

# Forecasting Occupancy Rate with Bayesian Compressed Methods

## Abstract

The curse of dimensionality is a challenge that researchers often face when dealing with large Vector Autoregressions (VARs). Different approaches have been proposed in the literature to address this issue. In this paper, we propose a new method based on the idea of compressed regression. In particular, we introduce two novel nonlinear compressed VARs to forecast the occupancy rate of hotels that compete within a narrow geographical area. We make the models more flexible through the introduction of neural networks, and compare their performance against several competing models. The empirical results show that the new compressed VARs outperform all other models, and their accuracy is preserved across nearly all forecast horizons from 1 to 36 months.

**Keywords:** Large Vector Autoregressions (VARs), Compressed Methods, Bayesian; Neural Networks; Hotel Occupancy Rate

## 1. Introduction

Forecasting tourism/hotel demand continues to be one of the most widely researched areas in tourism economics (Gunter and Önder, 2016; Wan et al. 2016; Hassani et al. 2017). A variety of methods have so far been introduced ranging from Autoregressive Moving Average Cause Effect (ARMAX) models (Akal, 2004) to combination of forecasts (Chan et al. 2010) and Bayesian methods (Wong et al. 2006; Assaf et al. 2018). Advanced techniques like support vector regression have been proposed and used early on (Chen and Wang, 2007). Nonlinear models like neural networks are also still being employed with some success (Li et al. 2018). Vector Autoregressions (VAR), in particular, have received strong attention in modelling tourism demand (Massidda and Mattana, 2013; Pan et al., 2012; Li et al., 2005; Witt et al., 2004). The method is popular because it relaxes the assumption of exogeneity and allows for reverse causality (e.g. between tourism demand and its determinants).

A problem with VARs, however, is that they suffer from the “curse of dimensionality,” that is, “the situation in which the number of parameters to be estimated grows exponentially with the number of variables in the system, while the size of the data set remains limited” (Assaf et al. 2018, p.3). Large VARs usually result in more parameters than individual data, and, as a result, forecast poorly out-of-sample (Koop et al., 2016). Several solutions to this problem have been proposed in a Bayesian framework, such as adopting prior shrinkage on the parameters (Koop et al. 2016), with some choices including the Minnesota prior, LASSO (least absolute shrinkage and selection operator), and SSVS (stochastic search variable selection) - see Park and Casella (2008), and George, et al. (2008) for more details. However, as discussed in Koop et al. (2016, p.2), most of these solutions will result in computationally intensive Markov Chain Monte Carlo (MCMC) and “their use in recursive forecasting exercises can be computationally infeasible”.

A more effective approach to tackle the problem is to use compression methods, where shrinkage is applied by compressing the data instead of shrinking the parameters towards a given value (for example, zero). In tourism studies, Li et al. (2018) used principal component Analysis (PCA) to reduce the dimensionality of the data and then used a neural network to forecast tourism arrivals. However, a problem with PCA is that it is an “unsupervised data compression method”, where “the data compression is done without reference to the dependent variable(s)” (Koop et al., 2016, p.3). An arguably more effective way to handle this issue is to use the idea of Bayesian compressed regression, introduced by Guhaniyogi and Dunson (2015), where compression is achieved using the concept of Bayesian model averaging (BMA). Koop et al. (2016) generalized the use of this method to the context of large dimensional VARs.

The objective of this study is take up this approach and propose and estimate flexible VAR models. We focus on a problem of forecasting monthly occupancy rates of hotels that compete within the same geographical area (also known as “tract”). Several studies have recently focused on forecasting hotel demand (e.g. Zakhary et al., 2011; Pan et al., 2012; Yang et al., 2014). This is rightly so as hotel management depends on the accuracy of forecasts which, in turn, can promote efficient resource allocation in the hotel industry. One of the contributions of the present study is to model occupancy rate as a function of prices of other hotels within the same tract (geographical area). The merits of this approach are rather obvious as the occupancy rate of a particular hotel within a tract is likely to be affected by the price offering of other hotels competing within the same tract (Schwartz et al., 2016).

The context is a perfect example of large VAR. The dimensionality of the predictor space increases dramatically when the number of hotels and lags increase. The question is then is how to select

predictors that are promising in terms of forecasting accuracy. The curse of dimensionality becomes even worse when we want to set up VARs for prices and occupancy rates of all 21 hotels resulting in 42 dependent variables (assuming we have only one lag) and a very large number of parameters. Standard Bayesian VARs (Wong et al., 2006) become computationally infeasible even when tight priors are adopted.

As the problem of dimensionality is endemic, one solution is to examine each hotel separately (that is, as a single equation in the VAR) and select the appropriate predictors.<sup>1</sup> However, if we do so, we ignore the correlation of errors with other hotels in the same tract and, of course, there is no strong theoretical justification for this. In this paper we use two techniques to deal with this problem. First, we reduce the VAR to the estimation of a single equation at a time, without ignoring the correlation of error terms. Second, for each single equation we use Bayesian Compression (Guhaniyogi and Dunson, 2015) to select the best set of predictors out of a universe of a potentially huge number of candidates. Bayesian compression attempts to estimate interesting linear combinations of the regressors which are promising in terms of fit by using randomly drawn weights for the linear combinations. The best linear combination is then selected using closed-form expressions for the marginal likelihood or “evidence”. The final model is a Bayesian VAR whose theoretical properties are not sacrificed and full information likelihood-based methods are used to address the forecasting problem. Our reduction of a potentially huge VAR to a series of univariate autoregressions (without sacrificing theoretical consistency with the full-blown VAR) along with compression, allows us to obtain good models at a fraction of the time that would be normally required to estimate medium-scale Bayesian VAR for each tract.

Additionally, recognizing that linear combinations are not necessarily ideal in our context, we augment the original set of predictors (lagged prices and occupancy rate from hotels within the same tract) with basis function used in neural networks (known as sigmoid function) whose parameters are also drawn randomly avoiding expensive global optimization or back propagation to select optimal weights. Artificial neural networks (ANN) can approximate arbitrarily well any given functional form (Stinchcombe and White, 1989). The marginal likelihood can still be used in the same fashion to select optimally the appropriate model.

Our techniques open up new possibilities for forecasting in tourism research. Even when researchers do not have access to contexts like ours, they can always construct huge sets of predictors using basis functions, lags, powers and other nonlinear functions of the predictors, etc. When the original problem is univariate, the problem reduces to finding the “best” predictors for a single equation / dependent variable. In the context of large VAR (i.e. multiple dependent variables) the VAR can be reduced to a series of single equations without any loss of generality. The VAR is further generalized using an ANN to a flexible form. Additionally, we also explore a fully non-parametric VAR based on Gaussian Process priors.

## 2. Statement of the problem

To state our forecasting problem, suppose we have a number of tracts and within each tract there is a number ( $n$ ) of hotels. Actually, the number of hotels varies by tract but in order to keep notation as simple as possible we use “ $n$ ”. Hotel occupancy rates are in the vector  $Y_t = [y_{1t}, \dots, y_{nt}]'$  where  $t$  is time period ( $t = 1, \dots, T$ ). For each hotel, we model occupancy rate as a

---

<sup>1</sup>An alternative is to impose some structural relationship across different dimensions. In this specific case, the distance between hotels can be structured. See for example, Pfeifer and Bodily (1990).

function of its own price as well as the prices of other hotels within the tract. Hotel price is an important component of guest satisfaction (Matilla and O'Neill, 2003), and is also an important criteria in the selection of hotel/motel accommodation (Lockyer, 2005).

As we will see below, we have 10 years of monthly data (i.e.  $T = 120$ ) for each hotel. Of course, it is possible to forecast hotel occupancy rates based on univariate models like the following, and we which we use later for comparison:

## 2.1. Univariate Model I

$$y_{it} = x'_{it}\alpha + \sum_{l=1}^L \beta_l y_{i,t-l} + \varepsilon_{it}, i = 1, \dots, n; t = 1, \dots, T, \quad (1)$$

where  $x_{it}$  is a vector of predictors (e.g. lagged prices),  $\alpha$ ,  $\beta$  are coefficients and  $\varepsilon_{it}$  is an error term. Hotels in the same tract are dependent in the sense that prices or price changes are correlated as they compete in the same market. Therefore, an alternative model could be:

## 2.2. Univariate Model II

$$y_{it} = x'_{it}\alpha + \sum_{l=1}^L \beta_l y_{i,t-l} + \sum_{l=1}^L \gamma_l \bar{y}_{t-l} + \varepsilon_{it}, i = 1, \dots, n; t = 1, \dots, T, \quad (2)$$

where  $\bar{y}_t$  is the average value of the dependent variable in the tract. For simplicity, we keep the number of lags ( $L$ ) to be the same. Such models can be estimated without much difficulty and their forecasting performance can be examined systematically.

Although such models<sup>2</sup> account for dependence within a tract, one can argue that dependence is deeper, and guests look at the prices of other hotels within the tract when deciding on a particular hotel. A model that accounts for such dependence is a VAR:

$$y_{it} = x'_{it}\alpha + \sum_{j=1}^n \sum_{l=1}^L \beta_{jl} y_{ij,t-l} + \varepsilon_{it}, i = 1, \dots, n; t = 1, \dots, T, \quad (3)$$

If we abstract from the presence of exogenous variables and restrict attention to  $L=1$ , a full VAR can be written as follows:

$$Y_t = BY_{t-1} + \varepsilon_t, t = 1, \dots, T, \quad (4)$$

---

<sup>2</sup> For all univariate models the number of lags is chosen using the Schwarz Bayesian Information criterion. We estimated for each tract individually. We have tried ARIMA but the forecast gains were trivial. We did not try vector ARMA because of computational difficulties. This means that MA terms do not add significantly to forecast performance, at least in this application. Given that the average of y's is included in each equation (except unit / tract i which we forecast) this means that common factors are accounted for, so it extremely unlikely that any forecast improvement will arise from vector ARMA processes as they only account for error correlation..

where  $B$  is  $n \times n$  matrix of unknown coefficients. For simplicity, we also omit the vector of intercepts. Here,  $\varepsilon_t = [\varepsilon_{1t}, \dots, \varepsilon_{nt}]'$  is a vector of error terms whose mean is zero and their covariance is  $\Omega$ .

Even in this simplified form, it is clear that if we have, for example,  $n=30$  hotels per tract, matrix  $B$  involves 900 unknown parameters and  $\Omega$  involves another 465 parameters.. With  $T = 120$  observations it is impossible to estimate these parameters<sup>3</sup>.

Hence, the important question is how to proceed when faced with such a large VAR. As mentioned, methods like principal components can be used as well. However, some new techniques have been proposed in the statistical literature based on the idea of ‘‘compression’’ (Guhaniyogi and Dunson, 2015). The potential number of explanatory variables is reduced to a much smaller number (determined by the data) automatically. We discuss the idea in the next section, and for discussion purposes, we confine our attention first to the case of the linear model.

### 3. The case of the linear model

Suppose we have a model of the form:

$$y_t = x_t' \beta_o + v_t, v_t \sim iidN(0, \sigma_v^2), t = 1, \dots, T, \quad (5)$$

where  $x_t$  is a  $k \times 1$  vector of predictors whose dimension can be large or even larger than  $T$ . Guhaniyogi and Dunson (2015) advocated compression methods in which the vector of regressors is replaced by  $\Phi x_t$  where  $\Phi$  is an  $m \times k$  compression matrix which, effectively, reduces the dimensionality to  $m$ . In turn, the regression model is:

$$y_t = (\Phi x_t)' \beta + u_t, u_t \sim iidN(0, \sigma^2), t = 1, \dots, T, \quad (6)$$

where  $\beta$  is the  $m \times 1$  parameter vector. Guhaniyogi and Dunson (2015) suggest to draw the elements of matrix  $\Phi$  randomly as follows:

$$\begin{aligned} Pr(\Phi_{ij} = \frac{1}{\sqrt{\varphi}}) &= \varphi^2, \\ Pr(\Phi_{ij} = 0) &= 2\varphi(1-\varphi), \\ Pr(\Phi_{ij} = -\frac{1}{\sqrt{\varphi}}) &= (1-\varphi)^2, \end{aligned} \quad (7)$$

---

<sup>3</sup> The VAR model can be estimated on a single-equation basis for each hotel (this is known to provide the maximum likelihood estimator under normality) - the matrix of regressors is the same for each hotel / equation. Bayesian methods can be used which impose prior information on the elements of  $B$ . For example, the well-known Minnesota prior assumes that the series are random walks and therefore the diagonal elements of  $B$  should be close to one while the off-diagonal elements should be close to zero. The question of estimating or imposing a prior on  $\Omega$  is left open or is dealt with using other techniques. Often, a reasonable alternative is to use a conditional Wishart prior so that  $\Omega$  can be directly integrated out of the posterior and then numerical techniques are used to explore the marginal posterior of  $B$ . How useful are such techniques in terms of forecasting is an empirical question. Many studies have been conducted to examine the effect of the prior in out-of-sample forecasting (Feldkircher, 2012; Assaf et al. 2018).

where  $\varphi$  and  $m$  are parameters. Matrix  $\Phi$  is orthonormalized before proceeding. Parameter  $m$  can be drawn from a uniform distribution in  $[2 \log k, \min(T, k)]$  and  $\varphi$  from a uniform distribution in  $(0.1, 0.9)$ . In turn, different draws for  $m$  and  $\Phi$  can be considered and we select the one with the highest values of the marginal likelihood<sup>4</sup>. In a simulation study and empirical work, Guhaniyogi and Dunson (2015) documented very good coverage properties of predictive intervals and large computational savings relative to other computational alternatives to the problem. In matrix notation, the problem  $y = X\beta_o + v$  is converted to  $y = X\Phi'\beta + u = \tilde{X}\beta + u$  and the LS estimate is  $\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y$ .

If we write the model as:

$$y_i = (\Phi x_i)' \beta + v_i, v_i \sim iidN(0, \sigma^2), i = 1, \dots, n$$

the marginal likelihood is available in closed form as follows:

$$M(y) = \frac{2^{n/2} \Gamma(n/2)}{|X\Phi'\Sigma_\beta\Phi X' + I|^{1/2} \left[ y' (X\Phi'\Sigma_\beta\Phi X' + I)^{-1} y \right]^{n/2} (2\pi)^{n/2}},$$

provided our prior is:  $\beta | \sigma \sim N(0, \sigma^2 \Sigma_\beta)$  and  $p(\sigma) \propto \sigma^{-1}$ . Moreover,

$$(X\Phi'\Sigma_\beta\Phi X' + I)^{-1} = I - X\Phi'(\Sigma_\beta^{-1} + \Phi X X \Phi')^{-1}\Phi X'.$$

Our interest in this study focuses on a different, more flexible model, an artificial neural network (ANN) of the form:

$$y_t = x_t' \beta_o + \sum_{g=1}^G \gamma_g \psi(x_t' \beta_g) + u_t, u_t \sim iidN(0, \sigma^2), t = 1, \dots, T$$

In this paper we focus on ANNs with a single layer, although “deep learning” techniques (relying on more layers) can be examined in future work. Such models also known as artificial intelligence methods and have been used widely in tourism forecasting as we mentioned in the Introduction.

<sup>4</sup> It is possible to average different models based on their posterior model probabilities which can be computed easily once we have the marginal likelihood. This is the method advocated in Guhaniyogi and Dunson (2015). Following the suggestion of an anonymous referee, we re-examined the forecasting problem but found that posterior model probabilities favour one model by a large margin (posterior model probability for one model was always above 0.9 and the other models received negligible probability). In turn, this suggests that, at least in our application, we can rely on model selection rather than model comparison or model averaging.

In equation (8),  $\beta_o$  and  $\beta_g$  are  $k \times 1$  parameter vectors,  $\gamma_1, \dots, \gamma_G$  are parameters,  $\psi$  is a given ‘‘activation function’’, for example  $\psi(z) = \frac{1}{1 + \exp(-z)}$ ,  $G$  is the number of nodes in the ANN. It is well known that as  $G$  increases, the model can approximate accurately any given functional form.

The problem is, of course, that with large  $k$ , the ANN is even more demanding in terms of data requirements. Therefore, we have to consider compression methods. There are two alternative compression techniques that we can consider in this context. **In Compression Method I, we compress the data prior to estimating and forecasting using an ANN. In Compression Method II, we apply a simpler procedure to select the nonlinear parameters of the ANN instead of full MCMC-based Bayesian inference. The method relies on the idea that the support of the nonlinear parameters consists of only three values, one of which is zero and the other two are placed symmetrically around zero. As this alternative reduces computational burden significantly, it is worth examining whether it performs well in terms of out-sample-forecasting relative to Compression Method I.**

### 3.1. Compression Method I

**In Compression Method I we compress the data in  $x_t$  to reduce the dimensionality, and after we apply the ANN. The data compression is achieved by using the Guhaniyogi-Dunson operator  $\Phi x_t$ . In turn, the number of variables is significantly reduced and the procedure is the same as in (6).**

$$y_t = (\Phi x_t)' \beta_o + \sum_{g=1}^G \gamma_g \psi((\Phi x_t)' \beta_g) + u_t, u_t \sim iidN(0, \sigma^2), t = 1, \dots, T, \quad (12)$$

The model is nonlinear in the parameters  $\tilde{\beta} = \{\beta_g, g = 1, \dots, G\}$ . Statistical inferences can be provided using MCMC algorithms, for example the Metropolis-Hastings or the Girolami and Calderhead (2011) technique. In this study we use a simple Metropolis-Hastings random-walk algorithm.<sup>5</sup>

### 3.2. Compression Method II

In Compression Method II we condition, first, on the nonlinear parameters  $\beta = [\beta_g', g = 1, \dots, G]'$ . Define  $\psi((\Phi x_t)' \beta_g) = z_{t,g}(\tilde{\beta})$  so that we have the model:

$$y_t = x_t' \beta_o + z_t'(\tilde{\beta}) \gamma + u_t, u_t \sim iidN(0, \sigma^2), t = 1, \dots, T, \quad (13)$$

where  $\gamma = [\gamma_1, \dots, \gamma_G]'$  and  $z_t(\tilde{\beta}) = [z_{t,1}(\tilde{\beta}), \dots, z_{t,G}(\tilde{\beta})]'$ . We can write the model as:

---

<sup>5</sup>The number of nodes ‘‘ $G$ ’’ is chosen using the marginal likelihood for each model (corresponding to different  $G$  from 1 to 10).

$$y_t = [x'_t, z'_t(\tilde{\beta})]' \delta + u_t = w'_t(\tilde{\beta}) \delta + u_t, t = 1, \dots, T, \quad (14)$$

where  $\delta = [\beta'_o, \gamma']'$  and  $w_t(\tilde{\beta}) = [x'_t, z'_t(\tilde{\beta})]'$  whose dimensionality is  $(k + G) \times 1$ . Conditional on  $\tilde{\beta}$ , we can apply compression directly to this model. To choose the  $\beta$  s we want to impose some sparsity and avoid proliferation, so we can draw randomly each  $\beta_j$  from a discrete distribution with support  $\{-\underline{b}, 0, \underline{b}\}$  with equal probabilities (1/3), where  $\underline{b}$  can be selected after some experimentation: The experimentation consists in treating the  $\beta$  s as parameters supported over continuous sets, apply MCMC to update them and then choose  $\underline{b}$  based on the MCMC samples to minimize a rough measure of predictive ability. Specifically, we try 1,000 different randomly selected values of  $\underline{b}$ , and we choose the one that provides minimum absolute forecast error at a 12-month horizon.. **To the best of our knowledge, this is a novel technique. Since the  $\beta$  s are nonlinear parameters and direct optimization to select them is quite costly, it was encouraging to see that this method works quite well in practice. The method can be extended in future work to allow for more points in the support. Intuitively, this method works because for the nonlinear parameters of an ANN, precisely knowledge is immaterial; what is more important is that we have a sufficient number of nodes and rough approximations for the values of the nonlinear parameters.**

#### 4. The case of Vector Autoregressions

VARs have been considered, from the point of view of compression, by Koop et al. (2016). The VAR model is:

$$Y_t = BY_{t-1} + \varepsilon_t, t = 1, \dots, T, \quad (15)$$

where  $Y_t = [y_{1t}, \dots, y_{nt}]'$  is a vector of variables,  $B$  is an  $n \times n$  matrix of coefficients, and for the error terms we assume:

$$\varepsilon_t \sim iidN(0, \Omega). \quad (16)$$

As the number of parameters is  $n^2$  (provided we have only lags of order one but the extension to more lags is straightforward) this can quickly exceed the sample size, so the need for compression is immediate. Our extension to the model is:

$$y_{it} = x'_t \beta_{oi} + \sum_{g=1}^G \gamma_{ig} \psi(x'_t \beta_{ig}) + \varepsilon_{it}, i = 1, \dots, n, t = 1, \dots, T, \quad (17)$$

where  $x_{it} = [y_{1,t-1}, \dots, y_{n,t-1}]'$ . Let  $z_{it}(\tilde{\beta}) = [x'_t, \psi(x'_t \beta_{i1}), \dots, \psi(x'_t \beta_{iG})]'$  be the  $(n + G) \times 1$  vector of regressors in the  $i$ th equation of the VAR, and  $\tilde{\beta}_i = [\beta'_{i1}, \dots, \beta'_{iG}]'$  is the vector of nonlinear parameters whose dimensionality is  $nG \times 1$ . In turn, we can write the  $i$ th equation of the VAR as follows:

$$y_{it} = z_{it}(\tilde{\beta})' \delta_i + \varepsilon_{it}, \quad (18)$$

where  $\delta_i = [\beta'_{oi}, \{\gamma_{ig}, g = 1, \dots, G\}]'$  is the  $(n + G) \times 1$  vector of parameters. This model is a

multivariate regression of the form<sup>6</sup>:

$$Y_t = Z_t(\tilde{\beta})\delta + \varepsilon_t, \quad (19)$$

where  $Z_t(\tilde{\beta})$  has  $z_{it}(\tilde{\beta})'$  as its  $i$ th row. We can apply any of the two compression methods in the previous section, as we can i) first, compress the  $x_t$ s in (14) or ii) treat the nonlinear parameters  $\beta_{ig}$  as given for the moment and compress directly in (15). In this form we can compress as:

$$Y_t = \left( Z_t(\tilde{\beta})\Phi \right) \delta + \varepsilon_t, \quad (20)$$

where  $\Phi$  is a  $k \times m$  compression matrix. In this form we still have to deal with i) the nonlinear parameters in  $\tilde{\beta}$ , and ii) a large covariance matrix  $\Omega$ .

Koop et al. (2016) exploit results from Primiceri (2005), Eisenstat et al. (2016) and Carriero et al. (2015) to express the VAR in a recursive form with independent errors facilitating equation-by-equation analysis. The representation for the standard VAR in (12) has the form:

$$Y_t = \Theta X_t + \Sigma E_t, \quad (21)$$

where  $A\Omega A' = \Sigma$ ,  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ ,  $A$  is a lower triangular matrix with ones on the main diagonal,  $X_t = [Y'_{t-1}, -Y'_t]'$ ,  $\Theta = [\Gamma, A - I_n]$ , and  $E_t \sim N(0, I_n)$ . In this representation, the first equation includes as regressors  $Y_{t-1}$ , the second includes  $(Y'_{t-1}, -y_{1t})$ , the third includes  $(Y'_{t-1}, -y_{1t}, -y_{2t})$  etc. Therefore the  $i$ th equation has the form:

$$y_{it} = \theta'_i (\Phi_i X_t^{(i)}) + \sigma_i E_{it}, \quad (22)$$

where  $X_t^{(i)}$  contains the regressors in the  $i$ th equation (as we described),  $\Phi_i$  is a compression matrix whose dimensionality is  $m \times \text{dim}(X_t^{(i)})$  and the  $i$ th row of  $\Theta$  is given by  $\theta'_i$ . This form suggests that the model can be made flexible through the introduction of ANNs as in the previous section (the case of the simple linear model). We call this the ‘‘Compressed VAR-ANN’’ or Compressed VAR (CVAR) in short.

## 5. Alternative Models

As an extension to the above, we also consider two alternative models. In the first, we consider a CVAR model using a non-parametric form. We label the model as a Gaussian Process VARs (GPVAR), where the implementation of GPVAR is based on the existing compression from linear VAR (so we maintain the same compression as in CVAR) and then the compressed variables are used in GPVAR. In other words, we maintain the advantage of single-equation-based estimation through non-parametric, Gaussian process regression, to implement our GPVAR. The non-parametric context has some obvious advantages. For example, we make no assumption on the functional form and thus we let the data speak for themselves in terms of in-sample fit. How useful

---

<sup>6</sup> In this multivariate regression unlike standard VARs the regressors are not the same in each equation.

they can be in terms of forecasting, is an empirical question. A priori, however, we do not expect them to perform well as non-parametric models tend to adapt too much to the idiosyncrasies of the sample and, as a result, may not always perform well in terms of out-of-sample forecasting.

In the second model, we consider a dynamic factor method, which also emerges naturally in the big-data context and has been used in previous work (Forni et al. 2000; Forni et al. 2015; Koop et al. 2016). From here, we believe it is an appropriate choice to assess the forecasting performance of Bayesian compressed methods.

## 5.1. Gaussian Process VARs

Once we have the single-equation representation of the VAR as in (19) it is possible to use a non-parametric version and write (19) as:

$$y_{it} = f_i(x_t^{(i)}) + \varepsilon_{it}, i = 1, \dots, n, \quad (23)$$

where  $\varepsilon_{it} = \sigma_i E_{it}$ , and  $x_t^{(i)} = \Phi_i X_t^{(i)}$ . In this formulation, we use a fully non - parametric version of each VAR equation instead of a flexible ANN representation. This formulation is novel in the sense that the regressors in the  $i$  th equation of the VAR are compressed before non-parametric regression. In turn, we use the Gaussian Process Prior (GP) formalism to implement non-parametric estimation (see O' Hagan and Kingman, 1978; Williams and Rasmussen, 1996, and Williams and Rasmussen, 2006). In the GP we assume:

$$f_i(x) \sim \text{GP}(0, K_i(\cdot, \cdot)), \quad (24)$$

where  $K_i(x, x)$  represents the covariance matrix of  $f_i(x)$  at different pairs of points, say  $(x, z)$ . Suppose  $f_i = [x_{i1}, \dots, x_{iT}]'$ . Then we have the representation:

$$\begin{aligned} y_i | f_i &\sim N(f_i, \sigma_i^2 I), \\ f_i &\sim N(0, K_i), \end{aligned} \quad (25)$$

where  $y_i$  and  $f_i$  are  $T \times 1$  vectors,  $f_i$  denotes the vector of unknown function values at the  $x_t^{(i)}$  s, and  $K$  is a  $T \times T$  matrix representing the covariance of the GP given by  $K_i(x, z) = E[(f_i(x) - f_i(z))(f_i(x) - f_i(z))']$ . The covariance matrix can be written as  $K_i(x_t, x_s) = \kappa_{\eta_i}(x_t - x_s)$  where  $\kappa_{\eta_i}(d) = \exp\{-d'd/2\eta_i^2\}$  is a kernel function and  $\eta_i$  is a hyperparameter. Often, an alternative kernel is used, viz.  $\kappa_{\eta_i}(d) = \eta_{i1} \exp\{-d'd/2\eta_{i2}^2\}$  where  $\eta_{i1}$  controls the variance and  $\eta_{i2}$  controls the decay of covariances.

Often, we want to evaluate the unknown function at a number of test points  $x_i^* = [x_{i,(1)}^*, \dots, x_{i,(M)}^*]'$  at which the function values are  $f_i^* = [f_{i,(1)}^*, \dots, f_{i,(M)}^*]'$ . Under the GP prior, the joint distribution of  $f_i$  and  $f_i^*$  is:

$$\begin{bmatrix} f_i \\ f_i^* \end{bmatrix} \sim \mathbf{N}\left(0, \begin{bmatrix} K_i(X, X) & K_i(X, X^*) \\ K_i(X^*, X) & K_i(X^*, X^*) \end{bmatrix}\right), \quad (26)$$

where  $X, X^*$  represent the vectors of observed and evaluation (test) points. Defining  $A_i = K_i(X, X) + \sigma_i^2 I_T$  we obtain the posterior mean and covariance matrix:

$$E(f_i^* | y_i) = K_i(X^*, X)A_i^{-1}y_i, \quad (27)$$

$$\mathbf{cov}(f_i^* | y_i) = K_i(X^*, X^*) - K_i(X^*, X)A_i^{-1}K_i(X, X^*). \quad (28)$$

These expressions can be derived if we notice that:

$$p(f_i^* | y_i) = \int p(f_i, f_i^* | y_i) df_i = \frac{1}{p(y_i)} \int p(y_i | f_i) p(f_i, f_i^*) df_i. \quad (29)$$

Finally, the hyperparameter(s)  $\eta_i$  can be chosen so that the marginal likelihood is maximized. The analytical expression for the log marginal likelihood is:

$$\log p(y_i | X_i, \eta_i) = -\frac{1}{2} y_i^T [K_{\eta_i}(X_i, X_i) + \sigma_i^2 I]^{-1} y_i - \frac{1}{2} \log |K_{\eta_i}(X_i, X_i) + \sigma_i^2 I| - \frac{T}{2} \log 2\pi. \quad (30)$$

For more details, see MacKay (1999), Williams and Rasmussen (2006), and Sundararajan and Keerthi (2001). The only expensive computation is the inversion of the  $A_i$  matrix (which depends on the hyperparameters). However, efficient methods have been proposed by Williams and Rasmussen (2006).

## 5.2. Dynamic Factor Model (DFM)

DFMs are popular in the big-data context and they have, more or less, the same interpretation as principal components. Principal components are linear combination of the variables which explain as much of the variation in the variables, as possible. Suppose  $y_t$  is an  $n \times 1$  vector of variables that we need to “explain” in terms of  $m \ll n$  factors, which we denote by  $f_t$ . Then we assume:

$$y_t = \Lambda f_t + u_t \quad (31)$$

where  $u_t$  is a vector error term with zero mean and covariance matrix  $\Sigma$ , and  $\Lambda = [\lambda_{ij}]$  is a matrix of parameters known as factor loadings. There are several possible normalizations that can be used for identification, for example  $\lambda_{ij} = 0, i < j, \lambda_{ii} = 1$ . In distinction to classical principal components or factor analysis, the factors have a dynamic structure:

$$f_t = A f_{t-1} + e_t \quad (32)$$

where  $e_t \sim N(0, \Omega)$ ,  $A$  is a square matrix ( $m \times m$ ) and  $\Omega$  is a diagonal matrix. The model is not unlike the factor analysis model and, in fact, some estimation techniques have been proposed which rely heavily on the static factor model. In terms of advantages, the DFM allows for serial dependence in the factors which is of interest in applied forecasting. For more details, see Forni

et al. (2000).

## 6. Data

The dataset for this study was obtained from Smith Travel Research (STR), an independent company that tracks lodging supply and demand data for most major hotels in the United States. We obtained data on occupancy rates and average daily rates (as a proxy of hotel price) for 1,508 hotel located in 50 different tracts. For each hotel, we have 10 years of monthly data ranging from the first month of 2008 to the last month of 2017. In total, we have 180,960 observations. Table 1 summarizes the number of hotels for each of these tracts. According to STR, a tract is defined as a “subset of a Metropolitan Statistical Area market and are defined by counties and/or zip codes. For example, the Boston MSA is divided into 10 tracts” (Canina et al. 2005, p. 571).

Hotels within the same tract naturally compete with each other as the tract represents the narrowest grouping available and is “the realistic options available to a consumer who desires to visit a particular location” (Canina et al. 2005, p. 571). For this reason, tracts have been used in some studies to represent “competitive clusters” (e.g. Assaf et al. 2016; Canina et al. 2005).

As mentioned, we model the occupancy rate for each hotel as a function of its own price (i.e. ADR) as well as the prices and occupancy rates of other hotels within the tract. Given that tracts represent a realistic option available to consumers, we believe that the occupancy rate of a certain hotel is affected by the price and occupancy rate of other hotels within the same tract. As discussed above, hotel price has always been an important criterion in the selection of hotel/motel accommodation (Lockyer, 2005).

## 7. Results

### 7.1. Forecasting

In Table 2 we report mean absolute forecast errors (MAFE) for all models we presented, under the baseline prior. Root Mean Squared Forecast Errors (RMSE) are reported in Table 3. Specifically, we present average MAFE across all tracts, and results for three tracts selected at random. The results are for the baseline prior (see subsection 7.2). Evidently, the best forecasts in terms of both MAFE and RMSE are delivered by CVAR-I and CVAR-II implying that compression provides accurate forecasts even at horizons of 36 months ahead.

Although, at some short horizons, GPVAR performs well, its behavior is not consistent across tracts and / or forecast horizons. Univariate and Dynamic Factor Models (DFM) do not perform nearly as well as compressed VARs, and the forecast errors are, invariably, quite large relative to the compression methods. This means that compression and (dynamic) factor analysis rely on different principles of selecting summaries of the data and, in practice, using compression delivers much better forecasts. Surprisingly, univariate models are not adequate in out-of-sample forecasts. This often happens and it is well known that more elaborate models (VARs for example), in particular Bayesian VARs sometimes perform better (Wong et al. 2006). In our application, it turns out that univariate models do not perform nearly as well as compression methods.

A formal criterion of selecting the best-performing model is the predictive Bayes factor (PBF). Loosely speaking, given the data  $Y_{1:T}$  where  $T$  is the available number of observations, we consider the predictive density:  $p(Y_{T+1:T+h} | Y_{1:T})$ , where  $h = 1, 2, \dots, H$  is the forecast horizon.

$$\begin{aligned}
p(Y_{T+1:T+h}|Y_{1:T}) &= \int_{\Theta} p(Y_{T+1:T+h}, \theta | Y_{1:T}) d\theta \\
&= \int_{\Theta} p(Y_{T+1:T+h} | \theta, Y_{1:T}) p(\theta | Y_{1:T}) d\theta.
\end{aligned} \tag{33}$$

Here,  $\theta \in \Theta$  represents the parameter vector. Given an ergodic sequence,  $\{\theta^{(s)}, s=1, \dots, S\}$ , which converges to the posterior,  $p(\theta | Y_{1:T})$ , the multivariate integral can be accurately approximated as:

$$p(Y_{T+1:T+h}|Y_{1:T}) \approx S^{-1} \sum_{s=1}^S p(Y_{T+1:T+h} | \theta^{(s)}, Y_{1:T}) \tag{34}$$

The ratio of any two predictive densities provides the predictive Bayes factor. PBFs are reported in Table 4. PBFs are computed in favor of CVAR-I and against all other models. Although PBFs are close to unity for CVAR-II (implying that the two models behave nearly the same) they are strongly in favor of CVAR-I (and therefore, CVAR-II as well) against all other models. For example, the odds against DFM at 12 months are overwhelming (292.15) and the same is true for most of the other models.

To evaluate the “statistical significance” of forecast superiority we rely on the Hansen, Lunde and Nason (2011) model confidence set (MCS)  $p$ -values. The MCS consists of models that are not clearly inferior in the light of the data. As these authors mention: “The interpretation of a MCS  $p$ -value is analogous to that of a classical  $p$ -value. The analogy is to a  $(1 - \alpha)$  confidence interval that contains the “true” parameter with a probability no less than  $1 - \alpha$ . The MCS  $p$ -value also cannot be interpreted as the probability that a particular model is the best model, exactly as a classical  $p$ -value is not the probability that the null hypothesis is true. Rather, the probability interpretation of a MCS  $p$ -value is tied to the random nature of the MCS because the MCS is a random subset of models that contains [the true model] with a certain probability.” (p. 463)

In Table 5, we report median  $p$ -values of the MCS test across all tracts. From the results reported in Table 5, it turns out that CVAR I and II are always in the MCS across all forecasting horizons and other models are clearly inferior in terms of forecasting performance. For the models in the MCS (viz. CVAR I and II) the  $p$ -values were never lower than 0.10. For the models that are not in the MCS they were never greater than 0.01.

## 7.2. Sensitivity Analysis

Sensitivity to prior assumptions is an issue in Bayesian analysis and, in principle, different priors may lead to different forecasting results. We address this issue in what follows. First, we describe the baseline prior. In (1), we have parameters  $\alpha, \beta$ . In (2) we have parameters  $\alpha, \beta, \gamma$ . In (8) we have parameters  $\gamma, \beta_0, \{\beta_g\}$ . In (30) we have the hyperparameters  $\eta_i = \exp(\xi_i)$ . In (31) – (32) we have parameters  $A$  and the different elements of  $\Lambda$ . The reparametrization from  $\eta_i$  to  $\xi_i$  facilitates the use of a normal prior for all parameters. All parameters are assumed *a priori* independent and for any parameter, say  $\lambda$  we assume:

$$\lambda \sim N(0, \omega^2 I) \tag{35}$$

In the baseline prior we have  $\omega = 1$ .. To examine alternative priors we vary parameter  $\omega$  between 0.01 and 100 randomly and, therefore, we generate 100 such priors. The reduction of the prior to

a single parameter ( $\omega$ ) facilitates exposition in the absence of specific information about the parameters.

Our results are reported in Table 6, where we present the percentage of cases in which CVAR-I delivers smaller RMSFE and MAFE relative to other models across all forecast horizons. Compared to CVAR-II, CVAR-I behaves about the same. Relative to other models, in particular Univariate-II and DFM, are occasionally better but in more than 90% of all cases, CVAR-I performs best. Whenever other models are better, a closer examination showed that MAFE and RMSE do not differ by more than 5%, showing that compression techniques are still competitive even when univariate or DF models are better.

## 8. Concluding remarks

In this paper we take up the issue of forecasting occupancy rates of hotels that compete within the same tract. Apart from univariate models, we propose the use of two novel nonlinear VARs based on the idea of compressing the regressors in each equation to avoid the curse of dimensionality. The VARs are extended to a flexible framework by using neural networks to specify the functional forms. In addition, we examined the forecasting performance of a nonparametric specification based on Gaussian processes. We also use dynamic factor models as an alternative method to summarize the data when the number of regressors is large. The empirical results show that the performance of the new compressed VARs is impressive, relative to the alternative models, and their accuracy is preserved across nearly all forecast horizons from 1 to 36 months.

One limitation of the study is that it does not account for other factors that may affect the occupancy rate of hotels within the tract. For instance, occupancy rate can be affected by seasonality, number of tourist arrivals, star rating, gross domestic product and consumer price index, among others. Unfortunately, obtaining some of these data at a consistent level for all tracks was not possible for this study. Future studies are encouraged to account for some of these variables. Nevertheless, the present study presents an important illustration on the power of compressed methods should more data become even more available.

Given the superiority of Bayesian compressed methods, future applications of the techniques are recommended. For example, future research may consider applying the methods to study other aspects of the tourism industry. The techniques, for instance, may open the door for more comprehensive applications that would account for a large number of variables to forecast important issues such as tourism demand and arrivals. It would be also interesting to compare the methods to other data reduction techniques currently used in the tourism literature (e.g. principal components).

Table 1. Sample Characteristics

| Tract #  | No. of Hotels | No of Obs. | Tract #  | No. Hotels | No. of Obs. |
|----------|---------------|------------|----------|------------|-------------|
| Tract 1  | 18            | 2160       | Tract 26 | 29         | 3480        |
| Tract 2  | 35            | 4200       | Tract 27 | 29         | 3480        |
| Tract 3  | 45            | 5400       | Tract 28 | 25         | 3000        |
| Tract 4  | 40            | 4800       | Tract 29 | 20         | 2400        |
| Tract 5  | 17            | 2040       | Tract 30 | 26         | 3120        |
| Tract 6  | 36            | 4320       | Tract 31 | 39         | 4680        |
| Tract 7  | 20            | 2400       | Tract 32 | 43         | 5160        |
| Tract 8  | 26            | 3120       | Tract 33 | 41         | 4920        |
| Tract 9  | 26            | 3120       | Tract 34 | 16         | 1920        |
| Tract 10 | 26            | 3120       | Tract 35 | 24         | 2880        |
| Tract 11 | 37            | 4440       | Tract 36 | 41         | 4920        |
| Tract 12 | 43            | 5160       | Tract 37 | 35         | 4200        |
| Tract 13 | 21            | 2520       | Tract 38 | 20         | 2400        |
| Tract 14 | 23            | 2760       | Tract 39 | 20         | 2400        |
| Tract 15 | 32            | 3840       | Tract 40 | 42         | 5040        |
| Tract 16 | 27            | 3240       | Tract 41 | 26         | 3120        |
| Tract 17 | 29            | 3480       | Tract 42 | 32         | 3840        |
| Tract 18 | 52            | 6240       | Tract 43 | 32         | 3840        |
| Tract 19 | 26            | 3120       | Tract 44 | 44         | 5280        |
| Tract 20 | 25            | 3000       | Tract 45 | 29         | 3480        |
| Tract 21 | 27            | 3240       | Tract 46 | 26         | 3120        |
| Tract 22 | 22            | 2640       | Tract 47 | 31         | 3720        |
| Tract 23 | 24            | 2880       | Tract 48 | 42         | 5040        |
| Tract 24 | 32            | 3840       | Tract 49 | 28         | 3360        |
| Tract 25 | 26            | 3120       | Tract 50 | 33         | 3960        |

Table 2. Forecasting Performance: Mean Absolute Forecast Errors (%)

|                           |               | 1 month | 3 months | 6 months | 12 months | 24 months | 36 months |
|---------------------------|---------------|---------|----------|----------|-----------|-----------|-----------|
| Average across all tracts | Univariate I  | 8.35    | 10.24    | 11.71    | 12.88     | 17.01     | 20.37     |
|                           | Univariate II | 5.52    | 7.12     | 9.77     | 12.81     | 14.35     | 19.13     |
|                           | CVAR I        | 3.21    | 3.48     | 3.90     | 4.21      | 4.30      | 4.34      |
|                           | CVAR II       | 2.89    | 3.12     | 3.25     | 3.90      | 4.10      | 4.17      |
|                           | GPVAR         | 2.77    | 3.22     | 3.71     | 4.10      | 4.47      | 4.55      |
|                           | DFM           | 7.62    | 9.44     | 10,33    | 11.45     | 13.17     | 15.05     |
| Random tract 1            | Univariate I  | 7.79    | 9.81     | 10.22    | 11.30     | 15.24     | 19.81     |
|                           | Univariate II | 5.68    | 6.90     | 9.35     | 11.33     | 13.81     | 18.55     |
|                           | CVAR I        | 3.19    | 3.20     | 3.79     | 4.12      | 4.27      | 4.22      |
|                           | CVAR II       | 2.81    | 2.89     | 3.10     | 3.24      | 4.51      | 4.23      |
|                           | GPVAR         | 7.82    | 9.77     | 10.32    | 11.53     | 14.89     | 21.20     |
|                           | DFM           | 6.34    | 7.33     | 9.81     | 10.44     | 14.07     | 18.61     |
| Random tract 2            | Univariate I  | 7.65    | 9.90     | 10.71    | 11.61     | 16.12     | 18.45     |
|                           | Univariate II | 5.50    | 7.05     | 9.44     | 10.43     | 12.85     | 19.20     |
|                           | CVAR I        | 3.44    | 3.45     | 3.78     | 3.57      | 4.03      | 4.36      |
|                           | CVAR II       | 2.70    | 3.11     | 3.17     | 3.39      | 4.12      | 4.20      |
|                           | GPVAR         | 7.55    | 9.81     | 10.02    | 11.79     | 14.81     | 17.10     |
|                           | DFM           | 6.32    | 8.55     | 9.33     | 10.16     | 11.39     | 19.72     |
| Random tract 3            | Univariate I  | 7.88    | 9.73     | 10.15    | 11.21     | 14.89     | 17.66     |
|                           | Univariate II | 5.50    | 6.36     | 9.10     | 12.25     | 13.77     | 18.90     |
|                           | CVAR I        | 3.01    | 3.12     | 3.89     | 4.33      | 4.58      | 4.71      |
|                           | CVAR II       | 2.77    | 2.93     | 3.35     | 4.02      | 4.13      | 4.20      |
|                           | GPVAR         | 2.65    | 3.15     | 3.88     | 3.97      | 4.40      | 4.66      |
|                           | DFM           | 6.57    | 8.12     | 9.81     | 10.83     | 12.67     | 17.32     |

Table 3. Forecasting Performance: Root Mean Squared Forecast Error (%)

|                           |               | 1 month | 3 months | 6 months | 12 months | 24 months | 36 months |
|---------------------------|---------------|---------|----------|----------|-----------|-----------|-----------|
| Average across all tracts | Univariate I  | 8.21    | 10.33    | 11.65    | 13.12     | 16.44     | 19.81     |
|                           | Univariate II | 5.33    | 6.89     | 9.13     | 11.45     | 14.20     | 19.30     |
|                           | CVAR I        | 3.15    | 3.47     | 3.87     | 4.10      | 4.22      | 4.55      |
|                           | CVAR II       | 2.71    | 3.20     | 3.35     | 3.86      | 4.30      | 4.81      |
|                           | GPVAR         | 2.82    | 3.15     | 3.89     | 4.40      | 4.68      | 4.91      |
|                           | DFM           | 4.51    | 8.30     | 8.71     | 10.44     | 13.16     | 17.12     |
| Random tract 1            | Univariate I  | 7.65    | 9.44     | 10.43    | 11.28     | 15.38     | 18.65     |
|                           | Univariate II | 5.58    | 6.88     | 9.13     | 11.71     | 14.30     | 18.22     |
|                           | CVAR I        | 3.17    | 3.28     | 3.75     | 4.19      | 4.35      | 4.81      |
|                           | CVAR II       | 2.85    | 2.91     | 3.15     | 3.22      | 4.47      | 4.20      |
|                           | GPVAR         | 7.77    | 9.81     | 10.45    | 11.44     | 14.35     | 21.30     |
|                           | DFM           | 5.89    | 6.51     | 9.22     | 11.35     | 14.05     | 17.89     |
| Random tract 2            | Univariate I  | 7.62    | 9.85     | 10.49    | 11.33     | 16.10     | 18.33     |
|                           | Univariate II | 5.48    | 6.91     | 9.35     | 10.40     | 12.79     | 19.33     |
|                           | CVAR I        | 3.42    | 3.55     | 3.90     | 4.02      | 4.15      | 4.31      |
|                           | CVAR II       | 2.67    | 3.10     | 3.12     | 3.22      | 4.57      | 4.30      |
|                           | GPVAR         | 7.49    | 9.74     | 10.21    | 11.35     | 14.71     | 16.30     |
|                           | DFM           | 5.41    | 6.72     | 9.89     | 10.81     | 15.32     | 18.10     |
| Random tract 3            | Univariate I  | 7.80    | 9.63     | 10.19    | 11.35     | 14.77     | 17.81     |
|                           | Univariate II | 5.32    | 6.55     | 9.31     | 12.12     | 13.65     | 19.12     |
|                           | CVAR I        | 3.13    | 3.65     | 3.92     | 4.30      | 4.51      | 4.65      |
|                           | CVAR II       | 2.72    | 2.85     | 3.30     | 4.17      | 4.21      | 4.35      |
|                           | GPVAR         | 2.60    | 3.09     | 3.71     | 4.09      | 4.45      | 4.60      |
|                           | DFM           | 5.10    | 6.14     | 8.86     | 11.81     | 13.32     | 17.40     |

Table 4. Predictive Bayes factors in favor of CVAR I and against stated model

|               | 1 month | 3 months | 6 months | 12 months | 24 months | 36 months |
|---------------|---------|----------|----------|-----------|-----------|-----------|
| Univariate I  | 85.32   | 91.62    | 144.30   | 181.552   | 281.14    | 315.87    |
| Univariate II | 115.44  | 122.35   | 189.30   | 293.40    | 377.14    | 415.12    |
| CVAR II       | 1.15    | 1.22     | 1.27     | 1.33      | 1.92      | 2.15      |
| GPVAR         | 17.82   | 25.34    | 36.16    | 42.55     | 71.12     | 93.71     |
| DFM           | 113.10  | 129.44   | 185.41   | 292.15    | 382.20    | 419.12    |

Table 5. MCS results

| Model         | 1 month | 3 months | 6 months | 12 months | 24 months | 36 months |
|---------------|---------|----------|----------|-----------|-----------|-----------|
| Univariate I  | 0.00    | 0.00     | 0.00     | 0.00      | 0.00      | 0.00      |
| Univariate II | 0.00    | 0.00     | 0.00     | 0.00      | 0.00      | 0.00      |
| CVAR I        | 0.25*   | 0.18*    | 0.27*    | 0.29*     | 0.32*     | 0.35*     |
| CVAR II       | 0.14*   | 0.19*    | 0.21*    | 0.23*     | 0.25*     | 0.20*     |
| GPVAR         | 0.00    | 0.006    | 0.00     | 0.00      | 0.00      | 0.00      |
| DFM           | 0.003   | 0.004    | 0.00     | 0.00      | 0.00      | 0.00      |

Notes: Reported are  $p$ -values for the MCS test. A value of 0.00 means that the  $p$ -value was less than 0.001. Forecasts are evaluated at the posterior means of parameters which were obtained from MCMC. An asterisk denotes that the model is included in the 95%-MCS. The MCS methodology is implemented using 1,000 bootstrap replications.

Table 6. Results of Prior Sensitivity Analysis

|               | Frequency of CVAR I forecasts with lower RMSFE, % | Frequency of CVAR I forecasts with lower MAFE, % |
|---------------|---------------------------------------------------|--------------------------------------------------|
| Univariate I  | 100                                               | 100                                              |
| Univariate II | 94.17                                             | 94.23                                            |
| CVAR II       | 45.32                                             | 51.12                                            |
| GPVAR         | 93.32                                             | 91.30                                            |
| DFM           | 92.14                                             | 93.12                                            |

## References

- Akal, M. (2004). Forecasting Turkey's tourism revenues by ARMAX model. *Tourism Management*, 25(5), 565-580.
- Assaf, A. G., Josiassen, A., & Oh, H. (2016). Internationalization and hotel performance: The missing pieces. *Tourism Economics*, 22(3), 572-592.
- Assaf, A. G., Li, G., Song, H., & Tsionas, M. G. (2018). Modeling and forecasting regional tourism demand using the Bayesian global vector autoregressive (BGVAR) model. *Journal of Travel Research*, in press.
- Canina, L., Enz, C. A., & Harrison, J. S. (2005). Agglomeration effects and strategic orientations: Evidence from the US lodging industry. *Academy of Management Journal*, 48(4), 565-581.
- Carriero, A., Clark, T. E., & Marcellino, M. (2015). Large vector autoregressions with asymmetric priors (No. 759). Working paper. London: School of Economics and Finance, Queen Mary University of London.
- Chan, C. K., Witt, S. F., Lee, Y. C. E., & Song, H. (2010). Tourism forecast combination using the CUSUM technique. *Tourism Management*, 31(6), 891-897.
- Eisenstat, E., Chan, J. C., & Strachan, R. W. (2016). Stochastic model specification search for time-varying parameter VARs. *Econometric Reviews*, 35(8-10), 1638-1665.
- Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics*, 82(4), 540-554.
- Forni, M., Hallin, M., Lippi, M., & Zaffaroni, P. (2015). Dynamic factor models with infinite-dimensional factor spaces: One-sided representations. *Journal of Econometrics*, 185(2), 359-371.
- George, E. I., Sun, D., & Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics*, 142(1), 553-580.
- Girolami, M., & Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2), 123-214.
- Guhaniyogi, R., & Dunson, D. B. (2015). Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512), 1500-1514.
- Gunter, U., & Önder, I. (2016). Forecasting city arrivals with Google Analytics. *Annals of Tourism Research*, 61, 199-212.
- Hansen, P. R., A. Lunde, & J. M. Nason (2011). The model confidence set. *Econometrica* 79 (2), 453-497.
- Hassani, H., Silva, E. S., Antonakakis, N., Filis, G., & Gupta, R. (2017). Forecasting accuracy evaluation of tourist arrivals. *Annals of Tourism Research*, 63, 112-127.
- Koop, G., Korobilis, D., & Pettenuzzo, D. (2016). Bayesian compressed vector autoregressions.

Working papers 103, Waltham: Brandeis University, Department of Economics and International Business School.

Li, S., Chen, T., Wang, L., & Ming, C. (2018). Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management*, 68, 116-126.

Li, G., Song, H., & Witt, S. F. (2005). Recent developments in econometric modeling and forecasting. *Journal of Travel Research*, 44(1), 82-99.

Li, S., Chen, T., Wang, L., & Ming, C. (2018). Effective tourist volume forecasting supported by PCA and improved BPNN using Baidu index. *Tourism Management*, 68, 116-126.

Lockyer, T. (2005). The perceived importance of price as one hotel selection dimension. *Tourism Management*, 26(4), 529-537.

MacKay, D. J. (1999). Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11(5), 1035-1068.

Massidda, C., & Mattana, P. (2013). A SVECM analysis of the relationship between international tourism arrivals, GDP and trade in Italy. *Journal of Travel Research*, 52(1), 93-105.

Mattila, A. S., & O'Neill, J. W. (2003). Relationships between hotel room pricing, occupancy, and guest satisfaction: A longitudinal case of a midscale hotel in the United States. *Journal of Hospitality & Tourism Research*, 27(3), 328-341.

O'Hagan, A., & Kingman, J. F. C. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-42.

Pan, B., Chenguang Wu, D., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*, 3(3), 196-210.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681-686.

Pfeifer, P. E., & S. E. Bodily (1990). A test of space-time ARMA modelling and forecasting of hotel data. *Journal of Forecasting* 9 (3), 255-272.

Primiceri, G., (2005). Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies*, 72, 821-852.

Schwartz, Z., Uysal, M., Webb, T., & Altin, M. (2016). Hotel daily occupancy forecasting with competitive sets: A recursive algorithm. *International Journal of Contemporary Hospitality Management*, 28(2), 267-285.

Stinchcombe, M. & White, H. (1989). Universal approximation using feedforward networks with non-sigmoid hidden layer activation functions. In Proceedings off the international Joint Conference on Neural Networks (pp. 1:613-618). San Diego: SOS Printing.

Sundararajan, S., & Keerthi, S. S. (2000). Predictive approaches for choosing hyperparameters in Gaussian processes. In *Advances in neural information processing systems* (pp. 631-637).

Wan, S., Song, H., & Ko, D. (2016). Density forecasting for tourism demand. *Annals of Tourism Research*, 60, 27-30.

Wong, K. K., Song, H., & Chon, K. S. (2006). Bayesian models for tourism demand

forecasting. *Tourism Management*, 27(5), 773-780.

Williams, C. K., & Rasmussen, C. E. (1996). Gaussian processes for regression. In *Advances in neural information processing systems* (pp. 514-520).

Williams, C. K., & Rasmussen, C. E. (2006). Gaussian processes for machine learning. *The MIT Press*, 2(3), 4.

Witt, S. F., Song, H., & Wanhill, S. (2004). Forecasting tourism-generated employment: The case of Denmark. *Tourism Economics*, 10(2), 167-176.

Yang, Y., Pan, B., & Song, H. (2014). Predicting hotel demand using destination marketing organization's web traffic data. *Journal of Travel Research*, 53(4), 433-447.

Zakhary, A., Atiya, A. F., El-Shishiny, H., & Gayar, N. E. (2011). Forecasting hotel arrivals and occupancy using Monte Carlo simulation. *Journal of Revenue and Pricing Management*, 10(4), 344-366.