Refinement of an Environmental Pollution Model for the Needs of the Electric Power Industry by Addition of Precipitation Attributes

Peter Krammer, Marcel Kvassay, Ondrej Habala, Ján Mojžiš,
Ladislav Hluchý, Ľuboš Pavlov, Ľuboš Skurčák

SACI conference

23.-26. 5. 2023

Pollution modelling in power Industry

- Power Industry is one of the key world industry
- Design and placement of high-voltage poles/pylons require to take account specific aspects of location.

Discharge phenomena at different discharge stages [3]

(a) Safety Zone, (b) Forecast Zone, (c) Danger Zone, (d) Flashover

- Pollution level ~ significant influence

  (can cause flashover)
- Require to model a pollution level
- Environmental reason for pollution modelling

Quantification of Pollution

- Pollution level - classification with IV classes (I. – IV.)
- Final pollution level is defined based on 3 partial numerical pollution criterion
  - S – total amount of trapped deposit (collected deposit of air pollution particles)
  - Sr – the amount of soluble substances of trapped deposit
  - g02 – the electrical conductivity of their 0.2% water solution of trapped deposit
- Measuring processes of these 3 criterions are difficult
  - Each measuring takes 6 weeks term
  - It require a special measuring device installed on pylon
  - Measuring Sr and g02 also require a special labolatory analyses
  - maintainance of devices

Measuring of these criterions are complicated and expensive. So it is tendency to modelling them based on another attributes / variables, which are measured and monitored by Slovak Hydrometeorological Institute by law from environmental and healthcare monitoring reasons.

Our Previous Research

- Computing and Informatics 2022 – Krammer, Kvassay, Forgáč, Očkay, Skovajsová, Hluchý, Skurčák, Pavlov: **Regression Analysis and Modeling of Local Environmental Pollution Levels for the Electric Power Industry Needs**

  { https://www.cai.sk/ojs/index.php/cai/article/view/2022_3_861 }

  - goal definition, analysis, possible approaches, problems

  [classification–strong class imbalance; regression–significant stochastic character]

- MDPI – Future Internet 2022 special section: Krammer, Kvassay, Mojžiš, Kenyeres, Očkay, Hluchý, Pavlov, Skurčák: **Using Satellite Imagery to Improve Local Pollution Models for High-Voltage Transmission Lines and Insulators**

  { https://www.mdpi.com/1999-5903/14/4/99 }

  - model improvement using extra attributes calculated from satellites information

It is still necessary to improve an accuracy of models for practical application and deployment of model in industry.

Overview of satellites spectral bands for definition of attributes

List of used satellite spectral bands including normalized difference indices.

Name	Scale	Pixel Size	Wavelength	Description
AOT	0.001	10 m		Aerosol optical thickness
B1	0.0001	10 m	443.9 nm	Aerosols
B2	0.0001	10 m	496.6 nm	Blue
B3	0.0001	10 m	560.0 nm	Green
B4	0.0001	10 m	664.5 nm	Red
B6	0.0001	20 m	740.2 nm	Red Edge 2
B8	0.0001	10 m	835.1 nm	NIR
L1 B10 cir	0.0010	60 m	1373.5 nm	Cirrus
B11	0.0001	20 m	1613.7 nm	SWIR 1
NDVI (normalized difference vegetation index)	0.0001	10 m		NDVI = (B8 − B4)/(B8 + B4)
NDWI (normalized difference water index)	0.0001	10 m		NDWI = (B3 − B8)/(B3 + B8)
NDSI (normalized difference soil index)	0.0001	20 m		NDSI = (B3 − B11)/(B3 + B11)
Moisture index	0.0001	20 m		moisture index = (B8 − B11)/(B8 + B11)

Group of attributes	Number of attributes in group	Description
SAT – satellite attributes	2340	13 spectral bands {B1,B2,B3,B4, B6, B8, B10, B11, NDVI, NDWI, NDSI, AOT,Mois } ▪ 5 representations {val, absroz, absdif, dif, roz} ▪ 6 time sequence calculation {min, max, avg, Q25, Q50, Q75} ▪ 6 space calculation {min, max, avg, Q25, Q50, Q75}.
RAD – Radar attributes about Rainfall evaluated for each day and then recalculated for 6 week period	580	Processed separately for temporal and spatial data. 6 attributes expressing the frequency of precipitation occurrences with graduated intensity of precipitation (up to 6 levels), 574 attributes = 7 correction methodologies radar, v1, v2 . . . v6 ▪ 82 attributes. The relevant 82 attributes consisted of 49 attributes, using functions (min, max, avg, stdev, Q25, Q50, Q75), to create pairs of functions for time aspect (7) ▪ spatial aspect (7) and 33 attributes using functions (min, max, avg, stdev, Q90, Q95, Q99) for temporal and spatial aspects separately (if the pair is not included in the group above).
RAD2 - Radar attributes about Rainfall calculated for 6 week period	42	The values of expected precipitation in the spatial and temporal surroundings were grouped into a set, to which one of 6 functions (avg, max, stdev, Q25, Q50, Q75) was applied, with 7 different correction methodologies (radar, v1, v2 . . . v6) for outliers removing.
SHMU attributes about air pollution (attributes from Slovak Hydrometeorological Institute about air pollution)	5	PM10 - yearly average of concentrations of dust particles with a diameter less than 10 µm. PM2.5 - yearly average of concentrations of dust particles with a diameter less than 2.5 µm. NO2 - yearly average of nitrogen dioxide concentrations. SO2 - yearly average of sulfur dioxide concentrations. O3 - yearly average of ozone concentrations.
Spatiotemporal Attributes	4	GPS-LON - GPS longitude; GPS-LAT - GPS latitude; ELEV - elevation value; Collecting number - represents average date of 6-week measuring proces.
Total original Input attributes	2971	

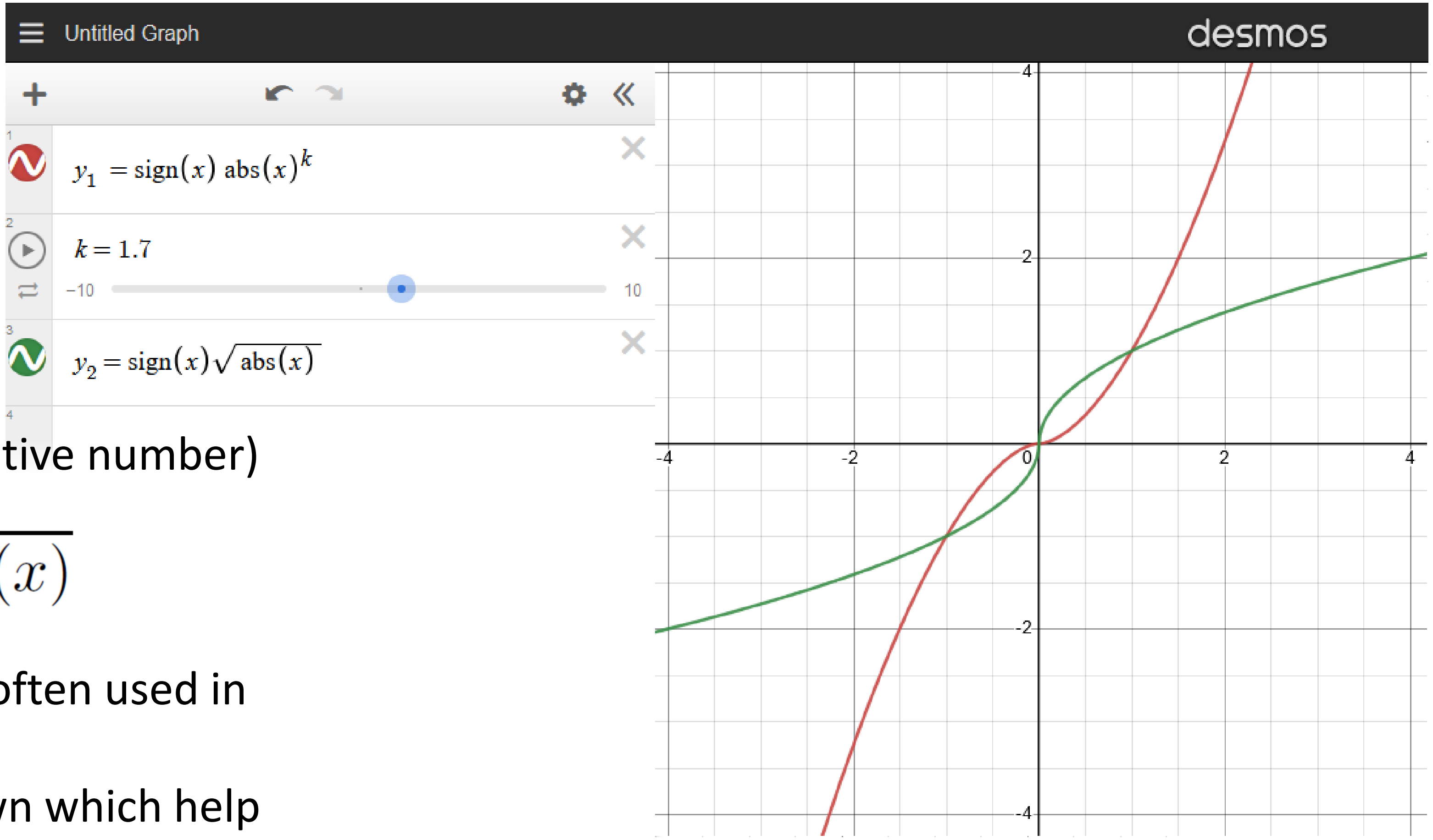
Non-linear transformation of input attributes

We also used non-linear transformed attributes from original
2971 attributes.

- Continuous function
- Differentiable function
- Defined for R (not only for positive number)

$$y = signum(x) \cdot \sqrt{abs(x)}$$

- Similar with sigmoid, which is often used in neural networks
- Change of slope is slowing down which help to represents multiple effects from nature (saturation effect – for example for radar reflectance)



Attribute Selection phase

- Used Forward selection method with model – linear regression
  - Without interactions
  - With interactions between attributes
- Calculated Criterions: Root Mean Square Error, Rsquared and LogLikelyhood
- Takes more than 8 hours
- Selected attributes has strong various; there are attributes from radar, satellites and also SHMU.
- We also tested another selection methods, but with worse results.

Overview of the most significant input attributes according to the forward selection method, using a linear regression model, for the target attribute Sr.

order	index	name	RMSE	Rsquare _Adj	LogLike H
1	81	RAD_v1_max _P95	0.003118	0.295244	1088.934
2	3003	SQRT_RAD_rad ar_max_P95	0.002861	0.406426	1110.900
3	1642	B11_val_Q25T _Q25S	0.002765	0.445616	1119.945
4	2776	RAD_v1_avg _Q50	0.002691	0.474955	1127.251
5	1698	B11_roz_minT _Q75S	0.002647	0.491888	1131.860
6	1875	NDVI_roz_minT _avgS	0.002592	0.512858	1137.642
7	3322	SQRT_RAD_rain _amount5	0.002555	0.526668	1141.752
8	3255	SQRT_RAD_v5 _avg_P95	0.002516	0.540954	1146.101
9	4610	SQRT_B11_val _avgT_minS	0.002473	0.556446	1150.912
10	3268	SQRT_RAD_v6 _P90_P90	0.002429	0.572210	1155.957
11	1701	B11_roz_maxT _avgS	0.002385	0.587447	1161.015
12	4814	SQRT_NDVI_val _minT_minS	0.002348	0.600270	1165.488
13	5379	SQRT_AOT_roz _Q50T_maxS	0.002322	0.609202	1168.841
14	2967	PM10	0.002293	0.618881	1172.507
15	5938	SQRT_PM10	0.002256	0.631146	1177.129
16	2628	mois_dif_Q75T _Q75S	0.002240	0.636214	1179.393
17	1	collecting_number	0.002223	0.641539	1181.774
18	3121	SQRT_RAD_v3 _P90_P90	0.002210	0.645983	1183.874
19	2408	AOT_roz_Q50T _maxS	0.002197	0.649955	1185.826
20	349	RAD_rain _amount3	0.002183	0.654597	1188.040

Overview of the most significant input attributes according to the forward selection method, using a linear regression model with interactions (mutual products of inputs), for the target attribute Sr.

order	index	name	RMSE	Rsquare _Adj	LogLike H
1	81	RAD_v1_max _P95	0.003118	0.295244	1088.934
2	2776	RAD_v1_avg _Q50	0.002593	0.512532	1136.025
3	1641	B11_val_Q25T _avgS	0.002480	0.553939	1148.654
4	1839	NDVI_val_minT _avgS	0.002401	0.582448	1158.985
5	5379	SQRT_AOT_roz _Q50T_maxS	0.002322	0.609011	1169.844
6	101	RAD_v2_P90 _P90	0.002257	0.630614	1180.196
7	1337	B8_roz_minT _Q50S	0.002165	0.660083	1194.486
8	2969	NO2	0.002078	0.687046	1209.426
9	4454	SQRT_B10_val _minT_minS	0.001989	0.713026	1225.655
10	2774	RAD_v1_max _Q75	0.001882	0.743208	1245.828
11	5550	SQRT_mois_roz _Q25T_Q50S	0.001782	0.769673	1266.721
12	939	B4_val_minT _avgS	0.001687	0.793614	1288.917
13	1672	B11_roz_avgT _Q25S	0.001609	0.812263	1310.640
14	11	RAD_radar_P95 _P95	0.001485	0.840193	1342.372
15	4929	SQRT_NDVI _absroz_maxT _maxS	0.001375	0.862946	1375.321
16	446	B1_roz_maxT _maxS	0.001255	0.885809	1414.687
17	3344	SQRT_GPS_lon	0.001141	0.905537	1458.774
18	227	RAD_v4_max _P90	0.001040	0.921579	1507.993
19	311	RAD_v6_P99 _P90	0.000914	0.939434	1575.185
20	3319	SQRT_RAD _rain_amount2	0.000717	0.962721	1687.594

Trained random forest models

Model:
Random Forest
with 100 trees

Validation:
40-fold Cross Val.

Process was repeated 20-times for different seeds (for stat. test)

Table shows average performance.

Forward Selection with Interaction										
Characteristic \ Number of inputs	20	19	18	17	16	15	14	13	12	11
Correlation coefficient	0.3264	0.4031	0.3625	0.3057	0.2997	0.2938	0.2829	0.2898	0.299	0.3034
Root mean squared error	0.0035	0.0034	0.0035	0.0035	0.0036	0.0036	0.0036	0.0036	0.0036	0.0036
Relative absolute error	88.63%	87.74%	88.05%	90.56%	91.49%	89.77%	92.83%	89.98%	92.69%	93.79%
Root relative squared error	94.38%	91.28%	92.96%	95.42%	95.66%	95.75%	96.74%	96.37%	96.00%	96.15%
R-squared	0.1065	0.1625	0.1314	0.0935	0.0898	0.0863	0.0800	0.0840	0.0894	0.0921
Characteristic \ Number of inputs	10	9	8	7	6	5	4	3	2	1
Correlation coefficient	0.3041	0.3482	0.3641	0.3447	0.3357	0.3971	0.3728	0.3128	0.334	0.2214
Root mean squared error	0.0036	0.0035	0.0035	0.0035	0.0035	0.0034	0.0034	0.0036	0.0036	0.0037
Relative absolute error	93.24%	92.27%	91.58%	89.69%	90.99%	88.83%	87.56%	93.19%	93.01%	91.87%
Root relative squared error	95.95%	93.85%	93.32%	94.10%	94.82%	91.60%	92.66%	96.30%	95.86%	98.32%
R-squared	0.0925	0.1212	0.1326	0.1188	0.1127	0.1577	0.1390	0.0978	0.1116	0.0490
Forward Selection linear (without Interactions)										
Characteristic \ Number of inputs	20	19	18	17	16	15	14	13	12	11
Correlation coefficient	0.4204	0.4226	0.3978	0.4473	0.4042	0.4165	0.4074	0.3986	0.3781	0.3689
Root mean squared error	0.0034	0.0034	0.0034	0.0033	0.0034	0.0034	0.0034	0.0034	0.0034	0.0034
Relative absolute error	83.56%	86.32%	85.48%	83.31%	85.91%	86.45%	86.64%	88.45%	88.75%	88.78%
Root relative squared error	90.46%	90.32%	91.41%	89.39%	91.25%	90.79%	91.08%	91.38%	92.24%	92.63%
R-squared	0.1767	0.1786	0.1582	0.2001	0.1634	0.1735	0.1660	0.1589	0.1430	0.1361
Characteristic \ Number of inputs	10	9	8	7	6	5	4	3	2	1
Correlation coefficient	0.3889	0.4087	0.4185	0.402	0.4084	0.4067	0.4007	0.2371	0.2217	0.2214
Root mean squared error	0.0034	0.0034	0.0034	0.0034	0.0034	0.0034	0.0034	0.0037	0.0037	0.0037
Relative absolute error	88.66%	86.53%	85.19%	85.95%	85.49%	85.71%	86.95%	93.35%	91.86%	91.87%
Root relative squared error	91.79%	90.97%	90.50%	91.25%	90.96%	91.11%	91.67%	98.45%	98.31%	98.32%
R-squared	0.1512	0.1670	0.1751	0.1616	0.1668	0.1654	0.1606	0.0562	0.0492	0.0490

Non-parametric Permutation test

- We statistically tested both of the resulting 20-element sets of achieved correlation coefficients to verify the significance of the achieved increase in accuracy.
- To verify the increase in accuracy, we used a statistical non-parametric permutation test, with a number of permutations of 200,000, and a Significance level alpha = 0.05;
- Xn represents the 20-element set of correlation coefficients, from the new most accurate model Xo represents a 20-element set of correlation coefficients obtained from the model in [15].
- Null hypothesis : Xn and Xo have the same mean value;
- Alternative hypothesis : Xn and Xo do not have the same mean value.
- Test statistic: abs (mean(Xn) - mean (Xo));
- A p-value of **0.000281** was achieved for the statistical test implemented in this way, which is significantly lower than the limit of alpha = 0.05. We therefore **reject the null hypothesis**, which **indicates a significant difference between the two trained models.** Overall, the new Random Forest regression model achieves a statistically significant increase in accuracy, compared to the model presented in our previous reseach [15].

Conclusions

- Confirmation of hypothesis of Precipitation attributes (Radar attributes) influence
- Identification of most relevant input attributes for experts in energetical and environmental domains
- Regression model improvement for 2 of 3 defined target attributes
- Statistical improvement confirmation by non-parametric permutation test
- Future: we prepare more sophisticated method for attribute selection, which could significantly improve model accuracy. Also we prepare another input attributes (for example wet deposit) which could better to detect production / propagation of dust.