Computers and Chemical Engineering 135 (2020) 106733

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Two complementary methods for the computational modeling of cleaning processes in food industry

Hannes Deponte[a,*], Alberto Tonda[b], Nathalie Gottschalk[a], Laurent Bouvier[c], Guillaume Delaplace[c], Wolfgang Augustin[a], Stephan Scholl[a]

[a] Technische Universität Braunschweig, Institute for Chemical and Thermal Process Engineering, Langer Kamp 7, Braunschweig 38106, Germany
[b] UMR GMPA, INRA, Université Paris-Saclay, 1 av. Brétignières, 78850 Thireval-Grignon, France
[c] CNRS, INRA, ENSCL, UMR 8207, UMET, Unité Matériaux et Transformations, University Lille, Lille F 59 000, France

ARTICLE INFO

Article history:
Received 28 August 2019
Revised 6 January 2020
Accepted 12 January 2020
Available online 15 January 2020

Keywords:
Statistical analysis
Symbolic regression
Dimensional analysis
Machine learning
Cleaning process
Food industry

ABSTRACT

Insufficient cleaning in the food industry can create serious hygienic risks. However, when attempting to avoid these risks, food-processing plants frequently tend to clean for too long, at extremely high temperatures, or with too many chemicals, resulting in high cleaning costs and severe environmental impacts. Therefore, the optimization of cleaning processes in the food industry has significant economic and ecological potential. Unfortunately, *in-situ* assessments of cleaning processes are difficult, and the multitude of different cleaning situations complicates the definition of a comprehensive approach.

In this study, two methodological approaches for the comprehensive modeling of cleaning processes are introduced. The resulting models facilitate comparisons of different cleaning processes and they can be scaled up for processes with similar conditions, using cleaning time as a response. A dimensional analysis is performed to obtain general results and to allow transfer of the approaches to other cleaning situations. The models are established according to the statistical rules for the deduction of multiple regression equations for the prediction of the response based on the input parameters. The terms of the model equation are confirmed with a significance analysis. A machine learning approach is also used to create model equations with symbolic regression. Both methods and the obtained model equations are validated.

The two applied approaches reveal similar significant terms and models. Significant dimensionless numbers are the Reynolds number, the density number that describes the ratio of the density of the soil to the density of the cleaning agent, and the soil number, which is a new dimensionless number that characterizes the properties of food soils. The methodology of both approaches is transparent; therefore, the resulting equations can be compared and similarities are found. Both methods are deemed applicable for the computational modeling of cleaning processes in food industry.

1. Introduction

1.1. Dimensional analysis in cleaning processes

This paper introduces two methodological approaches for the comprehensive modeling of data in the field of process engineering: statistical analysis and symbolic regression. The underlying experiments provide a database for cleaning processes from the food industry with different food soils. These soils are chemically and physically complex, which makes the cleaning situation challenging and any predictions on cleaning success and effectiveness quite difficult (Fryer et al., 2006).

Insufficient cleaning in food industry causes serious hygienic risks, affects the efficiency of the plant, and may influence the quality of the final product (Palabiyik et al., 2014). Cleaning procedures in food processing plants often involve the use of cleaning-in-place (CIP) technology to clean the product-contacting surfaces, such as heat exchangers, tanks, pumps, valves, and pipework (Goode et al., 2013). For this reason, these plants frequently apply automated cleaning protocols for these CIP processes to achieve the hygienic requirements of the food industry (Tamime, 2009). These food processing line are cleaned for a longer time as required as cleanliness of these closed surface is still nowadays difficult to assess by using a non-intrusive sensor for monitoring fouling state (Chen et al., 2019). Consequently, the design of these

[*] Corresponding author.
E-mail address: h.deponte@tu-braunschweig.de (H. Deponte).

https://doi.org/10.1016/j.compchemeng.2020.106733
0098-1354/© 2020 Elsevier Ltd. All rights reserved.

Nomenclature

A_i	dimensional matrix of a factor [-]
A_0	dimensional matrix of the response [-]
c	concentration of cleaning agent [kg·m^{-3}]
d	inner diameter of the channel [m]
f	model equation [-]
i,j	number of parameter [-]
k	number of physical dimensions of a process [-]
l_i	exponent for the dimension length [-]
L	dimension length [-]
m_i	exponent for the dimension mass [-]
M	dimension mass [-]
n	overall number of variables of a process [-]
p	number of independent parameters [-]
Q2	predictive power [-]
R	reproducibility [-]
R2	coefficient of determination [-]
Re	reynolds number [-]
t_{95}	cleaning time [s]
t_i	exponent for the dimension time [-]
T	dimension time [-]
T_{ca}	temperature of cleaning agent [K]
V	validity [-]
w	flow velocity [m·s^{-1}]
x_i	exponent for a physical quantity [-]
y	response of a model [-]
γ_{soil}	free surface energy of soil [N·m^{-1}]
α	factor [-]
β	regression coefficient [-]
δ_0	layer thickness [m]
η	dynamic viscosity of cleaning agent [kg·m^{-1}·s^{-1}]
η_{soil}	dynamic viscosity of soil [kg·m^{-1}·s^{-1}]
Π_i	independent parameters [-]
ρ	density of cleaning agent [kg·m^{-3}]
ρ_{soil}	density of soil [kg·m^{-3}]

protocols requires knowledge of the parameters affecting the cleaning process, so different approaches have been developed to classify the influencing parameters. For example, the well-known Sinner Circle assigns each factor to one of six categories: type and amount of soil, technical characteristics of the substrate, duration of the process, chemistry of the cleaning agent, mechanical parameters of the processing plant, and processing temperature (Lelieveld et al., 2005). Another common approach is to use a cleaning map to classify the cleaning problems based on the soil complexity and the type of cleaning chemical used (Fryer and Asteriadou, 2009). Both approaches are qualitative and offer no quantitative assessment for a given cleaning procedure.

Individual models for special cleaning situations can enable a more accurate prediction. Different developed models have been presented in the literature by Asteriadou et al. (2006), Dürr and Graßhoff (1999), and Jensen and Friis (2005), for example. However, the transfer of developed models to different cleaning processes is only possible to a certain extent because of the differences between each cleaning situation (Schöler et al., 2012). Nevertheless, the proposal has been made that a comprehensive and transferable consideration of cleaning parameters could be established with the use of dimensionless numbers (Delaplace et al., 2015).

Dimensionless numbers are obtained from the datasets of an observed process and are then used as input variables to affect the response. Dimensional analysis, when used in chemical engineering to develop dimensionally consistent models (Stichlmair, 2001; Woods et al., 2017), allows the comparison of physical phenomena and provides the possibility of scaling models for similar process conditions (Albrecht et al., 2013; Szirtes and Rózsa, 2007). Dimensionless numbers can also decrease the complexity of a model because they reduce the physical quantities of the coefficients and decrease a problem's degrees of freedom to the minimum (Sonin, 2004). Shen et al. (2014) describe the general properties of dimensional analysis and its application in the design of experiments.

The aim of the present study is therefore to develop statistical models based on dimensionless numbers for CIP processes. The fundamental rules for the construction of the set of dimensionless numbers describing food processes, established in (Delaplace et al., 2015), are used:

(1) Choosing the target variable, creating the relevance list.
(2) Determining the dimensions of the physical quantities in SI-units.
(3) Applying the Buckingham Π theorem.
(4) Constructing the dimensionless numbers.
(5) Rearranging the dimensionless numbers.

Dimensional analysis starts with the definition of the target variable, also known as the response of the system. For cleaning experiments, this can be the required cleaning time. All dimensional influencing parameters for the response are then identified, along with their dimensions, and are compiled in a relevance list. Application of the Buckingham Π theorem then derives the dimensionless numbers responsible of the evolution of the system for the given problem (Buckingham, 1914). Well known target dimensionless numbers encountered in cleaning processes could be the dimensionless particle diameter (Ziskind et al., 1995), the dimensionless cleaning time (Cole et al., 2010), and the dimensionless remaining mass of soil (Dürr and Graßhoff, 1999).

1.2. Modeling of cleaning processes

In this study, the cleaning processes in the food industry are modeled using two computational methods. The first approach is a statistical analysis, where the process is considered as a black box, with the influencing parameters as input factors and target variables as the response. The process is further influenced by noise and error, and the effect of a factor is deemed significant if it is greater than the noise (Babutzka et al., 2019). The result of this consideration is a multiple regression equation for the prediction of the value of the response based on the independent input parameters. The significance and influence of the model terms on the response is assessed by testing the null hypothesis, which states that no effect or difference exists or that no specific relationship is shared between a factor and the response. Therefore, the model terms are statistically significant if they are not likely to confirm the null hypothesis; the calculated probability value (p-value) estimating this likelihood has to be smaller than a defined threshold value, often defined as $p < 0.05$ (Siebertz et al., 2017). In practice, this guarantees that the influence of a factor on the response exceeds the noise. The model is validated via cross-validation to assess its predictive power. In this way, a high validity indicates good control over the experiment (Dennison et al., 2016).

The second method is a machine learning approach called *symbolic regression*. With this kind of evolutionary computation, computers are used to develop predictive models from a large database. The computers are not explicitly programmed for this task, but they can learn from the data (Shahriari et al., 2016). The symbolic regression algorithm, first described in 1992 by Koza (1992), has been implemented by different software tools (Schmidt and Lipson, 2009). Symbolic regression performs a stochastic exploration of the space of all possible equations, with

Table 1
Relevance list of parameters for cleaning processes arranged in the categories of the Sinner Circle.

Category	Parameter
Time	Cleaning time t_{95}
Mechanic	Pipe geometry d, flow velocity w
Temperature	Temperature of cleaning agent T_{ca}
Chemistry	Concentration of cleaning agent c, density of cleaning agent ρ, dynamic viscosity of cleaning agent η
Soil	Layer thickness δ_0, free surface energy of soil γ_{soil}, viscosity of soil η_{soil}, density of soil ρ_{soil}

bias toward candidate equations with good fit. The process begins with the generation of a set of equations, internally encoded as binary trees. Each equation is then evaluated on its error with respect to the available data. Equations with lower error are more likely to be selected to produce new equations by randomly changing some of their components. New equations are evaluated in turn, and the process iterates until a user-defined condition (for example, an error below a given threshold) is reached. Modern symbolic regression tools also take into account the size of the equations, because equations with greater complexity can more easily fit the available data, whereas simpler equations are less likely to overfit (Babutzka et al., 2019).

2. Dimensional analysis of food processes

2.1. Database of cleaning experiments

The underlying data for the modeling of CIP processes in the food industry was taken from Helbig et al. (2019), who conducted cleaning experiments in a flow channel. Overall, 88 experiments were performed, following a D-optimal design matrix with 3 replicates at the center point. A detailed description of the test rig is given by Schöler et al. (2012). The response of the system is the cleaning time, defined as the time at which 95% of the initial soiled surface is cleaned. Three different soils (starch, gelatin, and egg yolk) were assessed, and the cleaning time was measured with an optical analysis that determined the time-dependent surface coverage with soil (Gordon et al., 2014). In industrial cleaning plants the cleaning time is also determined qualitatively, even if other indirect methods such as turbidity or conductivity measurements are used there. The advantage of the test facility used here is, that the optical accessibility of the test plates can be used to apply a direct method specially developed for the determination of the declining surface coverage over time. A qualitative measurement of the cleaning time is sufficient, because the final decision criterion for the plant operator is whether a surface is clean or not.

The input variables for the cleaning experiments are the process parameters: temperature (*23°C* to *60°C*), flow velocity (*0.24* m/s to *2.12* m/s), and concentration of the cleaning agent, which was sodium hydroxide (*0* kg/m³ to *20* kg/m³).

The categories in the Sinner Circle indicate that the soil parameters also influence the cleaning process. Therefore, the temperature-dependent properties of the three test soils were measured in the range of the investigated process conditions (Gottschalk et al., 2019). The parameters were then sorted by their Sinner Circle categories and compiled into a relevance list, as reported in Table 1. The sixth category of the Sinner Circle, the technical characteristics of the substrate, is not represented by any factor, since previous evaluations of these experiments have shown that the substrate properties have for our operating conditions no significant effect on the cleaning time (Deponte et al., 2018).

This study is complemented by the dimensional analysis of the influencing factors from the relevance list. Since the temperature is the only parameter with the physical quantity *K*, it is not directly considered in the dimensional analysis. Therefore, the remaining units *kg, m* and *s* are all included with the mass M, the length L and the time T in the (M, L, T)-dimension system (Szirtes and Rózsa, 2007). Note that the influence of the temperature is not neglected in this approach: it is included in the temperature-dependent parameters, such as density and viscosity, which are considered with their corresponding temperature values.

2.2. Implementation of the Buckingham Π theorem

The Buckingham Π theorem, mentioned in Step 3 of the fundamental rules for the construction of dimensionless numbers, is applied to obtain the number p of dimensionless numbers Π_i, for a process with n variables and k physical dimensions (Buckingham, 1914).

$$p = n - k \tag{1}$$

With $n = 9$ for the relevant variables given in Table 1 and $k = 3$ for the dimensions from the given (M, L, T)-dimension system, the cleaning process is characterized by $p = 6$ dimensionless numbers. These dimensionless numbers are derived by compiling the dimensional matrix with all influencing factors in the columns and the dimensions in the rows. The 9 influencing factors and the 3 occurring dimensions result in a 3 × 9 dimensional matrix. The dimensionless numbers are generated by solving the equations due to transformation of the dimensional matrix. If necessary, the matrix, and thus the resulting dimensionless numbers, can be rearranged to be more applicable to the given cleaning process.

2.3. Computational dimensional analysis

In addition to the conventional approach, a more pragmatic method for dimensional analysis is used in this study. This approach was first introduced by Deponte et al. (2018) and is further improved here. The dimensional analysis is conducted with Matlab® (Version R2016a 9.0.0.341360, MathWorks, 2016). The identified physical quantities of all parameters from the relevance list are converted into their associated SI base unit from the (M, L, T)-system, as *kg, m* and *s,* and are placed in the dimensional matrix A_i with their corresponding exponents m_i, l_i and t_i.

$$A_i = M^{m_i} \cdot L^{l_i} \cdot T^{t_i} \tag{2}$$

For example, the matrix $A_{velocity}$ for the flow velocity w in m/s is defined as:

$$[w] = A_{velocity} = M^0 \cdot L^1 \cdot T^{-1} = L^1 \cdot T^{-1} \tag{3}$$

For the mathematical construction of the dimension of the response A$_0$, the parameters i to n are exponentiated with an integer value x_i and multiplied.

$$A_0 = \prod_i^n \left(M^{m_i} \cdot L^{l_i} \cdot T^{t_i} \right)^{x_i} \tag{4}$$

If the exponents of the dimensions are

$$m_i = l_i = t_i = 0 \tag{5}$$

then the dimension of the response A$_0$ is dimensionless, which mathematically results in

$$A_0 = 1 \tag{6}$$

Thus, the response is a dimensionless number.

For the computational dimensional analysis, Eq. (4) is split into the individual expressions for each dimension. For a dimensionless number, the sum of the exponents for each dimension should be equal to zero. This results in the elimination of the physical quantity.

$$\sum_{i}^{n} m_i \cdot x_i = 0 \tag{7}$$

$$\sum_{i}^{n} l_i \cdot x_i = 0 \tag{8}$$

$$\sum_{i}^{n} t_i \cdot x_i = 0 \tag{9}$$

Based on these three equations, the developed Matlab® script is testing different exponents x_i in the range of 3 to -3 on the input parameters from the relevance list with a for-loop. If the sum of the exponents for each resulting dimension (mass, length and time) is equal to zero, the input parameters are multiplied, taking into account their exponents. Thus, all parameters of the relevance list for cleaning processes appears in a fraction. The calculated number is dimensionless and is saved in the dimensional matrix. Multiples and reciprocals of already existing dimensionless numbers occurring in the matrix are automatically deleted from the list. Furthermore, already known dimensionless numbers, like the Reynolds number, are found in combination with other terms. The known numbers can be reduced from the results, so the remaining set of dimensionless terms potentially defines a new dimensionless correlation that characterizes the process. Their effect on the response is assessed by performing the significance analysis.

3. Computational modeling

3.1. Significance analysis

The dimensionless numbers derived from the computational dimensional analysis are selected with a significance analysis, performed with the statistical design-of-experiment software MODDE® Pro (Version 12.0.13948, Sartorius Stedim Data Analytics AB, 2017). The potential dimensionless numbers are defined as input factors for the response, and the multiple regression equation is formed. Each factor is considered with its linear and quadratic terms, and the interaction terms of two different factors are also taken into account. The regression equation including these terms has to be analyzed to create a mathematical expression, which is then used as a model equation for the given process. Therefore, the response y is calculated with the sum of the products of the model terms from the multiple regression equations with their corresponding regression coefficients β_i. The model terms for the influencing factors are linear a_i, quadratic $a_i{}^2$, or interactions of two factors $a_i a_j$.

$$y = \sum_{i}^{n} \beta_i a_i + \sum_{i}^{n} \beta_{ii} a_i{}^2 + \sum_{i}^{n} \sum_{j}^{n} \beta_{ij} a_i a_j \tag{11}$$

In some cases, the response requires adjustment due to transformation to achieve a normal distribution ("bell shaped") of the response distribution. If many values of the response are comparably small, the response distribution shows a positive skewness. In that case, a logarithmic transformation of the response is needed. The developed model therefore predicts the transformed response. If a function for the transformation is applied, the mathematical inverse function must then be applied on the model to achieve the values for y.

The model is further improved by analyzing the regression coefficients of the terms and their confidence intervals to verify whether they are significant. The significance is assessed by varying each factor from its minimum to its maximum, while keeping the others at their average values. A factor is assessed as significant if the regression coefficient deviates from *0* and the confidence interval does not cross zero. This is interpreted as an influence that exceeds the noise of the measured response. The calculated p-values are $p < 0.05$ for the significant factors. The coefficients of the terms are scaled and centered; therefore, their effect on the response can be compared. The value of each regression coefficient quantifies the influences on the response, while a negative deviation expresses the reduction of response and increases for positive coefficients.

The achieved model is assessed with four statistic model parameters. The fit of a model to the investigated variables is evaluated by the coefficient of determination R2. The predictive power Q2 defines the capability of the model to predict the response. The mathematical basis of the calculation of Q2 is the leave-one-out-cross validation. The lack-of-fit test compares the models' error with the pure error of the measurements. The pure error is calculated with the values of the replicates at the center point. The result is expressed in MODDE® Pro as the models' validity V. This is a software-specific parameter that represents a conversion of the p-value.

$$V = 1 + 0,57647 \cdot log_{10}(p) \tag{10}$$

A validity of 0.25 represents a p-value of $p = 0.05$, and a higher validity represents a lower probability that the null hypothesis applies. The last parameter is the reproducibility R of the model, which shows the variation of the measured values under the same conditions in comparison to the total variation of the response. If a good model for describing the cleaning process is developed with respect to all the quality measurements described before, the model equation can be extracted from the unscaled regression coefficients β_i of the significant factors.

The model equation is only valid in the observed design space and, thus, for the range of the given factors in the dataset. Predictions beyond the limits of the model are only possible to a limited extent, because the multiple regression equation only fits the data inside the design space. Physical phenomena outside the design space are not considered.

3.2. Symbolic regression

The symbolic regression approach is performed with the Eureqa Pro software (Version 1.24.0, Nutonian Inc., 2016). The dataset is inserted into the software tables and prepared for the modeling. The target expression, containing the response of the system and the parameters, has to be defined. This expression defines the parameters to be considered as factors for the model equation and the parameter that is the response of the system. Furthermore, the model building blocks for the generation of mathematical equations have to be selected. Next to the basic functions, like constants, addition, subtraction, or multiplication, the software can consider more complex terms, like trigonometric functions, logical operators, or exponential functions. The mathematical functions that can be potentially useful for the given problem should be defined before the modeling. The error metric is also selected; this specifies what type of error is calculated for the model and is used for error assessment. Available error choices include the mean squared error, mean absolute error, median absolute error, and similar metrics.

Following this definition phase, the search is initiated for model functions. The initial combination of parameters is formed randomly. Each equation is internally encoded with a binary tree. The evolutionary algorithm starts by generating random binary trees. While optimizing parts of the binary tree, Eureqa always preserves the best-fitting tree. The results are compiled in a list with the two parameters *size* and *fit*. The size represents the complexity of the model function. This parameter determines the number of constants, variables, and functions used within the model solution. The fit is calculated with *1-R2*. Thus, a good model should feature a small value for the fit and, in most cases, a small value for size. The selection of the best model is always a compromise between the minimum acceptable coefficient of determination R2 and a maximum complexity. These two parameters are plotted in an accuracy-versus-complexity diagram for localization and comparison of all the models. Each individual model solution can be analyzed in detail with different plots as the observed-versus-predicted plot, which compares the predicted response with the observed (i.e., measured) response from the dataset.

4. Results

4.1. Deduced characteristic dimensionless numbers

According to the Buckingham Π theorem, six dimensionless numbers were found. The dimensionless numbers obtained with the five fundamental rules (Delaplace et al., 2015) are:

$$\pi_1 = \frac{c}{\rho} \tag{12}$$

$$\pi_2 = \frac{\eta}{\rho \cdot w \cdot d} = \frac{1}{Re} \tag{13}$$

$$\pi_3 = \frac{\delta_0}{d} \tag{14}$$

$$\pi_4 = \frac{\gamma_{soil}}{\rho_{soil} \cdot w^2 \cdot d} \tag{15}$$

$$\pi_5 = \frac{\eta_{soil}}{\rho_{soil} \cdot w \cdot d} \tag{16}$$

$$\pi_6 = \frac{\rho_{soil}}{\rho} \tag{17}$$

Number Π_6 describes the ratio of the density of the soil divided by the density of the cleaning agent, so it is referred to as the density number. Number Π_2 is the inverse Reynolds number, found with the conventional approach. The significance analysis is performed with the six dimensionless numbers Π_1 - Π_6. All numbers are assessed as significant, except Π_4, the dimensionless number containing the free surface energy of the soil γ_{soil}. Since γ_{soil} appears only in this number, it is assumed that this parameter does not affect the cleaning time. Apart from the Reynolds number, numbers Π_3 and Π_5 predominantly affect the cleaning time. All these numbers contain the layer thickness *d*; therefore, this parameter is interpreted as the most significant for the prediction of cleaning time.

In parallel to this approach, the computational dimensional analysis presented in Section 2.3 is applied to find the dimensionless numbers. The computational dimensional analysis also identifies the Reynolds number as a dimensionless number describing the cleaning process. The significance analysis unveiled a high significance of the Reynolds number on the cleaning time ($p = 1.87 \cdot 10^{-10}$). Apart from the Reynolds number, four other dimensionless numbers were found by the computational dimensional analysis. While three of them are not significant for the cleaning time ($p > 0.05$), a soil-related number Π_{soil} showed a significant effect on the cleaning time. Its linear term is not significant ($p = 0.06$), but the influence of the quadratic term ($p = 2.30 \cdot 10^{-3}$) on the cleaning time exceeds even that of the Reynolds number.

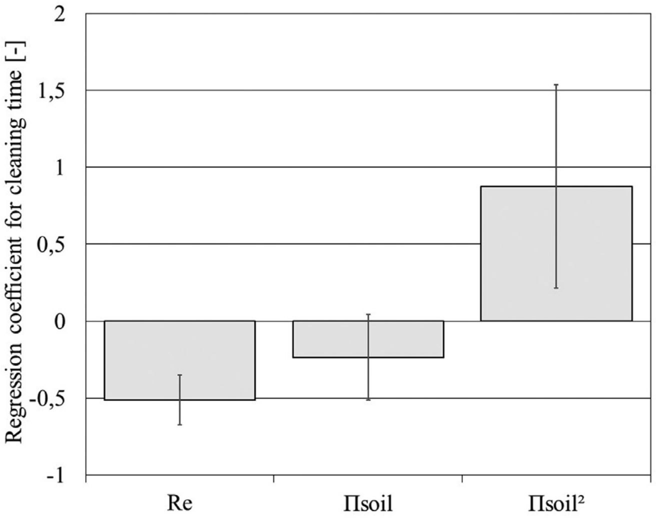

Fig. 1. Scaled and centered coefficient plot (columns) indicating the significance (error bar) of the model terms on the cleaning time.

The two dimensionless numbers are shown in Fig. 1, where the column for the regression coefficient for the significant terms exceeds the noise, represented by the error bar. The negative regression coefficient of the Reynolds number indicates that a higher Re reduces the cleaning time, a higher soil number is increasing it.

The soil number Π_{soil} contains the physical properties of the soil and the concentration of the cleaning agent, whose influence (e.g., swelling, dissolution) depends on the type of soil, as well as their derived exponents. Π_{soil} is expressed as:

$$\pi_{soil} = \frac{\delta_0 \cdot \gamma_{soil} \cdot \rho_{soil}^3}{c^2 \cdot \eta_{soil}^2} \tag{18}$$

The analysis of the soil-related number shows that it is a combination of all dimensionless numbers found with the five fundamental rules presented above, with the exception of the Reynolds number and Π_6:

$$\pi_{soil} = \frac{\pi_3 \cdot \pi_4}{\pi_1^2 \cdot \pi_5^2} \tag{19}$$

Except for the dimensionless number Π_6, the same influencing terms are found with both the fundamental rules and the computational dimensional analysis. Based on these findings, the significance of Π_6 on the cleaning time is also verified with the computational dimensional analysis. The result shows that the quadratic term of this dimensionless number is significant, albeit slightly less than the Reynolds number (see Fig. 2). The consideration of the density number increases the significance of the soil number because the effects of the density number were previously interpreted as the noise of the system and only the soil number described differences in response. Considering the density number as well gives a more accurate prediction of the response. Now, the linear term of Π_{soil} also affects the cleaning time.

The statistical significance of the model terms is proven by calculating the probability values. A p-value of $p < 0.05$ represents a significant model term. All model terms are statistically significant, except for the linear term of Π_6 (see Table 2).

Since the experimental design follows a D-optimal design matrix, but the model uses the dimensionless numbers as factors, their correlation is checked in a correlation matrix (see Table 3).

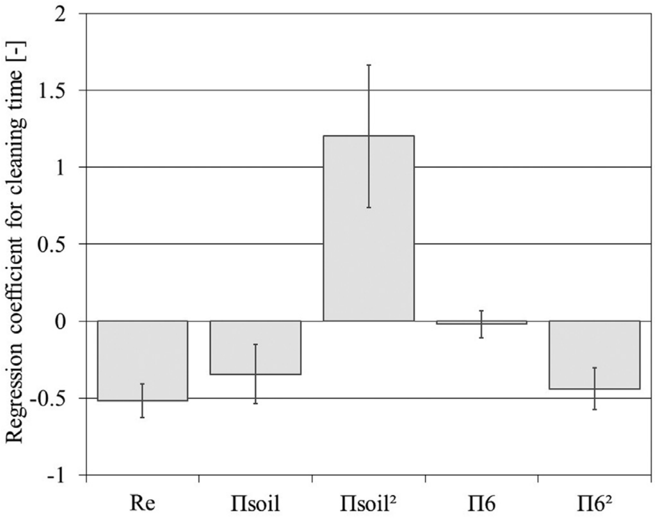

Fig. 2. Scaled and centered coefficient plot indicating the significance of the model terms Re, Π_{soil}, and Π_6 on the cleaning time.

Table 2
p-values of the model terms.

Model term	*p*-value
Re	$2.64\cdot10^{-14}$
Π_{soil}	$6.18\cdot10^{-4}$
Π_{soil^2}	$1.94\cdot10^{-6}$
Π_6	0.64
Π_{6^2}	$1.00\cdot10^{-8}$

Table 3
Correlation matrix with all correlation coefficients between the terms in the model.

	Re	Π_{soil}	Π_{soil^2}	Π_6	Π_{6^2}
Re	1.000				
Π_{soil}	0.094	1.000			
Π_{soil^2}	0.003	-0.182	1.000		
Π_6	-0.127	0.051	-0.136	1.000	
Π_{6^2}	-0.077	-0.211	0.256	-0.152	1.000

The correlation matrix displays the correlation coefficients between all the terms in the model. The value of the correlation coefficient represents the extent of the statistical relationship between two terms. The value of the coefficient ranges from -1 to 1. If the value is close to zero, the terms have no relationship. Here, all the correlation coefficients are near zero, so collinearity problems due to correlations of the terms can be excluded.

In all cases, no significant interaction terms are found between the dimensionless numbers in the multiple regression equations. Thus, the numbers can be described as independent and necessary for the comprehensive description of the cleaning process. Finally, three non-interacting dimensionless numbers were found:

- the Reynolds number, representing the conditions of the cleaning process,
- the density number Π_6, and
- the soil number Π_{soil}, describing the characteristic soil properties.

4.2. Resulting model

4.2.1. Multiple regression equation

With the knowledge of the significant terms, the model equation for the cleaning process can be defined. The resulting equations are obtained by both the statistical analysis and the machine learning approach. The model terms and coefficients of the multiple regression equations from the statistical analysis are extracted from the unscaled significance list created with MODDE®. Since the response distribution has a positive skewness (many values have short cleaning times, few values have long cleaning times), the response is transformed logarithmically to base 10 to achieve a normal distribution of the response. The model equation for the transformed response *y* is:

Table 4
Statistical model parameters for the model predicting the cleaning time.

Statistical model parameters	Value
Goodness of fit R2	0.67
Predictive power Q2	0.60
Validity V	0.55
Reproducibility R	0.68

$$log_{10}(y) = 0.561 - 1.101\cdot10^{-5}\cdot Re - 1.533\cdot10^{-15}\cdot\pi_{soil} + 3.736\cdot10^{-31}\cdot\pi_{soil}^2 + 3.159\cdot\pi_6 - 2.631\cdot\pi_6^2 \qquad (20)$$

The negative regression coefficient for the Reynolds number indicates a reduced cleaning time with increasing Reynolds number. The soil number influences the cleaning time with a linear and a quadratic term. The coefficient for the quadratic term is positive; therefore, the response decreases to a minimum with increasing soil numbers and then increases again with increasing soil numbers. The third significant number is the density number, which has a negative quadratic influence on the response.

The model is assessed with the statistical model parameters goodness of fit R2, predictive power Q2, validity V, and reproducibility R. The values for the model parameters are given in Table 4. The goodness of fit (ranging from *0* to *1*) is R2 = *0.67* for this model, which is above the threshold value R2 = *0.5* for defining a model that fits the data well, due to small variations of the predicted response. The goodness of fit is influenced by the reproducibility, which is also analyzed. The predictive power of the model can have a maximum value as high as R2; in this case, it is Q2 = *0.60*. The closer Q2 is to R2, the better is the predictive power of the model. The software code suggests that the predictive power be evaluated by two rules for a significant model: Q2 = *0.5* and Q2 > *0.8* · R2. Both rules are fulfilled in this case. Thus, the model can be interpreted as useful for the prediction of the response.

The predictive power is also influenced by the reproducibility. With a model validity of V = *0.55* > *0.25*, the model describes the measured values and the null hypothesis does not explain the observation, so the model is significant for the prediction of the response. Furthermore, the validity suggests that the model error is in the same range as, or is smaller than, the uncertainty of the measurements. The reproducibility $R = 0.68$ is interpreted as large enough to have a model with small uncertainties in the raw data and therefore with sufficient control of the cleaning experiments.

This model is used to calculate the predicted values for the cleaning time. The values for an exemplarily chosen set point (Helbig et al., 2019), the resulting numbers, and the predicted as well as the observed cleaning times, are given in Table 5.

Fig. 3 shows a parity plot of all the predicted versus observed (i.e., experimental) data. Even though outliers are present, the tendency is evident toward a higher predicted response with higher observed values. In general, the deviation of the data points is due to the uncertainty of the cleaning time measurement.

The assessment of the statistical analysis method leads to the conclusion that computational modeling of cleaning processes in the food industry is viable. As a first step, a model equation can be created with the multiple regression equation for the statistical prediction of the cleaning time. Nevertheless, an important subse-

Table 5
Observed and predicted cleaning times and calculated numbers for an exemplarily chosen set point from the multiple regression equations from the statistical analysis (parameters from Helbig et al., 2019).

Parameter	Setting	Unit
Pipe geometry d	0.026	m
Flow velocity w	0.40	$m \cdot s^{-1}$
Concentration of cleaning agent c	1.80	$kg \cdot m^{-3}$
Density of cleaning agent ρ	996.40	$kg \cdot m^{-3}$
Dynamic viscosity of cleaning agent η	$8.94 \cdot 10^{-4}$	$kg \cdot m^{-1} \cdot s^{-1}$
Layer thickness δ_0	110	µm
Free surface energy of soil γ_{soil}	33610	$N\ m^{-1}$
Viscosity of soil η_{soil}	56.1	$kg \cdot m^{-1} \cdot s^{-1}$
Density of soil ρ_{soil}	1000.0	$kg \cdot m^{-3}$
Reynolds number Re	11665	-
Density number Π_6	1.004	-
Density number Π_6 squared	1.008	-
Soil number Π_{soil}	$3.626 \cdot 10^{11}$	-
Soil number Π_{soil} squared	$1.315 \cdot 10^{23}$	-
Predicted cleaning time	8.95	min
Lower confidence limit	7.61	min
Upper confidence limit	10.29	min
Observed cleaning time t_{95}	9.84	min

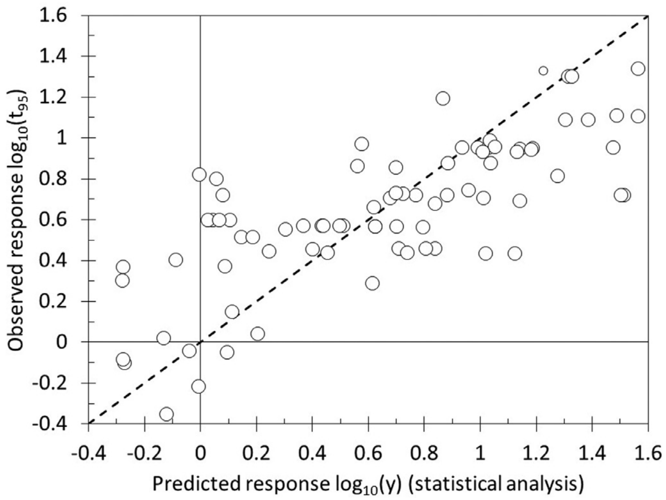

Fig. 3. Observed versus predicted values of the model transformed response from the multiple regression equation.

quent step is to evaluate this method against other approaches, like the development of a model with a machine-learning algorithm.

4.2.2. Symbolic regression

The machine learning approach of the Eureqa software uses symbolic regression algorithms to create models. Based on the previous results of the statistical analysis, a model is developed for the logarithmic response, as that distribution fits better than the normal distribution. The transformed response is chosen for a better comparison of the coefficients. Therefore, the target expression for the response $log_{10}(y)$, as a function of the Reynolds number Re, the soil number Π_{soil}, and the density number Π_6, is defined as follows:

$$log_{10}(y) = f(Re,\ \pi_{soil}, \pi_6) \tag{21}$$

The model equation f is obtained using all basic mathematical functions. In addition, the functions from the exponential group are included in the model formation, with the exception of factorial functions. Terms that are more complex, like trigonometric functions or logical operators, are not considered in the present study, since these functions are unlikely to represent the physical context. All functions used are listed in Table 6.

Table 6
Functions used by the symbolic regression approach.

Basic	Exponential
Constant	Exponential
Input Variable	Natural Logarithm
Addition	Power
Subtraction	Square Root
Multiplication	
Division	

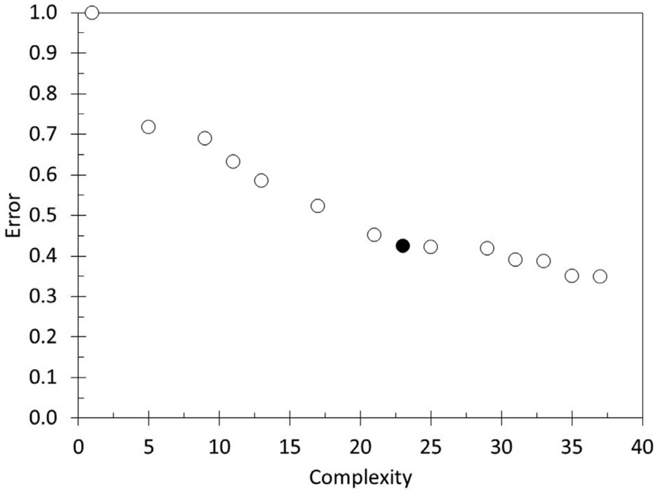

Fig. 4. Accuracy-versus-complexity diagram indicating the selected solution (●) for the model developed with the symbolic regression.

The R2 goodness of fit is selected as error metric, because it is easily comparable to the R2 calculated for the model achieved with the statistical analysis. With these general conditions, the search for equations is started in Eureqa.

Terms with a high significance have a greater response and are found earlier in the logical tree than are terms with minor effects. The evolutionary computation of the symbolic regression finds a candidate solution containing the Reynolds number very early and preserves this term for all further developed equations with higher complexity. The next dimensionless number appearing in the binary tree is the soil number; this number is first considered as a linear term and later as a quadratic one. The density number is then considered in a logical tree. This order provides an estimate of the significance of these terms.

All equations are compiled in the list with the best solutions of different complexity. Eureqa provides a set of candidate solutions, each offering the best solution for a specific size. From this list of results, a model is selected by analyzing the accuracy versus complexity diagram. The accuracy is the R2 goodness of fit and is calculated from the model's error. The first model found is just a constant value with a complexity of *1* and a very high error. Increasing the complexity of the model equation reduces the error. A model should be selected if a further increase in complexity is not rewarded by a corresponding reduction in error. As seen in Fig. 4, a model with the complexity of *23* and an error of *0.43*, which refers to a goodness of fit R2 = *0.57*, seems the best compromise (indicated by the black dot in the graph). This model is above the threshold value of R2 = *0.5* and the gain in accuracy is comparatively low for the subsequent models with increasing complexity of the model equation. The accuracy only increases slightly again from a complexity of *31*. In these more complex equations, the dimensionless numbers now also appear in the exponent of the constants or other dimensionless numbers.

Table 7
Performance parameters of the model equation found by symbolic regression predicting the cleaning time.

Performance parameter	Value
R2	0.57
Complexity	23
Correlation coefficient	0.76
Mean absolute error	0.26

Table 8
Confidence intervals for the regression coefficients.

Model term	Regression coefficient	Confidence interval
Re	$1.105 \cdot 10^{-5}$	$7.38 \cdot 10^{-8}$
Π_{soil}	$9.174 \cdot 10^{-17}$	$4.69 \cdot 10^{-19}$
Π_{soil^2}	$3.329 \cdot 10^{-31}$	$1.70 \cdot 10^{-33}$
Π_6	3.346	0.04
Π_{6^2}	2.903	0.04
$\Pi_{soil} \cdot \Pi_6$	$1.829 \cdot 10^{-15}$	$2.00 \cdot 10^{-31}$

The correlation coefficient of this developed model, with a value near 1, indicates that the model correlates well with the response. The mean absolute error (the mean difference of the predicted value from the observed value) is used to evaluate the exactness of the model. The value of 0.26 shows a slight deviation from the predicted values. The performance parameters of the model equation found by symbolic regression are given in Table 7.

The model equation found by symbolic regression is:

$$log_{10}(y) = 0.512 - 1.105 \cdot 10^{-5} \cdot Re - 9.174 \cdot 10^{-17} \cdot \pi_{soil} + 3.329 \cdot 10^{-31} \cdot \pi_{soil}^2 - 3.462 \cdot \pi_6 + 2.903 \cdot \pi_6^2 + 1.829 \cdot 10^{-15} \cdot \pi_{soil} \cdot \pi_6 \quad (22)$$

The regression coefficients are assessed with the 95% confidence interval (see Table 8). It can be interpreted, that the probability that the confidence interval contains the regression coefficient is 95%. The values for the intervals are very small, so the regression coefficients are considered well fitting. The confidence intervals are achieved by a bootstrap procedure, in which the symbolic regression is performed multiple times.

Notably, this equation includes the same significant terms previously found by significance analysis. The Reynolds number has a linear influence on the transformed response, while the soil number and the density number have both linear and quadratic influences. In addition, an interaction term is noted for Π_{soil} and Π_6. For both models, the values of the coefficients of the terms are comparable. The results of the model from the symbolic regression are assessed with the observed versus predicted plot (Fig. 5).

The observed versus predicted plot for the transformed response of the model developed with the symbolic regression has a similar pattern to that seen in the plot for the multiple regression equation (Fig. 3).

4.2.3. Comparison of both methods

Both methods led to the development of similar model equations. The same relevant terms are present in the equations and, due to numerically similar coefficients of the model terms, the predicted response is also very similar (see Figs. 3 and 5).

The procedure used for the development of the model equation is very different for each model. The machine learning approach is capable of handling mathematical operations of greater complexity, but its necessity must be questioned for the given problem. In the case of the prediction of cleaning time for cleaning processes in food industry, a practical approach should be based on mechanistic formulations. It is important to prepare the data and define how to follow the approach before initiating the development of model equations. With appropriate knowledge, the machine learning approach is easy to conduct with a software tool like Eureqa, and model equations can be obtained quickly. Nevertheless, the selection of the model equation among candidates of different size is, to some extent, subjective, and a compromise will always exist between the accuracy and size of a model.

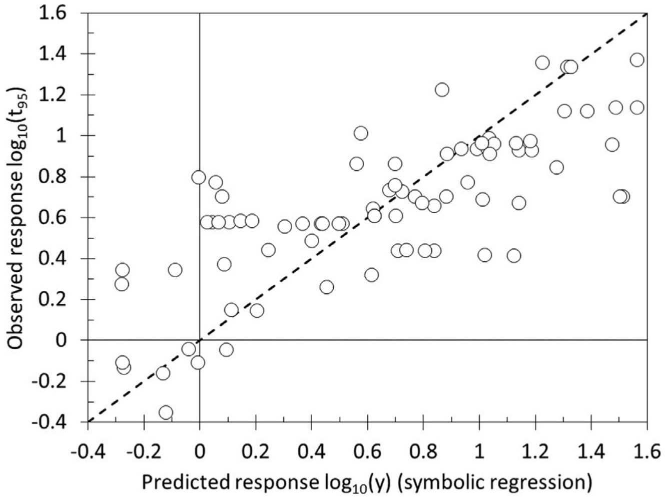

Fig. 5. Observed versus predicted values of the models transformed response from the symbolic regression.

The development of the model with statistical analysis follows the selection of significant terms assessed by the significance analysis. Therefore, the resulting equation is developed term by term and is evaluated with different statistical parameters. Statistical knowledge is required for this procedure. In the case of cleaning processes in the food industry, the developed model is objectively assessed as significant with the statistical model parameters, which is a substantial advantage of the statistical analysis.

5. Conclusion

In this study, two methods were compared for the creation of predictive models to predict the cleaning time in the food industry: a statistical analysis with MODDE® and a symbolic regression with Eureqa. These methods can be applied to cleaning processes in the food industry to create comprehensive models that compare and predict the cleaning success and effectiveness for different cleaning situations. In the case of the database used, with the cleaning time as the response, both methods are found applicable for computational modeling of cleaning processes in food industry, and comparable models can be developed. However, the identification of significant parameters in the model is more objective with the statistical analysis due to the individual assessment of the parameters.

The influencing factors were considered as dimensionless numbers and were obtained with a dimensional analysis. Apart from the known dimensionless numbers, a new dimensionless number characterizing the food soils properties was pointed out. Even though the soil number Π_{soil} is a dependent dimensionless number, since it is a combination of other numbers, it was very effective for modeling because it provides a comprehensive description of the soils.

Future work will aim to validate this approach with different cleaning databases. The accuracy of prediction still has to be assessed, especially for set points outside the design region of the database. Therefore, experiments with different process conditions or other food soils will be carried out.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Hannes Deponte: Conceptualization, Methodology, Formal analysis, Writing - original draft. **Alberto Tonda:** Conceptualization, Methodology, Software, Writing - original draft. **Nathalie Gottschalk:** Conceptualization. **Laurent Bouvier:** Writing - review & editing. **Guillaume Delaplace:** Conceptualization, Writing - review & editing. **Wolfgang Augustin:** Conceptualization, Writing - review & editing, Supervision. **Stephan Scholl:** Writing - review & editing, Supervision.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Albrecht, M.C., Nachtsheim, C.J., Albrecht, T.A., Cook, R.D., 2013. Experimental design for engineering dimensional analysis. Technometrics 55, 257–270. doi:10.1080/00401706.2012.746207.

Asteriadou, K., Hasting, A.P.M., Bird, M.R., Melrose, J., 2006. Computational fluid dynamics for the prediction of temperature profiles and hygienic design in the food industry. Food Bioprod. Process 84 (C2), 157–163. doi:10.1205/fbp.04261.

Babutzka, J., Bortz, M., Dinges, A., Foltin, G., Hajnal, D., Schultze, H., Weiss, H., 2019. Machine learning supporting experimental design for product development in the lab. Chem. Ing. Tech. 91 (3), 277–284. doi:10.1002/cite.201800089.

Buckingham, E., 1914. On physically similar systems; Illustrations of the use of dimensional equations. Phys. Rev. 4, 345–376. doi:10.1103/PhysRev.4.345.

Chen, B., Callens, D., Campistron, P., Moulin, E., Debreyne, P., Delaplace, G., 2019. Monitoring cleaning cycles of fouled ducts using ultrasonic Coda Wave Interferometry (CWI). Ultrasonics 96, 253–260. doi:10.1016/j.ultras.2018.12.011.

Cole, P.A., Asteriadou, K., Robbins, P.T., Owen, E.G., Montague, G.A., Fryer, P.J., 2010. Comparison of cleaning of toothpaste from surfaces and pilot scale pipework. Food Bioprod. Process. 88, 392–400. doi:10.1016/j.fbp.2010.08.008.

Delaplace, G., Loubière, K., Ducept, F., Jeantet, R., 2015. Dimensional Analysis of Food Processes. ISTE Press–Elsevier, London doi:10.1016/C2014-0-04744-2.

Dennison, T.J., Smith, J., Hofmann, M.P., Bland, C.E., Badhan, R.K., Al-Khattawi, A., Mohammed, A.R., 2016. Design of experiments to study the impact of process parameters on droplet size and development of non-invasive imaging techniques in tablet coating. PLoS One 11, e0157267. doi:10.1371/journal.pone.0157267.

Deponte, H., Helbig, M., Gottschalk, N., Augustin, W., Scholl, S., 2018. Dimensional analysis of cleaning-in-place processes for fouled organic material in food processes. In: Lipnizki, F., Wilson, D.I., Chew, Y.M.J., Jönsson, A.S. (Eds.), Fouling and Cleaning in Food Processing. Lund University, Lund, Sweden, p. 15.

Dürr, H., Graßhoff, A., 1999. Milk heat exchanger cleaning: modelling of deposit removal. Food Bioprod. Process. 77, 114–118. doi:10.1205/096030899532402.

Fryer, P.J., Asteriadou, K., 2009. A prototype cleaning map: a classification of industrial cleaning processes. Trends Food Sci. Technol. 20, 255–262. doi:10.1016/j.tifs.2009.03.005.

Fryer, P.J., Christian, G.K., Liu, W., 2006. How hygiene happens: physics and chemistry of cleaning. Int J Dairy Technol 59, 76–84. doi:10.1111/j.1471-0307.2006.00249.x.

Goode, K.R., Asteriadou, K., Robbins, P.T., Fryer, P.J., 2013. Fouling and cleaning studies in the food and beverage industry classified by cleaning type. Compr. Rev. Food Sci. Food Saf. 12, 121–143. doi:10.1111/1541-4337.12000.

Gordon, P.W., Schöler, M., Föste, H., Helbig, M., Augustin, W., Chew, Y.M.J., Scholl, S., Majschak, J.P., Wilson, D.I., 2014. A comparison of local phosphorescence detection and fluid dynamic gauging methods for studying the removal of cohesive fouling layers: Effect of layer roughness. Food Bioprod. Process. 92, 46–53. doi:10.1016/j.fbp.2013.07.010.

Gottschalk, N., Reuter, L.S., Zindler, S., Föste, H., Augustin, W., Scholl, S., 2019. Determination of cleaning mechanisms by measuring particle size distributions. Food Bioprod. Process. 113, 77–85. doi:10.1016/j.fbp.2018.10.003.

Helbig, M., Zahn, S., Böttcher, K., Rohm, H., Majschak, J.P., 2019. Laboratory methods to predict the cleaning behaviour of egg yolk layers in a flow channel. Food Bioprod. Process. 113, 108–117. doi:10.1016/j.fbp.2018.11.005.

Jensen, B.B.B., Friis, A., 2005. Predicting the cleanability of mix-proof valves by use of wall shear stress. J. Food Process Eng. 28, 89–106. doi:10.1111/j.1745-4530.2005.00370.x.

Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge.

Lelieveld, H., Mostert, T., Holah, J., 2005. Handbook of Hygiene Control in the Food Industry. Elsevier, London doi:10.1533/9781845690533.

Palabiyik, I., Olunloyo, B., Fryer, P.J., Robbins, P.T., 2014. Flow regimes in the emptying of pipes filled with a Herschel-Bulkley fluid. Chem. Eng. Res. Des. 92, 2201–2212. doi:10.1016/j.cherd.2014.01.001.

Schmidt, M., Lipson, H., 2009. Distilling free-form natural laws from experimental data. Science 324, 81–85. doi:10.1126/science.1165893.

Schöler, M., Föste, H., Helbig, M., Gottwald, A., Friedrichs, J., Werner, C., Augustin, W., Scholl, S., Majschak, J.P., 2012. Local analysis of cleaning mechanisms in CIP processes. Food Bioprod. Process. 90, 858–866. doi:10.1016/j.fbp.2012.06.005.

Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., De Freitas, N., 2016. Taking the human out of the loop: a review of Bayesian optimization. Proc. IEEE. 104, 148–175. doi:10.1109/JPROC.2015.2494218.

Shen, W., Davis, T., Lin, D., Nachtsheim, C., 2014. Dimensional analysis and its applications in statistics. J. Qual. Technol. 46, 185–198. doi:10.1080/00224065.2014.11917964.

Siebertz, K., van Bebber, D., Hochkirchen, T., Siebertz, K., Bebber, D., van, Hochkirchen, T., 2017. Kontrollverfahren. In: Siebertz, K., Bebber, D., Hochkirchen, T. (Eds.), Statistische Versuchsplanung. Springer, Berlin, pp. 61–85. doi:10.1007/978-3-662-55743-3_3.

Sonin, A.A., 2004. A generalization of the Pi-theorem and dimensional analysis. Proc. Natl. Acad. Sci. 101 (23), 8525–8526. doi:10.1073/pnas.0402931101.

Stichlmair, J., 2001. Scale-up Engineering. Begell House, United States.

Szirtes, T., Rózsa, P., 2007. Applied Dimensional Analysis and Modeling. Elsevier, London doi:10.1016/B978-0-12-370620-1.X5000-X.

Tamime, A., 2009. Cleaning-in-Place: Dairy, Food and Beverage Operations, Third Ed. Wiley-Blackwell, Hoboken NJ doi:10.1002/9781444302240.

Woods, D.C., Overstall, A.M., Adamou, M., Waite, T.W., 2017. Bayesian design of experiments for generalized linear models and dimensional analysis with industrial and scientific application. Qual. Eng. 29, 91–103. doi:10.1080/08982112.2016.1246045.

Ziskind, G., Fichman, M., Gutfinger, C., 1995. Resuspension of particulates from surfaces to turbulent flows-Review and analysis. J. Aerosol Sci. 26, 613–644. doi:10.1016/0021-8502(94)00139-P.