GROUND'2008

&

3^{rd} LPE

International Conference on Grounding and Earthing
&
3^{rd} International Conference on
Lightning Physics and Effects

Florianopolis - Brazil November, 2008

A HYBRID NON-LINEAR REGRESSION MODEL FOR THE ESTIMATION OF THE CORRELATION BETWEEN THE ELECTRICAL PARAMETERS OF SOIL AND THE SOIL CRITICAL ELECTRIC FIELD

Fani E. Asimakopoulou Ioannis F. Gonos George J. Tsekouras Ioannis A. Stathopulos
High Voltage Laboratory, National Technical University of Athens, Greece

Abstract – This paper aims to the verification of a correlation between the electrical parameters (soil resistivity and soil dielectric constant) of soil and the soil critical electric field. For that purpose, a hybrid non-linear regression model is described, implemented and applied on experimental data. This method takes into consideration the correlation analysis of the selected input values, which have been already transformed by implementing appropriate transformation functions, and finally forms an equation, between the dependent and independent values. The results are compared to those obtained from other regression methods. Furthermore, the parameters of the proposed equation have been optimized by using a Genetic Algorithm.

1 - INTRODUCTION

The injection of a high impulse current in a grounding system causes soil ionization phenomena, which lead to reduction of the impulse impedance of the grounding system. This fact was first introduced by Towne in 1929 [1].

However, various values for the soil critical electric field (E_c) have been suggested by numerous researchers. E_c ranges between 160-520kV/m in the experiments conducted by Towne. E_c was calculated by Bellaschi *et al*. [2], [3] in the range of 120-420kV/m. In 1974 Liew and Darveniza [4] used a value of 300kV/m for E_c. Loboda *el al.* [5], [6] investigated the electrical properties of different soils injected with current pulses and calculated E_c between 560-900kV/m. In impulse tests of several types of soil, which were conducted by Oettle [7], E_c varied in the range of 600-1850kV/m and 600-800kV/m for soils with higher moisture contents. The inhomogeneity of the soil also affects the value of E_c (in a homogenous soil E_c falls approximately 50%). Therefore, a value of 1MV/m was suggested. The value of 400kV/m is used by CIGRE[8]. Mousa [9] suggested that 300kV/m should be used for E_c. Gonos and Stathopulos [10] studied the variation of E_c against the soil resistivity and E_c was found to be approximately 200kV/m.

Since the electrical characteristics of soil vary among different soil types, the adoption of one value for E_c is not suggested. Therefore effort has been made by Manna and Chowdhuri [11] in order to study the influence of significant soil parameters to the value of E_c. The equation, that the researchers proposed, is the following:

$$E_c = 8.6083 \cdot k_g^{-0.0103} \cdot \rho_g^{0.1526} \qquad (1)$$

where E_c (kV/cm), k_g is the soil dielectric constant (dimensionless quantity) and ρ_g is the soil resistivity (kΩm).

2 - BASIC CONSIDERATIONS

In this paper a non-linear regression model [12] has been applied on the experimental results of [11], [13] in order for a relationship between the soil resistivity (ρ_g), soil dielectric constant (k_g) and the soil critical electric field (E_c) to be determined. The proposed model is capable of using more than one predictor variables and identifies the non-linear relationships between them and the response variable.

Furthermore, a Genetic Algorithm (GA) [14] has been used in order for the parameters of Manna's equation [11] and the one yielded by the regression model equation to be optimized. The optimization criterion, which has been used, is the mean absolute percentage error (MAPE).

2.1 - NON-LINEAR MULTIVARIABLE REGRESSION MODEL

A non-linear multivariable regression model was developed for the determination of the relationship between E_c (y) and the electric properties of soil (soil resistivity (x_1) and dielectric constant (x_2)). Through data processing the model selects a particular set of variables, which are examined for possible incorporation. The selected set of variables can be structured in vectors:

$$\vec{\boldsymbol{x}}_i = \left(\boldsymbol{x}_{i1},\ \boldsymbol{x}_{i2},\ \ldots,\ \boldsymbol{x}_{iN}\right)^T = \left(\boldsymbol{x}_{ij}, \boldsymbol{j} = 1,\ldots,\boldsymbol{N}\right)^T \qquad (2)$$

where x_{ij} is the i-th value of the j-th selected variable. There are m_1 vectors for training the model and m_2 for conducting the final estimation of E_c.

In Figure 1 the basic steps of the developed non-linear multi-variable regression model are outlined. As it can be observed, the main components of the model are the proper transformation of the variables, the correlation analysis and the model optimization.

The non-linear functions (x^a, $1/x$, $\ln(x)$, e^{-x}, where the parameter a belongs to set A and should be determined), which are used by the model, consist a selected function set $F = \left\{f_k\left(x_j\right) : k = 1 \ldots K\right\}$, where x_j is the j-th selected variable. Based on the selected function set F, the basic vector $\vec{X}$ is defined as:

$$\vec{X}_i = \left(1\ \ f_1\left(x_{i1}\right)\ldots f_k\left(x_{i1}\right)\ \ \ldots\ \ f_1\left(x_{iN}\right)\ldots f_k\left(x_{iN}\right)\right)^T \qquad (3)$$

having dimension w equal to $1+N \cdot K$, where N is the number of the input variables. Any linear combination of terms contained in vector $\vec{X}_i$ forms a basis for a candidate estimation model. The number of all possible combinations is $2^{N \cdot K}$ (the constant term always exists).

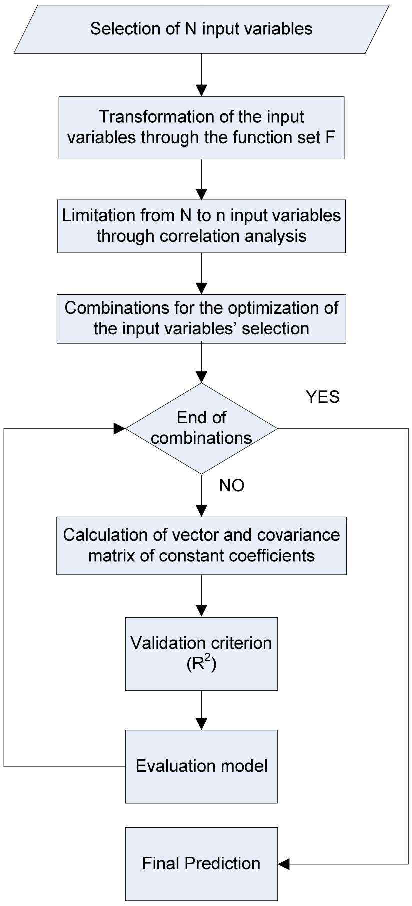
Selection of N input variables
Transformation of the input variables through the function set F
Limitation from N to n input variables through correlation analysis
Combinations for the optimization of the input variables' selection
End of combinations
YES
NO
Calculation of vector and covariance matrix of constant coefficients
Validation criterion (R^2)
Evaluation model
Final Prediction

Figure 1 - Developed non-linear multi-variable regression model

In order to reduce the number of candidate combinations, a correlation analysis is performed using the following basic steps:

- The correlation index between $f_k(x_j)$ and y is computed. If the index is greater than a pre-specified value cor_1, the term $f_k(x_j)$ is retained for further processing.
- For all terms being retained, a cross correlation analysis is performed. If the correlation index between two terms $f_k(x_j)$, $f_{k'}(x_{j'})$ (for $k \neq k'$ and $j \neq j'$ simultaneously) is smaller, than a pre-specified value cor_2, both terms are retained. In any other case, only the term with the largest correlation with respect to output y is retained. Therefore, all functions that have information overlap are eliminated through this cross correlation analysis and form the set F_c. The function having the smallest absolute correlation index with the variable y is removed and this process is repeated until the set F_c is empty.

Accordingly, a new vector $\vec{X}_{total}$ (with dimension w_1) is formed, which is comprised by those elements of the initial vector $\vec{X}_i$ that passed successfully through the above mentioned correlation analysis. Any linear combination of any of the w_1 elements that comprise vector $\vec{X}_{total}$ (2^{w_1} such combinations) is a candidate estimation model. The best estimation is determined by a thorough examination of the 2^{w_1} combinations.

Synoptically, the main steps of the developed regression model are the following:

1. The N input variables are selected.
2. Based on the selected variables and function set F, the basic vector $\vec{X}_i$ is determined.
3. Based on correlation analysis, a reduced subset of terms of the basic vector $\vec{X}_i$ is selected to form the vector $\vec{X}_{total}$.
4. Any combination of the terms in $\vec{X}_{total}$ is a vector $\vec{X}_m$ (m=1,…, 2^{w_1}) candidate to be examined as a basis for the model.
5. For each vector $\vec{X}_m$ the constant coefficients and the validation criterion R^2 are calculated.
6. The vector $\vec{X}_m$ with the overall maximum value of R^2 is selected.

2.1.1 - APPLICATION OF THE MODEL

The model will be used for the determination of the relationship between E_c, ρ_c and k_g. The measurements of the soil resistivity (x_1), soil dielectric constant (x_2) and E_c (y) of three different soil samples under four moisture contents are the data, which are used for the application of the model.

The vector $\vec{X}_i$ is defined (by equation 3) by applying the functions x^α, x^β, x^γ, $1/x$, $\ln(x)$ and e^{-x} on every input variable. Consequently the dimension of the input vector is 13 (including constant term). It should be noted that the values of α, β, γ belong to the sets A={1,1.1,…4}, B={0.2,0.201,…,0.990} and Γ ={0.001,0.002,…,0.190} respectively. For the variable x_1, the value $\beta = 0.2$ was chosen, since it leads to the highest absolute value of the correlation index between the input y and the function x^β. Therefore, the corresponding function that will be applied to the model is $x_1^{0.2}$. The following two steps of the correlation analysis were carried out:

- The correlation indices between $f_k(x_j)$ and y are estimated for each variable of the vector $\vec{X}_i$. Among these, the ones whose correlation index is higher than cor_1 (=0.2 in this case) were retained for further processing. For variable x_1, the correlation indices between the terms $f_k(x_1)$ and y are presented in Table 1 and since the absolute values of all indices are greater than 0.2, all the terms were kept for further processing.

$f_k(x_1)$	Correlation index $f_k(x_1) - y$
$x_1^{1.1}$	0.7596
$x_1^{0.2}$	0.9504
$x_1^{0.189}$	0.9504
$\ln x_1$	0.9354
$1/x_1$	-0.6634
e^{-x_1}	-0.8574

Table 1 - Correlation index $f_k(x_1) - y$

- The correlation indices between the retained terms $f_k(x_j)$-$f_{k'}(x_{j'})$ are computed. If between any two terms the correlation index is smaller than a pre-specified value cor_2 (=0.8), both terms are retained; otherwise, only the term with the largest correlation with respect to output y is retained. This procedure can be clarified through the illustrative example presented in Table 2. By examining the correlation index between $x_1^{0.2}$ and $x_1^{0.189}$, which is 1.0>0.8, it can be concluded that only $x_1^{0.2}$ is retained, since it presents the larger correlation index to y (Table 1).

$f_k(x_1)$	$x_1^{1.1}$	$x_1^{0.2}$	$x_1^{0.189}$	$\ln x_1$	$1/x_1$	e^{-x_1}
$f_{k'}(x_{j'})$	$x_1^{0.2}$	e^{-x_2}	$x_1^{0.2}$	$x_1^{0.2}$	$\ln x_1$	$x_1^{0.2}$
Correlation index $f_k(x_1)$-$f_{k'}(x_{j'})$	0.8199	0.7884	1.0000	0.9847	0.8037	0.9052
Which wins?	$x_1^{0.2}$	both	$x_1^{0.2}$	$x_1^{0.2}$	$\ln x_1$	$x_1^{0.2}$

Table 2 - Example of correlation index $f_k(x_j)$- $f_{k'}(x_{j'})$

Through the above procedure the terms that remain are the following:

$x_1^{0.2}$ and e^{-x_2},

which form the vector $\vec{X}_{total}$. Any combination of the above terms is a candidate model $\vec{X}_m$. All 2^2=4 combinations were examined. Finally, the following model was chosen:

$$E_c = 1.903+6.487 \cdot \rho_g^{0.2}\text{-}29.463 \cdot e^{-k_g} \qquad (4)$$

where E_c in kV/cm, ρ_g in kΩm and k_g dimensionless quantity.

The results of the developed model were compared to the results of the following models proposed by Manna [13]:

i) A power regression model expressing E_c as a function of soil's resistivity:

$$E_c = 8.432 \cdot \rho_g^{\,0.154} \qquad (5)$$

ii) A power regression model of soil's dielectric constant:

$$E_c = 39.298 \cdot k_g^{\,-0.751} \qquad (6)$$

iii) A multiple regression model, where E_c is expressed as a combination of ρ_g and k_g [11]:

$$E_c = 8.6083 \cdot k_g^{\,-0.0103} \cdot \rho_g^{\,0.1526} \qquad (7)$$

where E_c in kV/cm, ρ_g in kΩm and k_g dimensionless quantity.

As a criterion for the comparison, the value of MAPE (mean absolute percentage error) of every model was taken into consideration. MAPE is defined as:

$$MAPE = \frac{1}{N} \cdot \sum_{i=1}^{N} \frac{\left|E_{ci}^{c} - E_{ci}^{m}\right|}{E_{ci}^{m}} \qquad (8)$$

where N is the total number of data (108 in our case), E_{ci}^{c} is the value estimated by the model and E_{ci}^{m} is the experimental value.

In Table 3 the values of MAPE for each model are presented in order to be compared with the proposed model.

Model	Equation	MAPE (%)
i	$E_c = 8.432 \cdot \rho_g^{\,0.154}$	7.893
ii	$E_c = 39.298 \cdot k_g^{\,-0.751}$	17.354
iii	$E_c = 8.6083 \cdot k_g^{\,-0.0103} \cdot \rho_g^{\,0.1526}$	7.890
Proposed model	$E_c = 1.903+6.487 \cdot \rho_g^{0.2}\text{-}29.463 \cdot e^{-k_g}$	7.760

Table 3 - Comparison results of the models proposed in literature and the model proposed in this paper

2.1.2 - RESULTS

The results in Table 3 indicate that model ii presents the worst MAPE in comparison to the other models. Furthermore, the equation of the model ii approximates the experimental data better than model i but worse than the model proposed in this paper. Thus, the equation proposed in this paper presents the lowest value of MAPE and could be used as an approximation function for the estimation of E_c. However, for the determination of an equation estimating E_c, experimental data for various soil types are required. Moreover, the influence of parameters beyond soil resistivity and dielectric constant should be taken into consideration by the regression model. Such parameters could be the soil porosity and grain size. In conclusion it can be said, that soil critical electric field is a very important parameter for soil ionization phenomena and therefore further investigation is desirable.

2.2 – THE GENETIC ALGORITHM

Genetic algorithms (GA) are adaptive algorithms widely applied in science and engineering for solving practical search and optimization problems. Many problems can be efficiently tackled by using a GA approach because correlation between the variables is not a problem. The basic GA does not require extensive knowledge of the search space, such as solution bounds or functional derivatives.

In this paper a GA is developed using the software package Matlab and is applied for the optimization of the parameters of equation (1) and equation (4) in order to achieve a better fitting of the curve in the experimental data. The same GA produces excellent results in several optimization problems [14]-[18].

The applied GA starts with a generated population of P_s=40 random values for each parameter. Each parameter's value is converted to a 20-bit binary number. The next step is to form pairs of these points that will be considered as parents for reproduction. By crossover each pair of parents produces N_c=4 children. After crossover there is a P_m=2% probability of mutation. Through reproduction, the population of the "parents" is enhanced with the "children". By applying the process of natural selection only 40 members survive. These are the members with the lower values of the MAPE between the measured and optimized data, since a minimization problem is solved. By repeating the iterations of reproduction under crossover, mutation and natural selection, GAs can achieve minimum error. The best values of the population converge at this point. The termination criterion is fulfilled when the mean value of the optimization function in the P_s-members' population is no longer improved or the number of iterations is greater than the maximum defined number N_{max}.

2.2.1 – APPLICATION OF THE GENETIC ALGORITHM

The GA was applied to the experimental data, which were acquired by the figures presented in [11]. The equation for the estimation of E_c as proposed in [11] and [13] is the following

$$E_c = a \cdot k_g^{\,b} \cdot \rho_g^{\,c} \qquad (9)$$

while equation (4) can be written as:

$$E_c = A + B \cdot \rho_g^C - D \cdot e^{-k_g} \qquad (10)$$

where E_c is the critical soil electric field in kV/cm, k_g is the soil dielectric constant (dimensionless quantity), ρ_g is the soil resistivity (in kΩm), and *a*, *b*, *c*, *A*, *B*, *C* and *D* are constants to be determined.

2.2.2 - RESULTS

Given the above equation and the experimental data, the GA was implemented in order to optimize the parameters *a*, *b*, *c*, *A*, *B*, *C* and *D*. The input data of the GA, which calculates the values of the parameters, are the soil critical electrical field and the soil electrical parameters. In Tables 4 and 5 the initial and the optimized values of the parameters of equations (9) and (10) respectively are presented for the purpose of collation.

	a	*b*	*c*	MAPE (%)
Parameters of Equation (1)	8.6083	-0.0103	0.1526	7.890
Optimized parameters of Equation (9)	9.3309	-0.0533	0.2665	7.589

Table 4 - The parameters' values of the equation proposed by Manna et al. and the optimized parameters' values for (9) by using the GA.

	A	*B*	*C*	*D*	MAPE (%)
Parameters of Equation (4)	1.903	6.487	0.200	29.463	7.760
Optimized parameters of Equation (10)	1.661	8.123	0.156	22.493	7.477

Table 5 - The parameters' values of the equation derived by the regression model and the optimized parameters' values for (10) by using the GA.

Thus, the equation (1) for the optimized parameters is:

$$E_c = 9.3309 \cdot k_g^{-0.0533} \cdot \rho_g^{0.2665} \qquad (11)$$

and the optimized equation (4) is:

$$E_c = 1.661 + 8.123 \cdot \rho_g^{0.156} - 22.493 \cdot e^{-k_g} \qquad (12)$$

where E_c in kV/cm, ρ_g in kΩm and k_g dimensionless quantity.

Figures 2 - 5 illustrate the experimental data of the E_c along with equations (1), (4), (11) and (12) respectively. Comparing the values of MAPE of the equations (1) and (4) to the values of MAPE of the equations (11) and (12) it can be concluded that in each case the application of the GA has lessened MAPE's value. Furthermore, the values of MAPE of the equations (4) and (12) are lower than the MAPE values of equation (1) and (11) respectively.

From Figures 2 - 3 it is obvious that equation (11) fails to approximate the experimental data for high values of resistivity but offers a better approximation (than equation (1)) for lower values of resistivity. Nevertheless, equation (12) provides us with a MAPE lower than equation (1) due to the plethora of experimental data with ρ_g up to 10kΩm.

In conclusion, given the fact that equation (12):

- presents the lowest MAPE,
- approximates effectively the experimental data (Figure 5), when compared to the other equations (Figures 2 - 4),
- does not diverge from the experimental data even for high values of resistivity

it can be concluded that equation (12) is the best equation of all the equations examined in this paper.

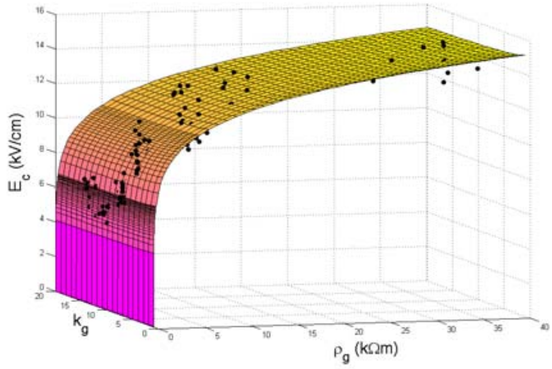

Figure 2 - Experimental data along with (1)

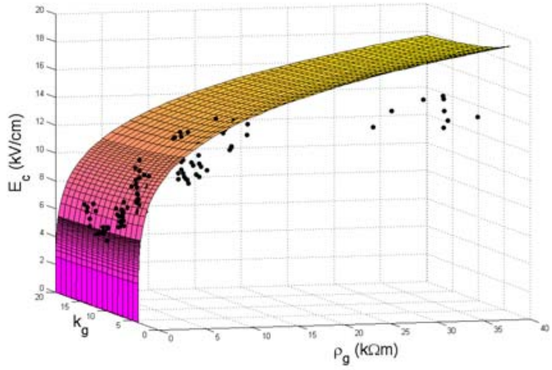

Figure 3 - Experimental data along with (11)

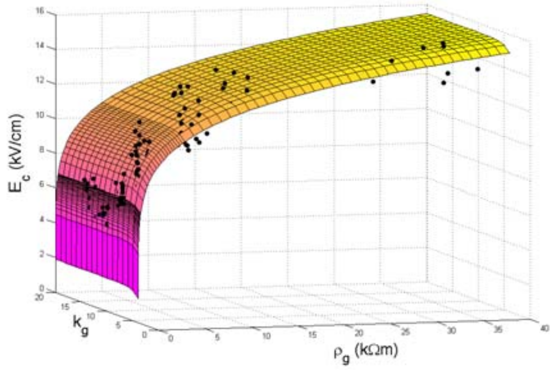

Figure 4 - Experimental data along with (4)

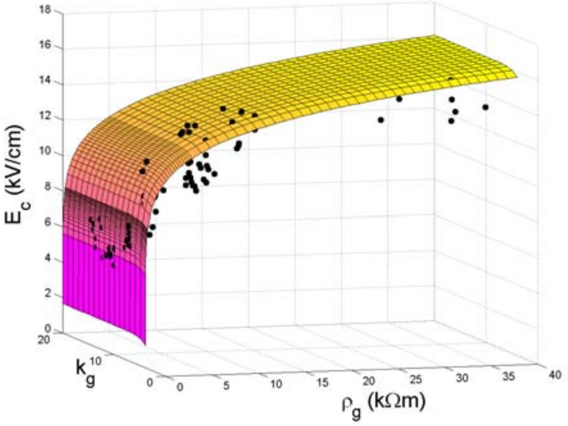

Figure 5 - Experimental data along with (12)

3 - CONCLUSION

By implementing a multi-variable regression model on experimental data from [13] an equation for E_c is derived. The parameters of this equation have been optimized through the application of a GA. Thus, equation (12) reduces the MAPE of equation (1) by 5.2%. These results show that not only the parameters of equation (1) require improvement, but also the form of the equation that implicates E_c with k_g and σ_g .

4 - REFERENCES

[1] Towne, H.M., “Impulse characteristics of driven grounds”, *Gen. Electr.Rev.*, pp. 605–609, November 1929

[2] Bellaschi, P.L., “Impulse and 60-cycle characteristics of driven grounds”, *Transactions. of American Institute of Electrical. Engineers*, vol. 60, pp. 123–128, March 1941

[3] Bellaschi, P.L., Armington, R.E., and Snowden, A.E., “Impulse and60-cycle characteristics of driven grounds II”, *Transactions of American Institute of Electrical. Engineers*, vol. 61, pp. 349–363, 1942

[4] Liew, A.C., and Darveniza, M., “Dynamic model of impulse characteristics of concentrated earths”, *Proc. Inst. Electr*. Eng., vol. 121, (2), pp. 123–135, 1974

[5] Loboda, M., and Pochanke, Z., “Experimental study of electric properties of soil with impulse current injections”, *18th International. Conference. on Lightning Protection Proceedings*, pp. 191–198, Munich, September 1985

[6] Loboda, M., and Scuka, V., “On the transient characteristics of electrical discharges and ionization processes in soil”, *23rd Int. Conference on Lightning Protection Proceedings*, pp. 539–544, Florence, September 1996

[7] Oettle, E.E., “A new general estimation curve for predicting the impulse impedance of concentrated earth electrodes”, *IEEE Transactions Power Delivery*, vol. 3, no. 4, pp. 2020–2029, 1988

[8] CIGRE Working Group on Lightning, ‘Guide to procedures for estimating the lightning performance of transmission lines”, CIGRE-Paris, France, October 1991

[9] Mousa, A.M., ”The soil ionization gradient associated with discharge of high currents into concentrated electrodes”, *IEEE Trans. Power Delivery*, vol. 9, no. 3, pp. 1669–1677, 1994

[10] Gonos I.F. and Stathopulos I.A., “Soil ionisation under lightning impulse voltages”, *IEE Proceedings Science,Measurement and Technology*, vol. 151, no. 5, pp. 343-346, September 2004

[11] Manna T.K. and Chowdhuri P., “Generalized equation of soil critical electric field Ec based on impulse tests and measured soil electrical parameters”, *IET Generation Transmission and Distribution*, vol. 1, no. 5, pp. 811-817, September 2007

[12] Tsekouras G.J., Elias C.N., Kavatza S., Contaxis G.C., “A hybrid non- linear regression midterm energy forecasting method using data mining”, *IEEE Bologna Power Tech. Conference*, Bologna, Italy, June 2003

[13] Manna T.K., “Impulse impedance of grounding systems and its effects on tower crossarm voltage”, Dissertation, Tennessee Technological University, August 2008

[14] Fotis G.P., Gonos I.F., Asimakopoulou F.E., Stathopulos I.A. ,“Applying genetic algorithms for the determination of the parameters of the electrostatic discharge current equation”, *Institute of Physics (IOP), Proceedings Measurement, Science & Technology*, vol. 17, pp. 2819-2827, 2006.

[15] Gonos I.F. and Stathopulos I.A., “Estimation of the multi-layer soil parameters using genetic algorithms”, *IEEE Transactions on Power Delivery*, vol. 20, no. 1, pp. 100-106, January. 2005

[16] Gonos I.F., Mastorakis N.E and Swamy M.N.S., “A genetic algorithm approach to the problem of factorization of general multidimensional polynomials”, *IEEE Transactions on Circuits and Systems*, Part I, vol. 50, no. 1, pp.16-22, January. 2003

[17] Gonos I.F., Topalis F.V. and Stathopulos I.A., “A genetic algorithm approach to the modelling of polluted insulators”, *IEE Proceedings Generation, Transmission and Distribution*, vol. 149, no. 3, pp 373-376, May 2002

[18] Asimakopoulou F.E., Fotis G.P., Gonos I.F. and Stathopulos I.A., “Parameter Determination of Heidler’s Equation for the ESD Current”, 15th *International Symposium on High Voltage Proceedings*, Ljubljana, Slovenia, August 2007

Main author
Name: Mrs. Fani E. Asimakopoulou
Address: High Voltage Laboratory,
School of Electrical and Computer Engineering,
National Technical University of Athens,
9, Iroon Politechniou Str., 15780 Zografou Campus,
Athens, Greece
Fax: +302107723504 ; Phone:+302107722523
E-mail: fasimako@mail.ntua.gr