

He YuanfangPhd student, Department of Information Systems and Technology, <https://orcid.org/0000-0002-6925-1540>

Taras Shevchenko National University of Kyiv, Kyiv

DEVELOPMENT OF A TREND FORECASTING MODEL FOR ENVIRONMENTAL POLLUTION MONITORING

Abstract. A complex model for forecasting time series of environmental pollution indicators is described, considering the aggregation of various forecasting models, which are formed based on predictive statistical analysis of pollution indicators and have an adaptive nature. The model differs from known models by providing the possibility of adapting the model parameters to changes in the state of the environment, which is especially important in the conditions of using such models in monitoring systems. The complex forecasting model includes higher-order exponential smoothing, Holt, Winters, moving average, weighted moving average, and autoregression models. All the parameters set in these models are related to the Hurst index, which is calculated based on predictive fractal statistical analysis of the time series. Relevant descriptions and justifications are given. Using such a model as part of the econometric system will help predict and respond more effectively to possible changes in the values of pollution parameters. In particular, the persistence of the time series of pollution parameters can mean a stable trend of increasing or decreasing pollution. Suppose the time series becomes close to random or ergodic. In that case, this may mean an emergency or additional erratic emissions in the region that must be monitored. The described model is a forecasting model that is part of the system for monitoring environmental pollution parameters. In the future, a model for forecasting the pollution level in different regions of the People's Republic of China will be implemented.

Keywords: forecasting model; monitoring of pollution parameters; predictive analysis; project management; information management; critical infrastructure; biomonitoring

Introduction

Ensuring a balanced solution to the problems of preserving a favorable environment, applying new approaches to environmental protection and respecting the economic interests of both enterprises and the entire population requires a purposeful scientific approach. Recently, a close relationship between the development of the economy and changes in the environment has been observed; the mutual influence of both the state of ecology on economic development and the results of economic activity on the state of the natural environment is increasing.

In conditions of constant deterioration of the ecological situation, the scientific basis for managing anthropogenic influence and multifactorial analysis of the formation of the pollution level in combination with the operational forecast of the pollution level is the only effective way to solve the problem.

Environmental pollution includes the study of air pollution, groundwater and surface water pollution, soil pollution, and the impact on the biosphere. Each type of pollution requires research models, methods, and forecasting.

The importance of environmental protection research is also confirmed by the governments of all the world's leading countries, which spend an average of

0.8% of their national budgets (more than \$600 billion) on environmental protection measures [1]. Among these expenses, R&D is in 3rd place.

Three basic approaches to forecasting the state of environmental pollution can be distinguished. Works [2 – 5] use an approach based on pattern recognition using neural networks. The possibility of applying methods based on regression analysis is shown in works [6 – 9]. The authors in [10; 11] use time series analysis methods, particularly trend forecasting. Peculiarities of environmental pollution assessment are also described in works [12; 13].

Problem statement

We will carry out a mathematical formalization of the problem considered in this dissertation. Let be a given discrete set

$$T = \{t_1, t_2, \dots, t_n\}, \text{ then}$$

$$Q = (q(t_1), q(t_2), \dots, q(t_n)),$$

where Q is a time series of the level of environmental pollution, reflecting the quantitative indicators of pollution parameters, which are fixed at moments of time t_1, t_2, \dots, t_n as a finite sequence of measurements, the initial moment of time is delayed by t_1 , the current moment of time will be denoted by t_n , $t_1, t_n \in T$, $n \in \mathbb{N}$.

Let's assume that the level of environmental pollution is determined at fixed points in time, such as a day, a week, a month, a year, etc. You can use the appropriate sensors for this. The measurement results are real numbers, $q(t_r) \in \mathbb{R}$, $r = \overline{1, n}$.

The forecasting method is a sequence of forecasting operations and actions, the execution of which ensures the construction of a forecasting model, taking into account the constructed estimates of the accuracy of the forecast values. The most common methods of forecasting include extrapolation, interpolation, the method of expert assessments, mathematical modeling, etc. The forecaster's task is to choose a method that fully meets the goal of forecasting and provides the required accuracy.

By the forecasting system, we will understand the system of methods that function by the forecasting principles; that is, they satisfy the following requirements: systematicity and interconnectedness of forecasts, continuity, adequacy of forecasts of the object of research, efficiency, variability, etc.

The analysis of time series involves the implementation of two main stages, in which the goal of the analysis is laid. The first stage is studying the structure of time series or predictive analysis. This stage involves pre-processing the data and identifying unique characteristics directly used to build an adequate forecasting model. The second stage is constructing and assessing the time series forecasting model.

A comprehensive model for forecasting time series of environmental pollution parameters

Fractal time series analysis is a universal tool that can be used for predictive analysis and integrated into the environmental monitoring system [14; 15].

Having calculated the Hurst index $H(Q^h)$ of the time series of environmental pollution parameters $Q = (q(t_1), q(t_2), \dots, q(t_n))$, it is possible to determine the presence of memory in the time series, that is, the presence of long-term dependence and the presence of cyclic components. The latter can be identified based on the analysis of V statistics. The growth of this statistic with an increase in the number of observations indicates the trend stability of the time series or persistence, and stabilization or decrease indicates the strengthening of the influence of random factors. A sharp change in the trend from rising to falling may indicate the transformation of the series into white noise, which may be characteristic of an emergency in changing the level of environmental pollution.

Let's consider a complex model for forecasting time series of environmental pollution parameters, considering the aggregation of various forecasting

models formed based on predictive statistical analysis of pollution indicators. A vital feature of the model is that it combines other well-known forecasting models and allows parameters to be adjusted to match the results of predictive series analysis based on fractal analysis.

The task is based on a time series

$$Q = (q(t_1), q(t_2), \dots, q(t_n))$$

to make the most accurate forecast, that is, to establish the behavior of the time series of pollution parameters a certain number of points ahead, that is, to find estimates of values

$$\hat{q}(t_{n+1}), \hat{q}(t_{n+2}), \dots, \hat{q}(t_{n+k}),$$

where $\hat{q}(t_{n+i})$ is time series forecast

$$Q = (q(t_1), q(t_2), \dots, q(t_n)) \text{ on } i = \overline{1, k} \text{ point forward.}$$

To find a solution to the problem, it is necessary to find a functional dependence that would approximate the required forecast value based on the known values of the time series. That is,

$$\hat{q}(t_{n+1}) = \Phi(q(t_c), q(t_{c+1}), \dots, q(t_n)), \quad c < n,$$

$$\hat{q}(t_{n+2}) = \Phi(q(t_{c+1}), q(t_{c+2}), \dots, q(t_n), \hat{q}(t_{n+1})),$$

...

$$\hat{q}(t_{n+k}) =$$

$$= \Phi(q(t_{c+k-1}), q(t_{c+k}), \dots, q(t_n), \hat{q}(t_{n+1}), \dots, \hat{q}(t_{n+k-1})).$$

To evaluate the quality of the forecast, you can use the average absolute error, average relative error, standard deviation, etc.

Let there be a set of models $\Phi_1, \Phi_2, \dots, \Phi_M$ that, based on the time series, allow you to make the most accurate forecast, that is, find the value:

$$\hat{q}_j(t_{n+1}) = \Phi_j(\alpha_1^j, \alpha_2^j, \dots, \alpha_m^j, q(t_c), q(t_{c+1}), \dots, q(t_n)),$$

$$c < n,$$

$$\hat{q}_j(t_{n+2}) =$$

$$= \Phi_j(\alpha_1^j, \alpha_2^j, \dots, \alpha_m^j, q(t_{c+1}), q(t_{c+2}), \dots, q(t_n), \hat{q}(t_{n+1})),$$

...

$$\hat{q}_j(t_{n+k}) = \Phi_j(\alpha_1^j, \alpha_2^j, \dots, \alpha_m^j, q(t_{c+k-1}), q(t_{c+k}), \dots,$$

$$\dots, q(t_n), \hat{q}(t_{n+1}), \dots, \hat{q}(t_{n+k-1})), \quad j = \overline{1, M},$$

where $\hat{q}_j(t_{n+i})$ is point-ahead forecast of the time series

Q , $i = \overline{1, k}$ based on the model $j = \overline{1, M}$, α_d^j are parameters of the forecasting model j , $d = \overline{1, m}$, m is the number of parameters of the forecasting model j .

Let the model Φ_1 be an ordinary exponential model of order $p \geq 0$. The exponential order model $p \geq 0$ is determined by the formula:

$$x_n^{[p]} = \alpha \cdot x_n^{[p-1]} + (1 - \alpha) x_{n-1}^{[p]},$$

where $x_n^{[0]} = q(t_1)$, $x_0, x_0^{[2]}, \dots$ is initial conditions of exponential averages of the corresponding order [9],

$\hat{q}_j(t_{n+1}) = x_n^{[p]}$, $\alpha \in [0, 1]$. That is, this model will be defined as $\Phi_1(\alpha, q(t_1), q(t_2), \dots, q(t_n))$.

Let the model Φ_2 be represented by Holt's exponential smoothing model, which is used to model time series with a pronounced trend component [9]:

$$\begin{aligned} \hat{q}_2(t_{n+1}) &= x_n + y_n, \\ x_n &= \alpha_1 q(t_n) + (1 - \alpha_1)(x_{n-1} + y_{n-1}), \\ y_n &= \alpha_2(x_n - x_{n-1}) + (1 - \alpha_2)y_{n-1}, \end{aligned}$$

where $\hat{q}_2(t_{n+1})$ is a forecast calculated one point ahead according to the Holt model of the time series Q , $\alpha_1, \alpha_2 \in [0, 1]$, accordingly, the model has the form $\Phi_2(\alpha_1, \alpha_2, q(t_1), q(t_2), \dots, q(t_n))$.

The Winters model Φ_3 is used for processes with an additive trend component and multiplicative seasonality. In this model, the value of the time series without a seasonal component, the trend and the seasonal component are smoothed separately [9]:

$$\begin{aligned} \hat{q}_3(t_{n+1}) &= (x_n + y_n) s_{n-P+1}, \\ x_n &= \alpha_1 \frac{q(t_n)}{s_{n-P}} + (1 - \alpha_1)(x_{n-1} + y_{n-1}), \\ y_n &= \alpha_2(x_n - x_{n-1}) + (1 - \alpha_2)y_{n-1}, \\ s_n &= \alpha_3 \frac{q(t_n)}{s_n} + (1 - \alpha_3)s_{n-P}, \end{aligned}$$

where $\alpha_1, \alpha_2, \alpha_3 \in [0, 1]$ is smoothing parameters, P is period of the seasonal cycle, s_n is assessment of the seasonal component of the model. We denote this model by $\Phi_3(\alpha_1, \alpha_2, \alpha_3, q(t_1), q(t_2), \dots, q(t_n))$.

The autoregressive model Φ_4 is determined by the formula:

$\hat{q}(t_{n+1}) = \gamma_0 + \gamma_1 q(t_{n-1}) + \gamma_2 q(t_{n-2}) + \dots + \gamma_c q(t_{n-c})$, the parameters are not determined in advance and are calculated based on the condition of minimizing the sum of root mean square errors. The number of points for applying the autoregressive model depends on the long-term memory, so we can write the model as: $\Phi_4(q(t_{n-c+1}), q(t_{n-c+2}), \dots, q(t_n))$.

The moving average model is defined by a parameter that determines the number of points used to calculate the predicted value. The larger the memory in the time series, the larger this parameter should be. That is, the model $\Phi_5(q(t_{n-c+1}), q(t_{n-c+2}), \dots, q(t_n))$ is defined as:

$$\hat{q}(t_{n+1}) = \frac{1}{\alpha} \sum_{j=0}^{c-1} q(t_{n-j}), \quad \alpha > 0.$$

A weighted moving average with a set of

normalized weights $\{\omega_1, \omega_2, \dots, \omega_c\}$, $\sum_{j=1}^c \omega_j = 1$,

determined by the formula:

$$\hat{q}(t_{n+1}) = \frac{1}{c} \sum_{j=0}^{c-1} \omega_{j+1} q(t_{n-j}), \quad c > 0.$$

We will denote this model by $\Phi_6(q(t_{n-c+1}), q(t_{n-c+2}), \dots, q(t_n))$. This list of models can be increased. However, the initial parameters must be chosen, considering the predictive analysis results. This will allow for a more minor forecasting error of the time series and a more accurate forecast of the time series of environmental pollution parameters. This is very important for building an effective environmental monitoring system.

We will build a general forecasting model that will consider all the described models and the results of the predictive fractal analysis of the time series. The model differs from known models by providing the possibility of adapting the model parameters to environmental changes. Each of the described models generates an error when applied. The minimum prediction error determines the accuracy of the model. Let on the section of the time series $Q^\circ = (q(t_z), q(t_{z+1}), \dots, q(t_n))$ the accuracy of forecasting models was observed, then the error can be calculated for each of the models $G_n^i(Q^\circ, Q)$:

$$G_n^i(Q^\circ, Q) = \sqrt{\frac{1}{n-z+1} \sum_{j=z}^n (\hat{q}_i(t_j) - q_i(t_j))^2}, \quad i = \overline{1, 6},$$

$G_n^i(Q^\circ, Q)$ is the prediction error of the point $q(t_n)$, which is made on the basis of the model Φ_i .

For each model, we will calculate the model selection criterion for forecasting according to the formula:

$$\bar{G}_n^i = \eta G_n^i(Q^\circ, Q) + (1 - \eta) \bar{G}_{n-1}^i, \quad \eta \in [0, 1],$$

where \bar{G}_n^i is the exponentially smoothed forecast error of the point $q(t_n)$, which is made on the basis of the model Φ_i .

If the error argument is the model from which it was calculated, ie $\bar{G}_n^i = \bar{G}_n(\Phi_i)$. Then the model for which the condition of the minimum forecasting error is fulfilled is selected for forecasting:

$$\Phi_i^* = \arg \min_i (\bar{G}_n^i).$$

We will use for the selected list of models their relationship with the results of predictive analysis. That is, with the calculated value $H(Q^n)$. For the first model, since the smoothing parameter α indicates the weight of the influence of the previous values of the series on the result. Moreover, if the value α is close to one, then the

forecast according to this model will resemble a naive one. On the other hand, the forecast may be naive if the series being forecast is highly persistent. That is, it is possible to derive a rational formula for determining the smoothing parameter α in the model $\Phi_1(\alpha, q(t_1), q(t_2), \dots, q(t_n))$:

$$\alpha = \begin{cases} H(Q^n), H(Q^n) \geq 0.5 \\ 0.5, H(Q^n) < 0.5 \end{cases}$$

In the second model $\Phi_2(\alpha_1, \alpha_2, q(t_1), q(t_2), \dots, q(t_n))$ the first smoothing parameter α_1 determines the trend, and the second α_2 – random component. Accordingly, these parameters can be determined by the formula:

$$\alpha_1 = \begin{cases} H(Q^n), H(Q^n) \geq 0.5 \\ 0.5, H(Q^n) < 0.5 \end{cases},$$

$$\alpha_2 = \begin{cases} \max\{1, 0.5 + 10|H_h^T - H(Q^n)|\}, 0.5 \leq H(Q^n) \leq H_h^T \\ 0.5, H(Q^n) > H_h^T \text{ or } H(Q^n) < 0.5 \end{cases}$$

In the third model $\Phi_3(\alpha_1, \alpha_2, \alpha_3, q(t_1), q(t_2), \dots, q(t_n))$, the first smoothing parameter α_1 determines the trend, and the second α_2 random component, the third α_3 seasonality. If the trend change points of the V statistics curve correspond to the seasonality of P, then these parameters can be determined by the formula:

$$\alpha_1, \alpha_3 = \begin{cases} H(Q^n), H(Q^n) \geq 0.5 \\ 0.5, H(Q^n) < 0.5 \end{cases},$$

$$\alpha_2 = \begin{cases} \max\{1, 0.5 + 10|H_h^T - H(Q^n)|\}, 0.5 \leq H(Q^n) \leq H_h^T \\ 0.5, H(Q^n) > H_h^T \text{ or } H(Q^n) < 0.5 \end{cases}$$

In models $\Phi_t(q(t_{n-c+1}), q(t_{n-c+2}), \dots, q(t_n))$, $t = \overline{4, 6}$, the value of parameter c is determined by the

presence of long-term memory in the time series. It can be determined by visual inspection of the V statistics curve. If the point that corresponds to a sharp change in the trend of this curve of growth and decline is equal to the value P, $P \in \square$, then $c=P$:

$$\Phi_t(q(t_{n-P+1}), q(t_{n-P+2}), \dots, q(t_n)), t = \overline{4, 6}.$$

This model's peculiarity is that, in addition to considering the results of predictive fractal analysis, it is adaptive. That is, it can adapt model parameters to environmental changes, which is especially important when using such models in monitoring systems.

Conclusions

1. A comprehensive model for forecasting time series of environmental pollution indicators is described, considering the aggregation of various forecasting models, which are formed based on predictive statistical analysis of pollution indicators and have an adaptive nature. The model differs from known models by providing the possibility of adapting the model parameters to changes in the state of the environment, which is especially important in the conditions of using such models in monitoring systems.

2. The complex forecasting model includes higher-order exponential smoothing models, Holt, Winters, moving average, weighted moving average, and autoregressive models. All the parameters set in these models are related to the Hurst index, which is calculated based on predictive fractal statistical analysis of the time series. Relevant descriptions and justifications are given.

It is indicated that using such a model as part of the econometric system will help to predict and respond more effectively to possible changes in the values of pollution parameters. In particular, the persistence of the time series of pollution parameters can mean a stable trend of increasing or decreasing pollution. Suppose the time series becomes close to random or ergodic. In that case, this may mean an emergency or additional erratic emissions in the region that must be monitored.

References

1. Government expenditure on environmental protection. (2019). Eurostat: Statistics Explained. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php/Government_expenditure_on_environmental_protection.
2. Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M. & et al. (2003). Extensive evaluation of neural network models for the prediction of no 2 and pm 10 concentrations, compared with a deterministic modelling system and measurements in central helsinki. *Atmos Environ*, 37 (32), 4539–550. doi: 10.1016/S1352-2310(03)00583-1.
3. Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., Kolehmainen, M. (2004). Evolving the neural network model for forecasting air pollution time series. *Eng Appl Artif Intell*; 17 (2), 159–67. doi: 10.1016/j.engappai.2004.02.002.
4. Kolehmainen, M., Martikainen, H., Ruuskanen, J. (2001). Neural networks and periodic components used in air quality forecasting. *Atmos Environ*, 35 (5), 815–25. doi: 10.1016/S1352-2310(00)00385.
5. Beckerman, B. S., Jerrett, M., Martin, R. V., van Donkelaar, A., Ross, Z., Burnett, R. T. (2013). Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in california. *Atmospheric Environ*, 77, 172–7. doi: 10.1016/j.atmosenv.2013.04.024.
6. Liu, B. C., Binaykia, A., Chang, P. C., Tiwari, M. K., Tsao, C. C. (2017). Urban air quality forecasting based on multi-dimensional collaborative support vector regression (svr): A case study of beijing-tianjin-shijiazhuang. *PLOS ONE*, 12(7), 1–17.

7. Bobb, J. F., Valeri, L., Henn, B. C., Christiani, D. C., Wright, R. O., Mazumdar, M. & et al. (2014). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, 16 (3), 058.
8. Gass, K., Klein, M., Chang, H. H., Flanders, W. D., Strickland, M. J. (2014). Classification and regression trees for epidemiologic research: an air pollution example. *Environ Health*, 13 (1), 17. doi: 10.1186/1476-069X-13-17.
9. Vercellis, C. (2009). *Business intelligence: data mining and optimization for decision making*. Cornwall: John Wiley & Sons Ltd. Publication, 417.
10. Kuchansky, A., Biloshchytskyi, A., Andrashko, Yu., Vatskel, V., Biloshchytska, S., Danchenko, O. & et al. (2018). Combined models for forecasting the air pollution level in infocommunication systems for the environment state monitoring. *2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS)*. Lviv, 125–130. DOI: 10.1109/IDAACS-SWS.2018.8525608.
11. Kuchansky, A., Biloshchytskyi, A., Andrashko, Y., Biloshchytska, S., Shabala, Y., Myronov, O. (2018). Development of adaptive combined models for predicting time series based on similarity identification. *Eastern-European Journal of Enterprise Technologies*, 1(4 (91)), 32–42. <https://doi.org/10.15587/1729-4061.2018.121620>.
12. Yuanfang, He. (2019). Fomalization of the problem of evaluation of pollution of the environment. *Management of development of complex systems*, 38, 168–172, <https://doi.org/10.6084/m9.figshare.9788702>.
13. He, Y., Biloshchytskyi, A. O. (2019). Hardware of the information system for environmental pollution monitoring. *Scientific Bulletin of Uzhhorod University. Series of Mathematics and Informatics*, 2(35), 143–148. [https://doi.org/10.24144/2616-7700.2019.2\(35\).143-148](https://doi.org/10.24144/2616-7700.2019.2(35).143-148).
14. Peters, E. E. (1994). Fractal market analysis: applying chaos theory to investment and economics. *John Wiley & Sons Inc*, 336.
15. Anis, A., Lloyd, E. (1976). The expected value of the adjusted rescaled Hurst Range of independent normal summands. *Biometrika*, 63, 111–116.
- 16.

The article has been sent to the editorial board 07.03.2024

Хе Юаньфанг

Аспірант кафедри інформаційних систем та технологій, <https://orcid.org/0000-0002-6925-1540>
Київський національний університет імені Тараса Шевченка, Київ

РОЗРОБКА ТРЕНДОВОЇ ПРОГНОЗНОЇ МОДЕЛІ ДЛЯ МОНІТОРИНГУ ЗАБРУДНЕННЯ НАВКОЛИШНЬОГО СЕРЕДОВИЩА

Анотація. Описано комплексну модель прогнозування часових рядів показників забруднення навколишнього середовища з врахуванням агрегації різних моделей прогнозування, що формуються на основі передпрогнозного статистичного аналізу показників забруднення і мають адаптивний характер. Модель відрізняється від відомих моделей забезпеченням можливістю адаптації параметрів моделі до змін у стані навколишнього середовища, що особливо важливо в умовах використання таких моделей в системах екомоніторингу. До складу комплексної моделі прогнозування включено моделі експоненціального згладжування вищого порядку, моделі Хольта, Вінтерса, плинної середньої, зваженої плинної середньої, авторегресійної моделі. Всі параметри, які задаються в цих моделях, пов'язані з показником Херста, який розраховується на основі передпрогнозного фрактального статистичного аналізу часового ряду. Наведено відповідні описання й обґрунтування. Вказано, що використання такої моделі в складі системи екомоніторингу допоможе ефективніше передбачати і реагувати на можливі зміни значень параметрів забруднення. Зокрема, персистентність часового ряду параметрів забруднення може означати стабільну тенденцію до зростання або спадання забруднення. Якщо ж часовий ряд стає близьким до випадкового або ергодичним, то це може означати надзвичайну ситуацію, або ж те, що в регіоні з'явилися додаткові непостійні викиди, які необхідно моніторити. Описано модель прогнозування з частиною системи моніторингу параметрів забруднення зовнішнього середовища. У майбутньому планується впровадити модель для прогнозування рівня забруднення в різних регіонах Китайської народної республіки.

Ключові слова: модель прогнозування; моніторинг параметрів забруднення; передпрогнозний аналіз; управління проектами; інформаційний менеджмент; критична інфраструктура; біомоніторинг

Link to publication

- APA He, Yuanfang. (2024). Development of a trend forecasting model for environmental pollution monitoring. *Management of Development of Complex Systems*, 57, 62–66, [dx.doi.org/10.32347/2412-9933.2024.57.62-66](https://doi.org/10.32347/2412-9933.2024.57.62-66).
- ДСТУ Хе Юаньфанг. Розробка трендової прогнозної моделі для моніторингу забруднення навколишнього середовища. *Управління розвитком складних систем*. Київ, 2024. № 57. С. 62 – 66, [dx.doi.org/10.32347/2412-9933.2024.57.62-66](https://doi.org/10.32347/2412-9933.2024.57.62-66).