

Genetic programming and artificial neural network modeling of PM10 emission close to a steel plant

Modeliranje emisij PM10 ob železarni z genetskim programiranjem in nevronske mreže

Miha Kovačič^{1,*}, Sandra Senčič², Uroš Župerl³

¹Štore Steel, d.o.o., Železarska cesta 3, SI-3220 Štore, Slovenia

²Kova, d.o.o., Teharska cesta 4, SI-3000 Celje, Slovenia

³University of Maribor, Faculty of Mechanical Engineering, Smetanova ulica 17, SI-2000 Maribor, Slovenia

*Corresponding author. E-mail: miha.kovacic@store-steel.si

Abstract

To implement sound air quality policies, regulatory agencies require tools to evaluate the outcomes and costs associated with various emission reduction strategies. The applicability of such tools can also remain uncertain. It is furthermore known that source-receptor models cannot be implemented through deterministic modeling. The article presents an attempt of PM10 emission modeling carried close to a steel production area with the genetic programming and artificial neural network method. The daily PM10 concentrations, daily rolling mill and steel plant production, meteorological data (wind speed and direction – hourly average, air temperature – hourly average and rainfall – daily average), weekday and month number were used for modeling during a monitoring campaign of almost half a year (23. 6. 2010 to 12. 12. 2010). The genetic programming modeling results show superior agreement with measured daily PM10 concentrations.

Key words: steel plant, PM10 concentrations, modeling, genetic programming, artificial neural network

Izvleček

Za implementacijo politike kakovosti zraka so od regulativnih agencij zahtevana orodja za ovrednotenje rezultatov in stroškov, povezanih s strategijami za zmanjšanje emisij. Uporaba takšnih orodij ostaja negotova. Velja, da za modele tipa vir-prejemnik težko uporabljamo deterministično modeliranje. V članku je predstavljen poskus modeliranja emisij delcev PM10 v bližini jeklarne z genetskim programiranjem in nevronske mreže. Pri skoraj polletni merilni kampanji (od 23. 6. 2010 do 12. 12. 2010) smo zbirali podatke o dnevni koncentracijah delcev PM10, dnevni proizvodnji valjarne in jeklarne, meteorološke podatke (urno povprečje smeri in hitrosti vetra ter temperature, dnevne padavine), o zaporednem številu dni v tednu in mesecu v letu. Rezultati, dobljeni z genetskim programiranjem, kažejo izjemno dobro ujemanje z izmerjenimi koncentracijami PM10.

Ključne besede: železarna, koncentracije PM 10, modeliranje, genetsko programiranje, nevronske mreže

Introduction

Particulate matter (PM) pollution is, especially in residential areas near industrial areas, a problem of great concern. This is not only because of the adverse health effects but also because of reduced visibility.^[1-3]

To reduce PM levels in the air a deep knowledge of the contributing sources, background emissions, the influence of the meteorological conditions, as well as of PM₁₀ formation and transport processes is needed.

However, current state-of-the-art PM₁₀ modeling does not allow us to quantitatively model the whole range of emissions behavior, which is why the dispersion modeling is thus increasingly connected with intelligent algorithms such as artificial neural networks^[4-9] and evolutionary computation.^[9]

The objective of this work was to model PM₁₀ emissions close to a steel plant area in Slovenia by means of a genetic programming and artificial network method. Genetic programming and neural network method have been proven to be an effective optimization tool for multicriterial and multiparametrical problems.^[10-13]

In paper is organized that the basic terms and experimental setup are stated in the beginning. Afterwards the idea of the proposed concept is presented. In the conclusion the main contributions of the performed research are summarized, while guidelines for further research are provided.

Experimental setup

Sampling sites

Figure 1 shows the locations of the sampling sites, rolling mill, steel plant and residential areas. Influencing PM₁₀ sources are rolling mill and steel plant, combustion and non-combustion traffic and urban background.

Sampling

Samples for this study were collected between 23. 6. 2010 and 12. 12. 2010. Sampling was performed 1.5 m above the ground. PM₁₀ samples were collected for 24 h on Mondays using low-volume samplers equipped with EPA-equiva-

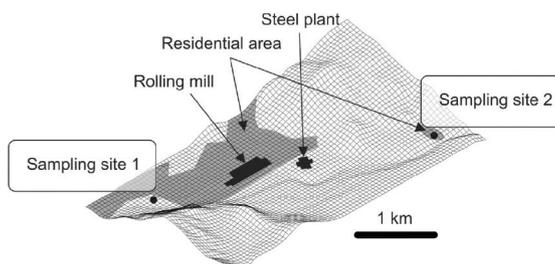


Figure 1: Topographic view of the study area.

lent size-selective inlets. Particles with diameter 10 μm (PM₁₀) were collected on cellulose esters membranes with high collection efficiencies (99 %). In total 172 PM₁₀ samples for each sampling site were available.

Before and after the samplings were made the filters were exposed for 24–48 h on open but dust-protected sieve-trays in an air-conditioned weighing room. The gravimetric determination of the mass was carried out using an analytical microbalance (precision 1 μg) located in the weighing room. In order to remove static electricity from filters the balance is equipped with a special kit in a Faraday shield.

The limit value of the EU directive – i.e. a daily mean PM₁₀ concentration – is 50 $\mu\text{g}/\text{m}^3$. At the sampling site 1 and 2 the measured PM₁₀ concentration exceeded limit value four times and five times, respectively.

Figure 2 shows the measured PM₁₀ concentrations during the study period for the sampling sites.

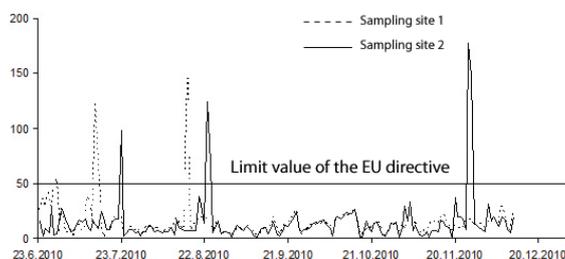


Figure 2: The measured PM₁₀ concentrations during the study period for the sampling sites.

Meteorological data

Hourly average air temperature, wind speed and direction and daily rainfall data were made available to the authors by the Slovenian Environment Agency.

Figure 3 shows the hourly average temperatures during the study period.

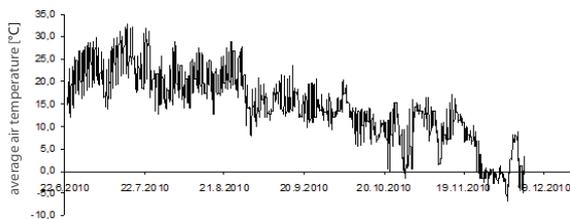


Figure 3: The hourly average temperatures during the study period.

Figure 4 shows the frequency distribution of wind direction and wind speed obtained based on wind direction and speed data measured every hour during the study period.

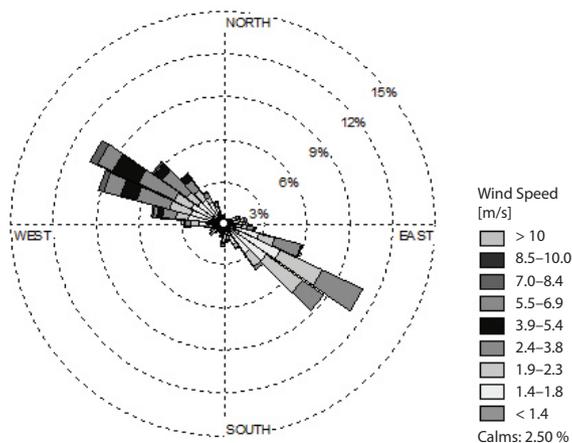


Figure 4: Frequency distribution of wind direction and wind speed.

Figure 5 shows the daily rainfall during the period of the study.

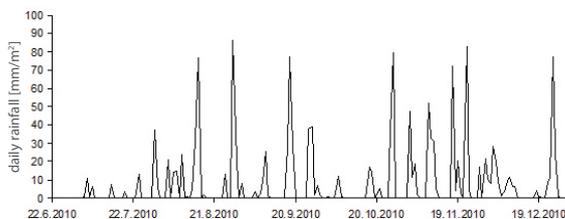


Figure 5: Daily rainfall during the study period.

The hourly data based on electric arc and rolling mill production was collected during the study period. During the study period, the electric arc furnace was stopped for 28 465 min and the rolling mill was stopped for 8 213 min. Figure 6 shows the minutes of stopping per day for the electric arc furnace and rolling mill during the study period.

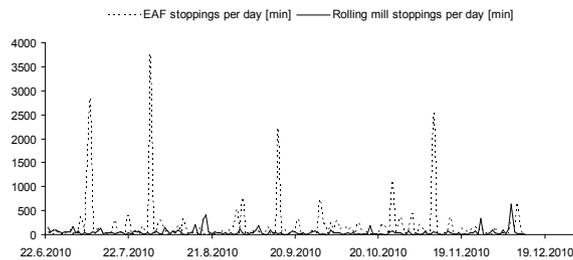


Figure 6: Minutes of stopping per day for the electric arc furnace and rolling mill during the study period.

Genetic programming modeling

Genetic programming is probably the most general evolutionary optimization method.^[14] The organisms that undergo adaptation are in fact mathematical expressions (models) for the PM10 concentrations prediction in the present work. The concentration prediction is based on the available function genes (i.e., basic arithmetical functions) and terminal genes (i.e., independent input parameters, and random floating-point constants). In the present case the models consist of the following function genes: addition (+), subtraction (-), multiplication (*) and division (/), and the following terminal genes: weekday (*WEEKDAY*) and month number (*MONTH*), wind speed [m/s] (*SPEED*), wind direction [°] (*DIRECTION*), air temperature [°C] (*TEMP*), rainfall [ml] (*RAIN*), electro arc furnace efficiency [min/hour] (*EAF*), rolling mill efficiency [min/hour] (*ROLLING*). In order to ascertain the influence of seasons and traffic during workday hours the weekday and month number were also added as terminal genes. One of the randomly generated mathematical models

$$\left(\frac{TEMP}{RAIN} + 2\right) \cdot \frac{EAF}{5.1} \quad (1)$$

is schematically represented in Figure 7 as a program tree with included function genes (*, +, /) and terminal genes (*TEMP*, *RAIN*, *EAF* and a real number constants 2 and 5.1). Random computer programs of various forms and lengths are generated by means of the selected genes at the beginning of the simulated evolution. The varying of the computer programs is performed by means of the genetic operations during several iterations, known as generations. After the completion of the variation of the computer programs a new genera-

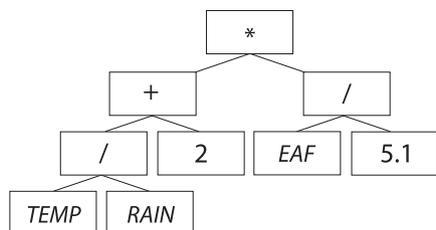


Figure 7: Randomly generated mathematical model for the PM10 concentrations prediction, represented in program tree form.

tion is obtained. Each generation is compared with the experimental data. The process of changing and evaluating organisms is repeated until the termination criterion of the process is fulfilled. The maximum number of generations is chosen as a termination criterion in the present algorithm.

The following evolutionary parameters were selected for the process of simulated evolutions: 500 for the size of the population of organisms, 100 for the maximum number of generations, 0.4 for the reproduction probability, 0.6 for the crossover probability, 6 for the maximum permissible depth in the creation of the population, 10 for the maximum permissible depth after the operation of crossover of two organisms, and 2 for the smallest permissible depth of organisms in generating new organisms. Genetic operations of reproduction and crossover were used. For selection of organisms the tournament method with tournament size 7 was used. 100 independent civilizations of mathematical models for prediction of the PM10 concentration were developed. The best evolution sequence of 100 generations was computed in 8 h and 41 min on 2.39 GHz processor and 2 GB of RAM by an AutoLISP based in-house coded computer program.

The model fitness f has been defined as:

$$f = \sum_{i=1}^n (P_i - M_i) + N \cdot 10\,000 \quad (2)$$

where n is the size of sample data and, P_i is predicted PM10 concentration, M_i is measured PM10 concentration and N is the number of all cases when:

$$P_i < 50 \wedge M_i > 50 \vee M_i < 50 \wedge P_i > 50$$

The limit value of the EU directive, i.e. a daily mean PM10 concentration, is $50 \mu\text{g}/\text{m}^3$. The

number N tells us when the prediction is above that limit value, when in order to assure PM10 concentration exceedance prediction by developed predictive model it should in fact be below the limit, and also when prediction by developed predictive model should be above the limit.

The simulated evolution in one run of the genetic programming system (out of 100) produced the following best model for prediction of PM10 concentration for sampling site 1 (cf. equation 3) with fitness of 1 019.95, number $N = 0$ and average deviation of $5.96 \mu\text{g}/\text{m}^3$.

The best evolutionary developed model (out of 100) for prediction of PM10 concentration for sampling site 1 (cf. equation 4) with fitness of 11 124.67, number $N = 1$ (on the 30. 6. 2010 the measured PM10 concentrations were $53.6 \mu\text{g}/\text{m}^3$ and predicted $21.41 \mu\text{g}/\text{m}^3$), and average deviation of $6.54 \mu\text{g}/\text{m}^3$.

Figures 8 and 9 show measured and predicted PM10 concentrations for sampling sites 1 and 2, respectively.

Artificial neural network modeling

Artificial neural networks consist of a large number of processing elements, called neurons that operate in parallel. Computing with neural networks is non-algorithmic. They are trained through examples rather than programmed by software. The Multi-Layer BP network is a supervised, continuous valued, multi-input and multi-output feedforward multi-layer network that follows a gradient descent method. [14]

The gradient descent method alters the weight by an amount proportional to the partial derivative of the error with respect to the weight in question. The backpropagation phase of the neural network alters the weights w_{ji} so that the error of the network is minimized. This is achieved by taking a pair of input/output vectors and feeding the input vector into the net. The net generates an output vector and than the output vector is compared to the output vector supplied. The comparison gives us the error value. The error is then passed back through the network (backpropagation process), modifying the weights due to this error using the

$$\begin{aligned}
& \left(DIRECTION + \left(DIRECTION + MONTH - WEEKDAY + \frac{SPEED (-1.99563 + 3WEEKDAY + 4MONTH WEEKDAY)}{MONTH - WEEKDAY} + \right. \right. \\
& \left. \left. + \frac{WEEKDAY (-1.99563 + MONTH WEEKDAY)}{MONTH - \frac{1}{2} + MONTH WEEKDAY} \right) \right) / \left(\frac{-1.99563 + MONTH WEEKDAY}{MONTH + WEEKDAY - \frac{-1.99563 + MONTH WEEKDAY}{MONTH - \frac{WEEKDAY}{SPEED}}} + \right. \\
& \left. + \left((-1.99563 + MONTH WEEKDAY) (-MONTH + (DIRECTION - WEEKDAY)) / (DIRECTION + WEEKDAY + \right. \right. \\
& \left. \left. + MONTH WEEKDAY - \frac{MONTH SPEED (-1.99563 + MONTH WEEKDAY)}{-DIRECTION + MONTH + WEEKDAY} - \frac{MONTH^2 (-1.99563 + MONTH WEEKDAY)}{MONTH - \frac{WEEKDAY}{DIRECTION}} \right) \right) / \\
& \left(\frac{MONTH - 2WEEKDAY + MONTH WEEKDAY}{DIRECTION WEEKDAY} \right) / \left(DIRECTION - RAIN + \left(MONTH - \frac{MONTH - WEEKDAY}{WEEKDAY} \right) + \right. \\
& \left. + MONTH WEEKDAY - \left(-MONTH + \left(MONTH - \frac{MONTH - WEEKDAY}{SPEED} \right) \right) (-1.99563 + MONTH WEEKDAY) \right) \quad (3)
\end{aligned}$$

$$\begin{aligned}
& \left(-5.76596 + MONTH + \frac{TEMP}{DIRECTION} \right) \left(\frac{-\frac{SPEED}{TEMP} + \frac{1 - \frac{SPEED}{TEMP}}{WEEKDAY}}{MONTH} + \frac{\left(EAF + \frac{MONTH + \frac{TEMP}{DIRECTION} - WEEKDAY}{SPEED} + \frac{TEMP}{DIRECTION TEMP} + \frac{TEMP}{EAF} \right) WEEKDAY}{EAF + \frac{SPEED}{DIRECTION TEMP} + \frac{TEMP}{EAF}} \right) \\
& \left(\frac{TEMP}{TEMP + \frac{1 - \frac{SPEED}{TEMP}}{WEEKDAY}} + \frac{EAF RAIN SPEED}{TEMP (-TEMP + \frac{TEMP}{WEEKDAY}) WEEKDAY} \right) \\
& \left. \right) / DIRECTION WEEKDAY + \\
& \frac{MONTH - \frac{WEEKDAY^2}{TEMP} + \frac{SPEED TEMP^2}{WEEKDAY \left(DIRECTION + \frac{TEMP - (TEMP + WEEKDAY)}{DIRECTION} \right)}}{DIRECTION + \frac{MONTH - WEEKDAY + \frac{TEMP + \frac{EAF WEEKDAY}{SPEED TEMP} + \frac{-TEMP + WEEKDAY}{DIRECTION}}{EAF + \frac{EAF RAIN}{TEMP WEEKDAY} + \frac{-TEMP + WEEKDAY + \frac{-TEMP + WEEKDAY}{DIRECTION}}{DIRECTION}}}{DIRECTION}} + 0.30872 \quad (4)
\end{aligned}$$

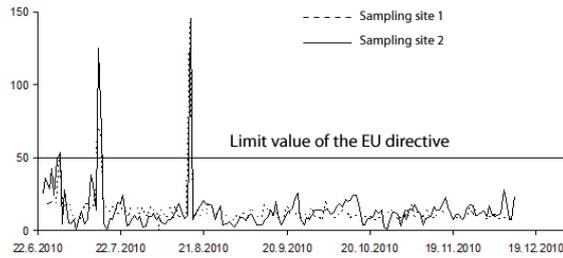


Figure 8: Measured and predicted PM10 concentrations [$\mu\text{g}/\text{m}^3$] for sampling site 1.

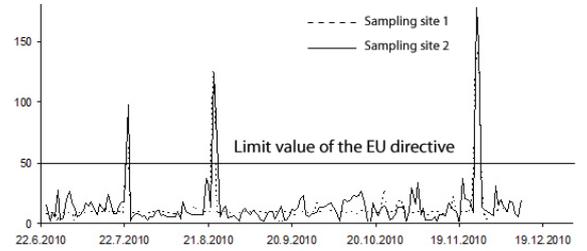


Figure 9: Measured and predicted PM10 concentrations [$\mu\text{g}/\text{m}^3$] for sampling site 2.

equations. Hence, if the same set of input/output vectors are presented to the network, the error would be smaller than previously found. For modeling the PM10 emission, three-layer feed-forward neural networks were used (Figure 10). They contained 9 neurons in the input layer, and 1 in the output layer. The number of

neurons in the hidden layer was varied in different experiments.

The detailed topology of the used ANN with optimal training parameters and mathematical principle of the neuron is shown on Figure 3. The ANN were trained with the following parameters: weekday ($WEEKDAY$) and month

number (*MONTH*), wind speed [m/s] (*SPEED*), wind direction [°] (*DIRECTION*), air temperature [°C] (*TEMP*), rainfall [ml] (*RAIN*), electro arc furnace efficiency [min/h] (*EAF*), rolling mill efficiency [min/h] (*ROLLING*).

The ANN registers the input data only in the numerical form therefore the information about the inputs must be transformed into numerical code. The learning method is error backpropagation. Signals passed through the neurons in the hidden and output layers are transformed on the basis of an Tangent (nonlinear) function which allows the identification of the nonlinear system. The data is automatically normalized in order to make the training process faster. This was done by mapping each term to a value between 0 and 1 using the Max Min method. This normalized data was utilized as the inputs and outputs to train the ANN. In other words, two vectors are formed in order to train the neural network: Input vector is [*WEEKDAY*, *MONTH*, *SPEED*, *DIRECTION*, *TEMP*, *RAIN*, *EAF*, *ROLLING*, *HOURL*]. The output vector is [PM10 concentration].

Training of the ANN is finished when the testing error is less than the tolerance limit. This tolerance limit is defined to 4 % at the beginning of the training. On average, the networks needed 63 iterations to achieve this goal. Approximately 8 min of training during machin-

ing are needed to set up the full prediction performance of ANN. After the neural network had been trained it was applied to 50 examples that did not take part in the training process. This time the solutions of the examples (PM10 concentration) were not supplied, so that the network had to estimate them.

To evaluate the individual effects of network topology and training parameters on the performance of neural network 40 different networks were trained, tested and analyzed. From the results the following conclusions can be drawn:

- LWrs.
- To minimize the estimation errors, momentum rates between 0.001 and 0.005 are good.
- It is found that there is an optimum number of hidden nodes beyond there is no significant change in the error prediction. In this instance, the optimum number of hidden layer nodes is 3.
- Networks trained with the tanh transfer function in all their processing elements give the least prediction errors, while those employing sigmoid and sine give the highest and next highest prediction errors respectively.
- By using a multi-layer perceptron with backpropagation training method, the neural network is trained to an accuracy of $\pm 3\%$ error.

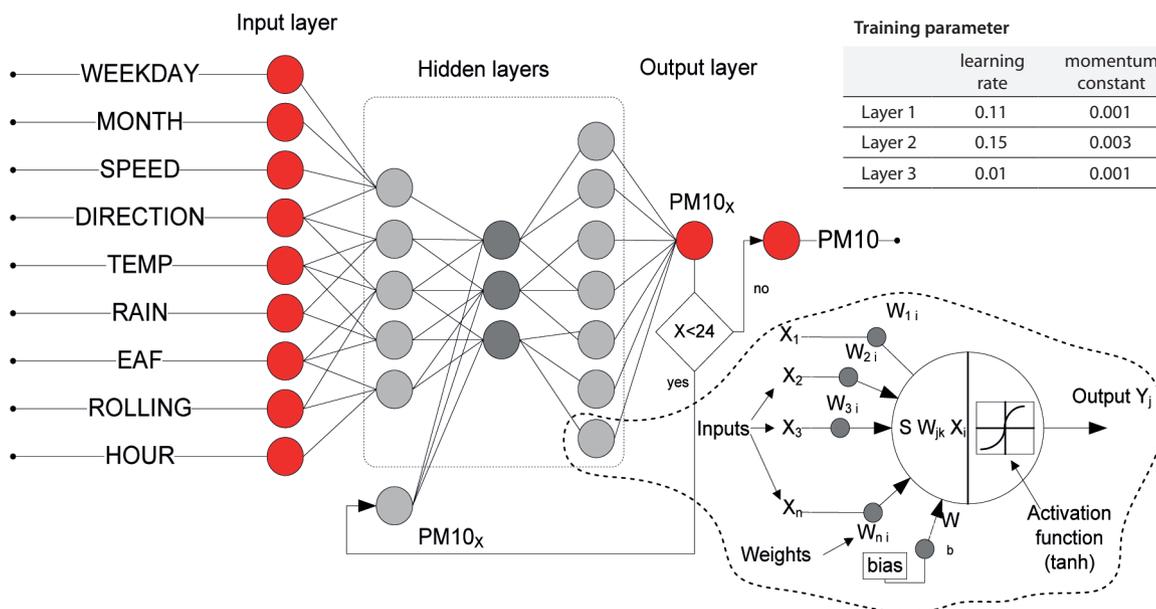


Figure 10: Predictive force model topology.

In testing the model, the PM10 concentrations for sampling site 1 and 2 were predicted with average deviation of $9.47 \mu\text{g}/\text{m}^3$ and $11.92 \mu\text{g}/\text{m}^3$, respectively. The number of cases N when the prediction is above the EU directive limit value, when in order to assure PM10 concentration exceedance prediction by developed predictive model it should in fact be below the limit, and also when prediction by developed predictive model should be above the limit is for both sampling sites 6.

Figures 11 and 12 show measured and predicted PM10 concentrations for sampling sites 1 and 2, respectively.

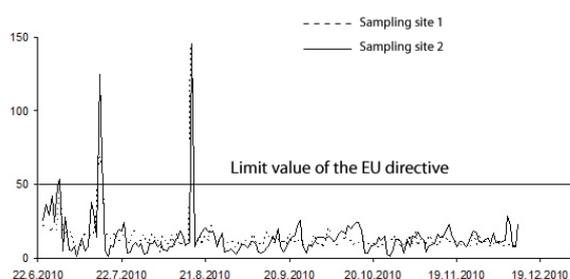


Figure 11: Measured and predicted PM10 concentrations [$\mu\text{g}/\text{m}^3$] for sampling site 1.

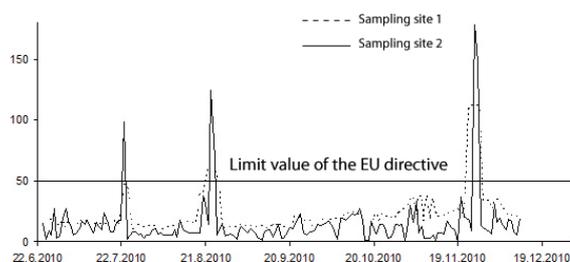


Figure 12: Measured and predicted PM10 concentrations [$\mu\text{g}/\text{m}^3$] for sampling site 2.

The distribution of concentrations deviation of training, test and verification data is presented in the Table 1.

Table 1: Distribution of concentrations deviation of training, test and verification data.

	Sampling site 1			Sampling site 2		
	Training data	Testing data	All data	Training data	Testing data	All Data
N	121	50	171	121	50	171
Average [$\mu\text{g}/\text{m}^3$]	5.85	18.23	9.47	3.41	14.23	6.58
St. dev. [$\mu\text{g}/\text{m}^3$]	4.53	14.63	10.39	3.17	9.72	7.66

Conclusions

This paper presented the possibility of the PM10 concentration prediction close to a steel plant area with genetic programming and artificial networks. The daily PM10 concentrations, daily rolling mill and steel plant production, meteorological data (wind speed and direction – hourly average, air temperature – hourly average and rainfall – daily average), weekday and month number were used for modeling during a monitoring campaign of almost half a year (23. 6. 2010 to 12. 12. 2010). The special fitness function for genetic programming system was designed in order to assure also PM10 limit value exceedance prediction. For each sampling site the best models for PM10 prediction were obtained from 100 runs of the genetic programming system. The model for sampling sites 1 and 2 predicts concentrations within an average error range of $5.96 \mu\text{g}/\text{m}^3$ and $6.54 \mu\text{g}/\text{m}^3$, respectively. All exceedances of the EU directive limit value ($50 \mu\text{g}/\text{m}^3$) were administered at sampling site 1, but only 4 out of 5 of these occurred at sampling site 2. In general it is also important to know how many times the prediction is above EU directive limit value when it should in fact (measured values) be below the limit and otherwise. The number of such cases at sampling site 1 and 2 are 0 and 2, respectively. Also the special artificial neural network topology adjustments were used. 40 different neural networks were trained, tested and analyzed. The best artificial neural network for sampling sites 1 and 2 predicts concentrations within an average error range of $9.47 \mu\text{g}/\text{m}^3$ and $11.92 \mu\text{g}/\text{m}^3$, respectively. 2 out of 4 EU directive limit value ($50 \mu\text{g}/\text{m}^3$) exceedances were administered at sampling site 1 and only 4 out of 5 of these occurred at

sampling site 2. The number, when the prediction is above that limit value, when in order to assure PM10 concentration exceedance prediction by developed predictive model it should in fact be below the limit, and also when prediction by developed predictive model should be above the limit, was 6 at both sampling sites. In the future we will carry out genetic programming based dispersion modeling according to the calculated wind field, air temperature, humidity and rainfall in a 3D Cartesian coordinate system. The prospects for arriving at a robust and faster alternative to the well-known Lagrangian and Gaussian dispersion models are optimistic.

References

- [1] Marcazzan, G. M., Ceriani, M., Valli, G., Vecchi, R. (2003): Source apportionment of PM10 and PM2.5 in Milan (Italy) using receptor modeling. *The Science of the Total Environment*, 317, 137–147.
- [2] Watson, J. G. (2002): Visibility: science and regulation. *Journal of the Air and Waste Management Association*, 52, 628–713.
- [3] Vrins, E., Schofield, N. (2000): Fugitive dust emission by an ironmaking site. *Journal of Aerosol Science*, 31, 524–525.
- [4] Kukkonena, J., Partanena, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Cawley, G. (2003): Extensive evaluation of neural network models for the prediction of NO₂ and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki, *Atmospheric Environment*, 37, 4539–4550.
- [5] Zhou, H., Cen, K., Fan, J. (2004): Modeling and optimization of the NO_x emission, characteristics of a tangentially fired boiler with artificial neural networks. *Energy*, 29, 167–183.
- [6] Hooyberghsa, J., Mensink, C., Dumont, G., Fierens, F., Brasseur, O. (2005): A neural network forecast for daily average PM10 concentrations in Belgium. *Atmospheric Environment*, 39, 3279–3289.
- [7] Perez, P., Reyes, J. (2006): An integrated neural network model for PM10 forecasting. *Atmospheric Environment*, 40, 2845–2851.
- [8] Grivas, G., Chaloulakou, A. (2006): Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. *Atmospheric Environment*, 40, 1216–1229.
- [9] Kovačič, M., Uratnik, P., Brezočnik, M., Turk, R. (2007): Prediction of the bending capability of rolled metal sheet by genetic programming. *Materials and Manufacturing Processes*, 22, 634–640.
- [10] Kovačič, M., Šarler, B. (2009): Application of the genetic programming for increasing the soft annealing productivity in steel industry. *Materials and Manufacturing Processes*, 24(3), 369–374.
- [11] Kovačič, M. (2009): Genetic programming and Jominy test modeling. *Materials and Manufacturing Processes*, 24(7), 806–808.
- [12] Kovačič, M., Senčič, S. (2010): Critical inclusion size in spring steel and genetic programming. *RMZ - Materials and Geoenvironment*, 57(1), 17–23.
- [13] Koza, J. R. (1999): Genetic Programming III. Morgan Kaufmann, San Francisco, 3–16.
- [14] Župerl, U., Čuš, F., Kiker, E. (2009): Adaptive network based inference system for estimation of flank wear in end-milling. *Journal of Materials Processing Technology*, 209(3), 1504–1511.