ARTICLE IN PRESS

Economic Modelling xxx (xxxx) xxx

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.journals.elsevier.com/economic-modelling

Forecasting tourism with targeted predictors in a data-rich environment

Nuno Lourenço[a,b], Carlos Melo Gouveia[a], António Rua[a,b,*]

[a] Banco de Portugal, Portugal
[b] Nova SBE, Portugal

ARTICLE INFO

JEL classification:
C53
C55
F47

Keywords:
Forecasting
Tourism
Factor models
Large datasets

ABSTRACT

Along with the deepening of globalization and economic integration, economic agents face the challenge on how to extract useful information from large panels of data for forecasting purposes. Herein, we lay out a modelling strategy to explore the predictive content of large datasets for tourism forecasting. In particular, we assess the role of multi-country datasets to nowcast and forecast tourism by resorting to factor models with targeted predictors to cope with such a data-rich environment. Drawing on business and consumer surveys for Portugal and its main tourism source markets, we document the usefulness of factor models to forecast tourism exports up to several months ahead. Moreover, we find that forecast performance is enhanced if predictors are chosen before factors are estimated.

1. Introduction

The past five decades saw an increasing interest in tourism econometric modelling and forecasting techniques. One of the reasons relates to the rapid growth of the tourism sector, which is often referred as one of the most prominent economic trends for many countries. According to the World Tourism Organization, international tourist arrivals attained 1323 million in 2017 and grew for the eighth consecutive year, a series of continuous growth not observed since the 1960s.

Given the increasing importance of tourism within the ongoing globalization process, it is natural that a lot of effort is being devoted to enhance and improve tourism forecasting models. Besides the interest of forecasting tourism developments, which is important *per se* for private and public managers, more accurate forecasts for tourism can also be valuable for improving the forecasting performance of economic activity as a whole. This turns out to be particularly relevant for central banks and international institutions or private professional forecasters when nowcasting and short-term forecasting GDP. In fact, there is evidence that a bottom-up approach may deliver a better forecasting performance than forecasting GDP directly. In this respect, see Perevalov and Maier (2010) for the United States, Esteves (2013) for the euro area and, more recently, Dias et al. (2018a) for Portugal.

Early contributions to the tourism forecasting literature date back to the 1960s, focusing mainly on static regressions or univariate models that build on previous values of the forecast variable. Recent empirical applications along these lines include Chu (2004) for Singapore, Coshall (2005) for the United Kingdom, Gil-Alana (2005) for the United States or Chu (2008, 2009) for several countries in the Asian-Pacific region. Notable progress has been made since then, with the rise of vector autoregressions or cointegration techniques. In this regard, multivariate forecasting models have received increasing attention (see, *inter alia*, González and Moral (1995) for Spain, Song et al. (2003) for Denmark, Veloce (2004) for Canada, Han et al. (2006) for the United States, Song and Witt (2006) for Macau, Athanasopoulos and Hyndman (2008) for Australia or Song et al. (2011) for Hong Kong). For a comprehensive review of the early literature on tourism forecasting see Witt and Witt (1995). More recently, Li et al. (2005), Song and Li (2008), Goh and Law (2011), Athanasopoulos et al. (2011) and Peng et al. (2014) provide an encompassing review of studies with emphasis on the latest advances on tourism econometric modelling and forecasting.

Notwithstanding the modelling techniques applied, all these studies operate within the framework of small datasets. However, with the enlargement and rapid dissemination of statistical information observed in the recent past, the information set available to private and public managers has become progressively larger. Such a data-rich environment poses challenges as to how all the available data can be taken into account, which can comprise a large number of series. In particular, qualitative surveys of economic activity conducted in the European

* Corresponding author. Banco de Portugal, Av. Almirante Reis, 71, 1150-012, Lisboa, Portugal.
E-mail address: antonio.rua@bportugal.pt (A. Rua).

https://doi.org/10.1016/j.econmod.2020.03.030
Received 29 October 2019; Received in revised form 25 March 2020; Accepted 30 March 2020
Available online XXX

Please cite this article as: Lourenço, N., et al., Forecasting tourism with targeted predictors in a data-rich environment, Economic Modelling, https://doi.org/10.1016/j.econmod.2020.03.030

Union have been widely available (e.g., business and consumer surveys released by the European Commission for different sectors).

A key advantage of using survey-based indicators arises from their timeliness, as in general surveys are published just a few days after (or even a few days before) the reference period, which contrasts with the considerable release lag of 'hard' data. These surveys usually encompass a wide range of sectors and as several questions are forward-looking in nature, they may signal future developments. Furthermore, qualitative indicators are not revised, thus, real-time reliability can also be granted. Previous literature has highlighted the importance of resorting to 'soft' data for forecasting macroeconomic variables (see, for example, Hansson et al. (2005) for an application to forecast GDP in Sweden, Schumacher (2007) for Germany, Rünstler et al. (2009) for several European countries, Angelini et al. (2011) and Bańbura and Rünstler (2011) for the euro area).

Forecasting macroeconomic variables in a data-rich environment corresponds to extract valuable information from a wide variety of series. The predominant framework to exploit the predictive content embedded in large datasets is through factor models. This type of models has proved to be effective to summarize the informational content of the dataset into a few factors used *ex-post* for forecasting. In essence, factor models allow to circumvent the curse of dimensionality in the presence of a large panel of series by reducing the number of variables to a manageable scale. On the use of factor models, one should mention, *inter alia*, the seminal contributions by Stock and Watson (1999b, 2002a,b) to forecast US macroeconomic aggregates, Marcellino et al. (2003) for the euro area, Artis et al. (2005) for the United Kingdom, Schumacher (2007) and Schumacher and Breitung (2008) for Germany, Giannone et al. (2008) for the United States, Rünstler et al. (2009) for a cross-country study comprising several European countries, Barhoumi et al. (2010) for France or den Reijer (2013) for the Netherlands.

We depart from previous literature on tourism forecasting by resorting to large datasets. However, enlarging a dataset for factor estimation might not enhance forecast accuracy. In fact, forecast performance can be mitigated if the additional predictors are noisy or if predictive power stems from a factor that is dominant in a smaller dataset but is a dominated factor in a larger one (see Boivin and Ng (2006)). Hence, it is important to reduce the influence of uninformative predictors. Bai and Ng (2008) suggest the use of penalized regression to target predictors namely by resorting to Least-Angle Regression with Elastic Net, henceforth LARS-EN, where a subset of variables is selected before factors are estimated. Their empirical application focuses on US inflation. The relevance of screening predictors prior to factor estimation is reinforced by the work of Schumacher (2007), who forecasts German GDP growth and Li and Chen (2014) who concentrate on the US economy. Such an approach may be particularly useful in the context of tourism forecasting as one can easily end up with datasets that include hundreds of series, especially if one intends to cover economic indicators regarding both the destination and origin countries.

In a data-rich setting, the number of studies assessing the importance of taking on board foreign data to forecast domestic macroeconomic series is rather limited. In this respect, it is worth mentioning the work by Brisson et al. (2003) who evaluate the usefulness of variables regarding the United States and other countries to forecast real GDP and inflation in Canada. Within the euro area, Schumacher (2010) assesses the role played by the euro area and the G7 economies to forecast activity in Germany. More recently, Dias et al. (2018b) forecast exports of goods in Portugal by resorting to data on Portuguese main trading partners. We also contribute to this literature by investigating the role of international data to forecast tourism exports.

Herein, we focus on Portugal which is a small open economy where tourism has become a major driver for GDP growth, namely in the aftermath of a severe economic and financial crisis. Besides considering domestic variables we extend the dataset to cover the country's main tourism source markets namely the United Kingdom, France, Spain, Germany, the United States and the Netherlands. Both for Portugal and these countries, we focus on 'soft' data and collect business and consumer surveys covering several sectors of economic activity. Given the large size of the dataset, we use the LARS-EN based pre-selection of variables and assess the usefulness of selecting series before the estimation of factors to improve forecast accuracy. We use timely monthly variables to nowcast and forecast monthly Portuguese tourism exports up to a 6-month ahead horizon. We find that the use of targeted predictors improves forecasting performance in such a data-rich environment. The results also reinforce the usefulness of taking on board economic data from the countries of origin for forecasting tourism.

The remainder of the paper proceeds as follows. In section 2, we present the econometric approach pursued. Section 3 describes the data considered in the empirical application. The empirical results are discussed in section 4. Section 5 concludes.

2. Econometric methodology

2.1. Factor model representation and estimation

In this subsection, we lay out the representation of factor models underlying the pursued method for forecasting with large datasets. Define X_t as a N-dimensional column vector containing the N predictors observed throughout time $t = \{1, \dots, T\}$. We assume that both the predictors and the forecasted variable, y, are stationary.

The factor model considers that each and every variable in X_t is represented as the sum of two orthogonal components: a common component, driven by a small number of unobserved common factors and an idiosyncratic component, driven by variable-specific shocks. Formally, the data generating process for X_t can be represented through a static factor representation given by

$$X_t = \Lambda F_t + \xi_t \tag{1}$$

with $F_t = (f_{1t}, \dots, f_{rt})'$ denoting an $(r \times 1)$ vector of latent factors, Λ corresponds to an $(N \times r)$ matrix of unknown factor loadings and ξ_t is a N-dimensional vector of idiosyncratic terms.[1]

The space spanned by the latent factors can be estimated through the principal components estimator which has been shown to be consistent under relatively general assumptions (see Stock and Watson (1998, 2002b), Bai and Ng (2002) and Amengual and Watson (2007)).

Once the factors are estimated, the estimation of the forecasting equation for the variable of interest follows. Hence, to obtain forecasts for variable y at horizon h, one should regress y_{t+h} on the r estimated factors and eventually on lags of y, that is

$$y_{t+h} = \alpha_0 + \sum_{i=1}^{r} \alpha_i \widehat{F}_{t,i} + \sum_{j=0}^{p} \delta_j y_{t-j} + \epsilon_{t+h} \tag{2}$$

where α_0 is a constant term, α_i denotes the coefficients associated with the estimated factors $\widehat{F}_t$, y_{t-j} denotes the autoregressive terms of the model, where δ_j are the respective coefficients and p is the number of autoregressive terms.

2.2. The Elastic Net optimization problem

Although a small set of r estimated factors may account for a considerable share of the communality of the series within the dataset, the estimation of the factors is completely independent of the series to be forecasted or the forecast horizon at stake. Hence, potentially useful information contained in the dataset may end up being disregarded. In this regard, Bai and Ng (2008) suggest to estimate the factor space from

[1] This representation is without loss of generality as it can be shown that the dynamic factor model representation has an equivalent static factor formulation (see, for instance Stock and Watson (2005a)). In addition, as argued by Bai and Ng (2007), such distinction is not relevant for forecasting purposes.

a set of targeted predictors. Drawing on the relationship between y_{t+h} and X_t, a subset of predictors $X_{t,\mathcal{A}} \subseteq X_t$ is selected prior to factor estimation. The proposed method relies on penalized regressions and conducts subset selection and shrinkage by removing uninformative predictors. Basically, the regression coefficients of less informative predictors to forecast the variable of interest are more penalized. In line with Zou and Hastie (2005), Bai and Ng (2008) consider the following EN optimization problem

$$\min_{\beta} \left\{ RSS + \lambda_1 \sum_{j=1}^{N} |\beta_j| + \lambda_2 \sum_{j=1}^{N} \beta_j^2 \right\} \tag{3}$$

where RSS denotes the residual sum of squares from a regression of y_{t+h} on all predictors, β_j is the regression coefficient of regressor j, and the parameters λ_1 and λ_2 control the penalties associated with the L_1- and L_2-norm of β, respectively.

The L_1 penalty solves

$$\hat{\beta} = \arg\min_{\beta} \left\{ RSS + \lambda_1 \sum_{j=1}^{N} |\beta_j| \right\} \tag{4}$$

where the tuning parameter λ_1 controls for the degree of shrinkage, or equivalently for the number of variables to be dropped. It augments the usual ordinary least squares regression with λ_1 regularization, leading to solutions that are sparse in terms of the coefficients. Such method is also known as the Least Absolute Shrinkage and Selection Operator (LASSO) (see Tibshirani (1996)).

The L_2 penalty leads to

$$\hat{\beta} = \arg\min_{\beta} \left\{ RSS + \lambda_2 \sum_{j=1}^{N} \beta_j^2 \right\} \tag{5}$$

where for $0 \le \lambda_2 < \infty$ shrinks the coefficients of uninformative regressors toward zero. This corresponds to the L_2 penalty of ridge regression.

The EN in (3) combines both penalties, i.e., the merits of LASSO and ridge regression and, thus, allows for shrinkage of regression coefficients, exclusion of regressors and efficient selection of predictors within the dataset.

2.3. The Least-Angle Regression algorithm

The EN optimization problem can be solved efficiently by resorting to the LARS algorithm (see Zou and Hastie (2005)). Conditional on the parameters λ_1 and λ_2, the LARS algorithm allows to estimate β and select the subset of predictors $X_{t,\mathcal{A}} \subseteq X_t$ corresponding to the minimization criterion in (3). It can also be shown that choosing the value for parameter λ_1 corresponds to setting the maximum number of regressors with non-zero β_j, i.e., the number of predictors $N_{\mathcal{A}} \le N$ to be included in $X_{t,\mathcal{A}}$.

The rationale of the LARS algorithm is the following. Firstly, with all coefficients set to zero, it finds the most correlated variable with the series of interest. Then, it considers the largest step possible towards this regressor until it finds another one that has as much correlation with the residual. Instead of proceeding towards the first variable, LARS moves in an equiangular direction between the two regressors, that is, along the least angle direction, until a third predictor is included in the subset of predictors. Then, it proceeds equiangularly between the three predictors until a fourth predictor enters and so on. In this way, the LARS algorithm estimates $\hat{\mu} = X\hat{\beta}$ in sequential steps, each step including one more regressor to the model. This implies that after k stages only k of the $\hat{\beta}_j$'s are non-zero.

Formally, following Efron et al. (2004), the LARS algorithm begins at $\hat{\mu}_0 = \mathbf{0}$ and builds up $\hat{\mu}$ by steps. Let $\hat{\mu}_{\mathcal{A}}$ be the current LARS estimate and

$$\hat{c} = X'(y - \hat{\mu}_{\mathcal{A}}) \tag{6}$$

the vector of current correlations. Define $\mathcal{A}$ as the set of indices corresponding to the variables with the largest absolute current correlations,

$$\hat{C} = \max_j \left\{ \left|\hat{c}_j\right| \right\} \text{ and } \mathcal{A} = \left\{ j : \left|\hat{c}_j\right| = \hat{C} \right\}. \tag{7}$$

Setting $s_j = sign\left\{\hat{c}_j\right\}$ for $j \in \mathcal{A}$, one computes $X_{\mathcal{A}}$, $A_{\mathcal{A}}$, $u_{\mathcal{A}}$ as well as the inner product vector

$$a \equiv X' u_{\mathcal{A}}. \tag{8}$$

For $\mathcal{A}$ a subset of indices, define the matrix

$$X_{\mathcal{A}} = (s_j X_j)_{j \in \mathcal{A}} \tag{9}$$

where the signs s_j equal ± 1.

Define

$$G_{\mathcal{A}} = X'_{\mathcal{A}} X_{\mathcal{A}} \text{ and } A_{\mathcal{A}} = (\mathbf{1}'_{\mathcal{A}} G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}})^{-1/2}, \tag{10}$$

where $\mathbf{1}_{\mathcal{A}}$ is a vector of ones of length equaling $|\mathcal{A}|$, the size of $\mathcal{A}$. The equiangular vector $u_{\mathcal{A}} = X_{\mathcal{A}} w_{\mathcal{A}}$, where $w_{\mathcal{A}} = A_{\mathcal{A}} G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}$ is the unit vector making equal angles less than 90°, with the columns of $X_{\mathcal{A}}$,

$$X'_{\mathcal{A}} u_{\mathcal{A}} = A_{\mathcal{A}} \mathbf{1}_{\mathcal{A}} \text{ and } \|u_{\mathcal{A}}\|^2 = 1. \tag{11}$$

Then, the LARS algorithm updates $\hat{\mu}_{\mathcal{A}}$ to

$$\hat{\mu}_{\mathcal{A}_+} = \hat{\mu}_{\mathcal{A}} + \hat{\gamma} u_{\mathcal{A}}, \tag{12}$$

where

$$\hat{\gamma} = \min_{j \in \mathcal{A}^c}{}^{+} \left\{ \frac{\hat{C} - \hat{c}_j}{A_{\mathcal{A}} - a_j}, \frac{\hat{C} + \hat{c}_j}{A_{\mathcal{A}} + a_j} \right\}. \tag{13}$$

The plus sign indicates that the minimum is taken over positive entries only within each choice of j.

3. Data

3.1. Tourism exports

The empirical application consists in forecasting the growth rate of nominal tourism exports for Portugal (see Fig. 1).[2] Tourism flows are released on a monthly basis by the Portuguese central bank, Banco de Portugal, without any seasonal or calendar adjustment. As depicted below, the series exhibits a high volatility, with year-on-year rates of change varying from −20 per cent to close to 40 per cent for the period under analysis. The spikes in June 2004 and April 2017 reflect the UEFA European Championship hosted in Portugal and the Pope's visit to the country, respectively.

We forecast the year-on-year growth rate of Portuguese tourism exports. By considering year-on-year growth rates, one purges the effect of deterministic seasonality and avoids the high volatility of the month-on-month growth rates of tourism flows. Furthermore, there is widespread evidence of larger resemblance between macro variables measured in changes from the previous year and the evolution of survey-based indicators. It also downplays the irregular component present in the series. Nevertheless, one should address calendar effects or moving holidays in model estimation and forecasting, as these are likely to influence the year-on-year growth rate of tourism exports.

3.2. The multi-country dataset

In this study, we compile a comprehensive dataset for Portugal which is augmented to account for its main tourism source markets, specifically the United Kingdom, France, Spain, Germany, the United States and the Netherlands. These countries explain two thirds of the

[2] In balance of payments data, this variable is recorded under the heading travel exports.

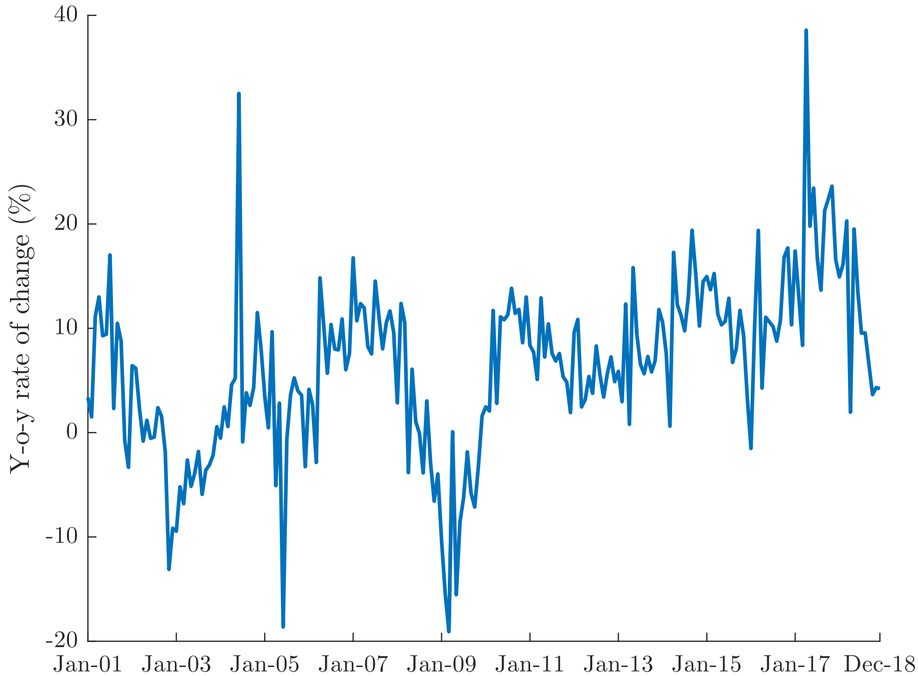

Fig. 1. Portuguese tourism exports. Source: Statistical Bulletin, Banco de Portugal.

inbound tourism revenues in Portugal. The United Kingdom and France account for the largest share, representing more than 16 per cent each in 2018. These are followed by Spain and Germany, which account for around 13 and 11 per cent of tourism exports, respectively. The share of the United States stands close to 6 per cent in 2018 whereas the Netherlands has a share slightly above 4 per cent.

Data for Portugal and its main European Union source markets draws on the business and consumer surveys released by the European Commission. The panel of variables encompasses qualitative data covering different sectors of the economies. Representatives of the industry (manufacturing), services, retail trade and construction sectors, as well as to consumers are asked on several domains. Questions in the industry survey include assessments of recent and future trends in production, of the current levels of order books and stocks, selling price expectations and employment. In the services survey, managers are asked about their assessment on the business situation, of the past and future changes in their company's turnover and employment and of their expecting selling prices. The retail trade survey is focused on assessments of recent developments in managers' business situation, of the current level of stocks, and their expectations regarding production, new orders, prices charged and employment. Similar questions are asked in the construction survey to infer on the short-term developments in this sector. Finally, the consumers' survey collects information on households' spending and savings intentions and measures their understanding of the factors that affect those decisions. Hence, questions are grouped around four topics: general economic situation, households' financial situation, savings and intentions with regard to major purchases. Questions concerning perceived and expected price changes are also included.

For the United States, the business surveys in the manufacturing and non-manufacturing sectors released by the Institute for Supply Management are used, in addition to the Conference Board and the University of Michigan consumer surveys. These surveys include, *inter alia*, questions on customer inventories, orders, prices paid, employment expectations as well as consumers' sentiment or expectations.

The sample period spans January 2000 to December 2018. On average, 40 series per country (20 regarding the United States) are covered, amounting to 257 series overall.[3]

For Portugal and Spain, we use the Expectation-Maximization algorithm suggested by Stock and Watson (2002a) to balance the dataset at the beginning of the sample, as a few series were not available for the full sample period. Following Stock and Watson (2005a), the series were also screened for outliers.

4. Empirical application

4.1. Design of the forecasting exercise

The forecasting exercise is performed in a fully recursive way. This means that, for each time period t, predictors are selected from the large dataset through the LARS-EN algorithm using data available up to t. Then, drawing on principal components, factors are estimated from the set of selected predictors in the previous stage. Afterwards, the forecasting equation in (2) is estimated with the number of estimated factors, r, chosen according to a modified version of the BIC criterion as in Stock and Watson (1998). Given the previous discussion on calendar effects, we also include in the model specification a deterministic variable to account for the number of working days in each month and dummy variables to control for the two moving holidays (Easter and Carnival) as well as for the above-mentioned events hosted in Portugal. Finally, the fitted model is used to produce h-step ahead out-of-sample forecasts.

One should note that, in this way, we are not imposing the same set of predictors over time and neither across forecast horizons. Since the model specification and estimation are allowed to be updated conditional on the information available up to time period t, we replicate what one could actually do at each point in time. Furthermore, to deal with the potential varying informational content of the dataset, we considered a rolling window estimation scheme so as to enhance model flexibility. In particular, we have chosen a window size such that the estimation period always encompasses a full cycle and therefore it is not influenced only by upward or downward movements. As it has been standard in the literature to consider as business cycles the fluctuations

[3] The list of series is available from the corresponding author upon request.

Table 1
Relative MSFE.

	Forecast horizon						
	0	1	2	3	4	5	6
Targeted predictors ($N_{\mathcal{A}}$)							
30	1.07	0.94	0.93	0.93	0.93	1.06	1.11
40	1.05	1.00	0.89	0.92	0.98	1.04	1.10
50	1.00	0.99	0.82	0.91	0.95	0.97	1.01
60	1.01	0.96	0.81	0.92	0.98	0.91	0.94
70	0.99	0.96	0.79	0.92	0.98	0.87	0.89
80	0.96	0.95	0.78	0.89	0.99	0.82	0.88
90	0.98	0.96	0.79	0.92	1.01	0.92	0.89
100	0.96	0.93	0.79	0.88	0.98	0.89	0.86
110	0.90	0.94	0.78	0.91	0.99	0.92	0.83
120	0.90	0.94	0.82	0.84	0.98	0.97	0.92
130	0.87	0.94	0.83	0.84	0.99	0.95	0.89
140	0.85	0.93	0.81	0.81	0.92	1.00	0.90
150	0.86	0.91	0.76	0.83	0.91	0.99	0.95
160	0.85	0.85	0.76	0.85	0.91	0.97	0.96
170	0.84	0.85	0.77	0.84	0.94	0.93	0.97
180	0.84	0.91	0.79	0.85	0.95	0.92	0.95
190	0.87	0.91	0.80	0.88	0.96	0.96	0.93
200	0.85	0.88	0.81	0.88	0.93	1.00	0.93
No pre-selection							
All series	0.87	0.87	0.80	0.87	0.95	0.99	0.87

that last up to 8 years (see, for example, Stock and Watson (1999a, 2005b)), we set the window size to 96 months.[4]

We also consider the case where no pre-selection of predictors prior to factor estimation is done, which boils down to the standard factor model approach. The usual AR(p) model, where p is the number of autoregressive terms chosen by the BIC criterion, is used as benchmark. Naturally, this model also includes the above mentioned deterministic and dummy variables.

To infer on the relative behaviour of the factor model *vis-à-vis* the benchmark, the out-of-sample forecasts are compared. The forecast evaluation period runs from January 2009 to December 2018, which corresponds to half of the sample period. We consider forecast horizons from 1 to 6-month ahead. However, since business and consumer surveys are released one month before tourism statistics, that is, data for time t is already available for the former whereas for the latter the last figure refers to time $t-1$, one can also consider nowcasting (i.e., $h = 0$).

Forecast accuracy is assessed through the Mean-Squared Forecast Error (MSFE) and the relative MSFE is calculated using the autoregressive model as benchmark. Thus, if this ratio is below one, the competing model outperforms the benchmark. We examine the statistical significance of the forecasting gains using the Clark and West (2007) test.

4.2. Results

In what follows, we consider 'soft' data driven forecasts, i.e., we resort to survey data for Portugal and its main source markets (amounting to a total of 257 series). As mentioned earlier, besides all the advantages inherent to 'soft' data, it allows one to assess model performance for both nowcasting and forecasting.

As discussed in section 2, the practical use of the LARS-EN procedure involves setting two parameters, λ_1 and λ_2. The parameter λ_1 controls the number of predictors to be selected, that is, $N_{\mathcal{A}}$. Given that there is no *a priori* about the optimal number of predictors, we considered a range of alternatives namely $N_{\mathcal{A}} = \{30, 40, \ldots, 200\}$. Regarding λ_2, which controls the importance of the penalty of the L_2-norm of β, we set $\lambda_2 = 0.25$ in line with Bai and Ng (2008) and Schumacher (2010).

In Table 1, we present the relative MSFE of the factor model *vis-à-vis* the univariate autoregressive model for the different forecast horizons ($h = \{0, 1, \ldots, 6\}$). At the bottom of the table, we also report the relative MSFE for the case of a factor model without pre-selection, i.e., considering all 257 series for factor estimation. The shaded entry denotes the minimum relative MSFE for each forecast horizon.

The empirical results obtained convey the following findings. Firstly, since most entries in Table 1 are below one, the factor model yields greater forecast accuracy than the univariate benchmark regardless of the horizon or the number of predictors considered.

Secondly, the results suggest that to forecast at shorter horizons it is preferable to take on board more predictors than at longer horizons. In particular, from $h = 0$ to $h = 4$, greater forecasting gains are delivered for a number of pre-selected predictors between 140 and 170, whereas for $h = \{5, 6\}$, a lower number of targeted predictors (between 80 and 110) yields greater forecasting accuracy gains. As the forecast horizon increases, the number of variables that convey informational content about future developments tends to decrease. We also find that around half of the selected predictors are common across adjacent horizons.

Thirdly, such selection of predictors leads to forecasting gains, on average, of 17 per cent when compared with the univariate benchmark. The statistical significance of such an improvement is corroborated by

[4] As a sensitivity analysis, we have also considered other rolling window sizes and the results are qualitatively similar.

Table 2
Model specification and goodness-of-fit for the best performing model at each horizon.

	$h = 0$ $N_{\mathcal{A}} = 170$	$h = 1$ $N_{\mathcal{A}} = 170$	$h = 2$ $N_{\mathcal{A}} = 160$	$h = 3$ $N_{\mathcal{A}} = 140$	$h = 4$ $N_{\mathcal{A}} = 150$	$h = 5$ $N_{\mathcal{A}} = 80$	$h = 6$ $N_{\mathcal{A}} = 110$
No. of factors	7	8	7	7	7	6	6
	(1.78)	(1.65)	(1.62)	(1.85)	(1.71)	(1.66)	(1.42)
No. of AR terms	1	1	1	1	1	1	1
	(1.76)	(1.57)	(1.78)	(1.73)	(1.64)	(1.59)	(1.28)
R^2	0.85	0.85	0.85	0.85	0.84	0.85	0.85
	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)	(0.05)

Note: Standard deviations in parentheses.

the Clark and West (2007) test procedure.[5]

To provide additional details about the best performing models (denoted by the shaded entries in Table 1) we report in Table 2 the average number of autoregressive terms and the average number of factors selected over the evaluation period (along with the standard deviation over time). Moreover, we report the R^2 to characterize the in-sample fit. In terms of specification, these models include, on average, one autoregressive term while the average number of factors ranges between six and eight with most cases including seven factors. The in-sample fit is quite noteworthy with a R^2 of around 0.85.

Finally, when no pre-selection is done, forecasting gains are lower than those obtained with targeted predictors, reaching 11 per cent. We have also computed the Clark and West (2007) test to compare the pre-selection and no pre-selection cases and found supporting evidence of statistically larger gains in the case of pre-selection. Hence, pre-selection of predictors before factors are estimated plays a role to forecast tourism exports.

We also examine the sensitivity of the results to the choice of λ_2. In particular, following Bai and Ng (2008), we consider $\lambda_2 = \{0.5, 1.5\}$. One can conclude that the main findings highlighted above do not seem to be sensitive to the choice of λ_2. In this respect, Bai and Ng (2008) and Schumacher (2010) also find that the choice for this parameter is not critical for the results.

Up to now the analysis has been based on 'soft' data. As a robustness check, we extend the dataset further to cover the main quantitative indicators of economic activity, namely industrial production, retail trade, activity in the services sector and labour market outcomes for Portugal and its main tourism source markets, amounting to 615 series overall. The results show that augmenting the dataset with quantitative data, or considering only 'hard' data, does not lead to an improvement of forecast accuracy. Moreover, the results with 'hard' data tend to deteriorate if one takes into account the publication lags. Hence, these results reinforce the usefulness of 'soft' data, in line with previous literature, and in particular for tourism forecasting.

4.3. Unveiling the targeted predictors

In this subsection, we intend to provide some insights regarding the selected predictors underlying the results presented in Table 1. Given the large dimension of the dataset, it is not feasible to detail the predictors. Hence, we focus on two important groupings of the 'soft' dataset. On the one hand, we have the country to which the variable belongs to, that is, if it refers to Portugal or to one of its source markets. On the other hand, we have the survey dimension, that is, from which survey comes the predictor. Beginning with the country analysis, we present in Fig. 2 plots for the average share of selected predictors country, for different number of targeted predictors ($N_{\mathcal{A}}$) and forecast horizons (h). A visual inspection immediately highlights that most of the series are selected from Portugal's source markets, which emphasizes the role of foreign data to forecast inbound tourism flows.

In the case of Portugal, when the number of selected predictors is small, the average share is low (close to 10 per cent), while increasing up to 15 per cent as more predictors are allowed to be selected. For the United Kingdom, the average share of series is high for smaller datasets (25 per cent). Although it shows some decrease with the number of predictors, the share of UK series is no less than 15 per cent. Such a role is grounded on the importance of the UK as the main source market for Portugal in what concerns tourism. Regarding the other main source countries, namely France, Spain and Germany, a similar pattern is observed, with average shares varying between 10 and 20 per cent. In turn, the United States presents a low share of targeted predictors, standing below 10 per cent. Finally, even though the Netherlands weighs less in Portuguese tourism exports, a noteworthy share of predictors belong to this country, particularly for a small number of predictors. This may reflect the fact that, as a small open economy, the Netherlands is exposed to the same drivers that affect Portugal's tourism flows. Lastly, the above mentioned shares do not seem to vary much with the forecast horizon, especially for larger datasets.

We now turn to the analysis by survey. Fig. 3 displays the average share of selected predictors by survey, for different number of targeted predictors and forecast horizons. The results clearly reveal that the consumers' survey accounts for the largest average share regardless of the size of the dataset or the forecast horizon. Notwithstanding a share usually above 30 per cent in the case of the consumers' survey, its share is even higher for shorter horizons and smaller datasets, being close to 50 per cent. Hence, among the several available surveys, the one that captures the current and prospective assessment by consumers is the more relevant to nowcast and forecast tourism flows. In turn, the industry survey shows an increasing share with the number of predictors, from less than 5 per cent to slightly above 20 per cent. The surveys regarding services and retail trade display a similar behavior, with shares between 10 and 20 per cent. The construction sector survey evidences a slightly lower importance standing above 10 per cent, whereas the miscellaneous category turns out to be unimportant.

To complement the above graphical analysis, we report in Table 3 the composition of the set of targeted predictors, by country and survey, for the best performing model for each horizon (which corresponds to a specific pair $(h, N_{\mathcal{A}})$ in each graph of Figs. 2 and 3). The results reinforce the above discussion and highlight the importance of foreign data to forecast tourism developments. In fact, domestic variables only account for around 15 per cent of the targeted predictors and for almost all horizons data regarding Portugal is surpassed by some other country. One can also see from Table 3 that the consumers' survey represents the main source of targeted predictors accounting for around one third of the set of predictors across all forecast horizons. The industry survey presents a share slightly above 20 per cent for $h = \{0, 1\}$ but loses importance for longer horizons. In contrast, the surveys concerning services and retail trade are more relevant for longer than shorter horizons.

[5] To save space these results and the ones discussed in the remaining part of this subsection are not included here but are available from the corresponding author upon request.

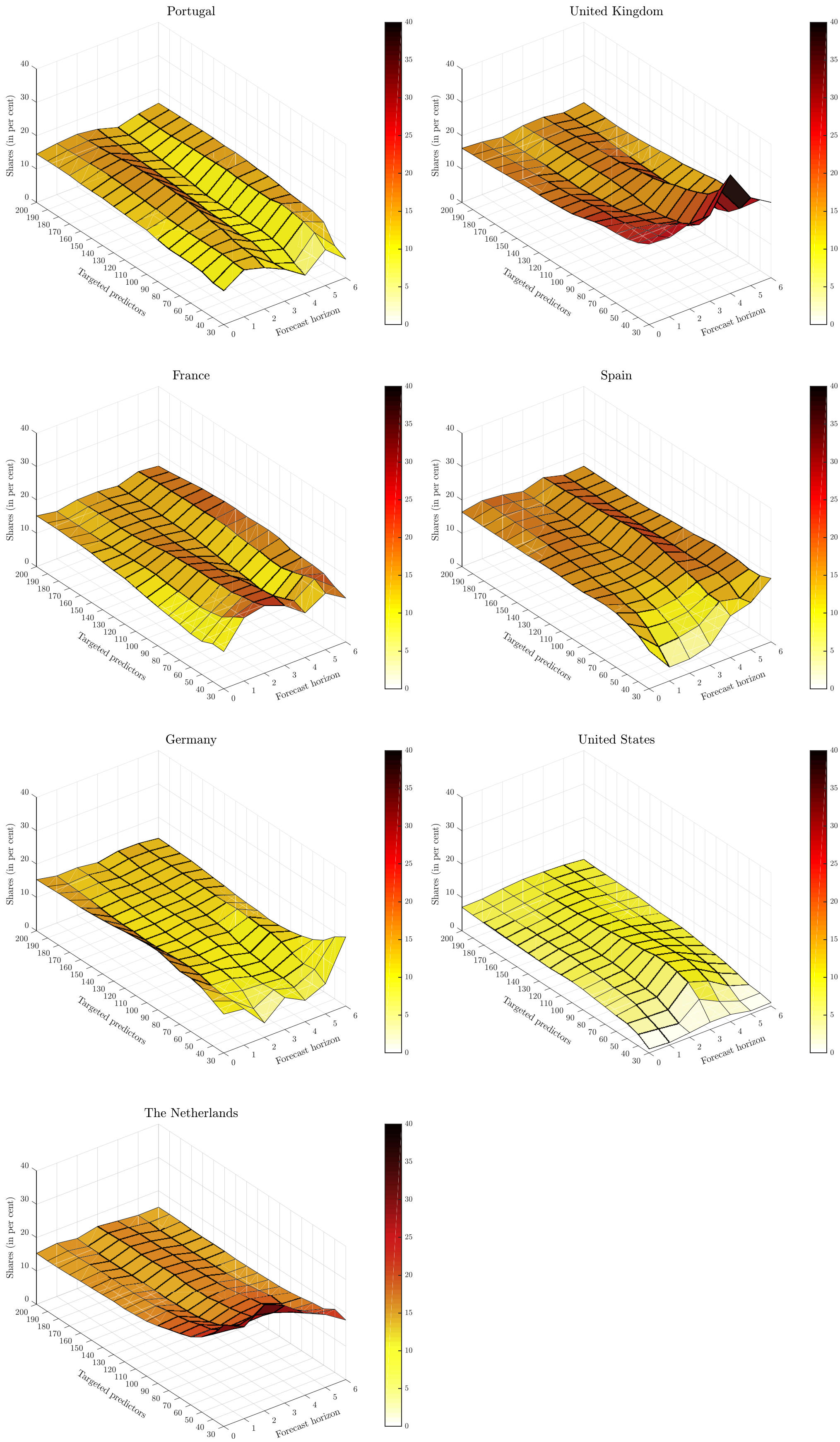

Fig. 2. Average share of selected predictors by country for different number of predictors and forecast horizons.

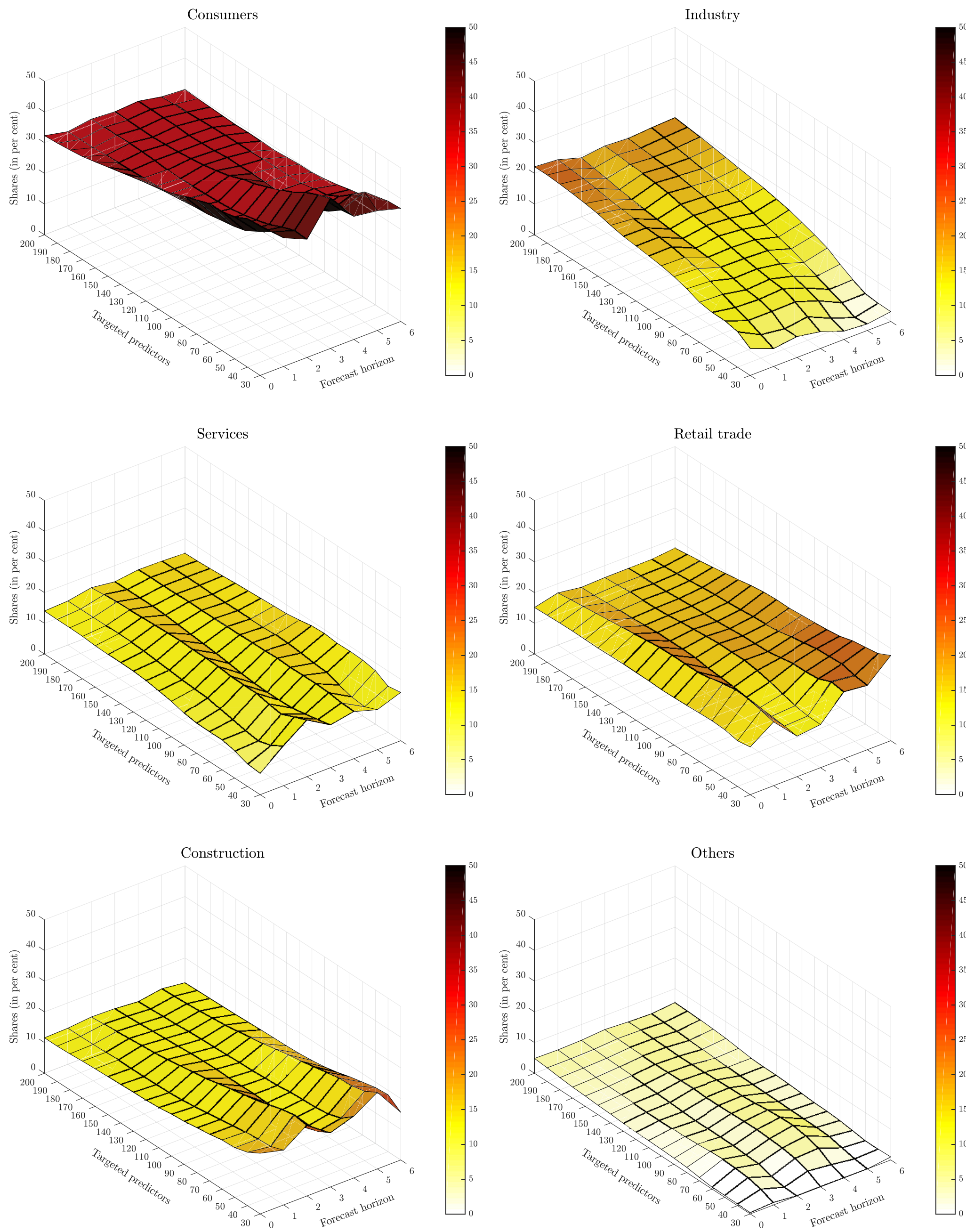

Fig. 3. Average share of selected predictors by survey for different number of predictors and forecast horizons.

In addition, we report in Table 4 the five most selected predictors throughout time for each horizon. The results highlight the value of resorting to international data to forecast tourism exports, in line with the previous discussion.

5. Concluding remarks

In the past decades, the tourism industry has paved the way in driving the prosperity of nations, with direct impact on economic growth, job creation or business investment. Given the importance of tourism worldwide, reinforced by the strong dynamics recently observed in several countries, it is of utmost interest to forecast its developments by private and public managers. Furthermore, more accurate forecasts for tourism can also be valuable to enhance the forecast accuracy of economic activity as a whole. In this respect, there is by now evidence that a bottom-up approach may lead to better forecasting performance than forecasting GDP directly. This is particularly important for central banks and international institutions or private professional forecasters when forecasting GDP.

Monitoring and forecasting tourism developments poses a challenge for economic agents in a context marked by increasing data availability. Hence, decision-makers require new methods and tools to take advantage of the informational content embedded in large datasets. In contrast with previous literature on tourism forecasting, we pursue an approach able to cope with such a data-rich environment. At the same time, our strategy allows to mitigate the influence of uninformative variables for forecasting purposes.

In our empirical application, we exploit the role of a multi-country dataset to nowcast and forecast Portuguese tourism exports on a monthly basis. We make use of factor models with targeted predictors

Table 3
Composition of the set of targeted predictors by country and survey for the best performing model at each horizon.

	$h = 0$ $N_A = 170$	$h = 1$ $N_A = 170$	$h = 2$ $N_A = 160$	$h = 3$ $N_A = 140$	$h = 4$ $N_A = 150$	$h = 5$ $N_A = 80$	$h = 6$ $N_A = 110$
By country							
Portugal	14.5	15.8	16.9	13.5	11.6	15.0	15.4
United Kingdom	16.7	16.1	14.9	16.7	17.3	15.2	16.9
France	15.1	14.1	15.5	15.4	14.4	16.1	18.3
Spain	16.2	17.4	16.7	15.4	18.5	16.3	17.0
Germany	15.6	13.4	13.4	13.0	13.2	13.6	11.4
United States	6.5	7.1	7.9	9.2	8.8	7.1	6.1
The Netherlands	15.3	16.0	14.7	16.8	16.3	16.8	14.7
By survey							
Consumers	33.6	31.1	32.0	32.2	31.6	32.7	29.6
Industry	21.5	20.3	16.4	17.4	12.8	8.7	14.7
Services	13.8	14.4	16.8	13.8	17.6	14.7	17.8
Retail trade	15.5	18.4	16.6	17.8	17.3	21.3	21.6
Construction	12.0	12.0	13.9	13.8	16.0	19.8	13.1
Others	3.6	3.8	4.2	5.1	4.7	2.8	3.1

Table 4
Most selected predictors for each forecast horizon.

Forecast horizon	Country	Survey	Series
h = 0	The Netherlands	Consumer	Financial situation of households over the last 12 months
	The Netherlands	Retail trade	Retail trade confidence indicator
	The Netherlands	Retail trade	Business activity over the last 3 months
	The Netherlands	Construction	Price expectations over the next 3 months
	United Kingdom	Construction	Overall order books
h = 1	Portugal	Services	Business situation over the last 3 months
	France	Retail trade	Employment expectations over the next 3 months
	The Netherlands	Retail trade	Orders placed with suppliers over the next 3 months
	The Netherlands	Retail trade	Retail trade confidence indicator
	The Netherlands	Construction	Price expectations over the next 3 months
h = 2	Portugal	Industry	Employment expectations over the next 3 months
	The Netherlands	Consumer	Financial situation of households over the last 12 months
	The Netherlands	Construction	Price expectations over the next 3 months
	United Kingdom	Construction	Overall order books
	United Kingdom	Services	Employment expectations over the next 3 months
h = 3	France	Consumer	General economic situation over last 12 months
	United Kingdom	Services	Employment expectations over the next 3 months
	The Netherlands	Services	Business situation over the last 3 months
	The Netherlands	Retail trade	Orders placed with suppliers over the next 3 months
	The Netherlands	Construction	Price expectations over the next 3 months
h = 4	The Netherlands	Consumer	Major purchases at present
	The Netherlands	Services	Business activity over the last 3 months
	United Kingdom	Services	Employment expectations over the next 3 months
	United Kingdom	Construction	Construction confidence indicator
	The Netherlands	Construction	Price expectations over the next 3 months
h = 5	France	Consumer	General economic situation over the last 12 months
	Germany	Consumer	Consumer confidence indicator
	The Netherlands	Consumer	Major purchases at present
	The Netherlands	Services	Business situation over the last 3 months
	United Kingdom	Construction	Construction confidence indicator
h = 6	Germany	Consumer	Consumer confidence indicator
	The Netherlands	Consumer	General economic situation over the last 12 months
	The Netherlands	Consumer	Major purchases at present
	Spain	Services	Business situation over the last 3 months
	Spain	Services	Services confidence indicator

to cope with such a large dataset. Drawing on business and consumer surveys for Portugal and its main tourism source markets, namely the United Kingdom, France, Spain, Germany, the United States and the Netherlands, we find significant forecasting gains up to 6-month ahead. Furthermore, we show that forecast performance is enhanced if predictors are pre-selected from the large dataset through the LARS-EN algorithm before factors are estimated. Hence, our results reinforce the usefulness of relying on survey-based data for tourism forecasting. Although the empirical exercise has focused on Portugal, where there has been a striking increase of tourism importance, the framework outlined in this study can be easily extended to other countries or regions.

Acknowledgements

The authors would like to thank two anonymous referees for valuable comments and suggestions.

References

Amengual, D., Watson, M., 2007. Consistent estimation of the number of dynamic factors in a large N and T panel. J. Bus. Econ. Stat. 25 (1), 91–96.

Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L., Rünstler, G., 2011. Short-term forecasts of euro area GDP growth. Econom. J. 14, C25–C44.

Artis, M.J., Banerjee, A., Marcellino, M., 2005. Factor forecasts for the UK. J. Forecast. 24, 279–298.

Athanasopoulos, G., Hyndman, R.J., 2008. Modelling and forecasting Australian domestic tourism. Tourism Manag. 29 (1), 19–31.

Athanasopoulos, G., Hyndman, R.J., Song, H., Wu, D.C., 2011. The tourism forecasting competition. Int. J. Forecast. 27 (3), 822–844.

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

Bai, J., Ng, S., 2007. Determining the number of primitive shocks in factor models. J. Bus. Econ. Stat. 25, 52–60.

Bai, J., Ng, S., 2008. Forecasting economic time series using targeted predictors. J. Econom. 146, 304–317.

Barhoumi, K., Darné, O., Ferrara, L., 2010. Are disaggregate data useful for factor analysis in forecasting French GDP? J. Forecast. 29 (1–2), 132–144.

Bańbura, M., Rünstler, G., 2011. 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, 333–346.

Boivin, J., Ng, S., 2006. Are more data always better for factor analysis? J. Econom. 132, 169–194.

Brisson, M., Campbell, B., Galbraith, J.W., 2003. Forecasting some low-predictability time series using diffusion indices. J. Forecast. 22, 515–531.

Chu, F.L., 2004. Forecasting tourism demand: a cubic polynomial approach. Tourism Manag. 25, 209–218.

Chu, F.L., 2008. Analyzing and forecasting tourism demand with ARAR algorithm. Tourism Manag. 29 (6), 1185–1196.

Chu, F.L., 2009. Forecasting tourism demand with ARMA-based methods. Tourism Manag. 30 (5), 740–751.

Clark, T.E., West, K.D., 2007. Approximately normal tests for equal predictive accuracy in nested models. J. Econom. 138, 291–311.

Coshall, J.T., 2005. A selection strategy for modelling UK tourism flows by air to European destinations. Tourism Econ. 11, 141–158.

den Reijer, A., 2013. Forecasting Dutch GDP and inflation using alternative factor model specifications based on large and small datasets. Empir. Econ. 44, 435–453.

Dias, F., Pinheiro, M., Rua, A., 2018a. A bottom-up approach for forecasting GDP in a data rich environment. Appl. Econ. Lett. 25 (10), 718–723.

Dias, F., Rua, A., Lourenço, N., 2018b. Forecasting exports with targeted predictors. Banco de Portugal Economic Studies 4 (2), 45–62.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32, 407–499.

Esteves, P.S., 2013. Direct vs bottom–up approach when forecasting GDP: reconciling literature results with institutional practice. Econ. Modell. 33, 416–420.

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

Gil-Alana, L.A., 2005. Modelling international monthly arrivals using seasonal long-memory processes. Tourism Manag. 26, 867–878.

Goh, C., Law, R., 2011. The methodological progress of tourism demand forecasting: a review of related literature. J. Trav. Tourism Market. 28 (3), 296–317.

González, P., Moral, P., 1995. An analysis of the international tourism demand in Spain. Int. J. Forecast. 11, 233–251.

Han, Z., Durbarry, R., Sinclair, M.T., 2006. Modelling US tourism demand for European destinations. Tourism Manag. 27, 1–10.

Hansson, J., Jansson, P., Löf, M., 2005. Business survey data: do they help in forecasting GDP growth? Int. J. Forecast. 21, 377–389.

Li, G., Song, H., Witt, S.F., 2005. Recent developments in econometric modeling and forecasting. J. Trav. Res. 44, 82–99.

Li, J., Chen, W., 2014. Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models. Int. J. Forecast. 30, 996–1015.

Marcellino, M., Stock, J., Watson, M., 2003. Macroeconomic forecasting in the euro area: country specific versus euro wide information. Eur. Econ. Rev. 47, 1–18.

Peng, B., Song, H., Crouch, G.I., 2014. A meta-analysis of international tourism demand forecasting and implications for practice. Tourism Manag. 45, 181–193.

Perevalov, N., Maier, P., 2010. On the Advantages of Disaggregated Data: Insights from Forecasting the US Economy in a Data-Rich Environment. Bank of Canada. Staff working papers 2010-10.

Rünstler, G., Barhoumi, K., Benk, S., Cristadoro, R., den Reijer, A., Jakaitiene, A., Jelonek, P., Rua, A., Ruth, K., Van Nieuwenhuyze, C., 2009. Short-term forecasting of GDP using large datasets: a pseudo real-time forecast evaluation exercise. J. Forecast. 28, 595–611.

Schumacher, C., 2007. Forecasting German GDP using alternative factor models based on large datasets. J. Forecast. 26 (4), 271–302.

Schumacher, C., 2010. Factor forecasting using international targeted predictors: the case of German GDP. Econ. Lett. 107, 95–98.

Schumacher, C., Breitung, J., 2008. Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data. Int. J. Forecast. 24 (3), 386–398.

Song, H., Li, G., 2008. Tourism demand modelling and forecasting – a review of recent research. Tourism Manag. 29, 203–220.

Song, H., Li, G., Witt, S.F., Athanasopoulos, G., 2011. Forecasting tourist arrivals using time-varying parameter structural time series models. Int. J. Forecast. 27, 855–869.

Song, H., Witt, S.F., 2006. Forecasting international tourist flows to Macau. Tourism Manag. 27 (2), 214–224.

Song, H., Witt, S.F., Jensen, T.C., 2003. Tourism forecasting: accuracy of alternative econometric models. Int. J. Forecast. 19, 123–141.

Stock, J., Watson, M., 1998. Diffusion Indexes. National Bureau of Economic Research. Working paper no. 6702.

Stock, J., Watson, M., 1999a. Business cycle fluctuations in U.S. macroeconomic time series. In: Taylor, J.B., Woodford, M. (Eds.), Handbook of Macroeconomics. Elsevier Science, Amsterdam.

Stock, J., Watson, M., 1999b. Forecasting inflation. J. Monetary Econ. 44, 293–335.

Stock, J., Watson, M., 2002a. Macroeconomic forecasting using diffusion indices. J. Bus. Econ. Stat. 20, 147–162.

Stock, J., Watson, M., 2002b. Forecasting using principal components from a large number of predictors. J. Am. Stat. Assoc. 97, 1167–1179.

Stock, J., Watson, M., 2005a. Implications of Dynamic Factor Models for VAR Analysis. National Bureau of Economic Research. Working paper no. 11467.

Stock, J., Watson, M., 2005b. Understanding changes in international business cycle dynamics. J. Eur. Econ. Assoc. 3 (5), 968–1006.

Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. J. Roy. Stat. Soc. B 58, 267–288.

Veloce, W., 2004. Forecasting inbound Canadian tourism: an evaluation of Error Correction Model forecasts. Tourism Econ. 10 (3), 263–280.

Witt, S.F., Witt, C.A., 1995. Forecasting tourism demand: a review of empirical research. Int. J. Forecast. 11, 447–475.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B 67, 301–320.