ISSN 0032-9460, Problems of Information Transmission, 2016, Vol. 52, No. 1, pp. 92–99. © Pleiades Publishing, Inc., 2016.
Original Russian Text © A.S. Lysyak, B.Ya. Ryabko, 2016, published in Problemy Peredachi Informatsii, 2016, Vol. 52, No. 1, pp. 101–109.

SOURCE CODING

Time Series Prediction Based on Data Compression Methods[1]

A. S. Lysyak[a] and B. Ya. Ryabko[b]

[a] *Novosibirsk State University, Novosibirsk, Russia*
e-mail: accemt@gmail.com
[b] *Novosibirsk State University, Novosibirsk, Russia*
Institute of Computational Technologies, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
e-mail: boris@ryabko.net

Received March 19, 2015; in final form, December 19, 2015

Abstract—We propose efficient ("fast" and low memory consuming) algorithms for universal-coding-based prediction methods for real-valued time series. Previously, for such methods it was only proved that the prediction error is asymptotically minimal, and implementation complexity issues have not been considered at all. The provided experimental results demonstrate high precision of the proposed methods.

DOI: 10.1134/S0032946016010075

1. INTRODUCTION

Time series prediction is of considerable practical interest, since it allows to solve prediction tasks for various phenomena in science, engineering, and economics. Apparently, that is why this problem attracts wide attention of researchers, and presently there exits many prediction methods, the most well known among which are prediction based on bilinear models [1], autoregression analysis of various types [2–5], Monte-Carlo based prediction [6], and methods based on constructing expert evaluation (so-called recursive strategies, whose description can be found in [7,8]). However, though numerous methods and approaches are known, the problem of increasing the accuracy of time series prediction is still actual.

In the present paper we continue the development of prediction methods based on using universal codes (or "data compression methods"). For the first time, the possibility of using universal codes for prediction was described in 1988 in [9], where random processes generating values from some finite set (alphabet) were considered. In [10] it was shown how universal codes can be applied to predict time series taking values in some continuous numerical interval, but no theoretical analysis of the proposed method was given at that time. Later, in [11], there were proposed asymptotically the most precise methods for predicting time series that take values in continuous intervals, but they were not implemented in practice, and there were no knowledge about (preasymptotic) prediction accuracy and complexity.

In the present paper we describe algorithms realizing the asymptotically optimal methods from [11] and experimentally investigate their accuracy on real-world time series. Comparison of the proposed method with known ones shows that its prediction accuracy is rather high.

[1] Supported in part by the Russian Foundation for Basic Research, project no. 15-07-01851.

2. UNIVERSAL CODES AND UNIVERSAL MEASURES AS A BASIS FOR PREDICTION METHODS

2.1. Prediction Problem

We first describe the prediction problem. Let us be given a stationary ergodic source generating a sequence of element $x_1x_2\ldots$ of some set A, referred to as an alphabet. The alphabet can be either finite or countably infinite, or it can be a real line segment. The prediction problem consists in estimating the probability distribution for the random variable x_{t+1} based on the values $x_1, x_2, \ldots, x_t$. In the case of a discrete alphabet, we will estimate conditional probabilities $P(x_{t+1} = a \in A \,|\, x_1x_2\ldots x_t)$. For a continuous alphabet, we only consider the case where there exist conditional probability densities $p(x_{t+1} \,|\, x_1x_2\ldots x_t)$ and the prediction problem reduces to constructing estimators for them (all other characteristics of interest used in prediction—mean value, variance of the process, etc.—are easily computed from the density). In particular (following many authors), we will estimate the mean value of the process x_{t+1} (for fixed $x_1\ldots x_t$) by computing it from the density estimation.

Let us briefly expose the relation between universal codes and so-called universal measures on one hand, and the problem of predicting stationary processes on the other. Let Ω be some set of stationary ergodic processes generating elements of an alphabet A. A predictor is a function γ defined on all words $x_1\ldots x_{t+1}$, $t \geq 0$, which for convenience will be denoted by $\gamma(x_{t+1} \,|\, x_1\ldots x_t)$ similarly to the conditional probability; we require the natural conditions $\sum\limits_{a\in A} \gamma(a \,|\, x_1\ldots x_t) = 1$ and $\gamma(x_{t+1} \,|\, x_1\ldots x_t) \geq 0$ to be satisfied for all $x_1\ldots x_{t+1} \in A^{t+1}$, $t \geq 0$.

Every predictor naturally defines a measure $\gamma(x_1\ldots x_t)$ on the set of words of length t:

$$\gamma(x_1\ldots x_t) = \gamma(x_1)\gamma(x_2 \,|\, x_1)\ldots\gamma(x_t \,|\, x_1\ldots x_{t-1}). \tag{1}$$

These measures are naturally matched for different t and define a probability distribution on the set of infinite sequences. Note that the converse is also true: every probability distribution π on infinite sequences over an alphabet A defines a predictor

$$\pi(x_t \,|\, x_1\ldots x_{t-1}) = \pi(x_1\ldots x_t)/\pi(x_1\ldots x_{t-1}). \tag{2}$$

As an example, consider the prediction method due to Laplace. He suggested to estimate the conditional probabilities $p(x_{t+1} = a \in A \,|\, x_1x_2\ldots, x_t)$ from known values $x_1\ldots x_t$ by the following predictor L:

$$L(a \,|\, x_1\ldots x_t) = (\nu_{x_1\ldots x_t}(a) + 1)/(t + |A|),$$

where $\nu_{x_1\ldots x_t}(a)$ is the number of occurrences of a symbol a in the word $x_1\ldots x_t$. Thus, for instance, $L(0\,|\,01010) = 4/7$ for $A = \{0, 1\}$.

A natural question is evaluating the prediction quality, or accuracy. One of the accuracy measures widely used in information theory and mathematical statistics is the Kullback–Leibler (KL) divergence. In our case the divergence between the probability distribution $P(x_{t+1} = a \,|\, x_1\ldots x_t)$ and a predictor $\gamma(a \,|\, x_1\ldots x_t)$ for fixed $x_1\ldots x_t$ is given by

$$\mathrm{KL}(P,\gamma)_{x_1\ldots x_t} = \sum_{a\in A} P(x_{t+1} = a \,|\, x_1\ldots x_t)\log(\gamma(a \,|\, x_1\ldots x_t)/P(x_{t+1} = a \,|\, x_1\ldots x_t) \tag{3}$$

(hereinafter, $\log x \equiv \log_2 x$), and the mean value (for a fixed series length t), by

$$\mathrm{KL}(P,\gamma)_t = \sum_{x_1\ldots x_t\in A^t} P(x_1\ldots x_t)\,\mathrm{KL}(P,\gamma)_{x_1\ldots x_t}. \tag{4}$$

Note that $\mathrm{KL}(P,\gamma)_{x_1\ldots x_t}$ is nonnegative, and equals zero if and only if $\gamma(a \,|\, x_1\ldots x_t)$ and $P(x_{t+1} = a \,|\, x_1\ldots x_t)$ coincide for all $a, x_1, \ldots, x_t$ (see [12]). In [9] it is shown that for any source P

generating independent and identically distributed symbols of A, the error of Laplace's predictor satisfies the inequality

$$\mathrm{KL}(P, L)_t \le \log e(|A| - 1)/(t + 1)$$

(here $e = 2.718\ldots$ is Euler's number). We see that the error of Laplace's predictor tends to zero for any source generating i.i.d. symbols. Unfortunately, there is no predictor possessing this property for any stationary ergodic source (for a proof, see [9]). However, for such sources there exist predictors for which a weaker property is satisfied: the Cesàro mean of errors (4) tends to zero. More precisely, there exists a predictor γ such that for any stationary ergodic source ω generating symbols of some finite alphabet A we have

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \mathrm{KL}(\omega, \gamma)_s = 0. \tag{5}$$

A predictor γ is said to be universal for a set of sources Ω if equality (5) holds for any $\omega \in \Omega$. The quantity $\frac{1}{t} \sum_{s=1}^{t} \mathrm{KL}(\omega, \gamma)_s$ will be denoted by $\overline{\mathrm{KL}}(\omega, \gamma)_t$ and called the average error of predictor γ on source ω. Form this definition, (3), and (4), one can easily obtain the equality

$$\overline{\mathrm{KL}}(\omega, \gamma)_t = \sum_{x_1\ldots x_t \in A^t} \omega(x_1 \ldots x_t) \log(\gamma(x_1 \ldots x_t)/\omega(x_1 \ldots x_t)), \tag{6}$$

where $\gamma(x_1 \ldots x_t)$ is defined in (1). Form this and (5), the meaning of source universality becomes clear: for any source $\omega \in \Omega$ the value of $\gamma(x_1 \ldots x_t)$ approaches $\omega(x_1 \ldots x_t)$.

2.2. Universal Codes and Measures

The notion of a universal measure relates the problems of prediction and universal coding. It is defined as follows: let a set Ω of stationary ergodic sources be given. A measure μ is said to be universal if for any source $P \in \Omega$ the equality

$$\lim_{t\to\infty} \frac{1}{t} \sum_{u\in A^t} P(u) \log(P(u)/\mu(u)) = 0$$

is valid. This equality shows that a universal measure μ is in a sense a nonparametric estimator for the unknown probability distribution P. By comparing the last equality with (1)–(6) it is seen that the universal measure and universal predictor are closely related and in fact coincide.

Universal measures and predictors are related with universal codes. Let us describe this interrelation, since it makes it possible to use for prediction the so-called archivers, i.e., programs designed for real text compression. Here it is important to note that modern archivers exploit for "compression" deviations in occurrence frequencies of different symbols and subwords, hidden periodicities, and a number of other regularities. This property of them is undoubtedly valuable from practical point of view, and one of the goals of the present paper is to show how a really working archiver can be used to construct a predictor, and also in the case where values of a random process are real numbers.

A detailed description of a nondistorting ("reversible") code can be found, e.g., in [12]; here we briefly note that a code is a mapping from words of length t over an alphabet A (i.e., A^t, $t \ge 1$) to a set of distinct words over the alphabet $\{0, 1\}$. A code U is said to be universal if for any stationary ergodic source P we have

$$\lim_{t\to\infty} E_P(|U(x_1 \ldots x_t)|)/t = H(P),$$

where $E_P(f)$ is the mean value of f with respect to the measure P, and $H(P)$ is the Shannon entropy of P, i.e.,

$$H(P) = \lim_{t\to\infty} -t^{-1} \sum_{u\in A^t} P(u)\log P(u).$$

Note that the entropy is an asymptotically tight lower bound on the average length of a nondistorting code; that is why such codes are called universal.

The following simple assertion (see, e.g., [9]) states that based on any universal code one can construct a universal measure.

Proposition. *Let U be a universal code for some set of sources Ω generating symbols of an alphabet A, and let a measure μ_U for each word v over A be given by*

$$\mu_U(v) = 2^{-|U(v)|} \Big/ \sum_{u\in A^{|v|}} 2^{-|U(u)|}. \tag{7}$$

Then μ_U is a universal measure for Ω.

To construct predictions for real-life processes, one should choose a particular universal measure (or fist a universal code, and then compute a measure from it according to (7)). For our prediction, we use the universal measure R (see [13]). The choice of this particular measure is due to the fact that it is constructed on the basis of a universal code whose redundancy is asymptotically minimal for the classes of Bernoulli and Markov sources [13].

To describe it, we first present a predictor, found in 1968, for which the error (6) is asymptotically minimal for the set of all Bernoulli sources [14, 15]. This predictor, which makes it possible to compute conditional probabilities for the next element of a series, is given by

$$K_0(a\,|\,x_1\dots x_t) = (\nu_{x_1\dots x_t}(a) + 1/2)/(t + |A|/2), \tag{8}$$

where $\nu_{x_1\dots x_t}(a)$ is the number of occurrences of an element a in a word $x_1\dots x_t$. It is interesting to note that the error for this predictor is asymptotically half as large as for Laplace's predictor.

Based on this predictor, we construct a measure K_0, which, as was first shown in [14], is universal for the class of Bernoulli sources:

$$K_0(x_1\dots x_t) = \prod_{i=0}^{t-1} \frac{\nu_{x_1\dots x_i}(x_{i+1}) + 1/2}{i + |A|/2}.$$

For example, $K_0(01010) = \frac{1/2}{1}\frac{1/2}{2}\frac{3/2}{3}\frac{3/2}{4}\frac{5/2}{5}$ for $A = \{0, 1\}$.

For Markov sources, an analogous measure is as follows [15]:

$$K_m(x_1\dots x_t) = \begin{cases} \dfrac{1}{|A|^t}, & t \le m, \\ \dfrac{1}{|A|^m} \displaystyle\prod_{i=m}^{t-1} \frac{\nu_{x_1\dots x_i}(x_{i+1-m}\dots x_{i+1}) + 1/2}{\nu_{x_1\dots x_{i-1}}(x_{i+1-m}\dots x_i) + |A|/2}, & t > m, \end{cases} \tag{9}$$

where $\nu_x(\vartheta)$ is the number of occurrences of a sequence ϑ in x. For example, $K_1(01010) = \frac{1}{2}\frac{1/2}{1}\frac{1/2}{1}\frac{3/2}{2}\frac{3/2}{2}$ for $A = \{0, 1\}$. This measure is universal for the set of Markov sources of order m (see [15]).

The measure R universal for the set of all stationary and ergodic sources is defined as follows:

$$R(x_1\dots x_t) = \sum_{i=0}^{\infty} \omega_{i+1} K_i(x_1\dots x_t), \tag{10}$$

where

$$\omega_i = 1/\log(i+1) - 1/\log(i+2). \tag{11}$$

2.3. Prediction Scheme for Time Series Generating Real Values

Prediction of time series that generate values from a finite alphabet is performed by the formula (see [9])

$$R(x_{t+1} = a \,|\, x_1 \dots x_t) = R(ax_1 \dots x_t)/R(x_1 \dots x_t).$$

In [11] it is shown how this method can be transferred to the case of series generating real values to obtain an asymptotically optimal predictor. Let us give necessary facts on this method.

Consider a time series generating a sequence x_t each element of which takes values in an interval $[A, B]$. Let $\{\Pi_n\}$, $n \geq 1$, be an increasing sequence of finite partitions of $[A, B]$ such that the maximum subinterval length in the partitions tends to zero (we refer to this process as quantization). Also, define $x^{[k]}$ to be the element of Π_k containing a point x.

In what follows, we consider only processes for which all multivariate densities exist. Denote by $p(x_1 x_2 \dots x_n)$ the probability density of the process with respect to the Lebesgue measure L.

Now define an estimator r for the probability density as follows:

$$r(x_1 \dots x_t) = \sum_{s=1}^{\infty} \omega_s \Big(R(x_1^{[s]} \dots x_t^{[s]})/L(x_1^{[s]} \dots x_t^{[s]}) \Big). \tag{12}$$

The factors ω_s are given by equation (11) and play the role of weight coefficients for partitions from Π_k. As is seen from (12), when computing the measure r each summand is normalized with respect to the Lebesgue measure L. Thus, we interconnect estimators for the probability densities for different increasing partitions, thus avoiding the dependence of a prediction result on a particular partition.

As is shown in [11], $r(x_1 \dots x_t)$ is an estimator for an unknown probability density $p(x_1 \dots x_t)$, the corresponding conditional density

$$r(a \,|\, x_1 \dots x_t) = r(x_1 \dots x_t a)/r(x_1 \dots x_t) \tag{13}$$

is an estimator for the density $p(a \,|\, x_1 \dots x_t)$, and both estimators are in a certain sense consistent [11].

3. DESCRIPTION OF THE ALGORITHM AND ITS COMPLEXITY EVALUATION

In [11] it is shown what can be used for predicting arbitrary sequences of finite partitions that define step functions estimating the density. We have experimentally found out that partitioning into equal subintervals gives, as a rule, the best prediction accuracy for real-life data; therefore, in experiments described below we use precisely this partition. As a preditive value of x_{t+1}, we take the mean value computed by the conditional density estimator (13).

Let us evaluate the complexity of the described prediction algorithm; by the complexity we mean the number of operations. The number of operation required to compute the density estimator over n subintervals is determined by the computation complexity for the measure R (see (10)) in an n-symbol alphabet, which in turn depends on the computation complexity for K_m (see (9) and (10)). It is easily seen that the computation complexity for expression (9) is $O(tn^{t+1})$. Hence we obtain that the computation complexity for the predicted value is $O(t^3 n^{t+2})$. A considerable reduction of computation complexity has been obtained due to the fact that for a large number n of subintervals (say $n > t$) many occurrence frequencies of the subintervals (ν in (9)) coincide, which allows to use the method of grouping alphabet symbols described in [16] to reduce the complexity. In this case the complexity reduction cannot be described analytically, since this quantity, in general, depends on values taken by the considered time series; however, experiments show that computation time becomes 3–5 times as small for series lengths from several tens to two thousand.

Table 1. Forecasting of fuel prices in USA. The R-method for 1 and 20 step(s) ahead

Partition size	R-method 1 step	R-method 20 steps
5	0.07260	0.44139
10	0.04658	0.57510
20	0.05202	0.29762
50	0.03915	0.12638

Thus, on one hand, the number n of partition subintervals determines the prediction error (which, obviously, cannot be less that half the subinterval length). On the other hand, the algorithm complexity substantially depends on n. By experiments we have found that $n = \lfloor \log_2 t \rfloor + 5$ is a reasonable trade-off, because large values of n have almost no effect on the prediction accuracy but considerably increase the computation time. Therefore, in out computations we used the following formula instead of (12):

$$r(x_1 \dots x_t) = \sum_{s=1}^{\lfloor \log_2 t \rfloor + 4} \omega_s R(x_1^{[s]} \dots x_t^{[s]}) / L(x_1^{[s]} \dots x_t^{[s]}) + \frac{1}{\lfloor \log_2 t \rfloor + 6} R(x_1^{[\lfloor \log_2 t \rfloor + 5]} \dots x_t^{[\lfloor \log_2 t \rfloor + 5]}) / L(x_1^{[\lfloor \log_2 t \rfloor + 5]} \dots x_t^{[\lfloor \log_2 t \rfloor + 5]}). \quad (14)$$

4. EXPERIMENTAL PREDICTION RESULTS

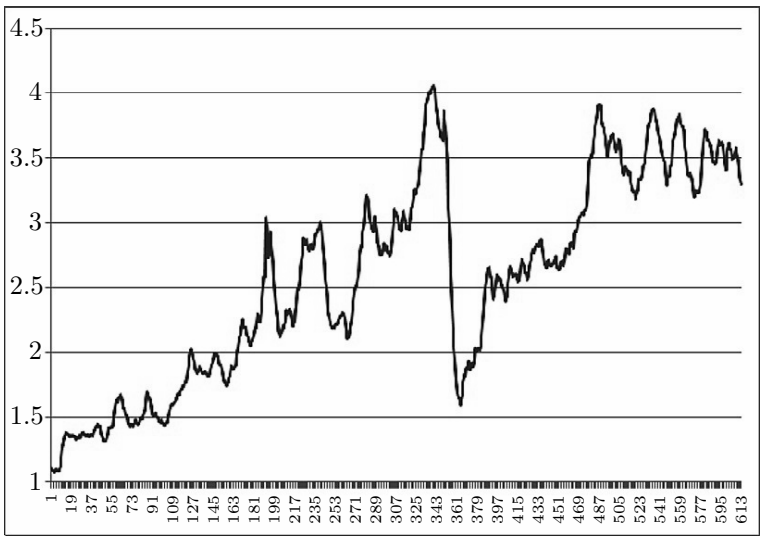
Description of the experimental analysis method. For a concrete series $x_1 \dots x_t$, from its segment $x_1 \dots x_{t-10}$ we computed the predicted value up to time $t-9$ (denote it by y_{t-9}), then from $x_1 \dots x_{t-9}$ we computed the predicted value y_{t-8} for $t-8$, etc.; from $x_1 \dots x_{t-1}$ we computed the predicted value y_t. After that we computed the average prediction error $\Big(\sum\limits_{i=1}^{10} |x_i - y_i| \Big)/10$, which was used for comparison of methods.

In forecasting, approaches are known in which data preprocessing is first performed (usually, this is filtering, smoothing, etc.), then the obtained series is used for prediction itself, and after that an inverse transformation is applied to predicted values. In examples considered below, we in some cases transformed the initial series $x_1 \dots x_t$ to a series $x_1^* \dots x_{t-1}^*$ by formula $x_i^* = x_{i+1} - x_i$ (in forecasting, this approach is usually applied to remove near-linear trends).

As an example, consider the problem of USA gasoline price forecasting (see [17]). Table 1 presents the results of predicting this series with interval of one week in the period from January 1, 2002, to October 1, 2013. The series length is 615 elements, the computation depth was taken to be 5, and the series interval length is 0.832. To clarify the scale of the varied quantity, we give the value of Δ, the maximum difference between two consecutive elements of the predicted series. The series is plotted in the figure. On the horizontal axis of the plot there are numbers of the time series elements in ascending order.

Prediction of this series was performed for the last ten points (over all the preceding), and the prediction accuracy was computed. The average error (over ten points) is 0.03915, whereas the difference between consecutive elements of the series comes up to 1. Thus, the prediction error is less than 4% of this value.

To compare the accuracy of the proposed method with previously known ones, we used data on USA economic time series, for which forecasting results by methods of the International Institute of Forecasters (IIF) are known. We took four time series from [17]: industry (industrial production index), finance (1) and finance (2) (USA financial activity index), and demographic (USA demographic index). These time series were taken in the following time periods: the Industry series



Plot of fuel prices (interval of 1 week).

Table 2. Predicting USA economic time series. The R-method, Autobox, ForecastPro, and PP-Autocast methods. Average prediction error

Time series	Sample size	Δ	R-Method	Autobox	ForecastPro	PP-Autocast
Industry	144	6050	706.52	340.72	301.86	303.64
Finance (1)	144	1550	164.48	680.49	794.42	793.03
Finance (2)	132	118	21.07	76.12	71.98	41.40
Demographic	134	2642	53.46	122.08	152.71	286.19

from Jan. 1982 to Jan. 1994, Finance (1) from Jan. 1962 to Jan. 1974, Finance (2) from Jan. 1965 to Jan. 1976, and Demographic from Jan. 1983 to Jan. 1994.

The experiments consisted in predicting one step ahead eighteen last elements of the presented time series. As competitors for the R-method, we took the following three most well-known methods whose results are given at the IIF website: AutoBox, ForecastPro, and PP-Autocast. The computation depth in all the considered cases was taken to be 3. The results are presented in Fig. 2.

It is seen from the presented data that the R-method in the cases of Finance (1), Finance (2), and Demographic series gives considerably better results as compared to the other known methods. As the performed experimental results have shown, the average prediction error for the R-method is about half as large as for the known methods.

5. CONCLUSION

The resented experimental results demonstrate high precision of the method based on the universal measure. The algorithm using the alphabet grouping method considerable reduces its complexity, which allows to use this method on standard computers for series lengths of several thousand. The high precision of the proposed method is justified by its comparison with other known methods. It is also important to note that the proposed method can easily be generalized to the case of predicting multivariate time series.

REFERENCES

1. Poskitt, D.S. and Tremayne, A.R., The Selection and Use of Linear and Bilinear Time Series Models, *Int. J. Forecasting*, 1986, vol. 2, no. 1, pp. 101–114.

2. Tong, H., *Non-linear Time Series: A Dynamical System Approach*, Oxford, UK: Clarendon, 1990.

3. Tong, H., *Threshold Models in Non-linear Time Series Analysis*, Lect. Notes Statist., vol. 21, Berlin: Springer, 1983.

4. Tong, H. and Lim, K.S., Threshold Autoregression, Limit Cycles and Cyclical Data, *J. Roy. Statist. Soc. Ser. B*, 1980, vol. 42, no. 3. 245–292.

5. Engle, R.F., Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, *Econometrica*, 1982, vol. 50, pp. 987–1007.

6. Bontempi, G., Local Learning Techniques for Modeling, Prediction and Control, *PhD Thesis*, IRIDIA, Université Libre de Bruxelles, Belgium, 1999.

7. Zhang, G., Patuwo, B.E., and Hu, M.Y., Forecasting with Artifcial Neural Networks: The State of the Art, *Int. J. Forecasting*, 1998, vol. 14, no. 1, pp. 35–62.

8. Cheng, H., Tan, P.-N., Gao, J., and Scripps, J., Multistep-Ahead Time Series Prediction, *Advances in Knowledge Discovery and Data Mining (Proc. 10th Pacific-Asia Conf. PAKDD'2006, Singapore, Apr. 9–12, 2006)*, Ng, W.K., Kitsuregawa, M., Li, J., and Chang, K., Eds., Lect. Notes Comp. Sci., vol. 3918, Berlin: Springer, 2006, pp. 765–774.

9. Ryabko, B.Ya., Prediction of Random Sequences and Universal Coding, *Probl. Peredachi Inf.*, 1988, vol. 24, no. 2, pp. 3–14 [*Probl. Inf. Trans.* (Engl. Transl.), 1988, vol. 24, no. 2, pp. 87–96].

10. Ryabko, B.Ya. and Monarev, V.A., Experimental Investigation of Forecasting Methods Based on Data Compression Algorithms, *Probl. Peredachi Inf.*, 2005, vol. 41, no. 1, pp. 74–78 [*Probl. Inf. Trans.* (Engl. Transl.), 2005, vol. 41, no. 1, pp. 65–69].

11. Ryabko, B., Compression-Based Methods for Nonparametric Prediction and Estimation of Some Characteristics of Time Series, *IEEE Trans. Inform. Theory*, 2009, vol. 55, no. 9, pp. 4309–4315.

12. Cover, T.M. and Thomas, J.A., *Elements of Information Theory*, Hoboken, NJ: Wiley, 2006, 2nd ed.

13. Ryabko, B.Ya., Twice-Universal Coding, *Probl. Peredachi Inf.*, 1984, vol. 20, no. 3, pp. 24–28 [*Probl. Inf. Trans.* (Engl. Transl.), 1984, vol. 20, no. 3, pp. 173–177].

14. Krichevskii, R.E., The Relation Between Redundancy Coding and Reliability of Information from a Source, *Probl. Peredachi Inf.*, 1968, vol. 4, no. 3, pp. 48–57 [*Probl. Inf. Trans.* (Engl. Transl.), 1968, vol. 4, no. 3, pp. 37–45].

15. Krichevsky, R., *Universal Compression and Retrieval*, Dordrecht: Kluwer, 1993.

16. Ryabko, B.Y., Astola, J., and Gammerman, A., Adaptive Coding and Prediction of Sources with Large and Infinite Alphabets, *IEEE Trans. Inform. Theory*, 2008, vol. 54, no. 8, pp. 3808–3813.

17. Gasoline and Diesel Fuel Update, Independent Statistics & Analysis, U.S. Energy Information Administration, `http://www.eia.gov/petroleum/gasdiesel/`.