Forecasting with model selection or model averaging: a case study for monthly container port throughput

Yan Gao, Meifeng Luo & Guohua Zou

To cite this article: Yan Gao, Meifeng Luo & Guohua Zou (2016) Forecasting with model selection or model averaging: a case study for monthly container port throughput, Transportmetrica A: Transport Science, 12:4, 366-384, DOI: 10.1080/23249935.2015.1137652

To link to this article: http://dx.doi.org/10.1080/23249935.2015.1137652

Accepted author version posted online: 04 Jan 2016.
Published online: 29 Jan 2016.

Submit your article to this journal

Article views: 16

View related articles

View Crossmark data
Forecasting with model selection or model averaging: a case study for monthly container port throughput

Yan Gao\textsuperscript{a,b}, Meifeng Luo\textsuperscript{c} and Guohua Zou\textsuperscript{b}

\textsuperscript{a}Department of Statistics, School of Science, Minzu University of China, Beijing, People’s Republic of China; \textsuperscript{b}Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, People’s Republic of China; \textsuperscript{c}Department of Logistics and Maritime Studies, Hong Kong Polytechnic University, Kowloon, Hong Kong SAR

ABSTRACT
An accurate short-term prediction of time series data is critical to operational decision-making. While most forecasts are made based on one selected model according to certain criteria, there are developments that harness the advantages of different models by combining them together in the prediction process. Following on from existing work, this paper applies six model selection criteria and six model averaging (MA) criteria to a structural change vector Autoregressive model, and compares them in terms of both the theoretical background and empirical results. A case study of the monthly container port throughput forecasting for two competing ports shows that, in general, the model averaging methods perform better than the model selection methods. In particular, the leave-subject-out cross-validation MA method is the best in the sense of achieving the lowest average of mean-squared forecast errors.

1. Introduction
Time series models are widely applied in analysing the underlying patterns of one or more economic variables based on past observations, in order to predict the possible value of these economic variables assuming that the data-generation process remains unchanged. In real world, mostly such predictions are done using one model selected from a set of potential candidates based on different criteria. These model selection (MS) methods often result in different models being selected due to data changes, which introduce uncertainties in prediction accuracy (Shen and Huang 2006). An alternative to selecting one model for all future predictions is known as model averaging (MA), which combines together the predictive power of a set of plausible models in order to produce more accurate and consistent predictions (Garratt et al. 2003; Kapetanios, Labhard, and Price 2008; Pesaran, Schleicher, and Zaffaroni 2009).

Understanding the properties of different prediction methods is important both for selecting different forecasting methods, and for adopting the forecast result in business

CONTACT Meifeng Luo \textsuperscript{c} meifeng.luo@polyu.edu.hk, lgtmluo@polyu.edu.hk

© 2016 Hong Kong Society for Transportation Studies Limited
decisions. However, such an understanding is only meaningful in a specific application context. There is no theoretical conclusion as to the preference order of different methods, nor any prior empirical evidence as to the dominance of one method over the others.

This paper demonstrates one particular application where different models have to be compared in order to adopt a prediction method. In recent years, container port throughput in China has increased rapidly, and ports play an important role in both regional growth and international trade. A long-term prediction of container throughput is critical for making a large capital investment in port infrastructure development, while for port operation and management, an accurate short-term forecast is essential.

Nevertheless, making an accurate and consistent prediction is not straightforward, even for a short-term prediction, when a port is in a competitive environment and there are structural changes in the time series data that has seasonal variations. Using the container port throughput in Hong Kong and Shenzhen as an example, we model the interactions of the two ports in a vector autoregressive (VAR) framework, taking into account any structural changes due to the financial crisis in 2008. In comparison with most existing studies on container port throughput forecasting, as shown in Section 2, we contribute to the literature by evaluating many potential VAR models, illustrate the use of MS and MA methods, and provide an accurate forecast of monthly container port throughput. In particular, compared with those structural models, our method relies only on the time series data of container port throughput, which is less demanding on the data availability, thus more practical in real application.

The rest of this paper is organised as follows. The following section is a review of the existing literature on forecasting container port throughput. Section 3 contains a description of the basic model. Section 4 introduces the MS criteria. Section 5 provides some frequentist model averaging (FMA) methods. Section 6 is an empirical comparison of different methods, and Section 7 concludes the paper.

2. Literature review on container throughput forecasting

Due to the importance of container ports in a national economy, and of container throughput forecasting in terminal operation and management, there have been many studies on how to make an accurate forecast of container throughput. Traditional methods include linear time series models, the Grey system and regression analysis. Guo et al. (2005) proposed the Grey Verhulst model of error-corrected time series data for port throughput forecasting. Fung (2002) developed a structural error correction model to forecast Hong Kong’s throughput, taking into account the substitution between different terminal operation types, such as midstream and ocean terminals, as well as the impact of Singapore port on Hong Kong throughput. The forecast results are claimed to be more accurate than those provided by the government. Hui et al. (2010) included real-estate prices, as a proxy of port charges, in the VAR error correction model for container port throughput forecasting. Peng and Chu (2009) compared six univariate models, namely the classical decomposition model, the trigonometric regression model, the regression model with seasonal dummy variables, the Grey model, the hybrid Grey model and the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, in forecasting the container import volume of three international ports in Taiwan, and found that the classical decomposition model is the best for forecasting container throughput with seasonal variation.
In addition to the traditional statistical methods, more complex methods have been applied in container throughput forecasting. Lam et al. (2004) developed a neural network (NN) model for the long-term (10 years) forecasting of Hong Kong port cargo throughput, and found that the NN model outperforms regression analysis. Xiao, Xiao, and Wang (2009) proposed a hybrid forecasting model using a novel feed forward NN based on an improved particle swarm optimisation with adaptive genetic operator (IPSO-FNN). The model was applied to forecast Tianjin port container throughput, and they found that the IPSO-FNN model performs better than the FNN and PSO-FNN models in terms of mean absolute error, mean absolute percentage error (MAPE) and root mean-squared error. Chen and Chen (2010) compared container throughput forecasts for major Taiwanese ports using genetic programming (GP), the decomposition approach (X-11), and the SARIMA model, and found that GP has much lower MAPE. Xie et al. (2013) proposed three hybrid approaches based on a least-squares support vector regression (LSSVR) model for container throughput forecasting. This method divided the data into linear and nonlinear parts, and can thus take the advantage of linear and nonlinear modelling simultaneously. Zhang and Cui (2011) developed a combination forecast method based on the Elman neural network to predict the container throughput in Shanghai port. They combined together the predictions of the cubic exponential smoothing method and the Grey model. Empirical results show that the combination method had a high forecast accuracy.

Compared with existing approaches, our model has two particular attributes. First, we include the impact of a competing port, considering that most ports operate in a competitive environment. Second, we compare MS with MA in terms of both the theoretical background and the empirical results. This provides readers with more information as to appropriate methods to choose in order to meet their own particular objectives.

3. Framework of the model

To illustrate the differing properties of MS and MA methods, we first provide a general time series model for container port throughput in Hong Kong and Shenzhen ports. As there are possible interactions between the two ports, a VAR(m) model is adopted. To take into account possible structural changes due to the 2008 financial crisis, the structural change VAR(m) (referred to hereafter as the SC-VAR(m) model) can be expressed as follows:

\[
y_i = \mu_i + e_i = \begin{cases} 
C^{(1)} + \Phi_1^{(1)} y_{i-1} + \cdots + \Phi_m^{(1)} y_{i-m} + e_i & 1 \leq i \leq n_1, \\
C^{(2)} + \Phi_1^{(2)} y_{i-1} + \cdots + \Phi_m^{(2)} y_{i-m} + e_i & n_1 < i \leq n,
\end{cases}
\]

where \( y_i \) is a 2 \times 1 vector of the time series variables, \( n \) is the total number of observation, and \( n_1 \) is the breakpoint that separates the sample into two periods. \( C^{(r)} \) is the 2 \times 1 intercept vector for period \( r \) (\( r = 1, 2 \)), \( \Phi_j^{(r)} \) (\( j = 1, \ldots, m \)) is the 2 \times 2 coefficient matrix, and \( m \) is the maximum lag length. The error term \( e_i \) is a 2 \times 1 vector, one for each port. They are assumed to be independent and identically distributed, that is, \( e_i \) is \( i.i.d. \sim N(0, \Sigma) \), where \( \Sigma \) is a 2 \times 2 covariance matrix. The diagonal elements in this matrix are the variances for each port (\( \sigma_1^2 \) for Hong Kong and \( \sigma_2^2 \) for Shenzhen), while the off diagonal elements are the covariances between the two ports (\( \sigma_{12} \) and \( \sigma_{21} \)).
Let \( x_i = (1, y'_{i-1}, \ldots, y'_{i-m})' \) be a \((2m + 1) \times 1\) vector and \( \Pi^{(r)} = [C^{(r)} \Phi_1^{(r)} \cdots \Phi_m^{(r)}] \) be a \( 2 \times (2m + 1) \) coefficient matrix. Thus, \( \{y_i, x_i\} \) is derived using the observations from \( m \) time-steps before \( i \) to time \( i \). The conditional mean at time \( i \) can be expressed as

\[
\mu_i = \begin{cases} 
\Pi^{(1)} x_i & 1 \leq i \leq n_1, \\
\Pi^{(2)} x_i & n_1 < i \leq n.
\end{cases}
\]  

(2)

Denote \( Y = (y'_1, \ldots, y'_n)' \), \( \mu = (\mu'_1, \ldots, \mu'_n)' \) and \( e = (e'_1, \ldots, e'_n)' \). The SC-VAR\((m)\) model can be written in the following form:

\[
Y = \mu + e.
\]

To take into account the structural changes, we split the data into two parts and estimate the coefficient matrices separately. For simplicity, we assume that the change point \( n_1 \) is known and that the lag lengths in the two regimes are the same. In this case, the estimator of the coefficient matrix, \( \hat{\Pi}^{(r)} \) can be easily obtained using Ordinary Least Squares (OLS). In the case that \( n_1 \) is unknown, it can be obtained using conditional least squares (Bai 1997).

A Wald statistic can be developed to test the significance of the structural change. Let \( \phi^{(r)} \) be the vectorised coefficient matrix, that is, \( \text{vec}(\Pi^{(r)}) \), and denote the estimator of \( \phi^{(r)} \) by \( \hat{\phi}^{(r)} \), the Wald statistic for the null hypothesis \( H_0 : \phi^{(1)} = \phi^{(2)} = \phi \) is

\[
W = n_1 (1 - n_1/n)(\hat{\phi}^{(1)} - \hat{\phi}^{(2)})' \hat{\Sigma}^{-1}(\hat{\phi}^{(1)} - \hat{\phi}^{(2)}),
\]

(3)

where \( \hat{\Sigma} = \hat{\Sigma} \otimes (X'X/n)^{-1} \), \( \hat{\Sigma} = \sum_{i=1}^n \hat{e}_i \hat{e}_i'/n \), \( \hat{e}_i \) is the full-sample OLS residual and \( X = (x_1, \ldots, x_n)' \). The rejection region is \( \{W > \chi^2_{4m+2}(\alpha)\} \). The details for the derivation of this test are given in the appendix.

The regression process provides a list of the potential time series models. In MS methods, one of them will be selected for forecasting. It has to be noted that for different time series data, the best MS method may be different. A comparison of the superiority of different methods in MS is only meaningful in a specific application with real data, which will be illustrated in Section 6. However, an introduction to the different MS criteria is necessary, and this will be provided in the next section.

4. MS methods

The objective of MS is to find the ‘best’ model among a given set of candidates. For VAR models, it is to find the ‘best’ by varying the length of lag. Some commonly used criteria for MS are information-based criteria, Mallows’ \( C_p \) criteria and cross-validation (CV) criteria.

4.1. Information-based criteria

Generally, these selection criteria, including Akaike information criterion (AIC), Bayesian information criterion (BIC) and Hannan–Quinn criterion (HQC), assume the form of a penalised log-likelihood function as follows:

\[
xC_k = -2 \ell_{n,k} + c_k,
\]

(4)

where \( x \) denotes AI, BI and HQ. In the first part, \( \ell_{n,k} \) is the log-likelihood function of model \( k \), which represents the goodness of fit. The second part, \( c_k \), is the penalty term, which is
proportional to the model complexity. Such a function takes into account the trade-off between good fit and model complexity. The model with the smallest $x_C$ value will be chosen.

Let $\hat{e}_i^{(k)}$ be the OLS residual of model $k$, and omit any irrelevant terms, then the log likelihood function of model $k$ can be expressed as

$$\ell_{n,k} = -\frac{n}{2} \log |\hat{\Sigma}(k)|,$$

(5)

where $|\hat{\Sigma}(k)|$ is the determinant of the estimated covariance matrix for model $k$ given by

$$\hat{\Sigma}(k) = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i^{(k)} \hat{e}_i^{(k)'}.$$

4.1.1. Akaike information criterion

AIC was proposed by Akaike (1973). The penalty term is $c_k = 2p_k$, where $p_k$ is the number of the estimated parameters in model $k$. Minimising AIC is equivalent to minimising an approximately unbiased estimator of the Kullback–Leibler (KL) distance (Kullback 1959). If $c_k = 2p_k(n/(n - p_k - 1))$, it transfers to the corrected AIC ($AIC_c$), which is an exactly unbiased estimator of the KL distance. It can be seen that when $n \to \infty$, $AIC_c$ will converge to AIC. Thus, in a large sample, the performance of these two methods is the same, whereas in a small sample case, $AIC_c$ is more accurate than AIC. The AIC score of the $k$th candidate model is

$$AIC_k = n \log |\hat{\Sigma}(k)| + 2p_k.$$  

(6)

4.1.2. Bayesian information criterion

Note that as the sample size $n$ increases, AIC will select a more complex model, because the log-likelihood will linearly increase with $n$, whereas the penalty term does not increase with $n$. BIC can avoid this problem by setting the penalty term as $c_k = p_k \log n$. For our problem, the BIC score is

$$BIC_k = n \log |\hat{\Sigma}(k)| + p_k \log n.$$  

Since the penalty of BIC is larger than that of AIC, BIC tends to select a smaller model, that is, a model with fewer parameters.

4.1.3. Hannan–Quinn criterion

Hannan and Quinn (1979) developed a consistent criterion, HQC, which does not penalise the model complexity as much as BIC. The penalty term is given by $c_k = cp_k \log \log n$, where $c$ is a constant larger than 1. Claeskens and Hjort (2008) pointed out that this choice of penalty might not be very useful, because $\log \log n$ remains small even for a large sample size. Then the penalty may be determined by the value of $c$. To make the penalty of HQC lie between those of AIC and BIC, we set $c$ to be 2. Thus, the QHC score is given by

$$HQC_k = n \log |\hat{\Sigma}(k)| + 2p_k \log \log n.$$
4.2. Mallows’ $C_p$ criterion

The Mallows criterion (1973), designed to be an approximately unbiased estimator for the in-sample fit $\|\mu - \hat{\mu}\|^2$, penalises on the sum of squared residuals. If the estimator $\hat{\mu}^{(k)}$ can be expressed as $\hat{\mu}^{(k)} = \Pi^{(k)}Y$, where $\Pi^{(k)}$ is a matrix independent of $Y$, the Mallows’ $C_p$ can be defined as follows:

$$C_p^{(k)} = \|Y - \hat{\mu}^{(k)}\|^2 + 2\operatorname{tr}(\hat{\Sigma}(k)).$$ (7)

where $\hat{\sigma}^2$ is an estimator of $\sigma^2$ and $\operatorname{tr}(\cdot)$ is the trace operator. Many commonly used estimators such as least squares, ridge regression, nearest neighbour estimators, kernel regression, local linear regression and spline estimators are included in this class. We can develop the Mallows criterion for the SC-VAR($m$) model as:

$$C_p^{(k)} = \|Y - \hat{\mu}^{(k)}\|^2 + p_k\operatorname{tr}(\hat{\Sigma}(k)).$$ (8)

Please see the appendix for details.

4.3. CV criterion

CV is often used in MS (Picard and Cook 1984). In CV, the data are divided into two parts, one for model fitting and the other for validation.

4.3.1. Leave-one-out CV criterion

The most common CV method is the leave-one-out CV (LooCV), where one observation is left out at a time. Let $\hat{\Pi}^{(i)}_{[-i]}(k)$ denote the OLS estimate of the coefficient matrix $\Pi^{(i)}(k)$ for period $r$ with the $i$th observation $\{y_i, x_i(k)\}$ deleted in the $k$th model. Following Equation (2), $\bar{\mu}^{(k)}_j$, the leave-one-out estimator of $\mu_j^{(k)}$, can be obtained:

$$\bar{\mu}^{(k)}_j = \begin{cases} 
\hat{\Pi}^{(1)}_{[-i]}(k)x_i(k) & 1 \leq i \leq n_1 \\
\hat{\Pi}^{(2)}_{[-i]}(k)x_i(k) & n_1 < i \leq n.
\end{cases}
$$

Then its residual can be calculated as $\bar{\varepsilon}^{(k)}_j = y_i - \bar{\mu}^{(k)}_j$. Denoting $\bar{\varepsilon}^{(k)} = (\bar{\varepsilon}^{(k)}_1', \ldots, \bar{\varepsilon}^{(k)}_n')'$, the LooCV criterion can be calculated by

$$\text{LooCV}_k = \bar{\varepsilon}^{(k)}'\bar{\varepsilon}^{(k)},$$ (9)

and the problem is to find the model with index $k$ that has the minimum $\text{LooCV}_k$.

4.3.2. Leave-subject-out CV criterion

Leave-subject-out CV (LsoCV) is commonly used in panel data models. Gao et al. (2015) used this criterion in time series models where the errors can exhibit autocorrelation. Instead of deleting one observation as in the previous method, LsoCV requires the deletion of a set of related observations, called a subject, for each observation, in order to find the best model. In this paper, a subject is a set of observations whose errors have autocorrelation. These subjects can be identified by autocorrelation function (ACF), partial ACF and mixing coefficients (White 1984). In this paper, ACF is used to determine which subject to delete.

For each candidate model $k$, we first calculate the ACF of the residuals to find the correlated error terms and record the lags $l$s with high ACF. Then a correlated index set $S_k$
containing 0 and \( l \) can be formed for each observation. Similar to the looCV, the leave-subject-out estimator of the coefficient matrix, \( \tilde{\mu}(k) \), can be calculated using the OLS method with data \( \{y_j, x_j \mid j \in \{1, \ldots, n\} \setminus \{i \pm S_k\}\} \). For example, if \( l = 1, 2, 3 \), and the correlated index set is \( S_k = \{0, 1, 2, 3\} \), then \( \tilde{\mu}(k) \) can be estimated using OLS with data \( \{y_j, x_j(k) \mid j = 1, \ldots, i - 4, i + 4, \ldots, n\} \), that is, the data where the current observation and three correlated observations from both before and after the current observation, a total of seven observations, are all deleted. Actually, the leave-one-out method is just a special case of the leave-subject-out method when there is no autocorrelation, that is, when \( S_k = \emptyset \).

The LsoCV estimator \( \tilde{\mu}(k) \) of \( \mu_i \) is obtained by Equation (2) with \( \tilde{\mu}(k) \) replaced by \( \tilde{\mu}(k) \). After calculating the residual \( \tilde{e}(k) = y_i - \tilde{\mu}(k) \), the LsoCV criterion can be obtained:

\[
\text{LsoCV}_k = \tilde{e}^{(k)'} \tilde{e}^{(k)},
\]

where \( \tilde{e}(k) = (\tilde{e}_1^{(k)}, \ldots, \tilde{e}_n^{(k)})' \).

### 4.4. Evaluating the MS: consistency and efficiency

A number of theoretical properties are available to evaluate the performances of MS criteria. Consistency and efficiency are two classical properties. Many researchers have studied these properties for specific models, and a fairly complete summary is available in Claeskens and Hjort (2008). Here, we briefly introduce these properties, and compare the above six model selection methods.

Weak consistency measures the ability of the MS criterion to select the best model among a set of potential models when the sample size is large. All the xC methods are of weak consistency, while \( \text{Cp} \) and CV are not.

Consistency of a criterion means that in a large sample, if there are more than one equally good models, this criterion will choose the model with the fewest parameters. Schwarz (1978) has shown that BIC and HQC are consistent. However, AIC and AIC\(_c\) are not consistent because they might pick a model that has more parameters than necessary, thus having a probability of over-fitting.

Efficiency measures whether the model can achieve minimum error in forecasting. The selected estimator is efficient if it can achieve the lowest possible mean square error. For autoregressive models and normal linear regression, AIC, AIC\(_c\) and \( \text{Cp} \) are asymptotically efficient, while BIC and HQC are not (Claeskens and Hjort 2008). The LooCV and LsoCV methods have also been proved to be asymptotically efficient (Li 1987; Xu and Huang 2012). Furthermore, Claeskens and Hjort (2008) pointed out that efficiency and consistency cannot occur together. The properties of different methods are listed in Table 1. From this, it is clear that no MS criterion is good in all three properties. Therefore, a tradeoff is necessary when deciding which criterion to adopt in the MS process.

<table>
<thead>
<tr>
<th>Weak Consistency</th>
<th>AIC</th>
<th>BIC</th>
<th>HQC</th>
<th>( \text{Cp} )</th>
<th>LooCV</th>
<th>LsoCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consistency</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Efficiency</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1. Comparison of different MS methods.

Downloaded by [RMIT University] at 05:39 03 March 2016
5. Frequentist MA

Instead of selecting one model for prediction, MA selects several models from the potential models and forecasts using a weighted average of these model predictions. There are two strands of MA approaches, Bayesian model averaging (BMA) and FMA. The BMA method requires a prior probability for each candidate model and usually assumes that the true model is in the candidate set. This may weaken the attractiveness of this method (Draper 1995; Hoeting et al. 1999; Clyde and George 2004). For real-world problems, such as modelling the monthly throughput for container ports, there may not be a ‘true model’, or it may not be possible to construct a ‘true model’. Just as the famous econometrician George Box once said, ‘All models are wrong, but some are useful’ (Box 1979, 202). Therefore, in this paper, we mainly focus on the FMA methods.

Let $w_k$ ($w_k \in [0, 1], k = (1, \ldots, K), \sum_{k=1}^{K} w_k = 1$) be the weights for $K$ selected models in MA, then the MA estimator of $\mu$ can be expressed as

$$\hat{\mu}(w) = \sum_{k=1}^{K} w_k \hat{\mu}^{(k)}.$$  

In FMA, the most important issue is the choice of weights, and many weight selecting criteria have been developed based on MS methods.

5.1. Information-based MA

Buckland, Burnham, and Augustin (1997) suggested using the following Smoothed AIC (S-AIC) weight:

$$w_{AIC}^k = \exp \left( -\frac{1}{2} AIC_k \right),$$

where $AIC_k$ is the AIC score for model $k$. Note that the weight is proportional to $\exp \left( -\frac{1}{2} AIC_k \right)$. If the penalties are equal for the two models, $w_{AIC}^k / w_{AIC}^{k'}$ is just the ratio of the likelihoods, which is the Bayes factor for comparing simple models. Further, if the prior odds ratio is one, this expression represents the posterior odds ratio of the respective models (Buckland, Burnham, and Augustin 1997). Similarly, by replacing AIC with BIC and HQC, one can get the S-BIC, S-HQC weights, respectively. These kinds of S-xC MA estimators are popular because of their simplicity in calculation.

5.2. Asymptotically optimal MA

Let $L_n(w) = \| \mu - \hat{\mu}(w) \|^2$ denote the loss function of $\hat{\mu}(w)$. An estimator is asymptotically optimal if it satisfies

$$\frac{L_n(\hat{w})}{\inf_{w \in \mathcal{H}} L_n(w)} \overset{p}{\longrightarrow} 1,$$

where $\hat{w}$ is the weight vector obtained by minimising a certain weight selecting criterion.
5.2.1. Jackknife MA method
Hansen and Racine (2012) developed a jackknife model averaging (JMA) estimator by minimising an LooCV criterion
\[
C_{n}^{\text{JMA}}(w) = \tilde{e}(w)^t \tilde{e}(w),
\]
where \( \tilde{e}(w) = \sum_{k=1}^{K} w_k \tilde{e}^{(k)}(w) \) and \( \tilde{e}^{(k)}(w) \) is the leave-one-out estimated error defined in Section 4.3. This is the first theoretical research for model averaging allowing for heteroskedasticity. Zhang, Wan, and Zou (2013) extended this method to models with serially correlated errors.

5.2.2. Leave-subject-out MA method
Gao et al. (2015) proposed a leave-subject-out cross-validation model averaging (LsoMA) method which allows for within-subject correlation in panel data. This criterion was also utilised in time series models with serially correlated errors. The criterion is
\[
C_{n}^{\text{LsoMA}}(w) = \tilde{e}(w)^t \tilde{e}(w),
\]
where \( \tilde{e}(w) = \sum_{k=1}^{K} w_k \tilde{e}^{(k)}(w) \) and \( \tilde{e}^{(k)}(w) \) is the leave-subject-out estimated error as described in Section 4.3.

The averaging estimators \( \hat{\mu}(\hat{w}) \) with the weight selected by the above three criteria have been proved asymptotically optimal. Note that the estimator of these methods needs to be expressed in the form of \( \hat{\mu}^{(k)} = P^{(k)} Y \). This may limit the kinds of models that can be combined. For example, the estimator of an ARMA model with a moving average part cannot be written in this form. However, since any invertible ARMA process can be approximated by an AR(\( \infty \)) process, one can use a high-order AR model instead of moving average or ARMA models.

5.3. Adaptive MA
Adaptive MA is proposed by Yang (1999). Compared with the asymptotically optimal methods, adaptive MA can combine different kinds of models. Yang (2004) proposed an Aggregated Forecast Through Exponential Reweighting (AFTER) algorithm for the one-step ahead prediction based on the adaptive approach. We apply this method to combine \( h \)-step ahead prediction. The \( h \)-step ahead combined forecast is given by
\[
\hat{y}_{n,h}(w) = \sum_{k=1}^{K} w_{n-h+1}^{(k)} \hat{y}_{n,h}^{(k)},
\]
The weights for model \( k \) are calculated by
\[
w_{i,j,h}^{(k)} = \frac{\pi_k \prod_{j=0}^{i-1} \tilde{v}_{j,h}^{(k)}}{\sum_{k'=1}^{K} \pi_{k'} \prod_{j=0}^{i-1} \tilde{v}_{j,h}^{(k')}}^{-1/2} \exp \left( -\frac{1}{2} \sum_{j=0}^{i-1} \hat{y}_{j,h}^{(k)} - y_{j+h} \right)^2 \tilde{v}_{j,h}^{(k)}\]
where \( i = 1, \ldots, n - h + 1 \), \( \hat{y}_{0,h}^{(k)} \) and \( \tilde{v}_{0,h}^{(k)} \) are arbitrary initial guesses, \( \hat{y}_{j,h}^{(k)} \) is the \( h \)-step ahead prediction of model \( k \) given \( y_{j}, \ldots, y_{1}, \hat{y}_{1,h}^{(1)}, \ldots, \hat{y}_{h,h}^{(k)} \) are variances of \( \hat{y}_{j,h}^{(1)}, \ldots, \hat{y}_{h,h}^{(k)} \) for \( j \geq 1 \), and \( \pi_k \) is the prior probability of selecting model \( k \). If there is no prior knowledge of the candidate models and the size of candidate models is not large, one can use uniform weights.
The weights can be calculated through the following iteration form:

\[
 w_{0,h}^{(k)} = \pi_k \\
 w_{i,h}^{(k)} = \frac{w_{i-1,h}^{(k)} (\hat{v}_{i-1,h}^{(k)})^{-1/2} \exp \left( -\frac{1}{2} (\hat{y}_{i-1,h}^{(k)} - y_{i-1+h})^2 / \hat{v}_{i-1,h}^{(k)} \right)}{\sum_{k'=1}^K w_{i-1,h}^{(k')} (\hat{v}_{i-1,h}^{(k')})^{-1/2} \exp \left( -\frac{1}{2} (\hat{y}_{i-1,h}^{(k')} - y_{i-1+h})^2 / \hat{v}_{i-1,h}^{(k')} \right)}.
\] (16)

and \( \hat{v}_{i,h}^{(k)} \) is given by

\[
 \hat{v}_{i,h}^{(k)} = \frac{1}{i} \sum_{l=0}^{i-1} (y_{l+h} - \hat{y}_{l,h}^{(k)})^2.
\] (17)

When \( h = 1 \), Yang (2004) pointed out that this weighting method has a Bayesian interpretation: if we view the weights \( w_{i,1}^{(k)} \) as the prior probability of model \( k \) before observing \( y_i \), then \( w_{i,1}^{(k)} \) is the posterior probability of model \( k \) after \( y_i \) is seen. Yang (2004) also proved the bound for the risk of the AFTER forecast:

\[
 \frac{1}{n} \sum_{i=0}^{n-1} E \left( \frac{(\hat{y}_{i,1}(w) - E(y_{i+1}))^2}{\text{Var}(y_{i+1})} \right) \leq C_0 \min_{1 \leq k \leq K} \left\{ \frac{-2 \log \pi_k}{n} + C_1 \frac{n}{n} \sum_{i=0}^{n-1} E \left( \frac{(\hat{y}_{i,1}^{(k)} - E(y_{i+1}))^2}{\text{Var}(y_{i+1})} \right) \right. \\
 + \left. C_2 \frac{n}{n} \sum_{i=0}^{n-1} E \left( \frac{(\hat{v}_{i,1}^{(k)} - \text{Var}(y_{i+1}))^2}{\text{Var}(y_{i+1})^2} \right) \right\},
\] (18)

where \( C_j \) (\( j = 1, 2, 3 \)) are constants.

For the structural change VAR model, let

\[
 f_{i,h}^{(k)} = |\hat{\Sigma}_{i,h}(k)|^{-1/2} \exp \left\{ -\frac{1}{2} (y_{i+h} - \hat{y}_{i,h}^{(k)}) (\hat{\Sigma}_{i,h}(k))^{-1} (y_{i+h} - \hat{y}_{i,h}^{(k)}) \right\},
\]

where for \( i = 0 \), \( \hat{\Sigma}_{i,h}(k) \) is any initial guess, and for the other \( i \)'s,

\[
 \hat{\Sigma}_{i,h}(k) = \sum_{j=0}^{i-1} (y_{j+h} - \hat{y}_{j,h}^{(k)}) (y_{j+h} - \hat{y}_{j,h}^{(k)}) / i.
\]

Then the weights are calculated by

\[
 w_{i,h}^{(k)} = \begin{cases} 
 1/K, & i = 0, \\
 w_{i-1,h}^{(k)} / \sum_{k'=1}^K w_{i-1,h}^{(k')}, & i \geq 1.
\end{cases}
\] (19)

Compared with the asymptotically optimal methods, the AFTER method can combine different kinds of models. On the other hand, the weights given by the AFTER method are based on the idea of forecasting. Furthermore, we can get different weights according to the forecast step \( h \), which is more flexible than the asymptotically optimal methods.
Figure 1. Container throughput of HKP and SZP (in thousand TEUs).

6. Empirical results

To illustrate the application of the MS and model combination methods, the monthly container throughput data of Hong Kong port and the Shenzhen port from January 1999 to December 2014, as shown in Figure 1, are used in the empirical analysis. The data on the Shenzhen port are from the Ministry of Transportation of Mainland China (www.moc.gov.cn). For the Hong Kong port, data are available from the Hong Kong Marine Department (www.mardep.gov.hk). Since the raw data display an increasing trend and seasonal fluctuations, we converted the data into annual growth percentage points of monthly throughput, that is, $R_t = \frac{y_t - y_{t-12}}{y_{t-12}} \times 100$. The new series is depicted in Figure 2.

Note that the container throughput of both ports sharply decreased after October 2008, reflecting the significant and long-term impact of the US subprime crisis. The SC-VAR($m$) model is used to simultaneously analyse the Hong Kong port and Shenzhen port. Since there are structural changes in the time series after October 2008, this time is set as the break-point.

Table 2 presents a summary of statistics for the annual growth rate of monthly throughput. The statistics in column ‘Total’ are obtained by the whole sample, while those in column ‘Period 1 (or 2)’ are calculated using the sample before (or after) October 2008. The means and standard errors (SD) are quite different in the two periods, which indicates that we should deal with the data separately. Compared with the Shenzhen port, the growth rate in the Hong Kong port is much smaller both before and after the financial crisis. The growth rate in the Hong Kong port is negative in the second period, while that in the Shenzhen port is still positive. Overall, the growth rate is more volatile in the Shenzhen port.

An augmented Dickey-Fuller (ADF) test is applied to check the unit roots. The results in Table 3 show that the annual growth rate is nonlinear stationary autoregressive. Set the
possible length of lag $m = 1, \ldots, 12$. The Wald test introduced in Section 2 is applied to check whether the break-point is significant for all the candidate models. The $p$-values in the test results (Table 4) are all smaller than 0.05, indicating that the SC-VAR($m$) model is appropriate for all 12 models.

To compare the mean-squared forecast error (MSFE) of different methods, starting from January 2013 ($t_0$), we estimate the candidate models, select the best model and calculate the weights for MA methods using the data before that month, then do $h$-step ahead out-of-sample prediction, where $h = 1, 2, \ldots, 12$, and calculate the prediction errors. This process is repeated 24 times where the $t_0$ is extended by one month in each iteration. Denoting the prediction set by $\{z_l\}$, $l = 1, \ldots, 24$, the formula below can be used to calculate the MSFE
Table 4. Wald test results.

<table>
<thead>
<tr>
<th>m</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>55.65</td>
<td>33.78</td>
<td>35.92</td>
<td>45.40</td>
<td>66.04</td>
<td>79.70</td>
</tr>
<tr>
<td>p-Value</td>
<td>3.42E−10</td>
<td>2.02E−14</td>
<td>1.07E−23</td>
<td>3.63E−24</td>
<td>2.74E−06</td>
<td>2.33E−07</td>
</tr>
</tbody>
</table>

Table 5. Mean-squared forecast errors of container throughput ($\times 10^{-2}$).

<table>
<thead>
<tr>
<th>h = 1</th>
<th>AFTER</th>
<th>LsoMA</th>
<th>JMA</th>
<th>S-AIC</th>
<th>S-BIC</th>
<th>S-HQC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6293(12)</td>
<td>0.4288(1)</td>
<td>0.4459(2)</td>
<td>0.5187(8)</td>
<td>0.5171(7)</td>
<td>0.4913(3)</td>
</tr>
<tr>
<td>h = 2</td>
<td>1.1534(12)</td>
<td>0.4935(1)</td>
<td>0.5246(2)</td>
<td>0.5784(7)</td>
<td>0.6152(9)</td>
<td>0.5757(5)</td>
</tr>
<tr>
<td>h = 3</td>
<td>1.2399(12)</td>
<td>0.5246(3)</td>
<td>0.5614(8)</td>
<td>0.7928(9)</td>
<td>0.5160(1)</td>
<td>0.5299(7)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>h = 12</td>
<td>0.9839(12)</td>
<td>0.5526(1)</td>
<td>0.5627(2)</td>
<td>0.7273(9)</td>
<td>0.5865(4)</td>
<td>0.6222(6)</td>
</tr>
<tr>
<td>Mean</td>
<td>1.0623(12)</td>
<td>0.4821(1)</td>
<td>0.5176(4)</td>
<td>0.7022(10)</td>
<td>0.5111(2)</td>
<td>0.5241(8)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>h = 1</th>
<th>Cp</th>
<th>LsoCV</th>
<th>LooCV</th>
<th>AIC</th>
<th>BIC</th>
<th>HQC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5073(6)</td>
<td>0.4919(4.5)</td>
<td>0.5901(11)</td>
<td>0.5188(9)</td>
<td>0.5226(10)</td>
<td>0.4919(4.5)</td>
</tr>
<tr>
<td>h = 2</td>
<td>0.5888(8)</td>
<td>0.5664(3.5)</td>
<td>0.7508(11)</td>
<td>0.5783(6)</td>
<td>0.6221(10)</td>
<td>0.5664(3.5)</td>
</tr>
<tr>
<td>h = 3</td>
<td>0.5255(4)</td>
<td>0.5298(5.5)</td>
<td>0.995(11)</td>
<td>0.7929(10)</td>
<td>0.5167(22)</td>
<td>0.5298(5.5)</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>h = 12</td>
<td>0.6034(5)</td>
<td>0.6296(7.5)</td>
<td>0.9395(11)</td>
<td>0.7274(10)</td>
<td>0.5849(3)</td>
<td>0.6296(7.5)</td>
</tr>
<tr>
<td>Mean</td>
<td>0.5185(5)</td>
<td>0.5235(6.5)</td>
<td>0.9675(11)</td>
<td>0.7021(9)</td>
<td>0.5116(3)</td>
<td>0.5235(6.5)</td>
</tr>
</tbody>
</table>

and to compare the performance of different methods:

$$\frac{1}{25-h} \sum_{l=1}^{25-h} \frac{(z_{l+h-1} - \hat{z}_{l+h-1})^2}{z_{l+h-1}^2}.$$ (20)

The results are shown in Table 5. To save space, we only list the results when $h = 1, 2, 3, 12$. For easy comparison, we also calculated the mean of the MSFEs of the 12 predictions. The number in brackets is the rank of the model selected by each method, according to its MSFE or its mean. The upper part of the table is the MSFEs of the six MA methods, while the lower part is the MSFEs of the best model selected by each MS method.

Although the ranks of $x_C$ and its corresponding averaging method $S-x_C$ are different, they actually perform quite similarly, because in the weight calculation the $S-x_C$ method puts almost all the weights on the model with the smallest $x_C$ value. For the CV methods, the MA estimators can achieve lower MSFE than its corresponding MS estimators. Note that LsoMA is better than JMA in all cases. From the numerical result, the average MSFE of LsoMA is only $0.4821 \times 10^{-2}$, much smaller than any other methods. Thus, LsoMA performs the best, then comes S-BIC.

To show the properties of the MA method, we also calculate the mean absolute percentage forecast error (APFE) using both the LsoMA method and that from a univariate SARIMA model (Peng and Chu 2009). The results are given in Table 6. The LsoMA method performs better in most cases, especially when $h$ is large. Based on these results, the LsoMA method is used to forecast the monthly container throughput.

The estimated coefficients using LsoMA are listed in Table 7. The first two columns are the estimated coefficients for the Hong Kong port following Equation (1). Period 1 (or 2) indicates before (or after) October 2008. The next two columns are that for the Shenzhen port.
Table 6. Mean absolute forecast errors of container throughput (%).

<table>
<thead>
<tr>
<th></th>
<th>$h = 1$</th>
<th>$h = 2$</th>
<th>$h = 3$</th>
<th>$h = 4$</th>
<th>$h = 5$</th>
<th>$h = 6$</th>
<th>$h = 7$</th>
<th>$h = 8$</th>
<th>$h = 9$</th>
<th>$h = 10$</th>
<th>$h = 11$</th>
<th>$h = 12$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HK</td>
<td>5.07</td>
<td>5.42</td>
<td>6.16</td>
<td>5.57</td>
<td>4.96</td>
<td>4.68</td>
<td>4.83</td>
<td>4.78</td>
<td>4.91</td>
<td>5.44</td>
<td>5.38</td>
<td>5.88</td>
</tr>
<tr>
<td>LsoMA</td>
<td>5.43</td>
<td>4.13</td>
<td>5.94</td>
<td>5.43</td>
<td>5.33</td>
<td>5.80</td>
<td>6.59</td>
<td>6.08</td>
<td>7.62</td>
<td>6.55</td>
<td>6.94</td>
<td>7.49</td>
</tr>
<tr>
<td>SARIMA</td>
<td>h = 1</td>
<td>h = 2</td>
<td>h = 3</td>
<td>h = 4</td>
<td>h = 5</td>
<td>h = 6</td>
<td>h = 7</td>
<td>h = 8</td>
<td>h = 9</td>
<td>h = 10</td>
<td>h = 11</td>
<td>h = 12</td>
</tr>
<tr>
<td>SZ</td>
<td>5.32</td>
<td>5.14</td>
<td>5.19</td>
<td>5.04</td>
<td>4.96</td>
<td>5.28</td>
<td>5.02</td>
<td>4.92</td>
<td>5.10</td>
<td>4.15</td>
<td>4.36</td>
<td>4.44</td>
</tr>
<tr>
<td>LsoMA</td>
<td>5.32</td>
<td>5.14</td>
<td>5.19</td>
<td>5.04</td>
<td>4.96</td>
<td>5.28</td>
<td>5.02</td>
<td>4.92</td>
<td>5.10</td>
<td>4.15</td>
<td>4.36</td>
<td>4.44</td>
</tr>
<tr>
<td>SARIMA</td>
<td>4.17</td>
<td>5.83</td>
<td>6.15</td>
<td>6.53</td>
<td>4.88</td>
<td>4.76</td>
<td>5.76</td>
<td>4.72</td>
<td>5.27</td>
<td>6.59</td>
<td>5.57</td>
<td>6.60</td>
</tr>
</tbody>
</table>

Table 7. Estimated coefficients of LsoMA.

<table>
<thead>
<tr>
<th></th>
<th>Eq. HK</th>
<th>Eq. SZ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Period 1</td>
<td>Period 2</td>
</tr>
<tr>
<td>Intercept</td>
<td>2.8010</td>
<td>−0.9554</td>
</tr>
<tr>
<td>HK.1</td>
<td>0.2217</td>
<td>0.4624</td>
</tr>
<tr>
<td>SZ.1</td>
<td>0.0758</td>
<td>0.0920</td>
</tr>
<tr>
<td>HK.2</td>
<td>0.2197</td>
<td>0.2673</td>
</tr>
<tr>
<td>SZ.2</td>
<td>−0.0575</td>
<td>0.0255</td>
</tr>
<tr>
<td>HK.3</td>
<td>0.0037</td>
<td>−0.1087</td>
</tr>
<tr>
<td>SZ.3</td>
<td>−0.0400</td>
<td>0.0340</td>
</tr>
<tr>
<td>HK.4</td>
<td>−0.0552</td>
<td>−0.0176</td>
</tr>
<tr>
<td>SZ.4</td>
<td>0.0661</td>
<td>−0.0216</td>
</tr>
<tr>
<td>HK.5</td>
<td>0.0811</td>
<td>0.0676</td>
</tr>
<tr>
<td>SZ.5</td>
<td>−0.0197</td>
<td>−0.0510</td>
</tr>
<tr>
<td>HK.6</td>
<td>−0.0330</td>
<td>−0.0082</td>
</tr>
<tr>
<td>SZ.6</td>
<td>−0.0099</td>
<td>−0.0181</td>
</tr>
<tr>
<td>HK.7</td>
<td>−0.0376</td>
<td>0.0153</td>
</tr>
<tr>
<td>SZ.7</td>
<td>−0.0030</td>
<td>0.0140</td>
</tr>
<tr>
<td>HK.8</td>
<td>0.0065</td>
<td>0.0328</td>
</tr>
<tr>
<td>SZ.8</td>
<td>0.0156</td>
<td>−0.0304</td>
</tr>
<tr>
<td>HK.9</td>
<td>0.0646</td>
<td>−0.1301</td>
</tr>
<tr>
<td>SZ.9</td>
<td>−0.0145</td>
<td>0.0613</td>
</tr>
<tr>
<td>HK.10</td>
<td>−0.0897</td>
<td>0.0729</td>
</tr>
<tr>
<td>SZ.10</td>
<td>−0.0125</td>
<td>0.0211</td>
</tr>
<tr>
<td>HK.11</td>
<td>0.0215</td>
<td>0.1074</td>
</tr>
<tr>
<td>SZ.11</td>
<td>0.0139</td>
<td>−0.0134</td>
</tr>
<tr>
<td>HK.12</td>
<td>−0.1338</td>
<td>−0.2184</td>
</tr>
<tr>
<td>SZ.12</td>
<td>0.0184</td>
<td>−0.0031</td>
</tr>
</tbody>
</table>

HK$_p$ (or SZ$_p$) denotes the $p$th lag of the annual growth rate of monthly throughput in the Hong Kong port (or the Shenzhen port). From this table, we can see that the absolute values of the coefficients of the 1st to 2nd and 12th lags are relatively high, which indicates that the annual growth rate of the previous two months and that of one year ago have greater impacts on the annual growth rate of the current month. The impacts of the previous two months are positive, while that for the last year are negative (except for the Shenzhen port in period 2). This reveals a positive correlation between the current growth rate and that of the last two months, and a negative one with that of one year ago.

The intercepts of both ports decrease after October 2008, indicating that the financial crises slowed down the economic growth. The estimated coefficients have both positive and negative signs. The positive sign indicates the influence of international trade on the relationship between the annual growth rate of the current month and the previous 12 months. In eq.HK, the value of the coefficient for the lag variable of the Hong Kong port is, in most cases, larger than that of the Shenzhen port. This indicates that the growth rate of monthly container throughput at the Hong Kong port has a higher correlation with its past
Table 8. Counts of negative coefficients.

<table>
<thead>
<tr>
<th></th>
<th>Eq. HK</th>
<th>Eq. SZ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Period 1</td>
<td>Period 2</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>HK</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>SZ</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 3. Prediction results using LsoMA.

growth rates than the Shenzhen port. However, in eq.SZ, the absolute value of the coefficient for the Hong Kong port is not very small compared with that of the Shenzhen port in eq.HK. This may be because both Shenzhen and Hong Kong ports serve the international trade to and from the Pearl River Delta (PRD) region. In addition, the Hong Kong port also has a large transshipment volume, but not the Shenzhen port. Therefore, any change in the trade volume in the PRD region will have a greater impact on the Shenzhen port than on the Hong Kong port. The Hong Kong port handles both direct cargo and transshipment. Therefore, the growth rate in the Shenzhen port has less of an influence on the Hong Kong port.

While the positive sign of the coefficients indicates the influence of international trade in the PRD region, the negative sign may be caused by the inter-temporal effect of the competitive relationship between the two ports. The number of negative signs for each equation in each period in Table 7 is summarised in Table 8. From this, we can see that the number of cross-port negative impacts is similar. The Shenzhen port has 7 (6) negative impacts to the Hong Kong port before (after) the financial crisis, while the Hong Kong port has 7 negative impacts on the Shenzhen port in both periods. Also, the cross-port negative impacts become very similar in period 2, maybe because of the similar competitive power between these two ports after the financial crisis.

The predictions for 2015 throughput are depicted in Figure 3, together with the actual throughput from January 2014 to April 2015 for the Hong Kong port and to June 2015 for the Shenzhen port. The dashed line and solid line represent the actual value and predicted value, respectively. Compared with the throughput in 2014, the Hong Kong port is decreasing, while the Shenzhen port is showing a slight increase. The Hong Kong port will reach a peak in the second quarter of 2015, while the Shenzhen port will reach a peak in the third
quarter. The actual throughput, the predicted value and the APFEs of the two ports are given in Table 9. The APFEs of the Hong Kong port in January and April 2015 are relatively small. For the Shenzhen port, most prediction results are quite good except for February 2015. This is because the prediction result is related to the value of the last year, and the container throughput of the Shenzhen port dropped sharply in February last year.

7. Discussion and conclusion

This paper demonstrates the application of six commonly used MS criteria, including AIC, BIC, HQC, \( C_p \), LooCV and LsoCV, and six FMA methods, namely S-AIC, S-BIC, S-HQC, JMA, LsoMA and AFTER. Taking into account any possible competition between the two nearby ports, and possible structural changes due to the 2008 financial crisis, we developed a SC-VAR model with a known break point, to model the changes in the annual increase rate of monthly container throughput in Hong Kong and Shenzhen ports. The SC-VAR model differs only in the length of the lag. The MSFE is calculated for every selected model and the MA method, for the prediction steps 1–12. The mean of the MSFE is used to rank the performance of each method. From the ranking, the LsoMA method is found to be the best.

The estimated coefficients of each equation show different patterns of cargo origin between these two ports. The Hong Kong port has both direct PRD cargo and transshipment cargo. Therefore, the annual growth rate of monthly throughput is more correlated with its own lagged variables. The negative signs in two equations reveal the inter-temporal effects of the competitive relationship between the two ports. In addition, the cross-port negative impact becomes similar after the financial crisis, indicating the similar competitive power of these two ports.

It has to be noted that the ranking of different models is valid just for this result. Time series data with different properties may generate a different ranking. In addition, this study does not include all possible MS or MA methods. There are many potential methods for MS (Konishi and Kitagawa 1996) and MA (Liang et al. 2011). For different time series, different selection criteria or averaging methods may be used by applying similar procedures to those used in this paper. To provide short-term prediction and satisfy the operational needs, it is better to repeat the MS and MA process whenever new data become available. Therefore, the value of this paper is not with regard to the prediction for any specific month, but with regard to the methodology.

For further research along this line, one could try to use VAR models with multiple unknown change points. In addition, some exogenous factors, such as the number of shipping berths, Hong Kong GDP and the value of foreign trade can be added into the model.

<table>
<thead>
<tr>
<th>Table 9. Predicted monthly throughput from January 2015.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HK</td>
</tr>
<tr>
<td>2015</td>
</tr>
<tr>
<td>January</td>
</tr>
<tr>
<td>February</td>
</tr>
<tr>
<td>March</td>
</tr>
<tr>
<td>April</td>
</tr>
<tr>
<td>May</td>
</tr>
<tr>
<td>June</td>
</tr>
</tbody>
</table>
Also, container port throughput includes export, import, transhipment and empty containers. These are all relatively independent, but may also be correlated. If the time series data on these variables are available, a VAR model could also be applied to study the trend of such variables explicitly, which would provide even better information for port operation and management. Finally, we revealed the inter-temporal effect of port competition, but not the contemporaneous effects, which could also be a direction for future research.

Acknowledgements
We thank the Editors-in-Chief Professor William H.K. Lam, Professor S.C. Wong and the anonymous referees for their constructive comments and suggestions which greatly improved the original manuscript.

Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
This research is supported by the Dean’s Reserve of Faculty of Business, The Hong Kong Polytechnic University (Project Code 1-ZVBS), the RGC general research fund [RGC/Gov. No.: PolyU 5463/12H], and the National Natural Science Foundation of China [11331011 and 11271355].

Notes
1. The critical value is set to $1.96/\sqrt{n}$ in the empirical study.
2. It is recognised that now, both Hong Kong and Shenzhen are subjected to the competition from Guangzhou port, in serving the imports and exports of Pearl River Delta (PRD) area. However, it was just a local port and only had the first international liner calling at 2006. Therefore, the number of observations for Guangzhou port is not enough for model training before 2008.

ORCID
Meifeng Luo http://orcid.org/0000-0001-5044-2737

References


**Appendix**

**Derivations of formulas (3) and (8)**

**The Wald test**

According to Hamilton (1994) and White (1984), under some regularity conditions, we have

\[
\sqrt{n}(\hat{\phi}^{(1)} - \phi^{(1)}) \xrightarrow{d} N\left(0, \frac{\Omega}{\tau (1 - \tau)}\right),
\]

where \(\tau = \lim_{n \to \infty} n_1/n\) lies in (0, 1), \(\Omega = \Sigma \otimes Q^{-1}\) and \(Q = E(x_t x'_t) > 0\).

Under \(H_0\), \(\phi^{(1)} = \phi^{(2)} = \phi\), then we have

\[
\sqrt{n}(\hat{\phi}^{(1)} - \hat{\phi}^{(2)}) \xrightarrow{d} N\left(0, \frac{\Omega}{\tau (1 - \tau)}\right).
\]

Let \(X'X/n\) be the estimator of \(Q\), then the Wald test statistic is

\[
W = n_1(1 - n_1/n)(\hat{\phi}^{(1)} - \hat{\phi}^{(2)})'\hat{\Omega}^{-1}(\hat{\phi}^{(1)} - \hat{\phi}^{(2)}) \xrightarrow{d} \chi^2_{4m+2}.
\]

One can reject the null hypothesis when \(W > \chi^2_{4m+2}(\alpha)\), where \(\alpha\) is the given significance level and \(\chi^2_{4m+2}(\alpha)\) is the upper \(\alpha\) quantile.

**The Mallows’ \(C_p\) criterion**

For regime one, denote \(X_{(1)} = (x_{11}, \ldots, x_{n1})'\), \(\mu_{(1)} = (\mu_{(1)}, \ldots, \mu_{n1})'\) and \(Y_{(1)} = (y_{11}, \ldots, y_{n1})'\). Let \(e_{(1)} = (e_{11}, e_{21}, \ldots, e_{n1})'\) and \(e_{(2)} = (e_{12}, e_{22}, \ldots, e_{n2})'\) be the error term of Equations (1) and (2), respectively. Similarly, we can define \(\mu_{(1)}\) and \(\mu_{(2)}\). Let \(\hat{\mu}_{(1)}, \hat{\mu}_{(1)},\) and \(\hat{\mu}_{(2)}\) be the corresponding OLS estimators and \(P_{(1)} = X_{(1)}(X_{(1)}'X_{(1)})^{-1}X_{(1)}'\). It can be easily shown that

\[
\|\mu_{(1)} - \hat{\mu}_{(1)}\|^2 = \|Y_{(1)} - \hat{\mu}_{(1)}\|^2 + e_{(1)}'e_{(1)} - 2e_{(1)}'(Y_{(1)} - \hat{\mu}_{(1)}) - 2e_{(1)}'(Y_{(1)} - \hat{\mu}_{(1)}) - \hat{\mu}_{(1)}
\]

\[
= \|Y_{(1)} - \hat{\mu}_{(1)}\|^2 + e_{(1)}'e_{(1)} - 2e_{(1)}'(I - P_{(1)})\mu_{(1)} - 2e_{(1)}'(I - P_{(1)})\mu_{(1)} + 2e_{(1)}'P_{(1)}e_{(1)} + 2e_{(1)}'P_{(1)}e_{(1)}.
\]

Hansen (2009) has shown that \(e_{(1)}'(I - P_{(1)})\mu_{(1)}\) and \(e_{(1)}'(I - P_{(1)})\mu_{(1)}\) have an approximate mean of zero and are therefore ignored and that

\[
e_{(1)}'e_{(1)} \xrightarrow{d} \sigma_Y^2 \chi^2_p \text{ and } e_{(1)}'P_{(1)}e_{(1)} \xrightarrow{d} \sigma^2 \chi^2_p,
\]

where \(p = \text{tr}(P_{(1)})\). Thus, an asymptotically unbiased estimate of \(\|\mu_{(1)} - \hat{\mu}_{(1)}\|^2 + e_{(1)}'e_{(1)}\) is given by

\[
\|Y_{(1)} - \hat{\mu}_{(1)}\|^2 + 2p \text{ tr}(\hat{\Sigma}).
\]

We can obtain a similar result for regime two. Based on the deviation above, we propose to use the following criterion for model \(k\)

\[
C_p(k) = \|Y - \hat{Y}^{(k)}\|^2 + p_k \text{ tr}(\hat{\Sigma}(k)).
\]