Time series in hydrogeological processes

Rashid Burtiev[1], Yuliia Semenova[2], Victor Jeleapov[1], Sergiu Troian[1], Vladlen Cardanets[1]

Abstract - In this work, the time series model is used to study the structure of hydrogeological data series to identify patterns in the change in the levels of the series and build its model in order to predict and study the relationships between the levels of hydrogeological data. Observations of the activity of hydrogeological processes showed that the periods of variations of these processes are scattered chaotically on the time axis. According to their schedule, it is impossible to definitely speak about the regularity in the duration of the periods of variations, and in the alternation of periods of calm with a period of high activity. The impetus for this study was the desire to analyze the structure of a number of formal methods to search for statistical patterns in the variations of hydrogeological parameters over time. Time series models were used to study the dynamics of hydrogeological processes. The accuracy of the forecast is indicated by the comparison of the forecast series with the actual data. The predicted values of hydrogeological data are within the confidence intervals. In this work, the time series model is used: to identify the statistical relationship between the frequency and depth of occurrence of earthquakes, as well as to identify the statistical dependence of these data on hydrogeological variations; determination of patterns in changes in the levels of the series and the construction of its model in order to predict and study the relationships between hydrogeological data.

Index Terms: SPSS, time series, hydrogeological data.

1 INTRODUCTION

Monthly samples from Lipcani well are under review. Observations of the activity of hydrogeological processes show that the periods of process variations are scattered chaotically on the time axis. According to their schedule, it is impossible to speak with certainty about the regularity in the duration of the periods of hydrogeological variations. The impetus for this study was the desire to analyze the structure of a number of formal methods to search for statistical patterns in the variations of hydrogeological parameters over time. Spatial and time series models can be used to study the dynamics of events. A spatial model describes a set of parameters at a given moment in time. A time series is a series of regular observations of a certain parameter at successive points in time or at intervals of time. In this work, the time series model is used to study the structure of hydrogeological events. In general, the purpose of studying a time series is to identify patterns in changes in the levels of a series and build its model in order to predict and study the relationships between hydrogeological phenomena.

$$x_t = m_t c_t s_t \varepsilon_t \quad (1)$$

$$x_t = m_t c_t s_t + \varepsilon_t \quad (2)$$

$$x_t = m_t c_t + s_t + \varepsilon_t \quad (3)$$

$$x_t = m_t + c_t + s_t + \varepsilon_t \quad (4)$$

The theory of time series is used to solve the following main tasks: determination of the nature of the series; determination of the main parameters of the series; prediction of future values of the time series from the available data. Time series components are trend, cyclic component, seasonal component and a random component. The time series consists of deterministic and random parts. Trend; the cyclical and seasonal components from the deterministic part, which is used to predict the future values of the series. The seasonal component is used to denote a non-random function that is formed based on fluctuations in the series under study that are periodically repeated at a certain time of the year.

In our case, it is assumed that the form of interaction of the listed components can be multiplicative (1), mixed (2, 3) additive (4), mt - trend, stable long-term trend; • ct - cyclical

[1]*Ministry of Education, Culture and Research, Institute of Geology and Seismology,*
[2]*Institute of Geophysics, National Academy of Sciences of Ukraine, Kiev, Ukraine*

ISSN 2229-5518

component, rather long irregular fluctuations, with a period of more than one year; st - seasonal component, fairly regular periodic fluctuations occurring in a time interval not exceeding one year; εt - random component (error). The series gt = mt + ct, consisting of the sum of the trend and the cyclical component, is called the trend - the cyclical component (conjuncture cycle). The presence of a trend means the presence of a long-term systematic component in the time series, which describes the main trend in the dynamics of the time series. The trend, as a smooth curve, reflects the influence of long-term non-periodic factors. Formally, a trend is understood as a sequence of conditional mean values determined by a time variable. Periodic fluctuations can be divided into seasonal fluctuations, in which the period does not exceed one year, caused by climatic conditions, and cyclical with a period of fluctuations of several years. Periodicity means xt + kp = xt; k = 1, 2, .., and is seasonal at p = 1 year, cyclical at p = 2.3,… years. Random noise makes it difficult to detect the regular components; usually, time series research methods include various noise filtering algorithms, which make it possible to more clearly determine the regular component. For the correct reflection of the real process by the time series, the data sampling must satisfy the following conditions: comparability; uniformity; stability, and sufficiency of the amount of data. In this work, we used the SPSS20 and Excel 2016 packages.

2 Time series in hydrogeological processes

The theory of time series can be useful in studying the temporal regime of hydrogeological events, to indicate periods of calm, as well as periods of moderate and increased activity of the occurrence of hydrogeological events. A trend is a non-random function formed under the influence of general or long-term trends. The cyclic component is also a non-random function. Hydrogeological data recorded in wells are subject to analysis. The preliminary stage of statistical processing should be the stage of checking the homogeneity of the sample, in the sense of the invariability of the probabilistic characteristics. A measure of the homogeneity of the statistical population is the coefficient of variation [Kramer, 1975]: [Anderson, 1971; Brillinger, 1981; Kendal, 1981; Hennan, 1974]: [Kramer, 1975]:

$$V_q = \frac{\sigma}{\bar{a}} 100\%, \quad (5)$$

The sample is considered homogeneous if the coefficient of variation Vq does not exceed 33%. In this task, sampling means regular monthly hydrogeological indicators of 12 wells. The coefficient of variation of the series does not exceed Vq = 26.4 (Table 1), confirms the homogeneity monthly samples from Lipcani well (Table.1).

Table 1.
The coefficient of variation of monthly samples from Lipcani well

January	February	March	April	May	June
1,898%	2,413%	1.917%	1.9669%	2.4018%	2.6419%
July	August	September	October	November	December
2.6022%	2.4089%	2.1054%	1.9114%	1.8116%	1.8146%

3 Autocorrelation

If the assumption about the random nature of the level fluctuations in the studied series is true, then there should be no connection between the levels. An alternative assumption assumes the existence of a relationship between successive levels, that is, in this case, the time series is not random. To assess the relationship between successive levels of the series, the autocorrelation function of the series can be used, for which, the values of autocorrelations do not go beyond 95% confidence intervals. Therefore, the data is "white noise". The purpose of time series analysis is to determine the model for the implementation of the series. Correlation analysis allows us to reveal the structure of the series, that is, to determine the presence in the series of this or that periodic components of unknown frequency. The autocorrelation

ISSN 2229-5518

function is determined by the formula [Brillinger, 1980]:

$$ACF(\tau) = \frac{\sum_{k=\tau+1}^{n}(y_k-\overline{y})(y_{k-\tau}-\overline{y})}{\sum_{k=1}^{n}(y_k-\overline{y})^2} \quad (6)$$

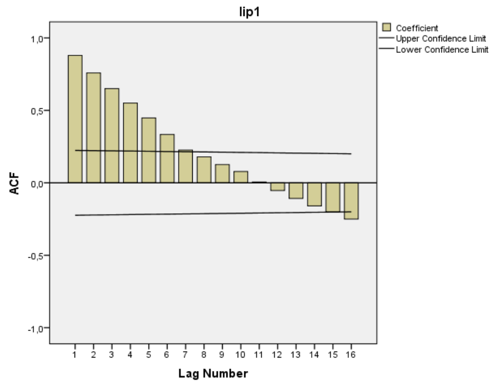


Figure 1a.

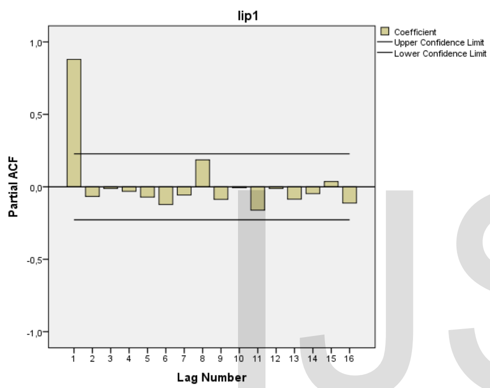


Figure 1b.

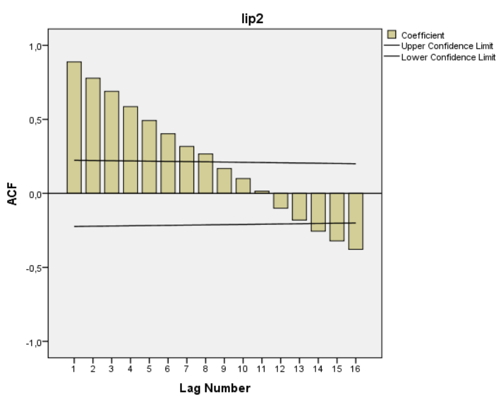


Figure 1c.

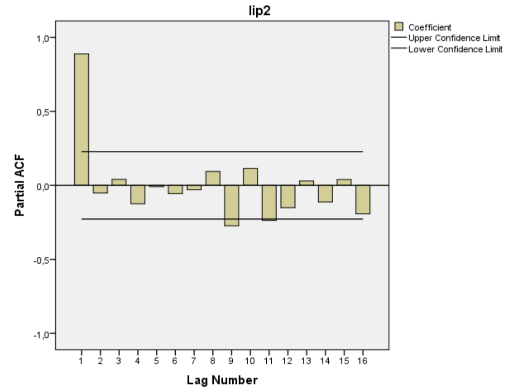


Figure 1d.

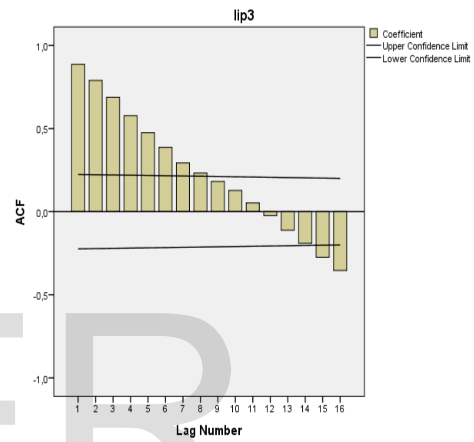


Figure 1e.

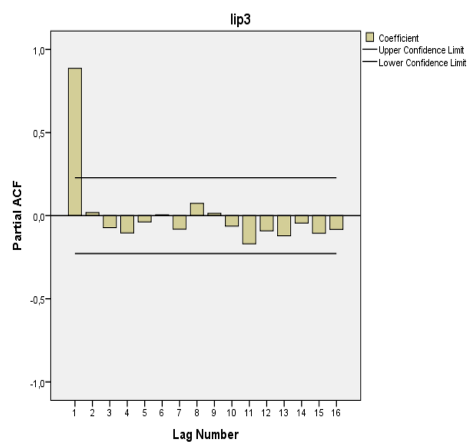


Figure 1f.

Fig. 1. Autocorrelation and partial autocorrelation functions

So, if the autocorrelation coefficient of the first order turned out to be the highest, then the series under study contains only a tendency [Anderson, 1971; Brillinger, 1980]. If in a series, none of the autocorrelation coefficients is significant, then this may mean: the series does not contain tendencies and cyclical fluctuations and has a random structure - "white noise" and the series contains a strong neutral tendency, which requires additional analysis to be identified. To determine the type and order of the processes that generate a stationary time series, the apparatus of autocorrelation functions is used: conventional - ACF (Fig. 1c) and private - PACF (Fig. 1d). If the time series is close to white noise, then the correlogram oscillates close to the horizontal axis, and its values are close to 0. At large values of τ, the ACF (τ) estimate for the autocorrelation coefficient contains errors. This is due to partial summation since observations are discarded from the summation. Therefore, the correlogram at large values of τ does not reflect the true structure of the series. For a stationary series, ACF (τ) decreases rapidly with increasing τ. In the presence of a trend, the autocorrelation function takes the form of a slowly falling curve. In the case of seasonal periodicity, the ACF graph contains peaks for lags that are multiples of the seasonality period. But these peaks can be hidden by the presence of a trend or a large variance of the random component. For a given significance level α (the default is usually α = 0.5 or 0.1), it is possible to calculate the boundaries of the confidence interval in which finding the value of the autocorrelation function, for a given lag τ, with probability 1-α does not contradict the assumption that there is no correlation of the cross-sections with this lag (fig. 1). When graphically depicting the autocorrelation function or its estimates, these intervals give two boundary curves (above and below the main graph). Going beyond these boundary curves is seen as an indication of the significance of the correlation with the corresponding lag.

The AR (Autoregressive) order is determined by the behavior of the PACF, and by the behavior of the ACF - the order of the MA (Moving Average) (Fig. 3). If there is a pattern in the ACF, only AR is included in the model, and if there is a pattern in the PACF, only MA is included in the model. As for the orders, they are determined by the number of the lag at which the border of the confidence interval is exceeded. Also, the presence of a statistically significant autocorrelation coefficient indicates that the series is not random, and there is a certain relationship between successive observations. A significant negative autocorrelation coefficient indicates high-frequency level oscillations. Linear autocorrelation coefficients characterize the tightness of only the linear relationship between the current and previous levels of the series. Therefore, according to the autocorrelation coefficients, one can judge only the presence or absence of a linear relationship between the levels of the series. To check the series for the presence of a non-linear trend, linear autocorrelation coefficients are calculated for the time series, consisting of the logarithms of the original levels. Non-zero values of autocorrelation coefficients indicate the presence of a non-linear trend.

Another useful method for studying periodicity is the study of PACF (Fig. 1), which is a deepening of the concept of the usual autocorrelation function. In PACF, the relationship between intermediate observations is eliminated. In other words, partial autocorrelation on a given lag is similar to regular autocorrelation, except that the calculation removes the influence of autocorrelations with smaller lags. [Anderson, 1971; Brillinger, 1980]. At lag 1 (when there are no intermediate values within the lag), the partial autocorrelation is equal to the normal autocorrelation. As noted above, the periodic component for a given lag k can be removed by taking the difference of the corresponding order. This means that the (i-k) -th element is subtracted from each i-th element of the row. There are two arguments in favor of the considered transformations: firstly, in this way, it is possible to determine the hidden periodic components of the series. Recall that autocorrelations on successive lags are dependent. Therefore, removing some autocorrelations will change other autocorrelations that may have suppressed them, and make some other seasonal components more noticeable. secondly, the removal of seasonal components transforms the time series into a stationary series, which is necessary for the use of ARIMA (Autoregressive Integrated Moving Average) and other methods, such as spectral analysis. The test criterion for checking the significance of autocorrelation coefficients is the Box – Pierce test [Shanchenko, 2008]:

$$\boldsymbol{Q} = \boldsymbol{n}\sum_{k=1}^{m} \boldsymbol{r_k} \qquad (7)$$

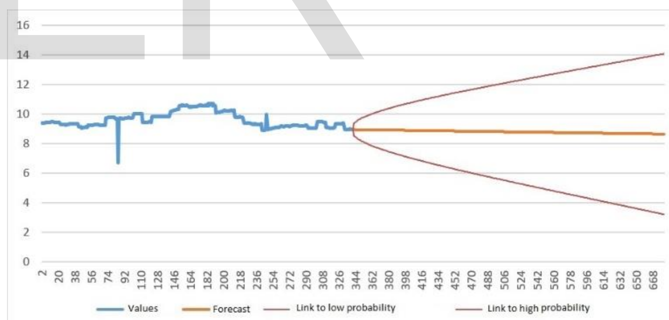
where r_k is the autocorrelation coefficient with lag k, m is the largest lag, n is the row length. Sample statistics of criterion (7) is a χ2 distributed random variable with m degrees of freedom. If the autocorrelation function of a series of residuals indicates that a trend and fluctuations still exist in it, and repeated application of the procedure does not change its nature. In such cases, it is recommended to build a moving average time series model and conduct a deeper analysis of periodic fluctuations [Spezialvorlesung Zeitreihenanalyse, 2007]. The criterion for the adequacy of the model to the time series is the indistinguishability of a number of residues from the process of "white noise" [Anderson, 1971; Brillinger, 1980; Boxing, Jenkins, 1974;

Marneau, 2008]. The Box-Ljung statistics calculated to test the significance of autocorrelations is less than the critical value corresponding to the probability of significance α = 0.05. A deeper analysis of periodic fluctuations in order to detect hidden periodicities is carried out using spectral analysis of the time series, shows the numerical characteristics of data regression for days. The Durbin-Watson coefficient shows autocorrelation, which takes values from the interval (0.4) [Seigno, 2007]. Values close to 0 indicate strong positive autocorrelation, close to 4, strong negative autocorrelation, and close to 2, no autocorrelation. In this problem, the value of the Durbin-Watson coefficient is $d = 1.876$, and the condition $1.5 < d < 2.5$ is satisfied. Consequently, there is no autocorrelation [Sidenko, Vishnyakov, Isaev, 2011]. According to Fisher's criterion, the regression model turned out to be insignificant for all the regressors included in the five models, since the probability of significance of the F-test sample statistics was used to test the null hypothesis H0: the regression equation is insignificant is = 0.562. The exploratory analysis makes it possible to describe subsets of observations using a variety of statistics (calculating frequencies and percentages, averages, etc.) and graphs. Kurtosis (Table 2) is a measure of the "smoothness" of the distribution. Positive kurtosis indicates a "flat-topped" distribution, in which the maximum probability is not as pronounced as in the normal distribution. The kurtosis values exceeding 5.0 indicate that there are more values at the edges of the distribution than around the mean. Negative kurtosis, on the contrary, characterizes a "peaked" distribution, the graph of which is more elongated along the vertical axis than the graph of the normal distribution. It is believed that a distribution with kurtosis in the range from -1 to +1 corresponds approximately to the normal form. In most cases, it is quite acceptable to consider a distribution with kurtosis not exceeding 2 in absolute value as normal. In most cases, distribution with an asymmetry ranging from -1 to +1 is accepted as normal. Skewness (Table 2) shows in which direction most of the distribution values are shifted relative to the mean. A zero value of skewness means asymmetry of the distribution relative to the mean, positive skewness indicates a shift in the distribution towards smaller values, and negative skewness towards larger values. In studies that do not require high accuracy of results, a distribution with an asymmetry not exceeding 2 in absolute value is considered normal. SPSS can compute a variety of existing correlation coefficients. Among them, both the simplest and one of the most commonly used are Pearson's linear correlation coefficient, and varieties of Spearman's and Kendall's rank correlation coefficients.

4 Time series forecasting hydrogeological data

Forecasting (Fig. 1) was carried out using EXCEL 2016. If you start forecasting before the last point, you can get an estimate of the forecast accuracy by comparing the forecast series with the actual data. But if you start forecasting too early, the forecast may differ from the forecast based on all statistical data. In this case, the forecast will be more accurate. If the data shows seasonal trends, it is recommended to start forecasting from the date before the last point of the statistical data. The test is based on calculating the maximum difference between the cumulative sample rates and the theoretical normal distribution function. This difference is denoted by the z-value, from which the probability of significance is calculated. According to the Kolmogorov – Smirnov criterion, the null hypothesis H0: the conformity with the normal distribution of the time series of data is confirmed [Seigno, 2007; Bernhardt 2007; SPSS Trends 14.0, 2006]. One of the indicators of the normality of the distribution of the sample is kurtosis and asymmetry, (Table 2)

Fig. 2. Kolmogorov – Smirnov test



5 Spectral analysis of the time series

Spectral analysis is used to determine the periodic component for a known period length. In fact, this is linear regression, where the dependent variable is the levels of the series, and the functions of sines and cosines are the regressors. Spectral analysis determines the correlation of regressors of different frequencies with the observed data. There is a well-known theorem [Boxing, Jenkins, 1974; Kolmogorov, Fomin, 1976; Piskunov, 1964], according to which among all trigonometric polynomials of order n the least root-mean-square deviation has a polynomial, the sought coefficients of which are the Fourier coefficients. One of the methods for modeling seasonal and cyclical fluctuations is based on the use of one-dimensional Fourier series. Fourier series is a heuristic algorithm that is one of the varieties of spectral analysis. With the help of spectral analysis in the structure of the time series, the peak of deviations from the trend is determined, which makes it possible to calculate the duration of the periodic component of the series. When the spectral analysis is applied, a random stationary process is represented as a sum of harmonic oscillations of different frequencies, called harmonics. The spectrum describes the distribution of the amplitudes of a random stationary process over different frequencies. Investigation of the frequency structure of the series is performed by the “Spectral Analysis” procedure of the SPSS package. As you know, almost any periodic function can be approximated by a Fourier series, the sum of sines and cosines [Box, Jenkins, 1974; Piskunov, 1964; Smirnov, Dunin-Barkovsky, 1965; Backhaus, 2011; Spezialvorlesung Zeitreihenanalyse, 2006]:

$$x_t = a\cos(\omega t) + b\sin(\omega t)\,;\ 0 \le w \le \pi \qquad (9)$$

When modeling a time series by the sum of sines and cosines, sinusoidal periodic components appear on the periodogram in the form of separate vertices, and non-sinusoidal ones in the form of a series of equally spaced vertices of different heights. The top corresponding to the lowest frequency indicates the frequency of the periodic component in the time series [Spezialvorlesung Zeitreihenanalyse, 2007]. The nature of the remaining peaks indicates that the shape of the annual periodic component is not sinusoidal. For non-stationary series, with a smooth trend, the periodogram contains a sharp rise in the low-frequency region. In general, the analysis of possible periodicities is best done using a smoothed periodogram (spectral density function). In the frequency domain, the trend manifests itself in the form of an oscillation with an infinitely long period, and, accordingly, with a very low frequency. This trend can affect the values of the spectral density function at the left end of the frequency range. Therefore, for the correct application of spectral analysis, the series should be decomposed into components and the trend removed. Different approaches to identifying a trend give different variants of it; accordingly, the values of the levels of the series cleared of the trend are also different.

Low-frequency oscillations are of particular interest since long-term trends in the dynamics of the series are contained in the low-frequency region. The better the trend component model, the more fluctuations it contains. By Dirichlet's theorem [Smirnov, 1974; Yushchenko, Yakubovich, 2008; Tsapaeva, 2011] a continuous function with a period of 2π is uniquely determined by its Fourier series:

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n\cos(nt) + b_n\sin(nt) \qquad (10)$$

The application of this method requires knowledge of the frequency structure of the series. This method of clearing the series from components with undesirable frequencies did not improve the forecast quality. The autoregressive integrated moving average (ARIMA) model is used to model non-stationary time series. This model is one of the methods for estimating unknown parameters and forecasting time series.

6 Descriptive statistics

These characteristics of the central trend, which describe the position of the distribution and the spread, are displayed by default. Central trend characteristics include mean, median, and 5% trimmed mean. The scatter characteristics reflect the degree of difference between the values of the studied data; they include standard error, variance, standard deviation, minimum and maximum values of variables, range, and interquartile range. Descriptive statistics also include characteristics of the shape of the distribution, such as skewness and kurtosis, which are reported along with their standard errors. The 95% confidence interval for the mean is also displayed; you can specify a different value for the confidence level.

TABLE 2.
DESCRIPTIVE STATISTICS

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
lip1	77	8,95	9,73	9,3514	,17849	-,208	,274	,837	,541
lip2	77	8,70	9,73	9,3664	,19120	-,769	,274	1,687	,541
lip3	77	8,96	9,72	9,3479	,17920	-,306	,274	,258	,541
lip4	77	8,97	9,71	9,3040	,18300	,271	,274	,184	,541
lip5	77	8,92	9,68	9,2318	,22173	,409	,274	-,820	,541
lip6	77	8,80	9,62	9,1817	,24257	,213	,274	-1,382	,541
lip7	77	8,80	9,70	9,1992	,23938	,155	,274	-1,209	,541
lip8	77	8,86	9,66	9,2168	,22200	,243	,274	-1,230	,541
lip9	77	9,00	9,72	9,2887	,19556	,481	,274	-,369	,541
lip10	77	9,06	9,72	9,3048	,17785	,618	,274	-,018	,541
lip11	77	9,06	9,73	9,3365	,16914	,659	,274	,027	,541
lip12	77	9,07	9,80	9,3416	,16951	1,076	,274	1,316	,541

You can choose one or more of the following subgroup statistics calculated for the variables within each separate category of each grouping variable: sum, number of cases, mean, median, group median, standard error of the mean, minimum and maximum values, range, grouping variable value for the first category, the value of the grouping variable for the last category, standard deviation, variance, kurtosis, standard error of kurtosis, skewness, standard error of skewness, percentage of the total, percentage of total N, percentage of the sum, geometric and harmonic mean. You can change the order in which subgroup statistics are displayed.

Acknowledgments

The authors are sincerely grateful to: Petr Pavlov for valuable comments and help in compiling the article.

Conclusions

Observations of the activity of hydrogeological processes show that the periods of process variations are scattered chaotically on the time axis. According to their schedule, it is impossible to definitely speak about the regularity in the duration of the periods of variations, and in the alternation of periods of calm with a period of high hydrogeological activity. The impetus for this study was the desire to analyze the structure of a number of formal methods to find statistical patterns in the variations of hydrogeological data over time. Spatial and time series models can be used to study the dynamics of events. A spatial model describes a set of parameters at a given moment in time. A time series is a series of regular observations of a certain parameter at successive points in time or at intervals of time. In this work, the time series model is used to study the structure of hydrogeological events. In general, the purpose of studying a time series is to identify patterns in the change in the levels of a series and build its model in order to predict and study the relationships between phenomena.

References

[1] Anderson, T., Statistical Time Series Analysis. Mir, Moscow, 1971, 746.

[2] Box J., Jenkins G., Time Series Analysis. Forecast and management. Issue I. Moscow. Peace. 1974.406. Brillinger D., Time series. Data processing and theory. Mir, Moscow, 1980, 532.

[3] Volodin I.N., Lectures on Probability Theory and Mathematical Statistics, Kazan, 2006, 270.

[4] Kendal M., Time series. Moscow. 1981.198.

[5] Kislyak N.V. Econometrics. Yekaterinburg. 2007, 157.

[6] Kolmogorov A.N., Fomin S.V., Elements of the theory of functions and functional analysis. Moscow, Nauka, 1976, 542.

[7] Cramer Harald. Mathematical methods of statistics. Moscow. Peace. 1975.648.

[8] Marno Verbic. A guide to modern econometrics. Moscow 2008, 616.

[9] Seigno P.S., Theory of Imovarities and Mathematical Statistics. Kiev, Knowledge, 2007, 558.

[10] Smirnov, I.V. Dunin-Barkovskiy N.V., Course of Probability Theory and Mathematical Statistics. Science, Moscow, 1965, 511

[11] Piskunov N.S., Differential and integral calculus. Moscow, Nauka, 1964, 312.

[12] Hennan E., Multidimensional time series. Mir, Moscow, 1974, 575.

[13] V. V. Khristianovsky Time series analysis in economics: application practice: 2011, 127.

[14] Tsapaeva S.A. Fourier Series, Veliky Novgorod, 2011.

[15] Shanchenko N.I., Lectures on econometrics. Ulyanovsk. 2008, 139.

[16] Yushchenko D.P., O.V. Yakubovich, Mathematical Analysis. Fourier series, Gomel, 2008, 148.

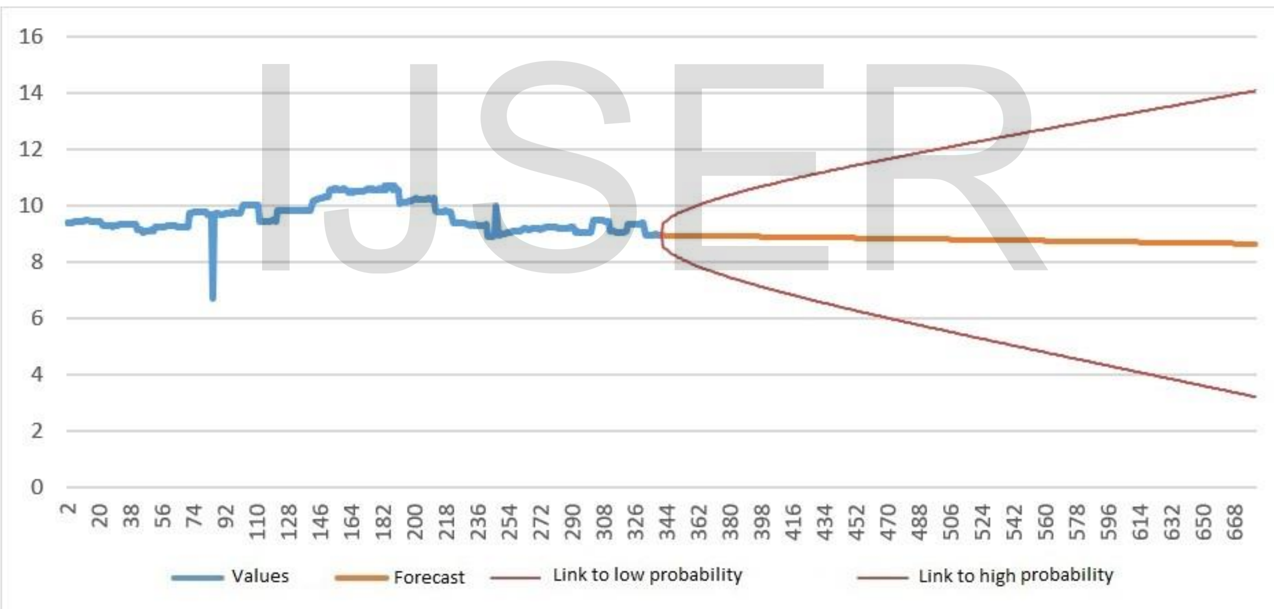
[17] Backhaus K. et al., Multivariate Analysemethoden, Springer – Verlag Berlin Heidelberg, 2011, 120–154.

[18] Bernhardt Christine. Modellierung von Elektrizitätspreisen durch lineare Zeitreihenmodelle und Value–at–Risk–Schätzung mittels Methoden aus der Extremwerttheorie Technische Universität München. Zentrum Mathematik. München. 2007, 97.

[19] Burtiev R. Z., Time Series in the Study of Seismic Regime of Vrancea (Romania) Seismic Zone. The Global Environmental Engineers, 2014, Volume. 1. N2, Karachi, Pakistan, 54–63

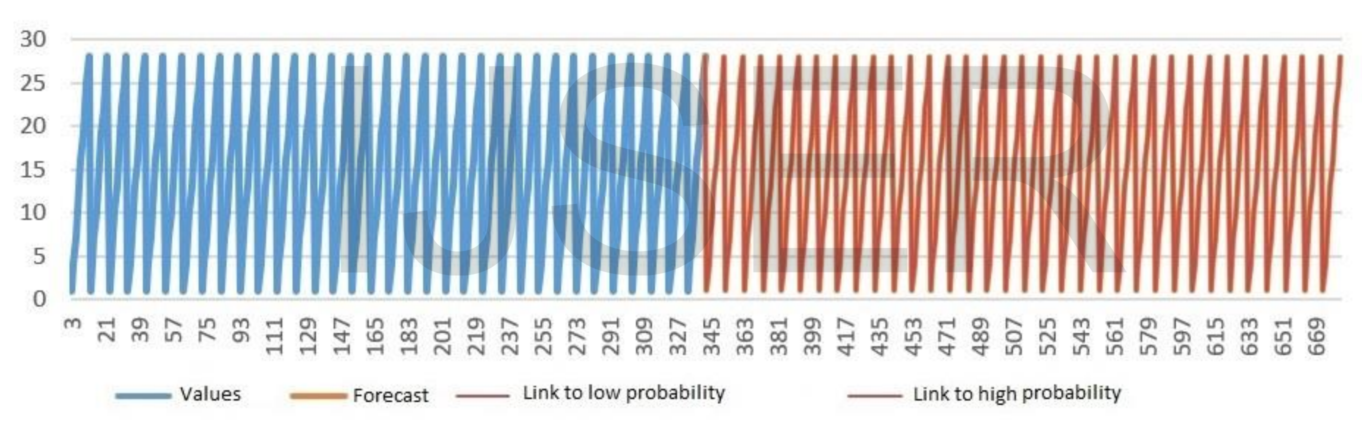
[20] Spezialvorlesung Zeitreihenanalyse–Mit Beispielen in Mathematica Institut fur Stochastik, Johannes Kepler Universität Linz. Linz, 2006, 277.

Forecast January

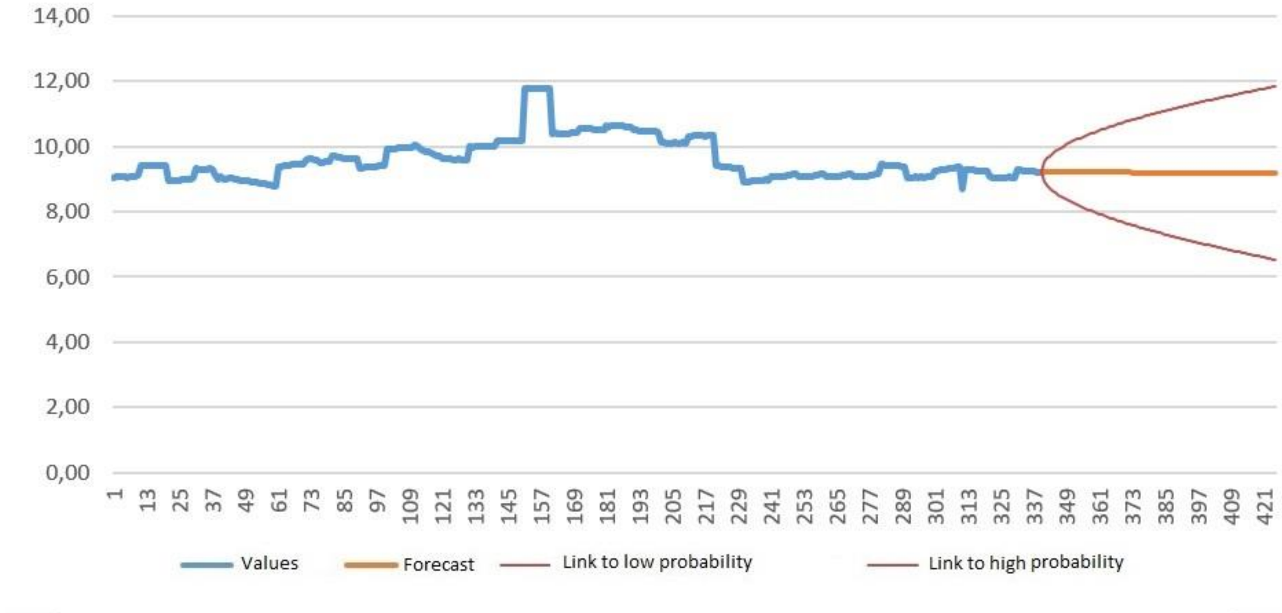


Forecast February

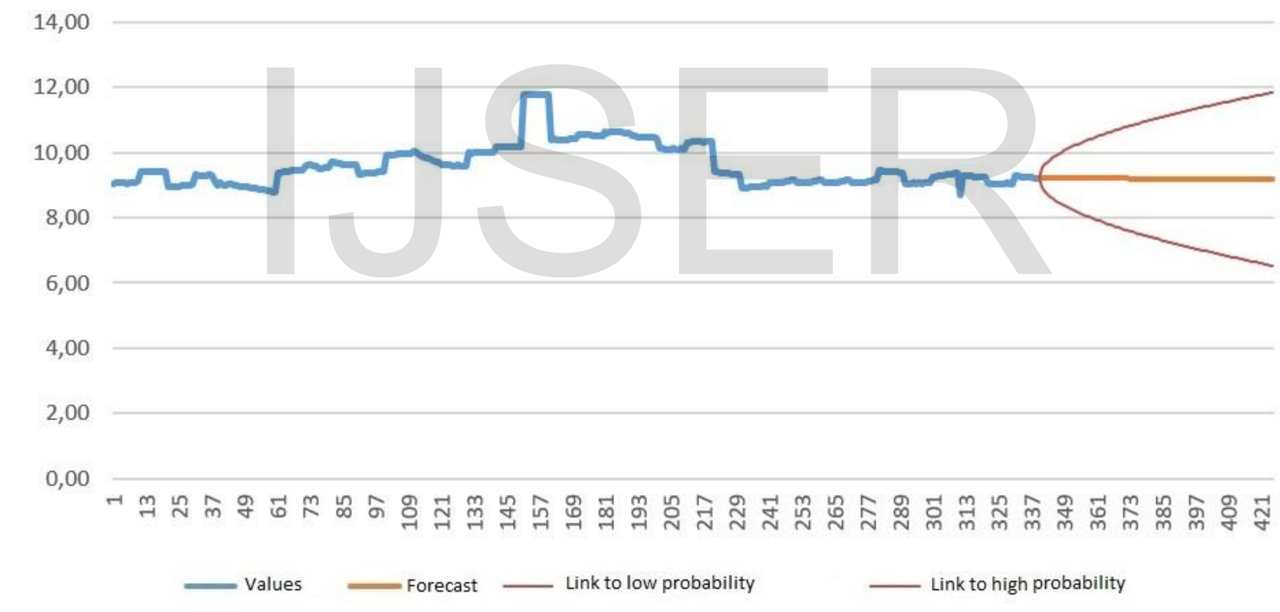
Forecast March



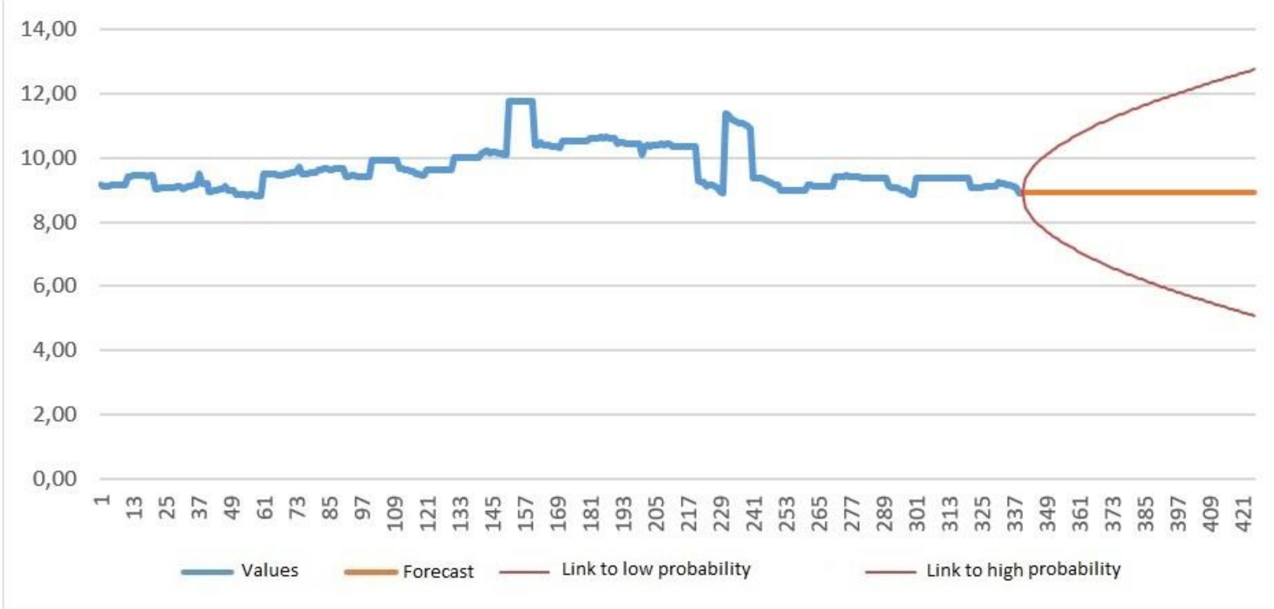
Forecast April



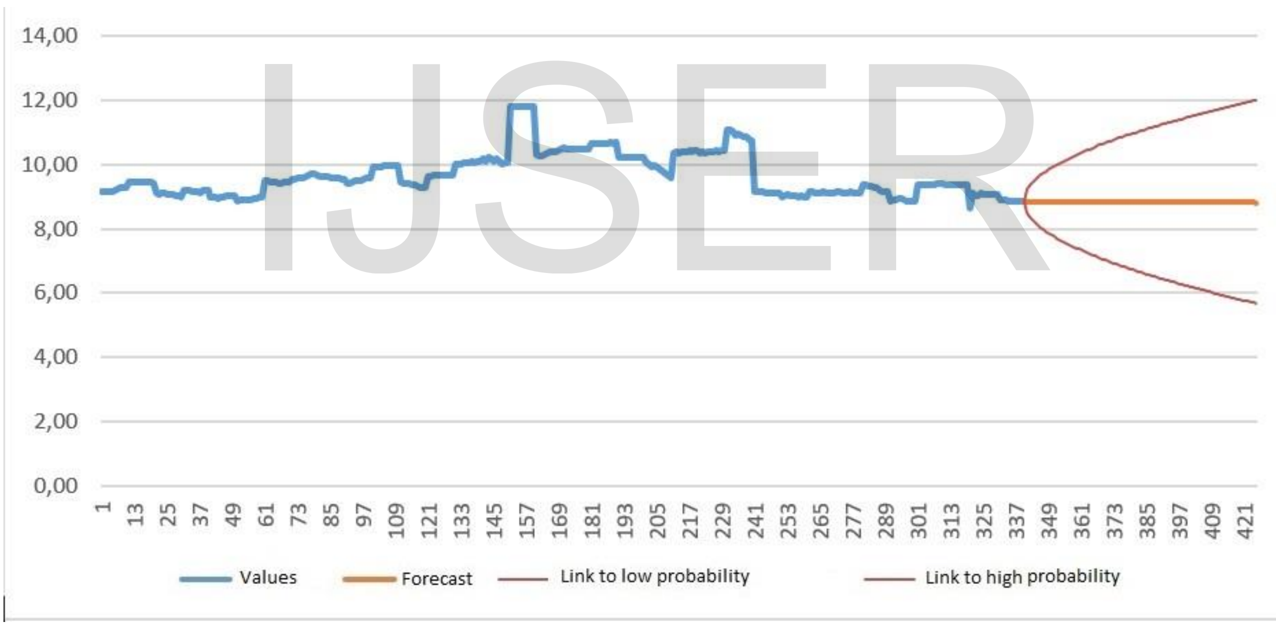
Forecast May



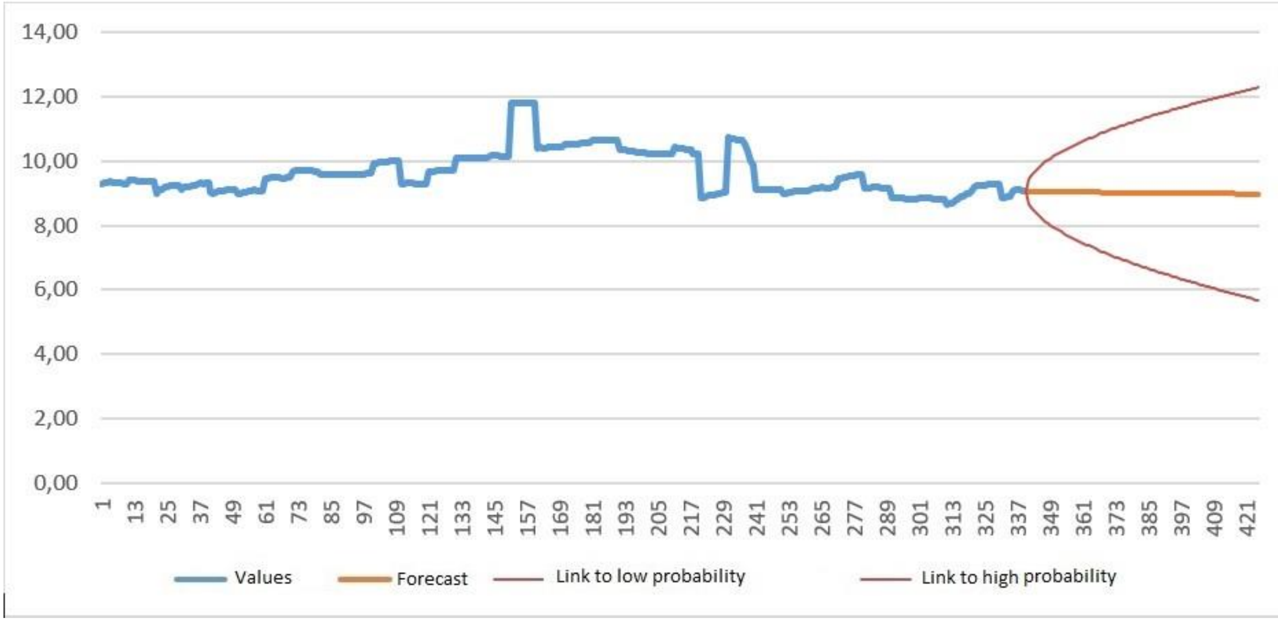
Forecast June



Forecast July

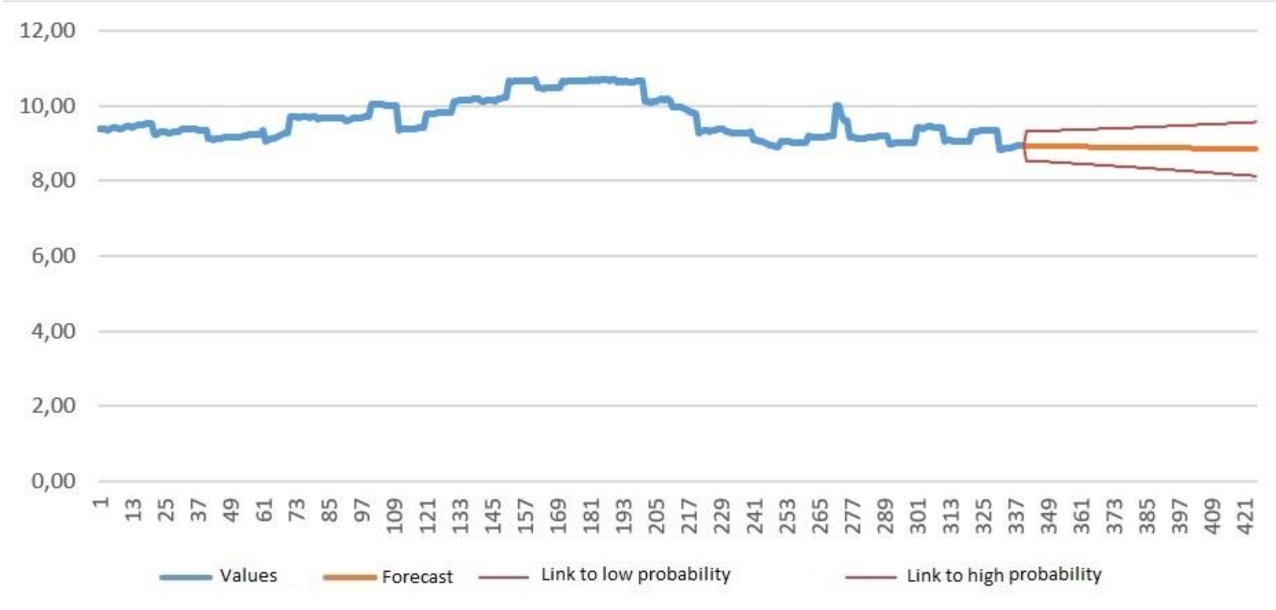


Forecast August

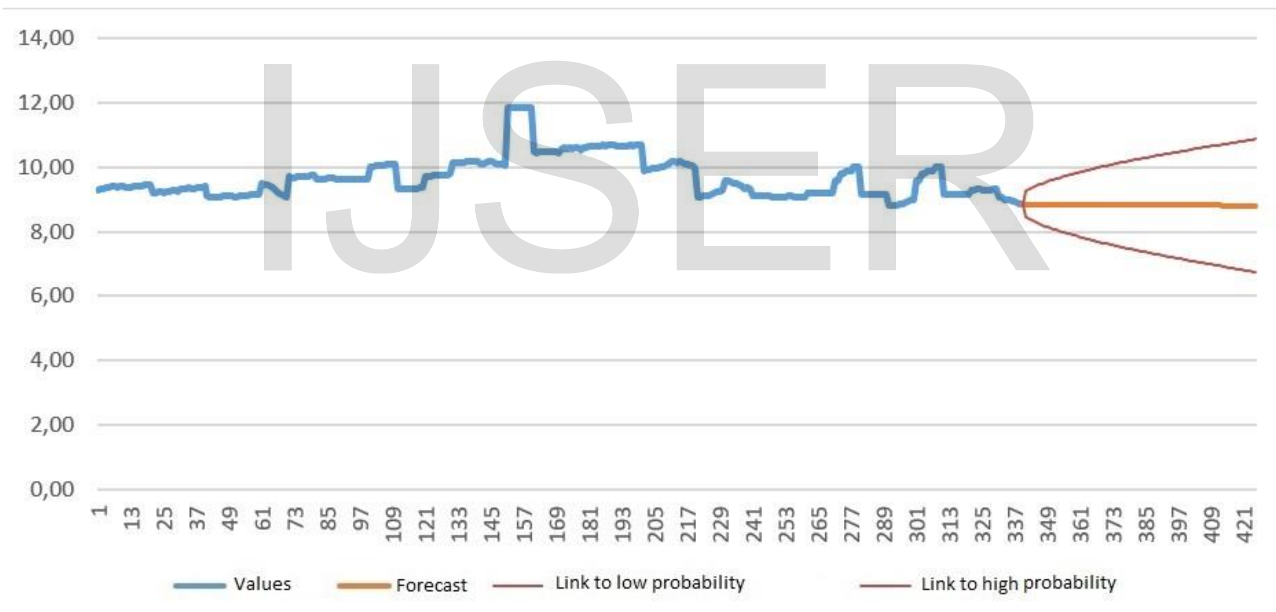


Forecast September

Forecast October



Forecast November



Forecast December