Chaotic Time Series Prediction: Run for the Horizon

Vasilii A. Gromov[1]

[0000-0001-5891-6597]

[1] National Research University Higher School of Economics,
Pokrovskii boulevard, 11, 109028, Moscow, Russian Federation
stroller@rambler.ru

Abstract. The present article reviews some recent papers concerned with chaotic time series prediction in the context of predictive clustering, and discusses in greater detail some novel techniques designed to avoid 'a curse of exponential growth' – errors grow exponentially depending on the number of steps ahead to be predicted. These techniques are non-successive observations, combined with a prognosis that employs already predicted values, the concept of non-predictable points, and a quality assessment of clusters used. The approach discussed, allows one to separate calculation into two parts: the first part, essentially larger, is performed off-line, the second, immediate prediction routine, is carried out on-line. This makes it possible to design fast and efficient prediction algorithms. A wide-ranging simulation, suggests that the error term associated with the prediction sub-model used, provided that clusters used to predict are chosen correctly, vanishes as the validation set size grows to infinity. Similarly, the error term associated with an incorrect choice of clusters used to predict, decreases when a validation set size increases.

Keywords: Time Series Prediction, A Chaotic Time Series, Predictive Clustering, Cluster Prognostic Value.

1 Introduction

Constant interest in chaotic systems and models expressed by researchers in various fields [2, 8, 23-26, 28], is due to both the fundamental importance of non-linear phenomena for natural and social processes description, and an inherent complexity of their forecasting. The overwhelming majority of information systems are complex and thereby tend to show chaotic behaviour. Mathematically, the problem to forecast such system characteristics is a chaotic time series prediction problem. One should emphasize that regular and chaotic time series, essentially differ, in that the latter features the prediction horizon, which is the maximum number of steps ahead that one can make a prognosis. Quite naturally, this quantity depends on a required maximum prediction error and a time series observation accuracy. Since the prediction horizon is finite for chaotic time series, and infinite for regular, it serves to distinguish the time series of these two types. The prediction horizon is attributed to an exponential divergence of initially close trajectories, due to the Lyapunov instability of chaotic time series [16, 22]. The exponent coefficient is called the highest Lyapunov exponent; for chaotic time series it is positive, and it is readily calculated from a time series [16, 22].

The exponential divergence mentioned above, is also responsible for the exponential error growth for multi-step prediction, and the highest Lyapunov exponent serves here as the exponent coefficient also. This explains the fact that most papers dealing with chaotic time series prediction, discuss results for a single step prediction only, whereas the problem of multi-step prediction for chaotic time series is still unresolved.

It is worthy to stress that the exponential growth of prediction error is intrinsic to a single prediction model for a chaotic time series, whereas, if one predicts using a set of sub-models, the reverse may hold true. As it is very difficult to develop a single model to predict a chaotic time series, it is quite natural to look for prediction methods that are able to combine, either explicitly or implicitly, a set of sub-models corresponding to various dynamic patterns observed in the series [3, 23, 24]. Of the known approaches, predictive clustering [4] stands out as the most robust; here, the sub-models are based upon clustered sequences of time series observations.

The present article reviews some recent papers concerned with chaotic time series prediction, in the context of predictive clustering, and discusses in greater detail some novel techniques designed to avoid ‘a curse of exponential growth’. These are non-successive observations combined with a prognosis that employs already predicted values, the concept of non-predictable points [7], and quality assessment of clusters used [10]. Actually, any prediction method to a significant number of steps ahead for chaotic series involves prediction based on the values that are predicted themselves. Consequently, it is extremely important to assess whether or not these predicted values are reliable. The aforementioned concepts constitute a set of tools to estimate the reliability – hence they are necessary to make a reliable multi-step prognosis.

The rest of the paper is organised as follows. The next section reviews recent advances in the field. The following sections three and four, presents the mathematical statement, the problems under study, and introduce basic concepts and non-successive observations. Section 5 outlines the prediction algorithm. Sections 6, 7 and 8 go on to describe the concepts of non-predictable points and quality assessment. The following sections outline a clustering method, and a method to estimate clusters' prognostic values, and provide the prediction results for the time series which are 1) generated by the Lorenz system, and 2) associated with the Australian energy market. The prediction results are obtained for a single prediction for various quality assessment techniques that makes possible to compare them. The last two sections compare the results to those obtained by other authors, and present conclusions.

2 Related works

One way to decrease the mean prediction error for predictive clustering algorithms, is to estimate prognostic values of the clusters at hand, with the employment of an additional validation set, distinct from the training (used to generate these clusters) and the testing (used to estimate ultimate prediction error). One may treat these cluster prognostic values as method hyperparameters, as introduced by Goodfellow et al. [11], with the sole difference that, for that case, the number of hyperparameters is equal to the number of clusters, and therefore is large. It is possible to estimate the cluster prognostic value using, say, the mean prediction error (on the validation set) induced by this cluster, or an invariant measure for the phase space region associated with this cluster; the latter alternative allows excluding clusters corresponding to the remote and unfrequented regions.

It is not necessary to present information about clusters prognostic values using scalars, quite the contrary. It is possible to employ, for example, logical rules indicating practicability of utilizing the cluster in question, to predict.

In either case, the prediction error is broken down into two terms. The first is associated with an incorrect choice of a cluster and, consequently, a sub-model to predict. The second is that caused by a discrepancy between predicted and actual values, provided the cluster is chosen correctly, that is, the chosen cluster is associated with the true space phase region where the time series (trajectory) portion to be predicted is situated. The latter term cannot be reduced for the clustering algorithm used – we consider it as a kind of theoretical minimum error for a given clustering algorithm, prediction sub-model and training samples size. However, one can delete clusters with lower prognostic values in order to reduce the former. The totality of procedures aimed at reducing the second summand (at estimating prognostic values of clusters [hyperparameters]) is termed 'quality assessment' for predictive clustering.

The present paper introduces a 'quality assessment' of clusters generated by a predictive clustering algorithm, and proposes several methods to solve this problem. To compare different methods, we utilize a contribution of the first summand to the total prediction error, as well as the number of non-predictable observations for a testing

set [7, 8], that is the observations the algorithm is unable to predict due to the lack of the appropriate cluster. Let us stress, that the ability of any algorithm to detect non-predictable observations is its great advantage. Actually, it is much better if an algorithm 'honestly' indicates that it is unable to predict properly at a certain point and does not try to predict 'forcibly' - without indicating risk to use such predicted values.

Of fundamental importance is the ability of, and necessity for, a predictive clustering algorithm to generate clusters, using not only the series to be predicted but also a group of similar series that contains it.

Conventionally, predictive clustering researchers pursue two avenues of inquiry [1]. The first proposes that a time series is a single entity and a set thereof may be clustered using various clustering techniques. The second one looks for typical dynamical patterns (known as typical sequences [7-10], motifs [27], chunks [29], shapelets [30], subsequences [1], etc.) either in a time series observed, or in a group of similar time series. In what follows, we restrict our attention to the second line of investigation.

As discussed by E. Keogh and J. Lin [18], it serves no purpose to use the single time series to be predicted to generate clusters; it is essential to utilize a set of all other similar series. Papers concerned with the algorithms of pattern discovery, usually explore techniques to generate a training set, using a series at hand, and to cluster it. These parts of a predictive clustering algorithm are associated with concepts of data-adaptation and algorithm-adaptation [21]. The data-adaptation concept allows use of raw data, feature-based transformation of the data, and model-based transformation of the data as well, to generate samples [1, 21]. Algorithm-adaptation places the primary emphasis on clustering algorithms and their adaptation to the forecasting problem: A large part of previous studies deals with k-means, c-means (crisp and fuzzy) and the like.

Huang et al. [12] employ k-means in order to adjust it to seek for similar sections in chaotic time series; the modified algorithm is dubbed TSkmeans (*Time Series k-means*). Martinez-Alvarez and his colleagues [24] also uses k-means to predict chaotic time series; the paper summarizes results by various investigators for forecasting of Australia's national electricity market prices – this bunch of series seems to become a sort of benchmark to test various prediction algorithms for chaotic time series; an extended version of the results may be found in [7]. Papers [13, 14] analyse spatio-temporal data using a clustering technique grounded on the modified Euclidean distance capable of taking into account hidden space and time patterns. Benitez et al. [3] examine ways to extract typical patterns from series amassed by generating company; it is aimed at designing algorithms of rational energy consumption; the authors use various modifications of k-means.

The trouble with such algorithms is that, on the one hand, the structure of clusters depend heavily on the metric used and, on the other hand, it, for most cases, requires knowledge of the number of clusters before clustering [6]. The methods that employ concepts and methods of graph/complex network theory are free, in some sense, of these drawbacks. Ferreira and Zhao [6] propose to map time series sections into graph vertices in order to apply then community detection algorithms. Gromov and Borisenko [7] employ the modified Wishart algorithm to cluster sequences of observations [20]; the authors point out to correlation between clusters obtained and phase space regions with higher values of invariant measure of the respective dynamical systems.

3 Time series prediction problem

Given a set S of chaotic time series $S = \{y^{(s)}\} = \{y_0^{(s)}, y_1^{(s)}, \dots, y_{t_s}^{(s)}\}, s = 1..|S|$, where t_s is the size of the s series, $y_i^{(s)}$ is i-th observation of s-th series, and a series $y = \{y_t\}$, estimate the value of an observation y_{t+K} to minimise the prediction error

$$I = minE(y_{t+K} - \hat{y}_{t+K})^2. \quad (1)$$

It is supposed that we know all observations of y up to and inclusive y_t. In particular, if $K > 1$, then the problem is called the multi-step (ahead) prediction problem.

If $S = \emptyset$, then one obtains a more conventional definition of prediction problems. The definition (1) appears to be more convenient for predictive clustering as it allows utilizing information from various time series. Actually, any predictive clustering algorithm implies that one seeks motifs in the time series considered. A motif is a typical sequence that emerges from time to time in a series. We assume here that all transient processes in the system that generate the time series in hand have been completed, and the time series reflects the trajectory movement in the neighbourhood of the attractor of the dynamical system that generates the series. It is worth emphasizing that neither the system nor its attractor is known, and the problem to reconstruct them is usually a much more complicated problem than the prediction problem. For chaotic time series, an attractor is usually a complex geometrical (fractal) set, called strange attractor. The second assumption is that the series meets Takens theorem conditions, and respectively, one can analyse the attractor structure, using time series observations [16, 22].

As the trajectory of the system moves along the same area of the attractor frequently, one can meet similar sequences in the time series. These sequences resemble the motif associated with the respective area. If one reveals these areas, describes corresponding motifs, and develops the simplest prediction models for each one, one makes it possible to predict chaotic time series up to a considerable time limit [9]. The clustering method presented below is employed to collect together sequences belonging to the same cluster. The motifs are usually centres of such clusters. It is straightforward to extend this approach to a set of time series S, just using all motifs that can be found in them.

4 Non-successive observations

Usually, to ensure that Takens theorem conditions are satisfied, vectors are composed from time series observations (z-vectors) [16, 22]: a d-dimensional z-vector is defined as $z_i^{(s)} = (y_i^{(s)}, y_{i+1}^{(s)}, \dots, y_{i+d-1}^{(s)})$. Conventional practice is to compose z-vectors from successive observations. Surprisingly, z-vectors composed of non-successive observations according to a certain pattern, proved more efficient [7]. For the best prediction,

one should run over all or, at least, over a considerable portion of all reasonable patterns, and single out the most appropriate clusters. Different attractor areas are associated with different clusters and corresponding motifs. The pattern is defined as a preset sequence of distances between positions of observations, such that these (non-successive) observations are to be placed on the successive positions in a newly generated sample vector.

The vector, thus concatenated, generalises a conventional z-vector [16, 22], which corresponds to the pattern $(1,1,\dots,1)$ (m times). Thus, each pattern is a $S-1$-dimension integer vector $(p_1,\dots,p_{S-1})$, $p_j \in \{1,\dots,P_{max}\}, j = 1..S-1$; the parameter P_{max} dictates the maximum distance between positions of observations that become successive in the vector to be generated. Thereby, the quantity $S \cdot P_{max}$ refers to a kind of a memory depth.

For predictive clustering, samples selected from the vectors of concatenated successive observations (z-vectors), prove less efficient than those based on the vectors concatenated according to various patterns [7]. This is attributed to the fact that vectors of non-successive observations are able to store information about salient observations: minima, maxima, tipping points and so on.

One should emphasize that each model mentioned above is an averaged representation of the clustered time series sequences, or alternatively, trajectories belonging to the respective attractor area. Consequently, it leads to a decrease in the prediction error due to averaging (the predicted values are obtained by using the cluster centres), and simultaneously, to its increase, in virtue of the fact that the 'chaotic' exponential growth is alleviated. The clustering method used strikes a compromise between these two tendencies.

5 Prediction algorithm

A predictive clustering algorithm is usually subdivided into three parts. The first part analyses a group of time series at hand in order to cluster sequences made of its observations, according to predefined patterns, and then to use cluster centres as typical sequences. The second, estimates clusters' prognostic values and deletes clusters with low values. Finally, the third provides a prognosis for the time series with the employment of the obtained typical sequences (cluster centres).

The series are considered to be normalized. We used two different normalization techniques. The first one suggests that an entire time series is normalized with the employment of its maximum and minimum values, whereas the second technique implies that sample vectors are normalized separately, using their own maxima and minima. Hereafter, we refer to these techniques as global (G) and local (L) respectively. The latter makes it possible to cluster, not typical amplitudes (as it takes place for the former), but rather typical profiles.

To cluster generalised z-vectors, we employ the Wishart clustering method [31] as modified by A. V. Lapko and S. V. Chentsov [20]. This method employs graph theory concepts and a non-parametric probability density function estimator, of k-nearest neighbours. Some problems associated with application of the algorithm to predict time series are discussed in [7]. The algorithms to estimate clusters' prognostic values are discussed in the next section.

To predict time series values in the framework of the third part of an algorithm, the centres of clusters (motifs) are calculated for all used patterns and obtained clusters. For a given position to be predicted (for the time series in question), and for a given cluster, one should take the following steps. Firstly, one composes a vector from time series observations, according to the pattern used to generate the cluster, with the position associated with the last vector element (respectively, undefined). Secondly, truncate the vector and the cluster centre - all elements but the last ones are included in the truncated vectors. Thirdly, calculate the Euclidian distance between the truncated observation vector and the truncated cluster centre. One searches over all patterns and clusters in order to find the cluster with the minimum distance. If the distance is less than a predefined vigilance threshold, then the centre of this cluster is employed to predict the observation, namely, the last element of the centre is used as a predicted value for the position in question. Otherwise, if the distance to any cluster available exceeds the threshold, the dynamics are considered unidentifiable, and the observation is appended to the set of non-predictable observations.

6 Non-predictable points

Employing clustering techniques to reveal typical sequences and to predict time series using the revealed sequences, the predictive clustering methods are sometimes unable to find, for a given point to be predicted, any appropriate typical sequences to predict value at this position. This happens when there are no cluster centres matching observations from the time series section preceding this position. Hereafter such observations are called unpredictable, and their number (related to the testing set size) is taken to be a measure of prediction quality, along with a prediction error averaged over all other (predictable) observations of the testing set. It is worth stressing that this feature is conventionally regarded as a limitation of predictive clustering, but it seems that it is much better if an algorithm 'sincerely' warns that the point is unpredictable, than it generates an erroneous prediction without warning.

7 Quality assessment

The total prediction error can be broken down into the two terms. The first term results from the incorrect choice of the active cluster, that is a cluster which centre is used to predict. Consequently, it is possible to state the problem of *estimating clusters' prognostic values* in order to minimize the term associated with incorrect choice of the active cluster, that is, the cluster engaged to predict the current observation (the first term). The problem involves selecting a subset of clusters such that the total pre-

diction error (on the testing set) corresponding to the forecasting routine that employs this subset only, is either minimal (the first statement) or less than a predefined threshold (the second statement). Mathematically, the problem is formulated as follows. Let Λ is the set of clusters employed to predict the time series in question; $\Im \equiv \{G : \Lambda \to R^1\}$; $\tilde{\Lambda}(G,\beta) = \{\lambda \in \Lambda : G(\lambda) \geq \beta\}$. The problem is to find the estimator $G^* \in \Im$ and the threshold value $\beta^* \in R^1$, $\beta^* > 0$ (the first statement) in order to minimize prediction error (on the testing set):

$$\min I\left(\tilde{\Lambda}(G,\beta)\right) \tag{2}$$

The second statement implies that one minimizes the number of clusters belonging to $\tilde{\Lambda}(G,\beta)$:

$$\min\left|\tilde{\Lambda}(G,\beta)\right| \tag{3}$$

subject to constraint

$$\left|I(\tilde{\Lambda}(G,\beta))\right| \leq \gamma, \tag{4}$$

where γ is a parameter of the algorithm

In the framework of the first statement, one places emphasis on the minimum prediction error, while the second statement is concerned primarily with the speed to obtain prediction results. In either case, this suggests reducing the number of clusters, or to put it differently, the overall complexity of the prediction model under study (while maintaining prediction accuracy). One cannot but make analogies of various methods to reduce the complexity of regression models (for instance, AIC, BIC, GIC, and so on [19]).

To solve the problem, an additional set (the validation) is introduced, under the assumption that it differs from both the training and testing ones, and all three of them are drawn from the same universal set.

8 The problem of estimating clusters' prognostic values (quality assessment)

Two techniques to estimate the values in question are considered. The first one suggests that the prognostic value of k-th cluster is calculated as follows:

$$Q_k(\beta) = \sum_{i \in S_k} \frac{\bar{e}_i}{e_{ik}} \frac{1}{|V_i|}, \quad (5)$$

where $\bar{e}_i = \frac{1}{|V_i|} \sum_{i \in V_i} e_{ij}$, V_i is a set of clusters able to predict i-th observation with an error less than β; S_k is the set of observations predicted by k-th cluster with an error less than β; e_{ij} is a prediction error for i-th observation if j-th cluster is used to predict.

The second method to perform quality assessment, offers not to use a single characteristic, but rather to extract knowledge from data about prediction errors for observations of the validation set.

Namely, we define for k-th cluster (over the validation set):

d_{ij} is the minimum Euclidian distance between i-th observation and elements of k-th cluster;

$S_i^{(d)}(\beta) = \left\{ y_i^{(s)} : d_{ij} \le \beta \right\}$ is the number of observations with the distance less than β from j-th cluster;

m_j is the number of times the cluster has been active;

n_j is the number of times the use of the cluster would lead to the minimum possible error.

Algorithm 2. The quality assessment routine with the replacement of the active cluster.

1. Initialization: For each $S_i^{(d)}(\beta) \neq \emptyset$, $S_i(\beta) \neq \emptyset$: $m_j \leftarrow 0$, $n_j \leftarrow 0$, $i \leftarrow 0$, $j \leftarrow 0$.
2. If $d_{ij} \le \beta$ then $S_i^{(d)}(\beta) = S_i^{(d)}(\beta) \cup y_i$.
3. If $e_{ij} \le \beta$ then $S_i(\beta) = S_i(\beta) \cup y_i$.
4. Find $d_{i\,min} = d_{ik} = \min\limits_j d_{ij}$; $m_k \leftarrow m_k + 1$.
5. Find $e_{i\,min} = e_{ip} = \min\limits_j e_{ij}$ and the distance of d_{ip}; $\mathrm{n_k} \leftarrow \mathrm{n_k} + 1$..
6. $j \leftarrow j + 1$. If the list of clusters is not exhausted, then go to step 3.
7. $\mathrm{i} \leftarrow \mathrm{i} + 1$. If the list of observations is not exhausted, then go to step 2.

In what follows, we refer to these algorithms as to 1st and 2nd.

9 Numerical results

The aforementioned clustering algorithm is applied to generate samples with the employment of all possible patterns of four elements with the maximum (minimum) distance between neighbouring positions in the pattern equal to 10. So the number of patterns used amounts to 10000. Each sample produces its own set of clusters, and then all sets of clusters are merged into a single set.

The method discussed in the previous section, is applied to a time series generated by the Lorenz system, to a set of noisy Lorenz series, and to a set of Australia's national electricity market price series too. Throughout the paper, we stick to single-step prediction. The highest Lyapunov exponent was calculated for all studied time series, with the employment of the analogue method [16, 22].

To measure prediction error, we used three measures. They are the root mean square error (*RMSE*), the mean average error (*MAE*), and the percentage of non-predictable observations. All three measures are averaged over the testing set, which is used neither for training nor for quality assessment.

The results obtained are presented in a uniform way for any series analysed. Namely, after introductory information about the series, we present prediction errors for different method versions in the form of a table. The table shows prediction errors corresponding to various choices of normalization, clustering, and quality assessment routines. The first column indicates a size of the validation set (the size of training set is usually the same); the next two columns present information about the method used.

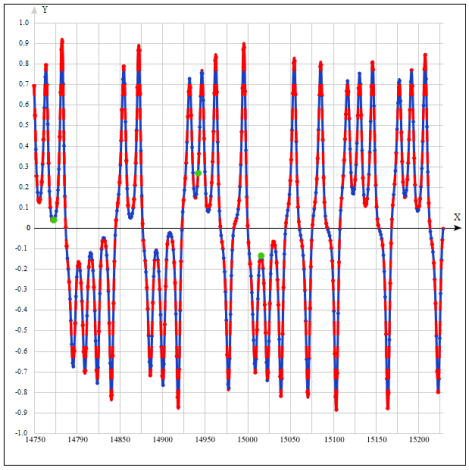

Fig. 1. Single-step prediction for Lorenz time series. Blue solid lines are associated with observed data, whereas red dashed lines are associated with predicted values. Green discs represent non-predictable points.

Namely, the second and third columns correspond to a normalization technique (G is global and L is local), and a method to estimate clusters' predictive values (quality assessment; 1 is the quality assessment method based upon a scalar estimate of clusters' prognostic value; 2 is the one based upon a replacement of the active cluster). The next three columns present RMSE, MAE, and the percentage of non-predictable observations. Finally, the last

two columns display (for comparison) MAE and RMSE for the case, when the true active cluster is known in advance (‘theoretical minimum’).

Table 1. Prediction errors for the Lorenz series

Size	*N*	*QA*	*RMSE* (*10^{-2})	*MAE* (*10^{-2})	*Non* (%)	*MMAE* (*10^{-2})	*MRMSE* (*10^{-2})
10^4	G	1	1.82	1.2	0.73	0.358	0.367
10^5	G	1	1.023	0. 89	0.61	0.358	0.364
10^6	G	1	0.89	0. 81	0.59	0.358	0.361
10^7	G	1	0.83	0. 79	0.52	0.358	0.359
10^4	G	2	1.45	1	0.74	0.358	0.367
10^5	G	2	1.027	0. 78	0.64	0.358	0.364
10^6	G	2	0.87	0. 73	0.6	0.358	0.361
10^7	G	2	0.78	0. 72	0.57	0.358	0.359
10^4	L	1	0.96	0.73	0.43	0. 207	0.229
10^5	L	1	0.79	0.69	0.34	0. 207	0.227
10^6	L	1	0.64	0.52	0.31	0. 207	0.224
10^7	L	1	0.48	0.48	0.3	0. 207	0.221
10^4	L	2	0.84	0.72	0.38	0. 207	0.229
10^5	L	2	0.78	0.63	0.36	0. 207	0.227
10^6	L	2	0.53	0.51	0.29	0. 207	0.224
10^7	L	2	0.46	0.45	0.26	0. 207	0.221

Size is the size of a training set; N is a normalization technique; QA is a quality assessment algorithm; RMSE is a root-mean-square error; MAE is a mean absolute error; Non is the percentage of non-predictable observations (for the testing set); MMAE is theoretically minimal mean absolute error; MRMSE is theoretically minimal root-mean-square error.

The method under study was applied to the series generated by the Lorenz system [15, 22]. The Lorenz system with standard ‘chaotic’ parameters $\sigma = 10, b = \frac{8}{3}, r = 28$ integrated with the employment of Runge–Kutta’s fourth-order method (integration step is equal to 0.05), yields a time series hereafter referred to as the Lorenz series.

The series in question, on the one hand, is a typical chaotic series (the highest Lyapunov equals to 0.92 that is in agreement with results by Malinetskii and Potapov [22] [see on p. 217]) and, on the other hand, is a conventional benchmark to test forecasting procedures for chaotic time series.

For the Lorenz series, the first 3000 observations are discarded in order to ensure that trajectory moves in the neighbourhood of the respective strange attractor. The

testing set for the series consists of 100000 observations, the training set consists of 100000, while a validation set size is varied and, actually, are crucial parameters for the method considered.

Figure 1 presents single-step prediction results for the Lorenz time series. The first figure displays a typical time series section (of the testing set) and the respective predicted values; blue solid lines are associated with observed data, whereas red dashed lines are associated with predicted values. Green circles represent non-predictable observations.

The size of the training set is 100000 observations, that of the validation set is 10^7. The percentage of non-predictable observations is 0.26%, RMSE is about 0.46%, while the average prediction error for predictable observations is equal to 0.0045 %. Table 1 shows prediction errors.

The Wishart clustering technique, in conjunction with a local normalization routine and the quality assessment method based upon a scalar estimate of clusters' prognostic values, proves the most efficient; however, it also proves the most time-consuming. Another point of interest is the fact that the percentage of the clusters to be discarded to obtain the best prediction, converges to a certain limit (around 19%) as the size of the validation set increases.

To explore the potential to use clustering results obtained for a certain group of series in order to predict distinct but similar series, we consider a set of noisy Lorenz series. The training set is generated with the employment of the standard Lorenz series (see above) of 100000 observations, while the validation and testing is generated using noisy series. To generate these series, we add the white noise to a normalized standard Lorenz series and then normalize again. The noise amplitude is a normal random variable with a mean equal to 0.0 and a variance equal to 0.3. The series prove chaotic with the highest Lyapunov varying from 0.98 to 1.23. The size of the training set is 100000; the size of the testing set is 100000.

Table 2. Prediction error for a noisy series

Size	*N*	*QA*	*RMSE* ($*10^{-2}$)	*MAE* ($*10^{-2}$)	*Non* (%)	*MMAE* ($*10^{-2}$)	*MRMSE* ($*10^{-2}$)
10^4	G	1	21.82	16.37	18.69	4.05	4.21
10^5	G	1	16.83	14.38	18.51	4.05	4.15
10^6	G	1	13.89	9.32	15.13	4.05	4.12
10^7	G	1	11.63	7.56	14.69	4.05	4.09
10^4	G	2	17.45	15.29	16.54	4.05	4.21
10^5	G	2	12.64	13.26	15.34	4.05	4.15
10^6	G	2	11.87	8.74	14.97	4.05	4.12
10^7	G	2	10.87	6.87	13.78	4.05	4.09

10^4	L	1	23.48	15.98	15.31	3.89	4.19
10^5	L	1	18.64	15.33	14.11	3.89	4.11
10^6	L	1	15.17	13.69	13.68	3.89	4.01
10^7	L	1	12.77	12.83	12.97	3.89	3.94
10^4	L	2	19.89	15.67	14.84	3.89	4.19
10^5	L	2	18.21	14.88	13.72	3.89	4.11
10^6	L	2	14.63	13.15	13.03	3.89	4.01
10^7	L	2	12.35	12.19	12.56	3.89	3.94

The abbreviations are the same as for Table 1.

For that case, the best combination of techniques appears to be that of Wishart clustering and the quality assessment by replacement of the active cluster. The optimal percentage of clusters to be deleted for the quality assessment routine based upon a scalar estimate, in contrast to the previous case, does not converge to a fixed value. This may be attributed to the fact that the training and the validation sets are of a different nature (usual and noisy Lorenz series).

Finally, the method under study is applied to time series generated by electricity prices in various settlements of the Commonwealth of Australia.

Table 3. Prediction error for Australia's national electricity market price

Size	*N*	*QA*	*RMSE* (*$*10^{-2}$)	*MAE* (*$*10^{-2}$)	*Non* (%)	*MMAE* (*$*10^{-2}$)	*MRMSE* (*$*10^{-2}$)
10^4	G	1	0.98	0.701	0.35	0.449	0.462
10^5	G	1	0.83	0.662	0.29	0.449	0.460
10^6	G	1	0.76	0.627	0.25	0.449	0.456
10^7	G	1	0.73	0.617	0.17	0.449	0.451
10^4	G	2	0.87	0.674	0.37	0.449	0.462
10^5	G	2	0.78	0.631	0.34	0.449	0.460
10^6	G	2	0.74	0.623	0.29	0.449	0.456
10^7	G	2	0.67	0.608	0.23	0.449	0.451
10^4	L	1	0.76	0. 587	0.29	0.287	0.314
10^5	L	1	0.72	0. 518	0.25	0.287	0.307
10^6	L	1	0.68	0. 509	0.19	0.287	0.301
10^7	L	1	0.66	0. 503	0.14	0.287	0.296
10^4	L	2	0.74	0. 521	0.27	0.287	0.314
10^5	L	2	0.72	0. 514	0.26	0.287	0.307
10^6	L	2	0.65	0.507	0.17	0.287	0.301
10^7	L	2	0.51	0. 492	0.15	0.287	0.296

The abbreviations are the same as for Table 1.

10 Comparison with published results

Tables 4 and 5 exhibit results obtained by various methods; the tables are partially borrowed from [24]; see also [7]. Let us stress, that prediction error for algorithms proposed is lower than that of conventional soft-computing algorithms, provided the points classified as non-predictable by the algorithm are excluded (their percentage is usually lower than 1%), and is comparable with it, if these non-predictable observations are predicted forcibly.

Table 4. MER for some days of the year 2004 (Australia's national electricity market – Price)

Day	5th June	17th June	20th June	21th June	Average
ARIMA(%)	32.31	29.09	33.73	24.18	29.82
SVM(%)	18.09	13.31	17.11	19.2	16.93
PSF(%)	16.72	8.31	14.23	18.93	14.55
PCW(%)	1.94	1.72	1.32	1.94	1.73(0.42%, 1.78)
PCW(1)(%)	0.87	0.78	0.64	0.84	0.74 (0.18%, 0.77)
PCW(2)(%)	0.76	0.72	0.58	0.83	0.69 (0.24%, 0,71)

ARIMA – the best ARIMA model; SVM – support vector machine; PSF – pattern sequenced-based forecasting; PCW – predictive clustering using the Wishart algorithm [31]; PCW(1) – predictive clustering using the Wishart algorithm with quality assessment based upon clusters' prognostic values; PCW(2) – predictive clustering using Wishart algorithm with quality assessment based upon active cluster replacement; last column in parentheses is a percentage of non-predictable observations and the error calculated provided the non-predictable observations are predicted forcibly.

Table 5. MER for some weeks of the year 2004 (Australia's national electricity market – Price)

Week	First of January	First of July	First of August	Third of December	Average
DWT(%)	12.94	12.2	16.17	10.01	12.84
SVM(%)	23.37	15.0	36.18	33.74	27.08
PSF(%)	15.62	9.12	13.98	10.23	12.23
PCW (%)	1.33	1.47	1.28	1.11	1.30 (0,38%, 1,34)
PCW(1)(%)	0.96	0.78	0.83	0.62	0.76 (0,21%, 0,79)
PCW(2)(%)	0.89	0.81	0.74	0.59	0.72 (0,19%, 0,74)

The abbreviations are the same as for Table 4.

11 Conclusions

1. Predictions that uses already predicted values, the concept of non-predictable points, and quality assessment of clusters employed, taken together, direct the way to solution of the multi-step chaotic time series prediction problem.

2. Quality assessment procedure aimed at estimating clusters' prognostic values and deleting clusters with low ones (in the framework of predictive clustering) decreases essentially predictive error both for benchmark and for real-word data.

3. A wide-ranging simulation suggests that the error term associated with prediction sub-model used (provided that clusters used to predict are chosen correctly) vanishes as a validation set size tends to infinity. Similarly, the error term associated with incorrect choice of clusters used to predict decreases when a validation set size increases.

4. Prediction error for algorithms proposed is lower than that of conventional soft-computing algorithms, provided the points classified as non-predictable by the algorithm are excluded (their percentage is usually lower than 1%), and is comparable with it, if these non-predictable observations are predicted forcibly. The best variant is Wishart clustering algorithm in conjunction with local normalization and replacement of the active cluster.

5. The approach discussed allows one to separate calculation into two parts: the first, essentially larger, is performed off-line, the second, immediate prediction routine, is performed on-line. This makes possible to design fast and efficient prediction algorithms.

Acknowledgements

The author is deeply indebted to Mr. Joel Cumberland, HSE for the manuscript proofreading and language editing.

References

1. Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.Y.: Time-series clustering – A decade review. Information Systems 23, 16–38 (2015).
2. Al Zoubi, O., Awad, M., Kasabov, N.K.: Anytime multipurpose emotion recognition from EEG data using a Liquid State Machine based framework. Artificial Intelligence in Medicine 86, 1–8 (2018).
3. Benítez, I., Díezb, J.L., Quijanoa, A., Delgado, I.: Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance. Electric Power Systems Research 140, 517–26 (2016).

4. Blockeel, H., De Raedt, L., Ramon, J.: Top-down induction of clustering trees. In: 15-th international conference on machine learning, pp. 55–63. (1998).
5. D'Urso, P., De Giovanni, L., Massari, R.: GARCH-based robust clustering of time series. Fuzzy Sets and Systems 305, 1–28 (2016).
6. Ferreira, L.N., Zhao, L.: Time series clustering via community detection in networks. Information Sciences 326, 227–42 (2016).
7. Gromov, V.A., Borisenko, E.A.: Chaotic time series prediction and clustering methods. Neural Computing and Appl 2, 307–15 (2015).
8. Gromov, V.A., Konev, A.S.: Precocious identification of popular topics on Twitter with the employment of predictive clustering. Neural Computing and Appl 28(11), 3317–22 (2017).
9. Gromov, V.A., Shulga, A.N.: Chaotic time series prediction with employment of ant colony optimization. Expert Systems with Appl 39(9), 8474–8 (2012).
10. Gromov, V.A., Voronin, I.M., Gatylo, V.R., Prokopalo, E.T.: Active Cluster Replacement Algorithm as a Tool to Assess Bifurcation Early-warning Signs for von Karman equations. Artificial Intelligence Research 6(2): 51–6 (2017).
11. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge, MA (2015).
12. Huang, X., Ye, Y., Xiong, L., Lau, R.Y.K., Jiang, N., Wang, S.: Time Series k-Means: A New k-Means Type Smooth Subspace Clustering for Time Series Data. Information Sciences 367–368, 1–13 (2016).
13. Izakian, H., Pedrycz, W.: Agreement-based fuzzy c-means for clustering data with blocks of features. Neurocomputing 127, 266–80 (2014).
14. Izakian, H., Pedrycz, W., Jamal, I.: Clustering spatiotemporal data: an augmented fuzzy c-means. IEEE Trans Fuzzy Syst 21(5), 855–68 (2013).
15. Jackson, E.A.: The Lorenz System: I. The Global Structure of its Stable Manifolds. Phys Scr 32, 469–75 (1985).
16. Kantz, H., Schneider, T.: Nonlinear Time Series Analysis. Cambridge University Press, Cambridge (2004).
17. Kattan, A., Fatima, S., Arif, M.: Time-series event-based prediction: An unsupervised learning framework based on genetic programming. Information Sciences 301, 99–123 (2015).
18. Keogh, E., Lin, J.: Clustering of time-series subsequences is meaningless: implications for previous and future research. Knowledge and information systems 8(2), 154–77 (2005).
19. Konishi, S., Kitagava, G.: Information Criteria and Statistical Modeling. Springer, N.-Y. (2008).
20. Lapko, A.V., Chentsov, S.V.: Nonparametric information processing systems. Science, Novosibirsk (2000).
21. Liao, T.W.: Clustering of time series data-a survey. Pattern Recogn 38(11), 1857–74 (2005).
22. Malinetskii, G.G., Potapov, A.P.: Modern problems of non-linear dynamics. Editorial URSS, Moscow (2002).
23. Martınez-Alvarez, F., Troncoso, A., Riquelme, J.C.: Data Science and Big Data in Energy Forecasting. Energies 11, 3224 (2018).
24. Martınez-Alvarez, F., Troncoso, A., Riquelme, J.C., Riquelme, J.M.: Energy time series forecasting based on pattern sequence similarity. IEEE Trans Knowl Data 23(8), 1230–43 (2011).
25. Obodan, N.I., Adlucky, V.J., Gromov, V.A.: Prediction and Control of Buckling: The Inverse Bifurcation Problems for von Karman Equations. In: Dutta, H., Peters, J.F.

(Eds),Applied Mathematical Analysis: Theory, Methods, and Applications Studies in Systems, Decision and Control, vol. 177, pp. 353–81. Springer, N.-Y. (2019).
26. Obodan, N.I., Adlucky, V.J., Gromov, V.A.: Rapid identification of pre-buckling states: A case of cylindrical shell. Thin-Walled Struct 124, 449–57 (2018).
27. Palit, A.K., Popovich, D.: Computational intelligence in time series forecasting. Theory and engineering applications. Springer, N.-Y. (2005).
28. Pérez-Chacón, R., Luna-Romera, J.M., Troncoso, A., Martínez-Álvarez, F., Riquelme, J.C.: Big Data Analytics for Discovering Electricity Consumption Patterns in Smart Cities. Energies 11(3), 683 (2018).
29. Phu, L., Anh, D.T.: Motif-Based Method for Initialization the *K*-Means Clustering for Time Series Data. In: Wang, D., Reynolds, M. (eds) AI 2011: Advances in Artificial Intelligence. AI 2011. Lecture Notes in Computer Science, vol 7106, pp. 11-20. Springer, N.-Y. (2011)
30. Widiputra, H., Kho, H., Pears, R., Kasabov, N.K.: A novel evolving clustering algorithm with polynomial regression for chaotic time-series prediction. Neural Inf Process 5864, 114–21 (2009).
31. Wishart, D.: A numerical classification methods for deriving natural classes. Nature 221, 97–8 (1969).
32. Zakaria, J., Mueen, A., Keogh, E.: Clustering time series using unsupervised shapelets. In: 12-th International Conference on Data Mining: IEEE Computer Society. pp. 785–94 (2012).