

STATISTICAL/MACHINE LEARNING FOR SURROGATE PARAMETERS AIR EMISSION MONITORING FROM INSTALLATIONS

Damir Rumenjak ^{1*}

¹ Domagojeva 9, Zagreb

*E-mail of corresponding author: damir.rumenjak@outlook.com

Abstract: Statistical/machine learning is discussed as a part of permit conditions for monitoring emissions into air using surrogate parameters. It is unavoidable step in establishing system of monitoring by models. Requirements for such learning, given by Directive for industrial emissions (IED) and Conclusions for best available techniques (BATC) are recognized. They are compared with requirements in standards for direct continuous emission monitoring and automated measuring systems and then use to broadly define statistical learning seeking the common principles that could be applied in permits. Such are found to be clear phases of learning introducing training, validation and testing, basic equations for learning, learning paths, blocking of observations and quality assurance based on statistical criteria. The findings are intended for monitoring practice using continuous monitoring of air emissions (mineral and energy sector and for waste incineration and co-incineration and even broader), needs permitting procedure and could use surrogate parameters models for monitoring emissions into air.

Keywords: 1st Statistical/machine learning, 2nd Emission monitoring, 3rd Surrogate parameters.

Received: 11.06.2025. / Accepted: 10.12.2025.

Published online: 17.12.2025.

Professional paper

1. INTRODUCTION

Statistical learning, according to definition (James et al. 2013), means finding statistically based relations between input and output variables and in case of surrogate parameters emission monitoring of relevant process parameters and emission. It is closely related with the term machine learning meaning developing algorithms on learning for specified goals.

Statistical learning has to establish the proper function between input and output of the models as supervised learning, obtaining models (commonly named learner) for purpose of prediction (Hastie et al. 2001). The approach is to use statistical learning function or loss function which in learning penalizes the error of prediction.

Models that could be proposed for surrogate parameters monitoring are on basic principles of chemical engineering modelling as mass balance, energy balance and process balance (Gomzi et al. 2019), regression models and artificial neural networks, on all statistical learning applies. Types of statistical learning, deep or shallow learning, are related to artificial neural networks, depending of numbers of layers for learning (Frana et al. 2024).

The obligation to put statistical learning in permit conditions arises from directive (Directive 2010/75/EU), if there was no document containing details for monitoring that could be invoked in permits. Documents on statistical learning still do not exist nor are demanded by existing legislation or standards.

Proposals for permit conditions for surrogate parameters air emission monitoring, containing conditions for statistical learning, are in February 2025 submitted to the Ministry of environmental protection for installations from energy and mineral industries. They have been primarily motivated by reasons that permits already using monitoring by surrogates' don't use statistical learning (Env. permit 2022), or use it in very general form (Env. permit 2021), such lacking reliabilities of monitoring comparing to direct emission monitoring and in this way, it is not in agreement with permit legislation.

BAT conclusions for waste incineration (BATC EU 2019/2010), as last issued conclusions supporting surrogate parameters monitoring, give even more possibilities for surrogate parameters monitoring in a regular and non-regular work that strengthens the importance of statistical learning. Perspectives of using surrogate parameters monitoring of emission from waste incineration and co-incineration are given in Rumenjak 2023.

2. BASIC REQUIREMENTS ON STATISTICAL LEARNING

2.1. Requirements of confidence limit interval for models

Requirements on confidence limit for air monitoring emission concentrations from large combustion plants

are given in Annex V Part 3 Point 9 and from waste incineration/co-incineration installations in Annex VI Part 6 Point 1.3. of Industrial emission directive ([Directive 2010/75/EU](#)). As statistical criteria they are also influencing the statistical learning.

Limit interval for single measured value for those cases is given as:

$$t_{95,df s_{im}} \cdot s_{im} \leq p \cdot ELV \quad (1)$$

where p is coefficient from directive, $t_{95,df}$, Student (t) distribution for 95% confidence level, df s_{im} degrees of freedom for t distributions, ELV emission limit values given in directive, s_{im} estimation of population standard deviation for the measurement. It could be demonstrated that a size of sample for **Equation 1** is $n = 1$.

In the case of surrogate parameters monitoring, limit interval should be defined differently and needs relation of variance of model s^2 with the variance of direct measurement (s_{im}^2) used in data pairs for learning. Taking into account the requirements of standard [HRN EN 14181 \(2014\)](#), part of quality assurance level QUAL 2 described in standards as variability test, it is defined for single calculation-measurement pair after testing as:

$$t_{95,df s_{im}} \cdot \sqrt{(s_{im}^2 + s^2 \cdot 2)} \leq p \cdot ELV \quad (2)$$

where the factor for model variance s^2 in **Equation 2** is approved for prediction models in statistical texts ([Mendenhall et al. 1988](#)).

In the case of difference between degrees of freedom for direct (control) measurement and phase of learning (here: validation phase) what is supposed as a general case in learning, equation holds:

$$t_{95,df s_{im}} \cdot \sqrt{\left(s_{im}^2 + \frac{s_{dfstest}^2}{s_{dfvalid}^2} \cdot s_{dfvalid}^2 \cdot 2\right)} \leq p \cdot ELV \quad (3)$$

where $s_{dfstest}^2$ is variance of model (equals s^2) with degrees of freedom equals degrees of freedom for direct control measurement, $s_{dfvalid}^2$ variance of model after validation phase, $\frac{s_{dfstest}^2}{s_{dfvalid}^2}$ ratio that could be derived from basic expression for Student (t) distribution enabling estimation of $s_{dfvalid}^2$ from variance of model after validation phase.

Correction for oxygen level (0%) (which is required by [Directive 2010/75/EU](#) for comparison with emission limit values (ELV) is:

$$t_{95,df s_{im}} \cdot \sqrt{\left(s_{im}^2 + s_{dfstest}^2 \Big|_{0\%} \cdot 2\right)} \leq p \cdot ELV \quad (4)$$

where $s_{dfstest}^2$ is variance of the model with degrees of freedom as for control measurement.

For mineral industries (glass and mineral wool and cement) instead of p , where p is not prescribed in Directive or BAT conclusions, factor k_v prescribed in [HRN EN 14181 \(2014\)](#) Annex I, is used in relation with measurement uncertainty (μ) ([Proposal 1](#), [Proposal 2](#)):

$$t_{95,df s_{im}} \cdot \sqrt{\left(s_{im}^2 + s_{dfstest}^2 \Big|_{0\%} \cdot 2\right)} \leq \mu \cdot k_v \cdot ELV \quad (4a).$$

When s estimation ($s_{dfstest}^2$) for model satisfies the conditions from **Equation 4 and 4a**, what is finally confirmed at the end of learning after testing, model reaches requirement comparable with direct continuous measurement of emissions given by Directive.

2.2. Requirements of monitoring results validation

The important requirements on learning are the surrogate parameters should be continuously measured ([Brinkmann et al. 2018](#)), to be in line with direct continuous emission monitoring. Model results are validated through sampling time for direct measurement by basic formula for validation required for continuous monitoring.

Equation for validated model results with one sided (lower) confidence interval for normal based distributions of results is:

$$V_{1/2 hr} = \hat{y}_{1/2 hr} - t_{95,df s} \cdot s \cdot \sqrt{\frac{n+1}{n}} \quad (5)$$

Symbol $\hat{y}_{1/2 hr}$ is used for models with predictive surrogates for concentrations (mostly 30 min average). **Equation 5** is written without part for random error of input justified by narrower confidence interval on the lower side of validation formula, but variants of **Equation 5** including that part have been also developed in **Proposal 1**. Symbol n denotes sample size used for sampling of surrogate parameters during the sampling time equals sampling size for direct measurement, where $V_{1/2 hr}$ are validated values for model results (30 min average), $t_{95,df}$ Student (t) distribution for 95% confidence level, s estimation of population standard deviation for models.

Situation with indicative surrogate parameters and their modelling is not different as long as serves for prediction, and results are validated the same, **Equation 5**. The indicative parameters give models for difference of emission concentrations (**Proposal 1**). Targeted functions for learning are generally indicated as \hat{y} or, in the case of indicative surrogate parameters, as $\Delta\hat{y}$.

3. DISCUSSION ON STATISTICAL/MACHINE LEARNING

3.1. Statistical learning models with mass balance models and regression models (predictive and indicative surrogate parameters)

Among models proposed for emission monitoring and use learning are mass balance models and linear regression models. Other models of balance types: energy and process balance models are still missing. Regression models are characterized by more parameters than other models.

Learning is basically provided by comparison of data pairs. Data pairs are pairs of calculated and measured values of output or other observables that must be supported by measured values of output as in validation phase. Basic learning phases are given in **Proposal 1** (for glass production) and **Proposal 2** (for aluminate cement and mineral wool production). Relatively many parameters have been used for input relations, especially in regression models in **Proposal 1**. The surrogate parameters entering learning are also summarized in those proposals.

Flow (block) diagram of learning is given in **Figure 1**. The learning paths for regular and work non-complied to requirements with phases and steps of learning contained blocks and observations are shown together with quality assurance (QUALs) comparable to standard **HRN EN 14181 (2014)**. The acceptance position of the model is also shown.

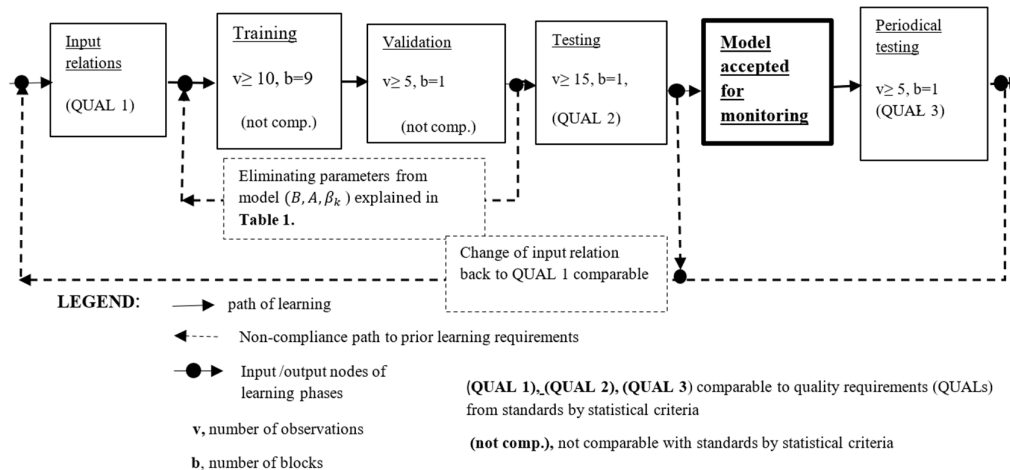


Figure 1. Flow (block) diagram of learning (with learning paths) in proposals for permit condition using balance and regression models-complete learning sequence including periodical testing

For eliminating the effects of noise factors, blocking strategy is used (**Jiji 2003**). Blocks are basically the steps of learning containing one or more observations (data pairs). Randomization of observations (in monitoring language it is a random selection of order of data pairs) is a common approach in experimental design, but only for a complete block design. Because of nature of monitoring process, randomization could be difficult and impractical. Although it is requested in proposals, it could be changed in permit conditions. Incomplete block design is far more convenient (**Dey 2010**). Some approaches to blocking e.g. replication and confounding, strictly mathematically speaking, are not convenient for monitoring emission, unless there is any possibility to control surrogate parameters in a way allowing those effects to be recognized in observations and compared with statistical errors. It can be observed that the blocking is mostly concentrated in training phase although it is possible in other phases of learning.

The lower number of observations in testing is determined by requirements of **HRN EN 14181 (2014)**, part QUAL 2, with minimum of 15 valid pairs of data. Providing QUALs or quality assurance level statistically

comparable to learning process in **Figure 1**, tends to assure quality of models as for the direct measurements if full adequacy with the standards for direct measurement is not possible because of technical reasons. QUAL 1 is to check uncertainties of input relations according simple formulas like **Equation 1**, but using in the equation total variances estimated during preparation of input relations instead of measurement variance. It is not possible to use QUALs from standards for training and validation phases of learning and number of observations there must be determined in other ways as it was done in proposals.

Requirements of quality assurance of **HRN EN 14181 (2014)**, level QUAL 3, are used for the formulation of the periodical testing with 5 data pairs of observations.

In the case of non-compliance results of testing (and periodical testing), it takes return to initial stage of the block diagram in **Figure 1**. Important for such commitment is position of authorities, who often seek for stricter solutions of those possible, so no other solution dealing with non-compliance has been considered.

The input relations for balance and linear regression in **B4 Control Solutions & FKIT (2022)** have been checked for uncertainties according to **Equation 1**, but checking has been using only model variances. **Proposal 1** qualifies them comparable to QUAL 1 requirements and leaves to be trained and tested further in learning.

For learning balance and regression models, **Equations 7** and **Equations 6**, together with the validation criteria and the reference on testing **Equation 4a**, are given in **Table 1**. They are used in **Proposal 2** and for the training and validation in **Proposal 1**. The testing in **Proposal 1** is to be on complex models trained models are part of as it has been described specifically in **Proposal 1**.

Table 1. Learning (balance models and regression models) recommended for NO_x, SO_x and dust emission monitoring

LEARNING PHASES	STATISTICAL FUNCTIONS FOR LEARNING
<p>Training phase of learning, Equations 7:</p> <ul style="list-style-type: none"> - for balance models: $(B, A) = \operatorname{argmin} \sum_{i=1}^M [y_i - \hat{y}(B, A)_i]^2$ - for regression (linear): $\min \operatorname{RSS}(\beta_o, \beta_k) = \min \sum_{i=1}^M \left[y_i - \beta_o - \sum_{k=1}^k x_{i,k} \beta_k \right]^2$ <p>and non-linear regression (non-linear functions f with linear basis expansion term θ (Hastie et al. 2001)): $\min \operatorname{RSS}(\theta) = \min \sum_{i=1}^M \left[y_i - \sum_{k=1}^k f(x_{i,k}) \theta_k \right]^2$ </p> <p>After training, linear regression models could be a subject to constraint shrinkage of coefficients (sh) for minimising variance of the model (Hastie et al. 2001): $\sum_{k=1}^k \beta_k^2 \leq sh$ </p> <p>Training targeting $B, A, \beta_o, \beta_k, \theta$: Methods for regression models: minimization of square error, other: downhill Simplex method in multidimension, direction set (Powell's method) in multidimension, conjugate gradient methods in multidimensions, variable metric methods in multidimensions, linear programming and the Simplex method, simulated annealing methods, etc. (Press WH et al, 2002)</p> <p>Validation phase of learning: Checking of sensitivity and importance of the individual surrogate parameter of model, is important for regression models (Smith J, Smith P (2007)). Data pairs for validation must be supported by the measured outputs of y_i through prediction error estimated for testing. From Equation 3 comparing variances in learning, with minimum number of observations in validation phase from Figure 1, and for 95% confidence level assuming $df_{sim} = 14$, non - Equation 8 stays as criteria for validation with condition: $s_{df\ valid.}^2 = 4 \cdot s_{df\ test.}^2$ </p> <p>and after arrangement: $s_{df\ valid.}^2 \Big _{0\%} \leq 2 \cdot \left(\frac{1}{t_{95, df_{sim}}} (\mu \cdot k_v \cdot ELV)^2 - s_{im}^2 \right) \quad (8)$ </p> <p>Then it goes to validation Equations 9.</p> <p>Testing: Testing and periodical testing is by Equation 4a</p>	<p>Learning functions (SLF) for training (loss function) Equations 6:</p> <p>For balance and regression models:</p> $L \text{ or } RSS = \sum_{i=1}^M [y_i - \hat{y}_i]^2,$ <p>where for regression models stays:</p> $\hat{y}_i = \beta_o + \sum_{k=1}^k x_{i,k} \beta_k,$ $\hat{y}_i = \sum_{k=1}^k f(x_{i,k}) \theta_k$ <p>Functions for validation, Equations 9:</p> <ul style="list-style-type: none"> - Sensitivity index for parameter: $S_{ind.} = \left \frac{\sum (y_i - \bar{y})}{\sum (x_i - \bar{x})} \right \cdot \frac{\bar{x}}{\bar{y}}$ - Importance index for parameter: $I_{imp.} = \left \partial \hat{y} / \partial x \right \Big _{\bar{X}} \cdot \frac{\sqrt{\sum (x_i - \bar{x})^2}}{\sqrt{\sum (y_i - \bar{y})^2}}$ <p>(to be noticed: importance parameters are always sensitive)</p> - Correlation coefficient for sensitivity of parameter: $r_{ii} = \frac{\sum_i^M x_i y_i}{\sqrt{\sum_i^M x_i^2 \sum_i^M y_i^2}}$

The symbols used in **Table 1** are: x_i input (surrogate) parameter for i th data pair, \hat{y}_i calculated values of the predictive model of i th data pair, y_i measured output values in data pair, i pair index of measured-calculated values for observations (data pairs), $\hat{y}(B, A)_i$ mass balance model function of coefficients from set B, A as arguments in i th data point, β_o, β_k coefficients of regression models, RSS residual sum of squares, $f(x_{i,k})$ function used in non-linear regression, θ_k linear basis expansion term of model parameter, M number of measured-calculated pairs, $argmin$ minimization of function on specified arguments, k number of parameters of regression model, \bar{X} point of calculation (in multidimensional space) for average values from set of input parameter averages \bar{X} , $s_{df\ valid}^2$ model variance from validation phase of learning, $s_{df\ test}^2$ model variance from testing phase.

The concrete values from equations in validation phase are not given in permit conditions proposals ([Proposal 1](#), [Proposal 2](#)) and it is left to installation operators to take right choice not risking return to training phase.

Elimination of non-sensitive input during the learning is proposed according to the **Table 1**. Only material balance models and linear regression have been proposed in proposals and no other balance or non-linear regression models have yet been considered.

Models with indicative surrogate parameters are also learned together with predictive parameters models ([Proposal 1](#)).

3.2. Statistical learning with artificial neural networks (ANNs)

Artificial neural network models (ANN) have been already proposed for the emission monitoring in [Metroalfa \(2022\)](#) and [Ekoneg \(2021\)](#). Proposed models have only two input (surrogate) parameters and two nodes with activation functions and its potential for the more surrogate parameters was not then according to that fact, fully utilized. The surrogate parameters are flow of natural gas and oxygen content (vol.%) in flue gases, which surrogates are already monitored in process.

Various types of modelling nonlinear functions ([Mesellem et al. 2021](#)) are available. For ANNs basically same as in [Figure 1](#), the sigmoid activation functions were used. That model, proposed for monitoring emission from hot water boiler ([Metroalfa 2022](#)) is:

$$\hat{y} = (y_{max} - y_{min}) \left\{ 1 + e^{-\left(\left[1 + e^{-(x_{N,1}w_{1,1} + x_{N,2}w_{2,1} + b_1)} \right]^{-1} \cdot w_1 + \left[1 + e^{-(x_{N,1}w_{1,2} + x_{N,2}w_{2,2} + b_2)} \right]^{-1} \cdot w_2 + b_3 \right)} \right\}^{-1} \quad (9)$$

where: y_{max}, y_{min} maximum and minimum real value of output, respectively, $x_{N,1}, x_{N,2}$ normalized input (surrogate) parameter of j th input ($j=1,2$), $w_{1,1}, w_{2,2}, w_{1,2}, w_{2,1}$ weight coef., b_1, b_2, b_3 biases given as in [Figure 2](#).

The normalization expressions for surrogate parameters used in model [Equation 9](#) is:

$$x_{N,j} = -C1(constant) + \frac{C2 \cdot x_j}{x_{max} - x_{min}} \quad (10)$$

where $C1$ are $C2$ are constants determined by process reasons.

The structure of proposed ANN ([Proposal 3](#), [Proposal 4](#)) with appropriate weights (w) and bias coefficients (b) is shown in [Figure 2](#).

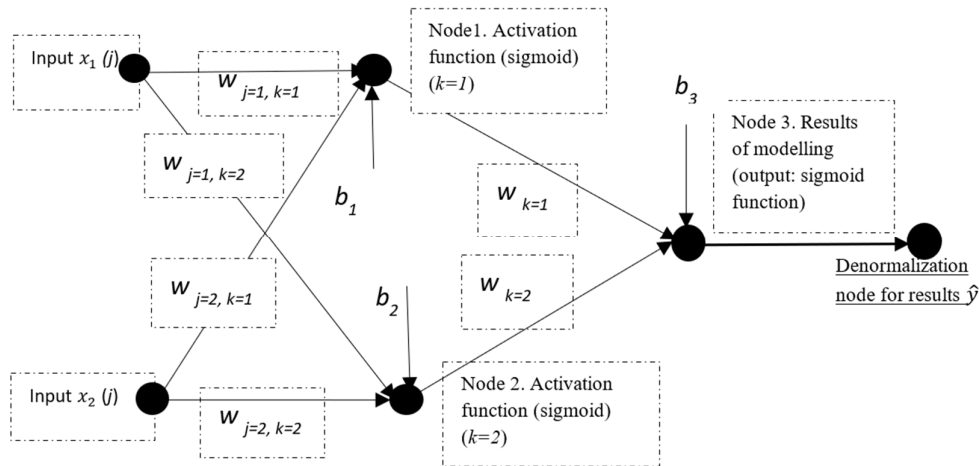


Figure 2. The structure of simple model of artificial neural network with one active (hidden) layer for monitoring emissions of NO_x and CO (for two inputs) used in [Metroalfa \(2022\)](#) and [Ekoneg \(2021\)](#)

Equation 9 could be applied on any “one hidden layer” model by extending the rows and columns of weight matrices for more nodes. To avoid more complex non-linear model outputs which could affect variance during testing phase, model (**Equation 9**) has been changed instead as:

$$\hat{y} = (y_{max} - y_{min}) \left(\sum_k \left[1 + e^{-\left(\sum_j x_{N,j} \cdot w_{j,k} + b_k \right)} \right]^{-1} \cdot w_k \right) + b_f \quad (11)$$

These changes also allow more input nodes and more active nodes for one hidden layer model.

Statistical learning for artificial neural networks for permit conditions proposals is given in **Table 2**, with learning functions as **Equations 12**, for training phase as **Equations 13** and **14**, for validation as **Equations 15** and reference to testing **Equation 4**.

Table 2. Learning with artificial neural network (ANN) for NOx and CO emission for models of **Equation 11** type

LEARNING PHASES	STATISTICAL LEARNING FUNCTIONS
<p>Training of ANN, Equations 13:</p> <p>Gradients of learning functions by coefficients w and biases b:</p> $\frac{\partial L}{\partial w}(w, b), \frac{\partial L}{\partial b}(w, b)$ <p>when (MSE) is applied, the first derivatives of L (obtained by chain rule) are included in the equation for corrected biases (b) and weight coefficients (w).</p> <p>Training using derivatives of loss function, Equations 14:</p> $w_{j,k}^{t+1} = w_{j,k}^t + \frac{2}{M} \cdot \eta \cdot (y_{max} - y_{min}) \sum_{i=1}^M (y_i - \hat{y}_i) \cdot w_k^t \cdot \sigma^t(z_{i,k}) \left(1 - \sigma^t(z_{i,k}) \right) \cdot x_{N,i,j}$ $w_k^{t+1} = w_k^t + \frac{2}{M} \cdot \eta \cdot (y_{max} - y_{min}) \sum_{i=1}^M (y_i - \hat{y}_i) \cdot \sigma^t(z_{i,k})$ $b_k^{t+1} = b_k^t + \frac{2}{M} \cdot \eta \cdot (y_{max} - y_{min}) \cdot \sum_{i=1}^M (y_i - \hat{y}_i) \cdot w_k^t \cdot \sigma^t(z_{i,k}) \left(1 - \sigma^t(z_{i,k}) \right)$ $b_f^{t+1} = b_f^t + \frac{2}{M} \cdot \eta \cdot (y_{max} - y_{min}) \sum_i (y_i - \hat{y}_i)$ <p>Validation phase of learning: Checking of importance of individual surrogate parameters of models and predictability, ensuring that calculation has support in the measured values of outputs. For two parameter models, as that in Equation 9, sensitive analysis is not fully adequate. It is given in the proposals for the case of more inputs. Data pairs for validation phase must be supported by measured outputs of y_i through prediction error giving criteria as well as of non-Equation 8. Then it goes to the validation Equation 15.</p> <p>Testing; Testing and periodical testing is by Equation 4, with degrees of freedom for testing.</p>	<p>Learning functions for training (loss function), equations 12:</p> $L(w, b) = \frac{1}{M} \cdot \sum_{i=1}^M [y_i - \hat{y}_i]^2, \text{ as mean square error (MSE).}$ <p>Other possible candidates for loss function: correlation coefficient (R), root mean square error (RMSE), mean absolute error (MAE), standard predicting error (ESP), error of prediction error model (EPM).</p> <p>Functions for validation phase of learning, Equation 15:</p> <p>The relative importance of the j_{th} input parameter (expressed through weight coefficients of artificial neural network) (Benyekhlef et al. 2021):</p> $I_{impj} = \frac{\sum_{k=1}^{N_k} \left(\frac{ w_{jk} }{\sum_{j,k=1}^{N_k} w_{jk} } w_k \right)}{\sum_{j=1}^{N_k} \left(\sum_{k=1}^{N_k} \left(\frac{ w_{jk} }{\sum_{j,k=1}^{N_k} w_{jk} } \right) w_k \right)} \quad (15)$ <p>The predictability of model after training phase, is also characterized by a cross - validated correlation coefficient Q_{ext}^2, and could be assessed as:</p> $Q_{ext}^2 = 1 - \frac{\sum_{i=1}^{train} (y_i^{train} - \hat{y}_{train,i})^2}{\sum_{i=1}^{train} (y_i^{train} - \hat{y}_{preced. train,i})^2} \quad (16)$ <p>as this is satisfied when $Q_{ext}^2 \geq 0.9$. (It is not actually a part of learning not been supported by the validation observations)</p>

The symbols used in **Table 2** are: L loss (error) function, k number of neuron in active (hidden) layer, w ($w_k, w_{j,k}$) weight coefficient, b (b_k, b_f) biases, $w_{j,k}^t, w_{j,k}^{t+1}$ weight coefficient of activation layer for j input and k neuron in step t and $t+1$ respectively (by analogy the same goes for w_k^t, w_k^{t+1} and for b_k^t, b_k^{t+1}), $x_{N,i,j}$ normalized input (surrogate) parameter j of i th data pair calculation, $\sigma^t(z_{i,k})$ calculated sigmoid function for k neuron and i th data pair, $train$, $train$ training phase of learning, $preced. train$ preceding training, η learning rate, M number of observations (data pairs), N_k number of neurons in active (hidden) layer.

The recommended learning rate (η) for one or two - layered networks, without error propagation correction system, is 0.2 and 0.3 but ≤ 0.5 (Sadadou et al. 2021). Initial trainings of ANNs were done in Metroalfa (2022) and Ekonerg (2021), the results being in same time part of request of operators for permit conditions. Number of training steps in Metroalfa (2022) was 5 pairs of calculated values - direct measurement. Learning rate (η) used was between 0,20 and 0,25. Number of training steps in Ekonerg (2021) was even less, 3 and adequately learning rate was 0,33.

Important to note is that number of training steps in studies was less than those recommended by proposals for permit conditions according to the **Figure 1**. The achieved uncertainties in both studies and not calculated as variances are low, but they have been estimated from the same set of data used for training what put them in question. Anyway, the proposals for permit conditions (**Proposal 3**, **Proposal 4**) qualify them as QUAL 1 comparable requirement which require further learning (training, validation, testing) to complete learning.

Blocking strategy proposed for ANN in learning is the same as in the **Figure 1**. Elimination of the non-sensitive input during the learning according to the **Table 1** for the mass balance and regression models, is considered also according to the **Table 2** but based on coefficients set (W). The validation for only two parameters could be also switch to testing instead but it is not discussed here.

From the examples could be concluded that shallow learning (Frana et al. 2024) is used for both proposals. Accordingly, they are still in competence of algorithmic machine learning.

Deep learning formula could be used for more hidden layers. It is based on general formula for any number of h layers, with the sign $-$ denoting the tendency of optimization (decreasing) for loss function (L gradient):

$$W_h^{t+1} = W_h^t + \left(-\frac{\partial L}{\partial W_h}\right) \quad (16)$$

where: W_h^{t+1} , W_h^t , matrices of weight coefficients in h layer of neurons (applicable for artificial neural networks with more hidden layers) (Ananthaswamy 2024). The back propagation of error through more hidden layers influencing learning should be met with adequate equations (Kelleher JD 2021).

3.3. Confirmation of learning and results of monitoring by models

Confirmation or final validation of learning is by variances (s^2) calculated for population of model results that must satisfy the type of conditions given by **Equation 4** or **4a** for models. This is naturally after the testing phase of learning. Also, the periodical testing is introduced according to requirements on learning.

For artificial neural networks models estimation of variances (s^2) for validation and testing is the same as for balance and regression models.

The permits containing learning conditions should be issued before allowed work according to permit, directing the way how learning is applied as a part of monitoring conditions. As a consequence, the learning should be provided and confirmed on-site and that is away how permits should actually function. Amendments of directive (**Directive 2024/1785/EU**) require all permit conditions to be a part of environmental management systems (EMS) with a special focus on monitoring including then the learning process.

The existing legislation (Directive and BAT conclusions) doesn't recognize distinction between surrogate and direct emission monitoring results validation, despite the fact that strictly allows application of surrogates for the continuous monitoring. The results of monitoring should be therefore expressed and validated in the way the standard direct continuous monitoring of emissions (**Equation 5**) as required in existing legislation.

4. CONCLUSIONS

Statistical /machine learning should be a part of modelling for the surrogate parameters emission monitoring and mandatory for those installations requiring permitting.

It should be noted that the motivation for discussion on the surrogate parameters monitoring is primarily not coming from actual industrial practice or their wider application (despite examples referred in this work), but from administrative possibilities to introduce such monitoring. Administrative considerations/obstacles have to be solved. Statistical learning is the most important among administrative priorities, because there is no piece of environmental legislation or standard in place that regulates it. Experience of the author shows, that only after solving the priorities among administrative options, the industry would accept surrogate parameters as approach to monitoring emissions and could benefit of it.

Following already existing documents for approving surrogate parameters monitoring in mineral wool, glass and aluminate cement production and considering legal possibilities of such monitoring for waste incineration and co-incineration, statistical learning procedures and rules have been developed for now actual and in future foreseeable cases. As such, they are ready for administrative regulation and for permitting. These procedures should be ready for easy changes of permit conditions and for more detailed or stricter conditions comparing to prior if necessary. That should be the rule for any permitting process by the very nature of it.

It is important to put the quality requirements on model learning, comparable to standards for direct measurement known as QUALs. It should start with the input relations for learning. ISO standards requirements

for the assurance quality of automated measuring systems are also responsible for the determination of number of observations in testing especially they are strongly positioned by the administrative authorities for air protection. They are not suitable for training and validations phases of learning because of the substantial difference in statistical techniques used. But they additionally strengthen the learning through testing.

Blocking of the observations is another set of the statistical techniques considered to minimise variances through learning together with the other statistical and modelling techniques and on importance for training. This is also a new quality assurance of monitoring not existing yet for the direct monitoring of emissions.

To put learning in permit conditions for the balance and regression models according to the **Figure 1**, in the **Table 1** are given learning functions (loss functions) as **Equations 6**, for the training as **Equations 7**, for criteria for validation and the validation as non-**Equation 8** and **Equations 9** respectively and reference to testing the models, **Equation 4a**.

For learning using ANN models, the learning functions are given in the **Table 2** as **Equations 12**, for the training as **Equations 13** and **14**, for the validation as **Equations 15**, and reference to **Equations 4** for the testing.

The learning with ANNs is still shallow learning. Shallow artificial neural networks represented by **Equation 11** and relatively simple, could be advantageous over the deep ANNs because of need for prior setting the basic technical issues and administrative items for statistical learning.

5. REFERENCES

- Ananthaswamy A (2024) Why machines learn, the elegant maths behind modern AI. Penguin Books. Dublin.
- B4 Control Solutions & Faculty of Chemical Engineering (FKIT) (2022) Derivation and validation of mathematical model for surrogate parameters monitoring for emission from flue gases of factory Vetropack Straža. Karlovac. (in Croatian).
- BAT Conclusions for waste incineration, Commission Implementing Decision (EU) 2019/2010 of 12 November 2019. Available on <https://BAT-reference-documents|EU-BRITE>. Cited 25 Nov 2025.
- Benyekhlief A et al. (2021) Kem. Ind. 70 (11–12), 639–650.
- Brinkmann T et al. (2018) JRC Reference report on monitoring of emissions to air and water from IED installations. Industrial Emissions Directive 2010/75/EU, EUR 29261. Available at: <https://eippcb.jrc.ec.europa.eu/reference>.
- Dey A (2010) Incomplete block designs. Hindustan Book Agency – World Scientific. Singapore.
- Directive 2010/75/EU of the European Parliament and of the Council of 24 November 2010 on industrial emissions (integrated pollution prevention and control). Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32010L0075>. Cited 25 Nov 2025.
- Directive 2024/1785/EU of the European Parliament and of the Council of 24 April 2024 amending Directive 2010/75/EU and Directive 1999/31/EC. Available at: <https://data.europa.eu/eli/dir/2024/1785/oj>. Cited 25 Nov 2025.
- Ekoneg d.o.o. (2021) Study on the establishment of surrogate parameters for monitoring emissions into air from hot water boiler K3, DS Smith Belišće Croatia d.o.o. Zagreb. (in Croatian).
- Environmental permit Class UP/I 351-02/20-45/05, No. 517-05-1-3-1-22-34 (2022). Permit conditions for producing aluminate cement in installation Calucem d.o.o., Ministry of Environmental Protection (MEPGT). Zagreb. Available at: <https://mzozt.gov.hr/>. Cited 25 Nov 2025. (in Croatian).
- Environmental permit Class UP/I 351-02/20-45/13, No. 517-05/20-45/03 (2021). Permit conditions for producing mineral wool installation Knauf Insulation, Ministry of Environmental Protection (MEPGT). Zagreb. Available at: <https://mzozt.gov.hr/>. Cited 25 Nov 2025. (in Croatian).
- Frana PL, Klein MJ (2024) Encyclopaedia of artificial intelligence – The past, present and future of AI. Bloomsbury Academic. New York.
- Gomzi Z, Kurtanjek Ž (2019) Modelling in chemical engineering – University textbook. HDKI/FKIT. Zagreb. (in Croatian).
- HRN EN 14181 (2014) Emissions from stationary systems – Assurance of quality of automated measurement systems. Croatian Standards Institute. Zagreb.
- Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer. New York.
- James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning with applications in R. Springer. New York.
- Jiji A (2003) Design of experiments for engineers and scientists. Butterworth-Heinemann. Amsterdam–Tokyo.
- Kelleher JD (2021) Deep learning. Mate d.o.o. Zagreb. (in Croatian).
- Mendenhall W, Sincich T (1988) Statistics for the engineering and computer sciences. Dellen Publishing Company. San Francisco.
- Mesellem Y et al. (2021) Kem. Ind. 70 (1–2), 1–12.
- Metroalfa d.o.o. (2022) Study on the establishment of surrogate parameters for monitoring emissions into air from hot water boiler VKLM-50, Gradska toplana d.o.o. Karlovac. Zagreb. (in Croatian).
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2002) Numerical recipes in C++ – The art of scientific computing, 2nd ed. Cambridge University Press. Cambridge.

Proposal 1 (2025) Permit conditions for glass factory Vetropack Straža – Annex with proposal for training, testing and validation of models in accordance with continuous measurement requirements (AMS). Class 351-03/25-01/453, No. 1. Submitted to Ministry of Environmental Protection (MEPGT), February 2025. Zagreb. (in Croatian).

Proposal 2 (2025) Change of permit conditions for Knauf Insulation and Calucem d.o.o. – Annex with proposals for training, testing and validation of models (AMS). Class 351-03/25-01/453, No. 1. Submitted February 2025. Zagreb. (in Croatian).

Proposal 3 (2025) Permit conditions for thermal plant Gradska toplana d.o.o. Karlovac – Annex with proposal for training, testing and validation of models (AMS). Class 351-03/25-01/453, No. 1. Submitted February 2025. Zagreb. (in Croatian).

Proposal 4 (2025) Permit conditions for DS Smith Belišće Croatia d.o.o. – Annex with proposal for training, testing and validation of models (AMS). Class 351-03/25-01/453, No. 1. Submitted February 2025. Zagreb. (in Croatian).

Rumenjak D (2023) Air emission monitoring from waste incineration/co-incineration installations using surrogate parameters. *Environmental Engineering* 10 (1–2), 24–29.

Sadadou A et al. (2021) *Kem. Ind.* 70 (5–6), 233–242.

Smith J, Smith P (2007) *Environmental modelling – An introduction*. Oxford University Press. New York.