INVESTIGATION OF DISCONTINUITIES SPACING HISTOGRAMS BY THE USE MACHINE LEARNING METHOD

Article · March 2012

CITATIONS
0

READS
532

3 authors:

Yusuf Uzun
Necmettin Erbakan Üniversitesi
88 PUBLICATIONS **45** CITATIONS
SEE PROFILE

Alparslan Turanboy
Necmettin Erbakan Üniversitesi
14 PUBLICATIONS **52** CITATIONS
SEE PROFILE

Gülay Tezel
Selcuk University
33 PUBLICATIONS **454** CITATIONS
SEE PROFILE

INVESTIGATION OF DISCONTINUITIES SPACING HISTOGRAMS BY THE USE MACHINE LEARNING METHOD

Yusuf Uzun[1], Alparslan Turanboy[2], Gülay Tezel[3]
[1,2]Selçuk University, Seydişehir Vocational School of Higher Education, Konya, Turkey, yuzun@selcuk.edu.tr, a.turanboy@selcuk.edu.tr.
[3]Selçuk University, Computer Engineering Department, Kampus, Konya, Turkey, gtezel@selcuk.edu.tr

ABSTRACT

Discontinuities are major geological features in the rock mass and discontinuity spacing is one of the important parameters in describing the rock mass. Relation between discontinuity spacing and relative spacing has described by different curve fittings. These curve fittings will show the type (negative exponential, log-normal or normal distribution) of the statistical distribution as histograms. Discontinuity spacing and frequency data obtained at a field site in southern Seydişehir (Turkey). Sampling methods vary from one study to another (core sampling, scan-line survey, aerial photography). In this study, the possible distributions of discontinuity spacing along a straight line through a rock mass are considered. In this study, 5 different drilling sampling have been used. We have examined discontinuity spacing and relative spacing relations that obtained from these core sampling with using machine learning method. Machine learning, a branch of artificial intelligence, is a scientific discipline that is concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. Minitab (LEAD Technologies, Inc) and Weka (Waikato Environment for Knowledge Analysis) software where preferred in all analyses and interprets. Results of study, different empirical equation for each histogram have been constituted. Machine learning method has been treated on the obtained equations and reached interesting outputs.

Keywords: *discontinuity, discontinuity analysis, discontinuity distribution, machine learning.*

1. INTRODUCTION

Discontinuity spacing is an important factor for rock mass description. This feature of rock mass is a perpendicular distance between two adjacent discontinuities, The spacing of discontinuity largely controls the size of individual rock blocks and therefore it is the most important block size parameter. (ISRM, 1978) It is largely used

conventional statistical and geostatistical methods for establishing of discontinuity spacing distributions in many researches.

For description of discontinuity spacing distributions, three samples of distribution rule can be summarized. (Rives et al.,1992) in statistic. Many researchers has been used these statistical rules to describing of discontinuity spacing distributions. Distributions of discontinuity spacing were described negative exponential (Hudson 1976; Villaescuse and Browri, 1990), log-normal (Sen and Kazi, 1984; Narr and Suppe, 1991), normal (Huang and Angelier,1989) distribution and combination of these (Senyur,1990).

Computer applications widely used in many statistical analyses recently. Especially in geotechnical analyses, many computer tools have been successfully applied. One of the important applications used in evaluated of discontinuity spacing analyzed that is a indispensable tools in many rock mechanics. Minitab is the one of important software will be able to use in these applications. Minitab software (LEAD Technologies, Inc) was used for the analysis of drilling data distribution (normal, log-normal, exponential). For linear regression analysis the Weka machine learning library was used. Weka stands for `Waikato Environment for Knowledge Analysis' and is open-source software issued under the GNU General Public license. In data mining, Weka is a program that evaluates the data analysis to develop models to support on decision making on business warehouse management. There are many data mining techniques for model developing. Among the most popular ones are Classification, Clustering and Association Rule Discovery which are applied in model developing (Witten and Frank, 2005).

2. MATERIALS AND METHODS

In this chapter, several statistical tools and machine learning methods used in the methodology will be briefly explained.

In probability theory, the normal distribution is a continuous probability distribution that is often used as a first approximation to describe real-valued random variables that tend to cluster around a single mean value. The graph of the associated probability density function is "bell"-shaped, or bell curve:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad (1)$$

where parameter μ is the mean (location of the peak) and σ^2 is the variance (the measure of the width of the distribution). The distribution with μ= 0 and σ^2= 1is called the standard normal. (Jagdish and Campbell, 1996)

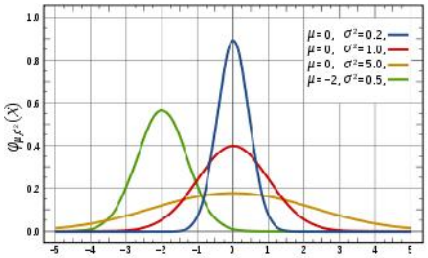

Figure 1. Standart Normal Distribution

In probability theory, a log-normal distribution is a probability distribution of a random variable whose logarithm is normally distributed. If *X* is a random variable with a normal distribution, then *Y*=exp(*X*) has a log-normal distribution; likewise, if *Y* is log-normally distributed, then *X*=log(*Y*) is normally distributed. (This is true regardless of

the base of the logarithmic function: if $\log_a(Y)$ is normally distributed, then so is $\log_b(Y)$, for any two positive numbers a, $b \neq 1$.) (Aitchison and Brown, 1957)

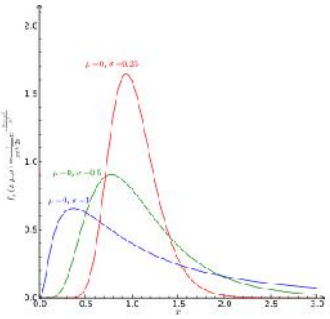

Figure 2. Log- Normal Distribution

In probability theory and statistics, the exponential distribution (a.k.a. negative exponential distribution) is a family of continuous probability distributions. It describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. (Schmidt and Makalic, 2009)

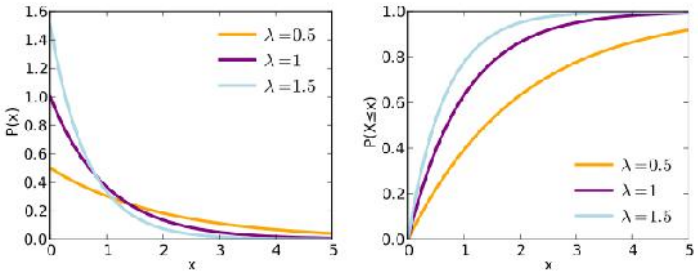

Figure 3. Exponential Distribution

3. MACHINE LEARNING

Mitchell (1997) provided a widely quoted definition: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Machine learning algorithms have proven to be of great practical value in a variety of application domains. They are especially useful in (a) data mining problems where large databases may contain valuable implicit regularities that can be discovered automatically (e.g., to analyze outcomes of medical treatments from patient databases or to learn general rules for credit worthiness from financial databases); (b) poorly understood domains where humans might not have the knowledge needed to develop effective algorithms (e.g., human face recognition from images); and (c) domains where the program must dynamically adapt to changing conditions (e.g., controlling manufacturing processes under changing supply stocks or adapting to the changing reading interests of individuals) (Mitchell, 1997).

4. CORRELATION COEFFICIENT

A single summary number that gives you a good idea about how closely one variable is related to another variable. (Mitchell, 1997) The formula for correlation coefficient:

$$r_{xy} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{(SS_x)(SS_y)}} \text{ or } r_{xy} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[(\sum X^2 - \frac{(\sum X)^2}{n_x})(\sum Y^2 - \frac{(\sum Y)^2}{n_y})\right]}}, \quad (2)$$

where ΣX is tells you to add up all the X scores, ΣY tells you to add up all the Y scores , ΣX^2 tells you to square each X score and then add them up, ΣX^2 tells you to square each Y score and then add them up, ΣXY tells you to multiply each X score by its associated Y score and then add the resulting products together (this is called a “cross-products”), and refers to the number of “pairs” of data you have.

All correlation coefficients range from -1.00 to +1.00. A correlation coefficient of -1.00 tells you that there is a perfect negative relationship between the two variables. This means that as values on one variable increase, there is a perfect predictable decrease in values on the other variable. In other words, as one variable goes up, the other goes in the opposite direction (it goes down).

A correlation coefficient of +1.00 tells you that there is a perfect positive relationship between the two variables. This means that as values on one variable increase, there is a perfect predictable increase in values on the other variable. In other words, as one variable goes up so does the other.

A correlation coefficient of 0.00 tells you that there is a zero correlation, or no relationship, between the two variables. In other words, as one variable changes (goes up or down) you can't really say anything about what happens to the other variable.

5. REGRESSION ANALYSIS

Regression analysis is a statistical methodology that is most often used for numeric prediction. In fact, many texts use the terms "regression" and "numeric prediction" synonymously. Regression analysis can be used to model the relationship between one or more *independent* or predictor variables and a *dependent* or response variable.

Straight-line regression analysis involves a response variable, y, and a single predictor variable, x. It is the simplest form of regression, and models y as a linear function of x. (Quinlan, 1987) That is,

$$y = b+wx \,, \tag{3}$$

where the variance of y is assumed to be constant, and b and w are regression coefficients specifying the Y-intercept and slope of the line, respectively. The regression coefficients, w and b, can also be thought of as weights, so that we can equivalently write,

$$y = w_0 + w_1 x, \tag{4}$$

These coefficients can be solved for by the method of least squares, which estimates the best fitting straight line as the one that minimizes the error between the actual data and the estimate of the line. Let D be a training set consisting of values of predictor variable, x, for some population and their associated values for response variable, y. The training set contains $|D|$ data points of the form (x_1, y_1), (x_2, y_2), ..., $(x_{|D|}, y_{|D|})$. The regression coefficients can be estimated using this method with the following equations:

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2}, \tag{5}$$

$$w_0 = \bar{y} - w_1 \bar{x}, \tag{6}$$

where $\bar{x}$ is the mean value of $x_1, x_2, \ldots, x_{|D|}$, and $\bar{y}$ is the mean value of $y_1, y_2, \ldots, y_{|D|}$. The coefficients w_0 and w_1 often provide good approximations to otherwise complicated regression equations.

6. PREDICTOR ERROR MEASURES

Let D^T be a test set of the form $(\boldsymbol{X_1}, y_1)$, $(\boldsymbol{X_2}, y_2)$, . . . , $(\boldsymbol{X_d}, y_d)$, where the $\boldsymbol{X_i}$ are the n-dimensional test tuples with associated known values, y_i, for a response variable, y, and d is the number of tuples in D^T. Since predictors return a continuous value rather than a

categorical label, it is difficult to say *exactly* whether the predicted value, y_i', for $\boldsymbol{X_i}$ is correct. Instead of focusing on whether y_i' is an "exact" match with y_i, we instead look at how far off the predicted value is from the actual known value. Loss functions measure the error between y_i and the predicted value, y_i'. (Mitchell, 1997) The most common loss functions are:

Absolute error: $|y_i - y_i'|$, (7)

Squared error: $(y_i - y_i')^2$, (8)

Based on the above, the test error (rate), or generalization error, is the average loss over the test set. Thus, we get the following error rates.

Mean absolute error: $\frac{\sum_{i=1}^{d} |yi - y_i'|}{d}$, (9)

Mean squared error: $\frac{\sum_{i=1}^{d} (yi - y_i')^2}{d}$, (10)

The mean squared error exaggerates the presence of outliers, while the mean absolute error does not. If we were to take the square root of the mean squared error, the resulting error measure is called the root mean squared error. This is useful in that it allows the error measured to be of the same magnitude as the quantity being predicted. Sometimes, we may want the error to be relative to what it would have been if we had just predicted *y*, the mean value for *y* from the training data, *D*. That is, we can normalize the total loss by dividing by the total loss incurred from always predicting the mean. Relative measures of error include:

Relative absolute error: $\frac{\frac{\sum_{i=1}^{d}|y_i-y_i'|}{d}}{\frac{\sum_{i=1}^{d}|y_i-\bar{y}|}{d}}$ (11)

Relative squared error: $\frac{\frac{\sum_{i=1}^{d}(y_i-y_i')^2}{d}}{\frac{\sum_{i=1}^{d}(y_i-\bar{y})^2}{d}}$ (12)

where y is the mean value of the y_i's of the training data, that is

$$\bar{y}=\frac{\sum_i^l l^{y_i}}{d} \quad (13)$$

We can take the root of the relative squared error to obtain the root relative squared error so that the resulting error is of the same magnitude as the quantity predicted. In practice, the choice of error measure does not greatly affect prediction model selection.

7. DRILLING DATA SET PARAMETERS

Sampling parameters that obtained from examined field are shown in table 1. Whole data set that attributed methodology is presented in this table.

Table 1. Discontinuity spacing (start and end) and frequency values of the drillings

Drilling 1			*Drilling 2*			*Drilling 3*			*Drilling 4*			*Drilling 5*		
Start	*End*	*Freq.*	*Start*	*End*	*Freq.*	*Start*	*End*	*Freq.*	*Start*	*End*	*Freq.*	*Start*	*End*	*Freq.*
0	7	2	0	10	9	0	12	7	0	10	18	0	8	0
7	14	5	10	20	7	12	24	10	10	20	17	8	16	5
14	21	15	20	30	3	24	36	19	20	30	6	16	24	0
21	28	5	30	40	6	36	48	14	30	40	9	24	32	3
28	35	5	40	50	2	48	60	5	40	50	5	32	40	3
35	42	2	50	60	2	60	72	9	50	60	7	40	48	3
42	49	3	60	70	1	72	84	1	60	70	3	48	56	2
49	56	1	70	80	0	84	96	0	70	80	1	56	64	2
56	63	3	80	90	0	96	10 8	1	80	90	1	64	72	2
63	70	2	90	10	1	108	12	2	90	10	2	72	80	1

				0			0			0				
				11			13			11				
70	77	0	100	0	0	120	2	1	100	0	2	80	88	2
				12			14			12				
77	84	2	110	0	1	132	4	1	110	0	0	88	96	1
				13			15			13			10	
84	91	1	120	0	2	144	6	0	120	0	0	96	4	3
				14			16			14			11	
91	98	1	130	0	0	156	8	0	130	0	0	104	2	1
	10			15			18			15			12	
98	5	1	140	0	0	168	0	0	140	0	0	112	0	1
	11			16			19			16			12	
105	2	0	150	0	0	180	2	0	150	0	0	120	8	1
	11			17			20			17			13	
112	9	0	160	0	0	192	4	0	160	0	0	128	6	0
	12			18			21			18			14	
119	6	0	170	0	0	204	6	0	170	0	0	136	4	0
	13			19			22			19			15	
126	3	0	180	0	0	216	8	0	180	0	1	144	2	0
	14			20			24			20			16	
133	0	0	190	0	1	228	0	0	190	0	1	152	0	0
	14						25						16	
140	7	1				240	2	1				160	8	1

8. STUDY RESULTS

Table 2 presents the obtained results from Minitab software. Here, at drilling 1, log-normal distribution with the best correlation coefficient value of 0,991 is shown. Also, at drilling 2, log-normal distribution with the best correlation coefficient value of 0,983 is presented. Additionally, at drilling 3 log-normal distribution with the best correlation coefficient value of 0,992 is shown. Last, at drilling 4 log-normal distributions with the best correlation coefficient value of 0,990 and at drilling 5, normal distribution with the best correlation coefficient value of 0,995 are presented.

Minitab software can analyze automatically which curves fit which distributions that have got start, end and frequency attributes data in drilling data sets. These probability plots and which curves belong to these plots that can be followed in figure 4, 6, 8, 10, 12.

In addition, classical descriptions of discontinuity spacing - frequency curves were constructed for each drilling samplings. These histograms can be followed in figure 5, 7, 9, 11, 13. Between suggested methodology and these classical illustrations will be commented in the conclusion.

Table 2. Drilling correlation coefficient values

Drillings	*Correlation Coefficient*		
	Log-Normal	*Exponential*	*Normal*
Drilling 1	0,991	-	0,943
Drilling 2	0,983	-	0,940
Drilling 3	0,992	-	0,946
Drilling 4	0,990	-	0,916
Drilling 5	0,950	-	0,995

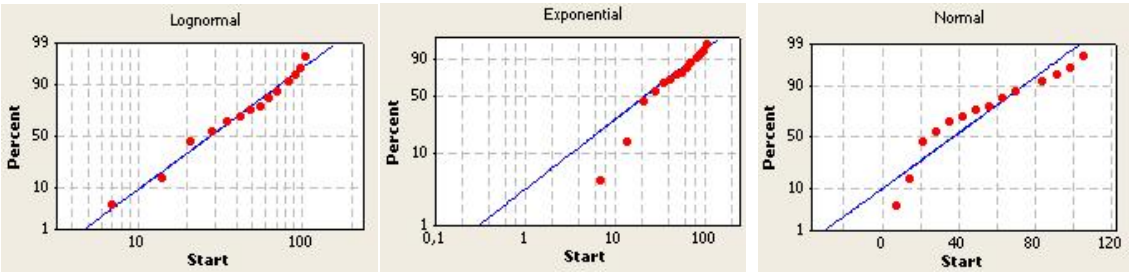

Figure 4. Lognormal, exponential and normal distribution curves for drilling 1

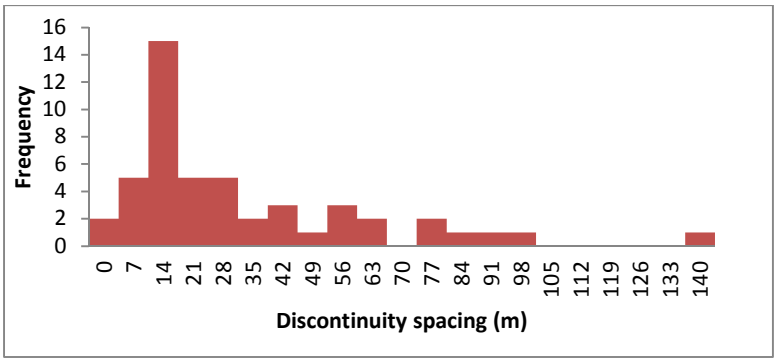

Figure 5. Discontinuity spacing histogram for drilling 1

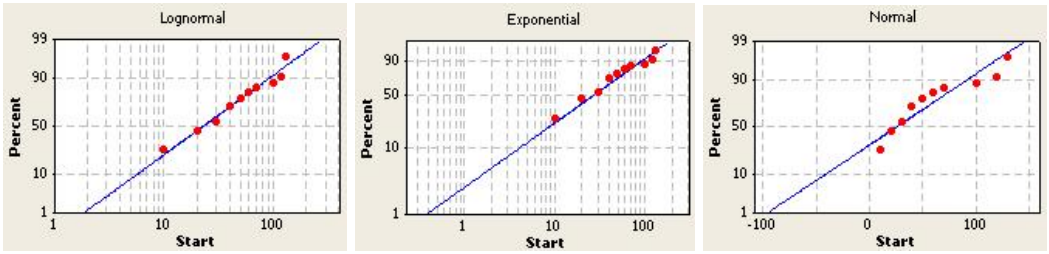

Figure 6. Lognormal, exponential and normal distribution curves for drilling 2

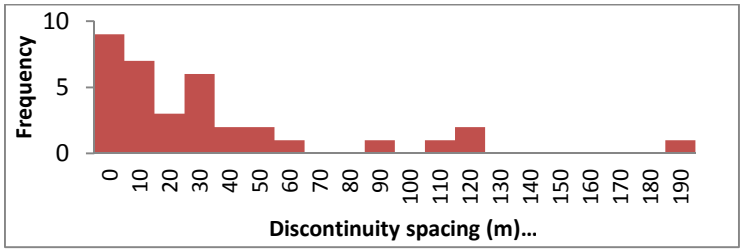

Figure 7. Discontinuity spacing histogram for drilling 2

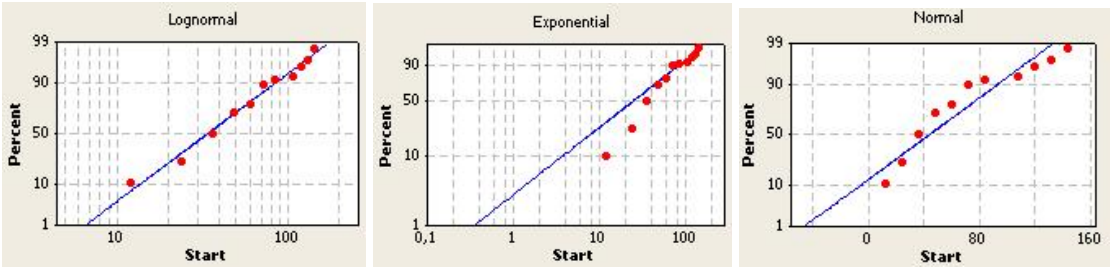

Figure 8. Lognormal, exponential and normal distribution curves for drilling 3

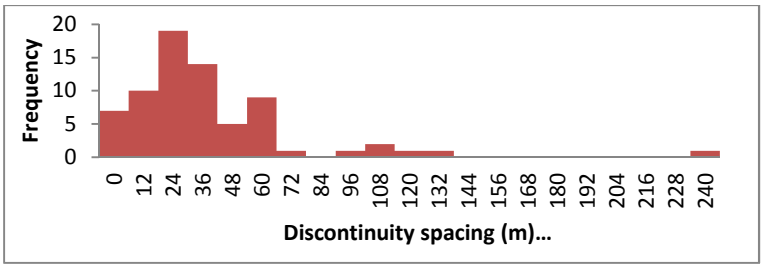

Figure 9. Discontinuity spacing histogram for drilling 3

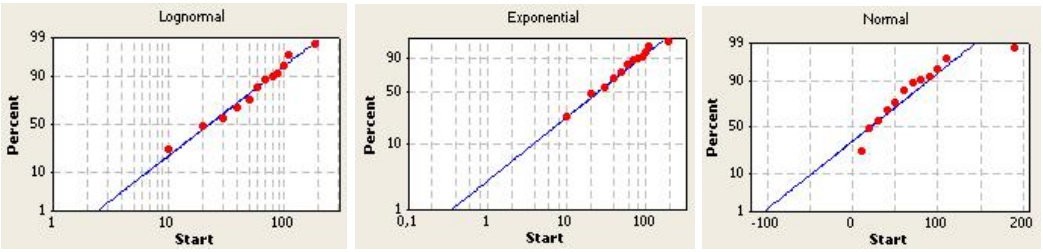

Figure 10. Lognormal, exponential and normal distribution curves for drilling 4

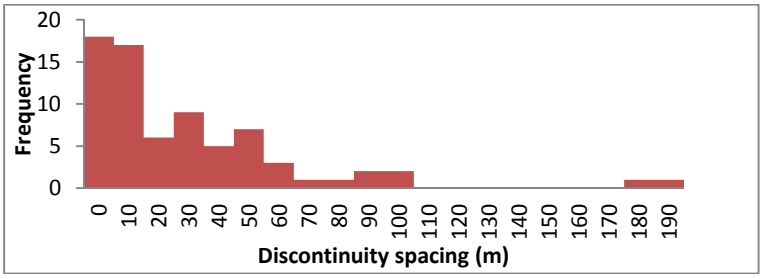

Figure 11. Discontinuity spacing histogram for drilling 4

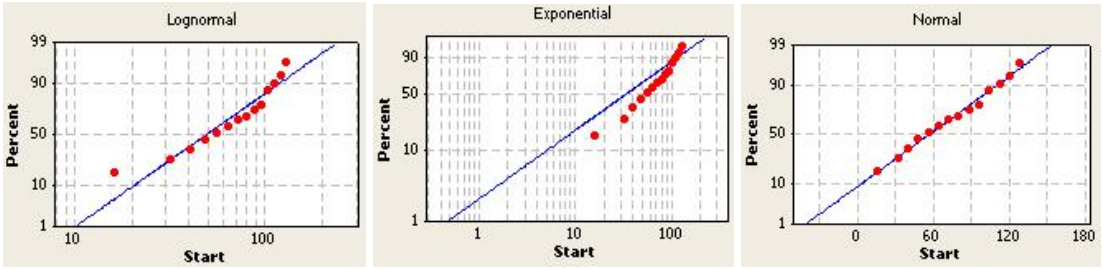

Figure 12. Lognormal, exponential and normal distribution curves for drilling 5

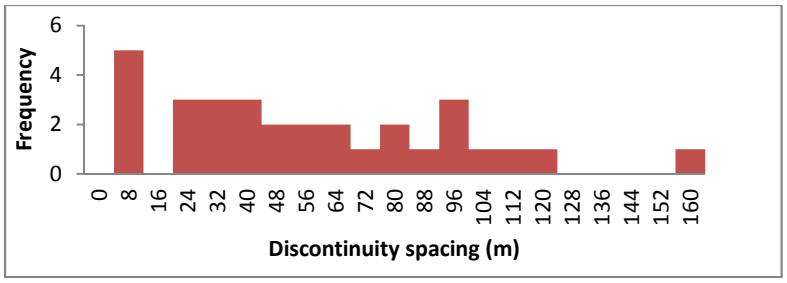

Figure 13. Discontinuity spacing histogram for drilling 5

9. SIMPLE LINEAR REGRESSION ANALYSIS

Simple linear regression analyses were performed with used Weka software. In the below, is showed regression models that produced by Weka for drillings at table 3 and numeric prediction values that produced by Weka for drillings at table 4.

Table 3. Regression models for drillings

Drillings	*Simple Linear Regression Models*
Drilling 1	Frequency= 5,74- 0,05*Start
Drilling 2	Frequency= 4,81- 0,03*Start
Drilling 3	Frequency= 9,64- 0,05*Start
Drilling 4	Frequency= 10,43- 0,07*Start
Drilling 5	Frequency= 2,63- 0,01*Start

Table 4. Numeric predictions for drillings

Drillings	*Correlation coefficient*	*Mean absolute error*	*Root mean squared error*	*Relative absolute error (%)*	*Root relative squared error (%)*
Drilling 1	0,473	1,571	2,943	70,32	85,22
Drilling 2	0,643	1,564	1,960	79,78	73,39
Drilling 3	0,657	3,009	3,987	70,15	72,99
Drilling 4	0,693	2,938	3,831	69,86	69,17
Drilling 5	0,286	0,946	1,322	80,22	94,85

The results of simple linear regression analyses can be followed as a WEKA interface for the drilling 1 in figure 14.

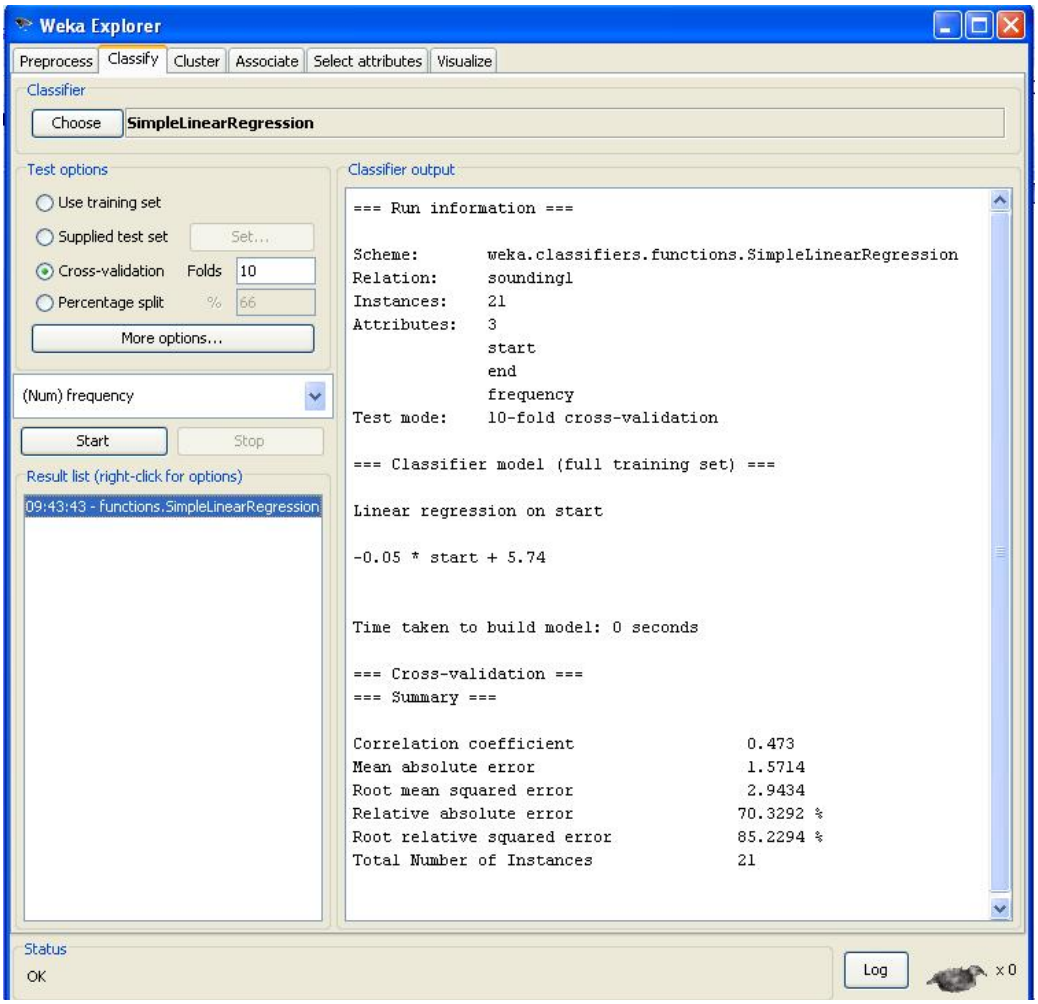

Figure 14. WEKA interface for simple linear regression analyses for drilling 1

This gives the error levels during a 10- fold cross validation. In order to get statistically meaningful results, the default number of iterations is 10. In case of 10-fold cross-validation this means 100 calls of one classifier with training data and tested against test data. In the above in table 4, is showed the best correlation coefficient value of 0,693 in drilling 4. Therefore, it can be seen a positive relationship between frequency and start

variables in drilling 4. It is clear that if correlation coefficient value increases, predictable will be naturally increase.

The Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE=MAE, then all the errors are of the same magnitude. Lower values are better. In the above in table 4, is showed the minimum MAE value of 0,946 and the minimum RMSE value of 1,322 in drilling 5. In the above at table 4, is showed the minimum relative absolute error value of 69,86% and the minimum root relative squared error value of 69,17% in drilling 4.

In the study in addition, relations between predicted frequency that obtained from Weka and start attributes in the data set were constructed as a new approach. In the below in figure 15, is showed the best linear regression distribution in drilling 4 and is showed the worst linear regression distribution in drilling 5.

In addition, as a seen there are strong relations between frequency and start attributes. These relations were supported with results obtained from simple linear regression algorithm. In simple linear regression method were also achieved the same strong relations. These relations can be seen in table 3, 4 and figure 15 for each drilling samples.

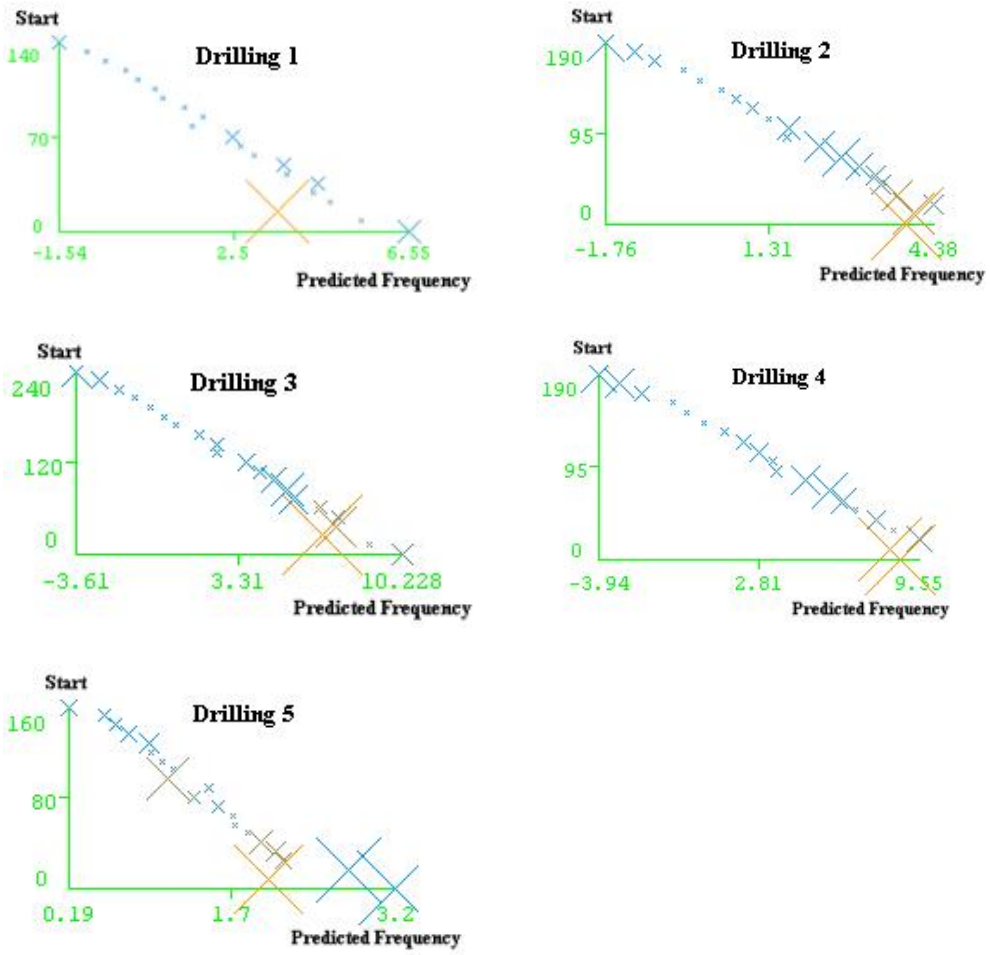

Figure 15. Linear regression graphics for drillings

10. RESULTS AND DISCUSSIONS

This paper presents distribution analysis and numeric predictions on the drilling data set. For numeric predictions is used linear regression algorithm that this is a kind of machine learning techniques. In here, in order to determine the relations between frequency properties and other properties, linear regression models were constituted.

In this resource we presented a new procedure for predictions distribution of discontinuity spacing with using Minitab and Weka software as a main tolls. Obtained results can be followed:

We used to three different probability plots (with distributions chosen from (negative exponential, normal and lognormal) to help determine which of these distributions best fits sampling data. Using the accurate testing methods, Discontinuity spacing curves always don't fit the same distribution rules. This inference has also been pointed out by many authors; we have been presented with using different procedure. This procedure is machine learning method. In our study, simple linear regression algorithm was used to accurate test results of well known classical methods alternatively. The results of these methods have more compatible and close relations. In the study we suggested that more realistically, fast, powerful, effective and simplest results can be obtained. Therefore, additionally, we suggest that must be used these methods together for many advantages in this area.

REFERENCES

Aitchison, J. & Brown, J.A.C. 1957. *The Lognormal Distribution*. Cambridge University Press.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, P.J. 1984. *Classification and regression trees.* Belmont, CA: Wadsworth International Group

Jagdish, K. & Campbell, B. 1996. *Handbook of the normal distribution* (2nd ed.). CRC Press. ISBN 0-824-79342-0.

Huang, Q. & Angelier, J., 1989, Fracture spacing and its relation to bed thickness, Geol. Mag., 126, pp:355-362

Hudson, J.A., Rock mechanics principles in engineering practice, pp:31-35

Mitchell, T.M. 1997. *Machine Learning*. McGraw-Hill Science

Quinlan, J.R. 1987. *Rule induction with statistical data- a comparison with multiple regression.* Journal of the Operational Research Society. 38: 347-352.

Rives, T., Razack, M., Petit, P. & Rawnsley, K.D., 1992, *Joint spacing: Analogue and numerical simulations*, Journal of Structural Geology. Volume. 14, No:8/9:925-937

Schmidt, D.F. & Makalic, E. 2009. *Universal Models for the Exponential Distribution*. IEEE Transactions on Information Theory. Volume 55: 3087-3090
Sen, Z. & Kazi, A.,1984, *Discontinuity spacing and RQD estimates from finite length scanlines*, Int. Jour. Rock. Mech & Mining Sci., 21:203-212.
Villaescuse, E. & Brown E.T., 1990, *Characterizing joint special correlation using geostatistical methods*, Rock Joints (edited by Barton, N. & Stephansson, O.), Balkema Rotterdam, 115-122
Witten, I.H. &Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques.* (2nd ed.). Morgan Kaufmann, San Francisco.