}| = 1$ then the m^{th} factor of the Google Trends and the change in unemployment, $R_t^{k,y}$, are cointegrated (i.e. they have the same source of error). This is why we need the elements of the vector in Equation (8) to have the same order of integration and it is via this correlation parameters that we exploit the auxiliary information.

The second step of the estimation procedure consists of estimating the remaining hyperparameters of the whole state space model (Equations 3.5–3.9) by maximum likelihood and applying the Kalman filter to re-estimate $\boldsymbol{f}_t^k$ and to nowcast the variables of interest, $\theta_t^{k,y}$, $L_t^{k,y}$ and $R_t^{k,y}$, providing unemployment estimates in real-time before LFS data become available: $\widehat{\theta}_{t\,\Omega_t^-}^{k,y}$, $\widehat{L}_{t|\Omega_t^-}^{k,y}$, and $\widehat{R}_{t\,\Omega_t^-}^{k,y}$ are the filtered nowcasts of, respectively, $\theta_t^{k,y}$, $L_t^{k,y}$ and $R_t^{k,y}$ based on the information set Ω_t^- available in month t. The information set in this case is $\Omega_t^- = \{\boldsymbol{x}_t^{k,GT}, \boldsymbol{y}_{t-1}^k, x_{t-1}^{k,CC}, \boldsymbol{x}_{t-1}^{k,GT}, \dots\}$. Note that, in contrast to Section 3.1, we now talk about 'nowcast' instead of 'forecast' of $\theta_t^{k,y}$ because a part of the data (the Google Trends) used in model (5)–(9) is now available in month t.

Some remarks are in place. First, although in Section 3.1 we mentioned that Statistics Netherlands publishes only $\widehat{L}_t^{k,y}$ and $\widehat{\theta}_t^{k,y}$ as official statistics for the unemployment, we are also interested in the estimation/nowcast accuracy of $R_t^{k,y}$ since it is the state variable of the labour force model that is directly related to the auxiliary series.

Second, note that in model (3) we do not make use of the superscript k, meaning that the first step of the estimation can be performed on the high frequency (weekly in our empirical case) variables. Since in each week we can aggregate the weekly Google Trends to the monthly frequency, we can use the information available throughout the month to update the estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. If the correlations between the factors and the trend's slope of the target variable are large, this update should provide a more precise nowcast of $R_t^{k,y}$, $L_t^{k,y}$ and $\theta_t^{k,y}$.

Third, we allow the factors of the Google Trends to be correlated with the change in unemployment and not with its level for two reasons: first, a smooth trend model is assumed for the population parameter, which means that the level of its trend does not have an innovation term. Second, it is reasonable to assume that people start looking for a job on the internet when they become unemployed and hence their search behaviour should reflect the change in unemployment rather than its level.

Fourth, while our method to include auxiliary information in a state space model is based on the approach proposed by Harvey and Chung (2000), the factors of the high-dimensional auxiliary series could also be included as regressors in the observation equation for the labour force. However, in such a model, the main part of the trend, $L_t^{k,y}$, will be explained by the auxiliary series in the regression component. As a result, the filtered estimates for $L_t^{k,y}$ will contain a residual trend instead of the trend of the unemployment. Since the filtered trend estimates are the most important target variables in the official monthly publications of the labour force, this approach is not further investigated in this paper.

Finally, we refer the reader to Sections 2.1, 2.2 and 2.3 of the supplementary material for a detailed state space representation of the labour force model when, respectively, a univariate, a high-dimensional or both type of auxiliary series are included. We further refer to the working version of this paper (Schiavoni et al., 2019) for an illustration on how to include the lags of the factors and how to model their cycle or seasonality, within our proposed high-dimensional state space model.

4 | SIMULATION STUDY

We next conduct a Monte Carlo simulations study in order to elucidate to which extent our proposed method can provide gains in the nowcast accuracy of the unobserved components of interest. For this purpose, we consider a simpler model than the one used for the LFS. Here, y_t^k is univariate following a smooth trend model and $\boldsymbol{x}_t^k$ represents the (100×1)-dimensional auxiliary series with one common factor ($r = 1$).

$$\begin{aligned}
\begin{pmatrix} y_t^k \\ \boldsymbol{x}_t^k \end{pmatrix} &= \begin{pmatrix} L_t^k \\ \Lambda f_t^k \end{pmatrix} + \begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix}, \\
L_t^k &= L_{t-1}^k + R_{t-1}^k, \\
\begin{pmatrix} R_t^k \\ f_t^k \end{pmatrix} &= \begin{pmatrix} R_{t-1}^k \\ f_{t-1}^k \end{pmatrix} + \begin{pmatrix} \eta_{R,t}^k \\ u_t^k \end{pmatrix}, \begin{pmatrix} \eta_{R,t}^k \\ u_t^k \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).
\end{aligned}$$

We allow the slope and factor's innovations to be correlated and we investigate the performance of the method for increasing values of the correlation parameter $\rho \in [0, 0.2, 0.4, 0.6, 0.8, 0.9, 0.99]$. The auxiliary variable $\boldsymbol{x}_t^k$ has the same frequency of y_t^k and it is assumed that all $\boldsymbol{x}_t^k$ are released at the same time without publication delays. The nowcast is done concurrently, that is, in real-time based on a

recursive scheme. This means that in each time point of the out-of-sample period, the hyperparameters of the model are re-estimated by maximum likelihood, extending the same used up to that period. This is done in the third part of the sample, always assuming that y_t^k is not available at time t, in contrast to $\boldsymbol{x}_t^k$. This implies that the available data set in period t equals $\Omega_t^- = \{\boldsymbol{x}_t^k, y_{t-1}^k, \boldsymbol{x}_{t-1}^k, y_{t-2}^k, \ldots\}$. The sample size is $T = 150$ and the number of simulations are $n_{\text{sim}} = 500$.

We consider three specifications for the idiosyncratic components and the factor loadings:

1. Homoskedastic idiosyncratic components and dense loadings:

$$\begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix} \sim N(\mathbf{0},\ 0.5\boldsymbol{I}_{n+1}), \quad \Lambda \sim U(0,\ 1).$$

2. Homoskedastic idiosyncratic components and sparse loadings. The first half of the elements in the loadings are set equal to zero. This specification reflects the likely empirical case that some of the Google Trends are not related to the change in unemployment:

$$\begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix} \sim N(\mathbf{0}, 0.5\boldsymbol{I}_{n+1}), \Lambda = (\Lambda_0', \Lambda_1')', \underset{50\times1}{\Lambda_0} = \mathbf{0}, \underset{50\times1}{\Lambda_1} \sim U(0, 1).$$

3. Heteroskedastic idiosyncratic components and dense loadings. The homoskedasticity assumption is here relaxed, again as not being realistic for the job-search terms:

$$\begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} 0.5 & \mathbf{0}' \\ \mathbf{0} & \text{diag}(H) \end{pmatrix}\right), \quad H \sim U(0.5,\ 10), \quad \Lambda \sim U(0,\ 1).$$

Let $\boldsymbol{\alpha}_t^k = (L_t^k, R_t^k, f_t^k)'$ denote the vector of state variables and $\hat{\boldsymbol{\alpha}}_{t|\Omega_t^-}^k$ its estimates based on the information available at time t. The results from the Monte Carlo simulations are shown in Table 2. We always report the MSFE, together with its variance and bias components, of the Kalman filter estimator of $\boldsymbol{\alpha}_t^k$ (since our goal is not to predict y_t^k, but its state variables), relative to the same measures calculated from the model that does not include the auxiliary series $\boldsymbol{x}_t^k$. Recall that the latter comes down to making one-step-ahead predictions.

$$\begin{aligned}
\text{MSFE}(\hat{\boldsymbol{\alpha}}_{t|\Omega_t^-}^k) &= \frac{1}{h}\sum_{t=T-h+1}^{T} \frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}} \left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-} - \boldsymbol{\alpha}_{jt}\right)\left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-} - \boldsymbol{\alpha}_{jt}\right)', \\
\text{var}(\hat{\boldsymbol{\alpha}}_{t|\Omega_t^-}^k) &= \frac{1}{h}\sum_{t=T-h+1}^{T} \left(\frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}} \left(\left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-} - \boldsymbol{\alpha}_{jt}\right) - \frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}}\left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-} - \boldsymbol{\alpha}_{jt}\right)\right)\right) \\
&\quad \left(\times\left(\left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-} - \boldsymbol{\alpha}_{jt}\right) - \frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}}\left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-} - \boldsymbol{\alpha}_{jt}\right)\right)'\right), \\
\text{bias}^2(\hat{\boldsymbol{\alpha}}_{t\,\Omega_t^-}^k) &= \frac{1}{h}\sum_{t=T-h+1}^{T} \left(\frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}} \hat{\boldsymbol{\alpha}}_{jt\,\Omega_t^-} - \boldsymbol{\alpha}_{jt}\right)\left(\frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}} \hat{\boldsymbol{\alpha}}_{jt\,\Omega_t^-} - \boldsymbol{\alpha}_{jt}\right)',
\end{aligned}$$

where h is the size of the of out-of-sample period.

TABLE 2 Simulation results from the three settings described in Section 4. The parameter ρ represents the correlation between the states' innovations. We indicate with $\widehat{L}^k_{t|\Omega_t^-}$ and $\widehat{R}^k_{t|\Omega_t^-}$ the filtered nowcasts of, respectively, the level and the slope of the trend of the target variable, y^k_t. The values are reported relative to the respective measures calculated from the model that does not include the auxiliary series; values <1 are in favour of our method. $n_{\text{sim}} = 500$

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.99$
	Homoskedastic idiosyncratic components and dense loadings						
$\text{MSFE}(\widehat{L}^k_{t\|\Omega_t^-})$	1.030	1.024	1.006	0.971	0.901	0.837	0.718
$\text{var}(\widehat{L}^k_{t\|\Omega_t^-})$	1.031	1.025	1.007	0.971	0.901	0.837	0.718
$\text{bias}^2(\widehat{L}^k_{t\|\Omega_t^-})$	0.775	0.767	0.756	0.733	0.692	0.659	0.567
$\text{MSFE}(\widehat{R}^k_{t\|\Omega_t^-})$	1.044	1.017	0.941	0.806	0.588	0.427	0.198
$\text{var}(\widehat{R}^k_{t\|\Omega_t^-})$	1.045	1.018	0.942	0.807	0.589	0.427	0.198
$\text{bias}^2(\widehat{R}^k_{t\|\Omega_t^-})$	0.650	0.633	0.583	0.492	0.350	0.252	0.122
	Homoskedastic idiosyncratic components and sparse loadings						
$\text{MSFE}(\widehat{L}^k_{t\|\Omega_t^-})$	1.031	1.026	1.011	0.981	0.920	0.862	0.744
$\text{var}(\widehat{L}^k_{t\|\Omega_t^-})$	1.031	1.026	1.012	0.981	0.920	0.862	0.745
$\text{bias}^2(\widehat{L}^k_{t\|\Omega_t^-})$	0.784	0.776	0.762	0.737	0.695	0.655	0.582
$\text{MSFE}(\widehat{R}^k_{t\|\Omega_t^-})$	1.044	1.019	0.946	0.817	0.605	0.446	0.208
$\text{var}(\widehat{R}^k_{t\|\Omega_t^-})$	1.045	1.020	0.947	0.817	0.606	0.446	0.209
$\text{bias}^2(\widehat{R}^k_{t\|\Omega_t^-})$	0.656	0.639	0.586	0.492	0.347	0.243	0.104
	Heteroskedastic idiosyncratic components and dense loadings						
$\text{MSFE}(\widehat{L}^k_{t\|\Omega_t^-})$	1.036	1.032	1.019	0.994	0.945	0.901	0.823
$\text{var}(\widehat{L}^k_{t\|\Omega_t^-})$	1.037	1.032	1.020	0.995	0.946	0.902	0.823
$\text{bias}^2(\widehat{L}^k_{t\|\Omega_t^-})$	0.707	0.645	0.579	0.521	0.484	0.483	0.543
$\text{MSFE}(\widehat{R}^k_{t\|\Omega_t^-})$	1.049	1.027	0.960	0.840	0.644	0.499	0.299
$\text{var}(\widehat{R}^k_{t\|\Omega_t^-})$	1.049	1.028	0.961	0.841	0.645	0.500	0.299
$\text{bias}^2(\widehat{R}^k_{t\|\Omega_t^-})$	0.805	0.697	0.556	0.397	0.230	0.161	0.237

In every setting, both the bias and the variance of the MSFE tend to decrease with the magnitude of the correlation parameter. The improvement is more pronounced for the slope rather than the level of the trend. For the largest value of the correlation, with respect to the model which does not include auxiliary information, the gain in MSFE for the level and the slope is, respectively, of around 25% and 75%. Moreover, for low values of ρ ($\rho \leq 0.2$), the MSFE does not deteriorate with respect to the benchmark model. This implies that our proposed method is robust to the inclusion of auxiliary information that does not have predictive power for the state variables of interest. In Section 4 of the supplementary material we report and examine additional simulation results with non-Gaussian idiosyncratic components, and draw the same conclusions discussed above for the MSFE and the variance of the state variables' nowcasts. The bias instead worsens while deviating from Gaussianity, but it does not affect the MSFE as it only accounts for a small part of the latter measure. We, therefore, conclude that the performance of our method is overall robust to deviations from Gaussianity of the idiosyncratic components.

The decision to focus the simulation study on the nowcast (rather than the in-sample) performance of our method, is motivated by the fact that the added value of the Google Trends over the claimant counts is their real-time availability, which can be used to nowcast the unemployment. Nonetheless, for completeness, in the empirical application of the next section we report the results also for the in-sample performance of our method.

5 | APPLICATION TO DUTCH UNEMPLOYMENT NOWCASTING

In this section, we present and discuss the results of the empirical application of our method to nowcasting the Dutch unemployment using the auxiliary series of claimant counts and Google Trends related to job-search and economic uncertainty.

As explained in Section 3.2, the Google series used in the model must be $I(1)$. We, therefore, test for nonstationarity in the Google Trends with the Elliott et al. (1996) augmented Dickey–Fuller (ADF) test, including a constant and a linear trend. We control for the false discovery rate as in Moon and Perron (2012), who employ a moving block bootstrap approach that accounts for time and cross-sectional dependence among the units in the panel.

Before proceeding with the estimation of the model by only including the Google Trends that resulted as being $I(1)$ from the multiple hypotheses testing, we carry out an additional selection of the $I(1)$ Google Trends by 'targeting' them as explained and motivated in what follows. Bai and Ng (2008) point out that having more data to extract factors from is not always better. In particular, if series are added that have loadings of zero and are thus not influenced by the factors, these will make the estimation of factors and loadings by PCA deteriorate, as PCA assigns a non-zero weight to each series in calculating the estimated factor as a weighted average. Bai and Ng (2008) recommend a simple strategy to filter out irrelevant series (in our case Google search terms) and improve the estimation of the factors, which they call 'targeting the predictors'. In this case an initial regression of the series of interest is performed on the high-dimensional input series to determine which series are (ir)relevant. The series that are found to be irrelevant are discarded and only the ones that are found to be relevant are kept to estimate the factors and loadings from. In particular, they recommend the use of the elastic net (Hastie & Zou, 2005), which is a penalised regression technique that performs estimation and variable selection at the same time by setting the coefficients of the irrelevant variables to 0 exactly. After performing the elastic net estimation, only the variables with non-zero coefficients are then kept. As we do not observe our series of interest directly, we need to adapt their procedure to our setting. To do so we approximate the unobserved unemployment by its estimation from the labour force model without auxiliary series. Specifically, we regress the differenced estimated change in unemployment from the labour force model without auxiliary series, $\Delta\widehat{R}_t^{k,y}$, on the differenced $I(1)$ Google Trends using the elastic net penalised regression method, which solves the following minimisation problem:

$$\boldsymbol{\beta}\{\widehat{min}\left[\frac{1}{2T}\sum_{t=1}^{T}(\Delta\widehat{R}_t^{k,y}-\boldsymbol{\beta}'\Delta\boldsymbol{x}_t^{k,GT})^2+\lambda P_\alpha(\boldsymbol{\beta})\right],$$

where

$$P_\alpha(\boldsymbol{\beta})=(1-\alpha)\frac{1}{2}\ \boldsymbol{\beta}\ _2^2+\alpha\ \boldsymbol{\beta}\ _1.$$

The tuning parameters λ and α are selected from a two-dimensional grid in order to minimise the Schwarz (1978) Bayesian information criterion (BIC). Notice that performing the penalised regression on the differenced (and, therefore, stationary) data, also allows us to avoid the inclusion in the model of Google Trends that have spurious relations with the change in unemployment. We both consider estimating the final model with all Google Trends included and with only the selected Google Trends included, thereby allowing us to assess the empirical effects of targeting. The final number of nonstationary Google Trends included in the model, n, may differ depending on whether we use the weekly Google Trends aggregated to the monthly frequency according to Equation (4), or the monthly Google Trends. Whenever we apply PCA, the Google Trends are first differenced and standardised.

We further need to make sure that the stationarity assumption of the idiosyncratic components is maintained. Therefore, after having estimated the factors by PCA in model (3), we test which of the idiosyncratic components $\boldsymbol{\varepsilon}_t$ are $I(1)$ with an ADF test without deterministic components, by controlling for multiple hypotheses testing as in Moon and Perron (2012). The $I(1)$ idiosyncratic components are modelled as state variables in Equation (5), with the following transition equation:

$$\boldsymbol{\varepsilon}_t^k = \boldsymbol{\varepsilon}_{t-1}^k + \boldsymbol{\xi}_t^k,$$

with usual normality assumptions on the $\boldsymbol{\xi}_t^k$. The covariance matrix of the idiosyncratic components $\boldsymbol{\Psi}$ is, therefore, estimated on the levels of the $I(0)$ idiosyncratic components and the first differences of the $I(1)$ idiosyncratic components. Section 2.2.1 of the supplementary material provides a toy example that elucidates the estimation procedure.

Finally, we here propose two different ways to obtain (possibly) more accurate estimates of $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$, since their estimation accuracy may affect the predictive power of the Google Trends. First, in Section 3.2 we mention that the first step of the two-step estimator, which estimates $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$ by PCA, can be carried out on the weekly Google Trends (which are, therefore, aggregated to the monthly frequency after the first step). Since the sample size of the high frequency data is larger, using weekly Google Trends might improve the estimation accuracy of $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$. Second, Doz et al. (2011) argue that from the Kalman filter estimates of the factors (obtained in the second step), it is possible to re-estimate $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$ (by least squares), which in turn can be used to re-estimate the factors and so on. Continuing the iterative procedure until convergence is equivalent to the Expectation–Maximisation (EM) algorithm which increases the likelihood at each step and, therefore, converges to the maximum likelihood solution. This however is very hard computationally. Therefore, we only iterate once as a trade-off between gaining estimation accuracy from iteration and computation cost. The iteration of the procedure provides an estimator that should be *only asymptotically* equivalent to the maximum likelihood estimator, since we start the procedure using consistent estimates of the parameters (Davidson, 2000, Chapter 9). Later in this section we check how sensitive our empirical results are to the different estimates of $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$.

We present empirical results for the in-sample estimates and out-of-sample forecasts. With the in-sample estimates we evaluate to which extent the auxiliary series improve the precision of the published monthly unemployment estimates after finalising the data collection. With the out-of-sample forecasts we evaluate to which extent the auxiliary series improve the precision of provisional estimates in a nowcast procedure during the period of data collection. We always estimate four different models: the labour force model without auxiliary series (baseline), the labour force model with auxiliary series of claimant counts (CC), of Google Trends (GT) and of both (CC & GT). We compare the latter three models to the baseline one with the in-sample and out-of-sample exercises. The period considered for the estimation starts in January 2004 and ends in May 2019 ($T = 185$ months). The out-of-sample nowcasts are conducted in real-time (concurrently) in the last 3 years of the sample

based on a recursive scheme: each week or month, depending on whether we use weekly or monthly Google Trends, the model, including its hyperparameters, is re-estimated on the enlarged sample now extended by the latest observations, while assuming that the current observations for the unemployed labour force and the claimant counts are missing. Analogously, when the Google Trends are first targeted with the elastic net, the targeting is re-executed in each week or month of the out-of-sample period on the updated sample.

We define the measure of in-sample estimation accuracy $\widehat{\mathrm{MSE}}(\widehat{\boldsymbol{\alpha}}^k_{t|\Omega_t}) = \frac{1}{T-d}\sum_{t=d+1}^{T}\widehat{\boldsymbol{P}}^k_{t|\Omega_t}$, where $\widehat{\boldsymbol{\alpha}}^k_{t|\Omega_t}$ is the vector of Kalman filter estimates of the state variables, $\widehat{\boldsymbol{P}}_{t|\Omega_t}$ is its estimated covariance matrix in month t, and d is the number of state variables that are needed to estimate the labour force model without auxiliary series, and that need a diffuse initialisation for their estimation ($d = 17$). The measure of nowcast accuracy, $\widehat{\mathrm{MSFE}}(\widehat{\boldsymbol{\alpha}}^k_{t|\Omega_t^-}) = \frac{1}{h}\sum_{t=T-h+1}^{T}\widehat{\boldsymbol{P}}^k_{t|\Omega_t^-}$, is the average of the nowcasted covariance matrices in the h prediction months. When weekly Google Trends are used, $\widehat{\boldsymbol{P}}^k_{t|\Omega_t^-} = \frac{1}{k}\sum_{j=1}^{k}\widehat{\boldsymbol{P}}^k_{j|\Omega_{j,t}^-}$, where $\widehat{\boldsymbol{P}}^k_{j|\Omega_{j,t}^-}$ is the nowcasted covariance matrix for the prediction in week j of month t and $\Omega_{j,t}^- = \{\boldsymbol{x}^{k,GT}_{j,t}, \boldsymbol{y}^k_{t-1}, x^{k,CC}_{t-1}, \boldsymbol{x}^{k,GT}_{t-1}, \ldots\}$ is in this case the available information set in week j of month t. This is because the nowcast is done recursively throughout the weeks of the out-of-sample period. We always report the relative $\widehat{\mathrm{MS(F)E}}$ with respect to the baseline model; values lower than one are in favour of our method. We note that nowcasting under the baseline model without auxiliary series and the baseline model extended with claimant counts comes down to making one-step-ahead predictions. Expressions for $\widehat{\boldsymbol{\alpha}}^k_{t|\Omega_t}$, $\widehat{\boldsymbol{\alpha}}^k_{t|\Omega_t^-}$ and their covariance matrices, $\widehat{\boldsymbol{P}}^k_{t|\Omega_t}$ and $\widehat{\boldsymbol{P}}^k_{t|\Omega_t^-}$, are given by the standard Kalman filter recursions, see for example, Durbin and Koopman (2012) Chapter 4).

The initial values of the hyperparameters for the maximum likelihood estimation are equal to the estimates for the labour force model obtained in van den Brakel and Krieg (2015). We use a diffuse initialisation of the Kalman filter for all the state variables except for the 13 state variables that define the autocorrelation structure of the survey errors, for which we use the exact initialisation of Bollineni-Balabay et al. (2017).

We use the three panel information criteria proposed by Bai and Ng (2002) which we indicate, as in Bai and Ng (2002), with IC_1, IC_2 and IC_3, in order to choose how many factors of the Google Trends to include in the model (in this paper, if for instance the information criterion IC_1 suggests to include two factors, we indicate as $IC_1 = 2$). When the Google Trends are targeted with the elastic net, the information criteria suggest to include one or two factors. In the empirical analysis we check the sensitivity of the results with respect to these two different numbers of factors included in the model.

We employ a Wilks (1938) likelihood ratio (LR) test to assess whether the correlation parameters are significantly different from zero and hence adding the auxiliary information might yield a significant improvement from the baseline model. Specifically, we indicate with $\rho_{CC} = 0$, $\rho_{1,GT} = 0$ and $\rho_{2,GT} = 0$ the null hypotheses for the individual insignificance of the correlation parameter with, respectively, the claimant counts, and the first and second factor (when present) of the Google Trends. With $\boldsymbol{\rho}_{GT} = \mathbf{0}$ and $\boldsymbol{\rho} = \mathbf{0}$ we instead indicate the null hypotheses for the joint insignificance of, respectively, the correlations with the Google Trends' factors and all correlation parameters. If the true distribution of the error terms is non-Gaussian, the LR test, based on the QML estimates, does not generally keep having, under the null hypothesis, an asymptotic χ^2 distribution with degrees of freedom equal to the number of restrictions. One exception is when the covariance matrix of the error terms from a regression involving observed variables, is replaced by a consistent estimator prior to the maximisation of the log-likelihood (Gourieroux & Monfort, 1993). In our case, if the idiosyncratic components of the Google Trends, $\boldsymbol{\varepsilon}^{k,GT}_t$, are the only error terms not being normally distributed, we may fall into this exception. The covariance matrix $\boldsymbol{\Psi}$ is indeed replaced, for the maximisation of the log-likelihood, by its consistent PCA estimator obtained in the first step of the two-step estimation procedure. Nonetheless, in the setting of Gourieroux and Monfort (1993) the regressors are observed,

whereas in our case the latter are the unobserved factors. Consequently, it is not trivial to assess whether our model specification indeed falls into the above-mentioned exception. A formal proof for this is beyond the scope of this paper, but in Section 4 of the supplementary material we conduct a simulation study in order to obtain the finite-sample probability density of the LR test under misspecifications of the distribution of the idiosyncratic components. We conclude that the distribution of the LR test is not affected by these misspecifications. At the end of this section we show that there is no evidence that the error terms other than $\boldsymbol{\varepsilon}_t^{k,GT}$, are not normally distributed. We should, therefore, be able to perform inference based on the usual asymptotic distribution of the LR test.

Table 3 reports the estimated hyperparameters for the four models, as well as the respective value for the maximised log-likelihood, the relative measures of in and out-of-sample performance, and the *p*-values from the LR tests, when the monthly Google Trends are used. The maximum likelihood estimates for the standard error of the seasonal components' disturbance terms tend to zero, indicating that the seasonal effects are time invariant. Recall from Equation (2) that the variances of the scaled sampling errors, $\sigma_{v_j}^2$, should take values close to one. Their estimates are divided by $(1-\widehat{\delta}^2)$ and are always slightly larger than one, which is an indication that the variance estimates of the GREG estimates, used to scale the sampling errors in Equation (2), somewhat underestimate the real variance of the GREG estimates.

The correlation with the claimant counts is estimated to be above 0.9 and remains large and significant when including the Google Trends. Similar conclusions can be drawn for the correlations with the Google Trends' factors, when the Google Trends are targeted with the elastic net and 39 of them are included in the model, although we do not attach any causal meaning to this finding. When the additional targeting is not applied and the 162 $I(1)$ Google Trends are directly included in the model, the correlation parameter with the first factor of the Google Trends is instead always small and insignificant (in this setting we do not include more than one factor). Moreover, for the same number of factors, targeting the Google Trends always yields a better performance in terms of estimation and nowcast accuracy of the state variables of interest, with respect to not targeting them. For this reason, we focus the remaining analysis of the empirical results only on the targeted Google Trends.

The best results in terms of both estimation and nowcast accuracy of all state variables, is achieved by the CC & GT model with one factor, yielding a gain of, respectively, around 40% and 20% for $\widehat{R}_t^{k,y}$, and around 20% and 23% for both $\widehat{L}_t^{k,y}$ and $\widehat{\theta}_t^{k,y}$, with respect to the baseline model. Note that this implies that the above-mentioned model outperforms also the model that contains only the claimant counts as auxiliary series. In particular, by comparing the relative $\widehat{\text{MSE}}$ and $\widehat{\text{MSFE}}$ of the CC & GT model to the CC one, it is possible to see that the gains of the former model over the latter are of 10% (for $\widehat{L}_{t|\Omega_t}^{k,y}$ and $\widehat{\theta}_{t|\Omega_t}^{k,y}$) and 30% (for $\widehat{R}_{t|\Omega_t}^{k,y}$) in terms of in-sample estimation accuracy, and 18% (for $\widehat{R}_{t|\Omega_t^-}^{k,y}$) and 6% (for $\widehat{L}_{t|\Omega_t^-}^{k,y}$ and $\widehat{\theta}_{t|\Omega_t^-}^{k,y}$) in terms of nowcast accuracy. Moreover, the Google Trends alone improve the estimation and nowcast accuracy, respectively, of 35% and 30% for $\widehat{R}_t^{k,y}$, and around 7% and 12% for $\widehat{L}_t^{k,y}$ and $\widehat{\theta}_t^{k,y}$, with respect to the baseline model. Finally, the CC model improves both the estimation and nowcast accuracy for $\widehat{L}_t^{k,y}$ and $\widehat{\theta}_t^{k,y}$ of 5%, with respect to the GT model, but the same measures for $\widehat{R}_t^{k,y}$ are around 25%–30% worse. In general, the models with Google Trends tend to achieve a better estimation and nowcast of the change in unemployment, $R_t^{k,y}$, rather than the other two state variables, with respect to the models that include the claimant counts.

Including two instead of one factor clearly increases the complexity of the model, which is reflected in smaller accuracy gains (in the CC & GT model probably also due to the decreased magnitude of the correlation parameter with the claimant counts), especially for the nowcast of the state variables, with respect to including only one factor. Nonetheless, the correlations with both factors are individually and jointly significantly different from zero, indicating that both factors bring additional information that helps in predicting the Dutch unemployment.

TABLE 3 Estimation and nowcast results for the labour force model with and without auxiliary series. The auxiliary series are the claimant counts and the monthly Google Trends about job-search and economic uncertainty. We denote with 'LF' the model that does not contain auxiliary series. 'CC' and 'GT' stand for the models that contain, respectively, the claimant counts and the Google Trends as auxiliary series, whereas 'CC & GT' is the model that contains both types of auxiliary series. The number of Google Trends and the number of their factors included in the model are denoted with n and r, respectively. The abbreviation 'all corr.' denotes that the correlation between the claimant counts and the Google Trends is also estimated. 'Targeted GT' indicates that the Google Trends have been targeted with the elastic net before including them in the model. The state variables $\theta_t^{k,y}$, $L_t^{k,y}$ and $R_t^{k,y}$ are, respectively, the Dutch unemployment and its trend's level and slope. Please refer to Section 3 for definitions of the model's parameters and for an explanation of the difference between estimates and nowcasts of the state variables

			$n = 162, IC_1 = 3, IC_2 = 1, IC_3 = 10$		Targeted GT, $n = 39, IC_1 = 2, IC_2 = 1, IC_3 = 2$				
			$r = 1$		$r = 1$			$r = 2$	
	LF	**CC**	**GT**	**CC & GT**	**GT**	**CC & GT**	**CC & GT, all corr.**	**GT**	**CC & GT**
$\hat{\sigma}_{R,y}$	2082.652	2776.030	1995.917	2704.918	3036.281	2608.394	3447.973	3587.985	3002.947
$\hat{\sigma}_{\omega,y}$	0.020	0.020	0.023	0.078	0.013	0.011	0.022	0.054	0.010
$\hat{\sigma}_{\lambda}$	3841.035	3883.658	3592.394	3715.303	3740.097	3740.748	3115.596	3670.943	3709.361
$\hat{\sigma}_{v_1}$	1.140	1.151	1.181	1.146	1.155	1.142	1.205	1.155	1.198
$\hat{\sigma}_{v_2}$	1.291	1.300	1.270	1.359	1.276	1.304	1.378	1.281	1.263
$\hat{\sigma}_{v_3}$	1.188	1.181	1.201	1.211	1.188	1.196	1.117	1.224	1.200
$\hat{\sigma}_{v_4}$	1.240	1.247	1.241	1.224	1.241	1.252	1.356	1.286	1.243
$\hat{\sigma}_{v_5}$	1.223	1.228	1.236	1.260	1.221	1.239	1.358	1.254	1.247
$\hat{\delta}$	0.384	0.381	0.378	0.395	0.377	0.384	0.390	0.383	0.384
$\hat{\sigma}_{R,CC}$		3490.261		3515.222		3503.077	3982.583		3979.232
$\hat{\sigma}_{\omega,CC}$		0.020		0.020		0.021	0.016		0.020
$\hat{\sigma}_{\varepsilon,CC}$		1318.691		1310.108		1309.136	1181.291		1052.729
$\hat{\rho}_{CC}$		0.918		0.913		0.803	0.935		0.755
$\hat{\rho}_{1,GT}$			−0.200	−0.003	−0.899	−0.509	−0.250	−0.785	−0.381
$\hat{\rho}_{2,GT}$								−0.591	−0.456

(Continues)

Table 3 (Continued)

			$n = 162, IC_1 = 3, IC_2 = 1, IC_3 = 10$		Targeted GT, $n = 39, IC_1 = 2, IC_2 = 1, IC_3 = 2$				
			$r = 1$		$r = 1$			$r = 2$	
	LF	CC	GT	CC & GT	GT	CC & GT	CC & GT, all corr.	GT	CC & GT
$\widehat{\rho}_{1,CC,GT}$							−0.093		
$\widehat{\mathrm{MSE}}(\widehat{L}^{k,y}_{\vert\Omega_t})$		0.868	1.003	0.863	0.919	0.796	0.895	0.861	0.849
$\widehat{\mathrm{MSE}}(\widehat{R}^{k,y}_{\vert\Omega_t})$		0.878	0.916	0.849	0.655	0.618	1.112	0.485	0.702
$\widehat{\mathrm{MSE}}(\widehat{\theta}^{k,y}_{\vert\Omega_t})$		0.889	1.009	0.888	0.941	0.835	0.916	0.899	0.881
$\widehat{\mathrm{MSFE}}(\widehat{L}^{k,y}_{t\vert\Omega_t^-})$		0.818	0.951	0.853	0.875	0.766	0.786	0.935	0.889
$\widehat{\mathrm{MSFE}}(\widehat{R}^{k,y}_{t\vert\Omega_t^-})$		0.983	0.878	0.981	0.705	0.801	0.755	0.839	0.869
$\widehat{\mathrm{MSFE}}(\widehat{\theta}^{k,y}_{\Omega_t^-})$		0.827	0.956	0.860	0.886	0.779	0.796	0.942	0.899
log-likelihood	−10160.378	−11835.779	−44726.153	−46392.985	−18712.139	−20379.780	−20384.560	−18719.515	−20388.495
$t\vert$		p-value from the LR test							
$H_0: \rho_{CC} = 0$		0.002		0.000		0.001	0.000		0.000
$H_0: \rho_{1,GT} = 0$			0.470	0.830	0.001	0.025	0.028	0.000	0.082
$H_0: \rho_{2,GT} = 0$								0.014	0.014
$H_0: \boldsymbol{\rho}_{GT} = \mathbf{0}$								0.000	0.017
$H_0: \rho_{1,CC,GT} = 0$							0.470		
$H_0: \boldsymbol{\rho} = \mathbf{0}$				0.001		0.000	0.000		0.000

Notice that in general all the relative measures of accuracy are below one, indicating that both the claimant counts and the Google Trends improve the estimation and nowcast accuracy of the unemployment and its change. Even when the Google Trends are not targeted and the correlation parameter with their factor is not significantly different from zero, the measures are never drastically above one, meaning that our method tends to ignore auxiliary series that are not helpful in predicting the target variable.

Finally, when we specified the covariance matrix (9) in Section 3.2, we did not let the claimant counts and the Google Trends be correlated because our goal is to improve the estimation/nowcast accuracy of the unobserved components of the labour force series, not of the claimant counts nor the Google Trends. Nonetheless, if the state variables of Equation (8) are all cointegrated (i.e. the correlation parameters are all equal to one) a more efficient estimation method would be to only estimate the variance of their common source of error. We, therefore, estimate the CC & GT model with one factor, when all series are correlated. We call this model 'CC & GT all corr'. Table 3 reports the empirical results also for this model. Although the nowcast accuracy is similar to that of the same model without the additional correlation between the claimant counts and the Google Trends (which we indicate as $\rho_{1,CC,GT}$), the in-sample accuracy deteriorates (even with respect to the baseline model) and $\rho_{1,CC,GT}$ is not significantly different from zero. We, therefore, conclude that the specification of the covariance matrix (9) is appropriate.

In Table 4 we report the empirical results for the GT and CC & GT models which employ the targeted Google Trends observed at the weekly frequency and aggregated to the monthly frequency according to Equation (4) in order to include them in the models. In this case we still look at the sensitivity of the results with respect to the number of factors included in the model, but also with respect to the two additional methods for the estimation of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ discussed at the beginning of this section. The measures of accuracy are again broadly lower than one, but the gains are not as large as observed for the monthly Google Trends. Including two factors improves the accuracy in the GT model, but not in the CC & GT model, except for a more precise nowcast of $R_t^{k,y}$. The correlation parameter with the claimant counts remains large and significant. In contrast, the correlation parameter with the first factor of the Google Trends is not significantly different from zero and there is a weak evidence for the correlation parameter with the second factor to be significantly different from zero. For this reason we continue the analysis by considering two factors in the model.

Estimating $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ on the weekly Google Trends improves the measures of accuracy only for the CC & GT model and not for the GT model. An additional iteration of the two step estimator, in order to obtain more accurate estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$, achieves instead better nowcasts for both the GT and the CC & GT models (and also better in-sample estimates for the latter model), and a similar performance to the models which employ the monthly Google trends and include two factors. Notice that the values of the log-likelihood for these two models increased with respect to the same model specifications that use the original two-step estimation (without the additional iteration). The latter result, as pointed out in the explanation of the iterated estimation of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ at the beginning of this section, is to be expected. Despite the above-mentioned improvements in estimation/nowcast accuracy, the correlation parameters with the Google Trends' factors are always insignificant. The aggregation of the Google Trends from the weekly to the monthly frequency yields time series that are more noisy with respect to the Google Trends that are directly observed at the monthly frequency and detecting significant results, therefore, becomes harder.

Finally, even though weekly Google Trends allow to perform the monthly nowcasts on a weekly basis, we notice that, in general, the precision of the nowcast does not monotonically improve with the number of weeks. If the high-dimensional state space model could be expressed and estimated on the highest frequency, the weekly gains in nowcast accuracy could be more evident. Nonetheless, we

TABLE 4 Estimation and nowcast results for the labour force model with auxiliary series of claimant counts and weekly Google Trends about job-search and economic uncertainty (aggregated to the monthly frequency according to Equation (4)). The number of Google Trends and the number of their factors included in the model are denoted with n and r, respectively,. 'Weekly $\hat{\Lambda}, \hat{\Psi}$' denotes that the latter estimates are obtained using the weekly Google Trends. 'Iterated $\hat{\Lambda}, \hat{\Psi}$' means that the latter estimates are obtained from an additional iteration of the two-step estimator. 'Targeted GT' indicates that the Google Trends have been targeted with the elastic net before including them in the model. Please refer to Section 3 for definitions of the model's parameters and for an explanation of the difference between estimates and nowcasts of the state variables

	Targeted GT, $n = 37$, $IC_1 = 1$, $IC_2 = 1$, $IC_3 = 2$							
	$r = 1$		$r = 2$					
					Weekly $\hat{\Lambda}, \hat{\Psi}$		**Iterated $\hat{\Lambda}, \hat{\Psi}$**	
	GT	**CC & GT**	**GT**	**CC & GT**	**GT**	**CC & GT**	**GT**	**CC & GT**
$\hat{\sigma}_{R,y}$	2020.195	2644.552	2590.937	3671.191	1995.064	2612.712	2238.557	2745.331
$\hat{\sigma}_{\omega,y}$	0.014	0.006	0.027	0.020	0.037	0.020	0.016	0.018
$\hat{\sigma}_{\lambda}$	3604.274	3738.299	3638.503	4281.357	3640.527	3568.508	3616.609	3635.421
$\hat{\sigma}_{v_1}$	1.146	1.151	1.142	1.181	1.161	1.147	1.155	1.148
$\hat{\sigma}_{v_2}$	1.295	1.286	1.292	1.376	1.294	1.294	1.278	1.312
$\hat{\sigma}_{v_3}$	1.203	1.171	1.208	1.211	1.167	1.204	1.207	1.199
$\hat{\sigma}_{v_4}$	1.253	1.225	1.248	1.358	1.247	1.274	1.252	1.267
$\hat{\sigma}_{v_5}$	1.240	1.179	1.234	1.227	1.244	1.225	1.231	1.243
$\hat{\delta}$	0.390	0.371	0.385	0.412	0.380	0.384	0.388	0.386
$\hat{\sigma}_{R,CC}$		3491.025		3635.234		3494.779		3508.248
$\hat{\sigma}_{\omega,CC}$		0.019		0.018		0.017		0.018
$\hat{\sigma}_{\varepsilon,CC}$		1342.202		1302.024		1280.971		1302.781
$\hat{\rho}_{CC}$		0.882		0.578		0.858		0.886
$\hat{\rho}_{1,GT}$	0.173	−0.054	0.441	−0.286	−0.101	−0.226	−0.245	0.275
$\hat{\rho}_{2,GT}$			0.539	−0.687	0.212	−0.030	0.371	0.015
$\widehat{\text{MSE}}(\hat{L}^{k,y}_{t\|\Omega_t})$	0.985	0.878	0.976	0.998	0.989	0.878	0.994	0.843
$\widehat{\text{MSE}}(\hat{R}^{k,y}_{t\|\Omega_t})$	0.936	0.872	0.904	0.996	0.912	0.836	0.961	0.799
$\widehat{\text{MSE}}(\hat{\theta}^{k,y}_{t\|\Omega_t})$	0.991	0.896	0.984	1.007	0.996	0.900	0.998	0.872
$\widehat{\text{MSFE}}(\hat{L}^{k,y}_{t\|\Omega_{t^-}})$	0.990	0.817	0.909	0.906	1.008	0.858	0.914	0.895
Week 1	0.988	0.811	0.928	0.890	1.005	0.860	0.909	0.897
Week 2	0.989	0.827	0.894	0.899	1.015	0.864	0.910	0.873
Week 3	0.993	0.811	0.901	0.932	0.993	0.847	0.895	0.943
Week 4	0.995	0.816	0.911	0.894	1.011	0.862	0.948	0.870
Week 5	0.969	0.823	0.920	0.932	1.032	0.858	0.897	0.894
$\widehat{\text{MSFE}}(\hat{R}^{k,y}_{t\|\Omega_{t^-}})$	0.965	0.982	0.833	0.843	0.930	0.840	0.830	0.819
Week 1	0.975	0.981	0.856	0.860	0.912	0.843	0.845	0.839
Week 2	0.972	0.991	0.835	0.832	0.948	0.862	0.824	0.812
Week 3	0.956	0.954	0.816	0.832	0.922	0.831	0.806	0.821

(Continues)

Table 4 (Continued)

	Targeted GT, $n = 37, IC_1 = 1, IC_2 = 1, IC_3 = 2$							
	$r = 1$		$r = 2$					
					Weekly $\hat{\Lambda}, \hat{\Psi}$		Iterated $\hat{\Lambda}, \hat{\Psi}$	
	GT	CC & GT	GT	CC & GT	GT	CC & GT	GT	CC & GT
Week 4	0.967	0.987	0.823	0.844	0.937	0.817	0.850	0.811
Week 5	0.934	1.021	0.834	0.852	0.931	0.867	0.818	.794
$\widehat{\text{MSFE}}(\hat{\theta}^{k,y}_{t\|\Omega_{i-}})$	0.991	0.825	0.917	0.937	0.994	0.873	0.897	0.902
Week 1	0.990	0.820	0.933	0.943	1.006	0.876	0.908	0.894
Week 2	0.991	0.835	0.928	0.963	0.980	0.873	0.890	0.886
Week 3	0.995	0.820	0.905	0.933	1.010	0.860	0.892	0.959
Week 4	0.996	0.825	0.905	0.908	1.008	0.884	0.891	0.882
Week 5	0.970	0.830	0.903	0.944	0.911	0.867	0.917	0.871
Log-likelihood	−17954.398	−19621.000	−17767.456	−19438.409	−18651.434	−20318.457	−17745.335	−19413.916
	p-value from the LR test							
$H_0: \rho_{CC} = 0$		0.001		0.000		0.001		0.000
$H_0: \rho_{1,GT} = 0$	0.813	0.514	0.689	1.000	0.555	1.000	0.685	1.000
$H_0: \rho_{2,GT} = 0$			0.133	0.070	0.604	1.000	0.221	1.000
$H_0: \boldsymbol{\rho}_{GT} = \mathbf{0}$			0.247	0.062	0.759	1.000	0.429	1.000
$H_0: \boldsymbol{\rho} = \mathbf{0}$		0.001		0.001		0.004		0.002

are limited by the transition equations for the RGB and the survey errors, to estimate the model on the monthly frequency.

Figures 2–4 compare the point nowcasts, respectively, of the change in unemployment, its trend, and the population parameter, obtained with the baseline, the CC, the GT and CC & GT models which employ monthly Google Trends and include two of their factors. From the first graph, it is evident that the models including claimant counts tend to deviate from the baseline model. The latter, in contrast, gives similar results as those of the GT model. The point nowcasts of $L_t^{k,y}$ and $\theta_t^{k,y}$ are more similar throughout the model specifications, with a slight and positive difference between the models that include the Google Trends and the ones that do not, at the beginning of the out-of-sample period.

Figures S2 and S3 show the selection frequency of, respectively, the monthly and weekly Google Trends in the out-of-sample period. Some of the most selected search terms in both cases are: werklozen (unemployed people), baan zoeken (job-search), curriculum vitae voorbeeld (curriculum vitae example), ww uitkering (unemployment benefits), ww aanvragen (to request unemployment benefits), resume, tijdelijk werk (temporary job) and huizenmarkt zeepbel (housing market bubble). Notice that the latter term (as well as 'economische crisis' (economic crisis) or 'failliet' (bankrupt), which are also frequently selected monthly Google Trends) is of economic uncertainty nature, rather than being job-search related. In additional analyses, that we do not report in this paper, we included only the latter type of search terms and we did not find them to have predictive power for the Dutch unemployment, which is now improved by the additional information contained in the search terms related to economic uncertainty.

The results of the empirical analysis can be summarised as follows. Targeting the Google Trends improves the predictive power of the latter series for the Dutch unemployment. Monthly Google

Trends improve the estimation and nowcast accuracy of the Dutch unemployment and its change, with both one and two factors. The largest gains are obtained when both the claimant counts and the Google Trends are included, and considering only one factor for the latter series. When two factors are considered the gains are smaller, but a LR test indicates that both of them should be included in the model. The sensitivity to the number of factors is somewhat similar for the weekly Google Trends, although there is a weak evidence only for their second factor to have predictive power for the unemployment. The weekly Google Trends yield in general less improvements in estimation and nowcast accuracy, with respect to the monthly Google Trends. The contributions of the two types of Google Trends are comparable only when the two-step estimator is additionally re-iterated for the weekly Google Trends (in order to obtain more precise estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$). This result suggests that iterating the two-step estimation can improve the predictive power of the Google Trends, and that the latter series are sensitive to the estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. Improvements are, instead, not always present when $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ are estimated on the weekly data. In general, the claimant counts mainly have a positive impact on the estimation and nowcast accuracy of $\theta_t^{k,y}$ and $L_t^{k,y}$, whereas the Google Trends on the estimation and nowcast accuracy of $R_t^{k,y}$. The point nowcasts of the latter state variable are more sensitive to the type of auxiliary series included, with respect to the ones of $\theta_t^{k,y}$ and $L_t^{k,y}$.

The assumptions of normality made and discussed throughout the paper can be tested on the standardised one-step ahead forecast errors (Durbin & Koopman, 2012, Chapter 7): $\tilde{\boldsymbol{v}}_t^k = \boldsymbol{B}_t^k \boldsymbol{v}_t^k$, for $t = d + 1, \ldots, T$ with $(\boldsymbol{F}_t^k)^{-1} = \boldsymbol{B}_t^{k'} \boldsymbol{B}_t^k$, where $\boldsymbol{F}_t^k$ is the covariance matrix of the prediction errors $\boldsymbol{v}_t^k$ estimated with the Kalman filter. The prediction errors for the labour force are defined as $\boldsymbol{v}_t^{k,y} = \boldsymbol{y}_t^k - \boldsymbol{Z}_t^y \hat{\boldsymbol{\alpha}}_{t|\Omega_{t-1}}^{k,y}$, for the claimant counts series as $v_t^{k,CC} = x_t^{k,CC} - \boldsymbol{Z}^{CC} \hat{\boldsymbol{\alpha}}_{t|\Omega_{t-1}}^{k,CC}$, and for the Google Trends series as $\boldsymbol{v}_t^{k,y} = \boldsymbol{x}_t^{k,GT} - \hat{\Lambda} \hat{\boldsymbol{f}}_{t|\Omega_{t-1}}^k$, for $t = d + 1, \ldots, T$ (the expressions for $\boldsymbol{Z}_t^y$ and $\boldsymbol{Z}^{CC}$ can be found in Section 2 of the supplementary material). We test the assumptions on the estimated CC & GT models when two factors of the Google Trends are included, and which employ, respectively, the monthly Google Trends, and the weekly Google Trends with the additional iteration of the two-step estimator (as they yield the best results in terms of estimation and nowcast accuracy of the state variables of interest, when two factors of the Google Trends are included).

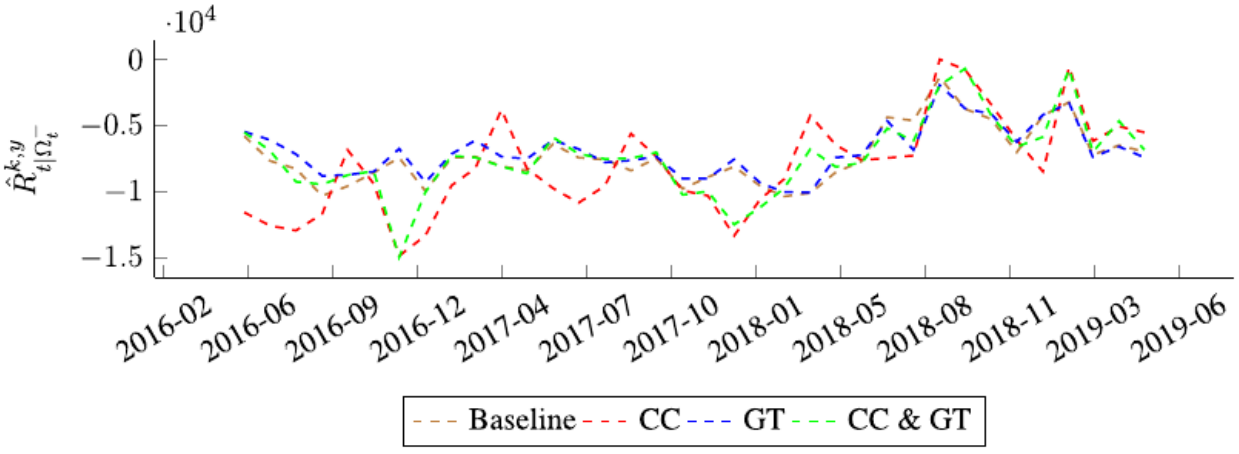

FIGURE 2 Nowcast of R_t^{ky} (which is the slope of the Dutch unemployment's trend) with the labour force models. We denote with 'Baseline' the model that does not contain auxiliary series. 'CC' and 'GT' stand for the models that contain, respectively, the claimant counts and the Google Trends as auxiliary series, whereas 'CC & GT' is the model that contains both types of auxiliary series. The results for the GT and the CC & GT models refer to the setting where the monthly Google Trends are used, and two of their factors are included in the model

We test the null hypothesis of univariate normality for each of the prediction error, with the Shapiro and Wilk (1965) and Bowman and Shenton (1975) tests, as suggested, respectively, in Harvey (1989) Chapter 5) and Durbin and Koopman (2012) Chapter 2). The former test is based on the correlation between given observations and associated normal scores, whereas the latter test is based on the measures of skewness and kurtosis. The p-values from the Shapiro–Wilk test are reported in Figures 5 and 6 for the two different model specifications discussed above, respectively. For both model specifications, there is no (strong) evidence against the normality assumptions for the error terms of the labour force and the claimant counts series, as their corresponding p-values are above the confidence level of 0.05. This result suggests that the model is correctly specified for these series. The test instead rejects the null hypothesis of normality for most of the idiosyncratic components of the Google Trends. The normality assumption seems, therefore, not appropriate for the latter series, but as discussed in Sections 3.1 and 5, and examined in the simulation study of the supplementary material, this type of misspecification does not affect the consistency of the estimators of the state variables and the hyperparameters, and does not seem to influence the performance of our method, nor the distribution of the LR test which allows to perform inference on the correlation parameters. Notice that we do not control for multiple hypotheses testing in this case. If we would control for it, we would obtain less rejections of the null hypothesis of normality for the error terms of the Google Trends, but the conclusions for the error terms of the labour force and the claimant counts series would stay the same. The conclusions from the Bowman–Shenton test are the same and the corresponding p-values are reported in Figures S4 and S5.

6 | CONCLUSIONS

This paper proposes a method to include a high-dimensional auxiliary series in a state space model in order to improve the estimation and nowcast of unobserved components. The method is based on a combination of PCA and Kalman filter estimation to reduce the dimensionality of the auxiliary series, originally proposed by Doz et al. (2011), while the auxiliary information is included in the state space

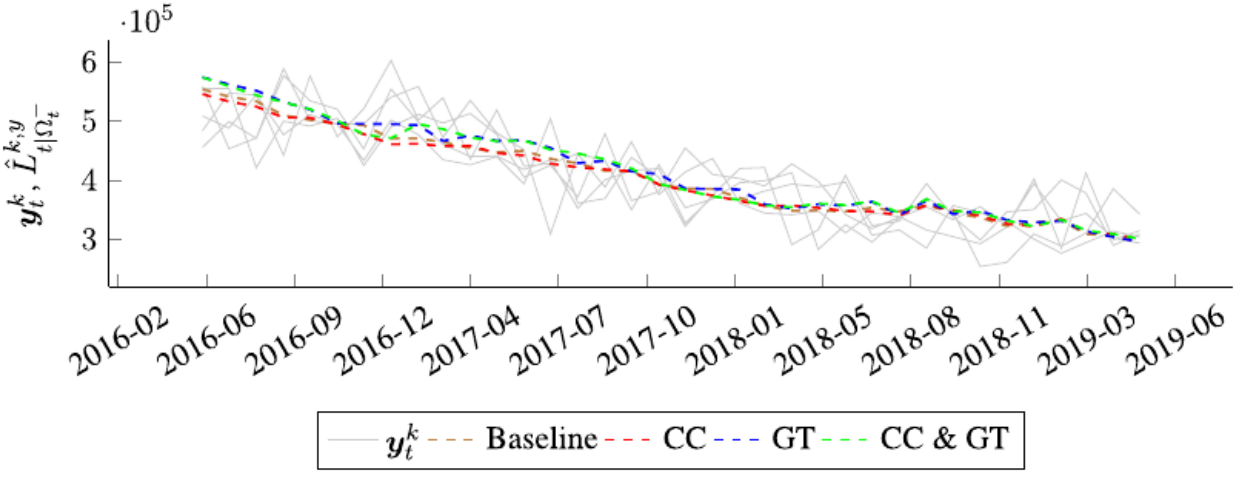

FIGURE 3 Nowcast of L_t^{ky} (which is the level of the Dutch unemployment's trend) with the labour force models, compared to the five waves of the unemployed labour force. We denote with 'Baseline' the model that does not contain auxiliary series. 'CC' and 'GT' stand for the models that contain, respectively, the claimant counts and the Google Trends as auxiliary series, whereas 'CC & GT' is the model that contains both types of auxiliary series. The results for the GT and the CC & GT models refer to the setting where the monthly Google Trends are used, and two of their factors are included in the model

model as in Harvey and Chung (2000). In this way we extend the state space model used by Statistics Netherlands to estimate the Dutch unemployment, which is based on monthly LFS data, by including the auxiliary series of claimant counts and Google Trends related to job-search and economic uncertainty. The strong predictive power of the former series, in similar settings, has already been discovered in the literature (see Harvey and Chung (2000) and van den Brakel and Krieg (2016)). We explore to which extent a similar success can be obtained from online job-search and economic uncertainty behaviour. The advantage of Google Trends is that they are freely available at higher frequencies than the LFS and the claimant counts, and, in contrast to the latter, they are not affected by publications delays. This feature can play a key role in the nowcast of the unemployment, as being the only real-time available information.

A Monte Carlo simulation study shows that in a smooth trend model our proposed method can improve the MSFE of the nowcasts of the trend's level and slope up to, respectively, around 25% and 75% when in the simulation data generating process the correlation between the auxiliary series and the series of interest is high. These results are robust to misspecifications regarding the distribution of the idiosyncratic components of the auxiliary series. Therefore, our method does have the potential to improve the nowcasts of unobserved components of interest.

In the empirical application of our method to Dutch unemployment estimation and nowcasting, we find that our considered Google Trends (when first targeted with the elastic net) do in general yield gains in the estimation and nowcast accuracy, (respectively, up to 35% and 30% alone, and up to 40% and 23% when the claimant counts series is also included in the model) of the state variables of interest, with respect to the model which does not include any auxiliary series. This result stresses the advantage of using the high-dimensional auxiliary series of Google Trends, despite involving a more complex model to estimate, which is especially relevant for countries that do not have any data sources related to the unemployment (such as the registry-sourced series of claimant counts), other than the LFS. We also find that, under certain model specifications, including both claimant counts and Google Trends outperforms the model which only includes the former auxiliary series. This result is explained by the fact that the two auxiliary series have a positive impact on the estimation/nowcast

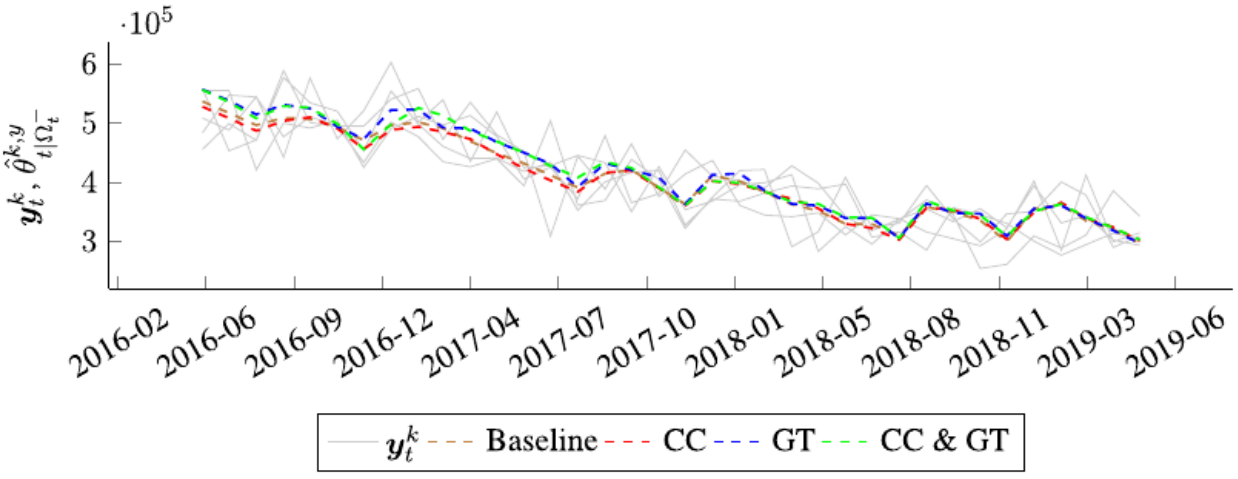

FIGURE 4 Nowcast of $\theta_t^{k,y}$ (which is the Dutch unemployment) with the labour force models, compared to the five waves of the unemployed labour force. We denote with 'Baseline' the model that doesnot contain auxiliary series. 'CC' and 'GT' stand for the models that contain, respectively, the claimant counts and the Google Trends as auxiliary series, whereas 'CC & GT' is the model that contains both types of auxiliary series. The results for the GT and the CC & GT models refer to the setting where the monthly Google Trends are used, and two of their factors are included in the model

accuracy of different unobserved components which constitute the unemployment, thus yielding an overall improvement of the fit of the model. This also indicates that claimant counts and Google Trends do not bring redundant information for the prediction of the Dutch unemployment.

The magnitude of the above-mentioned gains is, nonetheless, sensitive with respect to the following aspects of the data and the model specification. First, in our empirical application we employ both monthly and weekly Google Trends. The latter need to be aggregated to the monthly frequency in order to be included in the model, but allow to perform the nowcast on a weekly basis. We find that the former are less noisy and provide in general more accurate estimates/nowcasts of the state variables of interest, with respect to the latter. The predictive power of the monthly Google Trends for the Dutch unemployment is further corroborated by results from LR testing, which are in favour of their inclusion in the model. There is, instead, not strong and consistent evidence for this when the weekly Google Trends are employed.

Second, PCA involves the estimation of common factors that drive the Google Trends and in our method we relate these factors to the unobserved components that constitute the Dutch unemployment. Information criteria suggest that the Google Trends are driven by either one or two common factors. We find that including two factors yields, in general, less gains in accuracy, with respect to including one factor (due to the increased complexity of the model), but there is evidence that the second factor should also be included in the model in order to exploit all the predictive power that the Google Trends yield for the unemployment.

Finally, our estimation method is based on a two-step procedure. In the first step, the matrix of factors' loadings and the covariance matrix of the idiosyncratic components of the Google Trends are estimated by PCA. In the second step, these matrices are replaced by their PCA estimates, in order to re-estimate the Google Trends' factors and the unobserved components of the labour force series, with the Kalman filter. The estimation accuracy of these matrices might affect the predictive power of the Google Trends. We find that the predictive power of the weekly Google Trends can be improved (in order to yield similar gains as the ones obtained with the monthly Google Trends), with an additional iteration of the two-step estimation procedure, which should provide more accurate estimates of the two matrices.

As already mentioned, we generally find estimation/nowcast accuracy gains from the inclusion of the Google Trends, when they are first 'targeted', by selecting the ones that are relevant for the Dutch unemployment, based on the elastic net penalised regression. If the targeting is not first applied, we do

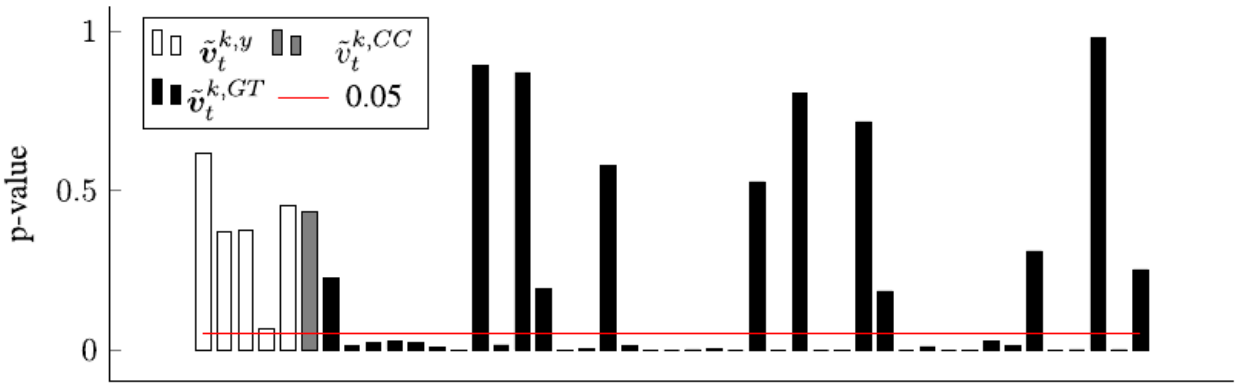

FIGURE 5 p-values from the Shapiro–Wilk test for individual normality, performed on each of the standardised prediction errors of the labour force, the claimant counts and the Google Trends series ($\tilde{\mathbf{v}}_t^k$). The standardised prediction errors are obtained from the CC & GT model which employs the monthly Google Trends and include two of their factors. The red line represents the confidence level of 0.05

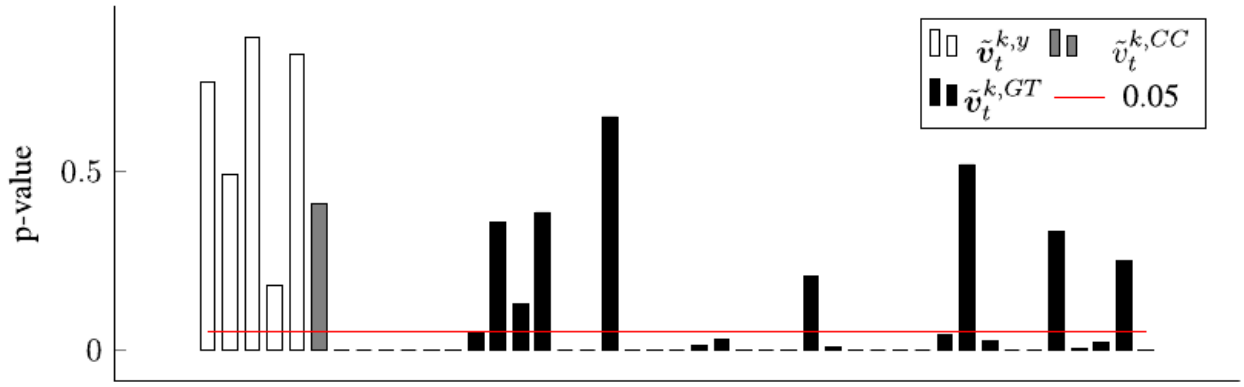

FIGURE 6 p-values from the Shapiro–Wilk test for individual normality, performed on each of the standardised prediction errors of the labour force, the claimant counts and the Google Trends series ($\tilde{\mathbf{v}}_t^k$). The standardised prediction errors are obtained from the CC & GT model which employs the weekly Google Trends and include two of their factors, and which iterates the estimation of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. The red line represents the confidence level of 0.05

not find these gains. Nonetheless, in this case the results do not deteriorate with respect to the model that does not include any auxiliary series, suggesting that our method is able to ignore the inclusion of irrelevant auxiliary series, in the estimation/nowcast of unobserved components of interest. This result is corroborated in our Monte Carlo simulation study. Hence, our proposed approach provides a framework to analyse the usefulness of 'Big Data' sources, with little risk in case the series do not appear to be useful.

One limitation of the current paper is that it does not allow for time-variation in the relation between the unobserved component of interest and the auxiliary series. For example, legislative changes may change the correlation between unemployment and administrative series such as claimant counts. Additionally, one can easily imagine the relevance of both specific search terms as well as internet search behaviour overall to change over time. While such time-variation may partly be addressed by considering shorter time periods, decreasing the already limited time dimension will have a strong detrimental effect on the quality of the estimators. Therefore, a more structural method is required that extends the current approach by building the potential for time variation into the estimation method directly, while retaining the possibility to use the full sample size. Such extensions are currently under investigation by the authors.

ACKNOWLEDGEMENTS

This work was funded by the European Union under grant no. 07131.2017.003-2017.596, the European Union's Horizon 2020 research and innovation program grant no. 770643 (MAKSWELL) and the Netherlands Organization for Scientific Research (NWO) under grant no. 452-17-010. The views expressed in this paper, are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. The authors thank the Joint Editor, the Associate Editor and two unknown referees for careful reading of former drafts of the manuscript and providing useful comments to improve our manuscript. Previous versions of this paper have been presented at CFE-CM Statistics 2017, The Netherlands Econometric Study Group 2018, Small Area Estimation Conference 2018, Methods for Big Data in Official Statistics, BigSurv 2018, the 29th $(EC)^2$ on Big Data Econometrics with Applications and at internal seminars organised by Maastricht University and Statistics Netherlands. The authors also thank the conference and seminar participants for their interesting comments. Additionally, the authors also thank Marco Puts and Ole Mussmann for their help with the data collection. All remaining errors are our own.

REFERENCES

Askitas, N. & Zimmermann, K.F. (2009) Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2), 107–120.

Bai, J. (2004) Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics*, 122(1), 137–183.

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

Bai, J. & Ng, S. (2008) Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2), 304–317.

Bailar, B. (1975) The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70(349), 23–30.

Bańbura, M., Giannone, D., Modugno, M. & Reichlin, L. (2013) Now-casting and the real-time data flow. Working Paper Series 1564, European Central Bank.

Barigozzi, M. & Luciani, M. (2017) Common factors, trends, and cycles in large datasets. Finance and economics discussion series 2017–111, Board of Governors of the Federal Reserve System (U.S.).

Bollineni-Balabay, O., van den Brakel, J. & Palm, F. (2017) State space time series modelling of the Dutch labour force survey: Model selection and mean squared errors estimation. *Survey Methodology*, 43(1), 41–67.

Bowman, K.O. & Shenton, L.R. (1975) Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. *Biometrika*, 62(2), 243–250.

Choi, H. & Varian, H. (2012) Predicting the present with Google trends. *Economic Record*, 88(suppl.1), 2–9.

Choi, H. & Varian, H.R. (2009) Predicting initial claims for unemployment benefits. *Google Inc*, pages 1–5.

D'Amuri, F. & Marcucci, J. (2017) The predictive power of Google searches in forecasting us unemployment. *International Journal of Forecasting*, 33(4), 801–816.

Davidson, J. (2000) *Econometric theory*. Hoboken: Blackwell Publishing.

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.

Durbin, J. & Koopman, S.J. (2012) *Time series analysis by state space methods*, second edition. Oxford Statistical Science Series. Oxford: OUP.

Elliott, G., Rothenberg, T.J. & Stock, J.H. (1996) Efficient tests for an autoregressive unit root. *Econometrica*, 64(4), 813–836.

Giannone, D., Reichlin, L. & Small, D. (2008) Nowcasting: the real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4), 665–676.

Gourieroux, C. & Monfort, A. (1993) Pseudo-likelihood methods. In: Maddala, G.S., Rao, C.R. & Vinod, H.D. (Eds.), *Handbook of statistics 11* (pp. 335–362), chapter 12. Amsterdam, North-Holland: Elsevier Science Publishers B.V.

Gourieroux, C., Monfort, A. & Trognon, A. (1984) Pseudo maximum likelihood methods: Theory. *Econometrica*, 52(3), 681–700.

Hamilton, J.D. (1994) *Time series analysis*. Princeton: Princeton University Press.

Harvey, A. & Chung, C.-H. (2000) Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(3), 303–309.

Harvey, A.C. (1989) *Forecasting, structural time series models and the Kalman filter*. Cambridge: Cambridge University Press

Hastie, T., Tibshirani, R. & Wainwright, M. (2015) *Statistical learning with sparsity: The Lasso and generalizations*. Boca Raton: CRC Press.

Hastie, T. & Zou, H. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

Hindrayanto, I., Koopman, S.J. & de Winter, J. (2016) Forecasting and nowcasting economic growth in the euro area using factor models. *International Journal of Forecasting*, 32(4), 1284–1305.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014) Supplementary materials for the parable of Google Flu: Traps in big data analysis. *Science*, 343(March), 1203–1206.

Maas, B. (2019) Short-term forecasting of the us unemployment rate. Mpra paper 94066, University Library of Munich, Germany.

Moon, H.R. & Perron, B. (2012) Beyond panel unit root tests: Using multiple testing to determine the nonstationarity properties of individual series in a panel. *Journal of Econometrics*, 169(1), 29–33.

Naccarato, A., Falorosi, S., Loriga, S. & Pierini, A. (2018) Combining Official and Google trends data to forecast the Italian youth unemployment rate. *Technological Forecasting and Social Change*, 130, 114–122.

Pfeffermann, D. (1991) Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9(2), 163–175.

Pfeffermann, D., Feder, M. & Signorelli, D. (1998) Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business & Economic Statistics*, 16(3), 339–348.

Rao, J.N.K. & Molina, I. (2015) *Small area estimation*. Wiley Series in Survey Methodology, 2 edition. Hoboken: John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992) *Model assisted survey sampling*. New York, NY: Springer-Verlag Publishing.

Schiavoni, C., Palm, F., Smeekes, S. & van den Brakel, J. (2019) A dynamic factor model approach to incorporate big data in state space models for official statistics. Working paper.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.

Shapiro, S.S. & Wilk, M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4), 591–611.

Stephens-Davidowitz, S. & Varian, H. (2015) A Hands-on Guide to Google Data. *Google, Inc*., pages 1–25.

Suhoy, T. (2009) Query indices and a 2008 downturn: Israeli Data. Discussion paper series no. 2009.06, Bank of Israel.

van den Brakel, J. & Krieg, S. (2009) Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology*, 35(2), 177–190.

van den Brakel, J.A. & Krieg, S. (2015) Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. *Survey Methodology*, 41(2), 267–296.

van den Brakel, J.A. & Krieg, S. (2016) Small area estimation with state space common factor models for rotating panels. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(3), 763–791.

Wilks, S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1), 60–62.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.
Data S1

How to cite this article: Schiavoni C, Palm F, Smeekes S, van den Brakel J. A dynamic factor model approach to incorporate Big Data in state space models for official statistics. *J R Stat Soc Series A*. 2020;00:1–30. https://doi.org/10.1111/rssa.12626