A combination selection algorithm on forecasting

Shuang Cang[a] & Hongnian Yu[b]

a: School of Tourism, Bournemouth University, Fern Barrow, Poole, Dorset BH12 5BB, UK
b: School of Design, Engineering & Computing, Bournemouth University, Fern Barrow, Poole, Dorset BH12 5BB, UK

ABSTRACT

It is widely accepted in forecasting that a combination model can improve forecasting accuracy. One important challenge is how to select the optimal subset of individual models from all available models without having to try all possible combinations of these models. This paper proposes an optimal subset selection algorithm from all individual models using information theory. The experimental results in tourism demand forecasting demonstrate that the combination of the individual models from the selected optimal subset significantly outperforms the combination of all available individual models. The proposed optimal subset selection algorithm provides a theoretical approach rather than experimental assessments which dominate literature.

Keywords: *Neural Networks; Seasonal Autoregressive Integrated Moving Average; Combination Forecast; Information Theory.*

1. Introduction

Forecasting has received the considerable research during the past three decades. Three main types of forecasting models (Li, Song & Witt, 2005; Song & Li, 2008) are *Time series model* (Cao, Ewing & Thompson, 2012; Cho, 200; Goshall & Charlesworth, 2011), *Causal econometric model* (Li, Song & Witt, 2006; Naude & Saayman, 2005; Page, Song & Wu, 2012; Roget & Gonzalez, 2006) and new emerging *Artificial Intelligence based model*, such as neural network, fuzzy time-series theory, grey theory, genetic algorithms, and expert systems (Cao, Ewing & Thompson, 2012; Carbonneau, Laframboise & Vahidov, 2008; Bodyanskiy & Popov 2006; Chen & Wang, 2007; Cho, 2003; Hadavandi, Ghanbari , Shahanaghi & Abbasian-Naghneh, 2011; Law & Au, 1999; Pai & Hong, 2005; Wong, Xia & Chu, 2010; Wu & Akbarov, 2011). From these studies, researchers often seek to identify the best individual model to generate a forecast. However, combination forecasting has proven to be a highly successful forecasting strategy in many fields, which has been demonstrated by empirical studies.

Forecast combination was pioneered in the sixties by Bates and Granger (1969). Since then it has been demonstrated that forecast combinations are often superior to their constituent forecasts in many fields (Greer, 2005; Hall & Mitchell, 2007; Holden & Peel, 1986; Lessmann et al. (2012); Li, Shi & Zhou, 2011; Newbold & Granger, 1974; Sánchez, 2008; Timmermann, Elliott & Granger, 2006; Winkler & Makridakis, 1983; Zheng, Lee & Shi, 2006). The most widely used and studied combination forecast methods are ensemble methods, such as bagging (Breiman, 1996) and boosting methods. The typical boosting methods are AdaBoost (Freund & Schapire, 1997), LogitBoost (Tibshirani, Friedman & Hastie, 2000) and MultiBoost (Webb, 2000). These methods which have the learning capability have two steps: step 1: construct a set of predication models; step 2: predicate a new pattern by taking a weighted vote of their predications. The average or median is used for the continuous outputs, and the majority voting is used for the categorical outputs of the set of predication models from step 1. The most applications are the categorical outputs from the set of predication models. For examples, Wezel & Potharst (2007) applied ensembles methods (bagging and boosting) to the customer choice modelling problem to improve customer choice predictions. Abellán and Masegosa (2010) proposed the ensemble method using credal decision trees, and showed the good percentage of correct classifications and an improvement in time of processing, especially for large data sets. Finlay (2011) applied bagging and boosting methods to the credit risk assessment to classify consumers as good or bad credit risks, and proposed a new boosting algorithm, 'error trimmed boosting'. Experiments showed that the bagging and boosting methods outperform other multi-classifier systems, and 'error trimmed boosting' outperforms bagging and AdaBoost by a significant margin.

For the continuous outputs from the set of predication models, Li, Wong and Troutt (2001) proposed an approximate Bayesian algorithm for combining forecasts using several examples. Zou and Yang (2004) developed an algorithm called 'AFTER' to calculate the weights in the combination forecasting with one-step-ahead forecasting, where the weights are updated for each additional observation. The results demonstrated the advantage of the 'AFTER' algorithm. He and Xu (2005) applied the self-organizing algorithm to combine the forecasting models, and demonstrated the superiority by an example of the total retail sales of consumer goods in Chengdu. All individual candidate models are used in the combination for these researches (Li, Wong and Trout, 2001; Zou and Yang, 2004; He and Xu, 2005).

For tourism demand forecasting, the outputs of the individual models are continuous variables. The most common combination forecasting models are linear combination of all available individual forecast models in tourism literature. The researchers (Andrawisa et al., 2011, Chan et al., 2010; Coshall & Charlesworth, 2011; Freitas & Rodrigues, 2006; Lessmann et al., 2012; Menezes, Bunn & Taylor, 2000; Shen, Li & Song, 2011) have demonstrated the efficiency of combination forecasts and the superiority of combination forecasts in contrast to individual forecasts. However all available individual models are used as inputs for the combinations. The question is whether we can optimally select a subset of all individual models instead of all individual models in constructing the combination model. If a subset of individual models as inputs for a combination model can improve performance over using all available individual models as inputs in terms of accuracy and robustness, then this subset of individual models is called as an 'optimal subset'.

One of the important issues is how to select the optimal subset of individual models from all those available individual models without having to try all possible combinations of the individual models. This poses an important challenge as examining all possible combinations of individual models only provides an experimental assessment which does not have a rigorous proof from a theoretical perspective. Furthermore, trying all possible combinations would involve intensive computation and is extremely time-consuming if the total number of individual forecasting models is large. The total number of all possible combinations is $\sum_{m=2}^{M} C_M^m / \Gamma(m+1)$ excluding the individual models for one combination method if there are *M* individual candidate models available, where $C_M^m = M \times (M-1) \times (M-2) \times \cdots \times (M-m+1)$ and $\Gamma(m+1) = m \times (m-1) \times \cdots \times 2 \times 1$. For example, there are 502 possible combinations for one combination method if *M* equals nine (nine individual models in total).

Combination selection forecasting is rarely studied in the literature. Costantini and Pappalardo (2010) and Kisinbay (2010) employed the encompassing test for combination forecasts algorithms. Costantini and Pappalardo (2010) proposed a hierarchical procedure for the combination, where the procedure was investigated using short-term forecasting models for monthly industrial production in Italy. Kisinbay (2010) demonstrated that the combination forecasts algorithm outperform the benchmark model forecasts using the US macroeconomic dataset, the algorithm developed by Kisinbay (2010) was adopted to analyse US data in the IMF working paper by Baba and Kisinbay (2010).

An optimal subset selection from all individual forecasting models is studied in this paper. The optimal subset may contain one individual model, up to a maximum of all individual models. If the selected subset contains only one single model, this means that the individual model gives the best performance out of all possible combinations of individual models.

An optimal subset selection algorithm using information theory (Mackay, 2003) is proposed in this paper. The linear combination models proposed by Shen, Li and Song

(2008, 2011) and Wong et al. (2007) are used to examine the optimal subset selection algorithm for this study. The information concepts have never been applied to the selection of individual models as combination models, and all available individual models are used as inputs for the linear combination methods in tourism demand forecasting literature. For this reason, it is useful to explain the developments in information theory that contribute to forecasting.

2. Methodological issues

2.1. Information theory

Traditionally, the best single forecasting model is selected from several individual models in terms of accuracy. In most cases, the best single model may not have extracted all the information that is relevant for the actual output values. The combination models may be able to offer more information to provide a better prediction compared with an individual model. Shannon’s information theory (Mackay, 2003) argues that we can select an optimal subset of all individual models, and this subset contains enough information to forecast the actual outputs. Optimal subset selection using information theory is widely used in other fields such as the pattern recognition and neural networks fields.

Sridhar, Bartlett and Seagrave (1999) proposed an algorithm using information theory for combining neural network models. This algorithm identifies and combines useful models regardless of the nature of their relationship to the actual output. The algorithm was demonstrated through three examples including the application to a dynamic process modelling problem. The obtained results demonstrated that the algorithm could achieve highly improved performance as compared with a single optimal network or the stacked neural networks based on a linear combination of neural networks.

Many algorithms on feature selection based on mutual information (MI) were developed. The algorithm ‘mutual information based feature selection’ (MIFS) based on MI between the individual and the class variables was developed by Battiti (1994) for selecting the features in the supervised neural net learning. However this algorithm can only calculate the MI between one single variable with another single variable. Kwak and Choi (2002) analysed the limitations of the MIFS algorithm (Battiti, 1994) and proposed an ‘MI feature selection uniform information distribution’ (MIFS-U) algorithm to overcome its limitations. Both MIFS and MIFS-U algorithms can provide better performance compared with the feature selection algorithms such as principal component analysis and neural networks, and have been successfully applied in many experimental design problems. However, both algorithms involve a parameter and it is difficult to determine the range of its value.

The fixed parameter is used in the MI based feature selection ‘minimal redundancy maximal relevance’ (mRMR) algorithm (Peng, Long and Ding, 2005). The ‘normalized mutual information feature selection’ (NMIFS) algorithm was proposed in the paper (Estévez, Tesmer, Perez and Zurada, 2009) based on the normalized MI by the minimum entropy of both features. The average normalized MI is used as a measure of redundancy of the individual feature and the subset of selected features. The experiments demonstrated that the NMIFS algorithm enhances the MIFS, MIFS-U and mRMR algorithms. The parameter is also fixed in the NMIFS algorithm, which is an advantage comparing with the algorithms MIFS and MIFS-U.

In term of speeding, ‘fast correlation based filter’ (FCBF) is fast due to that a few evaluations of bivariate mutual information are computed. The FCBF is a ranking method combined with the redundancy analysis (Yu & Liu, 2004). Fleuret (2004) proposed the forward selection and ‘conditional mutual information maximization criterion’ (CMIM) in term of binary feature selection and showed that CMIM is competitive with the FCBF in selecting binary features. Meyer, Schretter and Bontempi (2008) proposed a ‘matrix of average sub-subset information for variable elimination’ (MASSIVE) using variable complementarity for microarray data sets. Their experimental results demonstrated that MASSIVE is competitive with the FCBF and CMIM, and outperforms mRMR for some data sets. All these MI feature selection algorithms are based on nominal or binary feature selection. The continuous feature can be transformed to the nominal feature by dividing the variable domain into the finite number of regions with an equal size, where the variable is assumed to be a constant within the region. It is noted that a reasonable size of data should be used in order to transform the continuous feature to the nominal feature. The Kernel-based method (Christopher, 1995) which is based on the Parzen’s window (Parzen 1962) is employed in this study. The reasons are that 1) the data set used in this study has a small sample size; 2) features (outputs of individual models) are continuous variables; 3) the data set has a low dimension of input features comparing with the microarray data.

The mutual information (MI) (Mackay, 2003) which is symmetric is a measure of the dependence between random variables. The MI is a positive value and if and only if the variables are independent with the zero MI value. The MI between two discrete random vector variables *U* and *V* is defined as follows

$$MI(U,V) = \sum_{u \in U} \sum_{v \in V} p(u,v) \log \frac{p(u,v)}{p(u)\,p(v)} \qquad (1)$$

where $p(u,v)$ is a joint density function and $p(u)$ and $p(u)$ are the marginal density functions. The *MI* between two continuous random vector variables *X* and *Y* is defined as

$$MI(X,Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p(X,Y)\log\frac{p(X,Y)}{p(X)p(Y)}dXdY \qquad (2)$$

where $p(X,Y)$ is a joint density function, and $p(X)$ and $p(Y)$ are the marginal density functions. Using the entropy concept, (1) and (2) can be written as (3) and (4) below

$$MI(U,V)=H(U)+H(V)-H(U,V) \qquad (3)$$

$$MI(X,Y)=H(X)+H(Y)-H(X,Y) \qquad (4)$$

where $H(U,V)$ and $H(U)$ are defined in (5) for discrete random vector variables and (6) for continuous random vector variables, respectively.

$$H(U,V)=-\sum_{u,v}p_{u,v}\log p_{u,v}\ ,\ \ H(U)=-\sum_{u}p_u\log p_u \qquad (5)$$

where $p_{u,v}$ is the probability when $U=u$ and $V=v$, p_u is the probability when $U=u$.

$$H(X,Y)=-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p(X,Y)\log p(X,Y)dXdY,\ \ H(X)=-\int_{-\infty}^{\infty}p(X)\log p(X)dX \qquad (6)$$

where $p(X,Y)$ is a joint density function and $p(X)$ is the marginal density function. $H(X)$ is an entropy and a measure of the amount of uncertainty associated with the value of *X*. $H(X,Y)$ is a joint entropy which measures how much entropy is contained in a joint system of two random vector variables (*X* and *Y*). We need to work out the terms $\int_{-\infty}^{\infty}p(X)\log p(X)dX$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p(X,Y)\log p(X,Y)$ in (6) in order to calculate $MI(X,Y)$ in (4). There are no analytical solutions for these terms. Thus we use approximations of these terms presented in (7) according to the definition of 'Expectation' for continuous variable.

$$\int_{-\infty}^{\infty}p(X)\log p(X)dX \approx \frac{1}{N}\sum_{i=1}^{N}\log p(X^{(i)})$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p(X,Y)\log p(X,Y)dXdY \approx \frac{1}{N}\sum_{i=1}^{N}\log p(X^{(i)},y^{(i)}) \qquad (7)$$

where *N* is the size of the data *X*={$X^{(1)},X^{(2)},...,X^{(i)}...,X^{(N)}$}, $X^{(i)}(i=1,...,N)$ is a *d* dimensional vector and *Y*={$y^{(1)},...,y^{(i)},...,y^{(N)}$}.

The Parzen's window (Parzen, 1962) with the multivariate Gaussian Kernel-based function (Bishop, 2002) is the most popular construction method for computing the density function $p(X)$ and $p(X,Y)$. $p(X^{(i)},y^{(i)})$ in (7) is one more dimension of density function $p(X^{(i)})$. $p(X^{(i)})$ and $p(X^{(i)},y^{(i)})$ can be written as (8)

$$p(X^{(i)})=\frac{1}{N}\sum_{k=1}^{N}\frac{1}{(2\pi\sigma^2)^{d/2}}\exp(-\frac{\left\|X^{(i)}-X^{(k)}\right\|^2}{2\sigma^2})$$

$$p(X^{(i)}, y^{(i)}) = \frac{1}{N}\sum_{k=1}^{N} \frac{1}{(2\pi\sigma^2)^{(d+1)/2}} \exp(-\frac{\left\|X^{(i)} - X^{(k)}\right\|^2 + (y^{(i)} - y^{(k)})^2}{2\sigma^2}) \tag{8}$$

where || || is the Euclidean norm or Euclidean distance, σ is the kernel width smoothing parameter and can be determined by the training data. For this study, the range of the kernel width is from 0.1 to 1, step by 0.05, and the best σ is selected using the training data which is used to construct the model. *X* and *Y* in (8) are the inputs and actual output, respectively. It is noticed that for large dimension *d*, the density functions $p(X^{(i)})$ and $p(X^{(i)}, y^{(i)})$ in (8) tend to zero for $\sigma > 1/\sqrt{2\pi}$, infinity for $\sigma < 1/\sqrt{2\pi}$ and constant for $\sigma = 1/\sqrt{2\pi}$. However, in this study *d* is not large comparing with the other data sets such as microarray data sets.

2.2. Forecasting Error measurement

It is essential to introduce the 'forecasting error measurement' (FEM) when measuring the performance of a forecasting model. The Mean Absolute Percentage Error (MAPE) is recommended as the most appropriate error measurement (Hanke & Reitsch, 1995; Makridakis et al., 1982) and the MAPE formula is

$$\frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

where y_t is the true value and $\hat{y}_t$ is the predicted value at time *t*, *N* is the size of time series. The Mean Absolute Scaled Error (MASE) is suggested as the best available measure of forecast accuracy (Hyndman & Koehler, 2006) and the MASE formula is

$$\frac{1}{N-1}\sum_{t=1}^{N}\frac{|y_t - \hat{y}_t|}{\sum_{i=2}^{N}|y_i - y_{i-1}|}$$

These two popular forecast accuracy measures are used in this paper. The MAPE as an example is used as FEM in the MI algorithm of optimal subset selection in section 2.3.

2.3. Mutual information (MI) algorithm of optimal subset selection

To apply the MI algorithm, we first divide the whole data set *D* into two data sets: the training data D_{Train} and the test data D_{Test}. Each column of the whole data set *D* is the outputs of each individual forecasting model, the dimension of *D* is the total number of individual forecasting models. The training data D_{Train} is used to identify an optimal subset from all individual models using the MI theory, the test data D_{Test} is used to validate the optimal subset selection results. There are two common selection methods: forward selection and backward selection (Theodoridis & Koutroumbas, 1999). Both methods accept or reject features, one at a time, in order to construct an optimal subset. Here, the forward selection is applied and the MI algorithm is described as below.

MI Algorithm for optimal subset selection:

Step 1: Set the initial selected individual model set *M* = { } which is an empty set, the initial selected data set *S* = { } which is an empty data set, the training data D_{Train} = $[F_1, F_2, ...F_j .., F_M]_{N \times M}$, which is an *N* rows and *M* columns matrix, where $F_j = (f_{j1}, f_{j2}, f_{j3}, ..., f_{jN})^{'}, (1 \le j \le M)$ is the outputs (forecasting values) of the individual model f_j $(1 \le j \le M)$ on the training data set D_{Train}, *N* is the size of the training data D_{Train} and *M* is the total number of the individual models.

Step 2: For each column (F_j) of the training data set D_{Train} using equations (4), (6), (7) and (8) calculate the *MI* between $S \cup F_j$ and the actual outputs y, which is $MI(S \cup F_j, y)$, and find the maximum *MI* value among all *j* which is $\underset{j}{Max}\, MI(S \cup F_j, y)$, where $\cup$ indicates the combine set.

Step 3: Put the individual model f_j corresponding to the maximum MI value $\underset{j}{Max}\, MI(S \cup F_j, y)$ into *M*, *M* = { f_j}, put F_j into *S*, *S*=$\{F_j\}$ and delete F_j which is column *j* from the training data D_{Train}.

Step 4: Calculate the forecasting error measurement (FEM) described in section 2.2 for the data set *S*.

Step 5: Repeat **Step 2**, **Step 3** and **Step 4** until there is non-significant improvement of the FEM value on *S* (it implies that the current FEM value is bigger or very close to the previous FEM value). Thus, M is the optimal subset which contains some individual models excluding the current individual model.

The order of the individual models for the optimal subset is determined by the MI algorithm and the size of optimal subsets is determined by the FEM. Slightly different results may be obtained if a different FEM is used as the criteria in **Step 4**.

The individual models are the foundation of applying the MI Algorithm for optimal subset selection. Several different time series approaches as the individual forecasting models are adopted in this paper. The world 'GDP' and 'CPI' and other economic factors as proxies of the influencing factors can be used as inputs if we apply causal econometric models or new emerging artificial intelligence models. However this paper concentrates on forecasting UK inbounds tourism arrivals, not on the impact study of the factors on the UK inbounds tourism arrivals. Thus, the time series are employed and the adopted individual models are described in section 2.4.

2.4. Individual forecasting model

In this study, nine individual or single time series forecasting models are used, some of these individual time series are most popular forecasting models, and some of these individual time series are newly emerging techniques. The most frequently used time series in the tourism demand forecast literature (Song & Li, 2008) are adopted in this study. The nine individual models are described in the following sections.

2.4.1. Support Vector Regression (SVR) Neural Network

The foundations of SVR neural networks were first developed by Vapnik (1995, 1996). SVR are gaining popularity due to many attractive features and their promising empirical performance in the fields such as image processing and finance etc. The research has produced promising results that have been reported by Tay and Cao (2001) and Ni and Nguyen (2007). There are also applications using SVR for tourism demand (Chen & Wang, 2007; Pai et al., 2006). The experimental results revealed that the proposed models outperform the Autoregressive Integrated Moving Average (ARIMA) approaches.

In SVR, the training data (used to construct a forecasting model) is a subset of the whole available data and is considered as a set of pairs $(X^{(1)}, y^{(1)}),\ldots, (X^{(i)}, y^{(i)}),\ldots,$ $(X^{(N)}, y^{(N)})$ where $X^{(i)} \subset R^m$ denote the input space (*m* is the width or dimension of the inputs) and $y^{(i)} \subset R$ denote the corresponding actual target value for $i = 1,2,\ldots,N$, where *N* is the size of the training data set. For this study, $X^{(i)} = \{y_{t-1}\, y_{t-2} \ldots y_{t-m}\}^m \subset R^m$ are the vectors of the historical tourism demand observations at time *t* where *t* = *m*+1, *m*+2,…, *N* and *i* = *t-m*, and $y^{(i)} = y_t \subset R$ are the actual target values at time *t*. For example, $X^{(1)} = \{y_4\, y_3\, y_2\, y_1\}$, $X^{(2)} = \{y_5\, y_4\, y_3\, y_2\}$ and corresponding target values $y^{(1)} = y_5$, $y^{(2)} = y_6$ for *m* = 4. The purpose of the regression problem is to determine a function that can predict future values accurately. The generic SVR forecasting model with forecasting value $\hat{y}_t$ has the following general form

$$\hat{y}_t = f(X) = (W \cdot \Phi(X)) + b \qquad (9)$$

where *X* has the form $X^{(i)}$, $W \subset R^m, b \subset R$ are the best weights and base to be determined using the training data set, Φ denotes a nonlinear transformation from R^m to a high dimensional space and $\hat{y}_t$ is the forecasting value of y_t. The goal of SVR is to find the best values of W and b in (9) such that the nonlinear model (9) can be best fitted with the input data *X* and the output data y_t. The best values of W and b in (9) can be determined by the training data.

The data used in this paper is the quarterly data, thus the previous one year (inputs width *m* = 4), a year and a quarter (*m* = 5) up to the previous two years (*m* = 8) are used as inputs to construct the five different time series, respectively. The SVR model is generated using MATLAB (Version 2011b) software.

2.4.2. ARIMA Model

The Box-Jenkins forecasting time series model - ARIMA proposed by Box and Jenkins (1970) has become widely used in many fields for time series analysis including tourism demand forecasting (Chu, 2008). The quarterly inbound UK tourism arrivals data which has a seasonal time series feature is used in this study, thus the Seasonal ARIMA ARIMA(p,d,q)(P,D,Q)$_s$ with period *s* (*s*=4) is applied here due to the quarterly data. The ARIMA(p,d,q)(P,D,Q)$_s$ model is as follows

$$\phi_p(B)\Phi_P(B^s)[(1-B)^d(1-B^s)^D\hat{y}_t-\mu]=\theta_q(B)\Theta_Q(B^s)a_t \tag{10}$$

where *B* is a backward shift operator with $By_t = y_{t-1}$ and $Ba_t = a_{t-1}$. $\hat{y}_t$ is the value to be forecasted and a_t is the residual at time period *t*, μ is the overall mean of series which is a constant. $\phi_p(B)=1-\varphi_1 B-\varphi_2 B^2-...-\varphi_p B^p$ is a non-seasonal auto-regression of order *p*, $\theta_q(B)=1-\vartheta_1 B-\vartheta_2 B^2-...-\vartheta_q B^q$ is a non-seasonal moving average of order *q*, $\Phi_P(B^s)=1-\Phi_1 B^s-\Phi_2 B^{s2}-...-\Phi_P B^{sP}$ is a seasonal auto-regression of order *P*, $\Theta_Q(B^s)=1-\Theta_1 B^s-\Theta_2 B^{s2}-...-\Theta_Q B^{sQ}$ is a seasonal moving average polynomial of order *Q*. The value $\hat{y}_t$ in equation (10) is a forecasting value. The best fitted seasonal ARIMA model can be automatically generated using SPSS (Version 19) software.

2.4.3. Winters' Multiplicative Exponential Smoothing Model

The Winters' multiplicative exponential smoothing (Douglas, Lynwood & John, 1990) model is a popular time series forecasting method. Multiplicative decomposition considers the effects of seasonality to be multiplicative, which is, growing (or decreasing) over time. The model is presented in (11) below

$$\hat{y}_t = T_t \times S_t \times L_t + \varepsilon_t \tag{11}$$

where $\hat{y}_t$ is the forecasting at time t, T_t represents the trend component, S_t represents the seasonality and L_t is the long term cycles and ε_t is the error. This method requires at least two years of back data for forecasting. The Winter's additive exponential smoothing model is not considered as an individual model in this study, because it is a special case of the Seasonal ARIMA. The best fitted Winters' multiplicative exponential smoothing can be automatically generated using SPSS (Version 19) software.

2.4.4. Naïve 1 and Naïve 2 models

The Naïve 1 and Naïve 2 models (Chu, 2004; Oh & Morzuch, 2005) which are very popular models in tourism demand forecasting are adopted in this paper. A Naïve method simply states that future forecasts are equal to the most recently available value. The Naïve 1 model operates on the assumption that the number of tourists at time *t*, $\hat{y}_t$ is the same as the value at time *t*-4 denoted by y_{t-4} and is described as $\hat{y}_t = y_{t-4}$.

The Naïve 2 model operates on the assumption that the number of tourists at time *t*, $\hat{y}_t$ is equal to the value at time *t*-4 multiplied by a modification factor which includes the influence of the y_{t-8} (long range value) and can be written as $\hat{y}_t = y_{t-4}\{1+(y_{t-4}-y_{t-8})/y_{t-8}\}$.

2.5. Combination forecasting model

In general, there are three linear combination methods available in the literature for tourism demand. These three linear combination methods studied by Shen, Li and Song (2008) and Wong et al. (2007) are evaluated in this study. They are Simple Average (SA), Variance Covariance (VACO), and Discounted Mean Square Forecast Error (DMSFE) methods. Four individual models as inputs with one-step-ahead forecasting on these three linear combinations were evaluated by Wong et al. (2007), and seven individual models as inputs with multiple-step-ahead forecasting horizons were examined for these three linear combinations by Shen, Li and Song (2008).

The SA combination method can be expressed as $\hat{Y}_t^C = \sum_{j=1}^{M} w_j \hat{y}_t^{(j)}$ where $w_j = 1/M$, $\hat{y}_t^{(j)}$ is the forecast value (output) from the *j*th single forecasting model and $\hat{Y}_t^C$ is the combined forecast model at time *t*, *M* is the total number of individual forecasting models. In this simple average combination, each individual forecasting model makes an equal contribution (same weight) to the combined value $\hat{Y}_t^C$ with $\sum_{j=1}^{M} w_j = 1$. The VACO combination form is the same as the SA form, but the weight w_j is defined as $w_j = (\sum_{i=1}^{N}(y_i - \hat{y}_i^{(j)})^2)^{-1} \big/ \sum_{j=1}^{M}(\sum_{i=1}^{N}(y_i - \hat{y}_i^{(j)})^2)^{-1}$, where y_i is the *ith* true target value of the training data set, and $\hat{y}_i^{(j)}$ is the *ith* forecasting value of the training data set from the *j*th individual forecasting model. *N* is the total number of the training data set. The DMSFE combination form is the same as the form of SA, but the weight w_j is defined as

$w_j = (\sum_{i=1}^{N} \beta^{N-i+1} (y_i - \hat{y}_i^{(j)})^2)^{-1} / \sum_{j=1}^{M} (\sum_{i=1}^{N} \beta^{N-i+1} (y_i - \hat{y}_i^{(j)})^2)^{-1}$. The VACO method is a special case of the DMSFE when β = 1. β is chosen as 0.95, 0.9, 0.85 and 0.8, respectively which is the same as in the papers (Li, Song & Witt, 2005; Song & Li, 2008) in this study. It is noted that w_j computed in VACO and DMSFE also satisfies the constraint $\sum_{j=1}^{M} w_j = 1$.

All the above combination methods, SA, VACO and DMSFE, are linear combinations as discussed by Shen, Li and Song (2008) and Wong et al (2007). The advantage of these combinations is that they are simple and easy to apply. The parameters (weights) are fixed and are easy to calculate from the data set.

3. Experiment

3.1. Data set

The International Passenger Survey is available from the Office for National Statistics, UK and provides information on UK tourism arrivals and expenditure according to country of origin and purpose of visit. Tourists passing through passport control are randomly selected for interview. The results are based on face-to-face interviews with samples of passengers as they enter or leave the UK. Quarterly (Q) UK inbound visit numbers for Q1 1993 to Q4 2007 extracted from the IPS are used for this study, since the financial and economic crises began in 2008. The tourism industry is affected predominantly by the factors which are weather effect, festival effect, calendar effect in both the origin and destination countries (Lim, 2001). Figure 1 shows that arrivals for holiday and study purposes have a high degree of seasonality, arrivals for business purpose has least degree of seasonality, and the degree of seasonality for arrivals of visit friend/relatively (VFR) purpose is in the middle of holiday/study and business purposes.

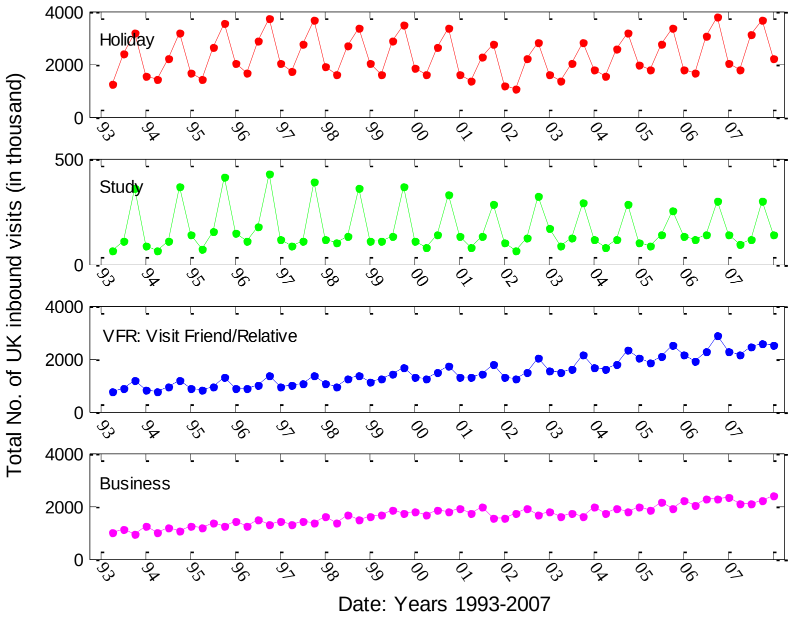

Figure 1: Characteristics of the data at different purposes

It is imperative to test for the presence of unit roots and seasonal unit roots in univariate series. The commonly used unit-root tests are the Augmented Dickey–Fuller (ADF) test (Dickey & Fuller, 1979), the Phillips-Perron (PP) test (Phillips & Perron, 1988), and the Hylleberg-Engle-Granger-Yoo (HEGY) test (Hylleberg et al., 1990) for a hypothesis of a seasonal unit-root which determines the nature of seasonal variation in the series. For examples, the ADF test is applied by Goh & Law (2002) and the PP test is applied by Gounopoulos, Petmezas and Santamaria (2012). The hypothesis ADF, PP and HEGY tests are presented in formula (12), and the results of the ADF, PP (using Eview) and HEGY (using R) are illustrated in Table 1.

$$\begin{cases} \text{ADF test}: X_t = \mu + \gamma t + \beta X_{t-1} + \alpha_1 \Delta X_{t-1} + \alpha_2 \Delta X_{t-2} + \ldots + \alpha_p \Delta X_{t-p} + e_t \\ \text{PP test}: X_t = \mu + \gamma t + \beta X_{t-1} + e_t \\ \text{HEGY test}:\ (1-L^4) y_t = \pi_1 z_{1,t-1} + \pi_2 z_{2,t-1} + \pi_3 z_{3,t-2} + \pi_4 z_{3,t-1} + e_t \end{cases} \qquad (12)$$

where $e_t \sim N(0,\sigma^2)$, $z_{1,t} = (1 + L + L^2 + L^3) y_t$, $z_{2,t} = -(1 - L + L^2 - L^3) y_t$, $z_{3,t} = -(1 - L^2) y_t$.

ADF and PP tests:

$$H_0 : \beta = 1 \text{(series has a unit root)}$$
$$H_1 : \beta < 1 \text{(series has no unit root)}$$

HEGY test (Ghysels, Lee & Noh, 1994):

$$H_0 : \pi_2 = \pi_3 = \pi_4 = 0 \text{(series has a seasonal unit root)}$$
$$H_1 : \pi_2 < 0, \pi_3 < 0, \pi_4 \neq 0 \text{(series has no seasonal unit root)}$$

Table 1: ADF, PP and HEGY tests for unit-root/seasonal unit-root

	ADF	PP	ADF	PP
Visit purpose	Level		First difference	
Holiday	p=0.4447	p=0.0000***	p=0.0216***	p=0.0001***
Study	p=0.1932	p=0.0000***	p=0.0003***	p=0.0001***
VFR	p=0.9999	p=0.2302	p=0.0001***	p=0.0001***
Business	p=0.8967	p=0.3995	p=0.0000***	p=0.0001***
	HEGY			
Visit purpose	Intercept & Seasonal dummies		Intercept & Trend & Seasonal dummies	
Holiday	t test: π_2 p=0.01*** F test: $\pi_3 \cap \pi_4$ p=0.01***		t test: π_2 p=0.01*** F test: $\pi_3 \cap \pi_4$ p=0.01***	
Study	t test: π_2 p=0.1* F test: $\pi_3 \cap \pi_4$ p=0.1*		t test: π_2 p=0.1* F test: $\pi_3 \cap \pi_4$ p=0.07*	
VFR	t test: π_2 p=0.1* F test: $\pi_3 \cap \pi_4$ p=0.05*		t test: π_2 p=0.1* F test: $\pi_3 \cap \pi_4$ p=0.064*	
Business	t test: π_2 p=0.032** F test: $\pi_3 \cap \pi_4$ p=0.046**		t test: π_2 p=0.028** F test: $\pi_3 \cap \pi_4$ p=0.01***	

***, **, *: Statistical significant difference at 1%, 5% and 10% level, respectively;

The time series is nonstationary if H_0 is accepted, which has a unit root or seasonal unit root. Otherwise, it is stationary. The ADF and PP test results in Table 1 show that some series have unit roots at level. However there is no unit root (H_1 is accepted at 1% significant level) with the first difference in all cases as expected. The rejection of H_0 for the HEGY test means that the series does not have a seasonal unit root. The test results support the application of Box–Jenkin model—Seasonal ARIMA in this study.

3.2. Framework

One to four quarters ahead forecasting from individual models that are described in the previous section are used for the optimal subset selection using the MI algorithm in this paper. The same process with the paper (Shen, Li & Song, 2008) is used for individual models generated here.

The individual forecasting models are constructed based on the data from Q1 1993 to Q4 1997 inclusive (training data). The out-of-sample forecasts (test data) are generated for Q1 1998 to Q4 2007 inclusive with one to four quarters ahead forecasting using the following recursive forecasting techniques.

Recursive forecasting:

1) Forecast one to four quarter ahead (Q1 1998 to Q4 1999) using the initial training data (Q1 1993 to Q4 1997)

2) Forecast one to four quarter ahead (Q2 1998 to Q1 2000) using the enhanced training data (Q1 1993 to Q1 1998) by adding one data point (Q1 1998) to the training data set
3) Continue step 2) until forecast the last point of test data set (Q4 2007) using the enhanced training data (Q1 1993 to Q3 2007)

The results from this process are 40 one quarter ahead forecasting values for each individual forecasting model, 39 two quarters ahead forecasting values, 38 three quarters ahead forecasting values and 37 four quarters ahead forecasting values that are generated from each individual model.

There are nine individual models in total for this study, five SVR with different dimensions of inputs from 4 to 8, Naïve 1, Naïve 2, Seasonal ARIMA and Winters' Multiplicative Exponential Smoothing models. There are 40 (Q1 1998- Q4 2007), 39 (Q2 1998- Q4 2007), 38 (Q3 1998- Q4 2007) and 37 (Q4 1998- Q4 2007) one to four quarters ahead forecasting values that are generated from each individual forecasting model. The first 24 (Q1 1998- Q4 2003), 23 (Q2 1998- Q4 2003), 22 (Q3 1998- Q4 2003) and 21(Q4 1998- Q4 2003) forecasting values (training data) from all individual forecasting models are used to select an optimal subset using the MI algorithm. The period from Q1 2004 to Q4 2007 (test data) is used to test this selected optimal subset. The framework of this case study is illustrated in the following (Figure 2).

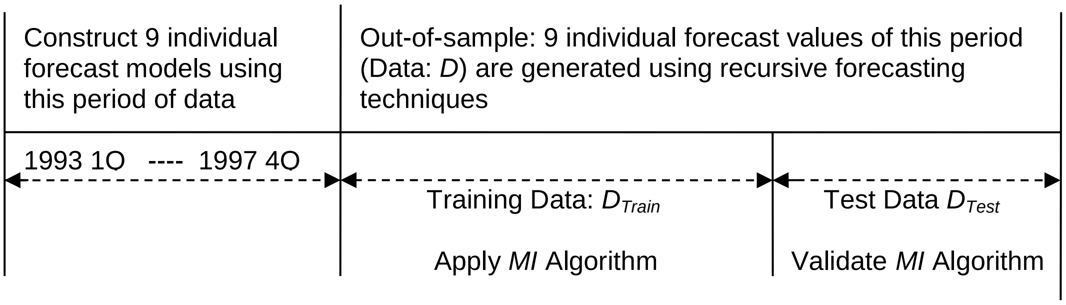
Construct 9 individual forecast models using this period of data
Out-of-sample: 9 individual forecast values of this period (Data: *D*) are generated using recursive forecasting techniques
1993 1O ---- 1997 4O
Training Data: D_{Train}
Test Data D_{Test}
Apply *MI* Algorithm
Validate *MI* Algorithm

Figure 2: Framework

3.3. Experimental results

Next, the optimal subsets from these nine individual models are selected by applying the MI algorithm using the training data. The MAPE values of the optimal subsets for the period from Q1 2004 to Q4 2007 inclusive at the different purpose of visits are presented in Table 2. The MASE values of the same optimal subsets as in Table 2 for the period from Q1 2004 to Q4 2007 inclusive at the different purpose of visits are presented in Table 3. For the simplicity, the mean of MAPE values for all six linear combination methods is used as the criteria in this case study.

Table 2: MAPE values of optimal subsets for different linear combination methods

Test (Ex-post) period: Q1 2004 to Q4 2007 inclusive								

Individual Model ID:	1: SVR (input dimension=4); 2: SVR (input dimension=5);
	3: SVR (input dimension=6); 4: SVR (input dimension=7);
	5: SVR (input dimension=8); 6: Naïve 1; 7: Naïve 2; 8: SARIMA; 9: WMES							

	Holiday				**Study**			
	1Q	**2Q**	**3Q**	**4Q**	**1Q**	**2Q**	**3Q**	**4Q**
Optimal Subset	[9,3,7,8]	[3,9]	[9,3,7]	[2,3,7,9]	[1,7,3,9]	[3,7,8,9]	[1,7,3]	[1,7,8,3]
	MAPE Value of Optimal Subset				MAPE Value of Optimal Subset			
SA	5.75	6.87	8.20	8.38	10.35	11.25	10.21	10.14
VACO	4.73	6.86	8.10	8.35	9.34	10.50	9.29	9.38
MSFE(β=0.95)	4.73	6.87	8.10	8.35	9.32	10.42	9.25	9.27
MSFE(β=0.9)	4.76	6.87	8.09	8.35	9.35	10.37	9.23	9.20
MSFE(β=0.85)	4.82	6.88	8.07	8.34	9.36	10.33	9.22	9.18
MSFE(β=0.8)	4.88	6.88	8.09	8.33	9.37	10.30	9.21	9.16
	VFR				**Business**			
	1Q	**2Q**	**3Q**	**4Q**	**1Q**	**2Q**	**3Q**	**4Q**
Optimal Subset	[9,7,5,8]	[9,7,5]	[9,7,5,1]	[2,9,7,1]	[7,9,8,1]	[7,9,8]	[7,9,8]	[4,7,8]
	MAPE Value of Optimal Subset				MAPE Value of Optimal Subset			
SA	4.89	5.14	5.51	5.84	4.57	4.78	5.14	6.61
VACO	4.69	4.95	5.33	5.85	3.99	4.51	4.82	6.58
MSFE(β=0.95)	4.68	4.95	5.33	5.84	3.97	4.51	4.87	6.61
MSFE(β=0.9)	4.68	4.96	5.33	5.83	3.96	4.58	4.92	6.77
MSFE(β=0.85)	4.69	4.97	5.34	5.81	3.95	4.64	4.98	6.93
MSFE(β=0.8)	4.71	4.99	5.35	5.78	3.94	4.68	5.04	7.04

Note: optimal subset [9, 3, 7, 8] means that the ID numbers 9, 3, 7 and 8 of individual models are selected as optimal subset and used in the combination model. Q=quarter; SARIMA: Seasonal ARIMA; WMES: Winters' multiplicative exponential smoothing

Table 3: MASE values of optimal subsets for different linear combination methods

Test (Ex-post) period: Q1 2004 to Q4 2007 inclusive								
	Holiday				**Study**			
	1Q	**2Q**	**3Q**	**4Q**	**1Q**	**2Q**	**3Q**	**4Q**
Optimal Subset	[9,3,7,8]	[3,9]	[9,3,7]	[2,3,7,9]	[1,7,3,9]	[3,7,8,9]	[1,7,3]	[1,7,8,3]
	MASE Value of Optimal Subset				MASE Value of Optimal Subset			
SA	0.1554	0.1827	0.2143	0.2154	0.1617	0.1676	0.1602	0.1560
VACO	0.1296	0.1825	0.2124	0.2158	0.1479	0.1556	0.1469	0.1490
MSFE(β=0.95)	0.1292	0.1826	0.2125	0.2164	0.1480	0.1551	0.1476	0.1485
MSFE(β=0.9)	0.1300	0.1827	0.2121	0.2171	0.1485	0.1548	0.1482	0.1484
MSFE(β=0.85)	0.1313	0.1828	0.2115	0.2180	0.1487	0.1547	0.1486	0.1486
MSFE(β=0.8)	0.1328	0.1829	0.2118	0.2190	0.1489	0.1546	0.1490	0.1487
	VFR				**Business**			
	1Q	**2Q**	**3Q**	**4Q**	**1Q**	**2Q**	**3Q**	**4Q**
Optimal Subset	[9,7,5,8]	[9,7,5]	[9,7,5,1]	[2,9,7,1]	[7,9,8,1]	[7,9,8]	[7,9,8]	[4,7,8]
	MASE Value of Optimal Subset				MASE Value of Optimal Subset			
SA	0.3657	0.3755	0.3988	0.4325	0.5806	0.6058	0.6460	0.8237
VACO	0.3516	0.3611	0.3834	0.4316	0.5089	0.5722	0.6066	0.8265
MSFE(β=0.95)	0.3511	0.3612	0.3832	0.4311	0.5056	0.5731	0.6123	0.8331
MSFE(β=0.9)	0.3511	0.3615	0.3833	0.4303	0.5050	0.5808	0.6188	0.8550
MSFE(β=0.85)	0.3516	0.3618	0.3837	0.4289	0.5041	0.5877	0.6261	0.8768
MSFE(β=0.8)	0.3526	0.3631	0.3848	0.4271	0.5030	0.5931	0.6338	0.8919

Note: optimal subset [9,3,7,8] means that the ID numbers 9, 3, 7 and 8 of individual models are selected as optimal subset and used in the combination model. Q=quarter

In order to validate the optimal subset selection approach, we compare the optimal subset selection results (MAPE and MASE values) with the results of the different linear combination methods using the test data for all possible combinations of *m* individual models ($2 \leq m \leq 9$). The MAPE and MASE values of all possible combinations with the best MAPE of the individual models are presented in Tables 4 and 5, respectively. The MAPE and MASE values of individual models are illustrated in Table A1 of Appendix.

Table 4: Best MAPE values for all possible combinations of *m* individual models on different combination methods

Test (Ex-post) period (Q1 2004 to Q4 2007 inclusive)																
m models	*m*=2	*m*=3	*m*=4	*m*=5	*m*=6	*m*=7	*m*=8	*m*=9	*m*=2	*m*=3	*m*=4	*m*=5	*m*=6	*m*=7	*m*=8	*m*=9
Total number of all possible combinations	(36)	(84)	(126)	(126)	(84)	(36)	(9)	(1)	(36)	(84)	(126)	(126)	(84)	(36)	(9)	(1)
Combine Model	**Holiday 1Q** (best MAPE of individual model is 6.39)								**Holiday 2Q** (best MAPE of individual model is 8.34)							
SA	5.17	4.79	**4.84**	5.13	5.48	5.64	5.83	6.20	**6.87**	7.13	6.83	6.96	7.07	7.35	7.58	7.76
VACO	5.22	4.98	**4.73**	4.91	5.09	5.21	5.41	5.81	**6.86**	7.07	6.94	6.88	6.96	7.26	7.50	7.77
MSFE(β =0.95)	5.22	4.97	**4.73**	4.91	5.10	5.22	5.42	5.85	**6.87**	7.11	6.90	6.90	6.99	7.32	7.57	7.88
MSFE(β =0.9)	5.20	4.96	**4.76**	4.95	5.14	5.27	5.49	5.95	**6.87**	7.14	6.85	6.89	7.03	7.39	7.66	8.01
MSFE(β =0.85)	5.17	4.94	**4.82**	5.01	5.20	5.34	5.61	6.12	**6.88**	7.02	6.80	6.88	7.06	7.47	7.80	8.15
MSFE(β =0.8)	5.19	4.91	**4.88**	5.07	5.27	5.41	5.79	6.32	**6.88**	6.86	6.75	6.90	7.10	7.57	7.97	8.30
	Holiday 3Q (best MAPE of individual model is 8.52)								**Holiday 4Q** (best MAPE of individual model is 8.81)							
SA	8.16	**7.99**	7.84	7.90	8.00	8.08	8.26	8.42	**8.18**	8.15	8.21	8.29	8.25	8.29	8.35	8.51
VACO	8.11	**7.89**	7.80	7.88	8.01	8.10	8.32	8.53	**8.06**	8.19	8.28	8.22	8.25	8.30	8.37	8.60
MSFE(β =0.95)	8.11	**7.88**	7.81	7.89	8.02	8.11	8.35	8.56	**8.10**	8.17	8.28	8.25	8.24	8.31	8.40	8.62
MSFE(β =0.9)	8.12	**7.86**	7.84	7.91	8.04	8.13	8.37	8.59	**8.13**	8.16	8.26	8.25	8.27	8.33	8.43	8.64
MSFE(β =0.85)	8.12	**7.83**	7.87	7.94	8.06	8.15	8.41	8.62	**8.15**	8.14	8.23	8.23	8.28	8.34	8.47	8.68
MSFE(β =0.8)	8.14	**7.84**	7.90	7.97	8.08	8.16	8.45	8.65	**8.18**	8.13	8.22	8.22	8.30	8.38	8.51	8.71
	Study 1Q (best MAPE of individual model is 10.67)								**Study 2Q** (best MAPE of individual model is 10.42)							
SA	9.36	9.26	**9.33**	9.47	9.62	9.66	9.52	9.71	10.29	**10.02**	10.22	10.33	10.63	10.87	11.03	11.27
VACO	9.44	9.29	**9.34**	9.45	9.39	9.51	9.63	9.81	10.32	**10.00**	10.33	10.34	10.51	10.74	11.03	11.27
MSFE(β =0.95)	9.43	9.28	**9.32**	9.42	9.44	9.54	9.66	9.84	10.29	**10.01**	10.28	10.34	10.45	10.72	11.04	11.29
MSFE(β =0.9)	9.43	9.28	**9.29**	9.38	9.47	9.55	9.69	9.88	10.29	**10.00**	10.22	10.36	10.42	10.73	11.06	11.32
MSFE(β =0.85)	9.44	9.28	**9.24**	9.33	9.49	9.56	9.70	9.90	10.30	**10.00**	10.17	10.39	10.44	10.74	11.07	11.36
MSFE(β =0.8)	9.43	9.27	**9.21**	9.31	9.49	9.58	9.71	9.92	10.31	**10.00**	10.14	10.40	10.45	10.74	11.09	11.39
	Study 3Q (best MAPE of individual model is 9.87)								**Study 4Q** (best MAPE of individual model is 9.68)							
SA	8.87	9.03	**8.93**	9.12	9.21	9.26	9.54	9.59	**8.96**	8.84	8.94	8.97	9.07	9.19	9.46	9.58
VACO	8.90	8.98	**8.89**	9.08	9.18	9.24	9.49	9.59	**8.90**	8.84	8.85	8.96	9.05	9.16	9.38	9.59
MSFE(β =0.95)	8.93	8.94	**8.87**	9.00	9.16	9.23	9.43	9.60	**8.81**	8.78	8.90	8.94	9.03	9.14	9.33	9.60
MSFE(β =0.9)	8.92	8.91	**8.86**	8.95	9.15	9.21	9.38	9.61	**8.73**	8.74	8.87	8.92	9.01	9.13	9.29	9.61
MSFE(β =0.85)	8.91	8.89	**8.85**	8.91	9.14	9.21	9.35	9.62	**8.67**	8.83	8.83	8.93	9.00	9.12	9.26	9.62
MSFE(β =0.8)	8.91	8.88	**8.84**	8.89	9.13	9.20	9.33	9.63	**8.63**	8.84	8.80	8.93	8.99	9.11	9.24	9.63
	VFR 1Q (best MAPE of individual model is 5.47)								**VFR 2Q** (best MAPE of individual model is 5.69)							
SA	4.84	**4.84**	4.89	4.95	5.03	5.10	5.16	5.39	4.79	**4.87**	4.92	5.05	5.12	5.23	5.37	5.42
VACO	4.84	**4.53**	4.66	4.75	4.86	4.95	5.02	5.19	4.99	**4.82**	4.72	4.83	4.92	4.98	5.07	5.24

MSFE(β =0.95)	4.84	**4.53**	4.66	4.75	4.85	4.94	5.02	5.18	5.00	**4.80**	4.75	4.85	4.94	4.99	5.09	5.24
MSFE(β =0.9)	4.84	**4.54**	4.67	4.76	4.87	4.95	5.02	5.18	5.00	**4.75**	4.81	4.88	4.96	5.02	5.14	5.26
MSFE(β =0.85)	4.84	**4.56**	4.69	4.80	4.89	4.97	5.05	5.19	4.95	**4.78**	4.82	4.92	4.99	5.08	5.21	5.27
MSFE(β =0.8)	4.84	**4.59**	4.70	4.83	4.92	5.00	5.08	5.23	4.87	**4.81**	4.85	4.97	5.03	5.14	5.25	5.31
	VFR 3Q (best MAPE of individual model is 5.94)								**VFR 4Q** (best MAPE of individual model is 5.86)							
SA	5.80	5.46	**5.43**	5.48	5.57	5.63	5.72	5.73	6.02	5.86	**5.78**	5.82	5.86	5.92	6.11	6.21
VACO	5.84	5.46	**5.32**	5.36	5.41	5.47	5.53	5.70	5.96	5.95	**5.85**	5.83	5.87	5.92	6.04	6.22
MSFE(β =0.95)	5.83	5.45	**5.32**	5.36	5.41	5.48	5.54	5.68	5.95	5.94	**5.84**	5.82	5.86	5.92	6.04	6.20
MSFE(β =0.9)	5.83	5.42	**5.32**	5.37	5.43	5.49	5.55	5.67	5.94	5.92	**5.83**	5.81	5.85	5.91	6.03	6.17
MSFE(β =0.85)	5.81	5.40	**5.32**	5.38	5.46	5.51	5.58	5.65	5.91	5.90	**5.79**	5.79	5.84	5.92	6.04	6.13
MSFE(β =0.8)	5.82	5.39	**5.34**	5.39	5.47	5.55	5.61	5.64	5.88	5.81	**5.76**	5.77	5.86	5.95	6.03	6.09
	Business 1Q (best MAPE of individual model is 3.94)								**Business 2Q** (best MAPE of individual model is 4.39)							
SA	4.13	4.10	**4.18**	4.28	4.20	4.19	4.33	4.43	4.68	**4.78**	4.76	5.03	5.24	5.41	5.60	5.76
VACO	4.22	4.06	**3.99**	4.06	4.21	4.38	4.51	4.71	4.65	**4.51**	4.98	5.34	5.56	5.70	5.86	6.03
MSFE(β =0.95)	4.22	4.06	**3.97**	4.07	4.23	4.40	4.53	4.72	4.63	**4.51**	5.07	5.42	5.62	5.77	5.93	6.09
MSFE(β =0.9)	4.20	4.05	**3.96**	4.08	4.24	4.41	4.54	4.74	4.59	**4.58**	5.16	5.48	5.67	5.81	5.97	6.13
MSFE(β =0.85)	4.16	4.04	**3.95**	4.08	4.24	4.41	4.55	4.75	4.54	**4.64**	5.22	5.53	5.71	5.84	6.01	6.16
MSFE(β =0.8)	4.13	4.02	**3.94**	4.07	4.24	4.41	4.55	4.75	4.48	**4.68**	5.25	5.56	5.73	5.86	6.02	6.18
	Business 3Q (best MAPE of individual model is 4.58)								**Business 4Q** (best MAPE of individual model is 5.01)							
SA	**4.52**	4.98	5.33	5.68	6.03	6.28	6.52	6.75	**5.69**	5.96	6.17	6.26	6.39	6.68	6.93	7.18
VACO	**4.56**	4.82	5.40	6.02	6.46	6.79	7.06	7.27	**6.07**	6.25	6.27	6.44	6.86	7.19	7.44	7.64
MSFE(β =0.95)	**4.55**	4.87	5.48	6.17	6.61	6.94	7.20	7.39	**6.09**	6.32	6.34	6.57	7.02	7.33	7.59	7.78
MSFE(β =0.9)	**4.54**	4.92	5.56	6.27	6.73	7.06	7.31	7.49	**6.10**	6.37	6.39	6.70	7.15	7.45	7.70	7.89
MSFE(β =0.85)	**4.52**	4.98	5.63	6.34	6.82	7.15	7.38	7.55	**6.09**	6.40	6.40	6.76	7.24	7.53	7.78	7.97
MSFE(β =0.8)	**4.50**	5.04	5.69	6.38	6.87	7.21	7.43	7.59	**6.06**	6.42	6.40	6.81	7.30	7.58	7.84	8.01

Table 5: Best MASE values for all possible combinations of *m* individual models on different combination methods

	Test (Ex-post) period (Q1 2004 to Q4 2007 inclusive)															
m models	*m*=2	*m*=3	*m*=4	*m*=5	*m*=6	*m*=7	*m*=8	*m*=9	*m*=2	*m*=3	*m*=4	*m*=5	*m*=6	*m*=7	*m*=8	*m*=9
Total number of all possible combinations	(36)	(84)	(126)	(126)	(84)	(36)	(9)	(1)	(36)	(84)	(126)	(126)	(84)	(36)	(9)	(1)
Combine Model	**Holiday 1Q** (best MASE of individual model is 0.1595)								**Holiday 2Q** (best MASE of individual model is 0.2206)							
SA	0.1356	0.1249	**0.1262**	0.1340	0.1434	0.1494	0.1530	0.1619	**0.1827**	0.1844	0.1803	0.1844	0.1867	0.1923	0.1975	0.2024
VACO	0.1355	0.1301	**0.1263**	0.1285	0.1335	0.1384	0.1425	0.1518	**0.1825**	0.1892	0.1816	0.1827	0.1838	0.1900	0.1965	0.2024
MSFE(β =0.95)	0.1355	0.1300	**0.1264**	0.1291	0.1341	0.1386	0.1427	0.1527	**0.1826**	0.1882	0.1810	0.1830	0.1846	0.1913	0.1982	0.2054

MSFE(β=0.9)	0.1355	0.1292	**0.1265**	0.1304	0.1360	0.1395	0.1440	0.1551	**0.1827**	0.1854	0.1806	0.1826	0.1856	0.1929	0.2003	0.2090
MSFE(β=0.85)	0.1355	0.1287	**0.1269**	0.1322	0.1386	0.1409	0.1469	0.1598	**0.1828**	0.1824	0.1799	0.1820	0.1866	0.1948	0.2041	0.2130
MSFE(β=0.8)	0.1368	0.1280	**0.1287**	0.1342	0.1400	0.1424	0.1514	0.1650	**0.1829**	0.1795	0.1792	0.1827	0.1876	0.1968	0.2083	0.2171
	Holiday 3Q (best MASE of individual model is 0.2351)								**Holiday 4Q** (best MASE of individual model is 0.2437)							
SA	0.2169	**0.2095**	0.2067	0.2064	0.2098	0.2104	0.2158	0.2138	0.2142	0.2154	**0.2136**	0.2145	0.2169	0.2197	0.2235	0.2138
VACO	0.2142	**0.2078**	0.2079	0.2080	0.2084	0.2118	0.2184	0.2113	0.2149	0.2156	**0.2128**	0.2156	0.2182	0.2210	0.2263	0.2113
MSFE(β=0.95)	0.2142	**0.2077**	0.2079	0.2085	0.2088	0.2121	0.2190	0.2121	0.2147	0.2152	**0.2128**	0.2157	0.2186	0.2217	0.2269	0.2121
MSFE(β=0.9)	0.2142	**0.2071**	0.2079	0.2086	0.2095	0.2127	0.2200	0.2127	0.2143	0.2146	**0.2127**	0.2163	0.2192	0.2227	0.2278	0.2127
MSFE(β=0.85)	0.2142	**0.2064**	0.2086	0.2089	0.2104	0.2134	0.2210	0.2133	0.2140	0.2138	**0.2126**	0.2163	0.2198	0.2239	0.2288	0.2133
MSFE(β=0.8)	0.2143	**0.2062**	0.2090	0.2092	0.2114	0.2143	0.2219	0.2139	0.2137	0.2145	**0.2126**	0.2164	0.2211	0.2253	0.2299	0.2139
	Study 1Q (best MASE of individual model is 0.1578)								**Study 2Q** (best MASE of individual model is 0.1596)							
SA	0.1466	0.1468	**0.1484**	0.1507	0.1526	0.1520	0.1508	0.1534	0.1546	**0.1532**	0.1552	0.1572	0.1602	0.1614	0.1631	0.1659
VACO	0.1480	0.1471	**0.1475**	0.1488	0.1492	0.1514	0.1531	0.1554	0.1541	**0.1527**	0.1545	0.1573	0.1576	0.1601	0.1635	0.1674
MSFE(β=0.95)	0.1479	0.1466	**0.1463**	0.1486	0.1499	0.1518	0.1535	0.1559	0.1532	**0.1523**	0.1523	0.1567	0.1571	0.1600	0.1636	0.1678
MSFE(β=0.9)	0.1480	0.1468	**0.1458**	0.1483	0.1503	0.1521	0.1538	0.1564	0.1528	**0.1520**	0.1519	0.1561	0.1569	0.1602	0.1639	0.1683
MSFE(β=0.85)	0.1479	0.1466	**0.1458**	0.1480	0.1506	0.1523	0.1541	0.1567	0.1527	**0.1518**	0.1516	0.1558	0.1572	0.1603	0.1642	0.1689
MSFE(β=0.8)	0.1476	0.1465	**0.1457**	0.1479	0.1506	0.1524	0.1543	0.1570	0.1527	**0.1518**	0.1514	0.1557	0.1573	0.1604	0.1644	0.1693
	Study 3Q (best MASE of individual model is 0.1568)								**Study 4Q** (best MASE of individual model is 0.1525)							
SA	0.1442	**0.1470**	0.1469	0.1458	0.1465	0.1479	0.1495	0.1515	0.1452	**0.1464**	0.1447	0.1429	0.1442	0.1452	0.1484	0.1503
VACO	0.1441	**0.1431**	0.1431	0.1447	0.1475	0.1487	0.1505	0.1534	0.1458	**0.1440**	0.1447	0.1456	0.1467	0.1477	0.1494	0.1520
MSFE(β=0.95)	0.1438	**0.1422**	0.1433	0.1445	0.1472	0.1489	0.1503	0.1538	0.1459	**0.1440**	0.1446	0.1459	0.1470	0.1478	0.1493	0.1524
MSFE(β=0.9)	0.1440	**0.1430**	0.1435	0.1443	0.1471	0.1490	0.1502	0.1542	0.1459	**0.1440**	0.1449	0.1462	0.1472	0.1479	0.1491	0.1527
MSFE(β=0.85)	0.1440	**0.1436**	0.1436	0.1442	0.1470	0.1490	0.1501	0.1546	0.1435	**0.1440**	0.1451	0.1464	0.1473	0.1480	0.1491	0.1530
MSFE(β=0.8)	0.1439	**0.1436**	0.1437	0.1442	0.1469	0.1492	0.1500	0.1549	0.1423	**0.1443**	0.1453	0.1465	0.1475	0.1481	0.1490	0.1533
	VFR 1Q (best MASE of individual model is 0.4135)								**VFR 2Q** (best MASE of individual model is 0.4123)							
SA	0.3661	**0.3597**	0.3626	0.3695	0.3743	0.3783	0.3829	0.3990	0.3512	**0.3504**	0.3585	0.3691	0.3734	0.3827	0.3926	0.3964
VACO	0.3663	**0.3444**	0.3476	0.3529	0.3601	0.3671	0.3732	0.3843	0.3659	**0.3497**	0.3417	0.3505	0.3563	0.3620	0.3693	0.3817
MSFE(β=0.95)	0.3662	**0.3447**	0.3479	0.3528	0.3598	0.3671	0.3729	0.3836	0.3661	**0.3486**	0.3442	0.3522	0.3580	0.3630	0.3713	0.3823
MSFE(β=0.9)	0.3661	**0.3456**	0.3486	0.3539	0.3609	0.3683	0.3734	0.3834	0.3653	**0.3478**	0.3469	0.3542	0.3602	0.3652	0.3751	0.3833
MSFE(β=0.85)	0.3661	**0.3473**	0.3494	0.3564	0.3633	0.3698	0.3748	0.3843	0.3630	**0.3466**	0.3509	0.3571	0.3624	0.3700	0.3802	0.3847
MSFE(β=0.8)	0.3662	**0.3493**	0.3502	0.3595	0.3660	0.3718	0.3768	0.3872	0.3569	**0.3453**	0.3540	0.3633	0.3653	0.3752	0.3836	0.3879
	VFR 3Q (best MASE of individual model is 0.4343)								**VFR 4Q** (best MASE of individual model is 0.4330)							
SA	0.4157	0.3912	**0.3886**	0.3958	0.4035	0.4101	0.4125	0.4157	0.4467	0.4301	**0.4227**	0.4276	0.4314	0.4361	0.4477	0.4546
VACO	0.4192	0.3919	**0.3834**	0.3871	0.3903	0.3942	0.3994	0.4116	0.4420	0.4381	**0.4284**	0.4276	0.4316	0.4344	0.4416	0.4546
MSFE(β=0.95)	0.4191	0.3911	**0.3832**	0.3870	0.3909	0.3945	0.4000	0.4106	0.4414	0.4382	**0.4280**	0.4272	0.4312	0.4342	0.4412	0.4532
MSFE(β=0.9)	0.4188	0.3893	**0.3833**	0.3877	0.3912	0.3958	0.4015	0.4095	0.4405	0.4375	**0.4266**	0.4267	0.4304	0.4341	0.4410	0.4512
MSFE(β=0.85)	0.4185	0.3875	**0.3837**	0.3877	0.3919	0.3983	0.4040	0.4086	0.4390	0.4337	**0.4240**	0.4265	0.4300	0.4346	0.4422	0.4488
MSFE(β=0.8)	0.4191	0.3864	**0.3848**	0.3896	0.3942	0.4012	0.4067	0.4084	0.4370	0.4262	**0.4223**	0.4258	0.4304	0.4372	0.4430	0.4462

	Business 1Q (best MASE of individual model is 0.5113)								**Business 2Q** (best MASE of individual model is 0.5635)							
SA	0.5354	0.5223	**0.5349**	0.5447	0.5370	0.5373	0.5559	0.5672	**0.5978**	0.6058	0.6056	0.6421	0.6698	0.6917	0.7164	0.7357
VACO	0.5372	0.5203	**0.5089**	0.5218	0.5396	0.5620	0.5790	0.6050	**0.5949**	0.5722	0.6376	0.6854	0.7132	0.7318	0.7504	0.7728
MSFE(β =0.95)	0.5352	0.5200	**0.5056**	0.5231	0.5419	0.5645	0.5816	0.6075	**0.5916**	0.5731	0.6501	0.6955	0.7214	0.7396	0.7585	0.7803
MSFE(β =0.9)	0.5329	0.5193	**0.5050**	0.5242	0.5432	0.5660	0.5834	0.6093	**0.5865**	0.5808	0.6612	0.7037	0.7279	0.7455	0.7645	0.7860
MSFE(β =0.85)	0.5296	0.5181	**0.5041**	0.5245	0.5442	0.5665	0.5842	0.6105	**0.5800**	0.5877	0.6676	0.7100	0.7327	0.7497	0.7686	0.7900
MSFE(β =0.8)	0.5270	0.5166	**0.5030**	0.5239	0.5441	0.5661	0.5842	0.6109	**0.5727**	0.5931	0.6715	0.7142	0.7357	0.7521	0.7709	0.7924
	Business 3Q (best MASE of individual model is 0.5828)								**Business 4Q** (best MASE of individual model is 0.6463)							
SA	**0.5664**	0.6291	0.6702	0.7142	0.7615	0.7961	0.8294	0.8593	**0.7148**	0.7478	0.7753	0.7859	0.8062	0.8450	0.8786	0.9117
VACO	**0.5691**	0.6066	0.6772	0.7613	0.8200	0.8635	0.9001	0.9287	**0.7592**	0.7824	0.7893	0.8110	0.8679	0.9122	0.9461	0.9734
MSFE(β =0.95)	**0.5686**	0.6123	0.6876	0.7806	0.8395	0.8835	0.9180	0.9447	**0.7617**	0.7912	0.7945	0.8295	0.8887	0.9319	0.9661	0.9919
MSFE(β =0.9)	**0.5677**	0.6188	0.6973	0.7926	0.8554	0.8996	0.9320	0.9573	**0.7625**	0.7978	0.7966	0.8444	0.9060	0.9475	0.9817	1.0064
MSFE(β =0.85)	**0.5663**	0.6261	0.7065	0.8018	0.8670	0.9114	0.9419	0.9662	**0.7612**	0.8022	0.7978	0.8530	0.9190	0.9585	0.9925	1.0167
MSFE(β =0.8)	**0.5644**	0.6338	0.7136	0.8079	0.8741	0.9188	0.9476	0.9714	**0.7575**	0.8047	0.7983	0.8586	0.9263	0.9652	0.9989	1.0229

Note: SA: Simple Average; VACO: Variance-Covariance; DMSFE(β): Discounted Mean Square Forecast Error method with different β , respectively.

m: The number of individual models for the combination, Bold denotes the best performance of combinations with *m* ($2 \leq m \leq 9$) individual models among all possible combinations models. Bold values corresponding to the best *m*. Q=quarter.

We can observe that the MAPE values presented in Tables 2 and 4 and the MASE values presented in Tables 3 and 5 for the combination of optimal subsets are significantly smaller than the best individual models for most cases apart from the business purpose of visit. If we use MAPE as an error measurement, only the cases which are the SA method for Q2-Q4 and the VACO method for Q2 at the study purpose of visit underperform the best individual models. If we use MASE as an error measurement, only the cases which are the SA method for Q1-Q4 of the study purpose of visit underperform the best individual models. The worst linear combination method is SA for this study. There are 72 cases (3X4X6=72) which are constructed by 3 purposes (holiday, study and VFR), 4 quarters (Q1-Q4), and 6 linear combination methods. There are only 4 out of 72 cases that the proposed combination model underperforms the best individual models. Therefore, the percentage of optimal subsets outperforming the best individual models is 94.4% (68 out of 72 cases) for all linear combinations and all purposes of visits except the business purpose of visit for both MAPE and MASE error measurements. For the business purpose of visit, the best individual model gives better performance than any subset of individual models that contains more than one individual model.

Tables 4 and 5 show that the combinations of two-four individual models give the best performance for all purposes of visits, which is similar to the results of applying the MI algorithm for optimal subset selection. These validate the results in the paper (Shen, Li & Song, 2011). Shen, Li and Song suggested that the highest frequencies of the best combination forecasts appear when the minimum number of individual models for the combination is two. This is unlikely to be effective if a combination of more than five individual forecasts.

The subset of individual models that gives the best performance among all possible combinations of individual models is called the best subset. In order to validate the optimal subset which is obtained using the MI algorithm, the Mann-Whitney test is employed, because the MAPE and MASE values do not satisfy normal distribution by using the Kolmogorov-Smirnov normality test. The results of Mann-Whitney tests and the mean values of MAPE and MASE for all linear combination methods except the SA method are presented in Tables 6 and 7 for the different sets of individual models. These different sets of individual models are the 'set of all individual models' (All), the 'best subset' (Best) presented in Tables 4 and 5 and the 'optimal subset' (Opt) presented in Table 2 and 3 at different purpose of visits. '=', '>' and '>>' in Tables 6 and 7 indicate non statistically significant (equally performance), statistically significant at 5% (performance better than), and very statistically significant at 1% (performance much better than), respectively.

Table 6: Mann-Whitney test to MAPE values for all linear combination methods (except SA)

	Q1-Q4	Q1	Q2	Q3	Q4	Q1-Q4	Q1	Q2	Q3	Q4
	Holiday (Mean value)					Study (Mean value)				
Best	6.91	4.78	6.87	7.86	8.12	9.22	9.28	10.00	8.86	8.75
Opt	7.02	4.78	6.87	8.09	8.34	9.55	9.35	10.38	9.24	9.24
All	7.82	6.01	8.02	8.59	8.65	10.10	9.87	11.33	9.61	9.61
Test	Best=Opt^{NS} Best>>All*** Opt>All**					Best>Opt** Best>>All*** Opt>>All***				
Order	Opt=Best>All					Best>Opt>>All				
	VFR (Mean value)					Business (Mean value)				
Best	5.12	4.55	4.79	5.32	5.81	4.79	3.96	4.58	4.53	6.08
Opt	5.20	4.69	4.96	5.34	5.82	5.06	3.96	4.58	4.93	6.79
All	5.57	5.19	5.26	5.67	6.16	6.54	4.73	6.12	7.46	7.86
Test	Best=Opt^{NS} Best>All** Opt>All**					Best=Opt^{NS} Best>>All*** Opt>>All***				
Order	Best=Opt>All					Best=Opt>>All				

***: Statistical significant difference at 1% level; **: Statistical significant difference at 5% level;
NS: No statistical significant difference; Q=quarter.

Table 7: Mann-Whitney test to MASE values for all linear combination methods (except SA)

	Q1-Q4	Q1	Q2	Q3	Q4	Q1-Q4	Q1	Q2	Q3	Q4
	Holiday (Mean value)					Study (Mean value)				
Best	0.1824	0.1270	0.1827	0.2070	0.2127	0.1464	0.1462	0.1521	0.1431	0.1441
Opt	0.1857	0.1306	0.1827	0.2121	0.2173	0.1500	0.1484	0.1550	0.1481	0.1486
All	0.1979	0.1569	0.2094	0.2127	0.2127	0.1579	0.1563	0.1683	0.1542	0.1527
Test	Best=Opt^{NS} Best>All** Opt=All^{NS}					Best>Opt** Best>>All*** Opt>>All***				
Order	Best=Opt>=All					Best>Opt>>All				
	VFR (Mean value)					Business (Mean value)				
Best	0.3759	0.3463	0.3476	0.3837	0.4259	0.6045	0.5053	0.5851	0.5672	0.7604
Opt	0.3817	0.3516	0.3617	0.3837	0.4298	0.6407	0.5053	0.5814	0.6195	0.8567
All	0.4073	0.3846	0.3840	0.4097	0.4508	0.8372	0.6086	0.7843	0.9537	1.0023
Test	Best:Opt^{NS} Best:All*** Opt:All***					Best:Opt^{NS} Best:All*** Opt:All***				
Order	Best=Opt>>All					Best=Opt>>All				

Note: ***: Statistical significant difference at 1% level; **: Statistical significant difference at 5% level;
NS: No statistical significant difference; Q=quarter

From Tables 6 and 7, we can see that the performance of the optimal sets is significantly better than that of the set that contains all individual models except holiday purpose with the MASE measurement, and no statistical significant difference with the best subset except one case (study). The optimal subset of individual models gives good enough performance, but not necessarily the best performance among all possible combinations of individual models. This suggests that we can use an optimal subset of individual models instead of all individual models in forecasting. The results of Tables 2, 3, 6 and 7 are suggesting following:

The optimal subset selection from individual models using the MI algorithm shows robust and good performance in general. However, optimal subset selection using the MI algorithm does not guarantee that the optimal subset is the same as the best subset.

For the seasonal pattern data (holiday, study, VFR purposes), the performance of optimal subsets is much better than that of the sets of all individual models, and the performance of optimal subsets is closer to that of the best subsets in general.

For the near linear mapping pattern data (business purpose of visits), the best individual models give better performance than that of all combinations of $m\,(m \geq 2)$ individual models in general. This is one of the future research issues to be investigated. The optimal subsets using the MI algorithm give good performance in general if we only consider two or more individual models as combination cases.

The number of individual forecasting models that are contained in the optimal subsets is similar for both MASE and MAPE forecasting error measurements for this study.

4. Conclusions and future work

4.1. Conclusions

This paper has proposed a novel optimal subset selection approach from all available individual models using information theory. This optimal subset from individual models shows good performance and robustness in general. The optimal subsets significantly outperform the non-optimal combination of all individual models as inputs and also give similar performance to the best subsets of individual model in most cases.

The assessment of finding an optimal subset using the MI theory reveals that we can avoid both using a combination of all individual models, and finding the optimal set by trying all possible combinations which involves huge calculations and is time consuming. The most important thing is that it is only an experiment by finding the optimal set using trying all possible combinations method. However, the proposed MI algorithm provides a theoretical approach for finding the optimal set. This paper reveals that the combination from the small size of individual models can achieve higher performance than the best individual model or the combination of all individual models. This significantly enhances the forecasts literature.

The optimal subset selection using the MI algorithm is by nature a ‘heuristic’ approach. It provides us with a good solution, i.e. it may not give a unique solution and it may not guarantee that the optimal subset is the same as the best subset. However, the optimal subset of individual models using the MI algorithm shows robust and good performance.

Two main results are observed: 1) The optimal subset forecast model performs statistically better than the combination model using all available individual models as inputs, and 2) the dimension of the optimal subset forecast model is in the range of two and five individual models. This research can help both government organizations and the tourism related industries, since accurate forecasting on tourism demand is critical for their policy and decision making. This can benefit the transportation, accommodation, catering, entertainment and retailing sectors. For examples, this research can help 1) make the appropriate government policies which can promote development of hospitality and tourism industries such as hotels, restaurants and attraction sites; 2) provide the guidance for both central governments and tourism related industries on capacity management, such as for the department of transportation on reducing congestion during the tourism seasons in order to achieve government 'public service agreement targets' (PSA targets). The tourism related industries can benefit to healthy run business by employing right number of staff and control business scale; 3) provide the guidance for both central governments and tourism related industries on investment such as airport, transport networks and tourism attraction sites etc.

4.2. Limitations and future researches

Time series individual forecasting models are used in this research. The causal econometric model using 'GDP' and 'CPI' influencing factors will be considered in the future research to see if the forecasting accuracy can be improved, and the combination selection algorithm can be enhanced.

This paper only used the data up to 2007 inclusive. The up to date data will be used in the future to test the robustness of the combination selection algorithm proposed in this study, in particular the impact of the financial crisis on the forecasting performance.

The linear combination methods are adopted in this research. The nonlinear combination methods can be applied to evaluate the combination selection algorithm.

The future work can also consider to exam the combination selection algorithm for the other tourism data sets. Another issue is to see whether the dimension of the optimal subset is still in the range of two and five for using a large number of available individual models.

Acknowledgements

The authors would like to thank editor and anonymous referees for their valuable comments and suggestions

References

Abellán J., & Masegosa R. A. (2010). An ensemble method using credal decision trees, European Journal of Operational Research, 205, 218–226.

Andrawisa R. R., Atiyaa A. F., & El-Shishiny H. (2011). Combination of long term and short term forecasts, with application to tourism demand forecasting. International Journal of

Forecasting, 27, 870–886.

Baba C., & Kisinbay T. (2011). Predicting recessions: A new approach for identifying leading indicators and forecast combinations. IMF Working Paper, WP/11/235.

Bates J. M., & Granger C. W. J. (1969). The combination of forecasts. Operational Research Quarterly, 20, 451–468.

Battiti R. (1994). Using mutual information for selecting features in supervised neural net l earning. IEEE Transaction Neural Networks, 5, 537-550.

Bishop M. C. (1995). Neural networks for pattern recognition, Oxford University Press.

Costantini M., & Pappalardo C. (2010). Hierarchical procedure for the combination of forecasts. International Journal of Forecasting, 26, 725-743.

Bodyanskiy, Y., & Popov, S. (2006). Neural network approach to forecasting of quasiperiodic financial time series. European Journal of Operational Research, 175, 1357-1366.

Box G., & Jenkins, G. (1970). Time series analysis: Forecasting and control. San Francisco: Holden-Day.

Breiman L. (1996). Bagging predictors, Machine Learning, 24 (2), 123–140.

Cao Q, Ewing B. T., & Thompson M. A. (2012). Forecasting wind speed with recurrent neural networks. European Journal of Operational Research, 221, 148-154.

Carbonneau R., Laframboise K., & Vahidov R. (2008). Application of machine learning techniques for supply chain demand forecasting. European Journal of Operational Research, 184, 1140-1154.

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

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

Cho V. (2001). Tourism forecasting and its relationship with leading economic indicators. Journal of Hospitality and Tourism Research, 25, 399–420.

Cho V. (2003). A comparison of three different approaches to tourist arrival forecasting. Tourism Management, 24, 323-330.

Christopher M. B. (1995). Neural Networks for Pattern Recognition. Oxford: Oxford University Press.

Chu F. l. (2004). Forecasting tourism demand: a cubic polynomial approach. Tourism Management, 25, 209-218.

Coshal J., & Charlesworth R. (2011). A management orientated approach to combination forecasting of tourism demand. Tourism Management, 32 759-769.

Dickey, D.A., & Fuller W.A. (1979), Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74, 427–431.

Douglas C. M., Lynwood A. J., & John S. G. (1990), Forecasting and time series analysis. McGraw-Hill, Inc, 2nd Edition. Columbus, OH, USA.

Estévez P.A., Tesmer M., Perez C.A., and Zurada J.M. (2009). Normalized mutual information feature selection. IEEE Transactions on Neural Networks, 20 (2), 189–201.

Finlay S. (2011). Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research, 210, 368–378.

Fleuret F. (2004). Fast binary feature selection with conditional mutual information. Journal of Machine Learning Research, 5, 1531–1555.

Freitas P. S. A., & Rodrigues A. J. L. (2006). Model combination in neural-based forecasting. European Journal of Operational Research, 173, 801-814.

Freund Y., & Schapire R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, 55 (1), 119–139.

Ghysels E., Lee H. S. & Noh J. (1994). Testing for unit roots in seasonal time series-some theoretical extensions and a Monte Carlo investigation. Journal of Econometrics, 62, 415-442.

Goh C., & Law R. (2002). Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. Tourism Management, 23, 499–510.

Gounopoulos D., Petmezas D., & Santamaria D. (2012). Forecasting tourist arrivals in Greece and the impact of macroeconomic shocks from the countries of tourists' origin. Annals of Tourism Research, 39 (2), 641-666.

Greer M. R. (2005). Combination forecasting for directional accuracy: An application to survey interest rate forecasts. Journal of Applied Statistics, 32, 6, 607–615.

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

Hall S. G., & Mitchell J. (2007). Combining density forecasts. International Journal of Forecasting, 23, 1–13.

Hanke, J. E., & Reitsch, A. G. (1995). Business Forecasting. (5th ed.). Englewood Cliffs, NJ7, Prentice-Hall.

He, C., & Xu X. (2006). Combination of forecasts using self-organizing algorithms. Journal of Forecasting, 24, 269-278.

Holden K., & Peel D. A. (1986). An empirical investigation of combinations of economic forecasts. Journal of Forecasting, 5, 229–242.

Hylleberg, S., Engle, R.F., Granger, C.W.J., & Yoo, B.S., 1990. Seasonal integration and cointegration. Journal of Econometrics, 44, 215-238.

Hyndman R. J., & Koehler A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22, 679– 688.

Li. K.-H., Wong H., & Troutt M. (2001). An approximate Bayesian algorithm for combining Forecasts. Decision Sciences, 32, 453-471.

Kisinbay, T. (2010). The use of encompassing tests for forecast combinations. Journal of Forecasting, 29, 715-727.

Kwak N., & Choi C. H. (2002). Input feature selection for classification problems. IEEE Transaction Neural Networks, 13, 143-159.

Law R., & Au N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. Tourism Management, 20, 89-97.

Lessmann S., Sung M. C., Johnson J. E.V., & Ma T. (2012). A new methodology for generating and combining statistical forecasting models to enhance competitive event prediction. European Journal of Operational Research, 218, 163-174.

Li G., Shi J., & Zhou J. Y. (2011). Bayesian adaptive combination of short-term wind speed forecasts from neural network models. Renewable Energy, 36(1), 352-359.

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

Li G., Song H., & Witt S. F. (2006). Time varying parameter and fixed parameter linear AIDS: An application to tourism demand forecasting. International Journey of Forecasting, 22, 57- 71.

Lim, C. (2001). Monthly seasonal variations: Asian tourism to Australia. Annals of Tourism Research, 28, 68–82.

Mackay David J. C. (2003). Information Theory, Inference, and Learning Algorithms. Cambridge: Cambridge University Press.

Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., & Lewandowski, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. Journal of Forecasting, 1, 111 – 153.

Menezes de L. M., Bunn D. W., & Taylor J. W. (2000). Review of guidelines for the use of combined forecasts. European Journal of Operational Research, 120, 190-204.

Meyer P. E., Schretter C. and Bontempi G. (2008). Information-theoretic feature selection in microarray data using variable complementarity. IEEE Journal of Selected Topics in Signal Processing, 2 (3). 261-274.

Naude W. A., & Saayman A. (2005). Determinants of tourist arrivals in Africa: A panel data regression analysis. Tourism Economics, 11, 365–391.

Newbold P., & Granger. C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. Journal of the Royal Statistical Society, Series A, 137, 131–149.

Ni Karl S., & Nguyen Truong Q. (2007). Image super resolution using support vector regression. IEEE Transactions on Image Processing, 16, 1596-1610.

Oh, C. O., & Morzuch, B. J., (2005). Evaluating time-series models to forecast the demand for tourism in Singapore: Comparing within-sample and post-sample results. Journal of Travel Research, 43, 404-413.

Page S., Song H., & Wu D. (2012). Assessing the impacts of the global economic crisis and swine flu on inbound tourism demand in the United Kingdom. Journal of Travel Research, 51(2), 142-153.

Pai P. F., & Hong W. C. (2005). An improved neural networks model in forecasting arrivals. Annals of Tourism Research, 32 (4), 1138-1141.

Pai, P. F., Hong, W. C., Chang, P. T., & Chen, C. T. (2006). The application of support vector machines to forecast tourist arrivals in Barbados: An empirical study. International Journal of Management, 23, 375–385.

Parzen, E. (1962). On estimation of a probability density function and mode. Ann. Math. Statist., 33: 1065.

Peng H., Long F., & Ding C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27 (8), 1226–1238.

Roget F. M., & Gonzalez X. A. R. (2006). Rural tourism demand in Galicia, Spain. Tourism Economics. 12, 21–31.

Sánchez I. (2008). Adaptive combination of forecasts with application to wind energy. International Journal of Forecasting, 24, 679–693.

Simon H. (1999). Neural networks: A Comprehensive Foundation. New Jersey: Prentice Hall.

Shen S., Li G., & Song H. (2008). An assessment of combining tourism demand forecasts over different time horizons. Journal of Travel Research. 47, 197-207.

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

Song H., & Li G. (2008). Tourism demanding modelling and forecasting – A review of recent research. Tourism Management, 29, 203-220.

Sridhar D. V., Bartlett E. B., & Seagrave R. C. (1999). An information theoretic approach for combining neural network process models. Neural Networks, 12, 915-926.

Tay, F., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. Omega: The International Journal of Management Science, 29(4), 309–317.

Tibshirani, R., Friedman J., & Hastie T. (2000). Additive logistic regression: A statistical view of boosting, Annals of Statistics, 28 (2), 337–407.

Timmermann A., Elliott G., & Granger C. W. J. (2006). Forecast combinations, Handbook of Economic Forecasting, Elsevier Pub., 135-196.

Theodoridis S., & Koutroumbas K. (1999). Pattern Recognition. San Diego: Academic Press.

Vapnik V., Golowich S., & Smola A. (1996). Support vector machine for function Approximation regression estimation, and signal processing. Advances in Neural Information Processing Systems, 9, 281-287.

Vapnik V. (1995). The nature of statistical learning theory. Springer, New York.

Webb G. I. (2000). MultiBoosting: A technique for combining Boosting and Wagging, Machine Learning, 40 (2), 159–196.

Wezel van M., & Potharst R. (2007). Improved customer choice predictions using ensemble methods. European Journal of Operational Research, 181, 436–452.

Winkler. R. L., & Makridakis. S. (1983). The combination of forecasts. Journal of the Royal Statistical Society, Series A, 146, 150–157.

Wong W. K., Xia M., & Chu W.C. (2010). Adaptive neural network model for time-series forecasting. European Journal of Operational Research, 207, 807-816.

Wong K. K. F., Song H., Witt S. F., & Wu C. D. (2007). Tourism forecasting: To combine or not to combine ?. Tourism Management, 28, 1068-1078.

Wu S., & Akbarov A. (2011). Support vector regression for warranty claim forecasting. European Journal of Operational Research, 213, 196-204.

Yu L., & Liu H. (2004). Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research, 5, 1205–1224.

Zheng W. Z., Lee D. H., & Shi Q. (2006). Short-term freeway traffic flow prediction: Bayesian combined neural network approach. Journal of Transportation Engineering, 132 (2) 114-122

Zouh H., & Yang Y. (2004). Combining time series models for forecasting. International Journal of Forecasting, 20, 69-84.

Appendix:

Table A1: MAPE and MASE values from individual models at different purpose of visits

Test data or Ex-post (Q1 2004 to Q4 2007 inclusive)									
		MAPE				**MASE**			
Holiday									
ID	**Model**	**1Q**	**2Q**	**3Q**	**4Q**	**1Q**	**2Q**	**3Q**	**4Q**
1	SVRNN (d=4)	8.76	10.26	9.54	9.47	0.2338	0.2717	0.2544	0.2573
2	SVRNN (d=5)	7.48	9.81	8.61	9.00	0.2018	0.2644	**0.2351**	0.2571
3	SVRNN (d=6)	7.78	8.91	**8.52**	**8.81**	0.2103	0.2444	0.2355	0.3023
4	SVRNN (d=7)	8.02	8.78	8.76	9.56	0.2147	0.2410	0.2408	0.3030
5	SVRNN (d=8)	8.10	9.20	9.29	9.52	0.2188	0.2553	0.2568	0.3166
6	Naïve 1	9.75	9.75	9.75	9.75	0.2540	0.2540	0.2540	**0.2437**
7	Naïve 2	14.26	14.26	14.26	14.26	0.3662	0.3662	0.3662	0.3380
8	SARIMA	**6.39**	11.41	12.21	9.99	**0.1595**	0.2788	0.2958	0.3418
9	WMES	6.49	**8.34**	9.47	9.44	0.1762	**0.2206**	0.2442	0.3371
Study									
ID	**Model**	**1Q**	**2Q**	**3Q**	**4Q**	**1Q**	**2Q**	**3Q**	**4Q**
1	SVRNN (d=4)	10.84	12.14	10.05	9.99	0.1656	0.1676	**0.1568**	0.1562
2	SVRNN (d=5)	**10.67**	12.34	12.26	12.01	0.1668	0.1658	0.1934	0.1787
3	SVRNN (d=6)	12.11	13.88	**9.87**	9.73	0.1872	0.1735	0.1636	0.1629
4	SVRNN (d=7)	12.67	14.22	10.26	**9.68**	0.1932	0.1796	0.1678	**0.1525**
5	SVRNN (d=8)	12.19	13.95	11.41	10.37	0.1882	0.1916	0.1819	0.1665
6	Naïve 1	11.58	11.58	11.58	11.58	0.1732	0.1732	0.1732	0.1732
7	Naïve 2	20.77	20.77	20.77	20.77	0.3140	0.3140	0.3140	0.3140
8	SARIMA	12.40	**10.42**	11.26	11.49	0.1852	0.1637	0.1695	0.1750
9	WMES	**10.67**	10.92	10.99	11.76	**0.1578**	**0.1596**	0.1613	0.1792
VFR									
ID	**Model**	**1Q**	**2Q**	**3Q**	**4Q**	**1Q**	**2Q**	**3Q**	**4Q**
1	SVRNN (d=4)	5.94	6.31	6.24	6.17	0.4406	0.4666	0.4620	0.4604
2	SVRNN (d=5)	5.68	5.93	**5.94**	**5.86**	0.4176	0.4333	**0.4343**	**0.4330**
3	SVRNN (d=6)	5.94	5.92	6.07	6.46	0.4433	0.4417	0.4517	0.4770
4	SVRNN (d=7)	6.17	6.09	6.26	6.39	0.4586	0.4531	0.4613	0.4722
5	SVRNN (d=8)	5.99	7.29	7.92	8.37	0.4317	0.5234	0.5696	0.5952
6	Naïve 1	9.58	9.58	9.58	9.58	0.6953	0.6953	0.6953	0.6953
7	Naïve 2	6.88	6.88	6.88	6.88	0.5070	0.5070	0.5070	0.5070
8	SARIMA	5.63	**5.69**	6.27	6.51	0.4222	**0.4123**	0.4489	0.4717
9	WMES	**5.47**	6.05	7.09	7.66	**0.4135**	0.4404	0.5105	0.5553
Business									
ID	**Model**	**1Q**	**2Q**	**3Q**	**4Q**	**1Q**	**2Q**	**3Q**	**4Q**
1	SVRNN (d=4)	4.99	6.42	8.00	8.55	0.6440	0.8237	1.0242	1.0922
2	SVRNN (d=5)	4.89	6.41	8.56	8.66	0.6338	0.8259	1.0981	1.1137
3	SVRNN (d=6)	5.81	7.12	8.51	8.86	0.7506	0.9202	1.0967	1.1383
4	SVRNN (d=7)	5.26	6.13	8.21	8.51	0.6795	0.7908	1.0531	1.0918
5	SVRNN (d=8)	5.22	6.64	8.55	9.46	0.6736	0.8524	1.0998	1.2143
6	Naïve 1	7.30	7.30	7.30	7.30	0.9218	0.9218	0.9218	0.9218
7	Naïve 2	9.08	9.08	9.08	9.08	1.1372	1.1372	1.1372	1.1372
8	SARIMA	5.15	5.33	5.13	6.97	0.6343	0.6806	0.6277	0.8635
9	WMES	**3.94**	**4.39**	**4.58**	**5.01**	**0.5113**	**0.5635**	**0.5828**	**0.6463**

Note: Bold denotes the best performance among all individual models for test data
ID indicates the id number of the individual model