Nowcasting Vietnam's RGDP Using a Kernel-Based Dimensional Reduction Method

Thanh Do Van(✉)

Faculty of Information Technology, Nguyen Tat Thanh University,
Ho Chi Minh City, Viet Nam
dvthanh@ntt.edu.vn

Abstract. The gross domestic product growth rate (RGDP for short) is one of the most important macroeconomic indicators often used for making economic policies and planning production and business development plans by government agencies and enterprise communities. Research to improve the forecast accuracy of this indicator has always been of interest to researchers. In Vietnam, this indicator is only released quarterly.

It is no longer appropriate to forecast the RGDP according to predictors at the same frequency as this indicator because, in the time interval between two quarters, there may be some political, socio-economic events occurring that have a substantial impact on many economic activities that cause the change of the RGDP in the current quarter and the next quarters. It is necessary to use another new forecast approach to overcome this limitation.

The purpose of this article is to build a model to nowcast the RGDP on a large dataset of predictors at higher frequencies than the quarterly frequency. Such a model is developed based on the dynamic factor model. Unlike previous studies, the factors in the built model are extracted from the input dataset by a variable dimension reduction method using kernel tricks and based on an RMSE-best model. The article also proposes a ragged-edge data handling method and reinforcement learning method, suitable for the regression method used to build the nowcasting model of the RGDP indicator.

Keywords: Data mining · Nowcasting · Big data · Dimensionality reduction · Kernel tricks · PCA

1 Introduction

The estimation of Vietnam's RGDP is only performed at quarterly and yearly frequencies. In Vietnam, the RGDP indicator, especially at the quarterly frequency, is now considered one of the most important macroeconomic indicators in the formulation of the economic regulation policies of the government. In the context of deep, broad international economic integration and unpredictable international changes, the forecast of RGDP and the update of forecasts of RGDP according to real-time data flows are essential and have very high practical significance. The nowcasting model of the RGDP indicator makes that possible.

P. Cong Vinh and N. Huu Nhan (Eds.): ICTCC 2021, LNICST 408, pp. 108–128, 2021.
https://doi.org/10.1007/978-3-030-92942-8_10

Nowcast is defined as the prediction of the present, the very near future, and the very recent past by using available, timely, and reliable information to formulate predictions for target variables of interest [1–4]. Nowcast aims to exploit the information published early and possibly at higher frequencies than the target variable of interest to obtain an early estimate before the official figures become available. A nowcasting model of macroeconomic indicators needs to have features to monitor many data releases, forming expectations about them and revising the assessment on the state of these indicators whenever realizations diverge sizeably from those expectations. The nowcasting approach is related to the datasets known as big data [2–5] and ragged-edge data [7–9].

Big Data has been variously defined in the literature. The work [2] introduced three ways to identify big data. First, big data is identified through several characteristics, of which the most important features are the four "V" [2, 9]: Volume (the quantity of generated and stored data), Variety (the type and nature of the data), Velocity (the speed at which the data is generated and processed to meet the demands and challenges), Veracity (the data quality and the data value). Big data can be identified by the number of variables and the number of observations. Then a dataset is called large if either the number of variables or the number of observations or both are large [2, 5]. Big data can also be identified through the content of the data. It is data from social networks, traditional business systems, or the internet of things [2].

There have been many studies on building the RGDP indicator nowcasting models or nowcasting this indicator. These models are built based on the factor state-space model (another version of the dynamic factor model developed according to the idea of Kalman filter) [10, 11], factor bridge equation model [12–14], and factor MIDAS model [12–15]. The factors are extracted from the predictors' dataset using the Principal Component Analysis (PCA) method in these models. The authors of the works [13, 16–18, 24] reviewed methodologies and econometric techniques to forecast macroeconomic indicators on large data sets. They are classified into two categories: statistical machine learning techniques and artificial intelligence machine learning techniques. Techniques based on regression such as the bridge equation, MIDAS, and logistic regression belong to the statistical technical group. The machine learning techniques such as artificial neural network (ANN), support vector machine (SVM), genetic algorithm, cluster analysis, k closest neighbors,… belong to the artificial intelligence technical group, in which the ANN and SVM techniques are the most common algorithms utilized for the forecast/classification purpose [19–22]. One found that ANN's ability in prediction is significantly better than SVM [20]. An essential aspect of deep models is that they can extract rich features from the raw data and make predictions. So, from this point of view, deep models usually combine both phases of feature extraction and prediction in a single phase. ANNs with different structures were tested, and the experiments proved the superiority of deep ANNs over shallow ones.

According to [2], the deep learning neural network method [23] is only appropriate in prediction exercises on datasets with a large number of observations and not a large number of variables. In other words, this method is not suitable for datasets with a large number of variables.

The essence of nowcasting economic-financial indicators is to forecast a target variable (or dependent variable) at a low frequency on a large dataset of time series predictors (or explanatory variables) at some higher frequencies. The works [2, 4, 5,

16–18, 24, 25] showed that the effective modeling method on mixed frequency macroeconomic big data is to use the dynamic factor model and the Kalman filter, in which the dynamic factor model is used more. The dynamic factor model includes the factor bridge equation model and the factor MIDAS model [13, 14, 17, 18], here the factors are extracted from the dataset of input predictors by the PCA method.

The bridge equation model approach [26] offers a convenient solution for filtering and aggregating variables characterized by different frequencies. However, aggregation may lead to the loss of valuable information [17]. This issue has led to the development of a mixed frequency modeling approach called MIDAS [27].

Bai, Ghysels, and Wright [28] studied the relationship between the MIDAS regression and Kalman filter when forecasting mixed frequency data. The authors examine how the MIDAS regression and the Kalman filter match under ideal cases, where the stochastic components, the lag of high and low-frequency variables are all assigned values exactly. The experimental results show that the forecast accuracy of the models built based on the Kalman filter and the MIDAS model is similar. In most cases, the Kalman filter gives a slightly more accurate result, but it requires much more computation. Ankargren, Sebastian, and Unn Lindholm [29] experimented and concluded that the MIDAS model and the bridge equation model achieve a lower forecast error than that of the factor state-space model (it is another version of the dynamic factor model where one used some equations containing factors at a high frequency in a regression model of variables at a low frequency) [11]. That article also showed that the bridge equation model using a small set of variables (≤ 6 variables) performs better than using a medium set of variables (around 14 variables) or large (about 34 variables). The best performance belongs to the MIDAS model when using a set of variables of medium size. However, this article has not shown that with a small set of predictors and in ideal cases, of the two models, the bridge equation model and the MIDAS model, which one has the small RMSE than.

The task of dimensionality reduction techniques/methods is to extract factors from an input data set to replace the predictors in this set. In economics and finance, the most commonly used factor extraction technique is the PCA and sparse PCA. The PCA method is a typical unsupervised learning method to transform a dataset in a high dimensional space into a much lower dimensional space while still preserving the maximum variance and covariance structure of the original dataset [30]. The dataset in the low dimensional space is the dataset of some principal component factors. Each factor is a linear projection of the mean-centered input dataset into an eigenvector of the covariance matrix created from the input dataset of predictors. The cumulative variance percentage of p first factors corresponding to p highest eigenvalues is also the percentage of information of the original dataset held by these p factors. In practice, one usually takes only p factors so that its corresponding cumulative variance percentage is in the range of 70%–90% to replace the original preditor variables.

The PCA method is very efficient for reducing the dimensionality of a dataset [31] if its data points are approximately a hyperplane and not efficient if that is not true. In our recent study, we have proposed a new variable dimension reduction method as a natural extension of the PCA method and called the KTPCA method. With nowcasting models built based on the dynamic factor model, the experiment of the variable dimension reduction methods PCA, SPCA, randomized SPCA, Robust SPCA, and the KTPCA methods on 11 large datasets of time series predictors showed that the

performance of the KTPCA method based on an RMSE-best model is always higher than that of the reported methods. Here the performance of a variable dimension reduction method is measured by the RMSE of a nowcasting model built based on the dynamic factor model, in which the factors are extracted using this variable dimension reduction method. Moreover, the dynamic factor model includes the factor bridge equation model, factor unrestricted MIDAS model (U-MIDAS), and factor restricted MIDAS models with the MIDAS weights to be Step, polynomial Almon, and exponential Almon functions. This study also compared the forecast accuracy of the factor bridge equation model, the factor U-MIDAS model, and some other factor-restricted MIDAS models. We received that the forecast accuracy of the factor U-MIDAS model is higher than that of other factor MIDAS models. With the number of factors being small (<=6), the factor bridge equation and U-MIDAS models are competitive in forecast accuracy.

This article applies the results of our recent research just mentioned to build a nowcasting model of Vietnam quarterly RGDP so that the forecast accuracy of the built model is the highest compared to models built based on other known so far. Based on the specific dataset to build the nowcasting model of the RGDP indicator, the nowcasting model is built based on the factor bridge equation model or the factor U-MIDAS model. Here, factors are extracted using the KTPCA method based on an RMSE-best model.

The article is organized as follows: following this section, Sect. 2 introduces some necessary content used in the following sections. Section 3 presents a dataset used to build and a method of building the Vietnam RGDP nowcasting model. Section 4 introduces some main results. Section 5 presents a ragged-edge data handling method and a reinforcement learning method and uses the built model to update the RGDP forecasts in a real-time data flow. The last Sect. 6, is some conclusions.

2 Preliminaries

2.1 Factor Bridge Equation Model

Bridge equations are linear regression ones that link variables at high frequency to lower frequency variables. This method allows early estimates of low-frequency variables by using higher frequency variables [2, 14, 17, 18].

Factor bridge equation model is defined as follows:

$$y_t^Q = \alpha + \sum\nolimits_{i=1}^{N} \beta_i x_{i,t}^Q + \sum\nolimits_{i=1}^{K} \gamma_i F_{i,t}^Q + \varepsilon_t \quad (1)$$

where y_t^Q is the RGDP indicator at the quarterly frequency Q; t is the time point of frequency Q; $x_{i,t}^Q$ are the indicators at the same frequency as the target variable y_t^Q; $F_{i,t}^Q$ are factors at the frequency Q aggregated from factors at higher frequencies $F_{i,t}^M$ and/or

$F_{i,t}^{D}$ (M, D are monthly or daily frequency, respectively), in which $F_{i,t}^{M}$ or $F_{i,t}^{D}$, respectively, are extracted from large sets of predictors $z_{i_j,t}^{M}$ and/or $z_{i_h,t}^{D}$ by using the KTPCA based on an RMSE-best model.

Due to the economic system has inertia, macroeconomic variables usually exist autocorrelation, so it is necessary to add a corresponding lag in a forecasting model of macroeconomic indicators, so the factor bridge equation model can be further extended to include lags of the target variable as well as of the predictors. Then the Eq. (1) can be written in the form:

$$y_t^Q = \sum_{k=1}^{q} b_k y_{t-k}^Q + \sum_{i=1}^{N} \sum_{j=0}^{r_i} \beta_{ij} x_{i,t-j}^Q + \sum_{j=1}^{p} \sum_{q=0}^{p_j} \gamma_{jq} F_{i,t-h}^Q + c + u_t \quad (2)$$

where r_i (i = 1, … , n), P_j (j = 1, … , m) and q, are the maximum lag of the variables $x_{i,t}^Q$, $F_{i,t}^Q$, and y_t^Q, respectively. The maximum lag can be determined using either AIC or BIC information criterion.

Model (2) can be rewritten as:

$$\psi(L) y_t^Q = \sum_{i=1}^{N} \beta_i(L) x_{i,t}^Q + \sum_{j=1}^{M} \gamma_j(L) F_{i,t}^Q + c + u_t \quad (3)$$

where L denotes usual lag operator, $\psi(L) = 1 - \sum_{k=1}^{q} b_k L^k$, $\beta_i(L) = \sum_{j=0}^{r_i} \beta_{i_j} L^j$, and $\gamma_j(L) = \sum_{h=0}^{p_j} \gamma_{j_h} L^h$.

In essence, model (2) is also the Autoregressive Distributed Lag ARDL(q, $r_1, \ldots, r_N, p_1, \ldots, p_M$) model [32].

2.2 The Factor MIDAS Models

The factor MIDAS model under consideration is [33]:

$$\psi(L) y_t^Q = \sum_{i=1}^{N} \beta_i(L) x_{i,t}^Q + f(\{F_{t/S}^M\}, \theta, \lambda) + u_t \quad (4)$$

$\{F_{t/S}^M\}$ is set of the factors extracted from a large set of predictors sampled at a higher frequency with S values for each low-frequency value; $\psi(L) = 1 - \sum_{k=1}^{q} b_k L^k$; $\beta_i(L) = \sum_{j=0}^{r_i} \beta_{i_j} L^j$; f is a function describing the effect of the higher frequency data in the low-frequency regression; b = (b_k), $\beta_i = \left(\beta_{i_j}\right)$, θ, and λ are vectors of parameters to be estimated.

If we like only to include each of the higher frequency components as a predictor in the low-frequency regression, then model (4) can be given by

$$\psi(L) y_t^Q = \sum_{i=1}^{N} \beta_i(L) x_{i,t}^Q + \sum_{\tau=0}^{k-1} F_{(t-\tau)/S}^{M\,T} \theta_\tau + u_t \quad (5)$$

where $F^{M}_{(t-\tau)/S}{}^{T}$ are the factors at the τ high-frequency periods before t. Then, a distinct θ_τ is associated with each of the k high-frequency lag factors. The number of θ_τ coefficients may be very large. If these coefficients are not constrained, then model (5) is called the unrestricted MIDAS model (or U-MIDAS). So, the U-MIDAS model offers the greatest flexibility but requires large numbers of coefficients.

2.3 The KTPCA Based on an RMSE-Best Model

Assume that the missing and outlier data in the dataset of the y^{Q}_{t}, $x^{Q}_{i,t}$, and $z^{M}_{j,t}$ variables have been processed. Without loss of generality, it can assume that all variables y^{Q}_{t}, $x^{Q}_{i,t}$, and $z^{H}_{j,t/S}$ are stationary time series. Improved from the KTPCA variable dimension reduction method based on an RMSE-best model for data sampled at the same frequency, the KTPCA method based on an RMSE-best model on mixed frequency datasets is shown in Fig. 1 below.

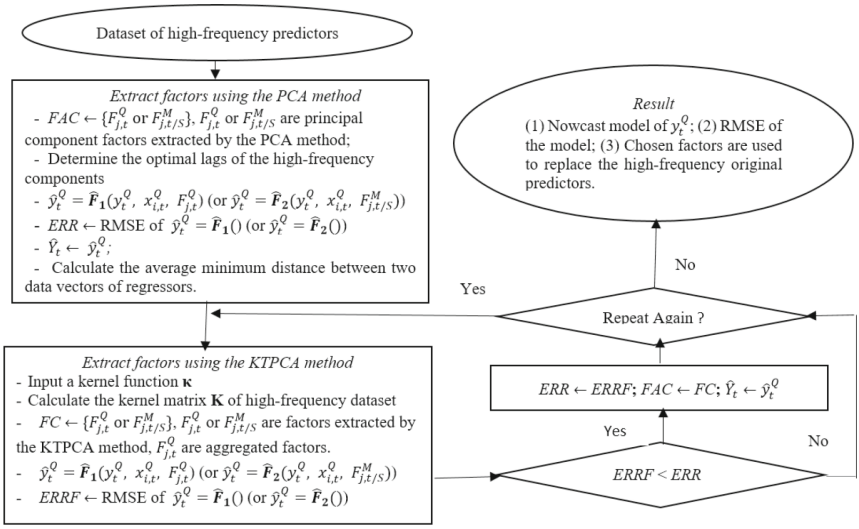

Fig. 1. The KTPCA method based on an RMSE-best model on mixed frequency datasets

In this flowchart, we denote $\hat{F}_1()$ and $\hat{F}_2()$ as nowcasting models of the dependent variable y^{Q}_{t} to be estimated based on the factor-bridge equation model (3) and the factor U-MIDAS model (5), respectively; FAC is the set of factors extracted from the dataset of high-frequency predictors using the KTPCA method based on an RMSE-best model where $F^{M}_{j,t/S}$ are factors extracted from the dataset of high-frequency predictors and are

sampled at a higher frequency with S values for each low-frequency value; moreover; $F_{j,t}^{Q}$ are aggregated from $F_{j,t/S}^{M}$ at the same low-frequency as the dependent variable. ERR is the standard mean forecast error of a nowcasting model and is measured by the root mean squared forecast error (RMSE for short) determined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{T} \cdot \sum_{t=1}^{T} \left(y_t^Q - \hat{y}_t^Q\right)^2} \tag{6}$$

here, $\hat{y}_t^Q$ is produced by the built nowcasting model and is called a fitted variable of y_t^Q; T is the number of observations. The process above is iterative, and in principle, we can choose a kernel so that the RMSE of its corresponding nowcasting model is small as possible. The above flowchart shows that the most suitable factor extraction and building a nowcasting model of a target variable are combined in one process called the factor extraction using the KTPCA method based on an RMSE-best model.

In practical applications, polynomial kernels $\boldsymbol{\kappa}(x_\text{i}, x_\text{j}) = (c_1\langle x_i, x_j\rangle + c_2)^d$ and Gaussian kernels $\boldsymbol{\kappa}(x_\text{i},x_\text{j}) = \exp(-\frac{\left\|x_i - x_j\right\|^2}{2.\rho^2})$, here $c_1 > 0, c_2 \geq 0$ and $\rho > 0$, are commonly used [34, 35]. According to some researchers, for Gaussian kernels, ρ^2 should be chosen around the value to be the average minimum distance between two column vectors of the dataset of the input predictors [36].

3 Dataset and Method

3.1 Dataset of Original Predictors

Following the approach of this article, the predictors at a higher frequency used to build the nowcasting model of the quarterly RGDP indicator are determined based on economic theories and as broad as possible. So, collected data of predictors used to build the model may contain redundant or irrelevant information for the RGDP indicator's forecasting. In other words, some original predictors can be potential predictors for the nowcasting purpose.

The list of candidate predictors collected data to build the nowcasting model of Vietnam's RGDP at the quarterly frequency is shown in Table 1 below.

Table 1. List of candidate predictors and selected predictors

Indicators	Units	Freq.	The number	Release dates	Source	The number of selected variables
The Gross Domestic Product Growth Rate (RGDP)	%, YoY	Q	01	A[1]	GSO[2]	Target variable
Total retail sales of consumer goods and services	Bil. VND, at current prices	M	01	A	GSO	01
Retail sales in some economic sectors	-	M	04	A	FiinPro[3]	04
Basic inflation, overall CPI, and CPI of 8 other main consumer goods baskets	%, YoY	M	10	A	GSO	08
Gold & US dollar price indices	%, YoY	M	02	A	GSO	0
Consumption index for the whole industry and 18 major production industries	%, YoY	M	19	A	FiinPro	16
The world price of Vietnam rice & Thailand rice; the world price of copper, coffee & rubber	USD/ton at current prices	M	5	B[4]	Fred[5]	04
The world price of Brent crude	USD/Cart on	M	1	B	Fred	01
Total imports & exports of goods and services	Mil. USD, at current prices	M	2	A	GSO	0

(*continued*)

Table 1. (*continued*)

Imports of 18 production industries	-	M	18	A	FiinPro	10
Exports of 25 production industries	Mil. USD, at current prices	M	25	A	FiinPro	14
Inventory index of the whole industry and 18 major production industries	%, YoY	M	19	A	FiinPro	08
Money M2, Deposits of financial institutions and residents	Bil. VND, at current prices	M	03	A	FiinPro	03
Total outstanding loans of the whole economy and five economic sectors of level 1	%, YoY	M	06	A	FiinPro	01
FDI includes implemented, registered, newly registered & additionally registered FDI	Mil. USD, at current prices	M	04	A	FiinPro	04
Industrial production index of the whole economy & of 9 main industries	%, YoY	M	10	A	GSO	08
Deposits from business organizations and residents	%, YoY	M	01	A	FiinPro	0
Manufacturing Purchasing Managers Index (PMI)	Point	M	01	C[6]	IHS-Markit[7]	01
Investment capital implemented from the State budget	Bil. VND, at current prices	M	01	A	GSO	0
Short-term and medium-term lending rates of state-owned commercial banks	%,	W	02	D[8]	FiinPro	0
Interest rates for short, medium, and long-term mobilization of state-owned commercial banks	%,	W	03	D	FiinPro	03

(*continued*)

Table 1. (*continued*)

Exchange rate of VND with USD	Nominal exchange rate	Day	01		FiinPro	0
Exchange rate of Yuan with USD	Nominal exchange rate	Day	01		Fred	01
Vietnam stock indexes: VN Index & HNX index	Point	Day	02		Cophieu 68[9]	01
Down Jones and S&P 5000 stock indexes	Point	Day	02		Fred	01

[a]A: within the last five days of a quarter
[b]https://www.gso.gov.vn/
[c]http://fiinpro.com/
[d]B: within the last three days of a month
[e]http://fred.stlouisfed.org/
[f]C: within the first four days of a subsequent month
[g]https://www.markiteconomics.com/
[h]D: the last day of the working week
[i]http://www.cophieu68.vn

Thus, there are 143 original predictors used to build the nowcasting model of Vietnam's RGDP at the quarterly frequency. The candidate predictors reflect the demand-side, supply-side, and market liquidity of the economy.

Data of variables at the monthly and quarterly frequencies are mainly collected from two sources (Table 1): General Statistics Office (GSO) and FiinPro - the company providing economic and financial data services. These data are usually released during the last five days of every month. The thing to note is data on total export and import, total retail sales of consumer goods and services, consumption index, and inventory index of the whole economy at a monthly frequency are released by GSO, but the FiinPro also releases the mentioned above and their more detailed data. With the same economic-financial indicators, FiinPro's data release date is usually about 2–3 days behind the General Statistics Office.

Interest rates on deposits and loans are released weekly on the first day of working weeks. Survey data on PMI conducted by IHS Markit is usually released on days 1 to 4 of the following month. Data collected from the Fed (stock indices, exchange rates,…) and Cophieu68 (stock indices) are daily figures released before the market closes. Besides, starting from 1/2013, the national account of Vietnam's economy is calculated according to the new economic subsectors and the base year of 2010, while the statistical data of previous years have not been adjusted accordingly. That means that it should only collect data from Q1, 2014 to Q1, 2020 for the RGDP indicator and other predictors at the same frequency as the RGDP, while data of other predictors at higher frequencies can be collected earlier. Specifically, in this article, data at monthly frequency was collected from January 2013 to March 2020. Data of the remaining variables were collected accordingly at the daily or weekly frequencies from January 1, 2013, to March 31, 2020.

3.2 Method

Figure 2 below briefly describes the process of building a nowcasting model of the RGDP indicator based on the most optimal model in the factor bridge equation model and the factor U-MIDAS model. Here, the factors are extracted using the KTPCA method based on an RMSE-best model.

The main content of the original dataset pre-processing is to add missing data, deal with outlier data and deal with the seasonality of the data. The missing or outlier data is overcome by the AR interpolation or smoothing method depending on the characteristics of each specific time series variable. It can be seen that the values of the dependent variable and other predictors in Table 1 include absolute and relative numerical values, in which relative numerical values (%) are compared with the same period last year. Macroeconomic data has often seasonality. The use of relative numerical values for economic variables implies dealing with the seasonality of its data. Predictors receiving absolute numerical values in Table 1 are converted into predictors of the same name that receive relative values compared to the same period last year. Assuming X_t is an absolute numeric value economic variable in Table 1, XS_t is the seasonally processed variable of X_t, the formula that converts the absolute numeric values of X_t into its relative numeric values is determined by:

$$XS_t = LOG(X_t/X_{t-12}) \text{ or } XS_t = LOG(X_t/X_{t-4}) \tag{7}$$

That depends on the indicator X_t at the monthly or quarterly frequency, respectively. When X_t is the GDP indicator, then the XS_t is called the RGDP indicator. To convert the absolute numerical values of financial predictors such as price, stock indices, and interest rates at daily or weekly frequency, we must aggregate these predictors at the monthly frequency using the arithmetic average formula and then apply the formula (7). The data conversion in the above way often makes non-stationary economic variables into stationary series (although this is not always the case), factors extracted by the KTPCA method based on an RMSE-best model to be stationary series, and avoids spurious regression when building nowcasting models using that variable dimension reduction method. This data conversion also facilitates the develop of software nowcasting automatically economic and financial indicators. Thus, the RGDP indicator has the observations from Q1/2014 to Q1/2020, while predictors at other frequencies have the observations from January 2014 to March 2020.

The converted dataset may still contain redundant and noisy information for the RGDP indicator's nowcasting purpose, so such information should be eliminated. The method of removing noisy or redundant information is to remove irrelevant or redundant predictors for nowcasting purposes using the Pearson correlation coefficient measure. Here the concepts of relevance and redundancy are defined as follows: suppose Y is the target variable at a low frequency and X is a higher frequency predictor. The predictor X is called redundant or irrelevant to Y if the aggregated variable of X at the same frequency as Y is redundant or irrelevant to Y, respectively. The removal of redundant or irrelevant predictors in such a way is essentially a way of selecting valuable predictors based on the filter approach using the Pearson correlation coefficient measure. The dataset of the selected predictors is called the input dataset.

Calculating the average minimum distance between two column vectors in the input dataset facilitates selecting a suitable Gaussian kernel when implementing the KTPCA method based on an RMSE-best model for extracting factors.

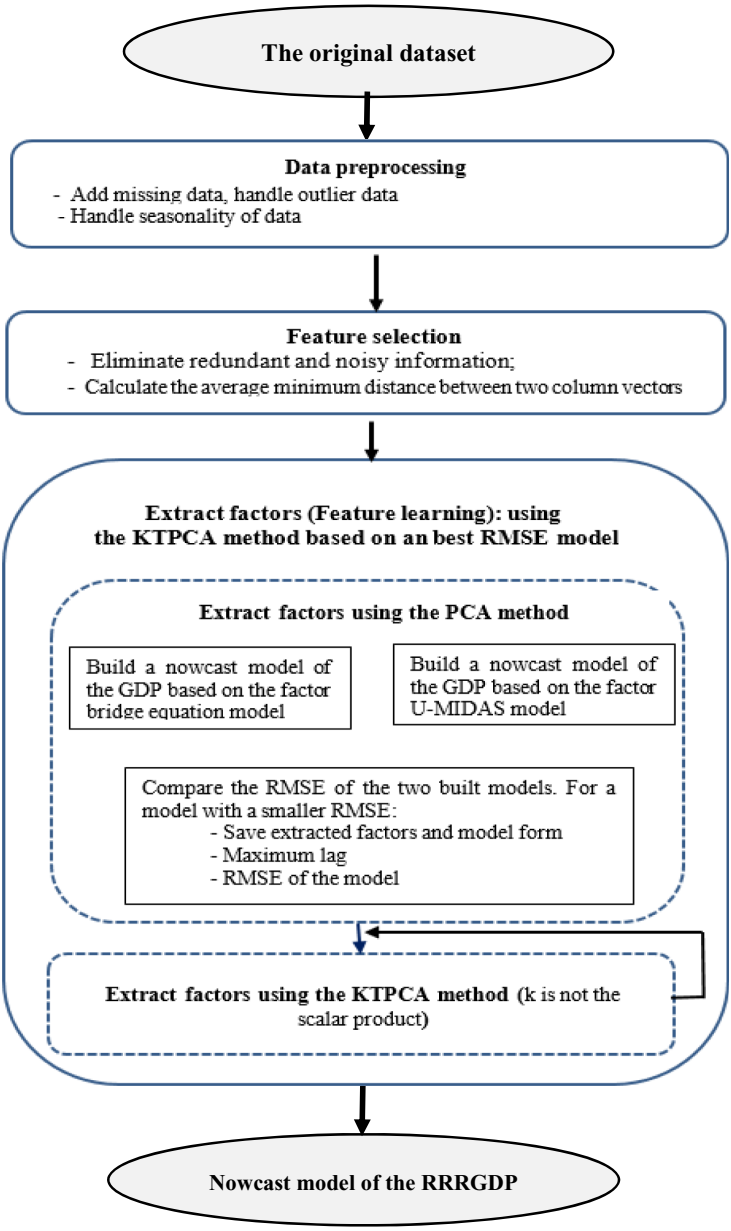

Fig. 2. Schema for building the nowcast model of the RGDP indicator

The extraction of factors from the input dataset of predictors at a frequency higher than the frequency of the target variable using the KTPCA method based on an RMSE-best model is started using the PCA method. On the chosen factors, two RGDP indicator nowcasting models are built based on the factor bridge equation model and the factor U-MIDAS model. Both models are estimated under an ideal condition, which is the maximum lag of all variables in each model to be precisely determined. For a nowcasting model with a smaller RMSE, save it, its RMSE, and the chosen factors. At the same time, the regression method used to build the nowcasting model with the smallest RMSE is also selected to build nowcasting models in the next stage.

The next stage is an iterative process of factor extraction using the KTPCA method, where kernels **k** are not the inner product of two column vectors in the mean-centered input dataset. The details of this process are shown in Fig. 1. At the end of the iterative process, we get the nowcasting model of the RGDP indicator.

4 Results

In this section, the article introduces some intermediate results obtained in the schema for building the nowcasting model of the RGDP indicator in Fig. 2.

With relevant and redundant thresholds of, respectively, 0.1 and 0.9, we get a set of 89 non-redundant and relevant predictors with the RGDP nowcasting purpose as shown in the last column of Table 1, in which the Manufacturing Purchasing Managers Index (PMI for short) and Vietnam Stock index (VN for short) are very highly relevant with the RGDP indicator. The article separates these indices from the set of the monthly frequency predictors and treats them as predictors at the quarterly frequency to facilitate monitoring and assessment of their impact on the RGDP. Their absolute numerical values at each quarter are the arithmetic average of the absolute numerical values at days or months in that quarter. Dataset of 87 remaining predictors at monthly frequency is considered as the input dataset. The average minimum distance between two column vectors of the mean-centered input dataset is $4.32226 = e^{1.463778}$.

Performing the ADF test for the RGDP, VN, and PMI at quarterly frequency shows that the RGDP is the first-order differential stationary while the VN and PMI are stationary series.

With the cumulative eigenvalue percentage threshold of 75%, using the PCA method to extract factors from the centered input dataset, we get the first two principal component factors with the commutative eigenvalue percentage of 77.496%. The results of building the nowcasting model of the RGDP indicator on the two indices VN and PMI and the two chosen factors based on the regression models (3) and (5) above are indicated, respectively, in Tables 2a and 2b below.

Table 2. The RGDP nowcasting model built based on the two dynamic factor models

2a. Nowcasting model is built based on the factor bridge equation model

D(RGDP) = -1.184*D(RGDP(-1))*** - 1.444*D(RGDP(-2))*** - 1.254*D(RGDP(-3))*** -
(0.101) (0.122) (0.126)
0.776*PMI*** - 0.324*PMI(-1)*** + 0.076*PMI(-3)* + 0.162*VN*** + 0.089*VN(-1)***
(0.074) (0.0438) (0.0358) (0.018) (0.013)
+ 0.080*VN(-3)***- 0.048*PC1*** - 0.026*PC1(-1)*** - 0.031*PC1(-2)*** + 0.012*PC2***
(0.011) (0.004) (0.006) (0.006) (0.002)
- 0.027PC2(-3)*** + 0.005*PC2(-2)*** + 0.017***.
(0.003) (0.001) (0.002)

R^2: 99.11; DW stat: 2.40; SMPL: 20 after adjustments; PCI (I =1, 2) are the chosen factors.
Significant codes: 0.0001: '****'; 0.001 '***'; 0.01 '**'; 0.05 '*'; 0.1: '. '
- RMSE (dynamic forecast): 0.000796;
- The common maximum lag for all variables: 03.

2b. Nowcasting model is built based on the factor U-MIDAS model

Formula D(RGDP) ~ mls(D(RGDP), 1:2, 1) + mls(VN, 0:2, 1) + mls(PMI, 0:2, 1) + mls(PC1, 0:7, 3) + mls(PC2, 0:3, 3)

	Estimate	Std. Error	Pr(>|t|)		Estimate	Std. Error	Pr(>|t|)
(Intercept)	-2.54E-03	4.58E-14	1.15E-11***	PC13	4.33E-02	5.81E-14	8.54E-13***
RGDP1	9.95E-01	5.14E-12	13.29E-12***	PC14	2.19E-03	3.13E-14	9.08E-12***
RGDP2	-5.20E-01	3.47E-12	4.25E-12***	PC15	2.82E-02	2.55E-13	5.77E-12***
VN1	1.31E-01	1.98E-13	9.61E-13***	PC16	-6.33E-02	2.04E-13	2.05E-12***
VN2	-2.19E-01	1.78E-13	5.18E-13***	PC17	-6.47E-03	1.45E-14	1.42E-12***
VN3	1.85E-01	3.23E-13	1.11E-12***	PC18	2.50E-02	6.01E-14	1.53E-12***
PMI1	-1.09E-01	3.96E-13	2.32E-12***	PC21	-2.71E-02	3.09E-14	7.27E-13***
PMI2	-1.25E-01	4.96E-13	2.53E-12***	PC22	2.27E-02	1.76E-14	4.94E-13***
PMI3	-2.74E-01	5.64E-13	1.31E-12***	PC23	-9.23E-03	2.69E-14	1.86E-12***
PC11	2.35E-02	1.25E-13	3.39E-12***	PC24	9.28E-03	2.16E-14	1.48E-12***
PC12	-5.50E-02	4.45E-14	5.16E-13***				

where Zi (i=1, 2, ...) is the variable Z lagged i-1 months. MLS() is the function in the 'MIDASu' package in R.CRAN that presents data to perform the MIDAS regression.
Significant codes: 0.0001: '****'; 0.001 '***'; 0.01 '**'; 0.05 '*'; 0.1: '. '; 1: ' '.
- RMSE (dynamic forecast): 0.00204 on 1 degree of freedom;
- The maximum lag for all factors at monthly frequency: 08;
- The maximum lag for all variables at quarterly frequency: 03.

Since the RMSE of the nowcasting model of the RGDP indicator built based on the factor bridge equation model is smaller than that based on the factor U-MIDAS model, the factor bridge equation model is selected to perform the following stages.

Table 3 below shows the number of chosen factors using the KTPCA method based on an RMSE-best model with kernels that are not the inner product, the cumulative eigenvalue percentage, and the RMSE of the RGDP indicator nowcasting model on the variables VN, PMI, and these two factors, where the nowcasting model is built based on the factor bridge equation model.

Table 3. Results of variable dimension reduction and nowcasting model building of the proposed method

Kernels $\kappa(x_i, x_j) =$	Number of chosen factors	Cumul. Eigen. percentage (%)	RMSE of the model	Maximum lag
$\langle x_i, x_j \rangle$ - (PCA)	2	77.496	0.000796	3
$(\langle x_i, x_j \rangle + 0.5)^2$	1	99.21	0.003023	3
$(\langle x_i, x_j \rangle + 0.5)^3$	1	99.97	0.003019	3
$\exp(-\frac{\|x_i - x_j\|^2}{2.e^{1.464}})$	1	80.60	0.003084	3
$\exp(-\frac{\|x_i - x_j\|^2}{2.e^{0.5}})$	6	76.14	0.002525	3
$\exp(-\frac{\|x_i - x_j\|^2}{2.e^{2.5}})$	1	90.47	0.003155	3

Table 3 shows that the inner product of two vectors is the most suitable kernel among the reported kernels to perform the variable dimension reduction by the KTPCA method based on an RMSE-best model. The model having the smallest RMSE among the built nowcasting models is the model in Table 2a. Figure 3 shows graphs of the actual RGDP, the fitted RGDP produced by that model, and its RMSE, where the RMSE graph lying between the two parallel lines implies that the model's forecast error (or RMSE) is relatively small and it can be considered to be approximately equal to zero and no observation is abnormal.

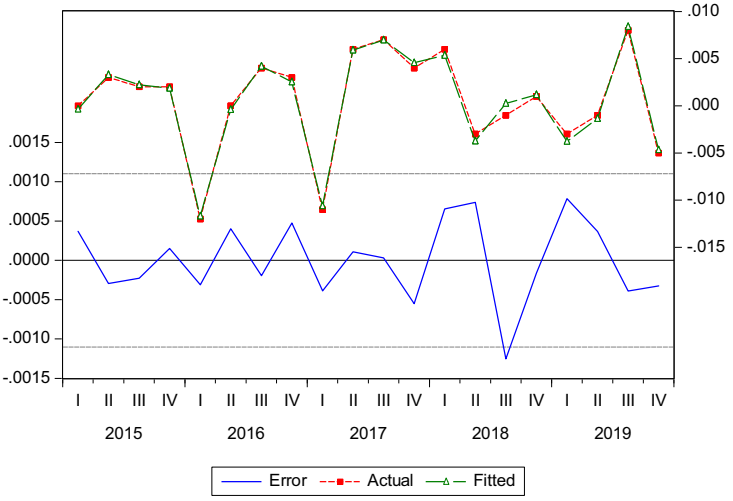

Fig. 3. Graphs of the Actual and Fitted RGDP indicator and RMSE of the nowcasting model

5 Update Forecasts of the RGDP Indicator in Real-Time Data Flow

One of the essential applications of nowcasting models is update forecasts by real-time data flows.

5.1 Handling Ragged-Edge Data in Nowcasting Model of Macroeconomic Indicators

Ragged-edge data often happens in mixed-frequency datasets due to missing values at the end samples for some predictors. There are three different methods to tackle it [6–8, 37]. The work [6] proposed realigning each time series in the sample to obtain a balanced dataset and then estimate the factors with the dynamic PCA method. The work [7] handled missing values in datasets using the EM algorithm and the traditional dynamic PCA method. The work [8] used the factor estimation approach based on a complete parametric representation of the factor model in state-space form. Developed on the idea of [6], the article proposes the realign of ragged-edge data to obtain a balanced dataset at the end samples of the predictors as follows:

Assuming T, t are the current quarter and month respectively, then t = 3T-2, 3T-1, or 3T (T = 1, 2, 3, 4). Our purpose is to forecast the quarter RGDP for the current and next quarter (here if (T + 1) > 5, then T = (T + 1) mode 4) whenever new data of the predictors is released on a date in month t.

- For predictors at the monthly frequency in Table 1, the new data updates are only made on the end days of the month (or the beginning days of the following month). Therefore, the forecasted value of these predictors in the following month (in this month) is the current value of the predictors in that month.
- Assume that X_t^M is a monthly frequency predictor aggregated from a daily frequency predictor X^D in a month t (for example, stock indices, exchange rates,…., in Table 1 are such predictors), X_t^{fM} is the forecast of the X^{TM} done on the most recent date. On a day in the month t, data of the X^D predictor is released, and its aggregated predictor at the monthly frequency X_t^M is updated as follows:

$$X_t^M = \frac{\left(X_{D,1} + X_{D,2} + \ldots + X_{D,m}\right) + (N - m) * X_t^{fM}}{N} \tag{8}$$

  where $X_{D,i}$ is the value of the X^D on the i[th] working day from the first working day in the month t, N is the number of working days in the month t, $m > 0$ is the number of working days from the first working day in the month t to the date on which the data of variable X^D is released.
- Assume that X_T^Q is a quarterly frequency predictor aggregated from the monthly frequency predictor X_t^M in the month t of the quarter T, here X_t^M is the monthly frequency predictor aggregated from the daily frequency predictor X^D (in Table 1, it is the VN predictor); X_T^{fQ} is the forecast of the X_T^{fM} done on the most recent date. On

a day in the month t, the data of X^D predictor is released, and the following formula updates its aggregated predictor X_T^Q at the quarterly frequency:

$$X_T^Q = \begin{cases} \frac{X_t^M + X_{t+1}^{fM} + X_{t+2}^{fM}}{3} \, if \; t = 3T - 2 \\ \frac{X_{t-1}^M + X_t^M + X_{t+1}^{fM}}{3} \, if \; t = 3T - 1 \\ \frac{X_{t-2}^M + X_{t-1}^M + X_t^M}{3} \; if \; t = 3T \end{cases} \tag{9}$$

where X_{t-k}^M is the X_t^{TM} predictor lagged k months, and X_{t+k}^{fM} is the out-of-sample k months forecast of the X_t^{TM} predictor from the date on which data of the X^D is released. The forecast of X_T^Q in the next quarter is based on the updated X_T^Q using an auxiliary model.

– Assume that Z_t^M is a monthly frequency predictor aggregated from a weekly frequency predictor Z^W in a month t (in Table 1, it is a lending rate or an interest rate). On a day in the month t, the data of Z^W predictor is released, and its aggregated predictor Z_t^M at the monthly frequency is updated by

$$Z_t^M = [m * Z_{t-1}^M + \sum\nolimits_{i=1}^{p} k_i * Z_{t,i}^W + (\mathrm{N} - m - \sum\nolimits_{i=1}^{p} k_i) * Z_t^{fM}]/N \tag{10}$$

here $Z_{t,i}^W$ is the value of the weekly frequency predictor Z^W at the i^{th} week in the month t; $1 \leq p \leq 5$; k_i is the number of days in the month t where the value of this variable is $Z_{t,i}^W$; $m > 0$ is the number of days from the first day in the month t to the date on which data of the weekly frequency predictor Z^W is firstly released.

In other words, on a day in a month t of the current quarter, if the data of predictors at the daily and weekly frequency is released, the value of predictors at the monthly frequency aggregated from the preditors at a higher frequency is updated and used to forecast values of the aggregated predictors in the following months. As for predictors at the monthly frequency, the value of the predictors at the month t is its forecasted value implemented at the latest update.

5.2 Sample Extension

It is necessary to develop a method of variable dimension reduction of sample extensions to reuse the variable dimension reduction previous results whenever new observations of the predictors appear. The method proposed in this article is similar to the method in [31].

Assuming $\mathbf{A}_{\mathrm{h}\times\mathrm{N}}$ is an h × N matrix of h new observations of N predictors, then the corresponding extension of p factors $\mathbf{PC}_{\mathrm{h}\times\mathrm{p}}$ of the h observations is determined by the following formula:

$$\mathbf{PC}_{\mathrm{h}\times\mathrm{p}} = \left(\boldsymbol{A}_{\mathrm{h}\times\mathrm{N}} - \overline{\mathbf{X}}\right) \times \mathbf{E}_{\mathrm{Nxp}} \tag{11}$$

here $\overline{\mathbf{X}} = [X_1 - \mu_1, X_2 - \mu_2, \ldots, X_N - \mu_N]$, where $X_i = (x_{ij}), i = 1, \ldots, m, j = 1, \ldots, N$ and $\mu_i = (\frac{1}{m+h} \sum_{j=1}^{m+h} x_{ij})_{1Xm}$, for every i = 1, … , N. $\mathbf{E}_{\mathrm{Nxp}}$ is the matrix of the first p eigenvectors of the kernel matrix of the original input dataset.

5.3 Update Forecast of the RGDP Indicator

Figure 4 below illustrates the updating of Vietnam's RGDP forecasts according to data release dates of the input predictors in Table 1.

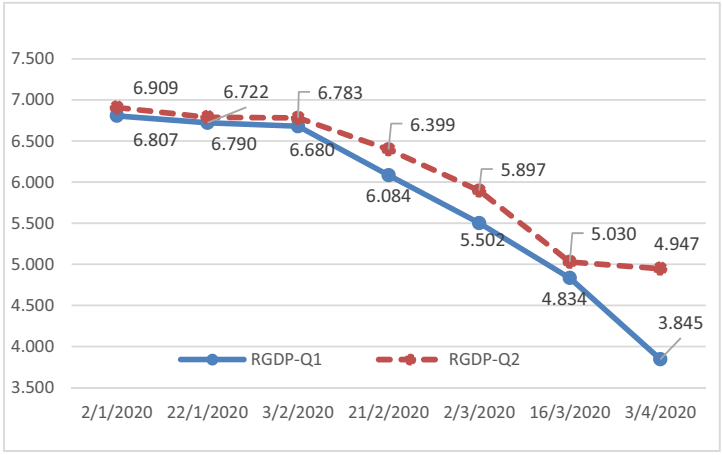

Fig. 4. Real-time forecasts of Vietnam's quarter RDGP

As known, 2019 is the year that Vietnam's economy has developed very well compared to the previous ten years: high economic growth, low inflation, high industrial production indexes, low unemployment rate, high total import and export, and the highest trade surplus ever, public debt decreased, Vietnam's stock index VNindex increased sharply. Vietnam's RGDP in the 1–4 quarters of 2019 was 6.8%, 6.8%, 7.31%, and 6.97%, respectively. The economy continued to develop exceptionally well until the end of January 2020, and since then, the implementation of many measures to combat the Covid-19 epidemic has strongly affected economic activities. The actual RGDP in the first quarter, 2020 is only **3.82%** (released by the GSO on March 30, 2020), much lower than in previous years.

Updating the real-time forecasts of RGDP in Fig. 3 shows that on January 2, 2020, Vietnam's quarter RGDP in the first and second quarters were forecasted to be relatively high, almost equal to the RGDP of the first and second quarters of the year 2019, respectively. However, from the beginning of February 2020, Vietnam's quarter RGDP in the first and second quarters decreased rapidly according to the extent of social distancing measures of Covid-19 epidemic prevention.

6 Conclusions

This article built the nowcasting model of the RGDP indicator based on a large number of potential predictors at many different frequencies using the most suitable model in the two models to be the factor bridge equation model and the factor U-MIDAS. Here, the factors are extracted in the set of original predictors using a combination of two-dimensional reduction techniques: feature selection using the Pearson correlation coefficient measure and feature learning technique using the KTPCA method based on an RMSE-best model. The built model is used to update forecasts of the RGDP indicator in real-time data follow.

The feature selection technique based on the Pearson correlation coefficient measure is used to eliminate noisy and redundant information, while the feature learning technique using the KTPCA method based on an RMSE-best model is used to extract a few new factors, but it still retains as much crucial information as possible of the dataset of original predictors. According to the proposed variable dimension reduction method, building forecasting and nowcasting models on large datasets are not divided into two phases but aggregated in one.

The paper proposes the method to handle ragged-edge data and the reinforcement learning method to reuse the previous variable dimension reduction results for updating forecasts in a real-time data flow. The approach of building the nowcasting model of the RGDP indicator in this article can be used to build nowcasting models for other macroeconomic indicators.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Giannone, D., Reichlin, L., Small, D.H.: Nowcasting: the real-time informational content of macroeconomic data. J. Monet. Econ. **55**(4), 665–676 (2008)
2. Kapetanios, G., Papailias, F., et al.: Big Data & Macroeconomic Nowcasting: Methodological Review. Economic Statistics Centre of Excellence, National Institute of Economic and Social Research (2018)
3. Bok, B., Caratelli, D., Giannone, D., Sbordone, A.M., Tambalotti, A.: Macroeconomic nowcasting and forecasting with big data. Ann. Rev. Econ. **10**, 615–643 (2018)
4. Baldacci, E., et al.: Big Data and Macroeconomic Nowcasting: From Data Access to Modelling. Luxembourg: Eurostat. http://dx.doi.org/10.2785/360587 (2016)
5. Doornik, J.A., Hendry, D.F.: Statistical model selection with 'big data.' Cogent Econ. Finance **3**(1), 1045216 (2015)
6. Altissimo, F., Cristadoro, R., Forni, M., Lippi, M., Veronese, G.: New Euro coin: tracking economic growth in real-time. Rev. Econ. Stat. **92**(4), 1024–1034 (2010)
7. Stock, J.H., Watson, M.W.: Forecasting using principal components from a large number of predictors. J. Am. Stat. Assoc. **97**(460), 1167–1179 (2002)
8. Doz, C., Giannone, D., Reichlin, L.: A two-step estimator for large approximate dynamic factor models based on Kalman filtering. J. Econ. **164**(1), 188–205 (2011)

9. Kitchin, R., McArdle, G.: What makes big data, big data? exploring the ontological characteristics of 26 datasets. Big Data Soc. **3**(1), 2053951716631130 (2016)
10. Giannone, D., Reichlin, L., Small, D.H.: Nowcasting RGDP and inflation: the real-time informational content of macroeconomic data releases. ECB Working Article no. 633 (2006). 51p. http://hdl.handle.net/10419/153067
11. Panagiotelis, A., Athanasopoulos, G., Hyndman, J.H., Jiang, B., Vahid, F.: Macroeconomic forecasting for australia using a large number of predictors. Int. J. Forecast. **35**(2), 616–633 (2019)
12. Kim, H.H., Swanson, N.R.: Methods for Pastcasting, Nowcasting, and Forecasting Using Factor-MIDAS with an Application to Real-Time Korean GDP. Mimeo, Rutgers University (2015). 51p.
13. Kim, H.H., Swanson, N.R.: Mining big data using parsimonious factor, machine learning, variable selection, and shrinkage methods. Int. J. Forecast. **34**(2), 339–354 (2018)
14. Chikamatsu, K., Hirakata, N., Kido, Y., Otaka, K., et al.: Nowcasting Japanese GDPs. Bank of Japan (2018)
15. Bragoli, D.: Now-casting the Japanese economy. Int. J. Forecast. **33**(2), 390–402 (2017)
16. Castle, J.L., Hendry, D.F., Kitov, O.I.: Forecasting and Nowcasting Macroeconomic Variables: A Methodological Overview. Discussion Paper No. 674. University of Oxford (2013). 73p. ISSN 1471-0498
17. Foroni, C., Marcellino, M.: A comparison of mixed frequency approaches for nowcasting Euro area macroeconomic aggregates. Int. J. Forecast. **30**(3), 554–568 (2014)
18. Foroni, C., Marcellino, M.G.: A Survey of Econometric Methods for Mixed-Frequency Data (2013). Available at SSRN 2268912
19. Guresen, E., Kayakutlu, G., Daim, T.U.: Using artificial neural network models in stock market index prediction. Expert Syst. Appl. **38**(8), 10389–10397 (2011)
20. Kara, Y., Boyacioglu, M.A., Baykan, Ö.K.: Predicting the direction of stock price index movement using artificial neural networks and support vector machines: the sample of the Istanbul stock exchange. Expert Syst. Appl. **38**(5), 5311–5319 (2011)
21. Wang, J., Wang, J.: Forecasting stock market indexes using principal component analysis and stochastic time-effective neural networks. Neurocomputing **156**, 68–78 (2015)
22. Hoseinzade, E., Haratizadeh, S.: CNNpred: CNN-based stock market prediction using a diverse set of variables. Expert Syst. Appl. **129**, 273–285 (2019)
23. Lauzon, F.Q.: An introduction to deep learning. In: 11th International Conference on Information Science, Signal Processing and Their Applications (ISSPA), pp. 1438–1439 (2012)
24. Bańbura, M., Rünstler, G.: A look into the factor model black box: publication lags and the role of hard and soft data in forecasting GDP. Int. J. Forecast. **27**(2), 333–346 (2011)
25. Urasawa, S.: Real-time GDP forecasting for Japan: a dynamic factor model approach. J. Jpn. Int. Econ. **34**, 116–134 (2014)
26. Baffigi, A., Golinelli, R., Parigi, G.: Bridge models to forecast the Euro area GDP. Int. J. Forecast. **20**(3), 447–460 (2004)
27. Ghysels, E., Santa-Clara, P., Valkanov, R.: The MIDAS Touch: Mixed Data Sampling Regression Models (2004). https://escholarship.org/uc/item/9mf223rs
28. Bai, J., Ghysels, E., Wright, J.H.: State space models and MIDAS regressions. Econ. Rev. **32** (7), 779–813 (2013)
29. Ankargren, S., Lindholm, U.: Nowcasting Swedish RGDP Growth. Working Article 154, Published by the National Institute of Economic Research (NIER) (2021). ISSN 1100-7818, 33p.
30. Shlens, J.: A tutorial on principal component analysis. ArXiv Preprint ArXiv:1404.1100 (2014)

31. Van Der Maaten, L., Postma, E.: Dimensionality reduction: a comparative review. J. Mach. Learn. Res. **10**, 66–71 (2009)
32. Greene, W.H.: Econometric Analysis. Prentice-Hall (2002). ISBN 0-13-066189-9
33. Ghysels, E., Kvedaras, V., Zemlys, V.: Mixed frequency data sampling regression models: the R package MIDAS. J. Stat. Softw. **72**(1), 1–35 (2016)
34. Kim, K.I., Franz, M.O., Scholkopf, B.: Iterative Kernel principal component analysis for image modeling. IEEE Trans. Pattern Anal. Mach. Intell. **27**(9), 1351–1366 (2005)
35. Schölkopf, B., Smola, A.: A short introduction to learning with Kernels. In: Mendelson, S., Smola, A.J. (eds.) Advanced Lectures on Machine Learning. LNCS (LNAI), vol. 2600, pp. 41–64. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-36434-X_2
36. Ma, X., Zabaras, N.: Kernel principal component analysis for stochastic input model generation. J. Comput. Phys. **230**(19), 7311–7331 (2011)
37. Marcellino, M., Schumacher, C.: Factor MIDAS for nowcasting and forecasting with ragged-edge data: a model comparison for German GDP. Oxford Bull. Econ. Stat. **72**(4), 518–550 (2010)