https://doi.org/10.1016/j.annals.2018.07.010

Forecasting Turning Points in Tourism Growth

Shui Ki Wan
School of Business
Hong Kong Baptist University, Hong Kong
Email: shuiki@hkbu.edu.hk

Haiyan Song
School of Hotel and Tourism Management
The Hong Kong Polytechnic University, Hong Kong
Email: haiyan.song@polyu.edu.hk

The second author would like to acknowledge the financial support of Natural Science Foundation of China (Grant No. NSFC71673233).

Forecasting Turning Points in Tourism Growth

ABSTRACT

Tourism demand exhibits growth cycles, and it is important to forecast turning points in these growth cycles to minimise risks to destination management. This study estimates logistic models of Hong Kong tourism demand, which are then used to generate both short- and long-term forecasts of tourism growth. The performance of the models is evaluated using the quadratic probability score and hit rates. The results show that the ways in which this information is used are crucial to the models' predictive power. Further, we investigate whether combining probability forecasts can improve predictive accuracy, and find that combination approaches, especially nonlinear combination approaches, are sensitive to the quality of forecasts in the pool. In addition, model screening can improve forecasting performance.

Keywords: tourism demand, combined probability forecast, quadratic probability score, Hong Kong

1. Introduction

Tourism has shown sustainable growth over the last few decades, leading to the development of multiple tourism-related industries. Accurate forecasts of tourism demand are crucial to decision making on tourism. In a recent review of the literature, Wu, Song and Shen (2017) find that non-causal time series models, causal econometric models and artificial intelligence (AI) based models are the most commonly used approaches to forecasting tourism demand in both research and practice. For example, Hassani, Webster, Silva and Heravi (2015) develop a singular spectrum analysis framework, a non-causal time series model, for forecasting tourist arrivals; and Yang, Pan and Song (2014), among others, use panel data to extend causal models identifying key determinants of hotel demand. The rapid development of AI and machine learning has also had an impact on the tourism literature. For instance, support vector regression (Cang, 2014; Chen and Wang, 2007), artificial neural networks (Claveria, Monte and Torra, 2015) and fuzzy systems (Hadavandi, Ghanbari, Shahanaghi, and Abbasian-Naghneh, 2011) are now used to forecast tourist arrival. Although the literature is rich in methodologies for forecasting tourism demand growth, few studies predict turning points using probability forecasting. Yet the latter may be more useful than level forecasting, as emphasised in a review article by Witt and Witt (1995), because tourism demand exhibits growth cycles over time. Butler (1980) categorises each cycle into six phases: exploration, involvement, development, consolidation, stagnation and either decline or rejuvenation. The differences between these phases fundamentally alter the key factors determining business success and the effectiveness of government policies, forcing businesses and governments to think and act very differently when developing investment strategies and establishing regulations. Therefore, accurate forecasts of phases of tourism demand are as important as accurate forecasts of tourism demand itself.

Another reason why probability forecasting is as important as level forecasting is that forecasts come with uncertainty. Even if forecasts are positive, tourism demand may turn out to be negative. Therefore, forecasting phases in terms of their probability has significant benefits for stakeholders. Such a probability not only indicates the chance of the occurrence of a particular scenario, but also reflects the forecaster's confidence in his/her estimates. Suppose that two analysts with equal prediction accuracy are asked to assess the probability of tourism growth in the next period. The first analyst opines that tourism demand in the next period has a 90% chance of being positive, whereas the second provides an estimate of 60%. Although both forecasts suggest that demand will be positive, the first analyst is more confident and more informative than the second. Probability forecasts also play an essential role in risk management. Investors and government officials can combine the probabilities assigned to different states with business knowledge about the costs and benefits of each scenario to calculate the expected profits and take action accordingly.

Given the importance of tourism demand growth cycles and the useful information gained through probabilistic assessment of each phase, more and more probability forecasts are being published and probability forecasting models being built for major economic and financial indicators. For example, the Survey of Professional Forecasts in the US is a survey of macroeconomic forecasts undertaken quarterly since 1968. The survey presents forecasters' estimates of the probability that the growth of a particular variable will fall into a certain range. The Monetary Authority of Singapore has also published quarterly probability forecasts of key macroeconomic variables since 1999. Probability forecasts not only have a high practical value, but have been shown to be theoretically important. In earlier studies, Witt and Witt (1989, 1991) forecast both directional change in tourism demand and its turning points by origin, and evaluate the resulting forecasts by the percentage of correct predictions. However, they find that forecasting accuracy varies considerably with origin and that none of the models outperform the no-change model, also called the naïve model, except in two of the origin countries. These studies have spawned a series of efforts to build forecasting models for turning points, such as Rossello-Nadal (2001) and Kulendran and Wong (2009), and research evaluating their performance, such as Witt, Song and Louvieris (2003). In general, these researchers find that models using leading indicators generally outperform univariate time series models in forecasting directional change or growth rate cycles. Kulendran and Wong (2011) make a first attempt to predict expansion and contraction periods using logit and probit models, and evaluate the predictive power of various leading indicators using the quadratic probability score (QPS), an analogue of mean squared forecasting error. They find that real income changes are the most important factor determining turning points.

Although the QPS scores for most of the models used by Kulendran and Wong (2011) are less than 0.55, the threshold below which a model can be said to have predictive power, as suggested by Chen (2009), the issues of real-time forecasting, long-horizon forecasting, instability in tourism demand growth and performance measurement have not been fully addressed. Therefore, in the following sections, we discuss the challenges involved in characterising phases in tourism demand growth, namely the information lag problem and variance in states. As the states of a time series are not observable, they can be defined in many ways. For example, the state at time t can be defined using two-sided averages. Therefore, the state at time t depends not only on past values, but also on future values. To avoid the information lag problem, we focus on predicting directional changes in real time rather than trends based on two-sided averages. In addition, the magnitude of the variance in states can have a huge impact on predictability. As we show later, variance in states is a component of the QPS. The chosen definition of states must account for this.

Second, as investment plans or policies are usually implemented over several quarters, it is best to produce both short- and long-term forecasts. We propose nine models that differ in their methods of information combination and in the leading indicators used, and evaluate their predictive ability over both horizons. Third, during the period under study, tourism growth in Hong Kong fluctuated considerably due to local and global crises and policy changes. To accommodate instability in the series, therefore, all of the models are estimated in a rolling framework, allowing the parameters to adjust to the recent economic environment in a timely manner. Fourth, although the QPS is widely used to measure the performance of probability forecasts, it can at best provide a coarse summary of out-of-sample predictability. To understand why one model outperforms another, it is best to break the QPS into separate components and examine them closely. Hence, in the evaluation section, QPS is decomposed into three components, namely calibration, sharpness and uncertainty, as proposed by Murphy (1973). As these components represent three dimensions of performance, they offer a more comprehensive understanding of why one model is better than another.

Forecast combination is a burgeoning field of tourism research. Chan, Witt, Lee and Song (2010) and Shen, Li and Song (2011) consider several linear combination methods, and Cang (2011) examines a nonlinear combination approach. They find that combined forecasts generally outperform the best individual forecast. However, the performance of combined probability forecasts has never been studied in the tourism context. In practice, forecasters often hold different or even opposing views on future states. For example, Moutinho and Witt (1995) report on a group discussion in which they invited a panel of tourism experts to give their opinions on the probabilities of various possible directions for tourism development, such as the use of AI for tourism programme design, underwater hotels and space travel, and to summarise their opinions to reach a consensus. However, from the perspective of forecast users, it is difficult if not impossible to gather experts and derive a consensus from their opinions. In this study, we introduce several combination approaches to forecasting turning points in tourism growth, and thus offer practitioners a new means of summarising opinions.

The paper is structured as follows. In Section 2, we briefly discuss the procedures for estimating individual probability forecasting models, describe the possible combination approaches associated with these models and define the performance measurements. Section 3 describes the data used in the study and discusses the issues involved in characterising different states of tourism growth rate. The estimation results and empirical analysis are also presented. Section 4 summarises our findings and provides recommendations for further research.

2. Econometric methods

First, we assume that y_t is a binary variable defining two states of growth, such as expansion and contraction or positive and negative growth. It takes a value of 1 when the tourism market is in the first state and 0 otherwise. Due to the latent nature of growth states, a wide variety of methods is used to characterise them. Issues concerning the definition of y_t are covered in the next section.

We further suppose that each of M analysts is given a set of information $\{x_{1,t}, \dots, x_{K,t}\}$ with which to predict y_{t+h}, where h is the forecasting horizon. As our data are given in quarterly frequencies, we consider a 1-quarter-ahead forecast for $h = 1$ and a 1-year-ahead forecast for $h = 4$. The h-step-ahead forecasts generated by model m using observations up to time t are denoted by $p_{m,t+h}$. The analysts then build their own logit models based on certain functions of K predictors, and estimate them using a maximum likelihood estimator. To accommodate the possibility of structural breaks in the time series, the model parameters β_s are estimated in a rolling framework with T_0 observations in each rolling window. Therefore, the first T_0 observations are used to estimate the parameters of model m, $m = 1, \dots, M$ and compute the h-step ahead forecasts $\{p_{1,T_0+h}, \dots, p_{M,T_0+h}\}$. Next, the same model using observations from 2 to $T_0 + 1$ is estimated to generate forecasts $\{p_{1,T_0+1+h}, \dots, p_{M,T_0+1+h}\}$. This procedure continues until the last set of forecasts $\{p_{1,T}, \dots, p_{M,T}\}$ has been generated.

2.1 Combined Probability Forecasts

Combining probability forecasts of states in tourism demand is more complicated than combining forecasts of tourism demand growth, as the former process is restricted to the unit interval [0,1]. In this paper, we evaluate several combination approaches based on either in-sample statistics or out-of-sample performance. Simple average (SA) combinations with relatively high success rates in the literature are also considered. Given the nonlinear nature of the probabilities involved, we also consider a nonlinear approach that takes a geometric average of all of the forecasts. To differentiate the combined models from the individual models, we denote the combined forecasts f_t. To obtain the training weights $w_{m,t}$ for the m^{th} forecast, we further divide the timeline at $T_1 = 2T_0$ to generate forecasts f_t for $t = T_1 + 1, \dots, T$ under a rolling scheme.

2.1.1 Linear Opinion Pool

The first combination method combines the forecasts linearly: $f_t = w_{0,t} + \sum_{m=1}^{M} w_{m,t}\, p_{m,t}$. As probabilities are often understood as expert opinions, this method is commonly termed 'linear opinion pooling' (LiOP). Standard ordinary least squares can be used to estimate the combination weights $w_{0,t}$ and $w_{m,t}$. However, the forecasts f_t are not guaranteed to fall inside the unit interval. Therefore, we impose constraints on the weights, namely $w_{0,t} = 0$ and $w_{m,t} \geq 0$ for $\sum_{i=1}^{M} w_{i,t} = 1$, and term the constrained method CLiOP. The SA method that sets $w_{0,t} = 0$ and $w_{m,t} = M^{-1}$ for all m is its special case.

2.1.2 Model Selection Criterion

The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are conventional goodness-of-fit statistics for measuring in-sample performance. If a model is stable, out-of-sample performance is expected to be consistent with in-sample performance, so forecasts weighted by these statistics, denoted by $w_{m,t}^{AIC}$ and $w_{m,t}^{BIC}$, respectively, should perform well. The weights are given by

$$w_{m,t}^{AIC} = \frac{\exp\left(-0.5AIC_{m,t}\right)}{\sum_{m=1}^{M}\exp\left(-0.5AIC_{m,t}\right)}$$
$$w_{m,t}^{BIC} = \frac{\exp\left(-0.5BIC_{m,t}\right)}{\sum_{m=1}^{M}\exp\left(-0.5BIC_{m,t}\right)},$$

where $AIC_{m,t}$ and $BIC_{m,t}$ are the respective information criteria for the mth model. They are also special cases of LiOP.

2.1.3 Geometric Average

The simple geometric mean (GM) method is the only nonlinear combination approach considered in this paper. It assigns equal weight to each forecast and is defined as the Mth root of the product of M forecasts, as follows:

$$f_t^{GM} = \left(\prod_{m=1}^{M} p_{m,t}\right)^{\frac{1}{M}}.$$

2.2 Model Screening

With a large sample, maximum likelihood estimation for the logit model is consistent. In practice, however, data are often limited. Therefore, including too many predictors results in the inefficient use of information and thus poor forecasts. Assigning equal weight to every model, as in the SA and GM methods, may increase sensitivity to the quality of the forecasting models in the pool. Recent literature suggests that model screening should be conducted before combination. Yuan and Yang (2005) propose using a screening procedure with the aid of the AIC and BIC to select a subset of M models, say M^*, which is less than M, for combination. In the following, we use the AIC derived from an initial regression to select the best $M^* = 2$ and the best $M^* = 3$ models. This allows us to focus on comparing the performance statistics for the combined forecasts with those for the best model in the pool.

2.3 Benchmark Models

The performance of the above methods is compared with two benchmarks. The first is the historical average (HA) method, which takes the SA over the past T_0 observations. Its forecasts are simply equal to $f_t^{HA} = \sum_{s=t-h+1-T_0}^{t-h} y_s / T_0$. The second is the naïve model. As the forecast generated for the next period is equal to $f_t^{Naïve} = y_{t-h}$, this model is also known as the no-change model.

2.4 Performance Evaluation

Traditional out-of-sample evaluation of the forecast $\hat{y}_t$ relies on the mean squared forecast error ($MSFE$), defined as $MSFE = \sum_{t=T_1+1}^{T}(y_t - \hat{y}_t)^2/(T - T_1)$. When $\hat{y}_t$ is a probability forecast, the QPS, as proposed by Brier (1950), is commonly used (see Estrella and Mishkin, 1998; Chen, 2009; and Kulendran and Wong, 2011):

$$QPS = \frac{2}{T - T_1}\sum_{t=T_1+1}^{T}(y_t - \hat{y}_t)^2 \qquad ,$$

where $\hat{y}_t, t = T_1 + 1, \dots, T$ are the out-of-sample forecasts generated by one of the M individual models, $p_{m,t}, m = 1, \dots M$, or the combined probability forecasts, f_t. It is similar to the $MSFE$ but scaled up by a factor of 2. As y_t is binary and $\hat{y}_t$ falls in the [0,1] interval, the QPS lies between 0 and 2, with a score of 0 indicating perfect accuracy. To gain a deeper understanding of the quality of the forecasts, Murphy (1973) decomposes the QPS into three components, as follows:

$$QPS = Uncertainty + Calib - Sharp.$$

As y_t is a zero-one dummy variable, its variance is equal to $\bar{y}(1 - \bar{y})$, where $\bar{y}$ is the mean of y_t. Uncertainty simply doubles the variance of y_t: $Uncertainty = 2\bar{y}(1 - \bar{y})$. Therefore, when y_t does not fluctuate, such that $\bar{y} = 0$ or 1, no uncertainty arises. However, when y_t is equally distributed between 0 and 1 such that $\bar{y} = 0.5$, uncertainty reaches its maximum level, equalling 0.5. As the uncertainty depends on the observation values for y_t only, it is the same for all of the forecasting models, and it thus cannot be used to differentiate model performance.

In contrast, calibration and sharpness are both important quantifiers of model performance. The calibration, denoted by $Calib$, of a probability forecast measures the reliability of a forecasting model. For example, when an event happens 70% of the time, a model of the event is said to be reliable if its probability forecast is 0.7. Sharpness, abbreviated as $Sharp$, measures the extent to which forecasts are followed by different realisations.

We use an example to explain how these quantifiers differ in terms of performance evaluation. Suppose that 100 probability forecasts are denoted by $f_t, t = 1, \dots, 100$. Each of these forecasts corresponds to its actual observation $y_t, t = 1, \dots, 100$, with an overall average value of $\bar{y}$. We assign f_t and their respective y_t to five equally sized bins in a histogram: (0,0.2), (0.2,0.4), (0.4,0.6), (0.6,0.8) and (0.8,1). If the first forecast f_1 is equal to 0.9, f_1 and its corresponding observation y_1, whether zero or one, will be assigned to the last bin. After assigning all f_t and y_t to each cell, we can calculate their average values, denoted by $\bar{F}_b$ and $\bar{Y}_b$, for each bin $b = 1, \dots, 5$.

Comparing $\bar{F}_b$ with $\bar{Y}_b$ and $\bar{Y}_b$ with $\bar{y}$ provides two ways of evaluating the models. The first comparison is captured by calibration measuring the proximity of $\bar{F}_b$ to $\bar{Y}_b$. Suppose that 20 forecasts with values between 0 and 0.2 fall in the first bin. The average value $\bar{F}_1$ must also fall in the same range, say 0.1. This number means that on average, the event only happens 10% of the time. Therefore, we expect 2 of the 20 realisations y_t in the first bin to take the value of 1 and 18 to take the value of 0 such that $\bar{Y}_1$ also equals 0.1 if the forecasting model is reliable. If $\bar{F}_b$ is the same as $\bar{Y}_b$ for every bin, the calibration value is equal to 0, and the model is said to be perfectly calibrated. In contrast, if $\bar{F}_b = 0$ and $\bar{Y}_b = 1$, then the most unreliable model or the least well calibrated model has a maximum calibration score of 2. Therefore, when comparing two models, a more reliable (better calibrated) model should have a smaller calibration value.

Sharpness measures the variation in $\bar{Y}_b$ around $\bar{y}$. Its values range from 0 to 0.5, with a large value indicating a model with high discriminatory power. Although both $\bar{Y}_b$ and $\bar{y}$ involve observations of y_t only, the forecasts generated by the model determine how to allocate the realisations y_t, which in turn determine the value of $\bar{Y}_b$. Therefore, if a forecasting model contains discriminatory information on two states, the variation in $\bar{Y}_b$ is expected to be large. First, consider a model with zero sharpness. This happens when forecasts cannot be used to distinguish between two states – observations y_t attach randomly to forecasts, regardless of forecast size, and forecasts randomly assign observations to bins. In this scenario, $\bar{Y}_b$ is approximately equal to $\bar{y}$ for every bin. In contrast, a model is said to be sharp if forecasts f_t are able to differentiate $y_t = 0$ from $y_t = 1$. Using the previously discussed example, the small forecasting values in the first bin suggest that the '$y_t = 0$' state is most likely to arise. A sharp model assigns more $y_t = 0$ than $y_t = 1$ to the first bin, such that $\bar{Y}_1$ is different from $\bar{y}$. Similarly, more $y_t = 1$ are assigned to the last bin, in which the large forecasts are located.

In sum, as the uncertainty is the same for all of the models, it cannot be used to evaluate the models' predictive performance. However, the definition of states can have a huge impact on the QPS, as it is predetermined by y_t. In contrast, a probability forecasting model with a low calibration value is reliable, and a model with a high sharpness score is able to differentiate between states, making it much more informative than less sharp models. Therefore, our objectives are to determine appropriate definitions of states in tourism demand growth and to find a model that is as sharp as possible, subject to calibration.

3. Data and Empirical Results

3.1 Data Description

The quarterly data comprise the number of visitor arrivals to Hong Kong by region ($visitor_{i,t}$), the real income ($income_{i,t}$) of the source markets and tourism price ($P_{i,t}$) relative to that of Hong Kong for $i = 1, \dots, N$. The latter is defined as follows:

$$P_{i,t} = \frac{CPI_i/EX_i}{CPI_{HK}/EX_{HK}},$$

where $CPI_{i,t}$ and CPI_{HK} are the consumer price indices and $EX_{i,t}$ and EX_{HK} are the exchange rates relative to USD of market i and Hong Kong, respectively. The data on visitor arrivals from different source markets are obtained from the United Nations World Tourism Organization. The data on consumer price index, real income and exchange rate are obtained from the International Monetary Fund's Statistics Yearbook of International Finance. The data cover the period 1995Q1-2017Q1, with 89 time points for each series related to Hong Kong's top $N = 10$ source markets: China, Taipei, Korea, the US, Japan, Macao, the Philippines, Singapore, Thailand and Australia.

We then use log transformation to obtain the year-on-year growth rate of visitor arrivals ($a_{i,t}$), the growth rate of real income ($gdp_{i,t}$) and the growth rate of price level relative to that in Hong Kong ($price_{i,t}$), leaving $T = 85$ observations for the period from 1996Q1 to 2017Q1.

3.2 Individual Forecasting Models

As $a_{HK,t}$, $a_{i,t}$, $gdp_{i,t}$ and $price_{i,t}$ at lags are common indicators of tourism demand, we also examine their ability to predict turning points in tourism demand growth rates for $h = 1,4$. We consider the following $M = 9$ models, which use the information set in different ways.

M1: $y_t = F\left(\beta_0 + \beta_1 a_{HK,t-h}\right)$

M2: $y_t = F\left(\beta_0 + \sum_{i=1}^{N} \beta_{1,i} gdp_{i,t-h} + \sum_{i=1}^{N} \beta_{2,i} price_{i,t-h}\right)$

M3: $y_t = F\left(\beta_0 + \beta_1 a_{HK,t-h} + \sum_{i=1}^{N} \beta_{2,i} gdp_{i,t-h} + \sum_{i=1}^{N} \beta_{3,i} price_{i,t-h}\right)$

M4: $y_t = F\left(\beta_0 + \beta_1 gc_{1_{t-h}} + \beta_2 pc_{1_{t-h}}\right)$

M5: $y_t = F\left(\beta_0 + \beta_1 a_{HK,t-h} + \beta_2 gc_{1_{t-h}} + \beta_3 pc_{1_{t-h}}\right)$

M6: $y_t = F\left(\beta_0 + \beta_1 \overline{gdp}_{t-h} + \beta_2 \overline{price}_{t-h}\right)$

M7: $y_t = F\left(\beta_0 + \beta_1 a_{HK,t-h} + \beta_2 \overline{gdp}_{t-h} + \beta_3 \overline{price}_{t-h}\right)$

M8: $y_t = F\left(\beta_0 + \beta_1 \sum_{i=1}^{N} \delta_{i,t-h} gdp_{i,t-h} + \beta_2 \sum_{i=1}^{N} \delta_{i,t-h} price_{i,t-h}\right)$

M9: $y_t = F\left(\beta_0 + \beta_1 a_{HK,t-h} + \beta_2 \sum_{i=1}^{N} \delta_{i,t-h} gdp_{i,t-h} + \beta_3 \sum_{i=1}^{N} \delta_{i,t-h} price_{i,t-h}\right)$

M1 assumes that the lagged tourism demand growth rate of a_{HK} already accurately reflects the economic conditions at time $t - h$, so including other variables cannot increase predictability.

M2 and M3 are two information combination models: M2 includes all of the lagged predictors, gdp_i and $price_i$, of the N cross-sectional units in the model, and M3 further adds $a_{HK,t-h}$ to M2. As the number of observations is limited, the forecasts are contaminated by the error arising from estimating the parameters β_s.

To address this issue, we consider three ways of reducing the dimensionality of the predictors. M4 uses the first principal components of real gross domestic product growth (gc_1) and relative price change (pc_1) as the two leading indicators. M6 summarises this information by taking SAs across the N cross-sectional units, $\overline{gdp}$ and $\overline{price}$. M8 weights each of the N units by the proportion of arrivals from the source destination i at time $t - h$, given by $\delta_{i,t} = visitor_{i,t} / \sum_j visitor_{j,t}$. The models M5, M7 and M9 are their dynamic counterparts, with $a_{HK,t-h}$ added to each. As logit and probit models yield similar results, our study reports the findings generated by logit models, where F is the cumulative distribution function.

3.3 Definition of Turning Points

Various methods can be used to characterise a turning point, ranging from the Markov switching model proposed by Hamilton (1989) to a rule-based method such as the algorithm proposed by Bry and Boschan (1971). The Markov switching model is usually used to study the lifecycle of tourism demand (see Moore and Whitehall, 2005), but its consistency largely hinges on the validity of the model specification for a_t. Alternatively, the rule-based method is often used in the forecasting literature because of its simplicity and replicability. For example, Zellner, Hong and Min (1991), Gouveia and Rodrigues (2005) and Kulendran and Wong (2009, 2011) identify the peaks (troughs) in tourism demand growth rate when it reaches a local maximum (minimum). That is, a downturn takes place when a_t is largest in the nearest time domain $[t - W, \dots, t + W]$, where W is an arbitrary positive integer, and vice versa. This means that time t forms a peak when $a_t > a_{t \pm w}$ for $w = 1, \dots, W$, and a trough when $a_t < a_{t \pm w}$. However, this method is not suitable for monitoring turning points in real time, as it requires knowledge of future states, a_{t+w}, at time t. This creates an information lag problem.

Another issue when considering the definition of states is the size of uncertainty. As uncertainty is a major component of the QPS, a state characterisation with volatile movement in y_t necessarily makes the indicators less predictive. For example, the rule-based method generates 14 turning points with $\bar{y} = 47\%$ of the observations falling in the expansion period. Its uncertainty is as high as 0.4982, which is close to the no-predictability threshold $QPS^* = 0.55$. Therefore, it is difficult to find sufficiently informative leading indicators for highly volatile trends.

To avoid the real-time information lag problem and minimise uncertainty, we focus solely on directional change in tourism demand growth, y_t, as follows:

$$y_t = \begin{cases} 1 & a_t > 0 \\ 0 & a_t \leq 0 \end{cases} \quad .$$

This characterisation is also adopted by Witt and Witt (1989, 1991), Witt et al. (2003) and Kulendran and Wong (2009). y_t has 10 turning points, of which 84.7% are positive, yielding an uncertainty score of only 0.26, almost half that generated by Bry and Boschan's (1971) procedure. In addition, y_t has a special relationship with a_t. Common predictors of a_t are likely also to predict y_t . To illustrate, suppose that $a_t = \alpha_0 + \alpha_1 x_{1,t-h} + \alpha_2 x_{2,t-h} - \varepsilon_t$, where $x_{1,t-h}$ and $x_{2,t-h}$ are the major determinants of a_t. For simplicity, $z_t = \alpha_0 + \alpha_1 x_{1,t-h} + \alpha_2 x_{2,t-h}$. Equivalently, $a_t = z_t - \varepsilon_t$. When $a_t > 0$, we have $z_t > \varepsilon_t$. According to our definition of y_t, the probability forecast for a positive state $P(y_t = 1)$ is equal to $P(a_t > 0)$. Therefore, we obtain the relationship $P(y_t = 1) = P(z_t > \varepsilon_t)$. The latter is simply equal to the cumulative distribution for ε_t, or $F(z_t)$. Due to the linkage $P(y_t = 1) = F(z_t) = F(\alpha_0 + \alpha_1 x_{1,t-h} + \alpha_2 x_{2,t-h})$, the leading indicators of a_t may also predict y_t.

Figure 1 plots the path of a_t (solid line) and highlights positive states in grey and negative states in white. Although the implementation of the Individual Visitor Scheme had a huge positive impact on Hong Kong's tourism market, the Asian financial crisis in 1997, the SARS outbreak in 2003, the mortgage subprime crisis in around 2008 and the Chinese government's tightening of the 'one trip per week' measure exerted strong downward pressure on local tourism.

[Insert Figure 1 here]

3.4 Empirical Results for Individual Models

A rolling window size of 30 quarters, i.e. $T_0 = 30$ and $T_1 = 60$, is applied to the individual models and combined approaches, leaving 25 quarters from 2011Q1 to 2017Q1 for the out-of-sample evaluation. We begin the analysis by presenting the in-sample performance of the individual models M1 to M9. As HA forecasts can be obtained by $p_t = F(\beta_{0,t})$, we determine whether the individual models outperform the HA method by testing at every time point whether all of the unknown parameters except the constant term are jointly equal to 0. The likelihood ratio (LR) statistic is used to test the null hypothesis

$H_0\colon \beta_{j,t} = 0, j \neq 0$. For example, the null hypothesis for M2 is $H_0\colon \beta_{1,i,t} = \beta_{2,i,t} = 0, i = 1, \dots, N$. When the p-value is small, we reject the null and conclude that the model outperforms the HA method.

The plots of the time-varying p-values for $h = 1$ in Figure 2 and $h = 4$ in Figure 3 reveal two issues. First, models M2 and M3, which do not use information effectively, severely underperform regardless of the forecasting horizon, and M1, which depends on past tourism growth, loses its predictability for $h = 4$. Second, the p-value paths are highly volatile, suggesting that the predictability of the models changes over time and justifying the use of a rolling framework. These models, using summarised information such as the principal components and weighted averages of the leading indicators, significantly outperform the HA method after 2011Q1 for $h = 1$ and after 2010Q2 for $h = 4$.

[Insert Figures 2 and 3 here]

Next, Table 1 presents the out-of-sample predictability of the individual models and the two benchmarks for $h = 1{,}4$. From 2011Q1 to 2017Q1, uncertainty equals 0.32 regardless of forecasting horizon or model.

[Insert Table 1 here]

In line with their in-sample performance, models M2 and M3, which combine all of the information provided, have the poorest forecasting performance over both horizons. Given a finite sample, incorporating all available information creates too much noise; a huge estimation error in the parameters results in a QPS larger than the no-predictability threshold ($QPS^* = 0.55$) for both models, with the exception of M3 for $h = 4$. These high scores can be attributed to two factors. First, M2 and M3 are the most unreliable models with the highest calibration values. Second, they are the least informative for $h = 1$, with close-to-zero sharpness scores (0.0187 for M2 and 0.0251 for M3). While they are not the least sharp models for $h = 4$, their sharpness scores are still close to the lowest boundary.

The models that summarise information are superior to M2 and M3. For $h = 1$, the dynamic models M5, M7 and M9 show an obvious marginal gain over the corresponding static models without $a_{HK,t-1}$. All of the dynamic models except M5, whose performance is the same as that of the naïve model, outperform the two benchmarks. This demonstrates that tourism demand retains a strong momentum for at least 1 quarter. However, for $h = 4$, the marginal gain generated by $a_{HK,t-4}$ shrinks, as events that occurred 4 quarters ago are usually less informative than events that happened in the previous quarter. Therefore, the QPSs for $h = 4$ are generally higher than those for $h = 1$. However, six models (M1, M4, M5, M6, M8 and M9) still outperform the benchmarks for $h = 4$. Another interesting insight gained from Table 1 is that the variation in QPS for $h = 4$ is much smaller than that for $h = 1$. The dynamic term $a_{HK,t-4}$ loses its predictive power for long-term forecasting. Therefore, model predictability relies almost entirely on whether the common predictors (income growth and relative price change) are used effectively. Short-term forecasting performance relies more on the dynamic term $a_{HK,t-1}$. Adding more predictors is beneficial only when the benefits arising from the extra information outweigh the increase in estimation error arising from estimating the parameters. This partly explains why the QPS difference between M1 (0.1288) and M3 (0.66) is so large.

The huge QPS improvement in the dynamic models for $h = 1$ can be attributed mainly to a substantial decrease in calibration and increase in sharpness. The inclusion of a dynamic term reduces the calibration score by 43%, to 90%. The close-to-zero calibration values, ranging from 0.008 to 0.027, suggest that the dynamic models (M5, M7 and M9) are more reliable than their static counterparts. In addition, their sharpness values show a remarkable increase, rising from 62% to 357% relative to those of their static analogues, reflecting the information hidden in previous tourism demand growth. Together, the high sharpness values and almost perfect calibration scores suggest that these models not only predict states precisely but also generate bold forecasts that are close to the boundaries. To visualise the usefulness of sharpness, we contrast the short-term forecasts generated by the naïve model (solid line) with the HA forecasts (dashed line), and the three dynamic models M5, M7 and M9 (solid lines) with their static counterparts (dashed lines) in Figure 4. The forecasts generated by the former models are clearly much closer to the probability boundary 0 or 1, especially for 2016, than those produced by the latter models.

[Insert Figure 4 here]

Overall, we find that the leading indicators of tourism demand growth rate can also predict positive and negative states in Hong Kong tourism demand. However, their predictive ability depends on how hidden information and the forecasting horizon are used.

A model's predictive power can be improved only by including crucial information. If too much information is combined, the calibration, sharpness and thus quality of a probability forecasting model is compromised. Our empirical studies show that M9 for $h = 1$ and M5 for $h = 4$, which summarise information, outperform the other models. Although forecasting far into the future is certainly useful for practitioners, it is not as accurate as forecasting the next period. As many changes in the economic environment are expected to occur over a year, for example, it is difficult for forecasters to make bold forecasts of events occurring a year later, resulting in a loss of sharpness.

3.5 Empirical Results for Combined Probability Forecasts

We consider four linear approaches (SA, AIC, BIC and CLiOP) and one nonlinear approach (GM) to combining the nine forecasting models. Table 2 shows their performance over short and long forecasting horizons. We find that regardless of the approach taken, the performance of these combination methods is less promising than indicated in the literature on forecast combination for tourism growth rate.

First, none of the four linear methods except CLiOP for $h = 4$ surpass the best individual model, although all four outperform the worst model in the pool. The performance of the short-horizon combined forecasts deviates more from that of the best model than the performance of the long-horizon combinations, which does not differ much. As mentioned earlier, the QPS varies more for $h = 1$ (from 0.0968 to 0.916) than for $h = 4$ (from 0.244 to 0.59). When the forecasts are combined, the worst model necessarily influences the quality of the combined forecast. This problem is exacerbated by combining the individual models in a nonlinear way. The GM approach, which assigns equal weight to every model, makes the combined forecasts highly unreliable and uninformative for future states. Not only does it generate the highest calibration value, but the resulting model almost entirely lacks sharpness. Its QPS for $h = 1$ is again larger than its QPS for $h = 4$.

Comparing the above mediocre results for the combination approaches with the success of traditional forecast combination approaches in the tourism growth literature, we hypothesise that the performance of combined forecasts depends predominantly on the quality of the forecasts in the pool, as the individual probability forecasts are now restricted to the [0,1] interval.

[Insert Table 2 here]

To test this hypothesis, we focus on a small-scale combination. We use the AIC[1] derived from the initial set of regressions to select the top M^* models where $M^* = 2{,}3$. The models with the lowest AIC (in ascending order) are M1, M7 and M5 for $h = 1$ and M1, M9 and M5 for $h = 4$.

Table 3 shows that the overall performance of these models is substantially better for both forecasting horizons, with the exception of a negligible increase in QPS for CLiOP for $h = 4$. Of all of the combination approaches, the GM model shows the greatest improvement. Its QPS is reduced by around 90% for $h = 1$ and 65% for $h = 4$. In general, such improvement is more significant for $h = 1$ than for $h = 4$, as the quality of the models with a short forecasting horizon varies more than that of the models with a long forecasting horizon. This outcome corroborates our earlier argument that combined forecasts are sensitive to the quality of forecasts in the pool. In particular, the AIC outperforms the best individual model for $h = 1$ and SA beats the best model for $h = 4$ when $M^* = 2$. When $M^* = 3$, the BIC combination involving M1, M5 and M7 is as good as the best individual model (M9) for $h = 1$, and the SA method retains its superiority over the other combination approaches for $h = 4$. This improvement derives mainly from the increase in sharpness.

Overall, our results yield several insights. First, combined probability forecasts are highly sensitive to the quality of the models in the pool. Therefore, it is vital to ensure high forecast quality to maximise the accuracy of a combined forecast. Sharpness generally increases when the quality of the models in the pool improves. Second, although combining probability forecasts does not necessarily improve their reliability, the calibration value of a combined forecast is generally very small.

3.6 Percentages of correct predictions

Lastly, in Tables 4 to 6, we report the hit rates, which are equal to the percentages of correct forecasts. A forecast is correct if the probability of a positive state is higher than the 0.5 threshold and the associated actual state turns out to be positive, or if the probability

[1] Assessment using the BIC yields the same results.

is lower than 0.5 and a negative state is realised. Our results show that the performance of the models as assessed by hit rate is broadly in line with their performance as assessed by the QPS.

Specifically, for $h = 1$, the dynamic models with $a_{HK,t-1}$ outperform their static counterparts. The hit rates of the three dynamic models (M5, M7 and M9) are between 92% and 96%, whereas their static counterparts (M4, M6 and M8) achieve hit rates between only 80% and 84%. M9 is again the best model, with 4% more correct forecasts than the naïve model and 16% more than the HA model. As discussed earlier, as the information content of $a_{HK,t-4}$ decreases, the dynamic models lose almost all of their marginal gain in predictability relative to the corresponding static models. Hit rates are not improved by including lagged terms for the weighted predictors (M7 and M9). The best model in terms of QPS (M5) remains the best model when hit rates are used for performance measurement. It also shows a 4% gain relative to its static version.

[Insert Table 4 here]

The results in Table 5 are based on combined forecasts created by pooling all of the models without screening, whereas those in Table 6 are based on forecasts using the top M^* models. Comparing the results shown in these two tables for $h = 1$, we observe again that model screening plays an important role in improving the percentages of correct predictions. Almost all of the combination approaches, especially the geometric approach, show a substantial increase in hit rate when the top $M^* = 2,3$ models are combined.

For $h = 4$, all of the linear combination methods perform better than the two benchmarks. Model screening does not necessarily increase the number of correct predictions. However, this reflects the difficulty of predicting the occurrence of an event far from the current time point, and hit rates provide at best a coarse performance measurement. They do not inform practitioners about the properties of forecasts, and thus cannot provide reliable guidance for model selection.

[Insert Tables 5 and 6 here]

4. Concluding Remarks

This study emphasises the importance of forecasting different phases of tourism demand, which has high practical value for both risk management and policy planning. We determine whether leading indicators of tourism demand can also predict states in the growth rate of tourism demand in Hong Kong over both short and long forecasting horizons. The indicators are the rate of growth of the number of visitor arrivals to Hong Kong; real income growth rate; and changes in relative real exchange rates in the top 10 source markets. We evaluate nine logit models that differ in their use of information, summarising the indicators by SAs, principal components or weighted averages. The models' out-of-sample predictive ability is evaluated using not only the QPS but also its three components, calibration, sharpness and uncertainty, helping analysts to characterise states in tourism demand and understand the sources of improvements to predictive ability.

Our empirical findings using quarterly data from 1996Q1 to 2017Q1 can be summarised as follows. First, we find that two key issues must be addressed when characterising states in tourism demand: real-time forecasting and uncertainty. Accordingly, we avoid using a definition that requires two-sided information and select a characterisation with acceptable uncertainty. Therefore, we choose to focus on directional changes in our analysis. Second, we find that putting all of the information into a single model does not improve forecasting accuracy, but rather creates noise. Models containing too much information are not only poorly calibrated; they also lack sharpness. Models that weigh information by either principal components or averaging across source markets have a much lower QPS, lower calibration values and greater sharpness.

We further evaluate the performance of combined probability forecasts using either all of the models or the top M^* models, and find that they yield substantially different results. When all of the models are included, the QPS for the linear approaches is only midway between that of the best and worst models, and the QPS for the nonlinear GM approach is even higher than that for the worst model. The findings indicate a major challenge to probability forecast combination, namely the quality of forecasts in the pool. As combination approaches are highly sensitive to forecast quality, we recommend applying the AIC to the initial set of observations to select the top two or three models before combination. We find that model screening substantially reduces the QPS of all of the combination approaches, especially the GM approach. The improvement in predictive ability is due mainly to a conspicuous increase in sharpness. We also observe that the combination approaches weighted by in-sample statistics and the SA method surpass the best individual model. Overall, the results show that combining probability forecasts can improve predictive power only when the quality of the forecasts in the pool is high, suggesting that model screening should be conducted before combination.

The analytical framework outlined here is particularly relevant to tourism policy makers and business stakeholders. The logit models will help practitioners to obtain warning signals in advance of turning points. The estimation results illustrate the importance of conventional leading indicators, and show that how we use information is crucial to predictive power. Our evaluation process also demonstrates that the components of the QPS can not only guide analysts in defining a state but also help them to gain a deeper understanding of the differences in predictive power between models, whereas hit rates provide at best a coarse summary. Further, if a high-quality spectrum of probability forecasts is available on the market, practitioners can obtain a useful consensus either by combining forecasts using in-sample statistics or simply by averaging forecasts.

This study has some limitations. Although tourism demand growth occurs in more than two phases, we focus on two states only. This simplifies not only the interpretation of tourism demand growth, but also the combination of probability forecasts. The study is also limited by the small scope of combination approaches. As governments start to collect opinions on various economic indicators, such as the health of the tourism industry, in terms of probability, further research evaluating a larger pool of linear and nonlinear combined probability forecasts will be critical. Knowing how to select and how to combine probability forecasts will help to ensure that effective investment and planning decisions are made.

REFERENCES

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3.

Bry, G., & Boschan, C. (1971). Cyclical analysis of time series: Selected procedures and computer programs. National Bureau of Economic Research, Columbia University Press.

Butler, R. (1980). The concept of a tourism area cycle of evolution: Implications for management resources. *The Canadian Geographer*, 24, 5–16.

Cang, S. (2011). A non-linear tourism demand forecast combination model. *Tourism Economics*, 17(1), 5–20.

Cang, S. (2014). A comparative analysis of three types of tourism demand forecasting models: Individual, linear combination and non-linear combination. *International Journal of Tourism Research*, 16(6), 595–607.

Chan, C.K., Witt, S.F., Lee, Y.C., & Song, H. (2010). Tourism forecast combination using CUSUM technique. *Tourism Management*, 31, 891–897.

Chen, S.S. (2009). Predicting the bear stock market: Macroeconomic variables as leading indicators. *Journal of Banking and Finance*, 33, 211–223.

Chen, K.Y., & Wang, C.H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management*, 28(1), 215–226.

Clark, T.E., & West, K.D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291–311.

Claveria, O., Monte, E., & Torra, S. (2015). Tourism demand forecasting with neural network models: Different ways of treating information. *International Journal of Tourism Research*, 17, 494–500.

Diebold, F.X., & Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253–263.

Estrella, A., & Mishkin, F.S. (1998). Predicting U.S recessions: Financial variables as leading indicators. *Review of Economics and Statistics*, 80(1), 45–61.

Gouveia, P.M., & Rodrigues, P.M.M. (2005). Dating and synchronizing tourism growth cycles. *Tourism Economics*, 11(4), 501–515.

Hadavandi, E., Ghanbari, A., Shahanaghi, K., & Abbasian-Naghneh, S. (2011). Tourist arrival forecasting by evolutionary fuzzy systems. *Tourism Management*, 32(5), 1196–1203.

Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57, 357–384.

Hassani, H., Webster, A., Silva, E.S., & Heravi, S. (2015). Forecasting US tourist arrivals using optimal singular spectrum analysis. *Tourism Management*, 46, 322–335.

Kulendran, N., & Witt, S.F. (2003). Leading indicator tourism forecasts. *Tourism Management*, 24, 503–510.

Kulendran, N., & Wong, K. (2009). Predicting quarterly Hong Kong tourism demand growth rates, directional changes and turning points with composite leading indicators. *Tourism Economics*, 15(2), 307–322.

Kulendran, N., & Wong, K. (2011). Determinants versus composite leading indicators in predicting turning points in growth cycle. *Journal of Travel Research*, 50(4), 417–430.

Moore, W., & Whitehall, P. (2005). The tourism area lifecycle and regime switching models. *Annals of Tourism Research*, 32(1), 112–126.

Moutinho, L., & Witt, S.F. (1995). Forecasting the tourism environment using a consensus approach. *Journal of Travel Research*, 33(4), 46–50.

Murphy, A.H. (1973). Hedging and skill scores for probability forecasts. *Journal of Applied Meteorology*, 12, 215–223.

Rossello-Nadal, J. (2001). Forecasting turning points in international visitor arrivals in the Balearic Islands. *Tourism Economics*, 7, 365–380.

Shen, S., Li, G., & Song, H. (2011). Combination forecasts of international tourism demand. *Annals of Tourism Research*, 38(1), 72–89.

Witt, S.F., Song H., & Louvieris, P. (2003). Statistical testing in forecasting model selection. *Journal of Travel Research*, 42, 151–158.

Witt, S.F., & Witt, C.A. (1989). Measures of forecasting accuracy: Turning point error v. size of error. *Tourism Management*, 10(3), 255–260.

Witt, S.F., & Witt, C.A. (1991). Tourism forecasting: Error magnitude, direction of change error, and trend change error. *Journal of Travel Research*, 20, 26–33.

Witt, S.F., & Witt, C.A. (1995). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting*, 11, 447–475.

Wu, C., Song, H., & Shen S. (2017). New developments in tourism and hotel demand modeling and forecasting. *International Journal of Contemporary Hospitality Management*, 29(1), 507–529.

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.

Yuan, Z., & Yang, Y. (2005). Combining linear regression models: When and how? *Journal of the American Statistical Association*, 100, 1202–1214.

Zellner, A., Hong, C., & Min, C. (1991). Forecasting turning points in international output growth rates using Bayesian exponentially weighted auto regression time varying parameters and pooling techniques. *Journal of Econometrics*, 49, 275–300.

Figure 1. Tourism demand growth rates

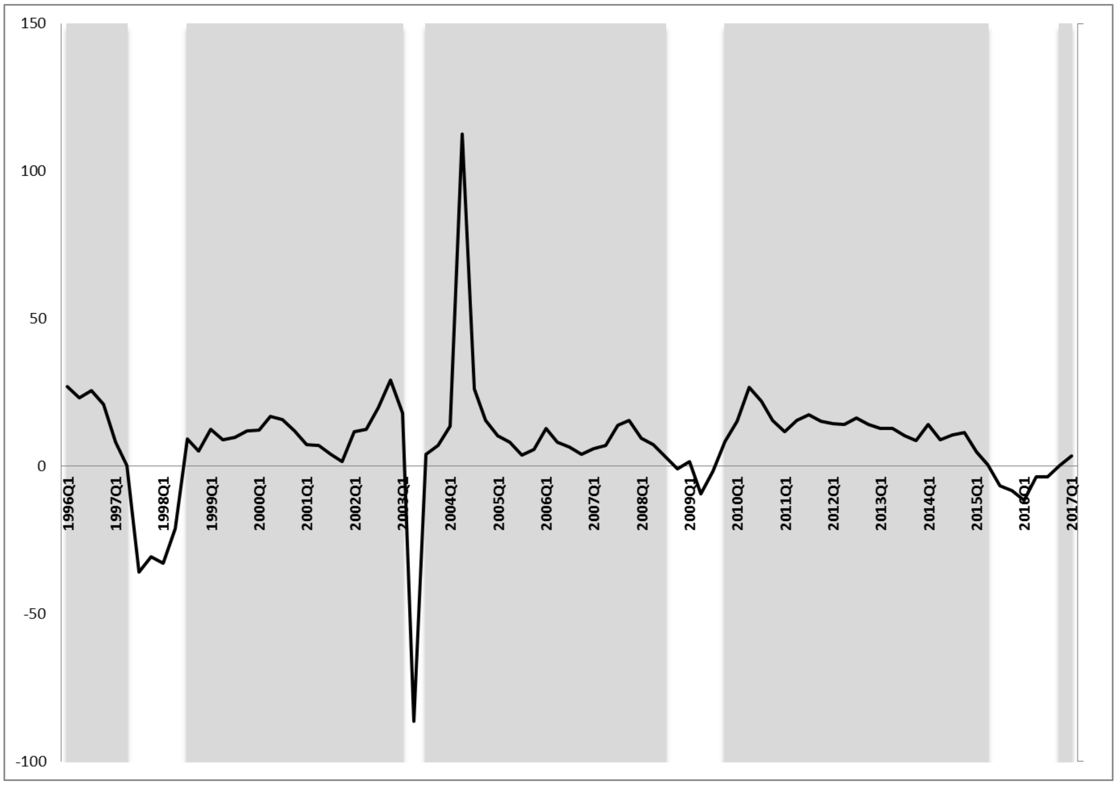

Figure 2. In-sample performance measured by p-values of LR statistics for $h = 1$

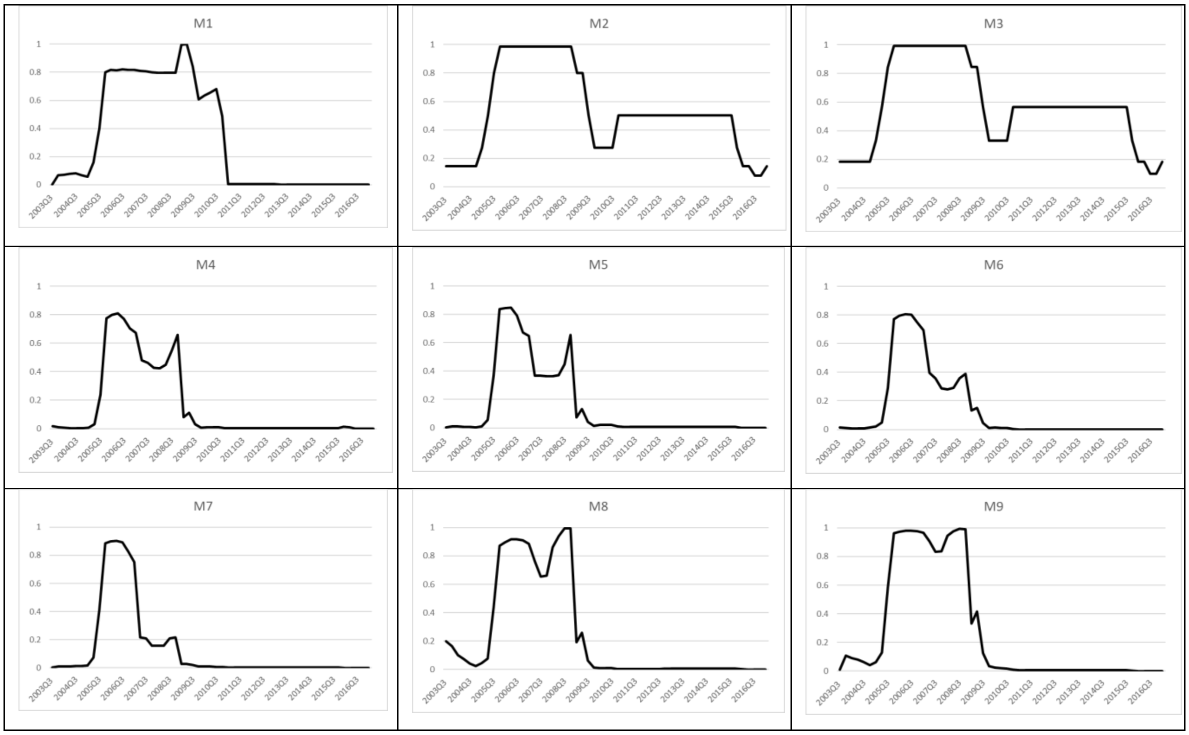

Figure 3. In-sample performance measured by p-values of LR statistics for $h = 4$

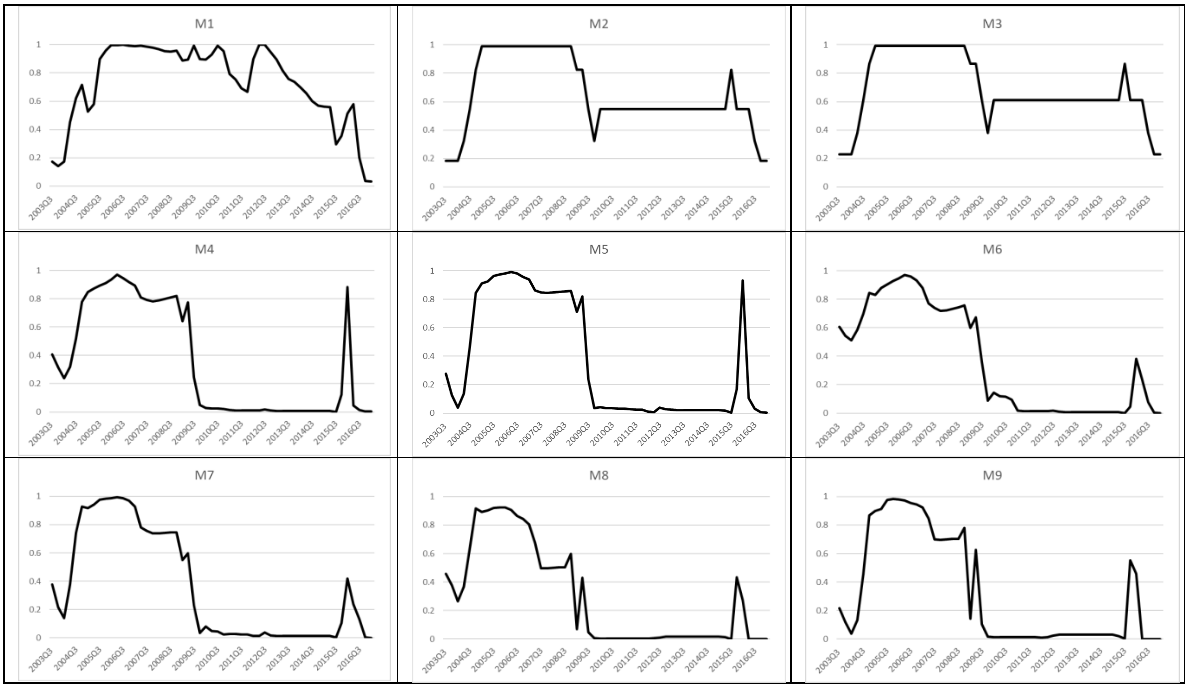

Figure 4. The out-of-sample one-step ahead forecasts

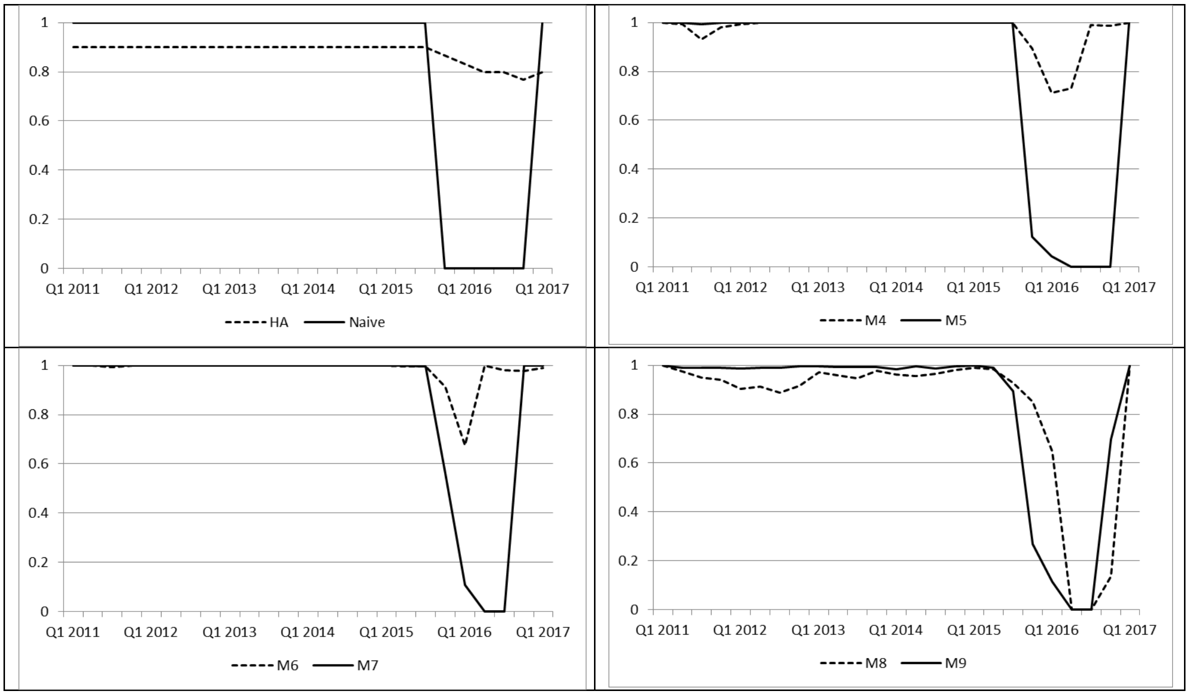

Table 1 Performance of individual models

	$h = 1$			$h = 4$		
	QPS	Calib	Sharp	QPS	Calib	Sharp
M1	0.1288	0.0248	0.2160	0.2760	0.2002	0.2442
M2	0.9160	0.6147	0.0187	0.5960	0.3088	0.0328
M3	0.6600	0.3651	0.0251	0.4680	0.2013	0.0533
M4	0.2888	0.0801	0.1113	0.2952	0.0346	0.0594
M5	0.1480	0.0080	0.1800	0.2440	0.0459	0.1219
M6	0.3144	0.0477	0.0533	0.3208	0.0165	0.0157
M7	0.1032	0.0270	0.2438	0.3592	0.0425	0.0033
M8	0.2504	0.0523	0.1219	0.2824	0.0346	0.0722
M9	0.0968	0.0208	0.2440	0.2824	0.0468	0.0844
HA	0.3016	0.0159	0.0343	0.3400	0.0200	0
Naive	0.1480	0.0080	0.1800	0.4040	0.0888	0.0048

Note: uncertainty is 0.32 for all models.

Table 2 Performance of combined probability forecasts

	$h = 1$			$h = 4$		
	QPS	Calib	Sharp	QPS	Calib	Sharp
SA	0.1608	0.0455	0.2047	0.2440	0.0418	0.1178
AIC	0.1480	0.032	0.2040	0.2632	0.0134	0.0702
BIC	0.1672	0.0512	0.2040	0.2632	0.0134	0.0702
CLiOP	0.1544	0.0184	0.1840	0.2376	0.0136	0.0960
GM	0.9928	0.6981	0.0253	0.8264	0.5440	0.0376

Table 3 Performance of combined probability forecasts using AIC model screening

(a) $M^* = 2$

	$h = 1$			$h = 4$		
	QPS	Calib	Sharp	QPS	Calib	Sharp
SA	0.1096	0.0336	0.2440	0.2376	0.042	0.1244
AIC	0.0904	0.0144	0.2440	0.2568	0.0482	0.1114
BIC	0.0968	0.0968	0.3200	0.2568	0.0482	0.1114
CLiOP	0.1224	0.0584	0.2560	0.2696	0.1405	0.1909
GM	0.1032	0.0632	0.2800	0.2696	0.0351	0.0855

(b) $M^* = 3$

	$h = 1$			$h = 4$		
	QPS	Calib	Sharp	QPS	Calib	Sharp
SA	0.1288	0.0128	0.2040	0.2248	0.0283	0.1235
AIC	0.1032	0.0272	0.2440	0.2952	0.0596	0.0844
BIC	0.0968	0.0968	0.3200	0.2568	0.0482	0.1114
CLiOP	0.1224	0.0584	0.2560	0.2696	0.0856	0.1360
GM	0.1224	0.0584	0.2560	0.2568	0.0218	0.0850

Table 4 Hit rates for individual models with 0.5 as threshold

	$h = 1$	$h = 4$
M1	0.92	0.76
M2	0.44	0.64
M3	0.60	0.72
M4	0.80	0.80
M5	0.92	0.84
M6	0.80	0.80
M7	0.92	0.80
M8	0.84	0.84
M9	0.96	0.84
HA	0.80	0.80
Naive	0.92	0.76

Table 5 Hit rates for combined approaches with 0.5 as threshold

	$h = 1$	$h = 4$
SA	0.92	0.88
AIC	0.92	0.84
BIC	0.88	0.84
CLiOP	0.92	0.84
GM	0.36	0.48

Table 6 Hit rates for combined approaches with 0.5 as threshold with model screening

(a) $M^* = 2$

	$h = 1$	$h = 4$
SA	0.96	0.84
AIC	0.96	0.80
BIC	0.92	0.80
CLiOP	0.92	0.84
GM	0.92	0.84

(b) $M^* = 3$

	$h = 1$	$h = 4$
SA	0.92	0.84
AIC	0.92	0.84
BIC	0.92	0.84
CLiOP	0.92	0.80
GM	0.92	0.88