ISSN 0005-1179, Automation and Remote Control, 2021, Vol. 82, No. 6, pp. 1049–1064. © Pleiades Publishing, Ltd., 2021.
Russian Text © The Author(s), 2021, published in Avtomatika i Telemekhanika, 2021, No. 6, pp. 149–168.

Randomized Machine Learning of Nonlinear Models with Application to Forecasting the Development of an Epidemic Process

A. Yu. Popkov[1*]

[1]*Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, 119333 Russia*

*e-mail: *apopkov@isa.ru*

Received October 15, 2020; revised January 12, 2021; accepted January 15, 2021

Abstract—We develop a discrete approach in the theory of randomized machine learning that is aimed at application to nonlinear models. We formulate the problem of entropy estimation of probability distributions and measurement noise for discrete nonlinear models. Issues related to the application of such models to forecasting problems, in particular, the problem of generating entropy-optimal distributions, are considered. The proposed methods are demonstrated on the solution of the problem of forecasting the total number of persons infected with novel coronavirus SARS-CoV-2 in Germany in 2020.

DOI: 10.1134/S0005117921060060

1. INTRODUCTION

A large number of various processes observed in various areas of human activity, nature, etc. cannot be effectively described by linear mathematical models. In this regard, the development of general approaches to nonlinear modeling is an urgent problem. However, it should be noted that developing and applying nonlinear models to specific problems encounters certain difficulties associated both with their training using real data and with the choice of the model structure.

Machine learning as a branch of applied science incorporates a large number of methods and approaches accumulated in various scientific disciplines [1, 2], a large contribution to which was made in such areas as probability theory and mathematical statistics [3, 4]. Machine learning methods are successfully applied to various problems, in particular, to classification and regression problems, which pertain to supervised learning problems [1], the main feature of which is the idea of tuning (training) the parameters of the required model using real data. The tuned (trained) model is supposed to be used for forecasting, i.e., producing a response after being presented with input data that have not been previously involved in training.

This approach is efficient and has been known since at least the 1960s [5–8]. However, most of the approaches developed in this area are focused on the use of linear models. In particular, one efficient approach to solving classification problems is linear classification, which consists in finding a linear dividing hyperplane and is used, for example, in the widely used Support Vector Machine method [9]. Regression problems have long been widely used in econometrics [3], and in this area most of them are also focused on linear regression. The main reasons for this consist mainly in the fact that, first, linear models are easier to investigate and interpret, second, numerical and analytical solutions of linear problems can be obtained either absolutely accurately (analytically) or numerically with high accuracy, and third, many practical problems can often be reduced to linear statements, and consequently, a better solution can be obtained taking into account the properties noted earlier.

At the same time, various nonlinear effects occur in some applied problems of classification and regression. These effects must be somehow resolved. An efficient approach to this problem is, for example, the kernel approach, which consists in a nonlinear transition to a high-dimensional space with the subsequent application of a linear method in this space to solve the classification or regression problem [1]. This approach proves efficient in many applied problems and has led to the emergence of various "kernel" versions of the known linear methods. Nevertheless, the issue of choosing a model (more precisely, a kernel function) continues to arise when applying these methods in practice. In addition, any transition to a high-dimensional space in the absence of a large amount of data in this space inevitably leads to undesirable effects such as, for example, overtraining, as well as a number of others.

Another approach to combating nonlinear effects is offered by methods that do not explicitly isolate the model, for example, methods based on decision trees [10], neural network models [1, 11, 12], etc.

Thus, on the one hand, we have a situation in which the presence of a "nonlinearity" that requires the use of nonlinear models is clearly observed in the practical problems of data analysis, and on the other hand, the existing methods often used in practice are not sufficiently developed to efficiently solve the problem in the nonlinear case and require reformulation or some adaptation of the problem for them to the applied.

In the present paper, we propose a universal approach to working with nonlinear models in data analysis problems, in particular, in problems of training regression models. This approach is based on the theory of randomized machine learning [13], the main idea of which is to artificially randomize the model parameters; this makes it possible to move from a model with deterministic parameters to a model with random parameters and determine their distributions rather than their point estimates as a result of training. The distributions are defined so as to maximize the entropy functional under the condition of balance with the average output of the model.

The main advantage of the proposed approach is independence of the real characteristics of the data used. Correct application of the method does not require confirmation or assumptions about the normality of the data (or their other probabilistic properties), and the distributions obtained as a result of training are calculated under conditions of maximum entropy, thus reflecting the "worst-case" scenario for the development of the research process, corresponding to its maximum uncertainty. These properties of the entropy estimation method go back to the works of Boltzmann [14], Jaynes [15, 16], and Shannon [17]. Another important feature of the method is obtaining the entropy-optimal distributions of the noise contained in the data together with the optimal distributions of the parameters. This property significantly distinguishes the method from the classical approaches, in which various assumptions are made about the characteristics of noise and data.

Model characteristics are used to predict the process being modeled. The standard forecasting approach is to evaluate a model with point estimates of parameters obtained by estimation procedure on real data for unknown ("future") data points [18–20].

Taking into account the fact that it is impossible to pick a model that exactly matches the data, an assumption is made about the stochastic nature of the data, more precisely, the assumption that the (unknown) mechanism of generating the observed data contains a stochastic component. The consequence of this assumption is that the observed data contains both a deterministic and a stochastic component with unknown probabilistic characteristics. In fact, the entire mathematical apparatus developed to date and used in this area is aimed at modeling this stochastic component. An essential part of this approach is the assumption of the normality of random data components. This assumption allows establishing the properties of the resulting estimates of the model parameters.

For example, the well-known and widely used maximum likelihood method is based on the idea of maximizing the distribution of model parameters given the observed data: it is necessary to determine the parameter values that deliver a maximum to the likelihood function of the data. This means that the probability that the existing data will be observed with the found parameter values is maximal. The assumption about the normality (or about the presence of another known distribution law) of random data components is the basis of this method; without this assumption, it will be impossible to obtain the likelihood function in closed form [3].

The least squares method is often used to calculate estimates maximizing the likelihood, but the estimates it produces correspond to maximum likelihood estimates only for linear models. In the nonlinear case, determining the properties of these estimates is associated with significant difficulties.

The assumption about the normality of the stochastic components of the models and hence of the data with which the model is associated is obviously not always completely correct, since it is not always possible to find out or prove the required facts about the probabilistic characteristics of the data based on the data available.

Entropy-optimal distributions obtained at the stage of model training can be used for forecasting in several ways; in particular, they can be generated to obtain an ensemble of model outputs with its subsequent analysis.

The present paper is devoted to developing a method of entropy-randomized learning and forecasting for nonlinear models with discrete parameters. The transition from continuous to discrete models makes it possible to overcome the difficulties of using continuous models under conditions of a large number of variables, leading to the problem of calculating multidimensional integrals, which cannot be solved exactly (analytically) and faces considerable computational difficulties when solved numerically.

The approach proposed in this paper is demonstrated by an example of the problem of forecasting the total number of people infected as a result of the development of the COVID-19 epidemic in Germany. The proposed approach is compared with the nonlinear least squares method [18, 21].

2. NONLINEAR DISCRETE RANDOMIZED MODEL

Consider a plant with n scalar inputs x_i, $i = 1, \dots, n$, and an output $\hat{y}$ whose transformation is described in the general case by a nonlinear function

$$\hat{y} = \Phi(\mathbf{x}, \mathbf{a}), \tag{2.1}$$

where $\mathbf{x} = (x_1, \dots, x_n)$ is the vector of inputs and $\mathbf{a} = (a_1, \dots, a_d)$ is the vector of model parameters.

The model output is measured with some noise ξ that acts additively upon the output, leading to a model of the form

$$v = \hat{y} + \xi = \Phi(\mathbf{x}, \mathbf{a}) + \xi. \tag{2.2}$$

Suppose that the values of each parameter are concentrated within the interval $\mathcal{A}_k = [a_k^-, a_k^+]$, $k = 1, \dots, d$, and the model output is calculated with noise ξ_j whose values are concentrated in the interval $\Xi_j = [\xi_j^-, \xi_j^+]$ for each given input $\mathbf{x}_j$, $j = 1, \dots, m$.

The parameters a_k are realized by a discrete random variable with M values on the interval $\mathcal{A}_k$, leading to the distributions

$$a_{k\ell} \in A_k, \quad p_{k\ell} \in [0, 1], \quad k = 1, \dots, d, \quad \ell = 1, \dots, m, \tag{2.3}$$

where the $a_{k\ell}$ are the values of the random variable and $p_{k\ell}$ are the probabilities of their realization.

The output measurement noise is realized by a discrete random variable ξ_j with L values on the interval Ξ_j for each input x_j. The output measurements are taken independently of each other, thus leading to the following distributions for m measurements:

$$\xi_{jh} \in \Xi_j, \quad q_{jh} \in [0,1], \quad j = 1,\dots,m, \quad h = 1,\dots,L, \tag{2.4}$$

where the ξ_{jh} are the values of the random variable and q_{jh} are the probabilities of their realization.

Considering m measurements, we obtain the ultimate form of the model (2.2),

$$\mathbf{v} = \hat{\mathbf{y}} + \boldsymbol{\xi} = \Phi(\mathbf{x}_j, \mathbf{a}) + \boldsymbol{\xi}, \quad j = 1,\dots,m, \tag{2.5}$$

where $\mathbf{v} = (v_1,\dots,v_m)$ is the vector of the measured model output, $\boldsymbol{\xi} = (\xi_1,\dots,\xi_m)$ is the noise vector, and $\hat{\mathbf{y}} = (y_1,\dots,y_m)$ is the model output vector.

The distribution of parameters and noise of model measurements is to be estimated using real data on the measurements of the output of the plant whose model is described by (2.5).

3. MODEL TRAINING USING REAL DATA

Randomized machine learning is based on the principles of entropy estimation of the model parameters and the output measurement noises. Entropy-optimal distributions reflect the most uncertain situation, which, in the absence of information about the real characteristics, is the only solution available under these conditions [13, 22, 23].

To calculate the optimal distributions, we need to solve the conditional maximization problem for the entropy of the parameter and measurement noise distributions under the conditions of normalization of the corresponding distributions and the validity of conditions on the balance of the average output of the model with the measured output of the plant. The problem is stated as follows:

$$H(P,Q) = -\sum_{k=1}^{d}\sum_{\ell=1}^{M} p_{k\ell}\ln p_{k\ell} - \sum_{j=1}^{m}\sum_{h=1}^{L} q_{jh}\ln q_{jh} \to \max_{P,Q}, \tag{3.1}$$

where P and Q are the parameter and noise distributions (2.3) and (2.4), under the conditions

$$\sum_{\ell=1}^{M} p_{k\ell} = 1, \quad \sum_{h=1}^{L} q_{jh} = 1, \quad k = 1,\dots,d, \quad j = 1,\dots,m, \tag{3.2}$$

$$\mathbb{E}[\mathbf{v}] = \mathbb{E}\big[\Phi(\mathbf{x}_j, \mathbf{a}) + \boldsymbol{\xi}\big] = \mathbf{y}, \tag{3.3}$$

where $\mathbf{y} = (y_1,\dots,y_m)$ is the plant output measurement vector (the real output data).

Condition (3.3) determines the balance between the average model output and the real output data,

$$\begin{aligned}\mathbb{E}[\mathbf{v}_j] &= \mathbb{E}\big[\Phi(\mathbf{x}_j, \mathbf{a}) + \xi_j\big] = \mathbb{E}\big[\Phi(\mathbf{x}_j, \mathbf{a})\big] + \mathbb{E}[\xi_j] \\ &= \sum_{\substack{\ell_k=1 \\ k=1,\dots,d}}^{M} \Phi(\mathbf{x}_j, a_{1\ell_1},\dots,a_{d\ell_d}) p_{1\ell_1}\cdots p_{d\ell_d} + \sum_{h=1}^{L}\xi_{jh}q_{jh} = \bar{\Phi}(\mathbf{x}_j) + \sum_{h=1}^{L}\xi_{jh}q_{jh} = y_j.\end{aligned} \tag{3.4}$$

The sum in the expression for $\bar{\Phi}$ contains M^d terms that are summed over all combinations of values of the random variables $a_{k\ell}$. The solution of problem (3.1)–(3.3), which is considered in detail in the Appendix, yields entropy-optimal parameter and measurement noise distributions, which is the ultimate goal of training the model using real data.

4. RANDOMIZED FORECASTING

As a result of training, the model is equipped with entropy-optimal estimates of the parameter and measurement noise distributions, thus forming a *randomized predictive model*. This model defines a special forecasting technique—*randomized forecasting*,—the elements of which were used for some applied problems [24–26].

Randomized forecasting is based on the generation of the entropy-optimal parameter distribution (A.4) and measurement noise distribution (A.5) with the subsequent construction of a model output ensemble for new model inputs, unknown during training.

Consider the set of randomized predictive model inputs for which we need to construct a forecast that can be represented as a block vector or a matrix whose columns are the indicated inputs,

$$X = \{\mathbf{x}_1, \dots, \mathbf{x}_s\} = \begin{bmatrix} x_{11} & \dots & x_{1s} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{ns} \end{bmatrix}.$$

Suppose that we have a sample of parameters from the distribution P of volume S. Then the model output ensemble for one input $\mathbf{x}$ is formed according to (2.1) and has the form

$$\hat{\mathcal{Y}} = \left\{\hat{y}_i = \Phi(\mathbf{x}, \mathbf{a}_i)\right\}, \quad i = 1, \dots, S,$$

where $\mathbf{a}_i$ is a realization of parameters with the distribution P. The ensemble contains S trajectories.

Now for each input $\mathbf{x}_j$, $j = 1, \dots, s$ and each realization $\mathbf{a}_i$, $i = 1, \dots, S$, of parameters we consider a sample of noise from the distribution q of volume U and form the final model output ensemble according to (2.2),

$$\mathcal{V} = \left\{\mathbf{v}_j = \Phi(\mathbf{x}_j, \mathbf{a}_i) + \boldsymbol{\xi}_j\right\}, \quad i = 1, \dots, S, \quad j = 1, \dots, s,$$

where $\boldsymbol{\xi}_j = (\xi_{j1}, \dots, \xi_{jU})$ is the vector of noise realizations for the jth input and $\mathbf{v}_j$ is the measured output vector of the model for the jth input.

Thus, when forecasting, U realizations of noise are generated for each input and each realization of the model parameters. As a result, the ensemble $\mathcal{V}$ consists of $W = SU$ trajectories, which together can be represented by a block vector or a matrix with rows corresponding to the predicted output of the model for each input,

$$\mathcal{V} = [\mathbf{v}_1, \dots, \mathbf{v}_W] = \begin{bmatrix} v_{11} & \dots & v_{1s} \\ \vdots & \ddots & \vdots \\ v_{W1} & \dots & v_{Ws} \end{bmatrix}.$$

The mean and median trajectories, the standard deviation area, and other sample probabilistic characteristics can be calculated to construct the final predicted trajectory of the modeled process by the ensemble $\mathcal{V}$.

As can be seen from the expressions of the ensembles, it is necessary to have a noise distribution for each input to form them. The distributions obtained during training cannot be directly used for an arbitrary number of predictive inputs, since they are obtained from the data known at this stage, but the amount and characteristics of the data during forecasting are not known in advance. There can be several ways out of this situation.

The first is to apply the distribution defined by the expression (A.5) for the mean value of the parameter λ (Lagrange multipliers) as the predicted noise distribution q.

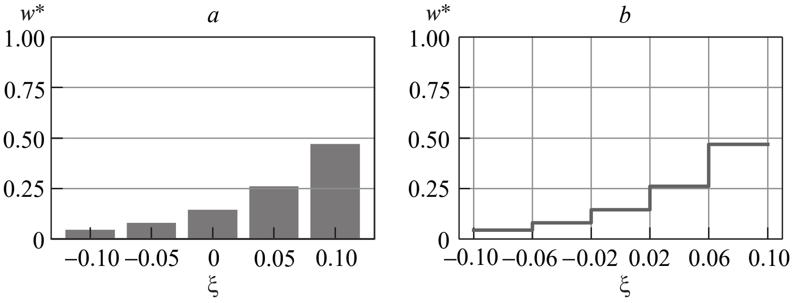

Fig. 1. Discrete and piecewise constant continuous distribution.

The second approach is to use the distribution for one of the inputs used in training, for example, the last one, as a predictive noise if the problem statement allows the order of inputs in the set. The idea of this method is as follows: if we proceed from the fact that measurements at each point of sequentially located data are taken by one "device," then it is logical to expect some stabilization of the characteristics of this measuring device, which is achieved by the moment of the last measurement in the sequence.

The third approach is based on the assumption that, as a result of entropy estimation of both parameters and noise, one can consider the pure output of the model without noise. Thus, we can say that entropy estimation performs filtering. In this case, the use of the model should be carried out in a pure form, without noise.

An important problem in the application of the entropy-randomized approach to forecasting is the generation of optimal distributions of the parameters and measurement noise obtained during the estimation (training) of the model. To solve this problem, two main approaches can be proposed using an evenly distributed random number generator.

The first approach is to generate a discrete distribution. To this end, a standard approach is used, which consists in randomly choosing the value of a random variable and then the corresponding probability. In this case, obviously, realizations made in this way will represent the set of values of the corresponding random variable, many of which will be repeated.

The second approach is based on the idea of representing the discrete distribution as a piecewise constant approximation of some continuous distribution on the corresponding interval. To this end, the interval of values of the corresponding random variable is divided into $L+1$ subintervals (where L is the number of values of the discrete random variable); the left and right boundaries of the finite subintervals correspond to the left and right boundaries of the distribution interval. Generation occurs evenly within each subinterval. As a result of the realization of this approach, it is possible to obtain a considerably larger number of different values of the corresponding random variables. An example of constructing continuous distributions for the random variable $w^*(\xi)$, where $\xi \in [-0.1, 0.1]$ and $L = 5$, is shown in Fig. 1.

5. FORECASTING THE CUMULATIVE NUMBER OF PEOPLE INFECTED WITH COVID-19 IN GERMANY

The approach proposed in this paper was used to model the dynamics of the COVID-19 epidemic in Germany based on data from Johns Hopkins University [27] starting from the fortieth day of the epidemic (March 8, 2020), when the total number of infected people exceeded 1000 people for the first time.

The data on the development of the epidemic (see Fig. 2) indicate that at first the infection is actively spreading in the population and its exponential growth is observed. Further, there is a decrease in the number of infected people, probably due to restrictive measures or to an increase

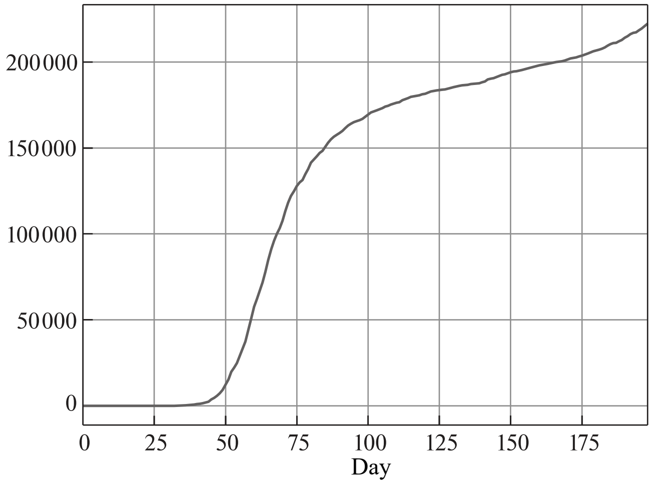

Fig. 2. Total number of infected persons in Germany on a daily basis starting from March 8, 2020.

in the number of immune members of the population. At the same time, as in most living systems, when examining them over a relatively short period of time, it can be assumed that the population size does not change (for example, migration, reproduction, and mortality can be neglected), which means that there is a limit on the number of infected persons. The dynamics of infected members of the population N in such a system can be described by the equation [28]

$$\frac{dN}{dt} = \lambda N \left(1 - \frac{N}{K}\right), \tag{5.1}$$

where λ is the growth rate of infected people, N is the number of infected people, and K is the population size. A solution of this equation is given by the Verhulst curve [29, 30]

$$N(t) = \frac{K}{1 + Be^{-\lambda t}}, \quad B = \frac{K - N_0}{N_0}, \tag{5.2}$$

where N_0 if the number of infected persons in the population at the initial time [30].

The model of the form (5.2) was actively used in early 2020s to predict the total number of cases [31–37] and has shown its efficiency at the initial stage of the epidemic. In this regard, it seems reasonable to use a similar model to apply the entropy-randomized approach to forecasting the total number of infected people. For such a model, we will use the three-parameter logistic growth model (LGM), which defines the transformation of the scalar input x into the output $\hat{y}$ using the logistic nonlinear function

$$\hat{y} = \Phi(x, \mathbf{a}) = \frac{a_3}{1 + a_1 e^{-a_2 x}}, \tag{5.3}$$

where $\mathbf{a} = (a_1, a_2, a_3)$ is the vector of model parameters. This model is a generalization of the model (5.2) and is considered here as an abstract model with parameters without using additional links between them.

In the context of the problem under consideration, the input is the ordinal number (or index) of the day, and the output is the accumulated (total) number of infected persons. The input and output are integers, but integers are converted to floating point numbers in calculations.

The randomized model whose output is distorted by an additive noise and the parameters and noise are realized by discrete random variables with values in appropriate intervals has the form

$$v = \hat{y} + \xi = \Phi(x, \mathbf{a}) + \xi,$$
$$a_{k\ell} \in A_k, \quad p_{k\ell} \in [0,1], \quad k = 1, \ldots, d, \quad \ell = 1, \ldots, m,$$
$$\xi_{jh} \in \Xi_j, \quad q_{jh} \in [0,1], \quad j = 1, \ldots, m, \quad h = 1, \ldots, L,$$

where $a_{k\ell}$ and ξ_{jh} are the values of the random variables realizing the parameters and noise, $p_{k\ell}$ and q_{jh} are the probabilities of their realization, m is the number of data points, and $d = 3$.

For training (estimating) the predictive model, we used data for several days (the training interval $\mathcal{T}_{\text{est}}$) starting from March 8, 2020 (the fortieth day of the epidemic in Germany), when the total number of infected persons over 1000 persons was recorded for the first time.

Forecasting is performed on the days following the estimation interval up to the 120th day of the epidemic (the forecasting interval $\mathcal{T}_{\text{pred}}$).

The resulting randomized predictions were compared with the curve fitting according to model (5.3) using the nonlinear least squares method implemented by the `curve_fit` function of the `scipy` library on the Python 3.7 platform.

After estimating the model, its implementation was carried out on the estimation (testing) interval and on the forecasting interval with the calculation of the following performance metrics for the true values of y and the predicted (model) values of $\hat{y}$:

1. The determination coefficient R^2 defined by the formula

$$R^2(y, \hat{y}) = 1 - \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2},$$

where $\bar{y}$ is the mean value based on real data, allows one to estimate the goodness of fit (GoF) and the predictive capability of the model in terms of the share of explained variance. The maximum of this indicator is 1; the closer its value is to one, the higher the performance of the model.

2. The mean square error (MSE) defined by the formula

$$\text{MSE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n-1}(y_i - \hat{y}_i)^2$$

shows the expected (mean) quadratic error.

3. The norm error (NE) defined by the formula

$$\text{NE}(y, \hat{y}) = \frac{\|y - \hat{y}\|}{\|y\| + \|\hat{y}\|} = \frac{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum\limits_{i=1}^{n} y_i^2 + \sum\limits_{i=1}^{n} \hat{y}_i^2}.$$

4. The rooted norm error (RNE) defined by the formula

$$\text{RNE}(y, \hat{y}) = \frac{\sqrt{\|y - \hat{y}\|}}{\sqrt{\|y\|} + \sqrt{\|\hat{y}\|}} = \frac{\sqrt{\sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{\sum\limits_{i=1}^{n} y_i^2} + \sqrt{\sum\limits_{i=1}^{n} \hat{y}_i^2}}.$$

According to the theory of the entropy estimation method, the parameter and noise distributions determined as a result of estimation are interval distributions. Thus, these intervals must be specified

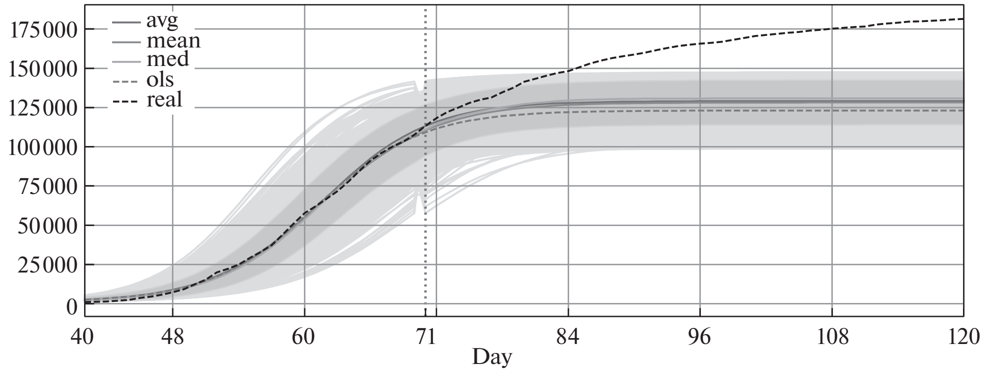

Fig. 3. Forecasting with no noise (NN) for $\mathcal{T}_{\rm est} = [40, 70]$ and $\mathcal{T}_{\rm pred} = [71, 120]$.

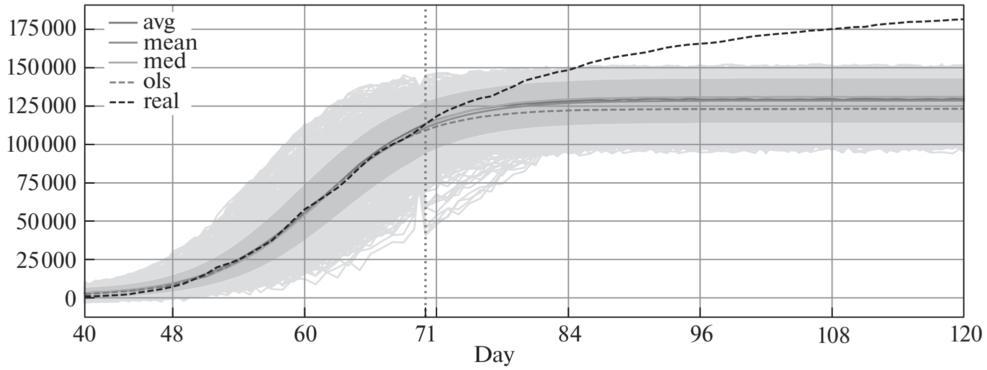

Fig. 4. Forecasting with noise within 10% (N1) for $\mathcal{T}_{\rm est} = [40, 70]$ and $\mathcal{T}_{\rm pred} = [71, 120]$.

for the method to be applied. In this paper, the intervals for the parameters were set on the basis of the optimal values obtained by the least squares estimation. The boundaries of the intervals were set within 20% of these values.

In the experiments, we used data scaled to the segment $[0, 1]$ on the estimation interval.

Estimation, testing, and forecasting were performed for several configurations:

- Without noise.
- With noise within 10%.
- With noise within 30%.

In the forecast, we used the noise distribution obtained for the last point in the estimation interval. The model was tested in the configuration corresponding to prediction.

The figures show the following results of modeling (trajectories):

- The least squares method (the dashed line marked ols).
- The real data (the dashed line marked real).
- Randomized forecasting with distribution-mean values of model parameters (the line marked mean_params).
- Randomized forecasting with ensemble mean (the line marked mean).
- Randomized forecasting with ensemble median (the line marked med).

Light gray marks the trajectories that make up the resulting ensemble, and dark gray marks the area of the standard deviation over the ensemble. All experiments were carried out for a sample from the model parameter distribution with a volume of 1000 and a sample from the noise distri-

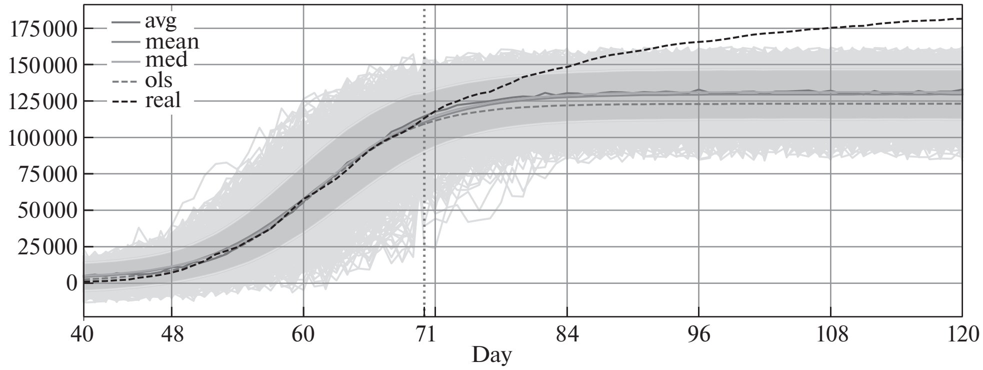

Fig. 5. Forecasting with noise within 30% (N3) for $\mathcal{T}_{\rm est} = [40, 70]$ and $\mathcal{T}_{\rm pred} = [71, 120]$.

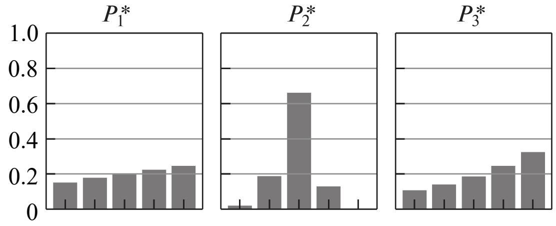

Fig. 6. Entropy-optimal distribution of the parameters P^*.

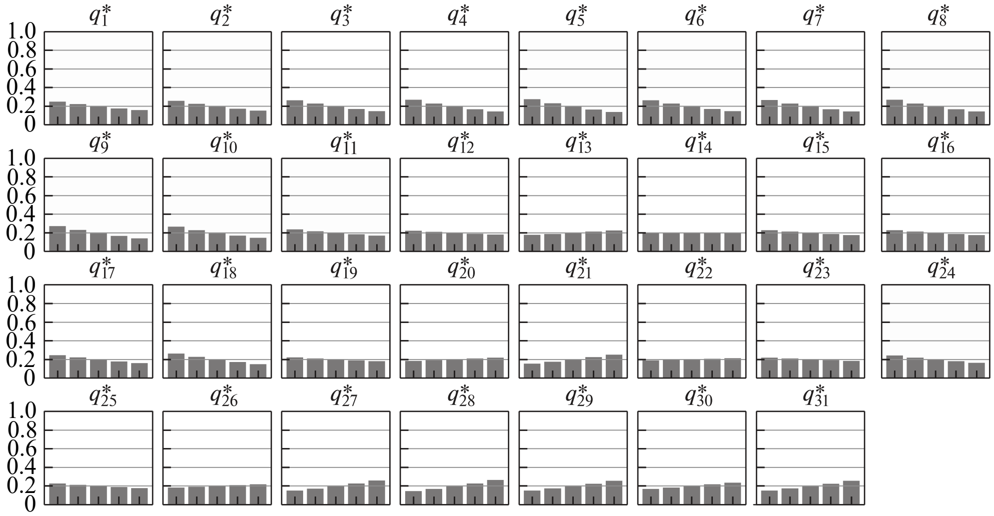

Fig. 7. Entropy-optimal distribution of noise Q^*.

butions with a volume of 100 for each parameter value. Noise distributions were generated for each point of the corresponding interval (testing and forecasting). Thus, the resulting ensemble consists of 10^5 trajectories. The vertical red dotted line is drawn at the start point of the prediction interval. The experiments were carried out on the Python 3.7 platform in the Windows 10 environment.

Figures 3–5 show the results of realization of the randomized predictive model on the intervals $\mathcal{T}_{\rm est} = [40, 70]$ and $\mathcal{T}_{\rm pred} = [71, 120]$ for three forecasting versions: no noise (NN), with noise 10% (N1), and with noise 30% (N3).

Figures 6 and 7 show the entropy-optimal parameter and noise distributions obtained as a result of training the model on the interval $\mathcal{T} = [40, 70]$ with noise 30%.

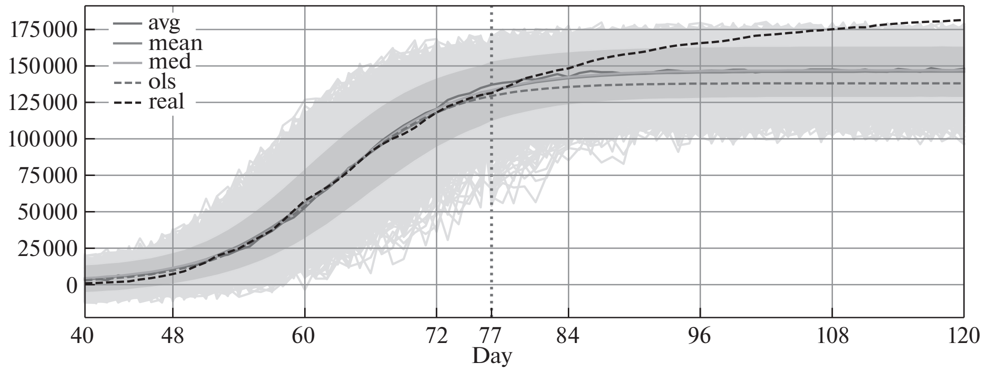

Fig. 8. Forecasting with noise within 30% (N3) for $\mathcal{T}_{\text{est}} = [40, 76]$ and $\mathcal{T}_{\text{pred}} = [77, 120]$.

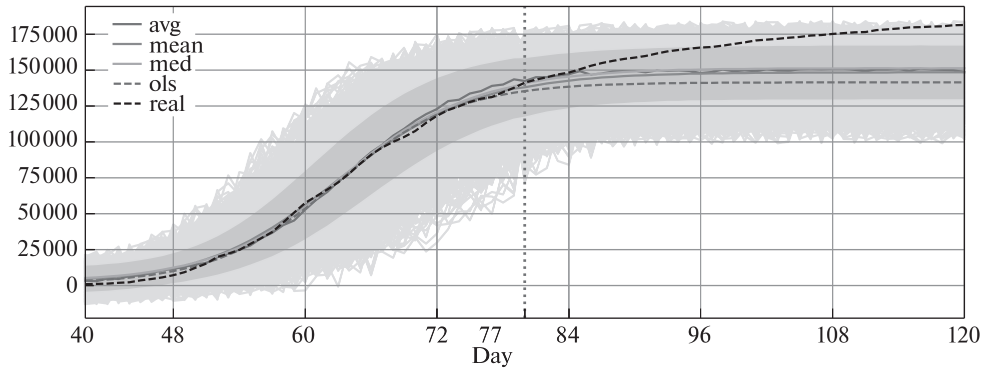

Fig. 9. Forecasting with noise within 30% (N3) for $\mathcal{T}_{\text{est}} = [40, 79]$ and $\mathcal{T}_{\text{pred}} = [80, 120]$.

Table 1. Estimates of parameters obtained by the least squares method

$\mathcal{T}_{test}$	a_1	a_2	a_3
[40, 70]	297749	0.2065	1.1448
[40, 76]	1086539	0.1856	1.0624
[40, 79]	83517	0.1803	1.0278

Figures 8 and 9 provide results of realization of the randomized predictive model on the intervals $\mathcal{T}_{\text{est}} = [40, 76]$ and $\mathcal{T}_{\text{est}} = [40, 79]$ and the corresponding forecasting intervals.

It is important to note that on the 77th day there was a slight increase (seen on the graph) in the number of cases. However, when training the model, this data had not yet been available. On the 79th day, there was still an increase in the number of cases, which began earlier, but in this case it was already possible to take these data into account in the training. The graphs show that the randomized model provides a more realistic forecast than the classical model in all cases.

The estimates of parameters obtained by the least squares method and the intervals of the parameters of the randomized model are indicated in Tables 1 and 2. In the configurations with noise, the noise intervals are given as $\Xi_j = [-0.1, 0.1]$ for N1 and $\Xi_j = [-0.3, 0.3]$ for N3, respectively.

The values of performance indicators when testing on three different intervals for the version of the model with noise 30% are listed in Table 3.

Analyzing the results obtained, it can be noted that the standard forecasting technique associated with the use of fitting the curve to the data by the least squares method, even if it demonstrates

Table 2. Configurations of parameters of the randomized model

$\mathcal{T}_{\text{pred}}$	A_1	A_2	A_3
$[40, 70]$	$[238199, 357299]$	$[0.1652, 0.2478]$	$[0.9158, 1.3738]$
$[40, 76]$	$[86923, 130384]$	$[0.1485, 0.2227]$	$[0.8499, 1.2749]$
$[40, 79]$	$[66813, 100220]$	$[0.1443, 0.2164]$	$[0.8223, 1.2334]$

Table 3. Performance metrics on the estimation interval

	R^2	MSE	NE	RNE
$\mathcal{T}_{test} = [40, 70]$				
ols	0.9984	0.0002	0.0004	0.0135
mean_params	0.9899	0.0011	0.0022	0.0335
mean	**0.9997**	0.0000	0.0001	0.0058
med	0.9980	0.0002	0.0004	0.0149
$\mathcal{T}_{test} = [40, 76]$				
ols	0.9982	0.0002	0.0004	0.0139
mean_params	0.9903	0.0012	0.0020	0.0314
mean	**0.9994**	0.0001	0.0001	0.0078
med	**0.9985**	0.0002	0.0003	0.0127
$\mathcal{T}_{test} = [40, 79]$				
ols	0.9982	0.0002	0.0004	0.0133
mean_params	0.9903	0.0012	0.0019	0.0306
mean	**0.9997**	0.0000	0.0001	0.0054
med	**0.9986**	0.0002	0.0003	0.0119

a certain efficiency, is not always able to solve the problem of constructing a correct forecast with high quality.

Due to the specifics of the COVID-19 epidemic under consideration here, there is a significant distortion of the data associated with it around the world. In this regard, the problem of predicting the number of infected people with a certain excess seems to be urgent. It can be seen from the results obtained that forecasting by a logistic model that had been evaluated using the data available at the time of the forecast considerably underestimated the real data. At the same time, the forecasts obtained using the approach proposed in this paper show an excess of the forecast values in comparison with the least squares method. It should also be noted that the use of noise, estimated point by point and then used in forecasting, in the model allows one to construct a more realistic forecast.

6. CONCLUSIONS

In the present paper, we have developed a method of randomized machine learning and forecasting based on the use of discrete random variables. This leads to problems more adapted to numerical solution with the use of modern computing technology. The proposed method was demonstrated using the problem of forecasting the total number of people infected with COVID-19 in Germany. The results obtained indicate the operability and efficiency of the method and its numerical implementation, which is determined by a smaller forecasting error in comparison with the standard technique based on the least squares method. It should also be noted that the constructed randomized model showed a good result on the training interval; however, the error was substantial in comparison with the real data on the forecast interval. This is most likely due to the fact that the logistic model is not efficient for forecasting at all stages of the development of the epidemic; in particular, to ensure an acceptable level of forecast performance, there must be signs of a slowdown in the epidemic during the training interval, as well as the presence of a real decline of the epidemic on the forecast interval. In the experiments, training was carried out using the data at the stage of the onset and active development of the epidemic; this explains the large error in forecasting.

APPENDIX

Consider the solution of problem (3.1)–(3.3) carried out by the method of Lagrange multipliers. The Lagrange function has the form

$$\begin{aligned} L(P,Q,\alpha,\beta,\lambda) = -H(P,Q) &+ \sum_{k=1}^{d}\alpha_k\left(\sum_{\ell=1}^{M}p_{k\ell}-1\right)\\ &+\sum_{j=1}^{m}\beta_j\left(\sum_{h=1}^{L}q_{jh}-1\right)+\sum_{j=1}^{m}\lambda_j\left(\bar{\Phi}(\mathbf{x}_j)+\sum_{h=1}^{L}\xi_{jh}q_{jh}-y_j\right).\end{aligned}$$

To seek the extremum of the Lagrange function, we calculate the derivatives with respect to the direct variables P and Q,

$$\frac{\partial L}{\partial P}=\frac{\partial L}{\partial p_{k\ell}}=\ln p_{k\ell}+1+\alpha_k+\sum_{j=1}^{m}\lambda_j\frac{\partial\bar{\Phi}_j}{\partial p_{k\ell}},$$

$$\frac{\partial L}{\partial Q}=\frac{\partial L}{\partial q_{jh}}=\ln q_{jh}+1+\beta_j+\sum_{j=1}^{m}\lambda_j\xi_{jh},$$

$$k=1,\dots,d,\quad j=1,\dots,m,\quad \ell=1,\dots,m,\quad h=1,\dots,L,$$

where $\bar{\Phi}_j=\bar{\Phi}(\mathbf{x}_j)$ and the derivative of the mean value of the model with respect to $p_{k\ell}$ is determined by the expression

$$\frac{\partial\bar{\Phi}_j}{\partial p_{k\ell}}=\sum_{\substack{\ell_s=1\\ s=1,\dots,d}}^{M}\Phi(x_j,a_{1\ell_1},\dots,a_{d\ell_d})\prod_{r\neq k}p_{r\ell_r}. \tag{A.1}$$

The sum in the expression for the derivative $\frac{\partial\bar{\Phi}}{\partial p_{k\ell}}$ contains $M(d-1)$ terms.

Equating the derivatives of the Lagrange function with respect to the direct variables with zero, we obtain an expression for the optimal probability distributions for the parameters and noise via the Lagrange multipliers,

$$p^*_{k\ell}(\alpha,\lambda)=\exp\left(-1-\alpha_k-\sum_{j=1}^{m}\lambda_j\frac{\partial\bar{\Phi}_j}{\partial p_{k\ell}}\right),$$

$$q^*_{jh}(\beta,\lambda)=\exp\left(-1-\beta_j-\lambda_j\xi_{jh}\right).$$

Let us transform these expressions as follows:

$$p^*_{k\ell}(\alpha,\lambda)=\exp\left(-\left(1+\alpha_k\right)\right)\exp\left(-\sum_{j=1}^{m}\lambda_j\frac{\partial\bar{\Phi}_j}{\partial p_{k\ell}}\right), \tag{A.2}$$

$$q^*_{jh}(\beta,\lambda)=\exp\left(-\left(1+\beta_j\right)\right)\exp\left(-\lambda_j\xi_{jh}\right); \tag{A.3}$$

substituting them into the normalization conditions (3.2), we obtain the expression

$$\exp(1+\alpha_k) = \exp\left(-\sum_{\ell=1}^{M} \lambda_j \frac{\partial \bar{\Phi}_j}{\partial p_{k\ell}}\right),$$
$$\exp(1+\beta_j) = \exp(-\lambda_j \xi_{jh}).$$

Let us substitute these expressions back into (A.2)–(A.3), thus eliminating the factors α and β, to obtain the ultimate expression of the entropy-optimal parameter and noise distribution depending on the factors λ,

$$p^*_{k\ell}(\lambda) = \frac{\exp\left(-\sum\limits_{j=1}^{m} \lambda_j \frac{\partial \bar{\Phi}_j}{\partial p_{k\ell}}\right)}{\sum\limits_{\ell=1}^{M} \exp\left(-\sum\limits_{j=1}^{m} \lambda_j \frac{\partial \bar{\Phi}_j}{\partial p_{k\ell}}\right)}, \quad k = 1,\dots,d,\ \ell = 1,\dots,m, \tag{A.4}$$

$$q^*_{jh}(\lambda) = \frac{\exp(-\lambda_j \xi_{jh})}{\sum\limits_{h=1}^{L} \exp(-\lambda_j \xi_{jh})}, \quad j = 1,\dots,m,\ h = 1,\dots,L. \tag{A.5}$$

The factors λ are determined by the solution of the system of equations obtained by the substitution of the expressions (A.4)–(A.5) into the balance relations (3.3),

$$\sum_{\substack{\ell_k=1 \\ k=1,\dots,d}}^{M} \Phi(\mathbf{x}_j, a_{1\ell_1},\dots,a_{d\ell_d}) \prod_{\substack{\ell_s=1 \\ s=1,\dots,d}}^{M} p^*_{s\ell_s}(\lambda) + \sum_{h=1}^{L} \xi_{jh} q^*_{jh}(\lambda) = y_j, \quad j = 1,\dots,m. \tag{A.6}$$

Thus, having solved system (A.6), we obtain the entropy-optimal parameter and measurement noise distributions, which is precisely the ultimate goal of training the model with the use of real data.

It should be noted that to solve this system in practice, it is necessary to involve some kind of numerical method, since solving it analytically is fraught with considerable difficulties.

Calculating the left-hand side of the system will require calculating the average value of the random function $\bar{\Phi}$ as well as its derivative with respect to the distribution P defined by the expressions (3.4) and (A.1). The summation in these expressions must be performed over all combinations of indices, and so the number of addition operations increases as a power of the number d of parameters.

FUNDING

This work was supported by the Russian Foundation for Basic Research, project no. 20-07-00683.

REFERENCES

1. Bishop, C.M., *Pattern Recognition and Machine Learning. Series: Information Theory and Statistics*, Berlin–Heidelberg: Springer, 2006.
2. Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning*, Berlin–Heidelberg: Springer, 2001.
3. Aivazyan, S.A. and Mkhitaryan, V.S., *Prikladnaya statistika i osnovy ekonometriki* (Applied Statistics and Fundamentals of Econometrics), Moscow: Yuniti, 1998.
4. Merkov, A.B., *Raspoznavanie obrazov. Vvedenie v metody statisticheskogo obucheniya* (Pattern Recognition. Introduction to Statistical Learning Methods), Moscow: URSS, 2010.

5. Arkad'ev, A.G. and Braverman, E.M., *Obuchenie mashiny raspoznavaniyu obrazov* (Training the Machine to Recognize Patterns), Moscow: Nauka, 1964.

6. Tsypkin, Ya.Z., *Osnovy teorii obuchayushchikhsya sistem* (Fundamentals of the Theory of Learning Systems), Moscow: Nauka, 1970.

7. Vapnik, V.N. and Chervonenkis, A.Ya., *Vosstanovlenie zavisimostei po empiricheskim dannym* (Recovering Dependences from Empirical Data), Moscow: Nauka, 1979.

8. Vapnik, V.N. and Chervonenkis, A.Ya., *Teoriya raspoznavaniya obrazov* (Pattern Recognition Theory), Moscow: Nauka, 1974.

9. Cristianini, N. and Shawe-Taylor, J., *An Introduction to Support Vector Machines*, 2000.

10. Breiman, J.H., Friedman, R., Olshen, A., and Stone, C.J., *Classification and Regression Trees*, 1984.

11. Rosenblatt, F., *The Perceptron, a Perceiving and Recognizing Automaton Project Para*, Cornell Aeronaut. Lab., 1957.

12. Rumelhart, D.E., Williams, R.J., and Hinton, G., Learning representations by back-propagating errors, *Nature*, 1986, vol. 323, no. 6088, pp. 533–538.

13. Popkov, Yu.S., Popkov, A.Yu., and Dubnov, Yu.A., *Randomizirovannoe mashinnoe obuchenie pri ogranichennykh naborakh dannykh: ot empiricheskoi veroyatnosti k entropiinoi randomizatsii* (Randomized Machine Learning with Limited Datasets: from Empirical Probability to Entropy Randomization), Moscow: LENAND, 2019.

14. Boltzmann, L., On the relationship between the second fundamental theorem of the mechanical theory of heat and probability calculations regarding the conditions for thermal equilibrium, in *L. Boltzmann. Selected Works. Classics of Science Series,* Shpak, L.S., Ed., Moscow: Nauka, 1984.

15. Jaynes, E.T., Information theory and statistical mechanics, *Phys. Rev.*, 1957, vol. 106, no. 4, pp. 620–630.

16. Jaynes, E.T., *Probability Theory: the Logic of Science*, Cambridge: Cambridge Univ. Press, 2003.

17. Shannon, C.E., Communication theory of secrecy systems, *Bell Labs Tech. J.*, 1949, vol. 28, no. 4, pp. 656–715.

18. Diebold, F., *Elements of Forecasting, 4th Ed.*, Ohio, US: Thomson, South-Western, 2007.

19. Gneiting, T. and Katzfuss, M., Probabilistic forecasting, *Annu. Rev. Stat. Its Appl.*, 2014, no. 1, pp. 125–151.

20. Hong, T. and Fan, S., Probabilistic electric load forecasting: A tutorial review, *Int. J. Forecast.*, 2016, vol. 32, no. 3, pp. 914–938.

21. Aivazyan, S.A. and Mkhitaryan, V.S., *Prikladnaya statistika. Klassifikatsiya i snizhenie razmernosti* (Applied Statistics. Classification and Dimension Reduction), Moscow: Finansy i Statistika, 1989.

22. Golan, A., Judge, G., and Miller, D., *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, New York: John Wiley & Sons, 1996.

23. Golan, A.etal., Information and entropy econometrics—a review and synthesis, *Found. Trends Econometrics*, 2008, vol. 2, no. 1–2, pp. 1–145.

24. Popkov, Y.S., Volkovich, Z., Dubnov, Y.A., Avros, R., and Ravve, E., Entropy 2-soft classification of objects, *Entropy*, 2017, vol. 19, no. 4, p. 178.

25. Popkov, Y.S., Popkov, A.Y., and Dubnov, Y.A., Elements of randomized forecasting and its application to daily electrical load prediction in a regional power system, *Autom. Remote Control*, 2020, vol. 81, pp. 1286–1306.

26. Popkov, Y.S., Popkov, A.Y., Dubnov, Y.A., and Solomatine, D., Entropy-randomized forecasting of stochastic dynamic regression models, *Mathematics*, 2020, vol. 8, no. 7, p. 1119.

27. Dong, E., Du, H., and Gardner, L., An interactive web-based dashboard to track covid-19 in real time, *Lancet Infect. Dis.*, 2020, vol. 20, no. 5, pp. 533–534.

28. *Mathematical Epidemiology. Lecture Notes in Mathematics,* Brauer, F., van den Driessche, P., and Wu, J., Eds., Berlin–Heidelberg: Springer, 2008. https://doi.org/10.1007/978-3-540-78911-6

29. Verhulst, P.-F., Notice sur la loi que la population suit dans son accroissement, *Corresp. Math. Phys.*, 1893, no. 10, pp. 113–126.

30. Singer, H.M., The COVID-19 pandemic: growth patterns, power law scaling, and saturation, *Phys. Biol.,* 2020, vol. 17, no. 5, p. 055001. https://doi.org/10.1088/1478-3975/ab9bf5

31. Kumar, J. and Hembram, K.P.S.S., Epidemiological study of novel coronavirus (COVID-19), 2020. http://arxiv.org/abs/2003.11376.

32. Yang, W., Zhang, D., Peng, L., Zhuge, C., and Hong, L., Rational evaluation of various epidemic models based on the COVID-19 data of China, 2020. http://arxiv.org/abs/2003.05666.

33. Tatrai, D. and Varallyay, Z., COVID-19 epidemic outcome predictions based on logistic fitting and estimation of its reliability, 2020. http://arxiv.org/abs/2003.14160.

34. Morais, A.F., Logistic approximations used to describe new outbreaks in the 2020 COVID-19 pandemic. 2020. http://arxiv.org/abs/2003.11149.

35. Shen, C.Y., Logistic growth modelling of COVID-19 proliferation in China and its international implications, *Int. J. Infect. Dis.,* 2020, vol. 96, pp. 582–589. https://doi.org/10.1016/j.ijid.2020.04.085

36. Wang, P., Zheng, X., Li, J., and Zhu, B., Prediction of epidemic trends in COVID-19 with logistic model and machine learning techniques, *Chaos Solitons & Fractals,* 2020, vol. 139, p. 110058. https://doi.org/10.1016/j.chaos.2020.110058

37. Chen, D.-G., Chen, X., and Chen, J.K., Reconstructing and forecasting the COVID-19 epidemic in the United States using a 5-parameter logistic growth model, *Global Health Res. Policy,* 2020, vol. 5, no. 1, p. 25. https://doi.org/10.1186/s41256-020-00152-5

This paper was recommended for publication by A.I. Mikhal'skii, a member of the Editorial Board