Analysis of multivariate time series predictability based on their features

Conference Paper · November 2020

DOI: 10.1109/ICDMW51313.2020.00055

2 authors, including:

Anton Kovantsev
ITMO University
14 PUBLICATIONS **40** CITATIONS

SEE PROFILE

Analysis of multivariate time series predictability based on their features

Anton Kovantsev and Peter Gladilin
ITMO University
Saint-Petersburg, Russia
ankovantcev@itmo.ru, peter.gladilin@gmail.com

***Abstract* — In this study we explore the features of time-series that can be used for evaluation of their predictability. We suggest using features based on Kolmogorov-Sinai entropy, correlation dimension and Hurst exponent to test multivariate predictability. Besides we use two new features such as 'noise measure' and 'random walk detection'. Then we experimentally test the accuracy of multivariate time series forecasting models, including vector autoregressive model (VAR), multivariate singular spectrum analysis (MSSA) model, local approximation (LA) model and recurrent neural network model with long short term memory (LSTM) cells. At last we test different causality methods for choosing additional time series as the predictors and claim that the relevance of taking into account additional predictors highly depends on the characteristics of the target time series and can be estimated using the developed method. The results of the work can be used as theoretical and experimental basis for the development of forecasting applications for the short time series using a combination of corporate and open source data as additional data predictors.**

Keywords—time series predictability, forecasting, causality, local approximation, clustering, noise measure

I. Introduction

The successful solution for the task of time-series forecasting hardly depends on their internal properties, and sometimes it's important to know in advance whether the predicted values would be of sufficient accuracy and should some additional efforts to be undertaken to improve the forecasting result. Dealing with additional concomitant factors is of increased importance with a lack of data, for example, when it is required to predict time series of complex structure and at the same time of small length. For complex models that have proven themselves in predicting complex time series, there is simply not enough data for the algorithm to converge and the model to start producing a high-quality forecast.

One of the troubles with forecasting accuracy of short time series is in their irregularity caused by some unobserved reasons. 'Short' means here that the series contains about 50-100 observations. In order to make these reasons more obvious we tried to get more information about 'environment' that can have some influence on the series we are dealing with.

We provide experiments with clustering the dataset of time-series according to their features and compare the result with the forecasting accuracy to evaluate their predictability and to set a benchmark for the accuracy that can be expected for each series before modeling. Then we analyze the applicability of different causality tests for making the forecast by different multivariate models. Vector Autoregression Model (VAR), Multivariate Singular Spectrum Analysis (MSSA) method, Recurrent Neural Network Model (RNN) and Local Approximation (LA) method are explored in this study.

As a secondary result we've got a number of metrics of the target time series that can help to separate the time series which forecasting can be improved by means of multivariate models from those that is probably useless. The work also examined the issue of justification for taking into account new factors (and, as a consequence, the complications of the basic forecasting model) for time series with different characteristics. We conducted experimental studies of the predictive ability of various models on the assembled dataset of time series, divided into two clusters in accordance with their features.

The rest of the paper is organized as follows: Section II represents related work on experiments with multivariate forecasting models and methods for choosing data predictors. Section III gives description of the features that can be used for predictability evaluation of the collected time series, Section IV presents the causality tests and Section V describes the forecasting models for experiments. Section VI is for the experimental results and their discussion. Finally, Section VII is for conclusions.

II. Related Works

With an increasing availability of data of various nature, it became possible to test various algorithms, including machine learning and deep learning models on large datasets in order to verify the quality of their prediction. The most comprehensive study on this subject is presented in [1], which compared the quality of prediction of both classical and, in part, modern machine learning models on the huge M4 dataset, containing more than 100 000 time series. These results show that the best accuracy on the M4 dataset can be achieved by the combination of both statistical and machine learning methods.

Often there are problems when one-dimensional forecasting is not enough. Examination of possible relationships between economic, financial and other factors play a role in optimizing business processes and operating more informatively in accordance with the available data. Knowing how the change in the predictor value will affect the target variable, i.e., the nature of this relationship and the delay (time lag between the trigger and the consequence) could significantly improve the quality of predictive models, especially if such dependence is directed in automatic mode. These tasks contribute to research for the architecture of the multidimensional predictive models and how to choose predictor series. One of the popular and simple multivariate forecasting models is VAR, which is the enhancement to the multidimensional case of autoregressive models [2]. VAR is

2020 International Conference on Data Mining Workshops (ICDMW) | 978-1-7281-9012-9/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICDMW51313.2020.00055

2375-9259/20/$31.00 ©2020 IEEE
DOI 10.1109/ICDMW51313.2020.00055

able to understand and use the relationship between several variables. This is useful for describing the dynamic behavior of the data and implementing VAR is rather simple [3].

But in order to effectively add time series to multivariate models such as VAR, it is necessary to choose them wisely from a set of potential predictors. The problem of the predictors selection for the forecasting is widely discussed in the literature [4]. It can be done not only with the classic causality tests (like Granger causality test or cross-correlation analysis) but with the approaches which is based on random walks with restarts on directed causality graphs. Another way is to explain causalities between variables in time series using the Granger causality graph with clustering as the result [5]. Choosing predictors from the known set of the historical measurements can be also done by applying the method of principal components to the targeted predictors selected using hard and soft threshold rules for reduction of the number of additional components in the model [6], [7].

Other researchers try to implement deep learning architectures for the multivariate forecasting together with investigating causality dependencies. For example, in [8] there was trained a model on multiple stock series, that allowed the network to exploit the correlation structure between these series so that model would learn the market dynamics in shorter sequences of data. In that case long-term temporal dependencies in and between financial time series was learned using deep convolutional neural network based on the WaveNet model [8]. In [9] the results of the initial LSTM model with multiple machine learning models, like Extreme Gradient Boosting model (XGB) and Random Forrest Regressor model (RFR) were compared. The prediction accuracy of the benchmark model was improved by incorporating features that describe the forward steps of the target variable.

However, it may occur that target time series is definite and structured enough to avoid using any additional predictors in order not to complicate the forecasting model unnecessarily. So, the first step in the analysis should be checking whether target time series can be predicted with acceptable accuracy without predictors. For example, in [10] and [11] the authors discuss the issues of predictability evaluation of the time-series. Our study differs in introducing the specific set of features including 'noise measure' and 'random walk detection' which can help to estimate time series predictability and at the same time justification of using additional predictors for forecasting. The experimental results of this research are presented in the following sections.

III. Time Series Features

For experimental setup, we have collected 110 short time series of 50 steps from the finance and economics, containing monthly data with the dynamics of various values, including sales of various goods and services, exchange rates, demand for goods and services, CPI, etc. Time series have different values for seasonality, trend strength, and signal-to-noise ratio. All the models for experiments here were developed in Python 3.6, using *keras* framework, *statsmodels* package, *pymssa* and *skccm* modules.

Given a time series $X = \{x_1, x_2, \ldots . x_m\}$ we test its predictability by evaluating specific characteristics. The following features for the time series were calculated: correlative dimension, estimated correlative entropy, Kolmogorov-Sinai entropy, Hurst's exponent, memory depth, noise measure and random walk detection. These features represent almost complete set of characteristics for dynamic systems that generate the time-series [12]. That's why they are considered to be useful for predictability exploration. Which of them actually can be used in time series ranging is found out as the result of the experiments with clustering (see Section VI). Below we provide a brief overview of these values.

A. Correlation dimension

Correlation dimension is a measure of the dimensionality of the space occupied by a set of random points, often referred to as a type of fractal dimension. It's less noisy when only a small number of points is available, and is often in agreement with other calculations of dimension. It can be calculated by means of the correlative integral for time series of finite length:

$$C(r) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} \frac{\theta\big(r - \rho(i,j)\big)}{m(m-1)}, \quad (1)$$

where $\rho_k(i,j) = \sqrt{\sum_{l=1}^{k}\big(x_{i-k+l} - x_{j-k+l}\big)^2}$, $\theta(x)$- Heaviside function, r - characteristic phase space cell's size.

The value of correlation dimension is the slope of the logarithmic graph of the correlation integral [12] and can be evaluated as the following limit:

$$d_k = \lim_{r\to 0} \lim_{m\to\infty} \frac{lnC(r)}{lnr}. \quad (2)$$

If the finite-dimensional embedding of the system represented by the time series X exists, then with an increase in k the tangent of the angle of inclination $ln\big(C(r)\big)[ln(r)]$will tend to a certain finite value. This value will be equal tod_k. The value of k at which the slope stops changing gives the dimension of the embedding. Figure 1 shows an example of the relation $ln\big(C(r)\big)[ln(r)]$ for different k. Thus, correlation dimension can be calculated as a linear coefficient of the first segment of the correlation integral graph between two first points for consequently increasing k until this coefficient's changes will occur to be relatively small.

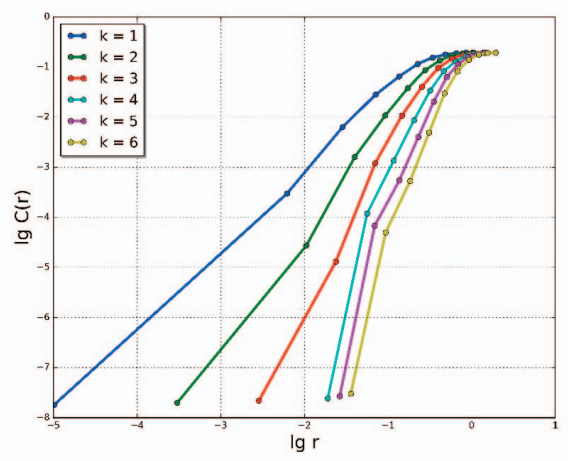

Fig. 1. Logarithmic graph of the correlative integral. For this example k=6, correlative dimension $d_k \approx 8$.

B. Correlation entropy

Correlation entropy helps to find out how many basic degrees of freedom are in a noisy series. To estimate the correlation entropy $E_{app}(n,r)$ in successive spaces of (n)- and (n+1)-dimensional embeddings, two attractors are generated based on the system represented by the time series. So that, the entropy can be evaluated as the logarithm of two sequential correlation integrals ratio [12]:

$$E_{app}(n,r) = -ln\frac{C^{n+1}(r)}{C^{n}(r)} \tag{3}$$

C. The Kolmogorov-Sinai entropy

The Kolmogorov-Sinai (KS) entropy is taken into account as a metric invariant of dynamical system that generates the time series. Following [12] this entropy can be calculated based on the entropy rates of finite partitions of the state space of the series. Given a finite k embeddings of phase space M of the of dynamical system that produces the time series $\xi = \{C_{(1)}, C_{(2)}, \dots, C_{(k)}\}$ one can calculate the probability measure of i-item $\mu(T^{-1}C_{i_1} \cap \dots \cap T^{-n}C_{i_n})$, where T - is a measurable transformation of phase space M onto itself preserving the measure μ, $n \in (-\infty, \infty)$, $i \in [1,k]$. The entropy rate $h_\mu(T, \xi))$ is the limit that can be evaluated as follows:

$$h_\mu(T,\xi) = -\lim_{n\to\infty}\frac{1}{n}\sum_{i_1,\dots,i_n}\mu(T^{-1}C_{i_1} \cap \dots \cap T^{-n}C_{i_n}) \times ln\,\mu(T^{-1}C_{i_1} \cap \dots \cap T^{-n}C_{i_n}) \tag{4}$$

To evaluate $h_\mu^{KS}(T)$ we take the supremum of the entropy rate $h_\mu(T, \xi)$ over all finite partitions:

$$h_\mu^{KS}(T) = \sup_{\xi} h_\mu(T, C). \tag{5}$$

So, we take the histogram divisions of the series' values by every number of bins from one to the half of the time series length as the finite partitions of phase space. Then we calculate the entropy rate for each of them and find the supremum that helps to estimate Kolmogorov-Sinai entropy.

D. Hurst exponent

Hurst exponent is used as a measure of long-term memory of time series. It can be estimated by the re-scaled range $R(\tau)$ and standard deviation $S(\tau)$ of the series following the formulas below [13]:

$$H(\tau) = \lim_{\tau\to\infty}\frac{log\frac{R(\tau)}{S(\tau)}}{log(\alpha\tau)}, \tag{6}$$

where $\tau \in [3, N], \alpha = 0.5$,

$$R(\tau) = \max_{1\le t\le\tau}\sum_{j=1}^{\tau}(x_j - \acute{x}_\tau) - \min_{1\le t\le\tau}\sum_{j=1}^{\tau}(x_j - \acute{x}_\tau) \tag{7}$$

and

$$S(\tau) = \sqrt{\frac{1}{\tau}\sum_{t=1}^{\tau}(x_t - \acute{x}_\tau)^2}. \tag{8}$$

These re-scaled range $R(\tau)$ and standard deviation $S(\tau)$ are used to represent the *R/S*-trajectory $RS(\tau) = \frac{R(\tau)}{S(\tau)}$, that helps to estimate the *memory depth* of the time series. In practice *H* is analyzed to estimate the type of the noise in the series. The memory depth of the system is determined by the element for which the R/S-trajectory receives a negative increment – one can define it as the first point where the curve changes from increasing to decreasing. For example, in Fig. 2 this point in log scale equals approximately *1.1*; so that, *H=0.46* and memory depth for this series is *4*).

For the evaluation of Hurst exponent, we simply applied the equations (6-8) limited by the length of time series.

E. Noise measure

Noise measure feature is based on comparing the standard deviation of a time-series with the standard deviation of its first-order differences as it's presented below. The idea is to detect high-frequency noise that increases the comparative deviation of value change on each time step; so, the higher value of this measure means the lower noise influence in time series and vice versa. The formula to evaluate the noise measure is the following:

$$F_N = 1 - \sqrt{\frac{N\cdot\sum_{i=1}^{N-1}(x'_i - \acute{x}')^2}{(N-1)\cdot\sum_{i=1}^{N}(x_i - \acute{x})^2}} \tag{9}$$

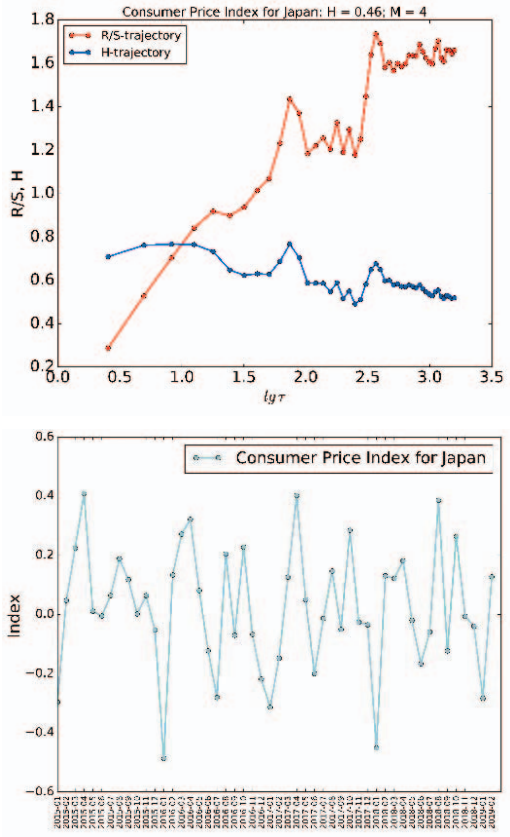

Fig. 2. Example of R/S-trajectory (upper) for the monthly Consumer Price Index in Japan series (lower)

where $x'_i = x_{i+1} - x_i$ - is the first-order differences, $\acute{x}, \acute{x}'$ - mean values of the initial time series and corresponding differenced series.

This measure is used to find out if the time series is the result of a random walk process. In this case the first-order differences should be noised. So that, the noise detection in first differences mean that the random walk occurs in the time series. This approach is based on ideas of Robert M. May firstly published in 1976 [14].

IV. Causality Analysis

The analysis of causation between potential predictor series and time-series of source data, as well as the analysis of predictability is usually based on cross-correlation analysis methods, Granger's causality test and a cross-convergent mapping method.

A. Method of Granger causality analysis

Granger's causality test is a statistical tool used to verify the existence of a causal relationship between two time series. Causality test according to Granger, which tests a null hypothesis of absence of causal relationship between the series $X_1(t)$ and $X_2(t)$ can be expressed as the following:

$$\begin{aligned} X_1(t) &= \sum_{j=1}^{p} A_{11,j} X_1(t-j) + \sum_{j=1}^{p} A_{12,j} X_2(t-j) + E_1(t) \\ X_2(t) &= \sum_{j=1}^{p} A_{21,j} X_1(t-j) + \sum_{j=1}^{p} A_{22,j} X_2(t-j) + E_2(t) \end{aligned} \quad (10)$$

where p – is the maximum number of lagged observations included in the model (the model order), the matrix A contains the coefficients of the model (i.e., the contributions of each lagged observation to the predicted values of $X_1(t)$ and $X_2(t)$, and E_1 and E_2 are residuals (prediction errors) for each time series. If the variance of E_1 (or E_2) is reduced by the inclusion of the X_2 (or X_1) terms in the first (or second) equation, then it is said that X_2 (or X_1) Granger-causes X_1 (or X_2) [15].

B. Convergent Cross Mapping Method

The analysis of causality between time-series-potential predictors and the series of source data with the use of Convergent Cross Mapping (CCM) method was performed based on the procedures of the *skccm* library, created in the laboratory of Dr. Sugihara [16]. The CCM method is a statistic test of causality between two time-series which, like Granger causality test, seeks to address the problem that correlation does not always imply causality. The CCM method is driven by the dynamic systems theory and relies on Takens's embedding theorem, which states that in the general case, the variety of attractors of a dynamical system can, under particular conditions, be reconstructed on the basis of the limited variety of this system [16].

C. Cross correlation

In spite the statement that "correlation does not imply causality" a time series closely correlated to target data can be useful as a predictor. Having in mind all warnings concerning cross correlation this test nevertheless have been included to the set of prediction ability tests [17].

V. Models

We test four multivariate forecasting models in order to check availability of the suggested approach for the clustering of initial time series in the sense of potential of increasing accuracy with adding predictors: vector autoregressive model (VAR), multivariate singular spectrum analysis (MSSA) model, local approximation (LA) model and recurrent neural network model with long short term memory (LSTM) cells. Below we present short summary of model's implementations.

A. Vector autoregression

The vector autoregression model (VAR) is arguably the simplest and most often used multivariate time series model for forecasting [2]. It is multidimensional case of the univariate autoregressive (AR) model. It describes the further values of a set of time dependent variables as a linear function of their past values:

$$R_t = \varphi_0 + \Phi R_{t-1} + \varepsilon_t \quad (11)$$

where R – is the vector of initial values,Φ – is the matrix of weights that can be evaluated during the training process, φ_0 – is a constant, ε_t –is an error process. As it follows from (10) VAR model only depends on lagged variables from the input time series.

B. Local approximation method

The idea of local approximation (LA) method is in taking not the nearest in time but closest in state space variables as arguments for regressive model. It's described in [12] and looks very close to the k-nearest neighbors' method with selection of local phase space is used instead of nearest neighbors selection. That's possible because a description of the phase space of a dynamical system can be obtained if we take p – dimensional vectors of delays from the values of the series at successive times instead of the real variables of the system according to Tackens' theorem. The algorithm of LA method can be described as follows. Target time series $\{x_i\}$ is transformed to delay matrix X (12), also known as Hankel trajectory matrix:

$$\{x_1, x_2, \cdots, x_N\} \to X_{p\times(N-p+1)} = \begin{pmatrix} x_p & \cdots & x_N \\ \vdots & \ddots & \vdots \\ x_1 & \cdots & x_{N-p+1} \end{pmatrix}, \quad (12)$$

where p – is the parameter of the method. Next step of the algorithm is the allocation of the specific region in the phase space (nearest neighbors' approach):

$$\{x_s\}: \sum_{s\in w_m} \left\| x_{N-p+1} - x_s \right\| \to \min_{w_{N-p+1}}, \quad (13)$$

where m – number of neighbors, w_m – indexes of neighbor vectors in delay matrix X; $m \geq p+1$; practically $m \approx 3\cdot(p+1)$.

Model of local representation is fitted with the method of least squares:

$$\{\hat{a}\}: \sum_{w_s} [x_{t+1} - (a_0 + x_t^T \cdot \hat{a})]^2 \to \min_{\hat{a}} \quad (14)$$

The prediction of the next step of time series can be evaluated as follows:

$$\hat{x}_{t+1} = f(x_t, \hat{a}) = (a_0 + x_t^T \cdot \hat{a}) \quad (15)$$

This method is also can be used for predicting the sequence of several forward steps if the vector of values x_t is used instead of scalar value.

The LA algorithm has one clear advantage over conventional auto-regression. It consists of the use of a piece-wise linear approximation instead of a globally linear (uniform for the whole series). This allows using LA to successfully predict irregular (means quasi-periodic, chaotic) time series for which a linear auto-regressive representation is not acceptable [12].

LA algorithm has been modified to give it an opportunity to use predictors. Predictor's series values are concatenated with x_t^T values in local representation (14) with the use of the

method of least squares (15) so that the predictor can produce its influence on the forecasting result.

C. Multivariate Singular Spectrum Analysis (MSSA)

Methodology of MSSA is described, for example, in [18] and is the enhancement of singular spectrum analysis to the multidimensional task. The classical SSA aims to decompose time series into several components, which are recognized as either a trend, periodic or quasi-periodic signal or residual noise. SSA approach basely consists of decomposition and reconstruction. The trajectory matrix of multi-channel time series consists of linked trajectory matrices of separate time series. The rest of the algorithm is the same as in the uni-variate case. Considering P time series with length N embedding is identified as a mapping step that transfers one-dimensional time series $X^{(p)} = \{x^{(p)}_1, x^{(p)}_2, \ldots, x^{(p)}_N\}$ into a multidimensional matrix $Y^{(p)} = [Y^{(p)}_1, \ldots, Y^{(p)}_M]$ with vectors $Y^{(p)}_j = (x^{(p)}_j, \ldots, x^{(p)}_{j+M-1})^T \in R^M$, where M corresponds to the window length for each channel and $K=N-M+1$. Therefore, we can obtain a trajectory Hankel matrix. Therefore, P trajectory matrices $Y^{(p)}$ $(p=1, 2, \ldots, P)$ can be obtained by operating embedding procedure. Then the blocked Hankel trajectory matrix Y is defined as follows:

$$Y = \left[Y^{(1)^T}, Y^{(2)^T}, \cdots, Y^{(P)^T}\right] \tag{16}$$

Then Hankel trajectory matrix is deconstructed into two matrices U and V:

$$Y = \alpha^{\frac{1}{2}} U \Lambda V^T, \tag{17}$$

where α - is the normalization factor, Λ - is the diagonal matrix of eigenvalues of U. The left and right singular vectors U and V^T reveal spectral features of the co-variance structure that are universal to all channels.

D. Recurrent neural network model

The recurrent neural network model with the elements of Long Short-Term Memory method (LSTM) was implemented with the use of the *keras* library procedures. Application of the second hidden recurrent layer in the model allows to significantly reduce the required total number of hidden nodes (hence, parameters for training). According to the work [6], two hidden layers should be used to compensate the loss of efficiency when using classic activation functions like *ReLU* and *tanh*. Experiments proved that the LSTM layer inserted before the predictor's part through the concatenation operation makes prediction more stable and therefore improves its quality [6], [8], [19].

The number of elements in hidden layers of the network has been selected experimentally. The best prediction accuracy was achieved with 128 LSTM cells in each hidden layer.

VI. Experimental Results

A set of real 110 time series with length of 50 observations with monthly periods was collected from open sources to fulfill the experiments. These series concern various global, regional and national economic and social parameters such as prices for goods, raw materials and natural resources, currency exchange rates, stock exchange indexes, sales of various stuff, wages, population in different social groups etc. For example, 'US Advance Retail Sales Excluding Food', 'Consumer Price Index in Germany', 'Latin America Electrical manufacturing import', 'Oil mining world', 'United States Market Total Auto' and so on. We tried to collect the various time series with different trends, seasonality and random components.

The 6-steps forward forecasting was made for each of the series by means of different models first separately then with additional predictors chosen from the rest of the set according to different causality tests. In Fig. 3 one can see the result of the mean absolute percentage error score (MAPE) evaluation on the forecasting. Different causality tests perform differently in improving the accuracy of the forecasting. As we can notice, the accuracy highly depends on the type of forecasting model, and the LA model performs best regardless of the type of causality test with a minimum spread of prediction accuracy values. As it can be seen in Fig. 3, it is not possible to trace a clear dependence of the quality of models on the inclusion of additional predictors in any of the causality tests.

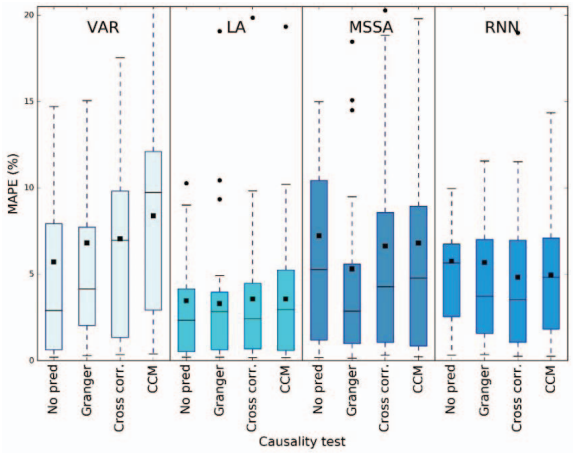

Fig. 3. Experimental results on the comparative study of the MAPE distribution for the four investigated multivariate models with one additional series-predictor. Four different causality approaches for choosing predictors were tested.

Then the features listed in Section 3 were calculated for each of the time-series. So we have got a data frame of the time series, the calculated features and the error scores of their forecasting with and without additional predictors. The initial dataset was divided into two groups regarding these features by k-means clustering method. Various combination of the features were tested, and it turned out that the sharpest division of the clusters occurred when the characteristics of dynamic system dimension and memory are excluded while the entropy and noise estimations are taken into account. Thus, for the clustering we used the following features: noise measure, correlative entropy, KS-entropy, Hurst's exponent and random walk measure. The other features either have no influence in clustering or reduce the difference between the clusters.

As the set of time series has been divided, we have compared the MAPE score of the predictions in the clusters. It turned out that the forecasting error in A-cluster mostly less than in B-cluster, so we can make a conclusion about time series predictability in different clusters and its accordance to applied features. It should be noted that for each of the

features used, the correlation with the forecast error and with the effect of improving the quality of prediction due to the use of predictors is rather weak, but the whole set of the features allows the efficient clustering.

The result of the clustering illustrated by means of 2D t-SNE dimension reduction is shown on Fig. 4 (a).

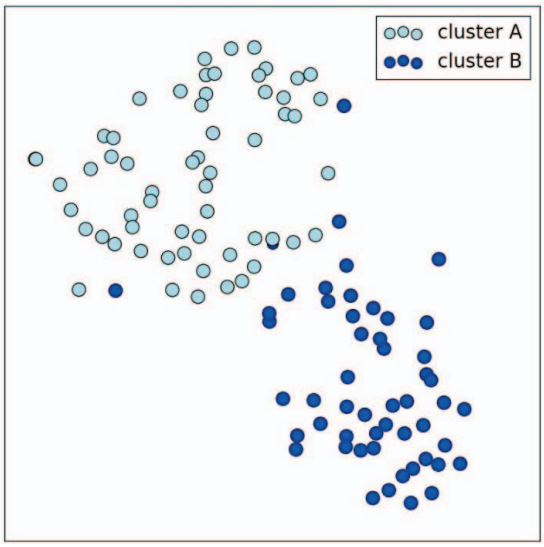

Fig. 4 (a). Experimental result of t-SNE visualisation for k-means clustering. Each point corresponds to a single time series.

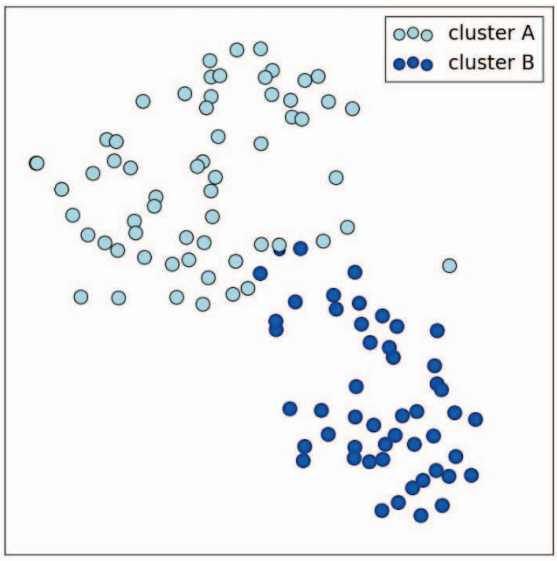

Fig.4 (b). Experimental result of t-SNE visualisation for DBSCAN clustering. Each point corresponds to a single time series.

It should be noted that the result of DBSCAN clustering with default parameters of *sklearn* library almost approves the result of k-means: only 6 of 110 points differ (see Fig. 4 (b)). One can see that two distinct clusters are distinguished, corresponding to a different set of features of time series.

The detected predictability in terms of MAPE distribution for all the time series in the dataset is illustrated with boxplots in Fig. 5. One can see that in B-cluster the accuracy is mostly worse than in A-cluster. The difference in median MAPE for the A and B cluster is statistically significant for all the models tested and reaches about 3% MAPE in the case of using LA model (the smallest) and up to 10% in the case of MSSA model (the largest). So we can evaluate the predictability of a time series in accordance of its features and calculate with some probability in advance if the prediction would be more or less accurate.

As the result of the unsupervised clustering each time series were marked as "A" or "B" in accordance with the label affixed by the algorithm. It turned out that there are 51 time series in the A cluster and 59 in the B cluster. In cluster B

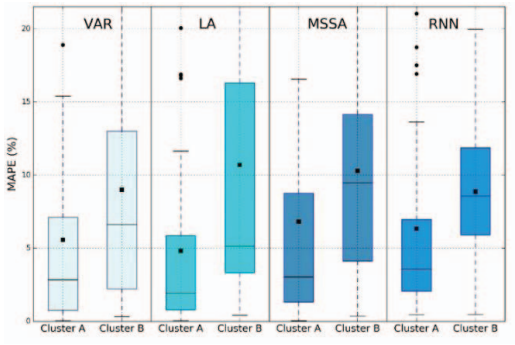

Fig. 5. MAPE distribution in clusters A and B for the one-variate forecasting task without predictors

time series have relatively large signal-to-noise ratio and unexpressed seasonal patterns than in cluster A, as it's illustrated with time series frames on fig 6.

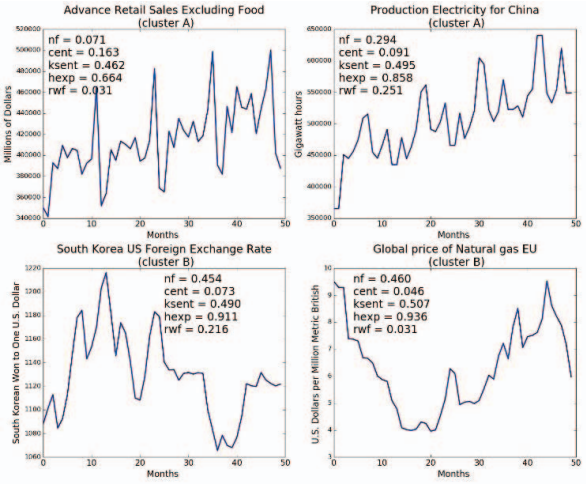

Fig.6. Examples of time series graphs for A and B clusters: 'nf' - noise measure; 'cent' - corr. entropy; 'ksent' - Kolmogorov-Sinai entropy; 'hexp' - Hurst exponent; 'rwf' - random walk detector.

In order to check in advance either additional data can be useful or not for A and B clusters we have evaluated relative MAPE alteration in the case of using predictors. The forecasting task was based on 6 steps forward as well by four models without predictors, with one predictor, with three and five predictors consequently. Each time, a MAPE value was determined as a metric. Because the MAPE value for the forecasting without predictors for different series is significantly different, the conclusion about the use of predictors for each model was made on the basis of the

relative change in MAPE for the best setup (with minimum MAPE) with the use of predictors, compared with the base model without predictors:

$$\Delta MAPE = \frac{MAPE_0 - \min\limits_{i=1,3,5}(MAPE_i)}{MAPE_0} \cdot 100\% \qquad (18)$$

In Fig. 7 one can see the relative MAPE alteration that was achieved by means of using predictors with each of the models for both A and B clusters. It's obvious that in A-cluster the effect of additional time series is less than in B-cluster for all the models except, local approximation model, which did not show a significant difference in the distribution of errors.

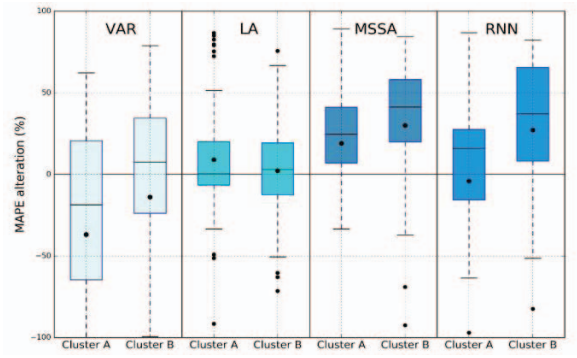

Fig.7. Relative MAPE alteration caused by using of the predictor time-series in two clusters.

So, we may conclude that using predictors for B-cluster would be more effective than for A-cluster series especially making prediction with MSSA and RNN models: median MAPE alterations for them are 30.0% and 27.2% correspondingly.

VI. Conclusion

We have presented a method of predictability evaluation of a time series and a comparative analysis of multivariate forecasting models in terms of using different causality tests for choosing the most suitable predictors. The presented method of unsupervised clustering of time series into two types have shown that characteristics of the initial series would significantly affect the forecasting accuracy and helps to define the expected forecasting quality. In addition, it can be useful for making decision about relevance of using additional series as predictors. Studies have shown that for noisier data with no pronounced patterns (cluster B), the use of additional predictors is more justified. Proposed methods for noise measuring and random walk detecting along with other time series features have their effect in the evaluation of time-series predictability and in investigation of predictor usefulness.

A forecasting model for the time series was developed with automatic selection of predictor series with a significant causal relationship with respect to the target variable. The selection of significant variables for improving the quality of the forecast of time series is performed in the algorithm using the cross-correlation method, the Granger's causality test and the convergent cross-mapping method. Experimental studies of the multivariate models showed that the suggested approach for the automatic selection of additional financial and economic indicators, collected from open sources, is able to reduce the relative error of benchmark models up to 30% if adding predictors is applicable for this data, that is, the series refers to cluster B.

Using the developed clustering method, it is possible to evaluate the expected predictability of the time series both within the framework of one-dimensional forecasting and within the framework of multidimensional forecasting using additional predictors. Results of the study can be used in evaluation of predictability and in prediction tasks in various business areas using the developed software. Also, the results of the work can be used as a theoretical and experimental basis for the development of applied services for forecasting financial time series on a combination of corporate and open source data.

Acknowledgements

This research is financially supported by The Russian Science Foundation, Agreement №19-71-10078.

References

[1] Makridakis S., Spiliotis E., Assimakopoulos V. The M4 Competition: Results, findings, conclusion and way forward // Int. J. Forecast. Elsevier B.V., 2018. Vol. 34, № 4. P. 802–808.

[2] Singh A. A Multivariate Time Series Guide to Forecasting and Modeling. 2018. P. 1–12.

[3] Jones S.S. et al. A multivariate time series approach to modeling and forecasting demand in the emergency department // J. Biomed. Inform. 2009. Vol. 42, № 1. P. 123–139.

[4] Przymus P. et al. Improving multivariate time series forecasting with random walks with restarts on causality graphs // IEEE Int. Conf. Data Min. Work. ICDMW. 2017. Vol. 2017-Novem. P. 924–931.

[5] Youssef H., Casali A., Lakhal L. A Causality Based Feature Selection Approach for Multivariate Time Series Forecasting // DBKDA 2017, Ninth Int. Conf. Adv. Databases, Knowledge, Data Appl. 2017. № July. P. 97–102.

[6] Bai J., Ng S. Forecasting economic time series using targeted predictors // J. Econom. 2008. Vol. 146, № 2. P. 304–317.

[7] Stock J.H., Watson M.W. Forecasting using principal components from a large number of predictors // J. Am. Stat. Assoc. 2002. Vol. 97, № 460. P. 1167–1179.

[8] Borovykh A., Bohte S., Oosterlee C.W. Conditional time series forecasting with convolutional neural networks // Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2017. Vol. 10614 LNCS. P. 729–730.

[9] Helmini S. et al. Sales forecasting using multivariate long short term memory network models. 2019.

[10] Garland J., James R., Bradley E. Model-free quantification of time-series predictability // Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys. 2014. Vol. 90, № 5.

[11] Xu P. et al. On predictability of time series // Phys. A Stat. Mech. its Appl. Elsevier B.V., 2019. Vol. 523. P. 345–351.

[12] Loskutov A.Y. et al. Local methods of time series forecasting. // Problems of nonlinear dynamics. Moscow Sta. Moscow, 2002.

[13] Mitra S.K. Is Hurst exponent value useful in forecasting financial time series? // Asian Soc. Sci. 2012. Vol. 8, № 8. P. 111–120.

[14] May R.M. Simple mathematical models with very complicated dynamics // The Theory of Chaotic Attractors. 2004. P. 85–93.

[15] Eichler M. Causal inference in time series analysis // Causality: Wiley Series in Probability and Statistics. 2012. P. 6–28.

[16] Clark A.T. et al. Spatial convergent cross mapping to detect causal relationships from short time series // Ecology. 2015. Vol. 96, № 5. P. 1174–1181.

[17] Dean R.T., Dunsmuir W.T.M. Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models // Behav. Res. Methods. 2016. Vol. 48, № 2. P. 783–802.

[18] Mao X., Shang P. Multivariate singular spectrum analysis for traffic time series // Physica A: Statistical Mechanics and its Applications. 2019. Vol. 526.

[19] Stathakis D. How many hidden layers and nodes? // Int. J. Remote Sens. 2009. Vol. 30, № 8. P. 2133–2147.