UDC 338.2
DOI: 10.15587/1729-4061.2022.259323

BUILDING A MODEL FOR CHOOSING A STRATEGY FOR REDUCING AIR POLLUTION BASED ON DATA PREDICTIVE ANALYSIS

This paper formalizes the model of choosing a strategy for reducing air pollution in an urban environment. The model involves determining the optimal location of biotechnological systems – biotechnological filter systems or smart air purification devices based on solving the problem of discrete optimization, taking into consideration the forecast of the air quality index. Two subtasks have been formalized, which make it possible to form a strategy for reducing air pollution. To solve one of the subtasks, a combined selective model for predicting the time series of the Air Quality Index (CSM) was built. The combined model software suite consists of the EMD-ESM hybrid model (Empirical Mode Decomposition-Exponential Smoothing Model), the HWM additive model (Holt-Winters Model), and the adaptive TLM (Trigg-Lich Model). To verify the proposed combined selective model, the time series of air quality indices (AQI) for the city of Nur-Sultan (data from 2010–2021, period 6 hours) were selected. As a result of verification, it was established that in the case of short-term forecasting of the air quality index time series, the EMD-ESM model has an advantage according to the criterion of a minimum root mean square error (RMSE), δ=0.11. For the case of medium-term forecasting of 3<τ≤5, the combined selective model (CSM) has the advantage. The results reported here are input data for the task of choosing strategies for reducing the volume of air pollution in the urban environment. The study's results make it possible to increase the flexibility of the formation of strategies for reducing air pollution since they avoid restrictions on the location of cleaners in specific urban areas. The consequence is the improvement of the environmental situation in the city and the development of the region in general

Keywords: air pollution, AQI, EDM, ESP, combined selective forecasting model, selection problem

Andrii Biloshchytskyi
Corresponding author
Doctor of Technical Sciences, Professor,
Vice-Rector for Science and Innovation*
Department of Information Technologies**
E-mail: bao1978@gmail.com

Alexander Kuchansky
Doctor of Technical Sciences, Head of Department
Department of Information Systems and Technology
Taras Shevchenko National University of Kyiv
Volodymyrska str., 60, Kyiv, Ukraine, 01033
Department of Cybersecurity and Computer Engineering**

Yurii Andrashko
PhD, Associate Pofessor
Department of System Analysis and Optimization Theory***

Alexandr Neftissov
PhD, Associate Professor
Research and Innovation Center «Industry 4.0»*

Vladimir Vatskel
CEO
IT-LYNX LLC
Raisa Bukina str., 18, Kyiv, Ukraine, 03164

Didar Yedilkhan
PhD, Associate Professor
Department of Computer Engineering*

Myroslava Herych
PhD, Associate Professor
Department of Theory of Probability and Mathematical Analysis***

*Astana IT University
Mangilik El ave., EXPO Business Center, Block C.1,
Nur-Sultan, Republic of Kazakhstan, 010000
**Kyiv National University of Construction and Architecture
Povitroflotskyi ave., 31, Kyiv, Ukraine, 03037
***Uzhhorod National University
Narodna sq., 3, Uzhhorod, Ukraine, 88000

Received date 17.04.2022
Accepted date 31.05.2022
Published date 29.06.2022

How to Cite: *Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Neftissov, A., Vatskel, V., Yedilkhan, D., Herych, M. (2022). Building a model for choosing a strategy for reducing air pollution based on data predictive analysis. Eastern-European Journal of Enterprise Technologies, 3 (4 (117)), 23–30. doi: https://doi.org/10.15587/1729-4061.2022.259323*

1. Introduction

Air pollution and the need to improve air quality are a problem for the whole world; this issue is especially relevant for megacities. According to the World Health Organization (WHO), air pollution is the most serious risk to the environment and can have negative consequences on the health of millions of people. For example, in 2021 alone, WHO determined that about 7 million deaths, one in eight of the total number of deaths in the world, were caused by air pollution [1]. Air pollution affects the environment and human life, and leads to a significant number of adverse effects on human health, climate, and the ecosystem in general. Air pollutants, both natural and anthropogenic, can extend over long distances and cover large areas in the form of wet and dry precipitation. This is a serious risk factor for human health when inhaling or getting these pollutants into the food chain. Significant health effects are associated with both short- and long-term exposure to air pollution.

The availability of standards and systemic measures to monitor the state of air pollution, as well as the creation of a long-term strategy to improve air quality, are important tasks that need to be addressed immediately. The result of solving these problems is to reduce the risks to human health. It is important to note that this issue is most relevant to the urban environment. Along with the introduction of standards, and the formation of strategies for solving these problems, one can use the internet of things technologies and predictive data analysis. This is due to the fact that the spread of various air pollutants occurs nonlinearly, in different regions it is concentrated in different ways, which affects the development of diseases of the population living in these regions. One needs to know in advance about the change in the concentration of pollutants to plan measures to reduce the negative impact on health. The effectiveness of the measures depends on the application of data analytics.

A new trend in solving the problems of air purification is the creation of a biotechnological filter system that uses special types of plant crops that remove pollutants and release oxygen. This area is known as Biotech. The proposed solution is implemented as an air filter that absorbs fine dust and other pollutants, thereby cleaning the air next to this device. Projects in this area are relevant and funded at the government level in the UK, Germany, Portugal, the Republic of Kazakhstan, etc. [2].

Smart air purification devices can be placed in almost any part of the city. Given the significant cost of devices and the lack of serial production of significant numbers of such devices, the question of their rational use is acute. In particular, the rational location of devices can increase the efficiency of their use. That is why the study of the choice of a strategy for reducing air pollution in the urban environment is an urgent task.

2. Literature review and problem statement

Study [2] notes that the use of a biotechnological system – a biotechnological filter system or a smart cleaning device CityTree – can effectively reduce air pollution from harmful particulate matter and gases, in particular NO_2 and CO_2. One such device replaces 275 trees. At the same time, to place a smart device, one needs four orders of magnitude less area than for the area on which green spaces of the corresponding volume are placed to obtain this effect. Another significant advantage of using air purification devices is much greater flexibility in their location.

As shown in [3], air pollutants in the city are unevenly distributed in space and time. This leads to a regular excess of the permissible concentrations of certain pollutants. Such an excess poses a threat to the health of city dwellers. Arranging smart air purification devices in such areas can reduce the concentration of pollutants and save the lives and health of urban residents.

Various strategies for reducing air pollution are being investigated. One class of strategies is to increase the number of green spaces in cities, the placement of which is the task of individual research. In particular, study [4] constructed a multicriterial problem of optimal arrangement of green areas in the city based on demographic indicators. However, that model does not take into consideration the dynamics of the city development. In [5], a multicriterial model of placement optimization maximizes air purification and minimizes the heating of street surfaces and the cost of green spaces. The proposed model makes it possible to choose the optimal one from predetermined plans for the location of green areas. Specifically, the authors listed the best of 30 potential plans for some hypothetical city in South Korea. The proposed model makes it possible to take into consideration the existing restrictions on the possibility of placing green areas but requires the use of expert opinion to prepare rational plans. The result of optimization significantly depends on the qualifications of experts and the quality of their consultations in the construction of plans.

To ensure the effective choice of a strategy for reducing air pollution, it is important to use the appropriate analytical apparatus. To solve the problem of discrete optimization for the arrangement of air purification devices within the city, it is necessary to understand the trend of changing air pollution in the relevant regions of the city. To do this, one can use predictive data analysis, in particular, to solve the problem of forecasting the time series of air quality indices for each individual region. For such time series, a common trend is the creation of hybrid and selective combined models of various types. In work [6], adaptive combined models are proposed, taking into consideration the similarity of time series, and their frequency is taken into consideration at the level of models in the software set. However, for the time series of air quality indices, it is important to understand the moments of changes in trends in improving or worsening air quality. This will make it possible to plan the location of air purifiers in the relevant regions. In [7], to predict the concentration of pollutants, hybrid intelligent forecasting models are used, taking into consideration models of multicriterial optimization. The models described in [7] are used for scientific purposes, as well as for the preparation of short-term corrective plans for air quality control in cities with high levels of pollution. However, to use the models described in [7] for other cities, it is necessary to adjust the parameters of the models each time.

To adjust forecasting models with an emphasis on determining changes in trends, one can use the EMD method to decompose a signal, which is the time series of the air quality index, into the functions of empirical modes. In [8], a combination based on the EMD method was used to predict the concentration of PM_{25}. The results of the study described in [8] showed that the EMD method can be used to improve the accuracy of the forecasting model when working with complex air quality data. Paper [9] considers the combination of models based on EMD with the model of neural networks. Neural network models require parameter adjustments to ensure the calculation of the high accuracy forecast. To predict nonlinear and non-stationary time series with a similar structure as the air quality index, hybridization of the EMD method and autoregression model (AR) is used in work [10]. This combination makes it possible to get a smaller forecasting error than using the AR model without taking into consideration EMD but requires a greater amount of computing resources. Paper [11] also considers a similar combination of the EMD method and the ARMA model, which makes it possible to obtain a forecast of sufficient accuracy only in the case of careful adjustment of the model parameters.

The structure of the air quality index time series (AQI) allows the use of a class of exponential smoothing (ESM) models for its forecasting, in particular those that adapt their own parameters to the time series generation mechanism. These models include the Trigg-Leach model (TLM) or models that take into consideration seasonality, the Holt-Winters model (HWM). In [12], it is proposed to fulfill the forecast of the time series of the air quality index based on the HWM model. However, for the effective use of this model for the time series discussed in this study, it is necessary to adjust the parameters of the model. Taking into consideration the selective principle of selection of forecasting models in combinations is described in [13]. The information technology that implements the principle of calculation based on combined forecasting models is described in [14]. Consequently, the study of combined selective type models for forecasting the time series of air quality indices based on the EMD method, and the class of exponential smoothing models with parameter adaptation is a promising direction. It will solve the problem associated with the choice of a strategy to reduce air pollution. The results of the study could serve as a theoretical and practical basis for the development of new combined models for predicting nonstationary time series.

3. The aim and objectives of the study

The aim of this study is to develop a model for choosing a strategy for reducing air pollution in the urban environment based on an analysis of air condition monitoring indicators. This will make it possible to improve the efficiency of management of the city's environmental policy.

To accomplish the aim, the following tasks have been set:

– to formalize the task of choosing a strategy for reducing air pollution in the urban environment;

– to develop a combined model for forecasting air quality indicators based on the results of monitoring at stationary observation stations and to verify the model.

4. The study materials and methods

Our paper explores the application of biotechnology and Internet of Things technologies, as well as data analysis for environmental tasks, in particular for air purification in a small location using a biotechnological system – a biotechnological filter system. Fig. 1 depicts the prototype of the system that we designed as one intermediate result on the topic «Development of the intelligent information and telecommunication systems for municipal infrastructure: transport, environment, energy and data analytics in the concept of Smart City».

The main hypothesis of this study assumes that the formation of strategies for reducing air pollution in urban environments, based on the analysis of air monitoring indicators, is effective.

To formalize the parameters of the model of choosing a strategy for reducing air pollution in the urban environment, the theory of discrete optimization is used. When constructing a combined model for forecasting air quality indicators based on the results of monitoring at stationary observation stations, models for analyzing time series were applied.

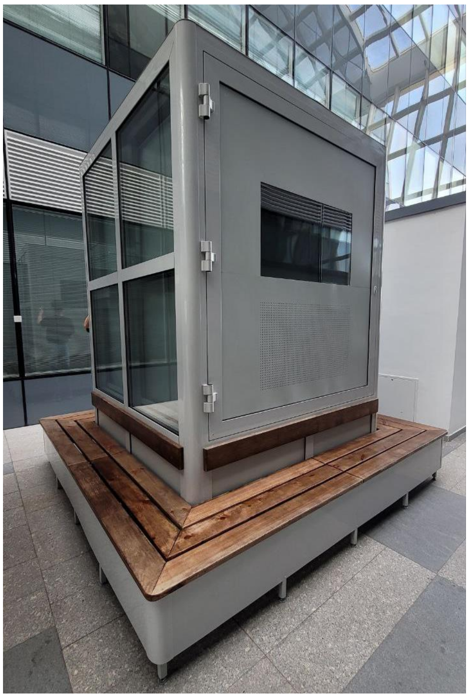

Fig. 1. The prototype of the biological system, a filter for air purification in an urban environment

The calculation of the air quality index (AQI) is carried out on the basis of air condition monitoring data based on air condition monitoring data in the city of Nur-Sultan (Republic of Kazakhstan): Astana Air Quality Dataset. Monitoring of the concentration (mg/cubic meter) of dust, sulfur and nitrogen dioxides, sulfates, carbon monoxide, and fluoride is carried out every 6 hours at 4 stations. Temperature, wind speed (m/s), wind direction, atmospheric pressure, etc. were also monitored. Data collection has continued from January 4, 2010, to the present. Pre-processing of data was carried out to exclude emissions that do not fall within the confidence interval 3σ. We also consider data without gaps. In the case of a gap fixed arising from a failure in the monitoring station, the missing values are replaced by the average values of the previous and subsequent observations. The percentage of emissions and omissions is 1.76 % of the total sample size and the emergence of two or more gaps of values in a row was not recorded.

Astana Air Quality Dataset air pollution monitoring data are available for on-demand research. The data analysis was conducted in the Jupyter Notebook environment using the Python programming language and Pandasta NumPy libraries.

During the research, a number of simplifications and assumptions were introduced. One of these simplifications is the uniformity of the spread of pollutants in the air. In addition, climatic factors such as the direction and strength of the wind, and the presence of precipitation are not taken into consideration.

5. Results of studying the construction of a model for choosing strategies and forecasting air quality indicators

5. 1. Formalization of the problem of choosing a strategy for reducing air pollution in the urban environment

The concept of reducing air pollution in an urban environment based on data analysis methods and IoT technologies is proposed. This study considers the approach that among the possible strategies to reduce air pollution in the region those will be selected that include the arrangement of biotechnological filter systems and monitoring of pollution in a given region.

We formalize the task of choosing a strategy to reduce air pollution in the urban environment using smart devices for air purification. Suppose one needs to place K smart devices to clean the air in a particular city. All devices are identical, and their characteristics are known and stable. The state of the air is determined by the concentration of pollutants ($P_1, P_2, ..., P_v$). Determine the functionality of air purification with one device C. Let the input of a smart air purification device be supplied with air containing pollutants of the appropriate concentration ($P_1, P_2, ..., P_v$). At the output, the device supplies air containing the $P_i - C(P_i)$ concentration of pollutants, $i = \overline{1, v}$. Then the air condition when using K devices can be described using the formula:

$$\widehat{P}_i = P_i - KC\left(P_i\right),$$

where $\widehat{P}_i$ is the concentration of the pollutant after air purification, $i = \overline{1, v}$.

It should be noted that such an assessment of the state of the air is fair only in a certain limited area around a smart device for air purification.

To assess air pollution, we introduce a function to find the air quality index (AQI) as an integral function of the concentration of pollutants $f(P_1, P_2, ..., P_V)$.

Consider ($t_1, t_2, ..., t_n$) moments that are placed evenly on the time axis. At each given time t_i, monitoring points ($R_1, R_2, ..., R_x$) monitor the concentration of various pollutants ($P_1, P_2, ..., P_v$). Let the city O be given, which is covered with a grid with cells ($o_1, o_2, ..., o_y$). In each cell of the grid, we calculate the pollution according to the model of the inverse distance weighting (IDW) [15]. As a result, we obtain multidimensional time series with geographical location ($P_{11}(t)$, $P_{12}(t)$, ..., $P_{1v}(t)$, $P_{21}(t)$, $P_{22}(t)$, ..., $P_{2v}(t)$, ..., $P_{yv}(t)$).

For each cell, calculate the air quality index (AQI) at each point in time. As a result, we define the time series:

$$z\left(t, o_j\right) = f\left(P_{j1}\left(t\right), P_{j2}\left(t\right), \ldots, P_{jv}\left(t\right)\right),\ j = \overline{1, y},$$

where $z(t, o_j)$ is the air quality index in cell o_j before cleaning.

If the air purification device is placed in the cell, the air quality index increases:

$$\tilde{z}\left(t, o_j\right) = z\left(t, o_j\right) + \Delta\left(z\left(t, o_j\right)\right),$$

where $\tilde{z}(t, o_j)$ is the air quality index in the cell o_j after cleaning, $\Delta(z(t, o_j))$ is the increase in the air quality index due to cleaning by one device.

Place in cells (o_1, o_2, ..., o_y) smart devices for air purification. Let K_1, ..., K_y be the number of smart devices for air purification that are placed in cells (o_1, o_2, ..., o_y). Then the choice of a strategy to reduce air pollution in the urban environment based on the use of smart devices is associated with the solution to the problem of maximizing air quality at times $n+1$, $n+2$, ..., $n+\tau$.

$$\sum_{j=1}^{y} \sum_{t=n+1}^{n+\tau} z\left(t, o_j\right) + K_j \cdot \Delta\left(z\left(t, o_j\right)\right) \to \max, \tag{1}$$

$$\sum_{j=1}^{y} K_j = K. \tag{2}$$

To solve problems (1), and (2), one needs to solve the following two subtasks:

1. Calculate the forecast of the time series of air quality indices at time $n+1$, $n+2$, ..., $n+\tau$.
2. Determine the increase $\Delta(z(t, o_j))$ taking into consideration the technical characteristics of the corresponding devices in the o_j cell and estimates of the forecast of the time series of the air quality index.

In this study, the first subtask is solved.

5. 2. Development and validation of the combined model for forecasting air quality indicators based on monitoring results (CSM)

Consider the combined model of forecasting air quality indicators for one cell o_j. For other cells, the model is used similarly with the corresponding parameter setting.

Let the time series $z\left(t\right) = \{z_t\}_{t=1}^{n}$, be set, based on the results of monitoring the state of the air at stationary observation stations at fixed moments of time t. The series $z(t)$ consists of the values of the air quality index (AQI), which are obtained by convoluting air quality indicators based on monitoring indicators of various indicators of pollution or taking into consideration modeling of atmospheric dispersion. Visually, the $z(t)$ series for the urban environment is quasi-periodic in nature with an increase in rush hour levels, due to increased pollutant emissions, and with a decrease in levels at night. However, for large agglomerations, the $z(t)$ time series quasi-period may not be visually fixed.

The z_M level of the time series $\{z_t\}_{t=1}^{n}$, $z_M \in \{z_t\}_{t=1}^{n}$, is called the local maximum of a given time series if the condition $z_{M+b} > z_{M+b+1}, z_{M-a} > z_{M-a-1}$ for $a = \overline{0, \alpha - 1}$, $b = \overline{0, \beta - 1}$ is met. The z_m level of the time series $\{z_t\}_{t=1}^{n}$, $z_m \in \{z_t\}_{t=1}^{n}$, is called the local minimum of a given time series, if $z_{m+b} < z_{m+b+1}$, $z_{m-a} < z_{m-a-1}$ for $a = \overline{0, \alpha - 1}$, $b = \overline{0, \beta - 1}$. The right arm for the local maximum point z_M is a series there is a number of β for the right shoulder point for the local maximum point z_M is the $\{z_M, z_{M+1}, \ldots, z_{M+\beta}\} = \{z_t\}_{t=M}^{M+\beta}$, series. β is the power of the right arm. The left arm for the point of local maximum z_M is the $\{z_M, z_{M-1}, \ldots, z_{M-\alpha}\} = \{z_t\}_{t=M}^{M-\alpha}$, series, α is the power of the right arm. The time series $\{z_t\}_{t=1}^{n}$, levels consisting of local maximum points with the left and right arm powers, respectively, α and β form a time series of local maximums $\{\widehat{z}_{t_j}^{\alpha,\beta}\}_{j=1}^{s}$, where the $\widehat{z}_{t_j}^{\alpha,\beta}$ levels satisfy the condition for a local maximum point with the arms α and β, $\{\widehat{z}_{t_j}^{\alpha,\beta}\}_{j=1}^{s} \subset \{z_t\}_{t=1}^{n}$.

Similarly, the right and left arms are formed for the local minimum point z_m. Then the $\{z_t\}_{t=1}^{n}$, time series levels, which consist of local minimum points with the powers of the left and right arms, α and β, respectively, form a time series of local minimums $\{\breve{z}_{t_j}^{\alpha,\beta}\}_{j=1}^{c}$, where the $\breve{z}_{t_j}^{\alpha,\beta}$ levels satisfy the condition for the local minimum point with the arms of the α and β, $\{\breve{z}_{t_j}^{\alpha,\beta}\}_{j=1}^{c} \subset \{z_t\}_{t=1}^{n}$. Any point in the series $\{z_t\}_{t=1}^{n}$ can be a local minimum point, a local maximum, or a regular point that does not meet the conditions for the local minimum or local maximum. Let's assume that the points of the local minimum and the local maximum alternate and their number is equal, $c=s=q$.

We use the EMD method to decompose the time series into empirical modes. Empirical mode (IMF) is a function for which the number of local minimums and maximums differs by no more than unity, and the average value on the upper and lower envelopes should be zero. The EMD method consists of the following steps:

Stage 1. For the time series $\{\widehat{z}_{t_j}^{1,1}\}_{j=1}^{q}$, we construct functions that are third-order polynomials for each of the segments $[t_{j-1}, t_j]$, $j = \overline{2, q}$:

$$\nu_j(t)=\alpha_j+\beta_j(t-t_j)+\gamma_j(t-t_j)^2+\zeta_j(t-t_j)^3, \tag{3}$$

$$\nu_j(t_j)=\alpha_j,\ \nu'_j(t_j)=\beta_j,$$

$$\nu''_j(t_j)=2\gamma_j,\ \nu'''_j(t_j)=6\zeta_j.$$

We shall find unknown polynomial coefficients and eventually build a cubic spline $\nu_j(t)$, given that the function $\nu(t)$ is twice continuously differentiable and the spline constructed will interpolate the function at points that are the $\{\hat{z}^{1,1}_{t_j}\}_{j=1}^{q}$ time series levels. Similarly, we shall build a cubic spline $\mu_j(t)$, which will interpolate the function at points that are the $\{\breve{z}^{1,1}_{t_j}\}_{j=1}^{q}$ time series levels.

Stage 2. Find the average for functions $\nu(t)$ and $\mu(t)$, $M(t)=(\nu(t)+\mu(t))/2$.

Stage 3. Let's highlight the difference $h(t)=z(t)-M(t)$.

Stage 4. Stages 1–3 are repeated until the empirical mode condition is met. According to the results of the stage, we get the first empirical mode for a fixed t, that is, $f_1(t)=h(t)$. The remainder in the first step $r_1(t)=z(t)-f_1(t)$. In the following steps, $r_k(t)=r_{k-1}(t)-f_k(t)$. Repeat stages 1–4, taking into consideration the residues according to the sifting procedure described in [16] until zero averages are provided or other stop criteria are met. As a result, we get the function of empirical modes and the input time series will be represented in the form:

$$\bar{z}(t)=\sum_{k=1}^{w} f_k(t)+r_w(t), \tag{4}$$

where $r_w(t)$ is the last residue, w is the number of empirical modes identified.

The process of constructing empirical modes is considered complete if the sifting process stops. Empirical mode is constructed if the number of local minimums and maximums differs by no more than unity, and the average value between the upper and lower envelopes is close to zero according to a certain criterion. In [17], it is proposed to introduce two threshold values θ_1 and θ_2. The $\sigma(t)$ function evaluation is defined as:

$$\sigma(t)=\left|\frac{2M(t)}{\nu(t)-\mu(t)}\right|, \tag{5}$$

moreover, the sifting procedure continues until $\sigma(t)<\theta_1$ for a certain part of the calculation duration $(1-\varepsilon)$ and $\sigma(t)<\theta_2$ for the rest of the part. In [17], the following parameters are suggested: $\varepsilon\approx 0.05$, $\theta_1\approx 0.05$, $10\cdot\theta_1\approx\theta_2$.

Taking into consideration the structure of the time series obtained based on the results of the EMD method, we use the exponential smoothing model class (ESM) to construct a $\{z_t\}_{t=1}^{n}$ time series forecast. Denote via $\hat{z}_\tau(n)$ the estimation of the $\{z_t\}_{t=1}^{n}$, time series forecast performed at point z_n with a period of τ, that is, $\hat{z}_\tau(n)=z_{n+\tau}$.

We shall build a combined selective forecasting model, which will be based on the following three models:

– The EMD-ESM Hybrid Model. Smoothing the time series (2), built according to the EMD method according to the exponential model of zero order, we obtain the following estimate of the forecast:

$$\hat{z}_1^1(n)=\xi_n, \tag{6}$$

$$\xi_n=\chi\bar{z}_n+(1-\chi)\bar{z}_{n-1},\ \xi_1=\bar{z}_1,$$

where $\hat{z}_1^1(n)$ is the prediction at point z_n with period 1 obtained under the EMD-ESM hybrid model, $\chi\in[0,1]$.

To construct a long-term forecast with a horizon of $\tau>1$, the calculated value $\hat{z}_1(n)$ is taken as the last value of the time series $z_{n+1}=\hat{z}_1(n)$ and the forecast is made according to the described model (4) at this point z_{n+1} with a period of 1. As a result, we obtain a forecast with a horizon of $\tau=2$. One can also use the exponential smoothing model of higher orders for this purpose.

– HWM model (the additive model by Holt-Winters). This model is used for processes with an additive trend component and multiplicative seasonality. Let q_i be the assessment of the seasonal component of the series $\{z_t\}_{t=1}^{n}$, and L is the period of the seasonal cycle, then the HWM model for the series $\{z_t\}_{t=1}^{n}$:

$$\hat{z}_1^2(n)=(s_n+m_n)\cdot q_{n-L+1}, \tag{7}$$

$$s_n=\alpha\frac{z_n}{q_{n-L}}+(1-\alpha)(s_{n-1}+m_{n-1}),$$

$$m_n=\beta(s_n-s_{n-1})+(1-\beta)m_{n-1},$$

$$q_n=\gamma\frac{z_n}{s_n}+(1-\gamma)q_{n-L},$$

where $\gamma\in[0,1]$ is the seasonal component smoothing parameter, $\alpha\in[0,1]$, $\beta\in[0,1]$, $\hat{z}_1^2(n)$ is the prediction at point z_n with period 1 obtained from the HWM model.

– TLM model (Trigg-Leach adaptive model). The Trigg-Leach model makes it possible to dynamically adjust the value of the smoothing parameter depending on the forecasting error that is obtained in the previous step:

$$\hat{z}_1^3(n)=s_n, \tag{8}$$

$$s_n=\alpha_n\cdot z_n+(1-\alpha_n)s_{n-1},\ \alpha_n=\left|\frac{d_t}{g_t}\right|,$$

$$d_n=\beta e_n+(1-\beta)d_{n-1},\ g_n=\beta|e_n|+(1-\beta)g_{n-1},$$

where $\alpha_n\in[0,1]$ is the smoothing parameter at n, $\beta\in[0,1]$, $e_n=z_n-\hat{z}_n$, $\hat{z}_1^3(n)$ is the prediction at point z_n with period 1 obtained from the adaptive TLM model.

The combined selective forecasting model. Let each model calculate the forecasts $\hat{z}_\tau^p(n)$ – forecasts at point z_n with a period of τ, calculated according to the hybrid model EMD-ESM (6), additive model HWM (7), and adaptive model TLM (8), $p=\overline{1,3}$. Then the selective combination of models involves the choice to calculate the forecast of the model that is most accurate in the current time series. The adaptability of the combined model is ensured by calculating the forecasting error for each model that is in its software set and comparing the estimate:

$$\lambda_n^p=\delta e_\tau^p(n-\tau)+(1-\delta)\lambda_{n-1}^p, \tag{9}$$

where $e_\tau^p(n-\tau)=|z_n-\hat{z}_\tau^p(n-\tau)|$ is the absolute forecasting error calculated at time $(n-\tau)$ with a period of τ, $0<\delta\leq 1$, $p=\overline{1,3}$.

At each point z_n, when the forecast is executed, the model for which estimate (9) is minimal is selected. That is,

$$\hat{z}_\tau^*(n)=\hat{z}_\tau^g(n),\ g=\arg\min_{p=\overline{1,3}}\{\lambda_n^p\}.$$

To verify the proposed combined selective model, the time series of air quality indices were selected, which are calculated for the city of Nur-Sultan (Republic of Kazakhstan), Astana Air Quality DataSet. The data were selected from

2010 to 2021 The data were recorded four times a day at five observation stations (more than 15 thousand points).

The EMD-ESM hybrid model (6), the HWM additive model (7), and the adaptive TLM (8) model, as well as their selective type combination (9), were implemented. To calculate the forecast error, an experimental section of the time series $\{z_t\}_{t=u}^{n}$, was selected, $U<u<n$. U is chosen taking into account the adaptation of models of exponential smoothing to the mechanism of time series generation, theoretically, it should be $U>40$, in the verification process the value $U=100$ was chosen. We calculated the standard deviation (RMSE):

$$RMSE_\tau=\sqrt{\frac{1}{n-u+1}\sum_{i=0}^{n-u}\left(\overset{\frown}{z}{}_{\tau}^{p}\left(u+i\right)-z_{u+i+\tau}\right)^2}.$$

The derived average standard deviations (Table 1, Fig. 2) show that the use of the combined selective model (CSM) allows for higher accuracy compared to the models included in the software set, for $\tau>3$. The software suite includes the EMD-ESM hybrid model (6), the HWM additive model (7), and the adaptive TLM model (8)).

Table 1

Average standard errors of the forecast of the time series of air quality with a period $\tau=\overline{1,5}$ for built models

Model \ τ	1	2	3	4	5
EMD-ESP	0.0207	0.0314	0.0506	0.0989	0.1563
HWM	0.0311	0.0462	0.0628	0.0917	0.1388
TLM	0.0515	0.0581	0.0733	0.1196	0.1859
CSM	0.0365	0.0426	0.0529	0.0715	0.1107

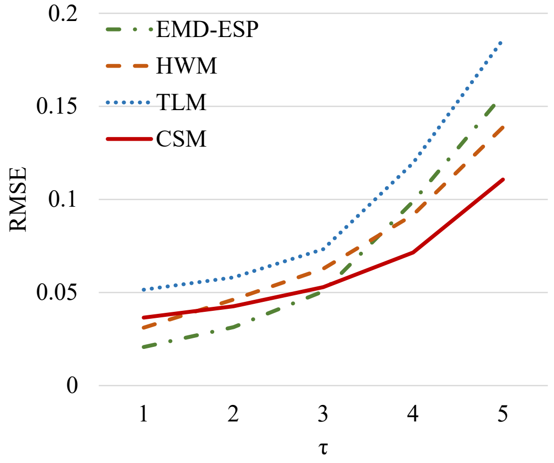

Fig. 2. Average root mean square errors (RMSE) of the air quality time series with a period $\tau=\overline{1,5}$ for constructed models

For small values $\tau\leq3$, the EMD-ESM hybrid model has an advantage based on RMSE. Taking into consideration the need to calculate the long-term forecast for the problem of forecasting air quality indicators, it is proposed to form time series taking into consideration the forecast period, thinning the value of the input series.

6. Discussion of the choice of parameters for the forecasting model and the choice of strategies for reducing air pollution

An important issue in ensuring the accuracy of the forecast is the choice of parameters for the described models. The choice of parameters for EMD-ESP, HWM, and TLM models is determined by the theoretical estimates from respective authors or is adjusted in the forecasting process, adapting to the structure of the time series. Then, to select the parameter δ in the combined selective model, thoroughness is required because the final accuracy of the forecast depends on it. The value of the δ parameter is determined by the type of time-series dynamics. The more persistent the input time series $\{z_t\}_{t=1}^{n}$, is, the error of forecasting $e_\tau^p\left(n-\tau\right)$ according to the described models for $p=\overline{1,3}$ will be less. Accordingly, in this case, one needs to reduce the δ parameter for (9). If the time series is close to random, then the errors $e_\tau^p\left(n-\tau\right)$ will increase, and, therefore, it is necessary to take them into consideration in adjusting the forecast according to the combined selective model, increasing the δ parameter for (9). The level of persistence, randomness, and anti-persistence of the time series is determined on the basis of fractal R/S-analysis. In addition, for an approximate assessment of the δ parameter in the combined selective model, one can use the ratio of the product of local maximum points to local minimum points with empirical arm power $\alpha\geq3$, $\beta\geq3$. To build a forecast of the level of air pollution, one can use the combined approach described in work [18].

Let the time series $\{\overset{\frown}{z}{}_{t_j}^{3,3}\}_{j=1}^{q}$ and $\{\breve{z}_{t_j}^{3,3}\}_{j=1}^{q}$ be constructed for the time series $\{z_t\}_{t=1}^{n}$. Then the evaluation of the δ parameter in the combined selective model for (9) is determined from the following formula:

$$\delta=\left(\prod_{j=1}^{q}\frac{\overset{\frown}{z}{}_{t_j}^{3,3}}{\breve{z}_{t_j}^{3,3}}\right)^{-1}.$$

For the series close to random, such an inverse ratio δ is close to unity, $\delta>1$. For persistent series, the δ value increases, $\delta\gg1$. Hearst's indicator for the input time series of the air quality index H=0.657 (calculated on the basis of normalized R/S-analysis), δ=0.11.

The result of our study is the constructed model for choosing strategies for reducing air pollution in the urban environment, solving which requires solving two subtasks. One of them can be solved by a combined model for forecasting the time range of air quality indicators.

The main limitation of this study is that for the use of a combined model for predicting the time series of air quality indices based on exponential smoothing models, it is necessary to have sufficient volume of data to adjust the parameters of the models.

A significant disadvantage of the study is that when predicting air quality indicators, meteorological indicators are not taken into consideration. Determining the increase in air quality, taking into consideration the technical characteristics of air purification devices and estimates of the time series of the air quality index, is not addressed in this study.

To apply the proposed model of choosing a strategy for reducing air pollution and assessing its effectiveness, it is necessary to obtain sequential solutions to two subtasks. One of them is solved in this study. To verify the second subtask, one needs to develop several biotechnological filter systems and test their work using an example of a particular region. That will be part of further research.

A potential effect is to reduce air pollution in a particular region and improve the health of the region's population, but this is a separate task of the study, which requires a long time and additional experiments.

We see further development of the study in two directions. In particular, the task of increasing air quality, taking into consideration the technical characteristics of air purification devices, requires additional research and will be considered in subsequent studies. The second area to advance our study is to supplement the combined selective model with additional parameters, in particular meteorological data.

7. Conclusions

1. The task of choosing a strategy for reducing air pollution in the urban environment has been formalized. A model has been built that provides for the optimal placement of air purification devices based on the results of solving the problem of discrete optimization, taking into consideration the forecast of the air quality index. Two subtasks that need to be solved have been formalized. To apply the proposed model of choosing a strategy to reduce air pollution, one needs to derive solutions to both subtasks. The problem of determining the growth $\Delta(z(t, o_j))$ taking into consideration the technical characteristics of the corresponding devices in the o_j cell and the estimates of the forecast of the time series of the air quality index will be solved in further studies.

2. To solve the problem of choosing strategies for reducing the air pollution in the urban environment, a combined selective model (CSM) for forecasting the time series of the air quality index was built. The combined model software suite consists of the EMD-ESM hybrid model (6), the HWM additive model (7), and the adaptive TLM model (8). To verify the proposed combined selective model, the time series of air quality indices were selected for the city of Nur-Sultan, Astana Air Quality DataSet. The data were selected from 2010 to 2021. The data were recorded four times a day at five observation stations (more than 15 thousand points). As a result of verification, it was established that in the case of short-term forecasting of the air quality index time series, the model EMD-ESM has an advantage according to the criterion of minimum root mean square error (RMSE). For medium-term forecasting $\tau \leq 3$, the EMD-ESM hybrid model has an advantage in terms of RMSE, with a parameter $\delta = 0.11$. Hearst's indicator is $H = 0.657$, for the air quality index time series (calculated on the basis of normalized R/S analysis). For the case of medium-term forecasting, $3 < \tau \leq 5$, the combined selective model (CSM) has the advantage. Our results are input data for the task of choosing strategies for reducing the volume of air pollution in the urban environment.

Acknowledgments

This research was funded by the Scientific Committee of the Ministry of Education and Science of the Republic of Kazakhstan under the program «Program-targeted financing of subjects of scientific and/or scientific and technical activities at the expense of the republican budget» on the topic «Development of the intelligent information and telecommunication systems for municipal infrastructure: transport, environment, energy and data analytics in the concept of Smart City», Grant No. BR10965311.

References

1. Roser, M. (2021). Data Review: How many people die from air pollution? Our World In Data. Available at: https://ourworldindata.org/data-review-air-pollution-deaths
2. CityTree: a Pollution Absorbing Innovation with the Power of 275 Trees (2018). Green City Solutions. Available at: https://urbannext.net/citytree/
3. Ung, A., Wald, L., Ranchin, T., Weber, C., Hirsch, J., Perron, G., Kleinpeter, J. (2002). Satellite data for the air pollution mapping over a city –The use of virtual station. In Proceedings of the 21th EARSeL Symposium, Observing our environment from space: new solutions for a new millenium, Paris, 147–151. Available at: https://www.researchgate.net/publication/42433064_Satellite_data_for_the_air_pollution_mapping_over_a_city_-_The_use_of_virtual_stations
4. Nyelele, C., Kroll, C. N. (2021). A multi-objective decision support framework to prioritize tree planting locations in urban areas. Landscape and Urban Planning, 214, 104172. doi: https://doi.org/10.1016/j.landurbplan.2021.104172
5. Yoon, E. J., Kim, B., Lee, D. K. (2019). Multi-objective planning model for urban greening based on optimization algorithms. Urban Forestry & Urban Greening, 40, 183–194. doi: https://doi.org/10.1016/j.ufug.2019.01.004
6. 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. doi: https://doi.org/10.15587/1729-4061.2018.121620
7. Heydari, A., Majidi Nezhad, M., Astiaso Garcia, D., Keynia, F., De Santoli, L. (2021). Air pollution forecasting application based on deep learning model and optimization algorithm. Clean Technologies and Environmental Policy, 24 (2), 607–621. doi: https://doi.org/10.1007/s10098-021-02080-5
8. Huang, G., Li, X., Zhang, B., Ren, J. (2021). PM2.5 concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition. Science of The Total Environment, 768, 144516. doi: https://doi.org/10.1016/j.scitotenv.2020.144516
9. Huang, Y., Yu, J., Dai, X., Huang, Z., Li, Y. (2022). Air-Quality Prediction Based on the EMD-IPSO-LSTM Combination Model. Sustainability, 14 (9), 4889. doi: https://doi.org/10.3390/su14094889

10. Duan, W., Huang, L. (2016). A hybrid EMD-AR model for nonlinear and non-stationary wave forecasting. Journal of Zhejiang University Science A, 17, 115–129. doi: https://doi.org/10.1631/jzus.a1500164
11. He, K., Zha, R., Wu, J., Lai, K. (2016). Multivariate EMD-Based Modeling and Forecasting of Crude Oil Price. Sustainability, 8 (4), 387. doi: https://doi.org/10.3390/su8040387
12. Wu, L., Gao, X., Xiao, Y., Liu, S., Yang, Y. (2017). Using grey Holt–Winters model to predict the air quality index for cities in China. Natural Hazards, 88 (2), 1003–1012. doi: https://doi.org/10.1007/s11069-017-2901-8
13. Kuchansky, A., Biloshchytskyi, A. (2015). Selective pattern matching method for time-series forecasting. Eastern-European Journal of Enterprise Technologies, 6 (4 (78)), 13. doi: https://doi.org/10.15587/1729-4061.2015.54812
14. Mulesa, O., Geche, F., Batyuk, A., Buchok, V. (2017). Development of Combined Information Technology for Time Series Prediction. Advances in Intelligent Systems and Computing, 361–373. doi: https://doi.org/10.1007/978-3-319-70581-1_26
15. Gelfand, A. E., Diggle, P., Guttorp, P., Fuentes, M. (Eds.) (2010). Handbook of Spatial Statistics. CRC Press, 619. doi: https://doi.org/10.1201/9781420072884
16. Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q. et. al. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 454 (1971), 903–995. doi: https://doi.org/10.1098/rspa.1998.0193
17. Rilling, G., Flandrin, P., Gonçalves, P. (2003). On empirical mode decomposition and its algorithms. Proceedings of IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing. NSIP-03. 3.
18. Kuchansky, A., Biloshchytskyi, A., Andrashko, Y., Vatskel, V., Biloshchytska, S., Danchenko, O., Vatskel, I. (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). doi: https://doi.org/10.1109/idaacs-sws.2018.8525608