Forecast Accuracy Evaluation of the Enterprise's Industrial Safety Integral Risk

Leyla M. Bogdanova [1], Sergey Ya. Nagibin [2], Aleksandr S. Chemakin [1]

[1] *Plekhanov Russian University of Economics, Stremyanny lane 36, Moscow, 117997, Russia*
[2] *Moscow Aviation Institute (National Research University), Volokolamskoe Highway, 4, Moscow, 125993, Russia*

Abstract **– Autoregressive models represent a time series as a linear dependence of the current value on the retrospective ones. Their feature is the mathematical and statistical base and formalization of the requirements for the parameters' selection, which makes them relevant and effective. The article describes an algorithm for analyzing time series representing changes in the integral risk indicator and its modeling using various autoregressive models with subsequent comparison of their adequacy and quality evaluation of the resulting forecast. It is shown that with the help of this class models, it is possible to build a forecast for a time period sufficient to make a decision on preventing accidents at complex infrastructure facilities.**

Keywords **– integral risk indicator, time series forecasting, industrial safety, mathematical modeling, time series analysis, risk-based approach, forecasting results evaluation.**

1. Introduction

When determining the methods and frequency of inspection (technical surveys) of objects working under pressure, as well as their maintenance, repair and other measures, two approaches can be conventionally distinguished:

- traditional (or prescriptive) is adopted in a number of countries, including Russia, based on the strict regulatory requirements fulfillment for the timing and methods of inspection;
- risk-oriented, which takes into account the technical devices' actual state and factors influencing the risk of their failure.

In the first approach, the inspection is associated with shutdown, run of equipment, tanks' inspection, which accelerates run-out and increases the risk of depressurization. In addition, the shutdown of the equipment entails losses from the enterprise's downtime and the government's budget revenues from the taxation of missed profits lost to the state. At the same time, this approach does not guarantee a reduction in the accident risk, but it can lead oil and gas companies to significant financial costs [1].

The second approach implementation requires the use of modern inspection methods based on monitoring, which involves the identification, analysis and prediction of industrial accidents hazards, risk evaluation and possible scale of accidents consequences in real time, to organize the necessary measures to prevent them, to prevent the threats' emergence of major industrial accidents and increasing the ensuring industrial safety's efficiency at a separate HPF and / or in the system of supervised facilities as a whole [2].

Different countries have different approaches to hazardous factors in production, while a systematic approach based on Process Safety Management, Dangerous goods control and Safety Engineering concept science-intensive prevails. In Russia, the industrial safety is designed to maintain protection from accidents and their consequences, the main concept of which is HPF [3], [4].

The risk's role in the HPFs' activities, which work is associated with the possibility of causing environmental damage, material and human losses, is very great. The need for risk management today is a complex urgent problem due to the risk management's holistic theory lack, as well as the ambiguity of using various risk evaluation methods

DOI: 10.18421/TEM101-06
https://doi.org/10.18421/TEM101-06

Corresponding author: Sergey Ya. Nagibin,
Moscow Aviation Institute (National Research University), Moscow, Russia.
Email: nagibinmai@mail.ru

Received: 14 September 2020.
Revised: 07 December 2020.
Accepted: 26 December 2020.
Published: 27 February 2021.

in building a risk management system, which is a relatively new direction in economic science.

Under risk, there is meant the probability of an event occurring that can have a negative impact on the set goals achievement and entail damage. The risk is in direct proportion to the objectivity and validity of the management decisions taken.

Therefore, the development of methods and tools for predicting and evaluating industrial safety risks is relevant and has scientific and great practical interest.

Industrial accidents have their own characteristic features, the main ones being the comparative rarity of accidents in comparison with the production life cycle and significant range of consequences. The danger of industrial accidents is usually evaluated by severe damage possibility or threat of their causing.

The modern accident risk's quantitative evaluation as a tool for analyzing industrial hazards was mainly formed in the West and in the USSR in the 1970s-1980s as a reaction to the major accidents phenomenon that swept through the industrialized and developing countries during these years.

Attempts to apply the well-known and well-developed methods of reliability theory to evaluate the rare unique events' frequency, and the theory of probability to determine the accident damage's random values in complex technical and social systems, have not brought satisfactory results. Therefore, for example, the theory of reliability operates with a random time value between successive failures, for "unique" accidents (catastrophes) this value tends to infinity. In addition, the accidents' causes are not only equipment failures, but also poorly formalized human errors and poorly predictable external influences.

Monitoring and risks' evaluation associated with the equipment's reliability that makes up the HPF process chain becomes especially important at the operation stage, since they significantly depend on the equipment's maintenance quality (repair), used consumables, catalysts and physical and chemical properties of the extracted / processed raw materials that can change over time. In addition, equipment (technical devices) during operation is subject to mechanical wear, environmental impact and aggressive processed materials (corrosion, erosion, etc.), which can affect their operation reliability.

Thus, the enterprise's industrial safety of complex technological processes is influenced by many factors of both external and internal nature, which, for a generalized evaluation, must be brought together into a single integral risk indicator [5].

Under the integral risk indicator there is meant the metric that is used to evaluate the risks to which HPF is exposed. The integral risk indicator formation is an independent and rather complex scientific and practical task. The integral risk metric should take into account the parameters most critical for the production business process, taking into account their weights that affect its evaluation. Moreover, this list of parameters cannot be too large, since the evaluation of industrial safety risk and its forecast should be carried out in real time, so that if the risk probability exceeds the standard level, there is time for making decisions and organizing measures to counter an emergency or catastrophic situation.

The data collection from all technical and production systems of the HPF landscape, preliminary processing and transformation of them into information, storage, analysis and transmission of a complex solution to control objects is entrusted to the intelligent monitoring systems. Under intellectual monitoring is meant not only the process of observing and registering the object's parameters, in comparison with the specified criteria for judging both the individual elements' state that make up the object's landscape, and about the object as a whole, based on the analysis of its characteristics. Such monitoring is designed not only to observe the object's state, but also to predict changes in its state to determine and predict the failure moment, both of its individual elements, and the transition to a critical state of the entire object, including the prevention of emergency situations leading to damage or object's destruction. In this case, the monitoring purpose is to prevent the negative event onset, and not to simply inform about an accident or catastrophe. At the stage of analyzing the data obtained, in this case, a predictive mathematical model is constructed that describes the indicators' behavior that depends on time, the so-called time series. This process makes it possible to evaluate the data dynamics and build confidence intervals for admitting deviations of the integral risk indicator's obtained actual values of the minimum and / or the maximum calculated ones. Forecasting, as an integral part of intelligent monitoring systems, provides information to prevent the occurrence of emergency situations at HPFs.

This article describes the process of constructing the integral risk indicator's autoregressive mathematical model of complex infrastructure facility.

2. Methodology

The task of the integral risk indicator forecasting is to determine the metrics' future values under consideration using machine learning methods based on historical data. In what follows, the integral risk metrics' obtained values will be called signals for simplicity. In general, the process of building a mathematical model includes the following steps:

1. considered signal analysis;
2. time series' main components determination;
3. building a forecast mathematical model;
4. forecast accuracy evaluation of the risk indicator.

2.1. Considered Signal Analysis

To build a model that accurately describes the signal's behavior, it is necessary to analyze the input data to establish the number of characteristics' values. In this work, the considered signal is the integral risk indicator, which mathematical model is the time series $\{u_i\}_{i=1}^{N}$ of the dynamic variable $u(t)$ [6]:

$$u_i = u(t_i), i = \overline{1,N}, \quad (1)$$

with current values' regression dependence from retrospective ones.

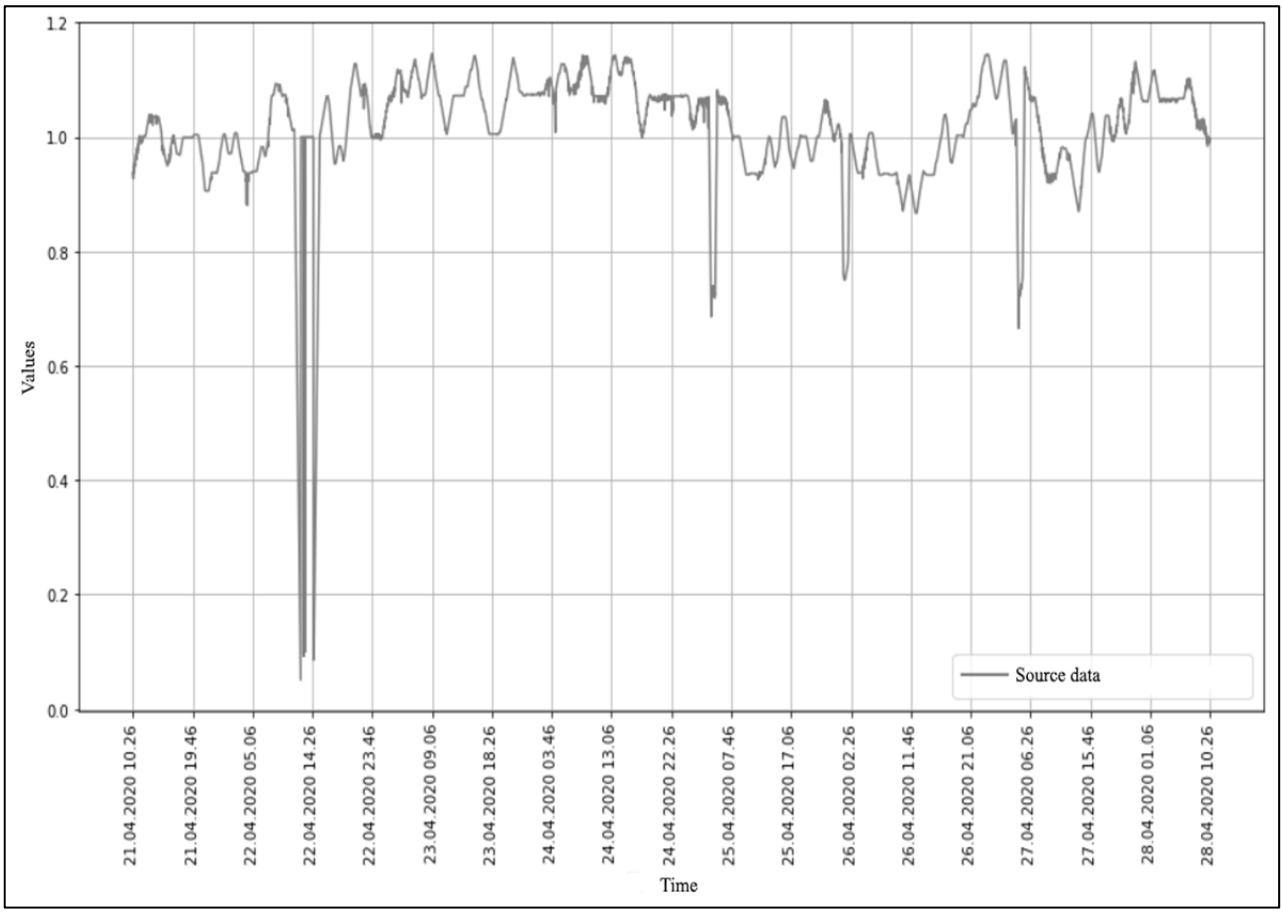

Figure 1. Integral risk's index for a weekly period

Figure 1. shows the integral risk's index for a weekly period with every 2 minutes frequency of readings. During data collection, due to various kinds of interference, the so-called noise, the data is distorted, including the received data homogeneity, that is, for specific points in time t_i, the values of the time series $u(t_i)$ are atypical, anomalous. In mathematical statistics, such values are called outliers and must be removed from the series and then replaced by an acceptable value by interpolation, since such values do not allow the model to concentrate on the series' key indicators, distracting to particular atypical cases.

To filter the integral risk indicator's outliers, it was proposed to apply an analytical method based on the interquartile range, which proved to be effective when considering stochastic processes at an industrial enterprise.

To determine the extreme values, the assumption is made that the main signal scattering is located between the 1st and 3rd quartiles, i.e. between the 25th and 75th percentiles (Figure 2.). It contains the center 50% of the observations in the ordered set, with 25% of the observations below the center point and 25% above:

$$IQR = Q_3 - Q_1, \quad (2)$$

Where Q_1 is the lower (first) quartile, Q_3 is the upper (third) quartile. Let us believe that the points with the i index are anomalous if they do not meet the condition:

$$(Q_1 - 1.5 * IQR) < u(t_i) < (Q_3 + 1.5 * IQR). \quad (3)$$

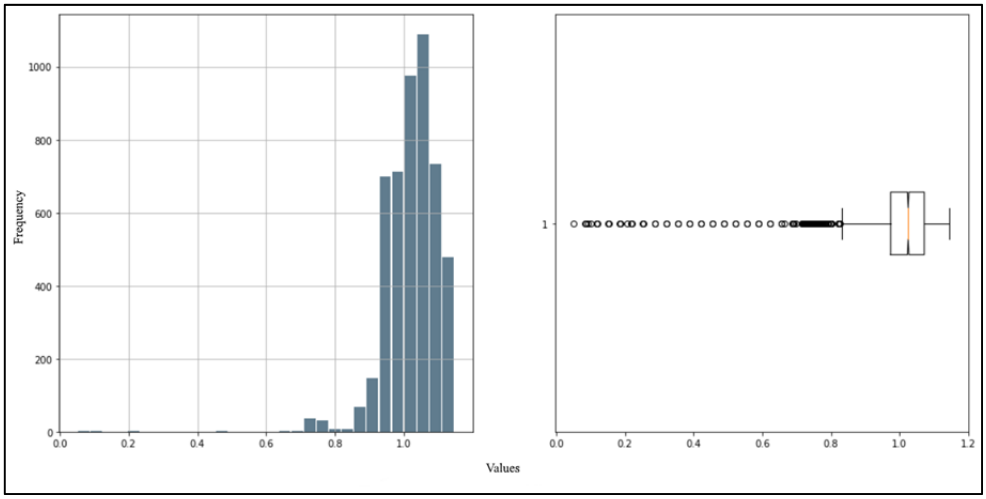

Figure 2. Histogram of frequency distribution

For points outside the specified range, the new value was calculated using local cubic spline interpolation, which for each interval $[t_i;\ t_{i+1}]$ constructs an interpolation polynomial of the third degree:

$$S_3(t) = \frac{(t_{i+1}-t)^2\cdot(2\cdot(t-t_i)+h)}{h^3}y_i + \frac{(t-t_i)^2\cdot(2\cdot(t_{i+1}-t)+h)}{h^3}y_{i+1} + \frac{(t_{i+1}-t)^2\cdot(t-t_i)}{h^2}m_i + \frac{(t-t_i)^2\cdot(t-t_{i+1})}{h^2}m_{i+1}, \quad (4)$$

where $t_i \le t \le t_{i+1}, i = \overline{1, n-1}, h = \frac{t_n - t_1}{n}$.

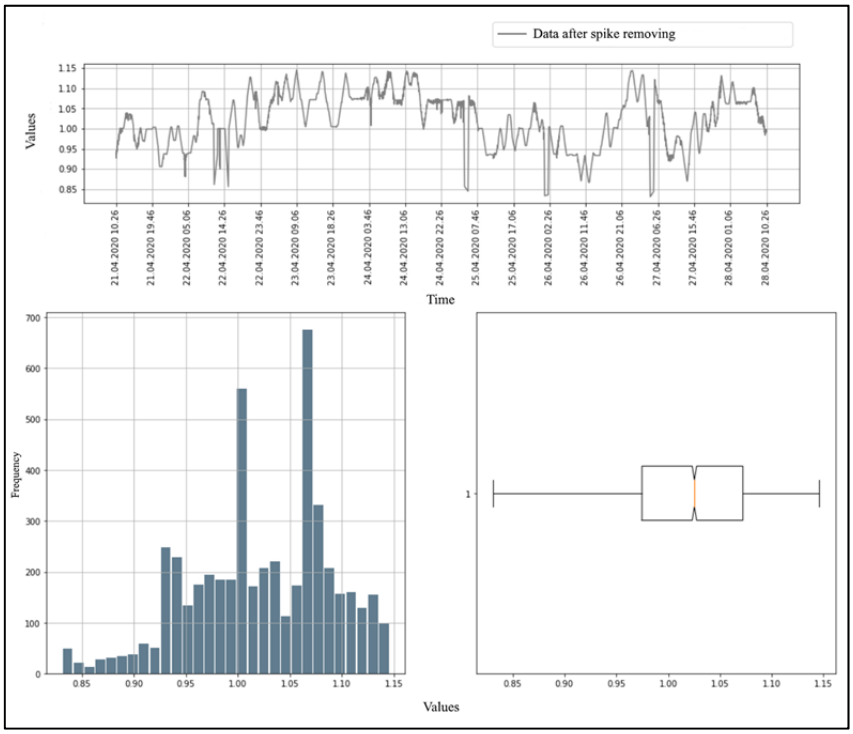

Figure 3. Input signal after the filtering outliers' process

It can be seen (Figure 3.) that the new series has no bright outliers.

2.2. Time Series' main Components Determination

After obtaining the time series, that meets the requirements for the homogeneity of the time step τ: $t_i = t_0 + (i-1)\tau$ and the outliers' absence, it is necessary to determine the main characteristics' values to create an adequate mathematical model describing the stochastic process. The time series is additive and consists of the following components [7]:

1. trend (systematic movement);
2. seasonality effect;

3. fluctuations in relation to the trend;
4. irregular component.

The trend is a component that changes systematically over a long time period. The data trend may increase, decrease, or be stable over time, but generally be upward, downward, or stable.

The least squares method is used to determine the trend component:

$$F(y) = \sum_{i=1}^{n}(u(t_i) - T(t_i))^2 \to min, \quad (5)$$

where $T(t_i), i = \overline{1,n}$ is a trend function.

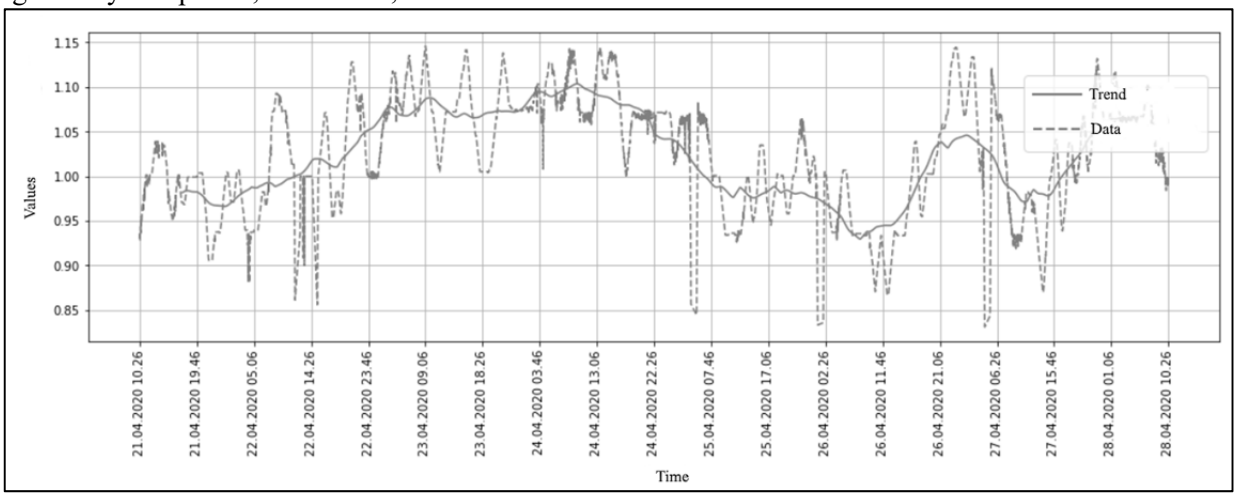

Figure 4. Trend component of the integral risk indicator

Seasonality effect is the changes imposed on the system by some external cyclical mechanism and leading to oscillatory processes. To be sure that seasonality is present in the consideration series, and fluctuations in the integral risk indicator's values are not caused only by random external factors' fluctuations, a study was carried out for the presence of a seasonal component.

This study is iterative: first, the rough trend is excluded and the seasonal factor is estimated, seasonality is excluded from the initial data, the trend is re-evaluated, etc.

For the additive model, the seasonal component is represented by the absolute deviation indicator Sa_k:

$$\sum_{k=1}^{l} Sa_k = 0, \quad (6)$$

where $k = \overline{1,l}$ is the season number within the considered time series period, l is the number of seasons in the reviewed period. Then for u_{kj} series, where $j = \overline{1,m}$ is the period number in the whole row, and m is number of periods, and $\tilde{u}_{kj}$ smoothed row's absolute deviation indicator Sa_k at some row calculated by the formula:

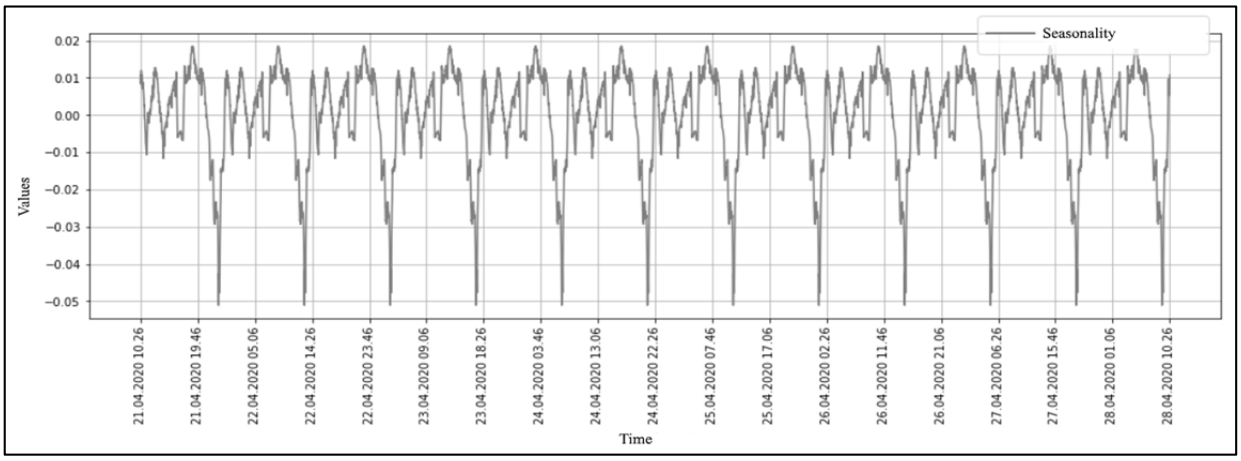

Figure 5. Seasonal component of the integral risk indicator

When the trend and periodic components are identified, a series is left representing fluctuations. Further research is aimed at identifying the residual series systematization, that is, at determining the dependence of the series on time values. If the series is systematic, then the phenomenon under consideration is fluctuations relative to the trend, which is a significant component and has a tangible effect on the subsequent series values; otherwise, the residual series is an irregular component (or error) and must be differentiated from the general time series for building a more accurate model.

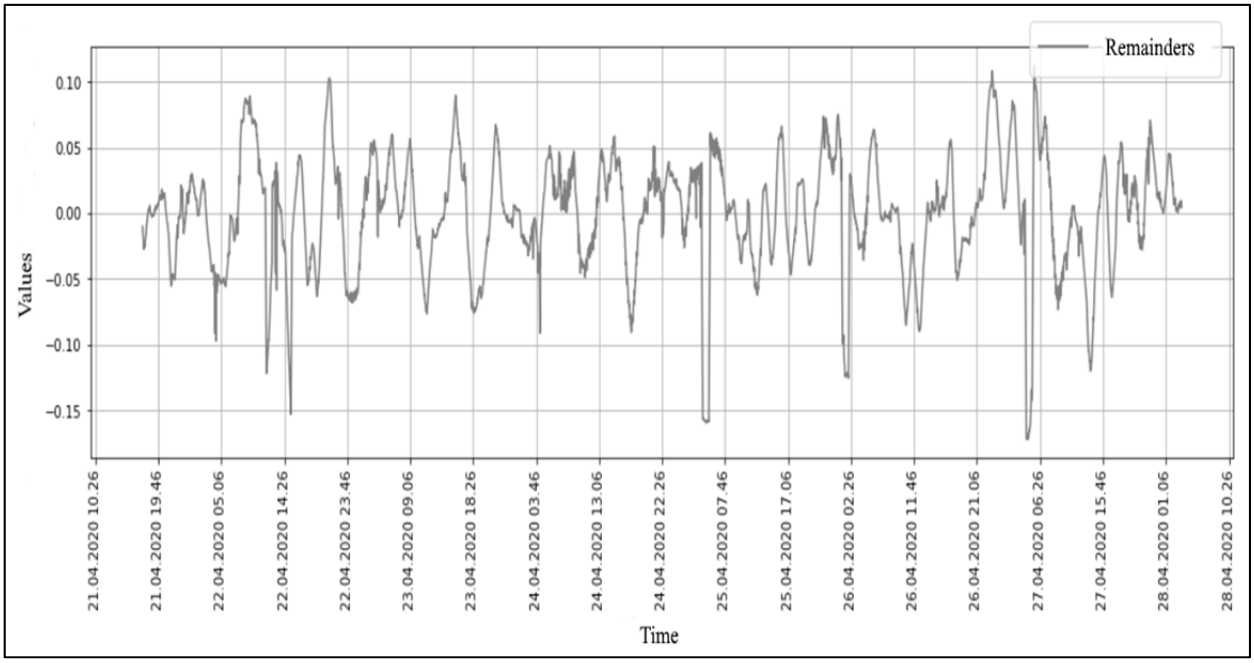

Figure 6. Remains of the integral risk indicator after extracting the trend and seasonal components

To study the systematization of the residuals, let us use the criterion based on Spearman's rank correlation [7]:

$$r_s = 1 - \frac{12V}{n(n^2-1)}, \quad (7)$$

where

$$V = \sum_{i<j}^{n}(j-i)H_{ij}, \quad (8)$$

$$H_{ij} = \begin{cases} 1, u_i > u_j \\ 0, u_i < u_j., \\ 1, u_i = u_j \end{cases} \quad (9)$$

For the considerate series, the r_s value was 0.67, which indicates the considered residual series' systematization, which is a significant component and cannot be differentiated from the general integral risk indicator.

2.3. Building a Mathematical Model

Time series models can have different forms, structures and represent different stochastic processes [8]. From current values' linear dependence assumption of the series on the retrospective ones, let us use the class of autoregressive models, the general view of which is given as follows:

$$Y_t = \propto_0 + \sum_{i=1}^{p} \propto_i Y_{t-i} + \varepsilon_t, \quad (10)$$

where $\propto_0$ is constant (often takes zero value);
p is the number of series' historical values;
$\propto_i$ are autoregressive coefficients;
ε_t is the random error with the following properties:

1. $E[\varepsilon_t] = 0$;
2. $E[{\varepsilon_t}^2] = \sigma^2$;
3. $E[\varepsilon_t, \varepsilon_s] = 0, t \neq s$.

This model describes a stationary process of order p, i.e. a stochastic process in which the probability distribution does not change over time, and the number of lags, the so-called delays, that is, the number of series' retrospective values, is equal to p.

The series is stationary in the broad sense if the average, variance and covariance Y_t do not depend on the time t [9].

To determine whether the process is stationary, let us use the Augmented Dickey-Fuller (ADF) test [10] for the presence of unit roots (Table 1.).

Table 1. ADF test results

ADF statistics	-4.455	
p value	0.0002	
Critical values of a given significance level	1%	-3.432
	5%	-2.862
	10%	-2.567

The test showed that the time series under consideration has a unit root, which means that the first differences of the series are stationary by definition, and the process under consideration is a first-order autoregression (Figure 7.).

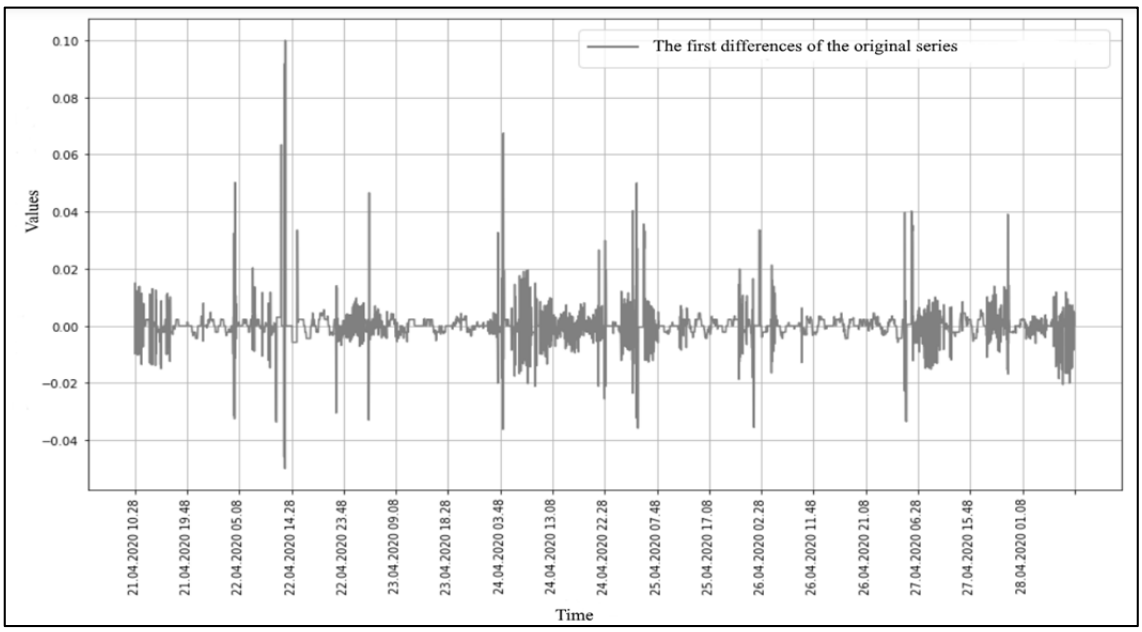

Figure 7. Time series after transformation by the first order difference operator

Based on the foregoing, the ARIMA (Auto Regressive Integrated Moving Average, Box-Jenkins model) model was chosen as a model describing the process of changing the integral risk indicator, autoregression model, an integrated moving average, a distinctive feature of which is the processes described interpretation by the integrated time series (difference-stationary time series). The $ARIMA(p,d,q)$ model is three-parameter and is expressed as follows:

$$\nabla^d Y_t = \propto_0 + \sum_{i=1}^{p} \propto_i \nabla^d Y_{t-i} - \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} + \varepsilon_t, \quad (11)$$

where: d is the non-negative integer characterizing the order of the model's integrated part;
∇^d is the difference order operator d, $\nabla^d = (1-L)^d$;
q is the random deviations' previous values' number taken into account in the model;
β_i are the moving average coefficients.

Above, there have been defined the parameter $d = 1$ as the order of the difference operator, under the influence of which the series becomes stationary. To determine the parameters p and q, consider the graphs of the autocorrelation functions (ACF) and partial autocorrelation function (PACF) of the integral risk indicator, differentiated once (Figure 8.).

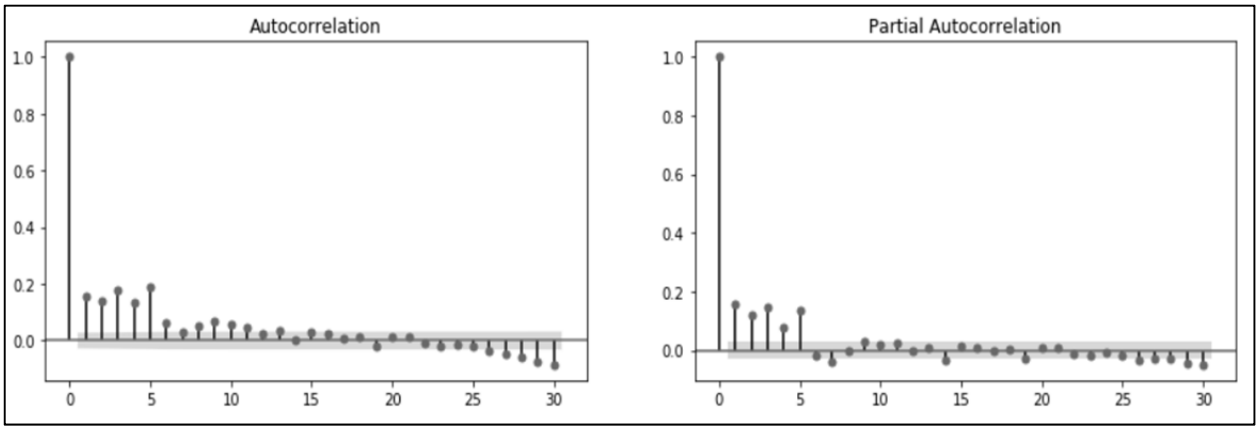

Figure 8. Graphs of the ACF and PACF functions for the integral risk indicator, differentiated once

For parameter p evaluation, let us consider the PACF function, which shows the correlation between the series' values and their lag, but without taking into account other dependencies (hence the name: partial). The maximum autocorrelation coefficient other than zero indicates an approximate parameter p evaluation, which for the considered time series is 1.

The values' determination of the parameters $\propto_i$ and β_i occurs at the training the model stage using the maximum likelihood method [11].

3. Results and Discussion

After obtaining all parameters' evaluations for the constructed model, the residuals are calculated:

$$e_t = Y_t - \hat{Y}_t, \quad (12)$$

by which the resulting model's adequacy is evaluated. Earlier there was introduced the assumption that the model error ε_t is the white noise. Consequently, if the process is successfully modeled, the residuals e_t under consideration should be uncorrelated normally distributed random variables. Figure 9. shows residual's graph ACF function. Since the outlier is present only at zero lag, it can be concluded that the error of the resulting model is random enough that it cannot be refined by another model.

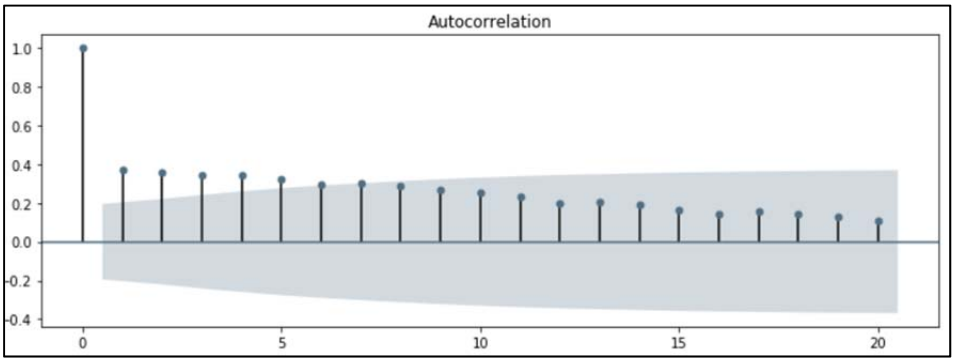

Figure 9. Plot of ACF residuals

A well-established method for evaluating the autoregressive time series model's quality is the Akaike's information criterion (AIC), which is used for models using the log likelihood function. This criterion shows how much information will be lost by decreasing the number of model parameters:

$$AIC = 2k - 2\ln(\mathcal{L}), \quad (13)$$

where k is the number of model parameters;
$\mathcal{L}$ is the likelihood function's maximum.

For the constructed model of the integral risk index, the logarithm of the likelihood function is equal to 10378.568, and $AIC = -25912.2$ with the initial values 18479.652 and -36943.304, respectively, which is a good modeling indicator.

The final stage in the building of a mathematical model for predicting the integral risk indicator is the building and forecast analysis. One of the tasks of this step is to determine the maximum length of the forecast period during which the considered model meets the quality requirements, for the calculation of which let us use the following indicators: the Kullback-Leibler distance ($KL\ D$), the determination coefficient (R^2); as well as visual analysis.

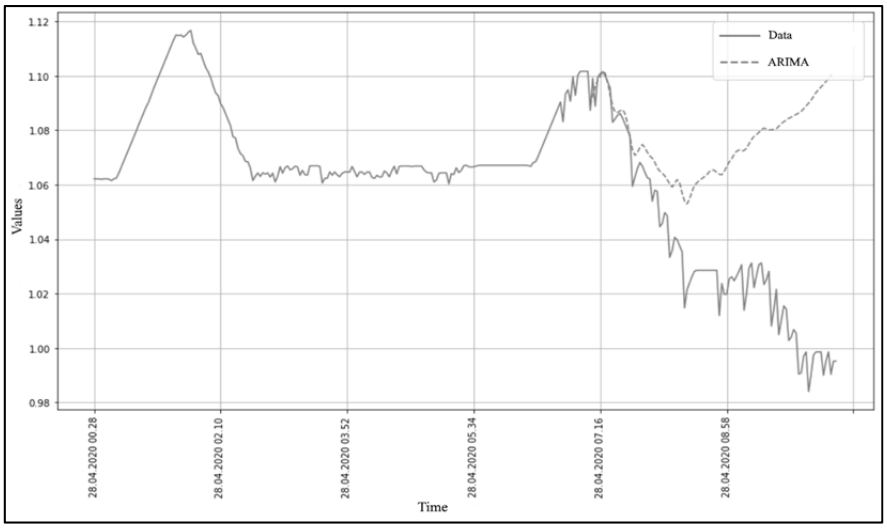

Figure 10. Result of integral risk indicator forecasting

Figure 10. shows the prediction result (dashed line) for a period of 200 minutes (100 points) and the previous values of the series (solid line) for 400 minutes (200 points). As can be seen from the graph, at first the forecast had a high accuracy and described the future behavior of the value under consideration quite well, but with an increase in the prediction step number, its quality deteriorated; moreover, the model predicted an uptrend in the interval when the trend was downtrend (after about an hour and a half from the start of forecasting).

Let us determine the period for which the built model is able to qualitatively predict the integral risk indicator's future values. To do this, consider the KL information criterion (16), which is a measure of the distance between two probability distributions [12], [13]:

$$KL = \sum_{i=1}^{l} z_i \log \frac{z_i}{\hat{z}_i}, KL \geq 0., \qquad (14)$$

The KL shows how much information will be lost when replacing the true series z_i with the assumed $\hat{z}_i$, the smaller the KL value, the better the coincidence of the initial and calculated series.

For two distributions, the validation dataset and the predicted one, the KL is $KL = 0.225$. Let us take this value as an acceptable level and plot the dependence of the KL indicator value on the period value on the basis that each interval included in the period contains 5 points (10 minutes), and the very first period contains two intervals.

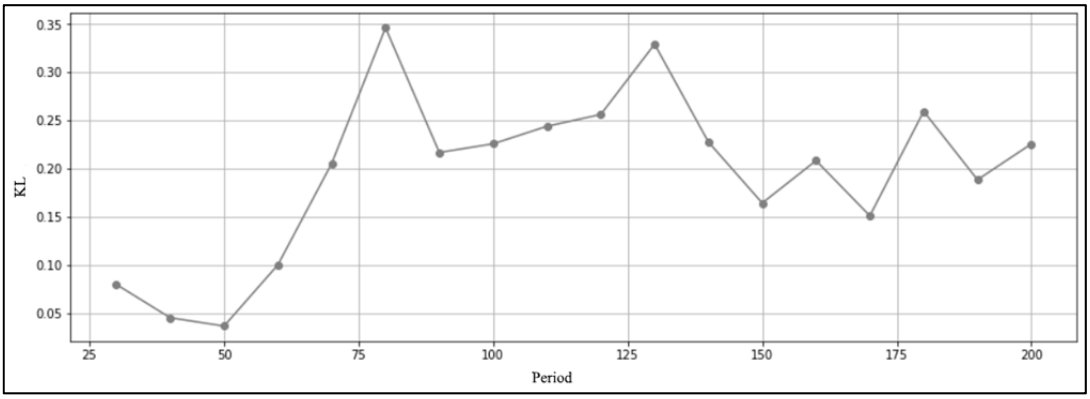

Figure 11. KL indicator dependence on the value of the period

As can be seen from the graph (Figure 11.), for small periods the indicator's value is within the acceptable level, but with an increase in the number of intervals included in the consideration, the KL value will first sharply increase, and then it is at high levels, periodically decreasing to an acceptable level. The time period for which the KL indicator has never exceeded the acceptable level is 70 minutes [14].

In order to consolidate the obtained result, let us conduct a similar test using the coefficient of determination indicator R^2, which characterizes how well the observed results are reproduced by the model, based on the share of the results explained by the model total deviation:

$$R^2 = 1 - \frac{SSR}{SST}, 0 \leq R^2 \leq 1, \qquad (15)$$

where $SSR = \sum_{i=1}^{l}(z_i - \hat{z}_i)^2$ is the sum of squares regression residuals;

$SST = \sum_{i=1}^{l}(z_i - \bar{z})^2$ is the total variance.

The coefficient of determination takes values from 0 to 1: the closer the coefficient value to 1, the stronger the dependence. For acceptable models, assume that R^2 must be at least 0.6.

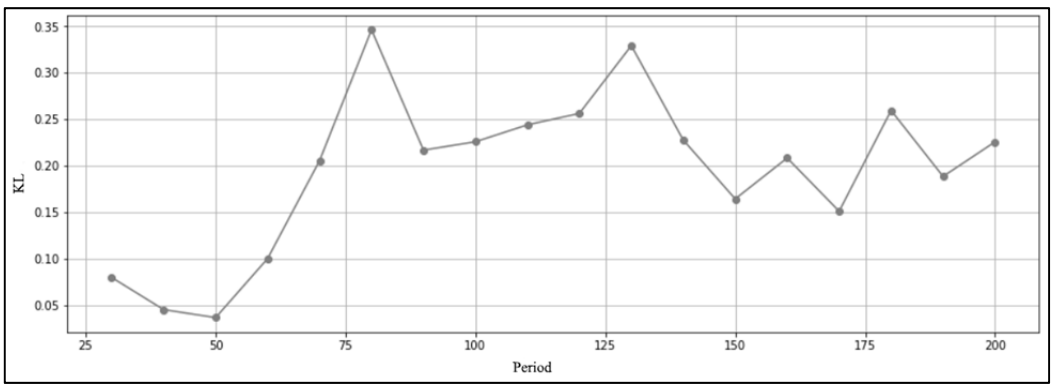

Figure 12. Dependence of the R^2 indicator on the value of the period

Figure 12. shows the dependence of the indicator R^2 value on the value of the period under consideration. The resulting graph confirms the results obtained earlier when considering the values of the KL indicator: the most qualitative forecast is obtained in short periods containing the first several intervals, after which the forecast quality drops significantly. The value of the allowable period when considering the R^2 index was 80 minutes [15].

4. Conclusion

The analysis of the integral risk indicator on real data of a large industrial facility and mathematical modeling of the obtained time series showed good results for the practical application of autoregressive models for forecasting in the industrial safety field. The forecast accuracy evaluation using various criteria made it possible to conclude that autoregressive models describe well the dynamics of changes in the integral indicator of enterprise's industrial safety risk and allow predicting its further behavior, which increases management efficiency. The test data against which the predictive data is compared is entirely within the 95% confidence interval. This allows asserting that the class of models under consideration can be used to predict the integral risk indicator.

The advantage of this model type is the absence of requirements for high computational costs, and therefore, for expensive equipment, with the goal of obtaining a prediction with a satisfactory quality level for a period sufficient for making a decision and organizing measures to prevent an accident.

The research and suggestions carried out in the article were used by the company "RKSS-Programming Systems" in the innovative software complex for intelligent monitoring "Zodiac" to predict industrial safety risks in a number of HPFs of the Fuel and Energy Complex of Russia.

References

[1]. Wintle, J. B., Kenzie, B. W., Amphlett, G. J., & Smalley, S. (2001). *Best practice for risk based inspection as a part of plant integrity management*. Great Britain, Health and Safety Executive.

[2]. Straub, D., Goyet, J., Sørensen, J. D., & Faber, M. H. (2006, January). Benefits of risk based inspection planning for offshore structures. In *International Conference on Offshore Mechanics and Arctic Engineering* (Vol. 47489, pp. 59-68).

[3]. Geary, W. (2002). Risk based inspection: a case study evaluation of offshore process plant. *Health and Safety Laboratory: Sheffield, UK*.

[4]. Law, F. (1997). On Industrial Safety of Hazardous Production Facilities. *Federal Law of*, (116-FZ).

[5]. Levin, S. Ye., Nagibin, S. Ya, & Shilov, V. V. (2018). Distance control of process safety of fuel and energy complex. *Engineering sciences – from theory to practice*, 30-36.

[6]. Loskutov, A. Y. (2013). Time series analysis. *Course of lectures. Faculty of Physics, Moscow State University.* Retrieved from: http://chaos.phys.msu.ru/loskutov/PDF/Lectures_time_series_analysis.pdf. [accessed: 12 July 2020].

[7]. Kendall, M., & Stuart, A. (1976). Multivariate statistical analysis and time series. *M.: Nauka*, 65-68.

[8]. Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2011). *Time series analysis: forecasting and control* (Vol. 734). John Wiley & Sons.

[9]. Magnus Ya.R., Katyshev P.K., Peresetskiy A.A. Ekonometrika. Nachal'nyy kurs (Econometrics. Initial course), Moscow, Delo Publ., 2004, 576 p. (in Russ.).

[10]. Fuller, W. A. (2009). *Introduction to statistical time series* (Vol. 428). John Wiley & Sons.

[11]. Bezruchko, B. P., & Smirnov, D. A. (2005). Mathematical modeling and chaotic time series. *Saratov, Russia: GosUNTs" Kolledzh*.

[12]. Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79-86.

[13]. Malinetskii, G. G., & Potapov, A. B. (2000). Modern problems of nonlinear dynamics. *Editorial URSS, Moscow*.

[14]. Plas, D. V. (2018). Python dlya slozhnykh zadach: nauka o dannykh i mashinnoye obucheniye. SPb.: Piter, 576.

[15]. Rashka, S. (2017). Python i mashinnoye obucheniye. M.: DMK Press.