A Causality-Based Feature Selection Approach For Multivariate Time Series Forecasting

Youssef Hmamouche*, Alain Casali* and Lotfi Lakhal*
*LIF - CNRS UMR 6166, Aix Marseille Université, Marseille, France
Email: firstname.lastname@lif.univ-mrs.fr

***Abstract*—The study of time series forecasting has progressed significantly in recent decades. The progress is partially driven by growing demand from different industry branches. Despite recent advancements, there still exist several issues that need to be addressed in order to improve the accuracy of the forecasts. One of them is how to improve forecasts by utilizing potentially extra information carried by other observed time series. This is a known problem, where we have to deal with high dimensional data and we do not necessarily know the relationship between variables. To deal with this situation, the challenge is to extract the most relevant predictors that will contribute to forecast each target time series. In this paper, we propose a feature selection algorithm specific to forecasting multivariate time series, based on (i) the notion of the Granger causality, and on (ii) a clustering strategy. Lastly, we carry out experiments on several real data sets and compare our proposed method to some of the most widely used dimension reduction and feature selection methods. Experiments illustrate that our method results in improved accuracy of forecasts compared to the evaluated methods.**

Keywords–Multivariate Time Series Forecasting; Granger Causality; Feature selection.

I. INTRODUCTION

Time series analysis incorporates a set of tools, methods, and models in order to describe the evolution of data over time. It has been developed primarily for the purposes of forecasting and business analysis. Time series analysis is an important component of any business intelligence system insofar as it generates new, valuable data by combining trends, forecasts, correlations, causalities *etc.* in intelligent ways. Consequently, time series analysis produces original, exploitable information that can then be used as a critical input to the decision-making process and, ergo, can contribute to more intelligent and effective decisions.

The first time series forecast models were introduced in the 1920s. These were followed shortly by the first application of the univariate Auto-Regressive model [1]. Advanced versions of these models are still in use today. Based on the Auto-Regressive principle, those models take into account data history in order to make forecasts. Nevertheless, despite their innovativeness, these first models only consider a single time series in their predictions and, thus, fail to utilize a significant amount of potentially-exploitable data. With this in mind, in the latter half of the last century, researchers began to lend greater attention to refining forecast models that exploit multiple time series [2]. Most of the algorithms used today for multivariate time series forecasting, which includes the algorithms most commonly used for economic forecasting, are based on concepts developed during this period.

Multivariate analysis is increasingly preferred by data scientists over univariate analysis. The latter is simpler than the former as it only takes into account the previous values of a respective time series. Multivariate models, on the other hand, seek to understand the behavior and characteristics of the time series in question by explaining each series based (i) on its previously observed values, Iin addition to (ii) the previously observed values of other series in the data set. This approach is particularly important when handling financial data because, indeed, the value of one variable often does not only depend on its previous values, but also on the past values of other variables in the same dataset. As such, in order to obtain the most accurate outputs, it is necessary to factor in as inputs all the relevant information from other variables when making forecasts [3]. Unfortunately, utilizing all the existing variables in a multivariate model in a way that achieves optimal results has yet to be perfected: (i) in some cases, existing models are simply not able to incorporate all variables; (ii) in other cases, models may not, for reasons that we will discuss, produce more accurate forecasts. For instance, the authors of [4], working with real data from Australia and the United States, were not able to improve accuracy of their forecasts when using more than $30\% - 60\%$ of the existing predictors.

In this paper, we propose a feature selection method specific to time series forecasting. We argue that our approach handle relatively the problem of dependencies between variables, which is a major drawback of many exisiting methods. Specifically, we are able to do so by explaining causalities between variables (i) using the Granger causality graph [5], and (ii) then clustering them. The proposed approach is currently being used in two industrial prototypes, which are to be used for different purposes: (i) the first one is designed to provide a tool for buyers, informing them when to purchase a product for their company; (ii) and the second prototype is used for detecting fraud in public markets. The objective of our work on both prototypes is the same: to forecast prices based on raw materials and/or finished products.

This paper is organized as follows: the first three sections are dedicated to related work: Section II is dedicated to prediction models, Section III is related to feature selection and dimension reduction methods and Section IV is devoted to the Granger causality. In Section V, we detail our approach. In Sections VI and VII, we perform experiments and comparison study on real data sets. And in Section VIII, we summarize our contributions and put forth possible future research.

II. PREDICTION MODELS

Many prediction models which are currently being developed are based on the idea of Auto-Regressive model $\mathrm{AR}(p)$ [6]. This model expresses a univariate time series as a linear function of its p precedent values:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \epsilon_t$$

 ISBN: 978-1-61208-558-6

Where p is the order of the model, $\alpha_0 \ldots \alpha_p$ are the parameters of the model, and ϵ_t is a white noise error term. The Moving Average model (MA) has the same expression, but for the error terms. The ARMA(p, q) model [6] combines these two models by considering both past error terms and values. For non-stationary time series, the ARIMA(p, d, q) model [6] is more preferable, it applies the ARMA(p, q) model after a differencing step, in order to obtain stationary time series, where d is the order of differenting (computing d times the differences between consecutive observations). In [7], the Vector Auto-Regressive VAR model is introduced as an extension of the AR model. Consider a k-dimensional time series Y_t, the VAR(p) system expresses each univariate variable of the multivariate time series Y_t as a linear function of the p previous values of itself and the p previous values of the other variables:

$$Y_t \quad = \alpha_0 + \sum_{i=1}^{p} A_i Y_{t-i} + \epsilon_t,$$

where ϵ_t is a white noise with a mean of zero, and $A_1, \ldots, A_p$ are $(k \times k)$ matrix parameters of the model. In [8], the Vector Error Correction (VECM) is introduced. This model transforms the VAR model by taking into account non-stationarity of the time series and by including cointegration equations. To simplify matters, let us consider two univariate time series (x_t, y_t) integrated of orger one, which means non stationary, but the first difference ($\Delta x_t = x_t - x_{t-1}$) is stationary. The VECM Model can be written as follows:

$$\Delta y_t = \alpha_{0y} - \gamma_y(\beta_0 y_{t-1} - \beta_1 x_{t-1}) + \sum_{i=1}^{p} v_{iy} \Delta y_{t-i} + \sum_{i=1}^{p} w_{iy} \Delta x_{t-i} + \epsilon_t$$

$$\Delta x_t = \alpha_{0x} - \gamma_x(\beta_0 y_{t-1} - \beta_1 x_{t-1}) + \sum_{i=1}^{p} v_{ix} \Delta y_{t-i} + \sum_{i=1}^{p} w_{ix} \Delta x_{t-i} + \epsilon_t$$

Where $\beta_0 y_{t-1} - \beta_1 x_{t-1}$ is stationary, the coefficients (β_0, β_1) are the cointegrating parameters, and (γ_y, γ_x) are the error correction parameters. If there exist no coefficients (β_0,β_1) such that $\beta_0 y_{t-1} - \beta_1 x_{t-1}$ is stationary, then x_t and y_t are not cointegrated and the VECM model is reduced to the VAR form.

III. FEATURE SELECTION AND DIMENSION REDUCTION METHODS

Feature selection refers to the act of extracting subset of the most relevant variables (features) of size k from a set of variables of size $n >> k$. While, dimension reduction methods consist in generating an artificial features with smallest dimension from the originals by combining them. Therefore, from a descriptive analysis point of view, feature selection is more interesting. However, both of them can be used to optimise the inputs of prediction models. Using all the existing variables in a multivariate model has two principal drawbacks. First, it can affect the rightness of the predictions computations. For example, in Auto-Regressive based models, if the number of regressors is proportional to the sample size, the ordinary least squares (OLS) forecasts are not efficient, and the challenge with these situtations is to reduce dimensionality of predictors [9]. Second, it prevents from detecting the most relevant variables, which can degrade forecasts accuracy [4].

The Principal Component Analysis (PCA) is one of the most common dimension reduction methods used [10]. Based on a set of variables, this method takes advantage of the inter-correlation between them. The idea is to generate the principal variables that describe as much as possible the original variables using a linear transformation. The Kernel PCA method is a non-linear principal component analysis proposed as an extension of PCA, by considering non-linear correlation between variables [11]. The Recursive Feature Elimination (RFE) technique works by recursively removing variables and building a model on those variables that remain [12]. These methods are widely used in forecasting time series, for example PCA and Kernel PCA have been adopted in two-step approach which reduces first the number of predictors, and then a applies a forecasting model [13]–[15]. Univariate appoaches are based on the principle of selecting variables by ranking them according to a statistical test or a similarity measure. For instance, in [16], a method based on causality is proposed. The algorithm selects variables that cause the target, and it shows good results compared with some dimension reduction methods.

IV. GRANGER CAUSALITY

The purpose of this section is to redefine the Granger causality [5], and to detail the statistical test used to estimate the bivariate causality between two time series. Let us consider two univariate time series x_t, y_t. The Granger definition of causality acknowledges the fact that x_t causes y_t if it contains information helpful to predict y_t. In other words, if by removing x_t from the available information used to predict y_t at the current time, the prediction results for y will be affected.

We detail here the standard Granger causality test [17], which uses the VAR model with a trend term. The test compares two models, (i) the first one only takes into account the precedents values of y_t and (ii) the second uses both x_t and y_t in order to predict y_t. If there is a significant difference between the two models, then it can be ascertained that the added variable, *i.e.*, x_t causes y_t:

$$\text{Model}_1 : \quad y_t = \alpha_0 + \alpha t + \sum_{i=1}^{p} \alpha_i y_{t-i} + \epsilon_t$$

$$\text{Model}_2 : \quad y_t = \alpha_0 + \alpha t + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=1}^{p} \beta_i x_{t-i} + \epsilon_t$$

The next step of the test is to compare the residual sum of squares (RSS) of these models using the Fisher test. The statistic of the test is expressed as follow:

$$F = \frac{(\text{RSS}_1 - \text{RSS}_2)/p)}{(\text{RSS}_2/n - 2p - 1)}$$

Where RSS_1 and RSS_2 are the residual sum of squares related to Model_1 and Model_2 respectively, n is the size of the predicted vector. Two hypotheses are tested, the null hypothesis H_0: $\forall i \in \{1, \ldots, p\}, \beta_i = 0$ (which means x does not cause y) and H_1: $\exists i \in \{1, \ldots, p\}$, $\beta_i \neq 0$. Under the null hypothesis H_0, F follows the Fisher distribution with $(p, n - 2p - 1)$ degrees of freedom, then the test is carried out at a level α in order to examine the null hypothesis of non causality.

 ISBN: 978-1-61208-558-6

V. OUR PROPOSAL

We focus here on the selection of the top predictor variables based on the Granger causality as a relationship between variables. Let us consider $Y = \{y_1, y_2, \ldots, y_n\}$ a multivariate time series and a target variable y. The idea is to choose a subset of Y, for which we have the more accurate forecasts. Let us underline that from a theoretical point of view, there are $\sum_{i=1}^{k} \binom{n}{i}$ possible partitions of size less than or equal to k. And in general, there are 2^n possible partitions, which means 2^n possible models [9]. In addition, the Granger causality is not a monotone function, as a consequence, finding the best subset of variables that maximizes the causality is a NP-hard problem. One solution is to choose a set of variables having strong causality regarding to the target y as investigated in [16]. However, this approach does not take into account hidden relationship between variables, which means that we could use the same information even when using many variables.

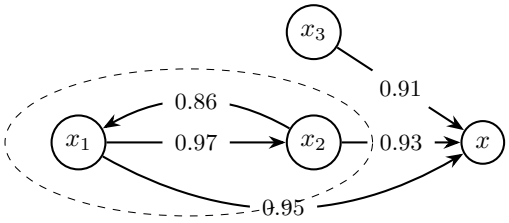

Figure 1. Illustration of dependencies between time series using Granger causality graph

In Figure 1, we show a small Granger causality graph describing dependencies between 4 variables. Let us try to select two variables as predictors for the target variable, *i.e.,* x. Selecting variables by ranking them according to causality leads to getting x_1 and x_2. However, x_1 and x_2 might provide the same information because x_1 causes x_2.

We propose a new method to deal with this problem based on clustering the Granger causality graph or the adjacency matrix using Partitioning Around Medoids (PAM) algorithm [18]. The p-value of the test is the probability to observe the given result under the assumption that H_0 is true, which means the probability of non causality. We consider so the causality as one minus p-value in order to express values of causalities in the range $[0, 1]$.

A. Algorithm of the proposed method

The algorithm of the proposed method can be divided into three steps:

- Building the adjacency matrix of causalities between variables.
- Clustering the set of all the possible predictors variables, by minimizing the causalities between clusters, and maximizing the causality within clusters, using the PAM method.
- Choosing one element from each cluster, the one that maximizes the causality on the target variable.

In Figure 2, the GSM (Granger Selection method) algorithm summarizes our approach. It generates for each target variable y, k variables that contributes to the prediction of y.

Input: Set of predictors time series $Y = \{y_1, y_2, \ldots, y_n\}$, y the target variable, MINCAUS Min-Causality threshold, k the selection size
Output: GSM-CL the selected variables associated to y

1: **for** $i = 1$ to n **do**
2: **if** $Y.size() \leqslant k$ **then**
3: $Y = Y \setminus \{y_i\}$
4: **end if**
5: **if** $Y.size() \leqslant k$ **then**
6: **return** GSM-CL $= Y$
7: **end if**
8: **end for**
/* The clustering step. */
9: Let Mc be the dissimilarity matrix of predictors
10: **for each** x_i, x_j in Y such that $i \neq j$ **do**
11: $Mc[i,j] = Mc[j,i] = 1 - max(causality(x_i \rightarrow x_j), causality(x_j \rightarrow x_i))$
12: **end for**
13: $Clusters = pam(Mc, k)$
/* The selection step. */
14: **for each** Cluster cl in $Clusters$ **do**
15: GSM-CL = GSM-CL $\cup$ $\arg\max_{cl_j \in cl} (causality(cl_j \rightarrow y)$
16: **end for**
17: **return** GSM-CL

Figure 2. The GSM Algorithm.

B. Example

Consider $Y = \{y_1, \ldots, y_8\}$ a set of predictors and a target variable y_9. Let's apply the GSM algorithm in order to select 4 predictor variables from Y that will contribute to forecast y.

1) The matrix of causalities: First, the algorithm computes the Granger causalities between variables in pairs. In this example, we take the matrix of causalities of our data sets corresponding to the dataset ts_1 described in Section VI-A:

$$MC = \begin{bmatrix} 1.00 & 0.935 & 0.999 & 0.999 & 0.832 & 0.998 & 0.998 & 0.933 & 0.998 \\ 0.28 & 1.00 & 0.877 & 0.87 & 0.224 & 0.785 & 0.801 & 0.999 & 0.868 \\ 0.033 & 0.656 & 1.00 & 0.106 & 0.479 & 0.944 & 0.775 & 0.082 & 0.905 \\ 0.028 & 0.647 & 0.239 & 1.00 & 0.483 & 0.944 & 0.776 & 0.096 & 0.905 \\ 0.7 & 0.457 & 0.977 & 0.978 & 1.00 & 0.343 & 0.031 & 0.398 & 0.901 \\ 0.808 & 0.417 & 0.818 & 0.817 & 0.906 & 1.00 & 0.997 & 0.431 & 0.722 \\ 0.274 & 0.742 & 0.992 & 0.992 & 0.942 & 0.959 & 1.00 & 0.906 & 0.788 \\ 0.327 & 0.999 & 0.998 & 0.998 & 0.427 & 0.895 & 0.996 & 1.00 & 0.900 \\ 0.304 & 0.071 & 0.581 & 0.584 & 0.205 & 0.448 & 0.999 & 0.754 & 1.00 \end{bmatrix}$$

2) Clustering and selecting the final variables: The algorithm partitions the variables based on the symmetrical matrix (as mentioned in the algorithm 2) using the PAM method. The idea behind symmetrizing the matrix of causalities is to build clusters where there is at least one causality between each pairs of variables, so it is logical to use the maximum. Let us underline also that the classical PAM algorithm partitions elements from a symmetric dissimilarity matrix, by minimizing dissimilarities within clusters. In our case, the algorithm maximizes causalities within clusters. That is why we use 1 minus the causality matrix as an input of the PAM method. Then, from each cluster, the algorithm chooses the element that has maximal causality on the target. The clustering vector associated to $\{y_1, \ldots, y_8\}$ obtained is $(1, 2, 1, 1, 3, 1, 4, 2)$. And based on the causalities to the target (last column of the adjacency matrix), the selected variables are $\{y_1, y_5, y_7, y_8\}$.

 ISBN: 978-1-61208-558-6

3) Evaluation of the clusters: The quality of the causalities founded depends on, first the type of the data. And second, on the evaluation of the clustering task. In our case, we evaluate the quality of the clusters using the following objective function:

$$\text{minimize } G(x) \quad = \sum_i^n \sum_j^n (1 - max(c_{ij}, c_{ji})) \times z_{ij},$$

where,

1) $z_{ij} = \begin{cases} 1 & \text{if } y_i, y_j \text{ belong to the same cluster} \\ 0 & \text{otherwise.} \end{cases}$
2) c_{ij} = causality($y_i \rightarrow y_j$).

This evaluation can be used in general as measure of causal relationships in multivariate time series. In the example, the value of G is 0.000168.

VI. EXPERIMENTS

We present in this part the methodologies adopted to carry out the experiments. We compare our method with four existing methods, selectKf: univariate feature selection method using the F-test statistical test, selectKc: univariate feature selection method using the Granger causality test [16], PCA: Principal Component Analysis [19], KERNEL PCA: Kernel Principal Component Analysis [11], and our proposal.

Vector Error Correction (VECM) [8] model is adopted to forecast the multivariate time series generated by the feature selection and dimension reduction methods. The univariate ARIMA model (see II) is also evaluated to show the forecasting results with no predictors variables. For our proposal, we use a p_value threeshold of the Granger causality test at 10%, and the lag parameters of the VECM and ARIMA models are determined according to the Akaike's Information Criterion (AIC) [20]. Experiments are made on an single computer with processor 2,2 GHz Intel Core i7 and 16Gb of RAM.

A. The used Data Sets

The first data set used comes from our current project. The second data set are taken from the Machine Learning Repository website [21], and the third one represents macroeconomic time series of United Sates [22]. A brief description of these data sets including the number of variables and observations and the target variables is presented in Table I.

B. Measuring forecast accuracy

The training step is carried out on the first 90% of the input series, and an evaluation on the last 10% real values is performed by one step ahead forecasts using rolling window VECM and ARIMA models. The measure of prediction accuracy used is the normalized root mean square error (NRMSE):

$$\text{NRMSE} = \frac{1}{\bar{y}} \sqrt{\frac{\sum_{i=1}^{h} (y_i - \hat{y_i})^2}{h}} \tag{1}$$

Where $(\hat{y_1}, \ldots, \hat{y_h})$ are the forecasts, $(y_1, \ldots, y_h)$ are the real values and $\bar{y}$ is the average value of y_t.

The comparison between methods will be the same if we use the mean squared error MSE or the root-mean-square RMSE. We use the NRMSE in order to have normalized and relative evaluations.

TABLE I. DESCRIPTION OF THE USED DATA SETS.

Data sets	Number of series	Number of observations	Description
$\mathbf{ts_1}$	9	1090	Our Dataset, expressing the prices of International Index containing Oil, Propane, Gold, euros/dollars, Butane, Cac40, and others, between $2013/03/12$ and $2016/03/01$, aiming to forecast the Cac40.
$\mathbf{ts_2}$	8	563	Data sets includes returns of Istanbul Stock Exchange (ISE) with seven other international index; SP, DAX, FTSE, NIKKEI, BOVESPA, MSCE_EU, MSCI_EM from Jun 5, 2009 to Feb 22, 2011.
$\mathbf{ts_3}$	36	360	Monthly coincident and leading economic indexes of economic activity in the United States, for forecasting four series: industrial production IP, real personal income less transfer payments GMYXP8, real manufacturing and trade sales MT82, and employee-hours in nonagricultural establishments LPMHU.

VII. COMPARATIVE STUDY

In a first time, we measure forecast accuracy using the univariate ARIMA model. The results obtained are shown in Table II. This will allow us to compare how much the reduced model; VECM with dimension reduction of the predictors will perform compared with the univariate model.

TABLE II. EVALUATING FORECAST ACCURACY OF THE ARIMA MODEL.

Data sets	Target series	NRMSE
ts1	**CAC40**	0.0136
ts2	**ISE**	0.1004
ts3	**IP**	0.0094
	GMXY	0.0094
	LPMHU	0.0103
	MT82	0.0175

We show in Table III the forecast evaluations of each data set, by considering different numbers of predictors variables for each experiment. Let us underline that for the dataset ts_3, which contains 36 variables, the performance accuracy decreased when we use more than 11 predictors. This is why the results in the Table III are shown for a number of predictors, *i.e.*, k, less or equal than 11.

For dimension reduction methods PCA and Kernel PCA, it is possible to have both automatic number of features k or a specific number given in the input, which is not the case for the univariate selection method using the Granger causality test [16] which selects features naturally. The number of features generated using this method can be seen in Figure 3c.

Our proposal can be extended to provide an automatic number of features by using some methods of selection for the optimal number of clusters of the PAM method. However, the number of variables computed in advance is generally not optimal in term of forecasting, since the optimal value must be determined according to the forecast accuracy. For this reason we evaluate different values of the number of predictors, *i.e.*, k.

 ISBN: 978-1-61208-558-6

TABLE III. EVALUATING FORECASTS ACCURACY WITH DIFFERENT REDUCTION SIZES k.
* INDICATES THAT THE REDUCTION SIZE K IS GREATER THAN THE NUMBER OF VARIABLES,
- IF AN ERROR OF RESOLUTION OCCURS.

Data sets	Targets series	Methods	Normalized root mean squared error (NRMSE)										
			Number of features k										
			1	2	3	4	5	6	7	8	9	10	11
ts1	CAC40	kpca	0.0136	0.0136	0.0137	0.0136	0.0136	0.0136	0.0136	0.0136	*	*	*
		pca	0.0137	0.0137	0.0135	0.0135	0.0135	0.0135	0.0135	0.0135	*	*	*
		selectKc	0.0134	0.0134	0.0134	0.0134	0.0134	0.0134	0.0134	0.0134	*	*	*
		gsm	0.0134	0.0134	0.0134	0.0135	0.0134	0.0134	**0.0133**	0.0134	*	*	*
		selectKf	0.0136	0.0136	0.0136	0.0136	0.0135	0.0134	0.0134	0.0134	*	*	*
ts2	ISE	kpca	0.0991	0.1093	0.1100	0.1145	0.1141	0.1109	0.1210	*	*	*	*
		pca	0.0991	0.1093	0.1101	0.1145	0.1141	0.1109	0.1210	*	*	*	*
		selectKc	0.1239	0.1239	0.1239	0.1239	0.1239	0.1239	0.1239	*	*	*	*
		gsm	**0.0983**	0.1128	0.1174	0.1127	0.1129	0.1208	0.1210	*	*	*	*
		selectKf	0.1020	0.1206	0.1247	0.1203	0.1298	0.1208	0.1210	*	*	*	*
ts3	IP	kpca	0.0090	0.0092	-	-	-	-	-	-	-	-	-
		pca	0.0102	0.0095	0.0111	0.0111	0.0102	0.0108	0.0104	0.0105	0.0178	0.0195	0.0215
		selectKc	0.0161	0.0161	0.0161	0.0161	0.0161	0.0161	0.0161	0.0161	0.0161	0.0161	0.0161
		gsm	0.0091	**0.0084**	0.0086	0.0090	0.0094	0.0095	0.0115	0.0169	0.0186	0.0223	0.0232
		selectKf	0.0093	0.0092	0.0093	0.0095	0.0123	0.0122	0.0103	0.0132	0.0141	0.0142	0.0164
	GMXY8	kpca	0.0189	0.0202	0.0228	-	-	-	-	-	-	-	-
		pca	0.0084	**0.0082**	0.0091	0.0100	0.0101	0.0102	0.0103	0.0109	0.0148	0.0162	0.0177
		selectKc	0.0082	0.0082	0.0082	0.0082	0.0082	0.0082	0.0082	0.0082	0.0082	0.0082	0.0082
		gsm	0.0089	0.0093	0.0098	0.0096	0.0094	0.0098	0.0122	0.0121	0.0135	0.0139	0.0146
		selectKf	0.0087	0.0099	0.0099	0.0100	0.0099	0.0098	0.0103	0.0107	0.0108	0.0162	0.0177
	LPMHU	kpca	0.0079	0.0127	-	-	-	-	-	-	-	-	-
		pca	0.0075	0.0073	0.0074	0.0080	0.0080	0.0082	0.0080	0.0082	0.0084	0.0100	0.0111
		selectKc	0.0104	0.0104	0.0104	0.0104	0.0104	0.0104	0.0104	0.0104	0.0104	0.0104	0.0104
		gsm	0.0076	0.0074	0.0073	**0.0071**	0.0078	0.0074	0.0078	0.0097	0.0096	0.0090	0.0085
		selectKf	0.0077	0.0078	0.0080	0.0082	0.0080	0.0087	0.0080	0.0094	0.0096	0.0103	0.0113
	MT82	kpca	0.0171	0.0171	-	-	-	-	-	-	-	-	-
		pca	0.0168	0.0178	0.0179	0.0169	0.0170	0.0166	0.0168	0.0173	0.0275	0.0277	0.0294
		selectKc	0.0206	0.0206	0.0206	0.0206	0.0206	0.0206	0.0206	0.0206	0.0206	0.0206	0.0206
		gsm	0.0162	0.0165	0.0154	0.0158	0.0154	**0.0154**	0.0161	0.0248	0.0234	0.0263	0.0240
		selectKf	0.0170	0.0170	0.0169	0.0171	0.0183	0.0188	0.0197	0.0203	0.0208	0.0205	0.0231

Evaluations in Table III show that overall, the GSM currently outperforms most of the target time series compared with the ARIMA model and the methods previously evoked. We can not show the statistical significance of forecast in all cases, since the differences between the obtained results are relatively small according to the NRMSE, but practically, by making more predictions, it is important to take into account any improvement. As a side note, it is worth to mention that some authors, such as [23], have argued that statistical significance testing of forecast accuracy should be avoided, as test results may be misleading and that practice may actually harm the progress of forecasting field. However, in Figure 3 we compare the number of features that provides the best accuracy for each method with the minimal number giving better or the same forecast accuracy by our proposal. We remark that the performance of those methods can be reached by our proposal using smallest number of features in most cases.

VIII. CONCLUSIONS

In the context of forecasting with many variables, the goal is to develop optimized models, performing both descriptive and predictive tasks [24]. That can be achieved, in (i) by optimizing the structure of the multivariate models, *i.e.*, reducing the number of predictors, while improving the forecast accuracy, and (ii) by providing an explanation of the dependencies between all variables. The application of feature selection and dimension reduction methods as a preprocessing step before the prediction is a reasonable solution to this issue, except that the former are slightly advantageous since they extract a subset of variables from the originals, while the latter reduce dimensionality by generating artificial variables. In the litterature, a considerable interest has been paid to correlation-based methods. That can be coherent regarding to regression or classification. But in forecasting, and especially with lags, the predictive aspect of the selected features is not negligible. In comparison, a little attention has been paid to the role of causality in feature selection. In the current reseach, we investigated its role in the context of time series forecasting and propose the Granger Selection Method.

Experiments on real data sets and a comparative study with others methods show an improvement of the forecast accuracy and a reduction of the number of input predictors. The measure adopted is the Granger causality, but the proposed algorithm is applicable for other measures of dissimilarity between time series. In the future, we aim to adopt a more deeper analysis on the graph of causalties than the clustering approach, in order to tackle dependencies between time series. We aim also to apply our approach on other prediction models, as well as study the applicability of feature selection methods according to the types of models (prediction, classification, regression, *etc.*).

 ISBN: 978-1-61208-558-6

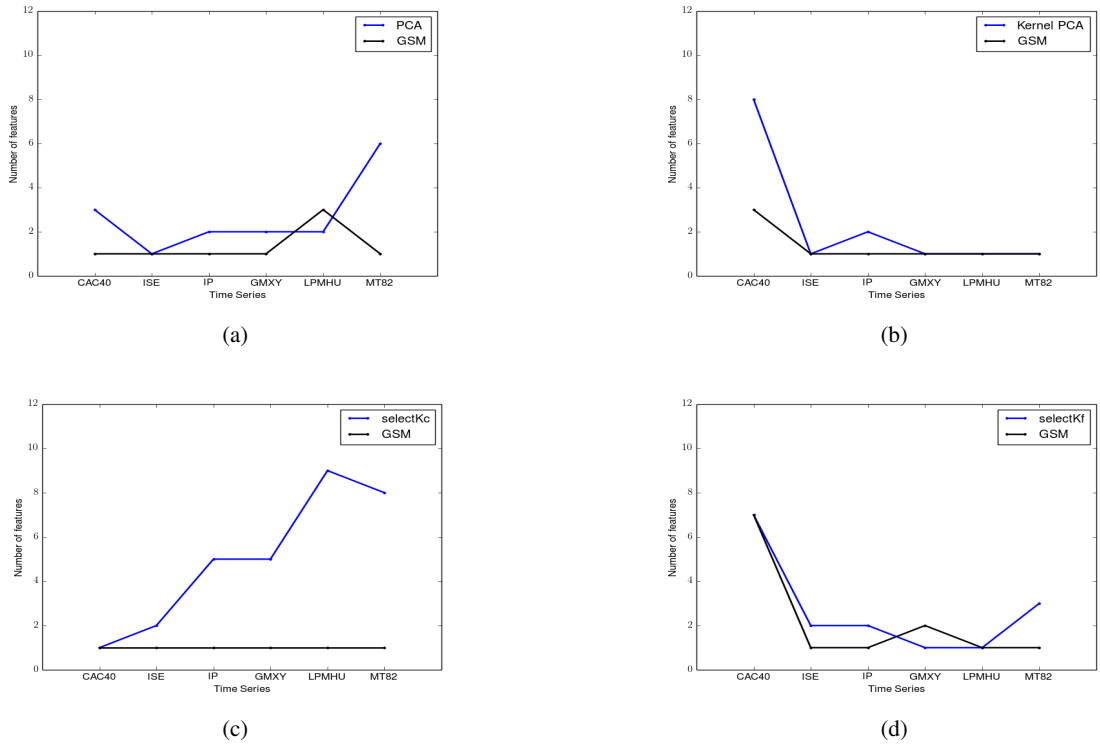

Figure 3. Comparison On The Number Of Predictors Providing The Same Or Better Forecast Accuracy By Our Proposal With The Methods Used.

REFERENCES

[1] G. Walker, "On periodicity in series of related terms," Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, vol. 131, no. 818, 1931, pp. 518–532.

[2] P. Whittle, "The analysis of multiple stationary time series," Journal of the Royal Statistical Society. Series B (Methodological), 1953, pp. 125–139.

[3] H. Lütkepohl, New introduction to multiple time series analysis. Springer Science & Business Media, 2005.

[4] B. Jiang, G. Athanasopoulos, R. J. Hyndman, A. Panagiotelis, F. Vahid et al., "Macroeconomic forecasting for Australia using a large number of predictors," Monash University, Department of Econometrics and Business Statistics, Tech. Rep., 2017.

[5] C. W. Granger, "Testing for causality: a personal viewpoint," Journal of Economic Dynamics and control, vol. 2, 1980, pp. 329–352.

[6] G. E. Box, G. M. Jenkins, and G. C. Reinsel, "Time series analysis: Forecastingand control," San Francisco: Holdenday, 1976.

[7] M. Quenouille, The analysis of multiple time-series, ser. Griffin's statistical monographs & courses. Griffin, 1957.

[8] S. Johansen, "Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models," Econometrica: Journal of the Econometric Society, 1991, pp. 1551–1580.

[9] J. H. Stock and M. W. Watson, "Forecasting with many predictors," Handbook of economic forecasting, vol. 1, 2006, pp. 515–554.

[10] H. Abdi and L. J. Williams, "Principal component analysis," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 4, 2010, pp. 433–459.

[11] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," Neural computation, vol. 10, no. 5, 1998, pp. 1299–1319.

[12] X.-w. Chen and J. C. Jeong, "Enhanced recursive feature elimination," in Machine Learning and Applications, 2007. ICMLA 2007. Sixth International Conference on. IEEE, 2007, pp. 429–435.

[13] X. Zhong and D. Enke, "Forecasting daily stock market return using dimensionality reduction," Expert Systems with Applications, vol. 67, 2017, pp. 126–139.

[14] P. C. Molenaar, "A dynamic factor model for the analysis of multivariate time series," Psychometrika, vol. 50, no. 2, 1985, pp. 181–202.

[15] B. Abraham and G. Merola, "Dimensionality reduction approach to multivariate prediction," Computational statistics & data analysis, vol. 48, no. 1, 2005, pp. 5–16.

[16] Y. Sun, J. Li, J. Liu, C. Chow, B. Sun, and R. Wang, "Using causal discovery for feature selection in multivariate numerical time series," Machine Learning, vol. 101, no. 1-3, 2015, pp. 377–395.

[17] C. W. Granger, B.-N. Huangb, and C.-W. Yang, "A bivariate causality between stock prices and exchange rates: evidence from recent asian-flu?" The Quarterly Review of Economics and Finance, vol. 40, no. 3, 2000, pp. 337–354.

[18] L. Kaufman and P. J. Rousseeuw, "Partitioning around medoids (program pam)," Finding groups in data: an introduction to cluster analysis, 1990, pp. 68–125.

[19] W. N. Venables and B. D. Ripley, Modern applied statistics with S-PLUS. Springer Science & Business Media, 2013.

[20] H. Akaike, "A new look at the statistical model identification," IEEE transactions on automatic control, vol. 19, no. 6, 1974, pp. 716–723.

[21] M. Lichman, "UCI machine learning repository," 2013, URL: http://archive.ics.uci.edu/ml [accessed: 2017-02-16].

[22] J. H. Stock and M. W. Watson, "New indexes of coincident and leading economic indicators," NBER macroeconomics annual, vol. 4, 1989, pp. 351–394.

[23] J. S. Armstrong, "Significance tests harm progress in forecasting," International Journal of Forecasting, vol. 23, no. 2, 2007, pp. 321–327.

[24] G. Shmueli et al., "To explain or to predict?" Statistical science, vol. 25, no. 3, 2010, pp. 289–310.

 ISBN: 978-1-61208-558-6