Computing and Informatics, Vol. 41, 2022, 861–884, doi: 10.31577/cai_2022_3_861

REGRESSION ANALYSIS AND MODELING OF LOCAL ENVIRONMENTAL POLLUTION LEVELS FOR THE ELECTRIC POWER INDUSTRY NEEDS

Peter Krammer, Marcel Kvassay, Radoslav Forgáč
Miloš Očkay, Lenka Skovajsová, Ladislav Hluchý

Institute of Informatics
Slovak Academy of Sciences
Dúbravská cesta 9
845 07 Bratislava, Slovakia
e-mail: {peter.krammer, marcel.kvassay, radoslav.forgac, milos.ockay, lenka.skovajsova, ladislav.hluchy}@savba.sk

Ľuboš Skurčák, Ľuboš Pavlov

VUJE, a. s.
Okružná 5
918 64 Trnava, Slovakia
e-mail: {lubos.skurcak, lubos.pavlov}@vuje.sk

Abstract. Reliability, longevity, and maintenance costs of electric power industry installations and equipment depend strongly on the extent to which their design reflects relevant environmental factors, such as expected levels of local environmental pollution. These factors guide the choice of specific types of components – insulators, towers, conductors, etc. – and are often estimated through complex and tedious long-term field measurements of pollution deposits. In Slovakia, such field measurements were mandated by the national standard STN 33 0405. This standard was retired in 2015 without replacement, which opened the way for developing alternative and less cumbersome methods. One such alternative is to apply artificial intelligence techniques to atmospheric pollution and other relevant data, which is already routinely monitored and collected in many countries. In this paper, we explore the strength of the relationships between the field measurements performed in various regions of Slovakia according to STN 33 0405 and atmospheric pollution data monitored and collected by the Slovak Hydrometeorological Institute (SHMÚ). The paper is focused on input attributes significance, in relation to output attributes. It represents the first phase of our long-term research aiming at the creation of reliable regression models of local pollution in order to replace the cumbersome field measurements mandated by STN 33 0405.

Keywords: Regression analysis, Weibull distribution, machine learning, neural networks

Mathematics Subject Classification 2010: 68T05, 68T07

1 INTRODUCTION

The smooth operation of electric power industry installations and equipment among other things also depends on the thorough knowledge of parameters and conditions of the ambient environment. Environmental factors are therefore important in their design and operation and significantly affect their construction and maintenance costs. The surrounding environment not only affects the choice of specific components (such as insulators, towers, conductors, etc.) but also plays a role in their design and maintenance to achieve high reliability and longevity of overhead power lines. Between the years 2010 and 2019, approximately 160 faults caused by unexpected flashovers were identified on the overhead power lines in the transmission system of the Slovak Republic (TS SR). The cause of the faults could be attributed, among other things, to excessive contamination of suspension and strain insulators, together with unfavourable weather conditions, such as fog, snow, rain, etc. These technological challenges still resonate in international research, as we demonstrate in our outline of the state of the art in the next section.

In Slovakia, the contamination assessment of overhead power line insulators was addressed by the technical standard STN 33 0405. It specified characteristics of evaluated areas according to the degree of pollution, the design requirements for the insulation of outer conduits, and the procedure to determine the degree of pollution for the considered areas. Consequently, depending on the degree of pollution, the standard STN 33 0405 mandated progressively more and more stringent rules for the design and operation of electrical power installations and equipment in terms of the required number of insulators, minimum lengths of their specific surface paths, and their recommended cleaning intervals. The most significant disadvantages of this procedure are:

1. Long interval required for the collection and analysis of pollution fallout, for at least two or more years.
2. Limited possibilities of processing the large number of samples. Samples need to be dried out.

3. Limited opportunity to assess new corridors for overhead power lines. Support structures for dust collectors need to be installed on the new corridors. The possible use of existing structures require the landowner permits. The number of suitable structures is limited, as the overhead power lines mostly pass over the inhabited areas.
4. High probability of discrepancies in the results, due to a low protection against unauthorized manipulation or accidental contamination of samples.

These factors have contributed to the need to replace this procedure with an advanced and efficient one, which would minimize the complexity of implementation while maintaining or even improving the accuracy of calculated attributes. The initiative to research a new innovative process originated at VUJE, a.s., a research institute with the extensive experience in the field measurements according to STN 33 0405. Its main idea is to replace the field measurements with relevant data that are already being continually monitored and archived by the Slovak Hydrometeorological Institute (SHMÚ). This institute has potentially useful data, such as hourly concentrations of pollutants PM_{10}, $PM_{2.5}$, O_3, SO_2, CO, NO_2 (shmu.sk) or the database of main polluters NEIS (National Emission Information System, air.sk).

The pollution level modeling task can be solved by the following machine learning approaches. The first approach is classification of output into the single or multiple discrete pollution levels. The second approach is based on regression. In this case pollution level is represented as a real number, which can be compared with the threshold values of individual classes. The presented paper is focused on the regression analysis of environmental pollution assessment in the field of power industry.

The rest of this paper is structured as follows. Section 2 describes related work on insulator pollution. The baselines for regression analysis are described in Section 3. Section 4 summarizes possible solution approaches focusing on prediction of 99.5 % quantile of Weibull distribution for output attributes. Section 5 evaluates the achieved results. And finally, Section 6 concludes the paper.

2 RELATED WORK

Insulator pollution is a current topic of international research. An ever-increasing number of publications on the topic of “insulator pollution” in the Scopus database over the past 10 years proves this fact. Studies such as [1, 2, 3, 4, 5] have investigated the occurrences of flashovers related to the shape of the insulator, its level of pollution and the chemical composition of the deposits. Pollution of insulators during their operation under high voltage is the subject of a research by Ferreira et al. [6], in which they proposed and validated a method for determining the degree of pollution by spectral analysis of acoustic emissions in the close proximity of polluted insulators. Maraaba et al. [7] analyzed pollution images by a digital camera in the HSV (hue, saturation, value) space. Wang et al. [8] proposed a non-contact method based on spectral analysis of plasma generated from settled contamination after the application of a short laser pulse (LIBS – Laser Induced Breakdown Spectroscopy). Jin et al. [9] used an information fusion of infrared, visible and ultraviolet spectrum. International research is progressing intensively with the modeling of the amount and chemical composition of pollution that settles on a given type of insulator in a given time under the specific weather conditions. Ferma et al. [10] showed a strong correlation between the amount of deposit and the concentration of dust particles in the air for different European locations, as long as they were far from large pollution sources. The same research suggested that seasonal fluctuations in the concentration of air pollutants are much higher in urban areas than in rural ones. He et al. [11] have established that high voltage insulators are polluted both by natural sources and by human activity. Their findings indicate that metal cations originate mainly from human activity and accumulate on the bottom side of the insulators. Chen and Zhang [12] proposed a dynamic model estimating the Non-Soluble Deposit Density (NSDD) based on selected meteorological parameters. Authors Qiao et al. [13] proposed a similar model estimating the Equivalent Salt Deposit Density (ESDD) based on the so-called grey system theory.

Vast majority of deposit and pollution models are only partially theoretically motivated and rely mainly on empirical studies. These are complex scientific problems, involving many variable and nonlinear factors, which lack a complete theoretical explanation. In a situation where the theoretical background of the problem seems unclear, but we have a sufficient dataset, methods of machine learning appear to be the modern and effective solution. The publications in the field of air pollution with dust particles $PM_{2.5}$ and PM_{10} support this claim. The paper by Deters et al. [14] proposes a machine learning approach based on six years of meteorological and pollution data analyses to predict the concentrations of $PM_{2.5}$ from wind (speed and direction) and precipitation levels. The results of their classification model show a high reliability in the classification of low ($< 10\,\mu g/m^3$) versus high ($> 25\,\mu g/m^3$) and low ($< 10\,\mu g/m^3$) versus moderate (10–25 $\mu g/m^3$) concentrations of $PM_{2.5}$. Wu et al. [15] presents a field-portable cost-effective platform for high-throughput quantification of particulate matter using computational lens-free microscopy and machine learning. The study demonstrates that $PM_{2.5}$ concentrations based on meteorological data can be predicted using machine learning statistical models. Implemented device rapidly screens 6.5 L of air in 30 seconds and generates microscopic images of the aerosols in the air. It provides the statistics of particle size and density distribution with a sizing accuracy of 93 %. Shahriar et al. [16] present the machine learning models such as Linear-Support Vector Machine (L-SVM), Medium Gaussian-Support Vector Machine (M-SVM), Gaussian Process Regression (GPR), Artificial Neural Network (ANN), Random Forest Regression (RFR) and use them for the prediction of the concentration of $PM_{2.5}$ and PM_{10} in Bangladesh. Meteorological variables from Dhaka, Chattogram, Rajshahi, and Sylhet for the period of 2013 to 2019, were utilized as exploratory variables. Lary et al. [17] combines machine learning, remote monitoring and meteorological data together with ground-based monitoring of $PM_{2.5}$ from 8 329 measuring sites in 55 countries between 1997 and 2014. The obtained results are related to the influence of particle concentration on some aspects of human mental health. Karimian et al. [18] compare different machine learning approaches, namely Multiple Additive Regression Trees (MART), the Deep Feedforward Neural Network (DFNN) and the Long Short-Term Memory (LSTM) hybrid model. The models were used to capture temporal dependencies in the time series data. From among the classic broad-spectrum publications with an emphasis on industrial applications, we could single out the handbook of Ablameyko et al. [19] and conference proceedings of Fuzzy Logic and Intelligent Technologies in Nuclear Science [20]. The work of Figedy [21] describes two possible approaches to the validation of machine learning models.

3 BASELINES FOR REGRESSION ANALYSIS

The procedure of pollution level prediction starts with the capture of pollution fallout into dust collectors and their subsequent analysis in specialized laboratories. In each considered location, STN 33 0405 mandated 6 to 12 consecutive dust collection measurements, with each measurement lasting for up to two months. Each collected sample was then analyzed in order to determine the total amount of trapped deposit S, the amount of soluble substances S_r and the electrical conductivity of their 0.2 % water solution $\gamma_{0.2}$. In this way, a collection of 6 to 12 values for each of the three attributes (S, S_r, and $\gamma_{0.2}$) was obtained for each evaluated location. Assuming that these values come from a Weibull distribution, the standard then required finding the upper bounds S_m, S_{rm}, and $\gamma_{0.2m}$ for these three attributes as 99.5 % quantiles of their respective Weibull distributions, which meant that the probability of their being exceeded in a given location was only 0.5 %. In the final step, the product of all the three upper bounds was calculated and, based on its value, each location was assigned into one of four ordinal levels or classes of environmental pollution labeled I–IV, with class "I" representing the least polluted areas and class "IV" the most polluted ones.

Analysed data consists of four groups of attributes:

1. Attributes which identify the specific measurement: measurement number (measurement location), GPS coordinates of the measurement location (latitude, longitude) and year of measurement.
2. Attributes which were monitored by SHMÚ: two measures of the annual average concentration of dust particles and three indicators of the annual average concentration of gaseous pollutants in the air for the monitored measurement locations:

   (a) $\mathrm{PM}_{2.5}$ – dust particles with diameter less than 2.5 micrometers;
   (b) PM_{10} – dust particles with diameter less than 10 micrometers;
   (c) SO_2 – sulfur dioxide;
   (d) NO_2 – nitrogen dioxide;
   (e) O_3 – ground-level ozone.

3. Upper bounds of trapped deposit amount are determined according to the retired standard STN 33 0405 as the 99.5 % quantile of Weibull distribution. This distribution represents the best description of the measured six-week deposit values, according to Section 2. It contains the following attributes:

   (a) S – total amount of trapped deposit;
   (b) S_r – soluble substance of trapped deposit;
   (c) $\gamma_{0.2}$ – electrical conductivity of 0.2 % water solution of trapped deposit.

4. The data records that were collected over the years 2008-2013. The dataset represented four measurement campaigns:

   (a) 2008-9 (8 six-week measurements);
   (b) 2010 (5 six-week measurements);
   (c) 2011 (8 six-week measurements);
   (d) 2012-13 (11 six-week measurements).

The first measurement campaign 2008-9 and the last measurement campaign 2012-13 exceeded a calendar year period. Therefore, two partial groups were merged into one two-year group. For input SHMÚ attributes $PM_{2.5}$, PM_{10}, SO_2, NO_2 and O_3, one weighted average value was calculated for the entire two-year period, using two corresponding annual values. The weight of each annual value reflected the number of six-week measurements that ended in a given calendar year.

4 POSSIBLE SOLUTION APPROACHES

The goal of this paper is to determine the significance of input attributes monitored by SHMÚ ($PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3 and their transformations). It is expressed by correlation coefficients with regard to the target attributes. In general, the following approaches come into consideration for a given prediction task on currently available data:

1. The first approach predicts the parameters of the Weibull distribution for the individual attributes S, S_r and $\gamma_{0.2}$. Prediction is followed by calculation of 99.5 % quantile and the quantiles' product. In final stage, pollution level for the measurement location is identified.
2. The second approach directly predicts the 99.5 % quantile of the Weibull distribution for the attributes S, S_r and $\gamma_{0.2}$, calculates their product and uses it for pollution level identification.
3. The third approach directly predicts the product of 99.5 % of quantiles for all three attributes S, S_r, $\gamma_{0.2}$ and uses it for pollution level identification.
4. The last approach directly predicts the pollution level of the measurement location using the scale of I to IV.

The first three approaches exhibit the regression task characteristics. The fourth approach represents a classification task. Preliminary data analysis and prediction experiments identified the second approach as the most promising option. Thus, we focused our research efforts on prediction of the 99.5 % quantile of the Weibull distribution for individual measured attributes S, S_r and $\gamma_{0.2}$.

4.1 Upper Bounds Predictions of S, S_r, $\gamma_{0.2}$

The following input attributes were used to model the target attributes:

1. Original SHMÚ input attributes (PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3);
2. Nonlinear transformations of SHMÚ input attributes, calculated by applying mathematical functions log(x), exp(x), sqrt(x), sqr(x);
3. Products of the SHMÚ attribute pair combinations;
4. Additional attributes (year and gps information (gps_lon, gps_lat) of the measurement location).

All the input attributes were normalized before the modeling process. The following models were used: Multivariate Linear Regression [22], Isotonic Regression [23, 24], Gaussian Processes [25], Multilayer Perceptron Regressor [22], Radial Basis Function Regressor [22], Regression Tree M5P [22, 26] and Random Forest [22, 27]. The minimum Root Mean Squared Error was used as the learning criterion. The 40-fold cross-validation [22] was used in order to obtain a sufficiently objective estimate of the error rate considering the small number of available records. In addition to the Root Mean Squared Error, the accuracy of the model was quantified in the cross-validation process also by the Correlation Coefficient, Mean Absolute Error and Relative Absolute Error. The accuracies of the validated models for the individual target attributes (S, S_r, and $\gamma_{0.2}$) are presented in Tables 1, 2 and 3. Numerical evaluation criteria (e.g. Mean Absolute Error, Pearson Correlation Coefficient) in these tables are calculated from the actual values of the target variable and predicted values of the target variable. SMO Reg is a Sequential Minimal Optimization algorithm for Support Vector Machine (SVM) regression, HN is number of hidden neurons, HL is number of hidden layers, sigmoid is sigmoid activation function, CGD (Conjugate Gradient Descent) and Epochs – maximum allowed epochs during the learning. The experiments were performed in the WEKA environment [28, 29], therefore the names of the used machine learning methods in the tables correspond to WEKA names (e.g. MLP Regressor vs. MultiLayerPerceptron).

Tables 1, 2 and 3 show that trained models do not achieve high modeling quality for any target attribute. This is especially clear for the Relative Absolute Error, where the errors reached 100 %. This is also evident for the correlation coefficient, which is in the limited range between -1 and 1. The models, in this form, are therefore not suitable for real-world applications.

Model	Pearson Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error
Multivariate Linear Regression	0.0588	0.0232	0.0296	100.54 %
Gaussian Processes	−0.0418	0.0244	0.0348	105.85 %
SMO Reg	−0.0557	0.0281	0.0806	121.94 %
Isotonic Regression	0.2808	0.0211	0.0276	91.41 %
MLP Regressor (HL = 1, HN = 2, sigmoid)	0.1126	0.0265	0.0373	114.70 %
MLP Regressor (HL = 1, HN = 4, sigmoid, CGD)	0.0763	0.0284	0.0380	122.85 %
MLP Regressor (HL = 1, HN=8, sigmoid, CGD)	0.1485	0.0445	0.0595	192.98 %
MLP Regressor (HL = 1, HN = 16, sigmoid, CGD)	0.1130	0.0482	0.0627	208.79 %
MultiLayerPerceptron (Epochs = 500, HL = 1)	0.1489	0.0299	0.0433	129.60 %
MultiLayerPerceptron (Epochs = 4 000, HL = 1)	0.1893	0.0384	0.0474	166.47 %
MultiLayerPerceptron (Epochs = 4 000, HL = 2)	0.0399	0.0408	0.0520	176.63 %
RBF Regressor (HN = 2, no CGD)	0.1838	0.0225	0.0297	97.47 %
RBF Regressor (HN = 4, CGD)	0.2844	0.0223	0.0288	96.55 %
RBF Regressor (HN = 8, CGD)	0.2053	0.0238	0.0318	102.99 %
RBF Regressor (HN = 16, CGD)	0.2288	0.0246	0.0329	106.77 %
Regression Tree M5P	0.0651	0.0237	0.0294	102.52 %
Random Forest	0.2847	0.0218	0.0281	94.25 %

Table 1. Validation results of trained regression models for the target variable S

Low accuracy of models trained so far, raises the following questions. How the individual input attributes (original and transformed) contribute to the accuracy of regression models and how significant are they?

For this purpose, a correlation analysis between individual input attributes and the target attributes was performed, using both Pearson's and Spearman's correlation coefficients. The Pearson's correlation coefficient indicates the strength of the linear association between input and target attributes. In contrast, Spearman's correlation coefficient captures the association more comprehensively. It can also recognize and take into account the nonlinear effect of input, as it tracks the degree of monotonicity and not only the degree of linearity between attributes.

Model	Pearson Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error
Multivariate Linear Regression	0.1470	0.0049	0.0070	104.66 %
Gaussian Processes	−0.0346	0.0068	0.0294	144.47 %
SMO Reg	0.1916	0.0043	0.0068	91.45 %
Isotonic Regression	0.0813	0.0048	0.0068	101.75 %
MLP Regressor (HN = 2, sigmoid, no CGD)	0.0871	0.0068	0.0128	146.07 %
MLP Regressor (HN = 4, sigmoid, CGD)	0.1002	0.0074	0.0122	159.11 %
MLP Regressor (HN = 8, sigmoid, CGD)	0.1102	0.0109	0.0174	232.44 %
MultiLayerPerceptron (Epochs = 500, HL = 1)	0.2372	0.0059	0.0083	126.79 %
MultiLayerPerceptron (Epochs = 4 000, HL = 1)	0.1509	0.0110	0.0177	235.69 %
MultiLayerPerceptron (Epochs = 4 000, HL = 2)	0.1126	0.0103	0.0146	220.24 %
RBF Regressor (HN = 2, no CGD)	0.1375	0.0046	0.0073	99.21 %
RBF Regressor (HN = 4, CGD)	0.1666	0.0049	0.0077	104.40 %
RBF Regressor (HN = 8, CGD)	0.2482	0.0046	0.0071	97.95 %
Regression Tree M5P	0.2170	0.0045	0.0067	97.18 %
Random Forest	0.3026	0.0045	0.0066	96.93 %

Table 2. Results of validation of trained regression models for the target variable S_r

4.2 The Results of Correlation Analysis

In the process of significance determination, we are primarily interested in statistical measures of the input attribute impact on target attribute. At this point, it is not decisive whether an increase in the value of the input attribute causes an increase or decrease in the value of the target attribute. The strength of association represents an essential property. For this reason, we applied the absolute value function to the individual values of the correlation coefficients. Resulting non-negative value captures the degree of significance/impact of the given input attribute on the target attribute. The maximum value of 1 would thus represent the maximum significance – an ideal case where the target attribute can be precisely modeled on the basis of a single input attribute.

The respective p-value was also calculated for the individual values of the Spearman's correlation coefficient. It describes the probability that there is no functional

Model	Pearson Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error
Multivariate Linear Regression	0.1095	0.1164	0.1540	101.75 %
Gaussian Processes	−0.0512	0.1681	0.6658	146.98 %
SMO Reg	−0.0240	0.1357	0.3291	118.69 %
Isotonic Regression	0.0601	0.1175	0.1532	102.75 %
MLP Regressor (HN = 2, sigmoid, no CGD)	−0.1587	0.1503	0.2055	131.45 %
MLP Regressor (HN = 4, sigmoid, CGD)	−0.0970	0.2251	0.3232	196.87 %
MLP Regressor (HN = 8, sigmoid, CGD)	0.0520	0.2837	0.3761	248.12 %
MultiLayerPerceptron (Epochs = 500, HL = 1)	0.0281	0.1677	0.2242	146.65 %
MultiLayerPerceptron (Epochs = 4 000, HL = 1)	0.0331	0.2462	0.3645	215.33 %
MultiLayerPerceptron (Epochs = 4 000, HL = 2)	−0.0291	0.2539	0.3228	222.07 %
RBF Regressor (HN = 2, no CGD)	0.1227	0.1197	0.1557	104.71 %
RBF Regressor (HN = 4, CGD)	−0.0644	0.1294	0.1703	113.15 %
RBF Regressor (HN = 8, CGD)	−0.0407	0.1437	0.1895	125.66 %
Regression Tree M5P	0.0057	0.1204	0.1579	105.33 %
Random Forest	−0.0263	0.1253	0.1625	109.56 %

Table 3. Results of validation of trained regression models for the target variable $\gamma_{0.2}$

dependency between the target and the input attributes. Desirable p-value value should be as close as possible to 0. The established convention considers p-values less than 0.05 to be a sufficient indicator of the statistical (and therefore unspecified functional) dependence of two attributes.

For the target values (annual upper bounds), the absolute values of the correlation coefficients listed in Tables 4 and 5 were determined. The p-value for each Spearman's correlation coefficient takes up corresponding space in Table 6.

S column, in the individual tables, represents the target attribute of the total amount of trapped deposit. S_r column represents the total amount of soluble substance and $\gamma_{0.2}$ column represents electrical conductivity. In all three cases, upper bounds are determined as the 99.5 % quantile of the corresponding Weibull distribution. The product column represents the product of these three upper bounds. The individual rows marked "PM_{10}", "$PM_{2.5}$", "NO_2", "SO_2", "O_3" represent the values of input attributes with the same name. The rows containing only a pair of original attributes represent the output attribute with the product of those two attribute values. E.g. "PM_{10} NO_2" is an attribute defined by the product of $PM_{10} * NO_2$. The rows starting with one of the following expressions EXP, LOG, SQR, or SQRT and followed by the attribute are defined by the appropriate mathematical function (ex, $\log_{10}(x)$, x^2, $x^{1/2}$) applied to the given attribute. The year attribute specifies the year in which the six-week campaign was measured. If a campaign of six-week measurements lasted for two consecutive years, the average value is used. E.g., if the campaign started in 2012 and ended in 2013, then 2012.5 is used as the value of the year. The attributes gps_lat and gp_lon represent the GPS coordinates of the measurement locations.

	S	S_r	$\gamma_{0.2}$	product
year	0.008036	0.015860	0.13030	0.031650
gps lat	0.128200	0.168200	0.03688	**0.170700**
gps lon	0.003652	0.005281	0.06603	0.075570
PM_{10}	0.036560	0.057840	0.10210	0.059730
$PM_{2.5}$	0.034380	0.008181	0.08608	0.004444
NO_2	0.091880	0.056850	0.02200	0.114100
SO_2	0.085310	0.109500	0.08177	0.085260
O_3	0.074330	0.055800	0.16210	0.061400
PM_{10} $PM_{2.5}$	0.009664	0.009213	0.10400	0.006945
PM_{10} SO_2	0.053630	0.080790	0.10760	0.062570
$PM_{2.5}$ SO_2	0.076900	0.098100	0.10540	0.080630
SO_2 O_3	0.094700	0.116600	0.04769	0.087750
LOG PM_{10}	0.047790	0.082800	0.09647	0.084770
EXP PM_{10}	0.020400	0.009963	0.11040	0.009413
SQRT PM_{10}	0.042220	0.071060	0.09945	0.073000
LOG $PM_{2.5}$	0.012310	0.039180	0.07674	0.036410
EXP $PM_{2.5}$	0.044260	0.019300	0.12470	0.026200
SQR $PM_{2.5}$	0.051100	0.020450	0.09287	0.025340
SQRT $PM_{2.5}$	0.023850	**0.236300**	0.08171	0.020420
LOG NO_2	0.073000	0.067350	0.02329	0.113900
EXP NO_2	**0.155200**	0.011100	0.02871	0.077290
SQR NO_2	0.108300	0.043480	0.02079	0.110700
SQRT NO_2	0.082690	0.062570	0.02271	0.114500
SQR SO_2	0.091100	0.099130	0.07653	0.893400
LOG O_3	0.069440	0.051760	**0.16760**	0.053940
EXP O_3	0.068500	0.046750	0.10910	0.073140
SQR O_3	0.077580	0.058120	0.15520	0.067390
SQRT O_3	0.072080	0.053990	0.16510	0.057850

Table 4. Significance of individual input attributes (expressed by the absolute value of the Pearson's coefficient) for individual target attributes. The best achieved values are highlighted.

	Correlation Coefficient			
	S	S_r	$\gamma_{0.2}$	product
year	0.047310	0.05969	0.14780	0.018600
gps lat	**0.117800**	0.12090	0.01257	**0.131100**
gps lon	0.040560	0.04278	0.02531	0.002185
PM_{10}	0.064870	0.06982	0.06097	0.059960
$PM_{2.5}$	0.003819	0.04242	0.03529	0.017820
NO_2	0.064890	0.02708	0.02359	0.036510
SO_2	0.096080	**0.14690**	0.03915	0.121600
O_3	0.062720	0.02946	**0.15090**	0.024250
$PM_{2.5}$ SO_2	0.082970	0.10910	0.04483	0.098790

Table 5. Significance of individual input attributes (expressed by the absolute value of the Spearman's coefficient) for individual target attributes. The best achieved value is highlighted.

	p-value			
	S	S_r	$\gamma_{0.2}$	product
year	0.54260	0.44210	**0.05589**	0.81080
gps lat	0.12830	0.11860	0.87150	0.09024
gps lon	0.60170	0.58190	0.74470	0.97760
PM_{10}	0.40340	0.36190	0.43240	0.44010
$PM_{2.5}$	0.96080	0.58510	0.64980	0.81860
NO_2	0.40330	0.72750	0.76150	0.63840
SO_2	0.21540	0.05745	0.61440	0.11650
O_3	0.41930	0.70460	**0.05095**	0.75500
$PM_{2.5}$ SO_2	0.28500	0.15920	0.56390	0.20270

Table 6. Matrix of p-values of Spearman's correlation coefficient for individual target attributes. Conventionally, those values of the Spearman's correlation coefficient for which the p-value is less than 0.05 are considered statistically significant. They reliably confirm the existence of an unspecified functional dependence between the respective input and target attribute. In our case, unfortunately, only two attributes came close to this significance level (the ones highlighted in the table), but did not actually cross it.

Correlation analysis showed that the input variables ($PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3 and their transformations) have unexpectedly low impact on target attributes. Low correlation between input and target attributes is very surprising, because of presumed existence of physical dependence. Very low impact of PM_{10} and $PM_{2.5}$ on the target attribute S is a good example. Low value of the Pearson's correlation coefficient could be explained by dependence nonlinearity between S and PM_{10}, $PM_{2.5}$. However, the Spearman's correlation coefficient, which takes into account the nonlinear dependence of the attributes, has also a low value. It confirms the overall low dependence of the mentioned attributes. The achieved values of this coefficient (PM_{10}, $PM_{2.5}$) are 0.06487 and 0.003819 in the case of attribute S. For comparison, correlation coefficient values above 0.7 have strong significance in the terms of attributes' dependence.

In contrast, the impact of the attribute year turned out to be higher than we expected. In particular, the Spearman's correlation coefficient with regard to the target attribute $\gamma_{0.2}$ indicates the higher impact of the year attribute. The correlation coefficient for attributes $\gamma_{0.2}$ and year reached 0.1478, but it is still one of the highest values. Particularly surprising is the fact that the significance of year attribute outperformed most of other attributes ($PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3) whose physical dependence appears to be obvious. Thus, there is an indication that the dependence of the target attributes on the input attributes, associated with the measurement of pollutants and dust particles (original and transformed), may not be completely stable. It may also depend on time or geographical location (region), because in different years the measurements also differ in the same location.

In comparison with SHMÚ attributes ($PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3), the attribute location has also comparable significance. Specifically, the attribute gps_lat, expressed by the Spearman's correlation coefficient, is one of the attributes with higher significance. For target attributes S and S_r, in comparison with input attributes PM_{10} and $PM_{2.5}$, the attribute gps_lat always reached a higher value of the Spearman's correlation coefficient. The correlation analysis suggests that the attributes year and gps location may have a certain impact on the modeling of target attributes. We therefore do not recommend neglecting or omitting them from datasets in the future. On the contrary, with availability of the larger dataset, the spatio-temporal analysis itself (distribution of pollution in space and time) can be very interesting and beneficial. It is also possible that the distribution of pollution depends on the attribute gps_lat (latitude) only on a small area of Slovakia. The specific position of measurement locations and polluters may contribute to this.

5 EVALUATION OF ACHIEVED RESULTS

The modeling of functional dependencies in data has to take into account several aspects. These aspects directly affect the choice of model type, structure, parameters etc. The following is a list of main aspects:

1. Dependence complexity between inputs and outputs in a real system;
2. The amount of noise present in the data;
3. Number of available data records;
4. Number of available input attributes;
5. Purposes and use of the model.

In our case, the data dependence complexity is not exactly known. However, we estimate that data will exhibit a medium or higher dependence complexity. Similarly, a medium or higher noise level is expected in the data, because the SHMÚ input attributes are interpolated and recalculated in various ways based on location and time.

5.1 The Problem of Appropriate Validation

Currently, only 168 data records are available, which contain 8 input attributes (5 – SHMÚ, 2 – GPS, 1 – year) and 3 target attributes. Due to such a low number of records, it is not possible to sufficiently divide the data into the training set and validation set and use the Hold-Out validation. The result of such validation would be inaccurate and extremely dependent on records assignment into the training set and validation set. The second option is the Bootstrap validation, which involves iterative resampling of the dataset. However, this validation significantly disadvantages some types of models, such as tree models and rules-based models, making it impossible to objectively compare the accuracy of models. For this reason, it is not one of the standard validation methods in the field of machine learning. It is rather a time consuming experimental method. Therefore, the best option is to use N-fold cross-validation, which is also suitable for smaller datasets. With a suitably determined parameter "N", it gives sufficiently stable accuracy estimates of the trained model. The disadvantage of this validation is the time aspect, especially for models where the training itself is a time consuming process. In our case, due to the small number of data records, we used up to 40-fold cross-validation.

5.2 The Problem of Class Imbalances

The available data records (samples) are very unevenly distributed among the different pollution levels (I-IV). Compared to the classification, the target attribute (pollution rate) as a continuous quantity looks like a more natural way of representation. Target attribute is calculated as a real number (as the product of the upper bounds for the three target attributes S, S_r and $\gamma_{0.2}$) and then it is transformed into a discrete class. Therefore, it is more appropriate to use these real numbers (upper bounds of the individual target attributes S, S_r, $\gamma_{0.2}$, as well as their product) as the target attribute of machine learning. This approach does not lead to the loss of information using discrete boundaries of individual classes, as the classification does.

5.3 The Problem of Overfitting

The most significant problem among the above mentioned problems is the effect of overfitting. The overfitting modifies the model to accurately capture the unique characteristics of specific records in the training set instead of following the general trends and characteristics of the dataset. This in turn results in a significant decrease in validation accuracy of unknown data. Overfitted model does not primarily follow the general trend of the data, but rather the specific records in the training set.

Overfitting effect arises in the following situations:

1. When the training set does not contain a sufficient number of records;

2. When the training set is not sufficiently representative (lack of diversity in records or insufficient space coverage of the input attributes);
3. When the trained model has too many parameters (degrees of freedom) in relation to the available records in the training set;
4. When the number of training attributes and the number of records are approximately equal.

For several trained models listed in Section 4.1, it is clear that an overfitting effect is present. This is due to two indicators: simpler models in some cases achieve higher accuracy, or alternatively, some models achieve worse accuracy with increased number of parameters (degrees of freedom).

Larger dataset or a simpler model type with fewer parameters are common ways to avoid the overfitting effect. N-fold cross-validation is also a valuable tool, which we therefore use in this study as well. Due to the fact that the number of data records available to us is currently limited (direct in-field measurements of fallout are logistically demanding and time consuming), the increase of data records is currently not possible. Therefore, the primary way to suppress the overfitting effect is the use of simpler models with fewer parameters. For this reason, it is impossible to use a deep learning approach, resp. more complex neural networks. Therefore, simpler neural networks like Multi Layer Perceptron (MLP) and Radial Basic Functions (RBF) based neural network were used.

It was observed that with an increasing number of hidden layers (resp. number of neurons) in MLP the accuracy of the model decreases. In some cases validated neural network achieved lower accuracy with the high number of training epochs. For one neural network this may be a fluke. But for 40-fold cross-validation, which consists of independent testing of 40 different sub-neural networks, this is extremely unlikely. Both cases of declining accuracy are typical for the overfitting effect.

In the case of the RBF model, the overfitting effect was not so significant, maybe because this model only contained one hidden layer. Nevertheless, there were cases where neuron increase resulted in the accuracy decreased. However, neural networks are the more complex types of models. In the majority of cases they fall behind the simpler models with fewer parameters.

In several cases, simpler models, such as multivariate linear regression or isotonic regression with fewer parameters performed better.

In some cases, interesting results were obtained using the Random Forest model (listed in Section 4.1). It is a complex model consisting of several tree models, combined using the principle of ensemble learning. This model largely withstands the effect of overfitting in terms of increasing the number of trees in the whole model. In other words, the overfitting effect is not enlarged by the addition of new trees to Random Forest. In this particular case, the risk of overfitting can be reduced by providing each tree with only a relatively small subset of input attributes and training them independently.

Overall, it should be noted that the achieved accuracy of the trained models was relatively low. Several factors may contribute to low accuracy: Insufficient representativeness of the dataset, the need for additional relevant input attributes, or a significant amount of noise in the attributes are discussed in more detail in Section 5.4.

5.4 The Problem of Achieving Higher Accuracy of Prediction Models

The analysis of the achieved results reveals the problems related to input data. The following problems appear to be the main factor preventing higher accuracy and quality of prediction models:

1. Small amount of data, few rows (records or measurement locations) in the dataset.
2. Low temporal resolution of data. For input attributes PM_{10}, $PM_{2.5}$, NO_2, SO_2 and O_3 only annual average values were provided. To determine the upper bounds of output attributes S, S_r and $\gamma_{0.2}$ it is necessary to know the variability of relevant input attributes over time, not just their long term average.
3. We do not know the total amount of dust particles in the air (TSP). Input attributes PM_{10} and $PM_{2.5}$ describe the amount of dust particles in the air only for particles with an aerodynamic diameter below 10 micrometers. Attribute S probably represents mainly larger particles with an aerodynamic diameter above 10 micrometers. These are relatively rarely measured as part of the total amount of particulate matter in the air (TSP).
4. Chemical composition of dust particles in the air is unknown. Reaction conditions to which they were exposed on the way to the measurement location are also unknown.

The upper bounds prediction of target attributes S, S_r and $\gamma_{0.2}$ based only on the long-term average values is problematic. This statement can be easily demonstrated by a simple example. Two different datasets can have the same average value, but different maximum, which in a sense determines the upper bounds.

Conceptual depiction of the complex atmospheric aerosol in [30] implies that input attributes PM_{10} and $PM_{2,5}$ are not a good indicator of the total amount of pollution in the air (TSP). PM_{10} includes all the particles with aerodynamic diameter below 10 micrometers, and thus also $PM_{2,5}$. $PM_{2,5}$ aerodynamic diameter is below 2.5 micrometers. These particles do not fall into dust collectors but remain suspended in the air as a mist for a long time. However, particles with aerodynamic diameter above 10 micrometers also float in the atmosphere for some time. Their relative proportion in the atmospheric aerosol is variable and does not correlate with PM_{10} or $PM_{2,5}$. This fact is also documented in [31]. There is a reasonable assumption that these larger and heavier particles (referred to as the TSP-PM_{10} fraction) represent the main component of the total amount of trapped deposit S collected in the VUJE dust collectors.

The hypothesis that the total amount of trapped deposit S consists mainly of large particles (TSP-PM_{10} fraction) is based on the fact that PM_{10} particles are

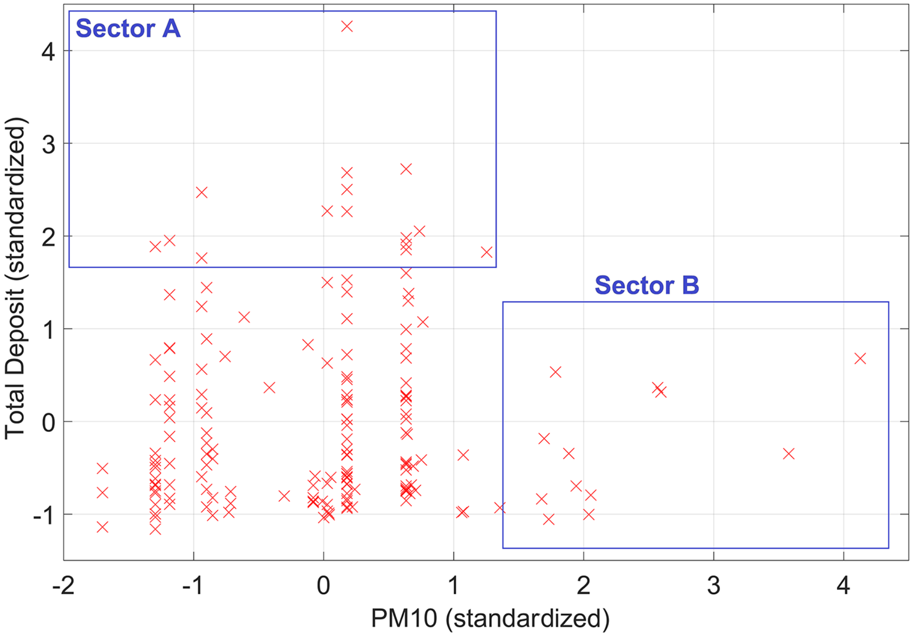


Figure 1. The missing correlation between the concentration of PM_{10} (X-axis) and the total amount of trapped deposit S (Y-axis) is visibly manifested. On the one hand, there are the samples with high S but a low PM_{10} concentration (Sector A). On the other hand, there are samples for which a small S was recorded in the air despite the high concentration of PM_{10} (Sector B).

carried by the wind over long distances. Gravity affects them minimally. Since VUJE dust collectors are constructed as gravity traps, the minimum of PM_{10} will fall in. Part of PM_{10} will undoubtedly get into the VUJE dust collectors, but rather due to wet deposition during rain and night dew. SHMÚ dust collectors use a different principle. They suck in the ambient air and pass it through the filter.

Hints of missing correlation between PM_{10} and S are displayed in Figure 1. For some VUJE samples a large amount of S was recorded at places with low concentration of PM_{10} (Sector A). While other samples recorded a small amount of S despite a large concentration of PM_{10} (Sector B). One of the possible reasons is the already mentioned hypothesis that total amount of trapped deposit S consists mainly of heavier particles with an aerodynamic diameter above 10 micrometers (TSP-PM_{10} fraction). Samples may appear in sector A or B, depending on different location of pollution sources and VUJE dust collectors.

These considerations lead to a qualitative conclusion that the attributes PM_{10} and the total amount of trapped deposit S are probably independent of each other, further research will be needed. In order to effectively predict the total amount of trapped deposit S, it will be necessary to identify other input attributes in addition to the PM_{10} and $PM_{2.5}$. This conclusion was also supported by Section 4.2, where low values of correlation coefficients between PM_{10} and S, both Pearson's (0.03656) and Spearman's (0.06487, p-value = 0.4034), were found.

6 CONCLUSIONS

We originally expected a significant impact of the five primary SHMÚ input attributes and their transformations on the target attributes. The values of all these attributes already take into account the location and a year of measurement, as well as other physical effects on the measured or calculated pollution, which should be more significant than the attributes of the location or year. For a given dataset, they appear to be less significant. From a statistical point of view, we can state that we still lack physical based input variables that would show signs of a more significant statistical impact on the target attributes.

Attempts to model target attributes only from currently available input attributes with relatively low significance are problematic. This has also been confirmed by the achieved accuracy of trained models. There are several possible statistical reasons for the low accuracy results:

1. The input attributes lack high relevance that would significantly influence the target attributes;
2. The chosen methodology is burdened with a degree of uncertainty and it will be necessary to examine the impact of this uncertainty on the overall results;
3. There may be a very complex relationships between the input and target attributes that cannot be reliably approximated from such a limited set of measured data (larger dataset is required);
4. Significant noise in available data (larger dataset is also required).

The main conclusion of this feasibility study is that the investigated machine learning models do not yet have the necessary input information. Thus, their accuracy and quality could not reach the required level. Addition of highly relevant input information, or alternatively, data sample increase could significantly improve the representativeness of the dataset and consequently increase the accuracy of the prediction models. This creates opportunities for the intervention of domain experts in the field of pollution measurement, meteorology (or related areas) on issues such as:

1. Identification of additional input attributes that could have a significant impact on the target attributes and without which more accurate modeling is not possible;
2. Reassessment of the target attributes dependence from the domain point of view (physical/chemical/meteorological);

3. Qualitative expert estimation of the degree of influence of individual input attributes on the target attributes from the domain point of view (physical/chemical/meteorological). Also comparison with the calculated values of correlation coefficients.

Acknowledgement

This work was supported by the Slovak Research and Development Agency under the Contract No. APVV-20-0548 (ARIEN) and by the Slovak Scientific Grant Agency VEGA 2/0125/20.

REFERENCES

[1] Zhang, Z.—Qiao, X.—Zhang, Y.—Tian, L.—Zhang, D.—Jiang, X.: AC Flashover Performance of Different Shed Configurations of Composite Insulators Under Fan-Shaped Non-Uniform Pollution. High Voltage, Vol. 3, 2018, No. 3, pp. 199–206, doi: 10.1049/hve.2018.0002.

[2] Mohammadi Savadkoohi, E.—Mirzaie, M.—Seyyedbarzegar, S.—Mohammadi, M.—Khodsuz, M.—Ghorbani Pashakolae, M.—Biazar Ghadikolaei, M.: Experimental Investigation on Composite Insulators AC Flashover Performance with Fan-Shaped Non-Uniform Pollution Under Electro-Thermal Stress. International Journal of Electrical Power and Energy Systems, Vol. 121, 2020, Art. No. 106142, doi: 10.1016/j.ijepes.2020.106142.

[3] Mahdjoubi, A.—Zegnini, B.—Belkheiri, M.—Seghier, T.: Fixed Least Squares Support Vector Machines for Flashover Modelling of Outdoor Insulators. Electric Power Systems Research, Vol. 173, 2019, pp. 29–37, doi: 10.1016/j.epsr.2019.03.010.

[4] Faramarzi Palangar, M.—Mirzaie, M.—Mahmoudi, A.: Improved Flashover Mathematical Model of Polluted Insulators: A Dynamic Analysis of the Electric Arc Parameters. Electric Power Systems Research, Vol. 179, 2020, Art. No. 106083, doi: 10.1016/j.epsr.2019.106083.

[5] Arias Velásquez, R. M.: Insulation Failure Caused by Special Pollution Around Industrial Environments. Engineering Failure Analysis, Vol. 102, 2019, pp. 123–135, doi: 10.1016/j.engfailanal.2019.04.034.

[6] Ferreira, T. V.—Germano, A. D.—da Costa, E. G.: Ultrasound and Artificial Intelligence Applied to the Pollution Estimation in Insulations. IEEE Transactions on Power Delivery, Vol. 27, 2012, No. 2, pp. 583–589, doi: 10.1109/TPWRD.2011.2178042.

[7] Maraaba, L.—Alhamouz, Z.—Alduwaish, H.: A Neural Network-Based Estimation of the Level of Contamination on High-Voltage Porcelain and Glass Insulators. Electrical Engineering, Vol. 100, 2018, No. 3, pp. 1545–1554, doi: 10.1007/s00202-017-0634-z.

[8] Wang, X.—Lu, S.—Wang, T.—Qin, X.—Wang, X.—Jia, Z.: Analysis of Pollution in High Voltage Insulators via Laser-Induced Breakdown Spectroscopy. Molecules, Vol. 25, 2020, No. 4, Art. No. 822, doi: 10.3390/molecules25040822.

[9] Jin, L.—Ai, J.—Tian, Z.—Zhang, Y.: Detection of Polluted Insulators Using the Information Fusion of Multispectral Images. IEEE Transactions on Dielectrics and Electrical Insulation, Vol. 24, 2017, No. 6, pp. 3530–3538, doi: 10.1109/TDEI.2017.006516.

[10] Ferm, M.—Watt, J.—O'Hanlon, S. M.—de Santis, F.—Varotsos, C. A.: Deposition Measurement of Particulate Matter in Connection with Corrosion Studies. Analytical and Bioanalytical Chemistry, Vol. 384, 2006, pp. 1320–1330, doi: 10.1007/s00216-005-0293-1.

[11] He, Z.—Gao, F.—Tu, Z.—Zhang, Y.—Chen, H.: Analysis of Natural Contamination Components and Sources of Insulators on ± 800 kV DC Lines. Electric Power Systems Research, Vol. 167, 2019, pp. 192–198, doi: 10.1016/j.epsr.2018.10.033.

[12] Chen, S.—Zhang, Z.: Dynamic Pollution Prediction Model of Insulators Based on Atmospheric Environmental Parameters. Energies, Vol. 13, 2020, No. 12, Art. No. 3066, doi: 10.3390/en13123066.

[13] Qiao, X.—Zhang, Z.—Jiang, X.—He, Y.—Li, X.: Application of Grey Theory in Pollution Prediction on Insulator Surface in Power Systems. Engineering Failure Analysis, Vol. 106, 2019, Art. No. 104153, doi: 10.1016/j.engfailanal.2019.104153.

[14] Kleine Deters, J.—Zalakeviciute, R.—Gonzalez, M.—Rybarczyk, Y.: Modeling $PM_{2.5}$ Urban Pollution Using Machine Learning and Selected Meteorological Parameters. Journal of Electrical and Computer Engineering, Vol. 2017, 2017, Art. No. 5106045, doi: 10.1155/2017/5106045.

[15] Wu, Y. C.—Shiledar, A.—Li, Y. C.—Wong, J.—Feng, S.—Chen, X.—Chen, C.—Jin, K.—Janamian, S.—Yang, Z.—Ballard, Z. S.—Göröcs, Z.—Feizi, A.—Ozcan, A.: Air Quality Monitoring Using Mobile Microscopy and Machine Learning. Light: Science and Applications, Vol. 6, 2017, Art. No. e17046, doi: 10.1038/lsa.2017.46.

[16] Shahriar, S. A.—Kayes, I.—Hasan, K.—Salam, M. A.—Chowdhury, S.: Applicability of Machine Learning in Modeling of Atmospheric Particle Pollution in Bangladesh. Air Quality, Atmosphere and Health, Vol. 13, 2020, No. 10, pp. 1247–1256, doi: 10.1007/s11869-020-00878-8.

[17] Lary, D. J.—Lary, T.—Sattler, B.: Using Machine Learning to Estimate Global $PM_{2.5}$ for Environmental Health Studies. Environmental Health Insights, Vol. 9, 2015, No. S1, pp. 41–52, doi: 10.4137/EHI.S15664.

[18] Karimian, H.—Li, Q.—Wu, C.—Qi, Y.—Mo, Y.—Chen, G.—Zhang, X.—Sachdeva, S.: Evaluation of Different Machine Learning Approaches to Forecasting PM2.5 Mass Concentrations. Aerosol and Air Quality Research, Vol. 19, 2019, No. 6, pp. 1400–1410, doi: 10.4209/aaqr.2018.12.0450.

[19] Ablameyko, S.—Goras, L.—Gori, M.—Piuri, V.: Neural Networks for Instrumentation, Measurement and Related Industrial Applications. IOS Press, NATO Science Series, III: Computer and Systems Sciences, Vol. 185, 2003.

[20] Ruan, D.—Abderrahim, H. A.—D'hondt, P.—Kerre, E. E.: Intelligent Tech-

niques and Soft Computing in Nuclear Science and Engineering. Proceedings of the 4th International FLINS Conference, World Scientific, 2000, doi: 10.1142/4466.

[21] Figedy, S.: In-Core Sensors Readings Diagnostics Based on Neuro-Fuzzy Techniques. Uncertainty Modeling in Knowledge Engineering and Decision Making, pp. 969–974, doi: 10.1142/9789814417747_0155.

[22] Hastie, T.—Friedman, J.—Tibshirani, R.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer New York, New York, NY, 2016, 767 pp.

[23] Wu, W. B.—Woodroofe, M.—Mentz, G.: Isotonic Regression: Another Look at the Changepoint Problem. Biometrika, Vol. 88, 2001, No. 3, pp. 793–804, doi: 10.1093/biomet/88.3.793.

[24] Kalai, A. T.—Sastry, R.: The Isotron Algorithm: High-Dimensional Isotonic Regression. Proceedings of the 22nd Annual Conference on Learning Theory (COLT), 2009, `https://www.microsoft.com/en-us/research/publication/isotron-algorithm-high-dimensional-isotonic-regression/`.

[25] Hida, T.—Hitsuda, M.: Gaussian Processes. American Mathematical Society, 1993, doi: 10.1090/mmono/120.

[26] Mohammed, A.—Rafiq, S.—Sihag, P.—Kurda, R.—Mahmood, W.—Ghafor, K.—Sarwar, W.: ANN, M5P-Tree and Nonlinear Regression Approaches with Statistical Evaluations to Predict the Compressive Strength of Cement-Based Mortar Modified with Fly Ash. Journal of Materials Research and Technology, Vol. 9, 2020, No. 6, pp. 12416–12427, doi: 10.1016/j.jmrt.2020.08.083.

[27] Biau, G.—Scornet, E.: A Random Forest Guided Tour. TEST: An Official Journal of the Spanish Society of Statistics and Operations Research, Vol. 25, 2016, No. 2, pp. 197–227, doi: 10.1007/s11749-016-0481-7.

[28] Witten, I. H.—Frank, E.—Hall, M. A.—Pal, C.: Data Mining: Practical Machine Learning Tools and Techniques. 4th Edition. Morgan Kaufmann, 2016, `https://www.amazon.com/Data-Mining-Practical-Techniques-Management/dp/0128042915`.

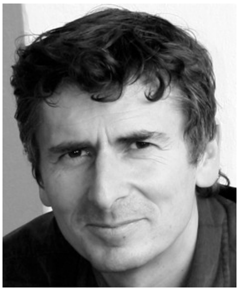
[29] Hall, M.—Frank, E.—Holmes, G.—Pfahringer, B.—Reutemann, P.—Witten, I. H.: The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter, Vol. 11, 2009, No. 1, pp. 10–18, doi: 10.1145/1656274.1656278.

[30] Phalen, R. F.: The Particulate Air Pollution Controversy: A Case Study and Lessons Learned. Springer Science and Business Media, 2002, doi: 10.1007/b101883.

[31] Rembges, D.—Kotzias, D.: Monitoring TSP, PM 10 and PM 2.5 at a Semi-Remote Area in Northern Italy – Relationships Between PM 10 and PM 2.5. Fresenius Environmental Bulletin, Vol. 12, 2003, pp. 402–405.

Peter Krammer is Researcher at the Institute of Informatics, Slovak Academy of Sciences. His main research topics include machine and deep learning, statistical modeling, statistical analysis and optimization. He graduated from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in Bratislava in 2010. He is author of more than 30 scientific papers and has participated in international and national research projects.

Marcel Kvassay graduated from the Faculty of Electrical Engineering of the Slovak University of Technology in Bratislava in 1991. After his professional career as a software engineer, software development couch and software process improvement manager he joined the Institute of Informatics of the Slovak Academy of Sciences in 2009 and earned his Ph.D. in applied informatics in 2017. His research interests include intelligent and knowledge-based technologies, NLP and semantic search, data mining and machine learning, as well as simulations.



Radoslav Forgáč is Researcher at the Institute of Informatics, Slovak Academy of Sciences. He graduated from the Armed Forces Academy of General Milan Rastislav Štefánik in Liptovský Mikuláš in 1993. He received his Ph.D. in artificial intelligence in 2006 from the Technical University of Košice. His research is focusing on machine learning, especially neural networks for classification, regression and clustering.

Miloš Očkay is Researcher at the Institute of Informatics, Slovak Academy of Sciences and Associate Professor at the Department of Informatics at the Armed Forces Academy in Liptovský Mikuláš. He holds his Ph.D. degree in the field of informatics from the Technical University of Košice, received in 2012. His research is focusing on parallel computing, and neural networks.

Lenka Skovajsová received her Ph.D. degree from the Slovak Academy of Sciences in information retrieval and neural networks in 2011. She is Researcher in the field of neural networks and machine learning at the Slovak Academy of Sciences. Her research interests include applied machine learning algorithms with the focus on the text and image processing.

Ladislav Hluchý is Senior Research Scientist and Manager with more than 20 years of experience in leading national and international research projects and teams of 5 to 20 researchers. He is a competent scientist in the area of high-performance computing, multicloud computing, parallel and distributed information processing and knowledge management. His research also focuses on data flow management through abstract language mechanisms. In the past, Ladislav Hluchý, Associate Professor has participated in several cooperations with industry, which will be beneficial for the transfer of the project results into practice.

Ľuboš Skurčák graduated from the Faculty of Electrical Engineering and Informatics of the Slovak University of Technology in Bratislava (FEI STU) in 1999, where he received the degree of engineer in the field of electrical power engineering. Since 1999, he has been working at VUJE, a.s., in the Division for the Support of Management and Operation of the Electricity System. In 2015, he successfully defended his dissertation on the topic "Influence of the configuration of external high-voltage and high-voltage distribution to the magnitude of electric and magnetic fields of the mains frequency from the point of view of personal protection" at FEI STU in Bratislava and received his Ph.D. degree. He has more than 20 years of experience in industrial research and experimental development. He is a long-term member of the international non-profit association CIGRE, which brings together experts in the field of electricity generation and transmission from more than 80 countries. He is the author or co-author of several papers focused on the issue of assessing the impact of energy networks, their reliability and development.

Ľuboš Pavlov is Specialist in modeling and calculations in the field of power line operation. He graduated from the Faculty of Electrical Engineering at the University of Žilina, Slovakia, in 2010, where he received the degree of engineer in the field of electrical power engineering. Since 2010, he has been working at VUJE, a.s., in the Division for the Support of Management and Operation of the Electricity System. In 2016, he successfully defended his dissertation on the topic "Monitoring of synchronous phasors in the power system" at FEI STU in Bratislava and received his Ph.D. degree. His research and development tasks are focused on energy efficiency of the electricity system. He is the author and co-author of several papers focused on current issues in electricity.