Container Demand Forecasting at Border Posts of Ports: A Hybrid SARIMA-SOM-SVR Approach

Juan Jesús Ruiz-Aguilar(✉), Daniel Urda, José Antonio Moscoso-López, Javier González-Enrique, and Ignacio J. Turias

Intelligent Modelling of Systems Research Group, Polytechnic School of Engineering (Algeciras), University of Cadiz, Avda. Ramon Puyol s/n, 11202 Algeciras, Spain
juanjesus.ruiz@uca.es

Abstract. An accurate forecast of freight demand at sanitary facilities of ports is one of the key challeng-es for transport policymakers to better allocate resources and to improve planning operations. This paper proposes a combined hybrid approach to predict the short-term volume of containers passing through the sanitary facilities of a maritime port. The proposed methodology is based on a three-stage process. First, the time series is decomposed into similar smaller regions easier to predict using a self-organizing map (SOM) clustering. Then, a seasonal auto-regressive integrated moving averages (SARIMA) model is fitted to each cluster, obtaining predicted values and residuals of each cluster. A support vector regression (SVR) model is finally applied in each cluster using the historical data clustered and the predicted variables from the SARIMA step, testing different hybrid configurations. The experimental results demonstrated that the proposed model outperforms other methodologies based on SVR. The proposed model can be used as an automatic decision-making tool by seaport or airport management due to its capacity to plan resources in advance.

Keywords: Container forecasting · Machine learning · Support vector regression · Self-organizing maps · Hybrid models

1 Introduction

The Border Inspection Posts (BIPs) were created in order to guarantee the security at border crossings and the quality of the import-export goods by inspecting them. BIPs are the approved facilities where the checks of goods (transported within containers by trucks or towing vehicles) are carried out before entering the Community territory. Thus, the BIPs are bottlenecks that must be necessarily considered by Port Authorities. In order to avoid time delays and congestion in the sanitary facilities, the port management must be able to accurately forecast

B. Dorronsoro et al. (Eds.): OLA 2020, CCIS 1173, pp. 69–81, 2020.
https://doi.org/10.1007/978-3-030-41913-4_7

the number of container passing through these sanitary facilities. An accurate prediction of this volume may become a useful tool to improve human resources, planning operations and the service quality at ports. In this paper, the forecasting techniques can be divided into three categories: single methods, combined methods and hybrid methods.

The first class comprises both linear and nonlinear techniques. On the one hand, linear techniques are based on the assumption of having a linear relationship between the future values and the current and past values of the time series. The well-known autoregressive integrated moving averages (ARIMA) [2] models have been constantly applied to solve forecasting tasks related to maritime transport [1,5]. On the other hand, nonlinear techniques have become a strength alternative against the weaknesses of linear models. In this subcategory, two techniques must be highlighted: artificial neural networks (ANNs) and support vector machines for regression (SVR). Due to its great generalization ability, SVR has been used in forecasting transport tasks with promising results. Some examples include a predicting of container throughputs, inspection freights and roll on-roll of freight traffic at ports [7–9]. Their findings showed that SVR makes more accurate predictions than ANNs.

The second category comprises the combined models. One of the most frequently approach consists of combining a single prediction technique with a clustering method. When the clustering method has divided the database into several clusters, a prediction technique is then applied in each cluster independently. Self-organizing maps (SOMs) [6] is probably the best-known clustering method. A combined SOM-ANN model was firstly introduced by Chen et al. [3] to predict traffic flows in transportation. Results showed that the SOM-ANN model outperformed the rest of the models. Due to the recent emergence of SVR in transportation, there is hardly any research related to transport combining SOM and SVR in a two-stage procedure. Nevertheless, it is a widespread solution in many other forecasting fields [4].

The third category includes hybrid models. Real-world time series are not completely linear or nonlinear, but rather contain both components. Thus, a methodology using linear and non-linear models in a hybrid way takes the capabilities of both models. Hybridizing linear and non-linear models have been proposed in recent years. ARIMA has been the most commonly used linear model in hybrid models literature. Several authors have proposed a hybridization of SARIMA and SVR to address several forecasting tasks in the transport sector. As an example, Xie et al. [11] proposed several hybrid approaches in a comparative way including the SARIMA-SVR model for container throughput forecasting. Authors pointed out that a hybrid strategy considering ARIMA and SVR models overcomes the performance of single models.

In this study, a combined-hybrid forecasting model is proposed in such a way that a hybrid model (SARIMA-SVR) is combined with a clustering method (SOM) to forecast the daily number of containers passing through a BIP, thus resulting in a new SOM-SARIMA-SVR strategy. This methodology unifies in a single model the strengths of clustering methods in decomposing the forecasting task into some relatively easier subtasks (using a SOM method) together with the strengths of hybrid models to fit linear and nonlinear components (using a SARIMA-SVR model).

2 Brief Introduction to SOM, SARIMA and SVR Models

Self-organizing Maps (SOM)

Within the unsupervised learning field, a SOM is a kind of neural network. First proposed by Kohonen [6], a SOM is a classification technique which groups objects of the systems into regions called clusters. In the process, neurons of the model organize themselves considering only those that play a similar role, forming a cluster. The topology of a SOM model is based on several neurons distributed into two layers. The first one (input layer) is formed by k neurons and each neuron correspond with one input. The output layer, called the competition layer, can consist on different topologies (2-D grid for this case) where the preprocessing is performed. All the neurons of the output layer are connected by weights with all neurons of the input layer. Different input vectors $x_i = [x_1, x_2, \ldots, x_k]^k$ are presented to the networks at each training iteration. During the network training, the Euclidean distance between x and all the weight vectors are computed as follows:

$$\|x - w_b\| = \min_i \left\{ \left\| x(t) - w\hat{i} \right\| \right\} \qquad i = 1, 2, \ldots, l \tag{1}$$

where l is the number of output neurons. According to Eq. (1), w_b is considered the winning neuron, i.e. the neuron that has the weight vector closest to x. In addition, the weight of the winning neuron and their neighbours are updated in a learning procedure by which the outputs become self-organised and the feature map between inputs and outputs is formed. It is worth mentioning that the neighbours will have their weights updated as well, although by not as much as the winner itself. The weight update equation, Eq. (2), has a time (epoch) dependent and descendent learning rate $\alpha(t)$, and a neighbour function N.

$$W(t+1) = W(t) + N(v,t)\alpha(t)(x - W(t)) \tag{2}$$

Auto-regressive Integrated Moving Averages (ARIMA)

ARIMA models were introduced by Box and Jenkins [2] and have been a widely used forecasting linear model during several decades. Three prediction terms compose this linear function: the autoregressive term (AR), the moving average term (MA) and the integration term (I). A SARIMA model can be obtained by extending the ARIMA model to incorporate seasonal features. In this way, the model is specified as SARIMA$(p, d, q)(P, D, Q)_S$, where q represent the order of the moving average terms, p denotes the order of the autoregressive terms and d is the degree of differencing. The parameters (P, D, Q) deals with the seasonal part and the capital letters corresponds to their counterparts for the seasonal models with the seasonal orders and the seasonality of the model is represented by the parameter s. Equation (3) depicts a typical expression of the SARIMA model.

$$\varphi_p(L)\Phi_P(B^S)\nabla^d\nabla_S^D y_t = \theta_q(B)\Theta_Q(B^S)a_t \tag{3}$$

where y_t is the observed value, ∇^d and ∇_S^D are the regular and seasonal differencing operators, respectively, p and P are the number of non-seasonal and seasonal autoregressive terms, q and Q are the number of non-seasonal and seasonal moving average terms, d and D are the number of regular and seasonal differences, φ and ϕ deptic the value weights of the non-seasonal and seasonal autoregressive term, θ and Θ represent the weights of the non-seasonal and seasonal moving average term, the seasonality is represented by S and at is the noise term.

Support Vector Machines for Regression (SVR) Models

Support vector machines (SVM) is a kind of machine learning system focused on the structural risk minimization. The main objective of this method is maximizing the margin distance [10]. First introduced for classification problems, the ε-insensitive loss function, has enabled its use in regression problems. The process is the following: first, the input data are mapped into a new space of higher dimensional features, called feature space, by a non-linear mapping a priori using a kernel transformation. The aim of this feature space is to detect a linear regression function that can be fit the output data with the input data. This linear regression corresponds to the nonlinear regression model in the original space and it can be expressed as in Eq. (4). The following Equation represents the problem that should be optimized:

$$\begin{aligned} \underset{w,b,\xi}{min}\,\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i^+ + \xi_i^-) \\ \text{subject to:} \\ w\cdot x_i + b - y_i \leq \epsilon + \xi_i^+ \\ y_i - w\cdot x_i - b_i \leq \epsilon + \xi_i^- \\ \xi_i^+, \xi_i^- \geq 0 \end{aligned} \tag{4}$$

with $i = 1,\ldots,l$ ξ_i^- and ξ_i^+ are slack variables that deal with the training error on the top and the bottom, respectively. The expression $\|w\|^2/2$ defines the structure risk concerning the flatness of the model and the parameter C is a correction factor which deals with the trade-off between the flatness and the error. Gaussian kernel was chosen as kernel function. The dual optimization problem can be solved with the Lagrangian multiplier method. The main reason for using Lagrange Multipliers is that it is not very difficult to setup the problem. The critical thing to note is that Lagrange multipliers only works with equality constraints and therefore it is necessary to rearrange them. The result is a fairly complicated system of equations, but there are methods to solve these. Using Karush-Kuhn-Tucker conditions, we can substitute these into the primal equation, rearrange and solve [10].

3 Forecasting Approach

The experimental database comes from the BIP of the Port of Algeciras Bay, located in the South of Spain. The Port of Algeciras Bay was the port with maximum throughput in the Mediterranean Sea and the fourth port in the European continent related to the total throughput in 2018. The database was provided by the Port Authority and contains daily records of the number of containers at the Algeciras BIP from 2010 to 2014, which makes a total of 1825 daily records.

3.1 The Proposed Hybrid Methodology

The proposed methodology consist on a three-step hybrid procedure to forecast the daily number of containers passing through a BIP. Different prediction horizons were assessed: one-day ($ph = 1$) and seven-day ($ph = 7$) ahead. The prediction was one-step ahead (y_{t+ph}), that is y_{t+1} and y_{t+7}. The estimation can be thereby modelled as a nonlinear function of the n preceding values of the series, called the autoregressive window (n) and its design is presented in Fig. 1. For the $ph = 7$ case, the autoregressive window is composed of values of the container series periodically sampled every seven days in the past. This is due to the weekly seasonality found in the analysis of the autocorrelation function of the time series. The main assumption here is that the best predictions are obtained when past inputs corresponding to the same day of the week are used (e.g., using several successive Mondays in the past to predict a future Monday).

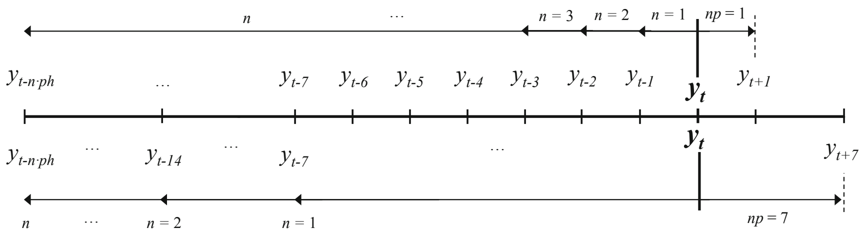

Fig. 1. Possible autoregressive window sizes in Steps II and III and their prediction horizons (ph): one-day prediction horizon (above the timeline) and seven-day prediction horizon (below the timeline). n is the size of the auto-regressive window.

Step I: SOM. A SOM model is first applied to the data in order to split the data-base in several disjoint groups, called clusters, with similar statistical distribution. Each cluster works independently in the second and third step. In such cases, a single SARIMA and SVR models are applied independently after decomposing the heterogeneous data into different homogeneous regions. An experimental framework was developed in order to select the optimal number of past values to be considered (nc) in the input vector of the SOM, which is described in Eq. (5):

$$x_i = [y_t, y_{t-1\cdot ph}, y_{t-2\cdot ph}, \ldots, y_{t-nc\cdot ph}]^T \quad (5)$$

where t is each sample (daily value). Each row is arranged recursively using different lagged terms (as in an autoregressive window).

Step II: SARIMA. A SARIMA model is fitted to each cluster generated by the SOM in Step I, obtaining different predicted and residual values of these clusters. A hold-out validation technique was applied during the process. The data (of each cluster) was divided into two groups: the training set containing two thirds of the dataset, and the test set comprising the rest of the samples. The parameters of the model were adjusted using the training set and the test set was used to validate the model. Different parameter ranges were tested using a trial-and-error procedure. The values of the parameters tested within each cluster are, for the non-seasonal part: $p = 0, 1, 2, 3, 4$; $d = 0, 1, 2$ and $q = 1, 2, 3, 4$; and for the seasonal part: $s = 2, 5, 7$; $P = 0, 1, 2$; $D = 0, 1, 2$ and $Q = 0, 1, 2, 3$. All the possible combinations of parameters were tested.

Step III: SVR. A SVR model is again applied to each generated cluster. The three different SVR parameters are determined by an iterative process (trial-and-error). For each cluster, the inputs of the SVR model are composed by the original data of the cluster and their forecasted values and residuals from the second (SARIMA) step. Thus, three different groups of variables compose the inputs of each cluster: the forecasted values and residuals from the SARIMA step, p_i and e_i respectively, and the original data y_i, where i denotes the cluster. The presence of these variables within the inputs leads to the proposed hybrid configurations. Each variable is sorted recursively in terms of an autoregressive window. The sizes of the original data, predicted values and residuals from the SARIMA step are denoted as ny, np and ne, respectively. The range of parameter tested in each cluster and each ph were $ne, ny, np = [1, 2, \ldots, 20]$ and, for the SVR parameters, $\epsilon, \gamma = [2^{(-12,-11,\ldots,-2)}]$ and $C = [1, 2, \ldots, 10, 50, 100, 200, \ldots, 1000]$. For each possible combination of the autoregressive parameters (ne, ny, np), all the possible combinations of the hyperparameters (C, ϵ, γ) were tested.

A twofold cross-validation (2-CV) technique was used. First, 2-CV divides the database into two sets (training and test) of equal sizes. The model determines the optimal hyperparameters with the training set. Then, the performance accuracy is computed by the training set. The sets are subsequently inverted and the process is computed again, obtaining the average of the two steps. This validation strategy was repeated 20 times and the final prediction performance was the average of these repetitions. The whole predicted time series is achieved by adding the predictions of each available clusters. Note that, as in the SARIMA model, the best SVR model may be different on each cluster. Two hybrid approaches were proposed and assessed. The prediction results were obtained for two prediction horizons, $ph = 1$ and $ph = 7$.

SOM-SARIMA-SVR-1 Model (Hybrid Approach 1)
The time series can be decomposed into two independent and additive terms: a linear component L_t and a nonlinear component NL_t. Then, a linear forecasting model such as SARIMA can be applied in order to model the linear component and thereby to obtain the predicted values denoted as $\hat{p}_t$ and the residual e_t. Subsequently, a SVR model is applied over the residuals to fit the nonlinear component NL_t:

$$\widehat{NL}_{t+ph} = f(e_t, e_{t-ph}, \ldots, e_{t-n\cdot ph}) + \varepsilon_t = \hat{e}_{t+ph} \tag{6}$$

where $\hat{e}_t$ is the predicted residual, f is the nonlinear function obtained by the SVR model, n is the size of the autoregressive window, ph is the prediction horizon and ε_t is the error term. Finally, the prediction is achieved by adding the two single components, that is:

$$\widehat{Y}_{t+ph} = \widehat{L}_{t+ph} + \widehat{NL}_{t+ph} \tag{7}$$

SOM-SARIMA-SVR-2 Model (Hybrid Approach 2)
The time series is considered a nonlinear function of the original data and the residuals and the predicted values from the second step:

$$\begin{aligned}\widehat{Y}_{t+ph} = f(&y_t, y_{t-1\cdot ph}, y_{t-2\cdot ph}, \ldots, y_{t-ny\cdot ph}, e_{t+ph}, e_t, e_{t-1\cdot ph}, \\ &\ldots, e_{t-ny\cdot ph}, \hat{P}_t, \hat{P}_{t-1\cdot ph}, \ldots, \hat{P}_{t-np\cdot ph}) + \varepsilon_{t+ph}\end{aligned} \tag{8}$$

where $\hat{p}_t$ is the predicted value from the SARIMA model and ne, ny and np represent autoregressive window sizes for e, y and $\hat{p}$ variables, respectively.

The proposed SOM-SARIMA-SVR procedure is graphically shown in Fig. 2:

3.2 Performance Indexes

Performance Criteria of Stage I (Clustering). Two clustering quality indices have been used, CQI_1 and CQI_2 (Eqs. (9–10)):

$$QI_1 = \left(\tilde{S}_i\right) \tag{9}$$

$$QI_2 = \sum S_i \tag{10}$$

where S_i is the silhouette function and its value for each pattern is between -1 to $+1$. This parameter is defined as $S_i = D_i - d_i/max(d_i, D_i)$, where D_i is the minimum average distance from one pattern of a cluster to another pattern in another cluster and d_i is the average distance in the own cluster from one pattern to the rest of patterns.

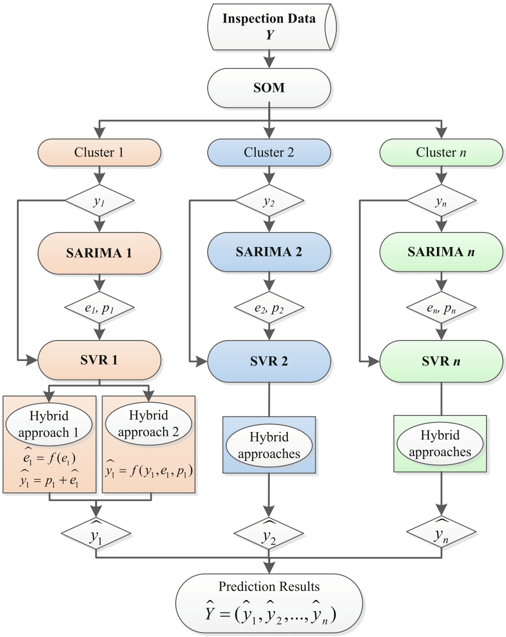

Fig. 2. The overall process scheme of the SOM-SARIMA-SVR approach.

Performance Indexes of Stages II and III (Prediction). The mean square error (MSE), the mean absolute percentage error (MAPE) and the mean absolute error (MAE) are the performance indices that have been considered to calculate the estimation of the generalization error in the prediction steps (I and III). Equations (11–13) shows these performance criteria and their calculation, where m is the sample size, y_t is the real value of the observation and y_t is the corresponding predicted value.

$$MSE = \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{m} \tag{11}$$

$$MAE = \frac{\sum_{i=1}^{m}|\hat{y}_i - y_i|}{m} \tag{12}$$

$$MAPE = \frac{\sum_{i=1}^{m}|y_i - \hat{y}_i)/y_i|}{m} \tag{13}$$

4 Experimental Results and Discussion

A comparison among the single SVR, the combined SOM-SVR, the hybrid SARIMA-SVR and the proposed SOM-SARIMA-SVR models was performed.

First, a SOM model was employed as a clustering technique. Testing different configurations of the SOM network, the most appropriate SOM size for the data was found to be the map size of 8×8 neurons in the output layer with a hexagonal grid topology and a three-dimensional input space. Results leads to consider that the SOM network has clustered the data into two groups. These results can be contrasted analytically and are collected in Table 1 which shows the best results obtained per cluster and their input vector configuration. Based on the two clustering performance indexes (CQI_1, and CQI_2), the two-classes clustering was the best choice for the time series, reaching the highest values of CQI_1 and CQI_2 (0.659 and 717.548, respectively). This result confirms the obtained previously with the SOM algorithm. Consequently, the database was also divided into two groups, hereinafter called Cluster 1 and Cluster 2. Best results were achieved using a three-element input vector ($nc = 3$) with a temporal leap of 7-day in the past.

Table 1. Clustering results of the SOM step, where c is the number of clusters tested and nc is the size of the input vector. The temporal leap in the past is 1 or 7 days.

Best configurations			Performance indices	
Clusters (c)	Temporal leap	nc	CQ_1	CQ_2
2	7	3	0.650	717.548
3	7	3	0.627	682.549
4	1	3	0.588	643.792
5	1	3	0.601	658.248

In the second step, a SARIMA model was independently applied to each cluster. Using an iterative trial-and-error procedure, the best-fitted models were ARIMA$(2,0,3)$ for Cluster 1 (without seasonal part) and SARIMA$(2,1,2)(2,1,3)_5$, with a seasonality of 5 days for Cluster 2. The requirements of a white noise process were satisfied to the residuals of the model.

Finally, in the third step, different SVR models were applied to each cluster considering the two proposed hybrid approaches which are formed depending on the input variables used. Focused on an individual hybrid configuration, a best SVR model was achieved in each cluster to fit the data. The parameter configuration of these SVR models is (generally) different in each cluster. The final prediction results of this hybrid approach were obtained by integrating the prediction values achieved in the two clusters as a single predicted time series. That is, $\hat{y}_{inspections} = \{\hat{y}_{cluster1}\} \bigcup \{\hat{y}_{cluster2}\}$.

The most accurate models for each hybrid approach are collected in Table 2. These prediction results were obtained considering the junction of the predictions of the two clusters. Table 2 is divided according to the prediction horizon used ($ph = 1$ or $ph = 7$ days). Furthermore, for each prediction horizon, results are collected depending on the hybrid configuration applied. For the hybrid approach 2 (SOM-SARIMA-SVR-2), results are presented considering the set of inputs used (y, e or/and p) in order to clearly show the most relevant inputs. The SOM-SARIMA-SVR-2 (without p as input) provides the best prediction results for one-day ahead prediction, followed by the rest of possible models presented in the hybrid approach 2 (considering different SVR inputs) and finally the hybrid approach 1, in that order. The SOM-SARIMA-SVR-2 achieved the best value in at least two performance indexes. In this case, more sophisticated models obtained no better results. Nevertheless, the classical approach considers an additive relationship between the linear and nonlinear component of the time series. Consequently, it can be concluded that this approach is less powerful than the other approach. For one-day ahead predictions, two different input variables (y and e) are proved to be sufficient to predict the time series accurately. However, there are not great differences among the prediction performances with the rest models.

Table 2. Mean prediction performance results of the proposed SOM-SARIMA-SVR models for one-day and seven-day ahead prediction horizons. The final number of the model indicates the hybrid approach (1 or 2). The column inputs indicates the inputs used in the SVR models, where y is the original data and e and p depict the residuals and predicted values from the second step, respectively. Best values in bold.

ph	Hybrid approach	Model	Inputs	Performance indices		
				MSE	MAE	MAPE
1	1	SOM-SARIMA-SVR 1	e	296.3551	11.8413	18.6794
	2	**SOM-SARIMA-SVR 2**	y, e	**287.0105**	**11.6790**	17.9391
		SOM-SARIMA-SVR 2	y, p	288.7795	11.7174	**17.7721**
		SOM-SARIMA-SVR 2	y, e, p	287.8423	11.8362	17.8603
7	1	SOM-SARIMA-SVR 1	e	299.6534	11.9761	18.4053
	2	SOM-SARIMA-SVR 2	y, e	290.8783	**11.8295**	18.1178
		SOM-SARIMA-SVR 2	y, p	306.1622	12.4121	18.3690
		SOM-SARIMA-SVR 2	y, e, p	**289.7224**	11.8394	**17.9783**

Similar results were obtained considering the behaviour of the models for 7-day ahead prediction, where better values of performance indexes were reached with the hybrid approach 2. The most complex approach (SOM-SARIMA-SVR 2 with all variables as inputs) obtained the best results, reaching four of the five best performance indexes. Better results were yielded using the more sophisticated models (hybrid approach 2) instead of the classical approach (hybrid approach 1). Particularly, SARIMA-SOM-SVR-2 with variables e and y as inputs of the SVR achieved the best results. The best-fitted network of Cluster 1 for this hybrid configuration 2 in the third step is composed by autoregressive window sizes of twelve for the y input variable ($ny = 12$) and two for the e input from SARIMA step ($ne = 2$), being the optimal SVR parameters $C = 200$, $\gamma = 2^{-4}$ and $\epsilon = 2^{-8}$. To model Cluster 2, the best parameter configuration was $ny = 12$,

$ne = 2$, $C = 50$, $\gamma = 2^{-2}$ and $\epsilon = 2^{-2}$. For this network architecture, the number and size of SVR inputs coincide in both clusters. The final prediction is reached by junction the predicted values of the two clusters.

To conclude, the most accurate single SVR model, the most accurate combined SOM-SVR model and the most accurate hybrid SARIMA-SVR model were also compared against the proposed model. These comparisons are summarized in Table 3. As this table shows, the proposed SOM-SARIMA-SVR model outperforms the rest of the models in both prediction horizons. This suggest that the "*divide-and-conquer*" principle, introduced with the usage of the clustering stage, can improve the performance of the hybrid models that consider the hybridization of linear and nonlinear forecasting techniques. Figure 3 represents a comparison point-to-point between the observed and predicted values for the best-fitted models concerning the $ph = 1$ case.

Table 3. Comparison of the best mean prediction performance results of the single models (SVR), the combined models (SOM-SVR), the hybrid model (SARIMA-SVR) and the proposed model (SOM-SARIMA-SVR) for one-day and seven-day prediction horizon. Best values in bold.

ph	Model	Performance indices		
		MSE	MAE	MAPE
1	SVR	389.1624	14.3328	23.0695
	SOM-SVR	381.6965	14.1565	21.8681
	SARIMA-SVR	302.0054	12.0024	19.4246
	SOM-SARIMA-SVR	**287.0105**	**11.6790**	**17.7721**
7	SVR	299.6534	11.9761	18.4053
	SOM-SVR	290.8783	**11.8295**	18.1178
	SARIMA-SVR	306.1622	12.4121	1 8.3690
	SOM-SARIMA-SVR	**289.7224**	11.8394	**17.9783**

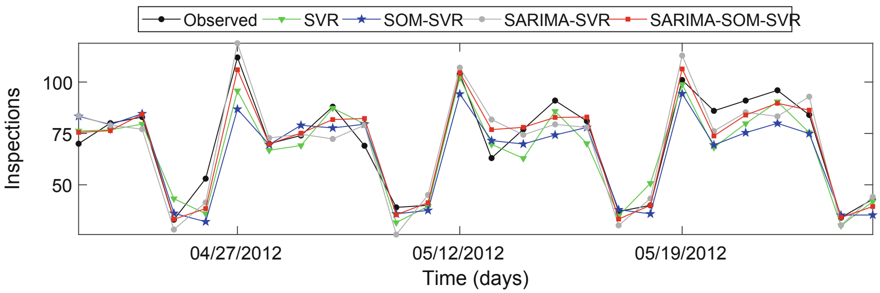

Fig. 3. Comparison of the observed and predicted value on number of containers checked with the most accurate models. $ph = 1$ case.

5 Conclusions

In this study, a combined-hybrid SOM-SARIMA-SVR forecasting model has been proposed based on a three-step procedure to predict the number of containers passing through a Border Inspection Post. A clustering SOM is first applied to obtain smaller regions with similar statistical features which may be easier to predict. A SARIMA model is then fitted within each cluster to obtain predicted values and residuals of the clustered database. Finally, a SVR model is used to forecast each cluster independently using the variables obtained from the second step together with the original data as inputs. The SOM-SARIMA-SVR model proposed has been developed and compared to other possible methodologies implied in the process (SVR, SOM-SVR and SARIMA-SVR). The results obtained indicate that the SOM-SARIMA-SVR model is the most competitive model, improving the forecasting performance of the rest of the models concerning the prediction of the container demand and outperforms these methodologies. This methodology can provide an automatic tool to predict workloads in inspection facilities avoiding congestion and delays. Therefore, it can be used as a decision-making tool by port managers due to its capacity to plan resources in advance.

Acknowlegements. Authors acknowledge support through grant RTI2018-098160-B-I00 from MINECO-SPAIN. The database has been kindly provided by the Port Authority of Algeciras Bay.

References

1. Babcock, M.W., Lu, X.: Forecasting inland waterway grain traffic. Transp. Res. Part E Logist. Transp. Rev. **38**(1), 65–74 (2002). https://doi.org/10.1016/S1366-5545(01)00017-5. http://www.sciencedirect.com/science/article/pii/S1366554501000175
2. Box, G.E.P., Jenkins, G.M.: Time Series Analysis: Forecasting and Control. Holden-Day, Oakland CA (1976). Revised edn
3. Chen, H., Grant-Muller, S., Mussone, L., Montgomery, F.: A study of hybrid neural network approaches and the effects of missing data on traffic forecasting. Neural Comput. Appl. **10**(3), 277–286 (2001)
4. Ismail, S., Shabri, A., Samsudin, R.: A hybrid model of self-organizing maps (SOM) and least square support vector machine (LSSVM) for time-series forecasting. Expert Syst. Appl. **38**(8), 10574–10578 (2011). https://doi.org/10.1016/j.eswa.2011.02.107. http://www.sciencedirect.com/science/article/pii/S0957417411003137
5. Klein, A.: Forecasting the Antwerp maritime traffic flows using transformations and intervention models. J. Forecast. **15**(5), 395–412 (1998)
6. Kohonen, T.: Self-Organising Maps. Springer, Berlin (1995). https://doi.org/10.1007/978-3-642-97610-0
7. Mak, K.L., Yang, D.H.: Forecasting Hong Kong's container throughput with approximate least squares support vector machines. In: Proceedings of the World Congress on Engineering, vol. 1, pp. 7–12. Citeseer (2007)

8. Moscoso-López, J.A., Turias, I., Jiménez-Come, M.J., Ruiz-Aguilar, J.J., Cerbán, M.D.M.: A two-stage forecasting approach for short-term intermodal freight prediction. Int. Trans. Oper. Res. (2016). https://doi.org/10.1111/itor.12337
9. Ruiz-Aguilar, J.J., Turias, I., Moscoso-López, J.A., Jiménez-Come, M.J., Cerbán, M.: Forecasting of short-term flow freight congestion: a study case of Algeciras Bay Port (Spain). Dyna **83**(195), 163–172 (2016). https://doi.org/10.15446/dyna.v83n195.47027. http://www.revistas.unal.edu.co/index.php/dyna/article/view/47027
10. Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)
11. Xie, G., Wang, S., Zhao, Y., Lai, K.K.: Hybrid approaches based on LSSVR model for container throughput forecasting: a comparative study. Appl. Soft Comput. https://doi.org/10.1016/j.asoc.2013.02.002, http://www.sciencedirect.com/science/article/pii/S156849461300046X