UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

MARINA RODRIGUES DE OLIVEIRA SEIBERT

OPTIMAL FEATURE SELECTION BASED ON CHEMICAL ENGINEERING CONCEPTS AND PROPOSAL SOFT SENSOR TO PREDICT *f*-CaO IN CLINKER USING INDUSTRIAL DATA

BELO HORIZONTE - MG
2023

MARINA RODRIGUES DE OLIVEIRA SEIBERT

OPTIMAL FEATURE SELECTION BASED ON CHEMICAL ENGINEERING CONCEPTS AND PROPOSAL SOFT SENSOR TO PREDICT *f*-CaO IN CLINKER USING INDUSTRIAL DATA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Química da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do Grau de Mestre em Engenharia Química.

Linha de Pesquisa: Simulação e Otimização de Processos.

Orientador: Esly Ferreira da Costa Junior.

BELO HORIZONTE – MG
2023

Seibert, Marina Rodrigues de Oliveira.
S457o Optimal feature selection based on chemical engineering concepts and proposal soft sensor to predict *f*-CaO in clinker using industrial data [recurso eletrônico] / Marina Rodrigues de Oliveira Seibert. - 2023.
1 recurso online (95 f. : il., color.) : pdf.

Orientador: Esly Ferreira da Costa Junior.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Apêndices: f. 91-95.

Bibliografia: f. 86-90.
Exigências do sistema: Adobe Acrobat Reader.

1. Engenharia química - Teses. 2. Mineração - Teses. 3. Cimento - Indústria - Teses. 4. Óxido de cálcio - Teses. 5. Polinômios - Teses. 6. Sensor virtual - Teses. 7. Fenomenologia - Teses. 8. Variáveis (Matemática) - Teses. I. Costa Junior, Esly Ferreira da. II. Universidade Federal de Minas Gerais. Escola de Engenharia III. Título.

CDU: 66.0(043)

Ficha catalográfica elaborada pela bibliotecária Ângela Cristina Silva CRB/6 2361
Biblioteca Prof. Mário Werneck, Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

FOLHA DE APROVAÇÃO

"OPTIMAL FEATURE SELECTION BASED ON CHEMICAL ENGINEERING CONCEPTS AND PROPOSAL SOFT SENSOR TO PREDICT F-CAO IN CLINKER USING INDUSTRIAL DATA"

Marina Rodrigues de Oliveira Seibert

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Química da Escola de Engenharia da Universidade Federal de Minas Gerais, como parte dos requisitos à obtenção do título de **MESTRE EM ENGENHARIA QUÍMICA**.

307ª DISSERTAÇÃO APROVADA EM 25 DE AGOSTO DE 2023 POR:

Documento assinado eletronicamente por **Gustavo Matheus de Almeida**, **Professor do Magistério Superior**, em 25/08/2023, às 14:15, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Andrea Oliveira Souza da Costa**, **Coordenador(a)**, em 25/08/2023, às 14:30, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Esly Ferreira da Costa Junior**, **Professor do Magistério Superior**, em 25/08/2023, às 14:59, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2537823** e o código CRC **C8B670C2**.

Referência: Processo nº 23072.248868/2023-45 SEI nº 2537823

ACKNOWLEDGMENTS

Maybe this is the most complicated part of this work. I am writing these words in the middle of a trip, between the mountains, in Peru Andes, this thesis was written in more than five countries. I do not know how many people I have thank!!! Without you, this work would not be possible.

I have one phrase for this, “Talent wins games, but teamwork and intelligence win championships”. (Michael Jordan)

I had the help of many people to accomplish this work. Here are some of them:

Professors Andréa Oliveira Souza da Costa and Esly Ferreira da Costa Jr. did not allow me to give up during the most impossible moments in this journey.

Unexpected friends who I met during this process, Bruno Scalia and Ana Esther saved me in the dark moments with the data. There is some sarcasm here.

Diego Galarza, a friend, and tutor, who always believed in me and did not allow me to give up my engineering career.

Erick Oliveira, who I met in the while working on this research and helped me so much with my software doubts and troubles. Python sent thanks to you.

Thanks to a special friend, Breno Reis, my brother, Antônio Henrique and mother, Eugênia Oliveira. They do not have any idea about what I did here, but this was not an excuse to not support me as much as they could.

My friends also have a space here. They are the family that I chose: Alessandra Limão, Michelle Carine, Janaína Vasconcellos, Thiago Vasconcellos, Juliana Midori, Thiago Francês, Leandro Oliveira, famous as ‘os malas’. Thank you for the words, support, advice, time, and everything that someone could hope for from realistic friendships.

Finally, I would like to say thank you to this institution, UFMG, Universidade Federal de Minas Gerais, which is responsible for a great change in my way of my life.

The present work represents much more than one academic work, it is a symbol of a personal and professional journey, of which the goal only was achievable because the correct people were beside me. Thanks to everyone that did not give up and stayed together with me.

RESUMO

As indústrias de cimento e mineração estão diretamente relacionadas com o desenvolvimento econômico mundial e fornecem materiais essenciais para a transição para energias limpas. Contudo, a indústria cimenteira é responsável por 7% das emissões mundiais de CO_2. O desenvolvimento de um sensor virtual para previsão de cal livre (*f*-CaO) representa uma melhoria, pois diminui o consumo específico de energia térmica para a produção do clínquer, principal componente do cimento. Qualquer nova tecnologia na indústria cimenteira deve considerar as reações químicas heterogêneas envolvidas na produção do clínquer. Demonstrar a importância da seleção de variáveis do sistema que impactam o *f*-CaO e interpretar o significado das variáveis do ponto de vista fenomenológico, que incluiu conceitos relacionados à termodinâmica, mecânica dos fluidos, cinética química, é um diferencial, pois não é uma abordagem usual na literatura. Seis meses de dados industriais foram analisados para determinar o conjunto ótimo de variáveis; com o subconjunto definido, a segunda etapa foi aplicar as técnicas de aprendizado de máquina nos dados para desenvolver um sensor virtual capaz de prever *f*-CaO no clinquer; softwares como MATLAB, Microsoft Excel®, códigos em linguagem C++, bibliotecas escritas para a linguagem Python como pandas, NumPy foram utilizados para implementar a análise de dados, que inclui pré-processamento, regressão linear múltipla (MLR) entre outros. Algoritmos de aprendizado de máquina foram utilizados para modelos de previsão. Na etapa de pré-processamento, as variáveis inconsistentes foram eliminadas e o conhecimento sobre a operação do forno foi o alicerce para a tomada de decisão. O filtro para o conjunto final de dados foi as variáveis relacionadas a qualidade do clínquer, uma vez que não há amostragem do clínquer se o sistema do forno falhar. Em seguida, variáveis da amostragem anterior da qualidade do clínquer foram inseridas juntamente com a amostra atual. Dois conjuntos de dados foram gerados, pois cerca de 50% dos dados possuíam valores contínuos da qualidade da farinha. O primeiro conjunto de dados, DATASET01, contém todos os dados SEM as variáveis das análises químicas online. O segundo conjunto de dados, DATASET02, são os dados COM as variáveis das análises químicas online. Para cada conjunto de dados foram realizadas várias simulações com MLR combinada com a metodologia forward-stepwise para selecionar o conjunto de variáveis. Algoritmos robustos não podem compensar a falha na escolha das

variáveis; a seleção de variáveis é tão importante quanto a aplicação dos próprios algoritmos de previsão. Os resultados obtidos na MLR demonstraram a importância da etapa de seleção de variáveis. As variáveis relacionadas à composição química e operação do resfriador têm influência substancial nos modelos de predição do *f*-CaO. Os modelos estatísticos complexos (XGBoost, CatBoost, SVM ; RDF) tiveram baixo desempenho; a otimização dos hiperparâmetros combinada com a metodologia apresentada neste trabalho é sugerida para trabalhos futuros. Finalmente, os modelos polinomiais multivariados tiveram resultados satisfatórios, com R^2=0.78 e R^2=0.75 nos modelos de terceiro e quarto grau respectivamente. Há oportunidades para melhorar o desempenho do modelo polinomial para um conjunto de dados maior com a análise química da farinha crua online disponível.

Palavras-chave: Quantidade de óxido de cálcio no clinquer. Sensor virtual. Indústria cimenteira. Polinômios multivariados. Seleção de variáveis. Parâmetros fenomenológicos.

ABSTRACT

Mineral and cement industries are directly related to the world's economic development and supply essential materials for the clean-energy transition. However, the cement industry is responsible for 7% of worldwide CO_2 emissions. The development of a soft sensor to predict free lime (*f*-CaO) represents an improvement due to reducing the specific thermal energy consumption to produce clinker, the main and most expensive cement component while maintaining the desired cement quality. Any new technology involving the cement industry has to consider the complex heterogenous chemical reactions involved in clinker production. A demonstration of the importance of an optimal feature selection regarding the kiln system, which impacts the *f*-CaO, and interpreting the meaning of these variables from a phenomenological point of view, including concepts related to thermodynamics, fluid dynamics and chemical kinetics is the main distinguisher of this work, for it is not a common approach in literature. Six months of industrial data were investigated using a combination of deep system knowledge , as well as statical tools to determine the optimal operational and quality features that impact the *f*-CaO at clinker; with the subset of features defined, the second step was applying machine learning techniques on the data to develop a soft sensor to predict *f*-CaO; MATLAB, Microsoft Excel®, self-written code using C++ language and Python libraries, such as pandas, NumPy and statsmodels, were used to implement the data analysis, which includes data pre-processing, multiple linear regression (MLR), and prediction models using standard machine learning algorithms. In the pre-processing step, inconsistent features were deleted, and knowledge about kiln operation was the basis for decision-making. The filter for the final dataset was based on clinker quality features because there is no clinker sampling if the kiln system fails. Then, previous clinker quality features were input together with the time series sample. Two datasets were generated, due to approximately 50% of the data being continuous raw meal quality values (online chemical analyses). The first dataset, DATASET01, consists of all the data *WITHOUT* the online chemical analyses' features. The second dataset, DATASET02, is the data *WITH* the online chemical analyses' features. For each dataset, various simulations were carried out using MLR combined with the forward stepwise methodology to select the feature set. Robust algorithms cannot compensate for an incorrect variable setup; therefore the feature selection step is as important as the application of the prediction algorithm. The results

obtained in the MLR demonstrated the importance of the feature selection step. The variables related to the chemical composition and cooler operation have a substantial influence on the prediction models for *f*-CaO. The complex statistical models (XGBoost, CatBoost, SVM and RDF) had poor performance and hyperparameter optimization combined with the methodology present in the current work is suggested for future research. Finally, the multivariate polynomial models had satisfactory results, with R^2=0.78 in the fourth-degree model and R^2=0.75 in the three-degree model. There are opportunities to improve the polynomials model's performance for a bigger dataset size with raw meal chemical analysis online available.

Keywords: Free calcium oxide content. Soft sensor. Cement industry. Multivariate polynomials. Feature selection. Phenomenological parameters

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI Artificial Intelligence

ANN Artificial Neural Networks

CatBoost Gradient Boosting on Decision Trees

f-CaO Free lime; Calcium oxide

ML Machine learning

MLP Multi-Layer Perceptron

MLR Multiple Linear Regression

PolR Polynomial Regression

RDF Random Decision Forests

SVM Support Vector Machine

XGBoost Extreme Gradient Boosting

LIST OF SYMBOLS

$\%h_H$	Water of crystallization in the clay	[g/100g clinker]
%F	Kiln filling	[%volume]
$\%FL_{1450^\circ C}$	% liquid phase, mass basis, at 1450°C	[%mass]
ρ_{cl}	Clinker apparent density	[t/m^3]
∅	Kiln slope	[%m/m]
$\Delta HF_{clinquer}$	Heat of formation for clinker at 25ºC	[kcal/kg]
A	Total grate area	[m^2]
AR	Alumina Ratio	[-]
BF	Burnability factor	[-]
BI	Burnability index	[-]
b_i	Polynomial coefficient with index *i*	[-]
C_{pgas}	Specific heat capacity at constant pressure	[kJ/ Nm^3.ºC]
C_2S	Belite	[%mass]
C_3A	Aluminate	[%mass]
C_3S	Alite	[%mass]
C_4AF	Ferrite	[%mass]
D_i	Rotary kiln internal diameter	[m]
CaO_{free}	Calcium oxide (lime)	[%mass]
H	Clinker height at grate	[m]
kW1	Initial power required by the fan	[kW]
kW2	New power required by the fan	[kW]
L	Rotary kiln length	[m]
LSF	Lime saturation factor	[-]
$\dot{m}$	Hot gas flow	[Nm^3/hour]
$\dot{m}_1$	Initial volume of air or gas	[m^3/hour]
$\dot{m}_2$	New volume of air or gas	[m^3/hour]
MRE	Mean relative error	[%]
Na_2O, K_2O	Oxides values obtained from raw meal x-ray fluorescence analysis	[%mass]
N	Kiln speed	[rpm]
n	Number of regression variables	[-]

Naeq	Total alkalis as Na_2O	[%mass]
nf	Number of regression variables after applying the null-hypothesis significance	[-]
ni	Number of regression variables *before* applying the null-hypothesis significance	[-]
nv	Number of process variables	[-]
ns	Number of sampling points	[-]
Pr_{cl}	Clinker production	[tph]
P1	Initial pressure	[mbar]
P2	New pressure	[mbar]
Q	Sensible heat from hot gases	[kJ/hour]
R^2	Coefficient of determination	[-]
r_i	Correlation coefficient for each variable with index *i*	
RMSE	Root mean squared error	[-]
SM	Silica Modulus	[-]
t_{cooler}	Cooler retention time	[min]
t_i^*	test statistic for the variable with index *i*	[-]
t_{kiln}	Retention time in the rotary kiln	[min]
T_0	Gas at reference temperature, 20℃	[℃]
T_{gas}	Gas temperature,	[℃]
U1	Initial fan impeller rotation	[rpm]
U2	New fan impeller rotation	[rpm]
W1	Initial power required by the fan	[kW]
W2	New power required by the fan	[kW]
w_i	Functions of *nv* and depend on the degree of adjusted polynomial	[-]
$\hat{y}_j$	Predict value of model (*f*-CaO) with index j	[%mass]
$\bar{y}$	Average for real values (*f*-CaO)	[%mass]
y_j	Real value of model (*f*-CaO) with index *j*	[%mass]
x_i	Process variable with index *i*	[-]
$\bar{x_i}$	Average for process variable with index *i*	[-]
$x_{i,j}$	Measured value of process variable x_i with index *j*	[-]

SUMMARY

1 INTRODUCTION

1.1 The importance of cement industry in the world development

World economic development is directly related to the consumption of some materials produced by mineral and cement industries. They are essential for the development of society as we know it: buildings, hospitals, bridges, stadiums, roads, and every majestic construction around us have materials such as cement, concrete and steel in their basis. It is impossible to think of our society without the crucial mineral and cement industry.

The statical relationship between cement production and economic development was found by Uwasu, Hara, Yabar (2014), and can be summarized in two important points:

1. Cement production increases with the level of development until a certain point in which a transition occurs and cement production begins to decline to a low level, tending to convergence.

2. Where the peak occurs and at what level it converges change by country and region. For example, in Europe, the peak occurs when per capita is USD 4210, with a corresponding per capita cement production of 531kg. On the other hand, in East Asia and North America it is close to 500 kg in regions of the same income level.

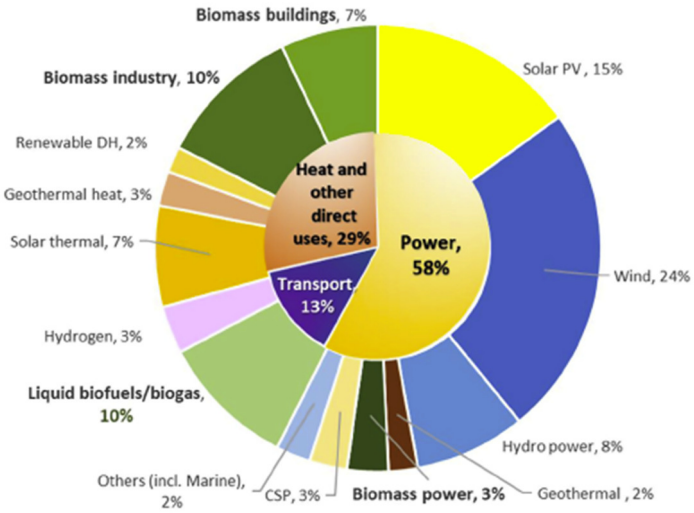
When the Sustainable Development Goals (SDGs), adopted by the United Nations General Assembly (UNGA) in 2015, are also considered, the central role of the cement industry in the development is again highlighted.

There are 17 SDGs, and their 169 targets are central in the “2030 Agenda”, which defines a guideline to end extreme poverty, fight inequality and injustice, and protect the environment. In this topic, sustainable energy is central to the success of “Agenda 2030” (Dolf, et al., 2019), and here the cement industry shows its importance.

An ambitious estimate for the renewable energy sector in 2050 is in Figure 1. In total, 222 EJ (EJ) of renewable energy is deployed in final energy terms and the power sector accounts for 58% (GIELEN et al., 2019) of this total. For all these sectors (solar

thermal plants biomass buildings, hydropower) cement and minerals are critical to build new renewable energy plants, and enormous amounts are consumed. The wind energy industry is a good example to understand the magnitude of the cement industry role in the clean-energy transition. In Figure 2, some numbers are shown associated with the erection of the wind turbine machines. 1.200 tons of concrete are necessary for a single 3-megawatt turbine.

Figure 1 - Breakdown of renewable resource used in regards to total final energy consumption, REmap 2050. Note: Excludes non-energy use.



Source: (GIELEN et al., 2019).

Figure 2 - Main materials used for wind turbine erection.

Source: (FLSmidth, 2020).

At the same time, the cement industry consumes approximately 12–15% of the total industrial energy use. Moreover, this industry releases CO_2 emissions to the atmosphere as a result of burning fossil fuels (mainly petroleum coke, natural gas, coal, lignite) to produce thermal energy needed for the pyro process to obtain clinker, the main and most expensive component of Portland Cement; as a consequence, the cement industry is responsible for around 7% of the total worldwide CO_2 emissions (ALI; SAIDUR; HOSSAIN, 2011).

On the other hand, a lot of technological improvements were achieved in the last decades in this industrial sector along with a substantial decrease in the electrical/thermal energy consumption in cement manufacturing. As a comparison, the 'oldest' common technology, the wet kiln process, consumed around 6 GJ/ton of clinker as opposed to less than 3 GJ/ton clinker consumed by the most modern technology, 6-Stage pre-heater *plus* calciner *plus* high-efficiency cooler (ALI; SAIDUR; HOSSAIN, 2011). This significant reduction in thermal energy consumption to produce clinker throughout the last decades cannot be ignored and is an incentive for continuous research and innovation in this essential industrial sector. A short chronology with the main technological improvements in the cement industry in regard to thermal energy consumption and its reduction is in Table 1.

Table 1 - Thermal energy consumption for different kiln process and notes about the developed technologies.

Kiln process	Thermal energy (GJ/ton clinker)	% reduction (related to previous technology)	Chronology / Notes
Wet process	5.86-6.28	-	First rotary kiln was patented by Frederik Ransome (1885), with 2m diameter and 25m length. Production: 30 to 50 ton clinker/day.
Long dry process	4.6	-24.2%	Long dry kilns are similar to long wet kilns. These kilns were developed and became popular particularly in North America.
1-Stage cyclone pre-heater	4.18	-9.1%	Cyclone preheater was patented in 1934 in Czechoslovakia by an employee of FLSmidth. The first preheater was built and commissioned in 1951 by KHD.
2-Stage cyclone pre-heater	3.77	-9.8%	
4-Stage cyclone pre-heater	3.55	-5.8%	The most common preheater is the 4-stage suspension preheater.
4-Stage cyclone pre-heater plus calciner	3.14	-11.5%	In the late 1960s, several precalciner systems began the precalciner revolution. By 1984, there were over 20 different types marketed by 28 manufacturers.
5-Stage cyclone pre-heater plus calciner plus high efficiency cooler	3.01	-4.1%	In the late 1990s, a new type of clinker cooler was introduced (Cross-BarTM cooler by FLSmidth) and the generation of high efficiency coolers began, with a thermal efficiency higher than 70% (75% can be achieved for the current high-efficiency cooler).
6-Stage cyclone pre-heater plus calciner plus high efficiency cooler	less than 2.93	-2.7%	
Total reduction (comparing 6-stage cyclone to wet process)		51.7% -	Today, some kilns are producing as much in a day (>10.000ton clinker/day) as the wet kiln produced in a year, with less than half the amount of thermal energy.

Source: Adapted from (ALI; SAIDUR; HOSSAIN, 2011) (PCA-Portland Cement Association, 2004).

1.2 What is cement and why is *f*-CaO a crucial parameter?

"CÆMENT, among builders, a strong sort of mortar, used to bind bricks or stones together. There are two sorts, 1. Hot cæment, which is the most common, made of resin, beeswax, brick-dust, and chalk, boiled together. 2. Cold cæment, made of Cheshire-cheese, milk, quicklime, and whites of eggs. This cæment is less used than the former and is accounted a secret known but to few bricklayers." Encyclopædia Britannica, 1771 (PCA-Portland Cement Association, 2004).

The use of cements as hydraulic binders goes back thousands of years; however, the modern term for cement that we know by the name 'Portland cement' appeared the first time in 1824, in a patent (No. 5.022) that was granted to Joseph Aspdin, a bricklayer from Leeds (PCA-Portland Cement Association, 2004)(HEWLETT, 2006). Portland Cement production, preparation, including chemical

composition and physical properties changed a lot throughout history. Nowadays, a better definition is: “Portland cement—a hydraulic cement produced by pulverizing clinker consisting essentially of hydraulic calcium silicates, usually containing one or more forms of calcium sulphate as an interground addition“ (ASTM C150 / C150M, 2021).

There are many types of 'Portland Cement', and the composition can vary in an enormous way, but, in general, the Portland Cement composition is as described in Table 2.

Table 2 - Main cement components.

Main types	Main constituents (% by mass)						
				Pozzolana		Fly ash	
	Clinker	Blast-furnace slag	Silica fume	Natural	Natural calcined	Siliceous	Calcareous
CEM I	95-100	-	-	-	-	-	-
CEM II	65-94	06-35	06-10	06-35	06-35	06-35	06-35
CEM III	05-64	36-95	-	-	-	-	-
CEM IV	45-89	-	----------11-55----------				
CEM V	20-64	18-49	-	----------18-49----------			-

	Burnt shale	Limestone	Gypsum
CEM I	-	-	0-5
CEM II	06-35	06-35	0-5
CEM III	-	-	0-5
CEM IV	1,5	-	0-5
CEM V	-	-	0-5

Source: Based on (EUROPEAN COMMITTEE FOR STANDARDIZATION, 2011).

As can be observed in Table 2, the main and most important component of Portland Cement is clinker, the intermediate cement product which is obtained by rotary kilns and is also the object of the current academic study. Clinker is composed mainly of 4 calcium silicates and *f*-CaO, as described in Table 3 with respective typical compositions for European clinkers:

Table 3 - Main calcium silicates and *f*-CaO present in clinker with typical values for European standards.

Symbol	Chemical formula	Mineral name	Values (%mass)	
			Standard	Interval
C_3S	3 CaO.SiO_2 OR Ca_3SiO_5	Alite	57	45-65
C_2S	2CaO.SiO_2 OR Ca_2SiO_4	Belite	16	10-30
C_3A	3 CaO.Al_2O_3 OR $Ca_3Al_2O_6$	Aluminate	9	5 - 12
C_4AF	4Cao.Al_2O_3.Fe_2O_3 OR $Ca_4Al_2Fe_2O_{10}$	Ferrite	10	6 - 12
F-CaO	CaO	Calcium oxide	1,0	1,0 - 1,5

Source: Adapted from (NEWMAN; CHOO, 2003).

In addition to the four main calcium silicates, another crucial component of clinker is *f*-CaO, *cf.* Table 3, which is the unreacted calcium oxide that remains in the clinker. The amount of *f*-CaO in clinker is a parameter in the cement industry that is related to the quality of the pyro process:

- Optimal value for *f*-CaO is between 1.0 and 1.5% in mass;

- Below the value of 1.0%, the clinker is overburned and more thermal energy than necessary was consumed;

- Above value of 1.5%, the clinker is unburned, not achieving the quality requirement and has to be discarded.

The control of optimal *f*-CaO content in clinker is important to avoid the waste of thermal energy in both undesired scenarios (over burned and unburned clinker), to maintain the required quality in clinker, and for the relative ease to perform the *f*- CaO chemical analysis. However, the *f*-CaO chemical analysis is a bottleneck in the cement industry for two reasons: (1) the average residence time in the kiln system (around 1 hour); (2) The frequency of *f*-CaO chemical analysis in the cement industry that varies between 1 and 2 hours.

As a result, the kiln operator has to wait the time between analyses to be sure if the change in the kiln operational parameters to correct an off-standard *f*-CaO had the expected effect. The delay time between the events will decrease substantially with the modelling of one effective soft sensor to predict *f*-CaO in clinker. This will, consequently, decrease the thermal energy consumption in the process. In the current days, this time is on the scale of hours. With a good soft sensor, this scale can achieve the range of minutes.

2 OBJECTIVE

The main objective is developing an effective soft sensor to predict *f*-CaO in clinker based on optimal feature selection using industrial data that describes the system in the real world and demonstrates the importance of the optimal feature selection from the kiln system which impacts the *f*-CaO in clinker, interpreting the meaning of variables from a phenomenological point of view.

The specific objectives are:

1. Pre-processing six months of industrial data which included more than one hundred quality and operational variables, and applying process knowledge as a tool for decision-making;

2. Applying a multiple linear regression algorithm (MLR) combined with forward stepwise methodology for a feature selection;

3. Interpreting the feature selection from a phenomenological point of view that included concepts related to thermodynamics, fluid dynamics, heat transfer, chemical kinetics, chemical heterogeneous reactions, and other concepts related to chemical engineering science;

4. Modelling the soft sensor to predict *f*-CaO in clinker using the regression models: multiple linear regression (MLR) and multivariate polynomial regression (PolR);

5. Modelling the soft sensor to predict *f*-CaO in clinker with machine learning algorithms: Multi-Layer Perceptron (MLP); Extreme Gradient Boosting (XGBoost), Gradient Boosting on Decision Trees (CatBoost); Random Decision Forests (RDF) and Support Vector Machine (SVM);

6. Comparing, analysing, and interpreting the performance of the prediction models with statistical and phenomenological approaches.

3 LITERATURE REVIEW

The literature review covers and discusses topics such as the mathematical models, feature selection methods and all necessary key points to understand the pyro process system, including the flowsheet description, chemical reactions, modules, proprieties, kinetic, thermodynamics considerations, also heat transfer and fluid dynamics concepts involved in the process. This interdisciplinarity is essential for understanding the features that describe the complex system.

3.1 Mathematical models for chemical engineering systems

Modelling a chemical engineering system, such as a clinker kiln system, with the physical and chemical laws governing the process is complicated. As discussed in the next topics, there are many phenomena, in this dynamic system, such as heat, mass, and momentum transfer, additionally many chemical reactions, phase transition, multiphase flow in a no-isothermal system with complex interactions.

Mathematical models applied in this kind of system could be classified in some categories as linear *versus* nonlinear, steady-state *versus* non-steady-state, lumped parameters *versus* distributed parameters, continuous *versus* discrete variables, deterministic *versus* stochastic and so on. However, an essential classification is if the model is mechanistic, empirical, or semi-empirical (RASMUSON et al., 2014).

As described by Rasmuson et al. (2014), “mechanistic means that models are based on the underlying physics and chemistry governing the behaviour of a process; empirical means that models are based on correlated experimental data. Empirical modelling depends on the availability of process data, whereas mechanistic modelling does not; however, a fundamental understanding of the physics and chemistry of the process is required. Mechanistic models are preferably used in process design, whereas empirical models can be used when only trends are needed, such as in-process control. Semi-empirical models cover the range in-between”. In any case, the physics and chemistry involved in experimental data must not be ignored.

3.1.1 Feature selection

There are more than one hundred features that described the kiln system, variables such as temperatures, pressures, speeds, flows, powers, chemical parameters and so on. The first step is identifying all input variables to achieve the best prediction of the output variable when any mathematical model is developed. There are many methodologies to find the best set of variables that describe the phenomenon being modelled, and they can be classified into two groups:

1. *General empirical models*: Also named black-box models. Features are selected by statical models that do not consider prior and deep knowledge about the physical and chemical laws governing the process. In this methodology, a strong correlation between features can be found, and it does not mean causation. As cited by Altman and Krzywinski (2015), “Correlation implies association, but not causation. Conversely, causation implies association, but no correlation.”, in the same article, the authors gave the example: “suppose we observe that people who daily drink more than 4 cups of coffee have a decreased chance of developing skin cancer. This does not necessarily mean that coffee confers resistance to cancer; one alternative explanation would be that people who drink a lot of coffee work indoors for long hours and thus have little exposure to the sun, a known risk”. Consequentially, misunderstandings about any process and/or phenomena can occur when this kind of methodology is applied.

2. *Mechanistic and empirical models applied to chemical engineering systems*: according to Rasmuson et al. (2014), in the mechanistic model, the governing equations, as mass and heat balances, are formulated even if there is not sufficient knowledge or computer resources to solve the equations. The features are not selected as a prior. This approach does not apply in the current study, due to the high complexity, and it is normally used for kiln system design and audit. The second methodology, empirical, used by experienced engineers and correlated professionals, is to list all variables that are believed to be important according to the deep knowledge about the process. The final model is then obtained by experimenting and model fitting using experimental data to reach a reliable result in the model.

A semi-empirical model, or hybrid model, covers the range in-between mechanistic and empirical models applied to chemical engineering systems.

In this work, the *empirical models applied to chemical engineering systems* is used to select a good set of variables to model the soft sensor. The goal is not only to have a mathematical model to predict the *f*-CaO at clinker, but also to understand the meaning of features based on a mechanistic point of view, which included the concepts involved in phenomes as heat transfer, chemical kinetics, thermodynamics and so on.

3.1.2 Multiple linear regression for feature selection

The high-dimensional data that described the kiln system is an issue that must be handled for modelling the system. Additionally, a central premise to any mathematical model is to select independent and relevant features as inputs. Many statistical algorithms were developed with this goal. Algorithms based on Filter Feature Selection adopt the strategy to select the sub-set of features INDEPENDENT of supervised learning model is used some examples are Information Gain Attribute Ranking and Consistency-based Feature Selection. For the Wrapper Methods, various sub-set of features are tested into the supervised learning model and the sub-set of features with high-performance is adopted in the final model, Recursive Feature Elimination is an elaborate algorithm with this approach. Finally, there are the Embedded Methods, where the feature selection is an INTEGRATED PART of supervised learning model, some algorithms that follow this technique are Lasso Regression, Gradient Boosting, and many others (BOMMERT et al., 2019).

Simple and Multiple Linear Regression (MLR) are the most studied and understood statistical models. The concept of linear regression was proposed by Sir Francis Galton in 1894, ever since has been largely used in all science fields. The first proposal for any empirical or semi-empirical model is fitting the available data with linear regression. Optimized algorithms with powerful statistical tools to minimize residuals with good stability and low computational costs are available on common software, as MATLAB, Statistica™ and libraries for programming languages such as as Python, R (MAULUD; ABDULAZEEZ, 2020) (OZGUR et al., 2017). As a result, linear regression is a supervised learning model widely applied to find the best subset of features, as explained previously, it is a Wrapper Method. There are two approaches to select the subset of features that will be tested in MLR: Forward and Backward

Stepwise. At Forward Stepwise, the model starts with zero feature (null model) and the most statistically significant features are added one by one until a stopping criterion is achieved or all features are added. On the other hand, at Backward Stepwise, the model starts with all features (Full Model) and the least statistically significant features are removed one by one until a stopping criterion is achieved or all features are removing (KHAIRE; DHANALAKSHMI, 2022).

Additionally, the effects of non-linear features can be investigated by a linearization process before the application of MLR. In engineering systems, the multiplication of continuous variables to analyse the cross effect or consider the inverse value of features on MLR are common linearization approaches. Linear equations are easy to solve and have low computational costs, so, with some simple mathematical manipulations, non-linear effects can be analysed in a faster and more effective way when MLR is applied in combination with the linearization process (RASMUSON, et al., 2014) (HASTIE; TIBSHIRANI; FRIEDMAN, 2017).

3.1.3 Machine learning algorithms for prediction models

Machine learning (ML) is the field of Artificial Intelligence (AI) responsible to develop methodologies for ‘machine learning’ in a similar way to humans. In the engineering field, ML is related to the development of robust statistical models to predict output variable that the formulation by the usual way, based on phenomenological laws, is not possible due to the complexity of the phenomenon itself (RASMUSON et al., 2014). ML algorithms are divided into many types and groups. The choice of what technique has to be applied depends on the problem. Normally, for engineering applications, the variables are correlated by chemical and physics laws, they are continuous and structured, and the prediction model is a goal, considering these points, some algorithms with application in the industry's real problems are:

1. *Multiple Linear Regression (MLR) and Polynomial Regression (PolR)*: the polynomial regression can be as simple as an MLR since PolR models are linear in the coefficients (b_i) (NIQUINI et al., 2019). It is the most well-known technique; the over-fitting issue does not occur, and it is good for continuous values as output when there are relationships between variables (RAY, 2019) (SARKER, 2021);

2. *Multi-Layer Perceptron (MLP)* is a class of Artificial Neural Networks (ANN). The logistic function is applied to estimate the probability of the event occurring or not based on inputs. Logistic regression algorithms, such as MLP, are predominantly used in industrial problems (RAY, 2019);

3. *Extreme Gradient Boosting (XGBoost), Gradient Boosting on Decision Trees (CatBoost) and Random Decision Forests (RDF)* are algorithms based on building a series of individual models, typically decision trees, that the approach is splitting the data continuously according to specific parameters. These algorithms use the gradient to minimize the loss function and are robust models from a statistical point of view. On the other hand, overfitting is a common issue and defining the optimal values for the hyperparameters is challenging when the models are applied to engineering problems;

4. *Support Vector Machine (SVM)* applies the kernel functions to find hyperplanes which separate the data according to the class. It is a good approach for non-linear problems and high-dimensional data. The overfitting phenomenon is minimized. Disadvantages include a poor performance with large datasets due to the difficulty finding the correct kernel function. It is hard to understand the final model and does not work well when the dataset has noisy (HASTIE; TIBSHIRANI; FRIEDMAN, 2017) (RAY, 2019) (SARKER, 2021).

3.2 Pyroprocessing system – clinker formation

In the next topics, an overview of the process for cement production in an industrial scale and the main chemical reactions involved in the clinker formation are presented.

3.2.1 Process flowsheet

The process for cement production on the industrial scale can be divided into six fundamental steps, as described below:

1. *Quarries and crushing plant:* They are the first steps in cement production. Rocks are mined and crushed (in a primary crusher) where the granulometry is reduced by a ratio of 4-8 and raw materials with granulometry below 30 mm (95%) are produced. The most important rock is limestone, since $CaCO_3$ is the main component and source of CaO. It is obtained from sedimentary rocks and composes ≅ 87% (mass basis) of raw material used in the following cement production step. The second component is clay, source of Al_2O_3 and SiO_2. It is essentially hydrated aluminium silicate, whose origin is sedimentary rock decomposition (≅ 11%, mass basis) (Holcim Group Support Ltd, 2008) (PCA-Portland Cement Association, 2004);

2. *Raw material storage and pre-blending*: Correctives are stored in common piles. They are components used to correct the chemical composition of the raw meal (final material mix). Usually, sand is used to correct the SiO_2 amount (≅1,5%, mass basis) and iron ore, as a source of Fe_2O_3, to correct the amount of this component in the raw meal (≅ 0,5%, mass basis). The storage of limestone and clay is a little more sophisticated, which occurs in longitudinal or circular piles that intercalate the material in horizontal layers to increase the homogenization. As a consequence, there is a reduction in chemical fluctuations and standard deviation for quality parameters. For example, $CaCO_3$ content in the limestone and Al_2O_3/ SiO_2 content in clay. The layers have different shapes depending on the method. They form triangles in the Chevron Method or rhombuses in the Windrow Method. Limestone and clay are reclaimed in a vertical direction to

increase the homogenization effect (Holcim Group Support Ltd, 2008) (PCA-Portland Cement Association, 2004);

3. *Raw meal grinding and homogenization silo*: Limestone, clay, and correctives (iron ore and sand) are ground and dried in a mill which can be horizontal (ball mill) or vertical mill (more recent technologies). The mix of these materials are known as “raw meal” and has a determined chemical composition and granulometry, between 12-18% residue 170# (90 μm). The raw meal is stored in a specific silo, known as homogenization silo, where a second blending step occurs. The homogenization silo is designed with an air aeration system, mix chamber, and different technological strategies to mix the raw meal and significantly decrease the LSF (limestone saturation factor) standard deviation in it. The blending effect, that is defined as the ratio between LSF standard deviation inlet and outlet the silo, can achieve values around 6 in the system that is running without any issues. As a result, an LSF standard deviation < 1 is obtained in the raw meal silo outlet, which is an excellent value for a good and stable operation in the kiln system, which is the next step in the process (Holcim Group Support Ltd, 2008);

4. *Clinker formation in the pyro process system*: The most important step in the cement industry. The pyro process system could be defined in 3 stages. The first is in the preheater tower with calciner, where the raw meal is fed. Firstly, the residual water that remains in the material is dried, then the decarbonatization happens. The temperature profile is between 250 and 900°C in this stage. The hot meal, common name for the raw meal after the first step, enters in the rotary kiln and the second stage starts, where the main chemical reactions to form the clinker crystals occur. In the kiln hood, the temperature profile achieves the highest value, around 1.450°C for hot clinker and higher than that in the flame, the maximum temperature depending on the fuel used. For the reactions to occur, the necessary energy is supplied from fuel combustion that happens in the main burner, located in the kiln hood and in additional and smaller burners in the calciner. The last step consists in cooling the hot clinker with ambient air in the cooler. In a modern cement plant, the final

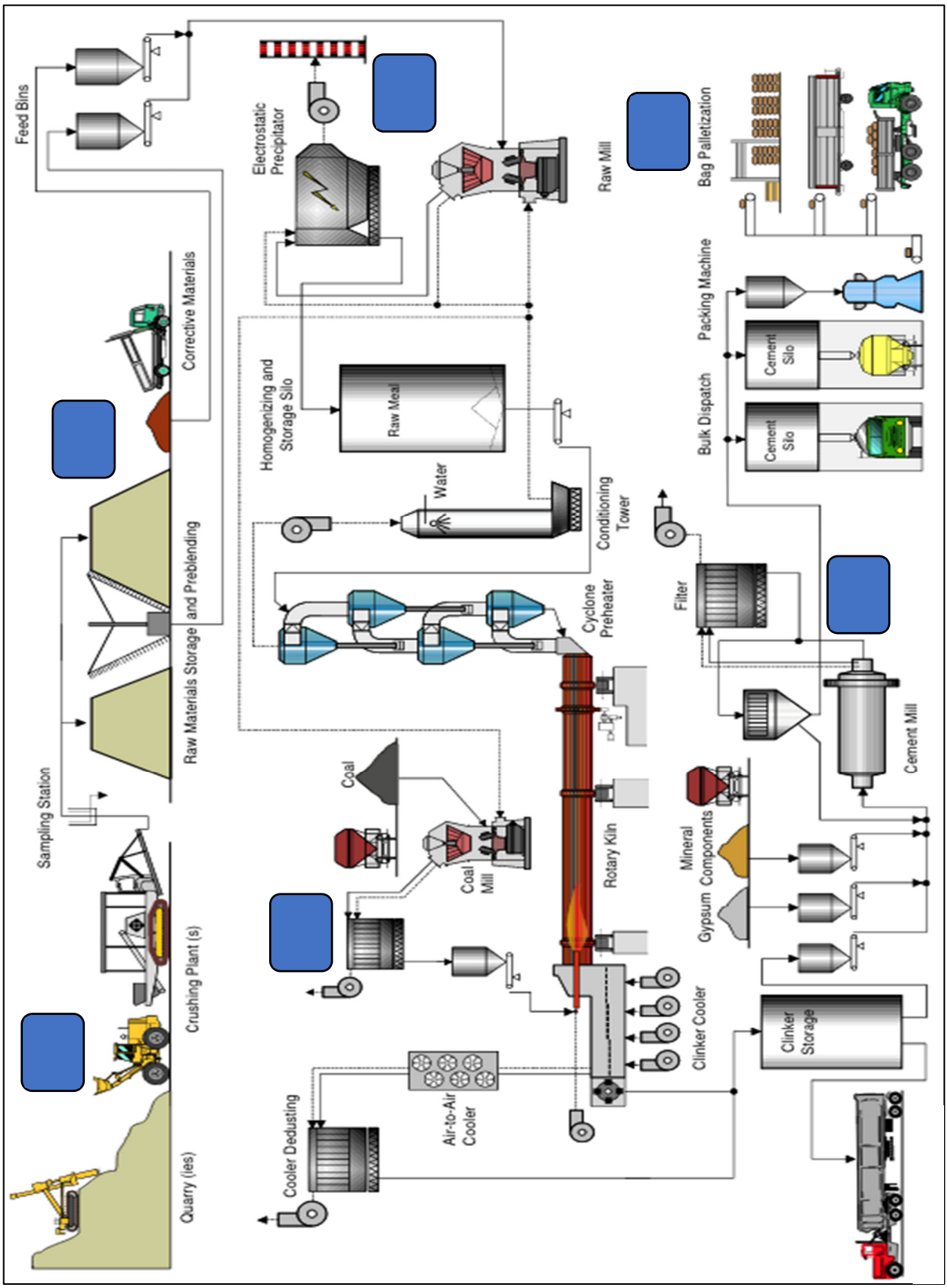
clinker temperature is around 95°C with high-efficiency cooler equipment. Finally, clinker is stored in standard silos (PCA-Portland Cement Association, 2004);

5. *Cement grinding:* Clinker is ground with additional components (gypsum, limestone, pozzolana) to produce a large range of cement types, *cf.* Table 2. Cement ball mills are the most common ones, but in the last decade, high-pressure roller press (HPGR) and vertical mills for cement have begun to be a technological option to decrease specific energy consumption and maintain cement quality (ALI; SAIDUR; HOSSAIN, 2011). Cement is store in silos;

6. *Cement dispatch*: Cement is consumed in two forms, the first one is bulk dispatch in trucks, which is the best option for cement consumption in the big building projects (roads, stadiums, bridges, wind farms). The second option is cement bags. In this case, cement is packed in packing machines and stored in this form. Usually, this option is for the public that buys cement in the building supply store or similar (Holcim Group Support Ltd, 2008).

The cement process flowsheet with these six main steps identified is in Figure 3.

Figure 3 - Cement process flowsheet.



Source: (Holcim Group Support Ltd, 2008).

3.2.2 Chemical reactions

Formation of clinker minerals occurs in many steps, in a multiphase system with a considerable temperature range and with intermediate product formation. Some important features of this system are:

- Temperature range: 20 to 1450 ºC;

- Many distinct reaction types: decomposition, heterogeneous reactions (solid-gas, solid-solid, solid-liquid, solid-solid-liquid);

- The first chemical reaction that takes place in the system, limestone decarbonisation, occurs in a gas-solid system, with the contact of the raw meal with hot gas from fuel combustion; Other chemical reactions occur in a heterogeneous system (liquid phase + solid) and the surface of this system also has contact and heat exchange with the hot gas from combustion;

- The most important intermediate products are metakaolinit, calcium silicate ($CaO.SiO_2$) and calcium aluminate ($CaO.Al_2O_3$) (TELSCHOW, 2012).

Chemical reactions involved in the clinker mineral formation with respective temperatures and brief descriptions are in Table 4.

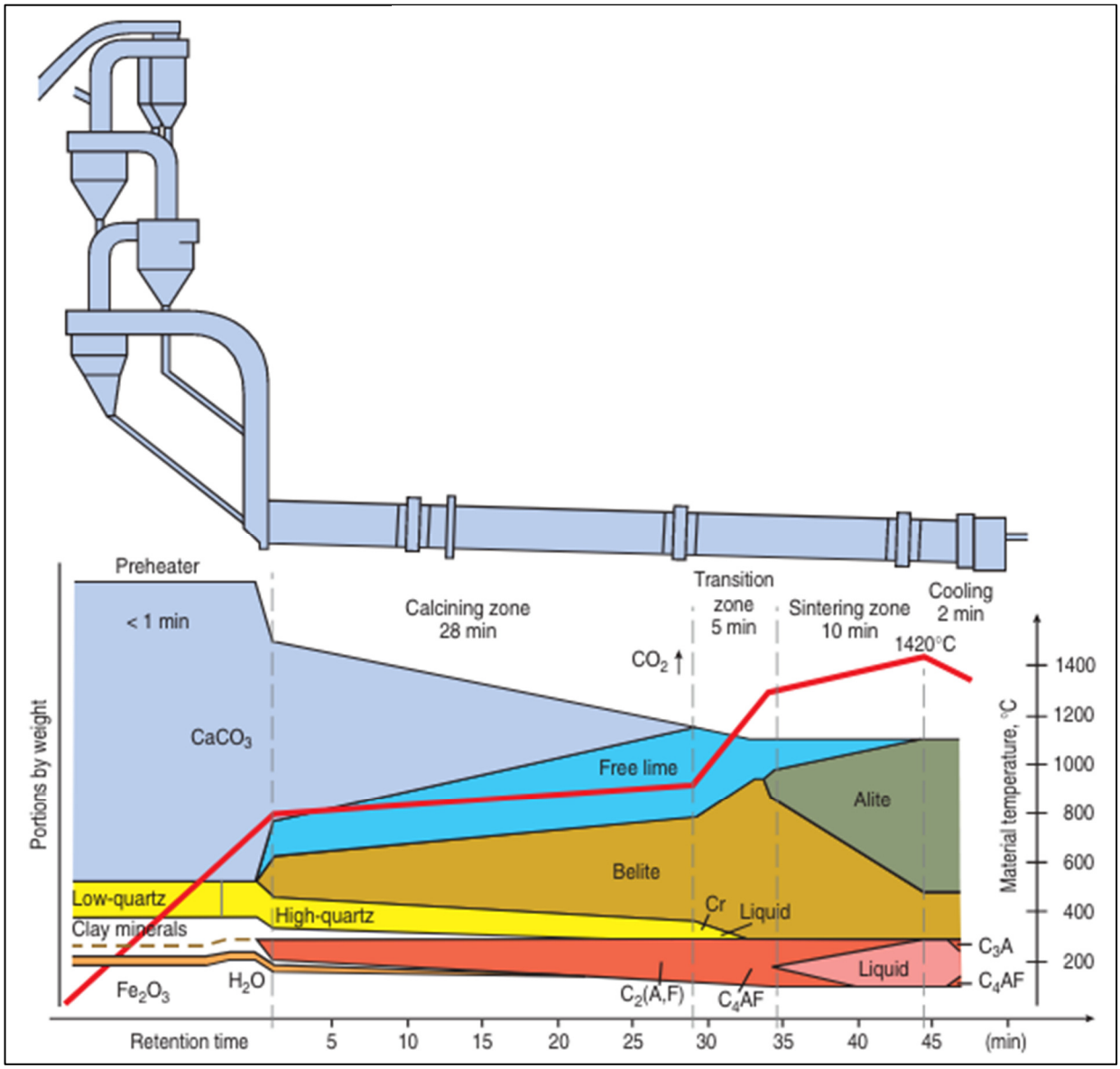
Table 4 - Chemical reactions involved in the clinker minerals formation with respective temperatures and kinetic considerations.

Temp. (ºC)	Process	Chemical reaction	Reaction/ kinetic notes
20 - 200	Drying	$H_2O \uparrow$	-
200 - 450	Elimination of adsorbed water	$H_2O \uparrow$	-
450 - 600	Decomposition of clay, formation of metakaolinit	$Al_4(OH)_8Si_4O_{10} \rightarrow 2(Al_2O_3.2SiO_2) + 4H_2O$	Decomposition
600 - 950	Decomposition of metakaolinit	$Al_2O_3.2SiO_2 \rightarrow Al_2O_3 + 2SiO_2$	Decomposition; Liquid phase formation
800 - 1000	Limestone decomposition, CS and CA formation	$CaCO_3 \rightarrow CaO + CO_2$ $3CaO + 2SiO_2 + Al_2O_3 \rightarrow 2(CaO.SiO_2) + CaO.Al_2O_3$	Decarbonation; heterogeneous reaction: solid (CaO) / liquid phase (Al_2O_3 + Fe_2O_3)
1000 - 1300	Formation of clinquer minerals	$CaO.SiO_2 + CaO \rightarrow 2CaO.SiO_2$ $CaO.Al_2O_3 + 2CaO \rightarrow 3CaO.Al_2O_3$ $CaO.Al_2O_3 + 3CaO + Fe_2O_3 \rightarrow 4CaO.Al_2O_3.Fe_2O_3$	Heterogeneous reactions: solid-solid (C_2S, C_3A formation), solid-solid-liquid (C_4AF formation), all reactions occur in the liquid phase
1300 - 1450	C_3S formation	$2CaO.SiO_2 + CaO \rightarrow 3CaO.SiO_2$	Heterogeneous reaction: solid-solid (C_3S formation), reaction occur in the liquid phase

Source(adapted): (PCA-Portland Cement Association, 2004) (TELSCHOW, 2012).

A schematic drawing is in Figure 4. The red line is the temperature profile with the values in the y-axis, on the left. The qualitative portion of the weight to the main species is also in the y-axis, on the right. The approximate retention time is shown in the x-axis. Finally, the main species is also identified, and the graph information is correlated with kiln system zones (calcining zone, transition zone, sintering zone, and cooling). It is important to observe the multiphase behaviour of the system.

Figure 4 - Transformation of raw meal to clinker. Red line shows the temperature profile.



Source: (PCA-Portland Cement Association, 2004).

3.3 Thermodynamics considerations about chemical reactions to form clinker

When the chemical reactions to form clinker on an industrial scale are analysed carefully, a lot of important concepts from distinct areas of science have to be taken into account to understand the phenomena as a whole. From thermodynamics science, two of them are the heat of formation and lower heat value (LHV) from fuels that are used in kiln systems to obtain the necessary heat to make the reactions possible.

Heat of reaction (HF), also known as enthalpy of formation, is the enthalpy change when 1 mol of compound is formed at standard state (25°C, 1 atm) from its constituting elements in their standard state. When this amount of energy is taken out of the system, it is an exothermic reaction (-), and when this amount of energy is absorbed by the system, it is an endothermic reaction (+).

If the compound is formed through multiple steps, as happens in the kiln system, the HF is the sum of the enthalpy change in each process step (BASU, 2018).

A summary of endothermic and exothermic reactions that happen in the kiln system are in Table 5.

Table 5 - Endothermic and exothermic reactions that occur in kiln system.

Endothermic reactions (25ºC)	**kJ/kg**
Dehydration of clays	170
Calcination reaction	1990
Heat of melting	105
Heating raw meal (1450°C)	2050
Sub total	**+4315**
Exothermic reactions (25ºC)	**kJ/kg**
Reaction related to the crystallization dehydrated clay	-40
Heat of formation, reactions related to clinker minerals formation	-420
Crystallization of melt	-105
Cooling of clinker	-1400
Cooling of CO_2	-500
Cooling of water	-85
Sub total	**-2550**
Clinker - heat formation (≈)	**1765 (420 kcal)**

Source: (PCA-Portland Cement Association, 2004).

The heat for clinker minerals formation is around 1765 kJ/kg (420 kcal/kg) and can be estimated by Equation (1) (PCA-Portland Cement Association, 2004).

$$\Delta HF_{clinquer} = 2{,}22 \times \%Al_2O_3 + 5{,}86 \times \%h_H + 6{,}48\%MgO + 7{,}646\,\%CaO - 5{,}116\%SiO_2 - 0{,}59\%Fe_2O_3 \quad (1)$$

Note: If alkalis are present, the heat of formation is reduced by about 2 kcal/kg.

Where:

$\Delta HF_{clinquer}$ = heat of formation for clinker at 25ºC, [kcal/kg];

$\%h_H$ = water of crystallization in the clay, [g/100g clinker];

%SiO_2, %Al_2O_3, %Fe_2O_3, $\%MgO$,$\%CaO$ = oxides values obtained from clinker x-ray fluorescence analysis, [%mass];

The heat required for clinker formation increases with the amount of liquid phase, burning zone temperature and C_3S. On the other hand, the heat decreases by increasing C_4AF. HF can achieve 1900 kJ/kg (MOHAMED et al., 2018). It is important to note that the total amount of energy spent in the process, total heat consumption, including the heat losses due to inefficiencies like radiation, convection, heat losses by gas, clinker and dust depend on the kiln system technology, as described in Table 1. Consequently, the temperature profile in the system is also changed according to the heat of formation and system technology (RODRIGUES et al., 2017).

Another important concept from thermodynamics that appears when the kiln system is analysed is the low heating value (LHV). LHV is defined as the heat produced by the combustion of a substance at a constant pressure of 0.1 MPa (1 atm), with any water formed remaining as vapour. The values of LHV for distinct materials, fuels in the cement industry, are obtained by an adiabatic bomb calorimeter as a device. The procedure for it has to follow recognized international standards for reaching confidence values to LHV for distinct fuels (ASTM D5865/D5865M, 2019).

In the cement industry, the energetic matrix is very varied, can be solid fuels (pet coke, coal), liquid fuels (oils), gas (natural gas) and alternative fuels that wide range included biomass, liquid waste materials, tyres, every kind of plastics and so on. For a specific kiln system, it is normal to use a combination of different fuels as the energetic matrix. Consequently, the LHV values in a kiln system are inconstant and unstable. LHV values fluctuations are common in cement plant operational routine due to different fuels adopted and fluctuations also occur for the same fuel type. It is a

quality parameter controlled by the quality department in each cement plant (ZIERI; ISMAIL, 2019).

3.4 Kinetic considerations about chemical reactions to form clinker

The main kinetic parameters involved in the chemical reactions to form clinker are discussed in the next topics.

3.4.1 Clinker crystal description and properties

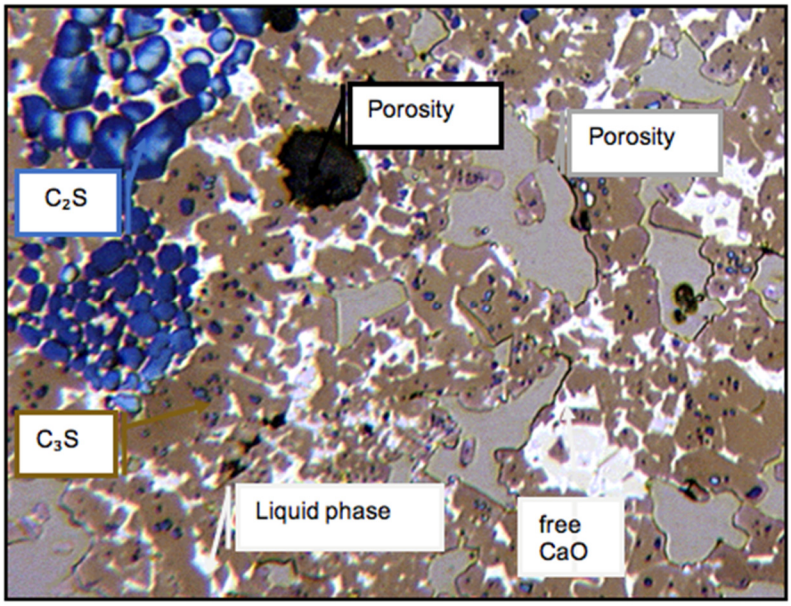
Clinker is a multiphase mixture, but a degree of homogeneity is expected for good and predictable properties also in a multiphase system. In microscopy, the clinker phases are identified for different colours under acid attack (nitric acid alcohol or acetic acid, CH_3COOH) (THEISEN, 2010).

Clusters of *f*-CaO and belite are easily identified; however, the brown alite crystals are the most abundant. The colours that identify each phase are in Table 6. A clinker microscopy example, with phases identified, is in Figure 7. The microscopy for a homogeneous and desirable clinker is in Figure 6. The clinker with higher homogeneity degree is hard to be obtained in the industrial conditions, but a certain degree of homogeneity is achieved.

Table 6 - Colours for clinker crystals identification in the microscopy.

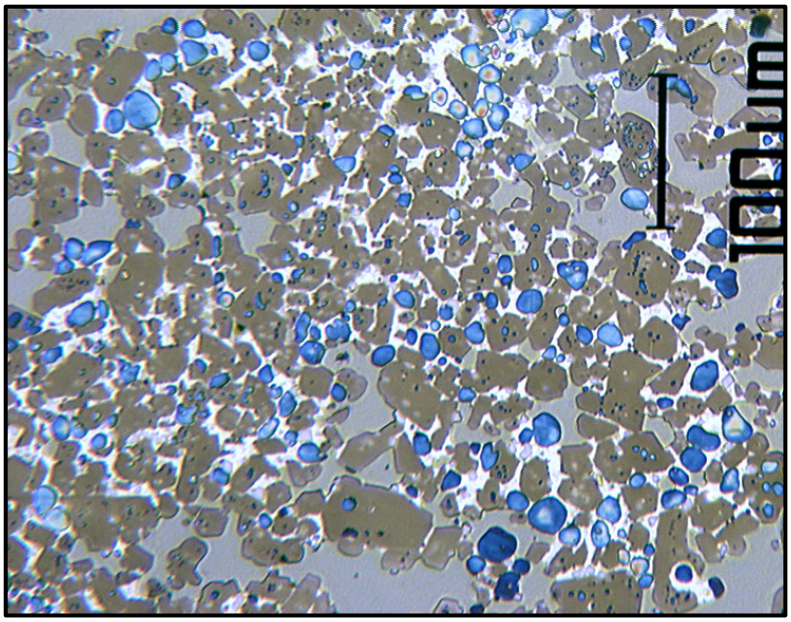
Colour	Phase	Symbol
Brown/grey	Alite	C_3S
Blue	Belite	C_2S
White/Grey	Ferrite/Aluminate	C_4AF/C_3A
White round	F-CaO	CaO
Dark grey	Alkali sulphates	$(Na/K)_2SO_4$
Black/grey	Porosity	
Grey between clinker pieces	Porosity	

Figure 5 - Clinker microscopy with phases identified.



Source: (THEISEN, 2010).

Figure 6 - Microscopy of clinker with high homogeneity degree.



Source: (THEISEN, 2010).

Alite (C_3S) is the most important clinker component, responsible for the initial cement resistance. With a high value for heat hydration, this component is formed in a very high temperature (> 1300°C), but it is also unstable at high temperature (<1275°C). The ideal crystal size is small, because it is more reactive and easier to grind in the cement mills. The crystal size depends on parameters such as liquid phase viscosity and amount, raw meal granulometry, temperature profile and retention time in the kiln

sintering zone, *cf.* Figure 4 (TELSCHOW, 2012) (PCA-Portland Cement Association, 2004) (HEWLETT, 2006).

Belite (C_2S) is the second most important clinker component. With a low heat hydration value, it has a low development rate for initial cement resistance, but has the same final cement resistance than alite. On the other hand, the belite grindability is much higher than alite, which could have a prejudicial impact in the cement mills. It is hard to identify belite crystals in x-ray diffraction and the crystals have a spherical shape, *cf.* Figure 7 and Figure 8. (TELSCHOW, 2012) (PCA-Portland Cement Association, 2004) (HEWLETT, 2006).

Ferrite (C_4AF) is the first component to form. It also has a low heat hydration value and is responsible for cement stability under chemical attack due to the ferrite capacity to incorporate alkalis between three to five times more than alite and belite. Ferrite does not have any contribution to cement mechanical resistance and with a fast crystallization process, a higher percentage of this component in clinker increases the total clinker grindability value. (TELSCHOW, 2012) (PCA-Portland Cement Association, 2004) (HEWLETT, 2006).

Aluminate (C_3A) is the component responsible for the early cement setting times. As alite, it has a high hydration heat value and contributes for a high initial cement resistance rate; however, it is susceptible to alkaline chemical attack that modify the shape of the crystals (cubic to orthorhombic to monoclinic) and, consequently the alkalis modify the aluminate reactivity. It is an important parameter to optimize the gypsum percentage used in the cement grinding process (TELSCHOW, 2012) (PCA-Portland Cement Association, 2004) (HEWLETT, 2006).

f-CaO is related to the combustion process quality in the clinker formation. The unreacted oxide calcium is a quality parameter controlled in the cement industry. The ideal value is between 1.0 and 1.5% on a mass basis. It is also responsible for the expansibility in the concrete when catalysing the C_2S hydration reaction in it. The expansibility in concrete can cause undesired fissures in it (TELSCHOW, 2012) (PCA-Portland Cement Association, 2004) (HEWLETT, 2006).

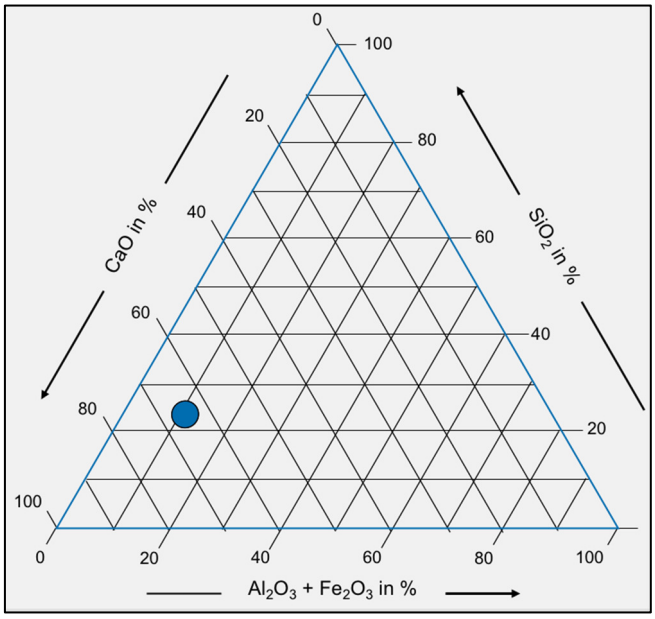
Minor components as MgO, alkalis (Na_2O, K_2O, chlorides) and sulfur are also present in the raw meal/clinker composition. The impact of these components varies enormously. A high value for MgO impacts cement quality, high values for sulfur and alkalis also impact kiln operational stability and final cement quality. The raw meal and clinker composition are quality parameters controlled in the cement industry routines (TELSCHOW, 2012) (PCA-Portland Cement Association, 2004) (HEWLETT, 2006).

3.4.2 Chemical modules

The chemical modules can be described as the practical approach to control the chemical composition in the necessary steps to produce clinker.

The concept is used to define the raw meal recipe, the amount of each material that has to be fed in the raw mill. The modules are ratios of main oxides required to produce clinker (CaO, Fe_2O_3, Al_2O_3 and SiO_2). These ratio and module values are based on the ternary diagram of the oxides and the region where clinker is formed. The ternary diagram with the small region where clinker exists highlighted is in Figure 7 (HEWLETT, 2006).

Figure 7 - Tertiary diagram for oxides involved in the clinker formation. Region where clinker exists is highlighted with blue colour.



Source: (HEWLETT, 2006).

There are three modules that define raw meal and clinker composition: Lime Saturation Factor (LSF), Alumina Ratio (AR) and Silica Modulus (SM).

Lime Saturation Factor (LSF) is the ratio between CaO and the other three oxides (Fe_2O_3, Al_2O_3 and SiO_2). A raw meal with high LSF value is more difficult to burn (high burnability index). Consequently, it leads to a high *f*-CaO value in the clinker, which induces instability in the cement hydration volume. On the other hand, a low LSF value negatively affects cement resistance. The formula for LSF, in raw meal basis, is in Equation (2).

$$LSF = \frac{100 \times Ca0}{2{,}80 \times .SiO_2 + 1{,}18 \times Al_2O_3 + 0{,}65 \times Fe_2O_3} \quad (2)$$

Where:
LSF= lime saturation factor, [-];
CaO, SiO_2, Al_2O_3 and Fe_2O_3 = oxide values obtained from raw meal x-ray fluorescence analysis, [%mass];

Silica Modulus (SM), *cf.* Equation (3), is the ratio between SiO_2 and (Fe_2O_3 + Al_2O_3). It is important to highlight that the metal oxides Fe_2O_3, Al_2O_3 are responsible to form the liquid phase, where the calcium silicates reactions take place. A high value for SM, as a high value for LSF, is associated with high burnability index, and similar issues. For a low SM value, the issue is the increase in the risk to build a coating in the kiln, which causes operational instabilities.

$$MS = \frac{SiO_2}{Al_2O_3 + Fe_2O_3} \quad (3)$$

Where:
SM= Silica Modulus, [-];
SiO_2, Al_2O_3 and Fe_2O_3 = oxide values obtained from raw meal x-ray fluorescence analysis, [% mass];

Alumina Ratio (AR) is the ratio between Al_2O_3 and Fe_2O_3, *cf.* Equation (4). It is related to the liquid phase properties: for a high AR value, the liquid phase amount is higher for the same temperature reference. The liquid phase dependence with Al_2O_3 and Fe_2O_3 present in the raw meal is in (5) (HEWLETT, 2006).

$$AR = \frac{Al_2O_3}{Fe_2O_3} \quad (4)$$

$$\%FL_{1450°C} = 3{,}00 \times Al_2O_3 + 2{,}25 \times Fe_2O_3 + MgO + Na_2O + K_2O + SO_3 \quad (5)$$

Where:

AR= Alumina Ratio, [-];

$\%FL_{1450°C}$ = % liquid phase at 1450°C, [%mass];

MgO, Al_2O_3 , Fe_2O_3, Na_2O,SO_3 and K_2O = oxide values obtained from raw meal x-ray fluorescence analysis, [% mass]. Alkalis are impurities present in the raw meal.

The reference values for Lime Saturation Factor (LSF), Alumina Ratio (AR), Silica Modulus (SM) and liquid phase are in Table 7.

Table 7 - Modulus reference values for a desire raw meal.

Parameter	Unit	Range or limit	Standard values
Lime Saturation Factor	-	90 - 105	95 - 98
Silica Modulus	-	1.8 - 3.9	2.2 - 2.6
Alumina Ratio	-	1.5 - 2.9	1.6 - 2.0
Liquid phase	%mass basis	-	24 - 28

Source: (HEWLETT, 2006).

3.4.3 Temperature profile and cooling effect in the C_3S crystal formation

As described in Table 4, the chemical reactions to produce clinker occur in a heterogeneous phase (liquid and solid). Because of that, the effect of pressure in the system can be disregarded in a chemical kinetic point of view. The flowing gas in the system is from the combustion process and participates in the heat exchange steps, but, not in the chemical reactions. Consequently, it is possible to simplify and say that the temperature profile is directly interconnected with the combustion gas flow in the system; however, this gas flow and pressure profile are parameters that do not have a direct role in chemical reactions in liquid and solid systems. (TELSCHOW, 2012).

The kiln system can be considered a non-isothermal, non-homogeneous and non-ideal reactor. The 5 main regions with respective temperature profile which are identified in this non-ideal reactor are in Table 8.

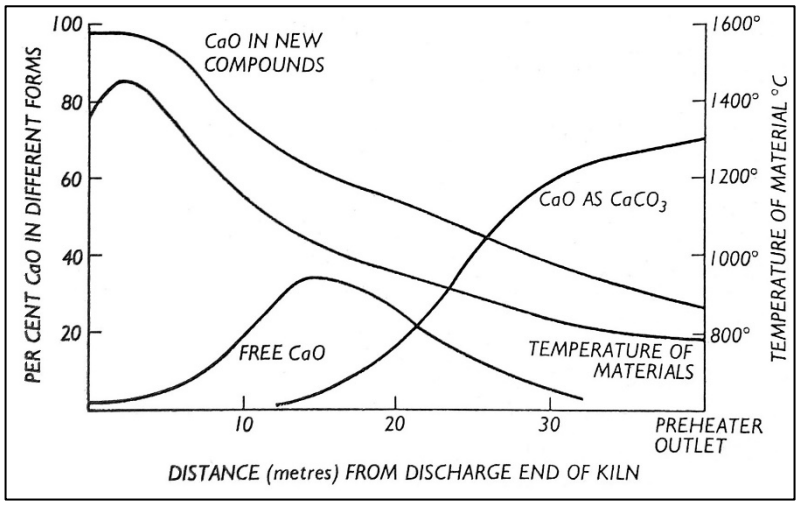
Table 8 - Temperature profile description for kiln system based on Figure 4.

Region	Temperature (℃)	Temperature profile notes
Preheater	~20 -850	Positive linear temperature profile
Calcining zone	~850 - 950	Approximately constant
Transition zone	~950 - 1250	Positive linear temperature profile
Sintering zone	~1250 - 1420	Positive linear temperature profile
Cooling zone	~1420 - 1300	Negative linear temperature profile

Source: author.

The calcium component (as carbonate, *f*-CaO or new components) dependence with temperature and position in the kiln system is in Figure 8. It is important to note that 0% of *f*-CaO is not desired.

Figure 8 - Calcium components in a dry-process kiln (40 m) with preheater.



Source: (HEWLETT, 2006).

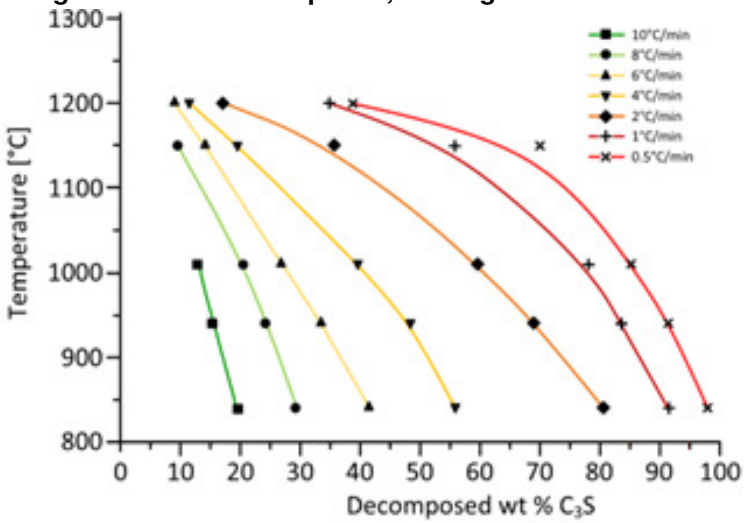
As described before, the alite crystal, C_3S, is formed at very high temperatures (>1300°C). However, this crystal is unstable between 1000 and 1275 °C, in this temperature interval, the alite is decomposed into belite and *f*-CaO as in Equation (6).

$$C_3S \rightarrow C_2S + CaO\ free \quad (6)$$

The %C_3S in clinker is directly related to *f*-CaO. High *f*-CaO values lower C_3S content in the clinker. One reason for the high *f*-CaO value in clinker is the decomposition of C_3S that occurs when the operational parameters in the kiln are not optimal. Time is a primordial factor in this reaction; when the clinker is cooled quickly, the decomposition reaction is interrupted. Therefore, this is the main strategy in the cement industry to maintain the C_3S amount formed in the sintering zone in clinker.

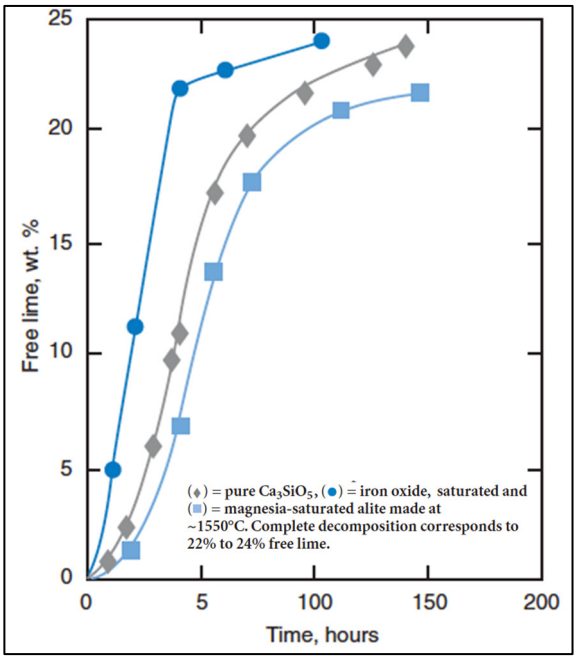
The equipment responsible for the fast clinker cooling is the high-efficiency cooler (PCA-Portland Cement Association, 2004) (HEWLETT, 2006). The relationship between C_3S content, temperature, cooling rate and *f*-CaO is in Figure 9. Time influence at C_3S decomposition is in Figure 10, where pure alite is the grey curve and the blue with square symbols curve is for magnesia- saturated alite.

Figure 9 - Alite decomposed, cooling from 1200 ℃ to 870 ℃.



Source:(TENÓRIO, et al., 2008).

Figure 10 - Decomposition of pure and doped alite in steam, 1 bar total pressure, at 1055℃.



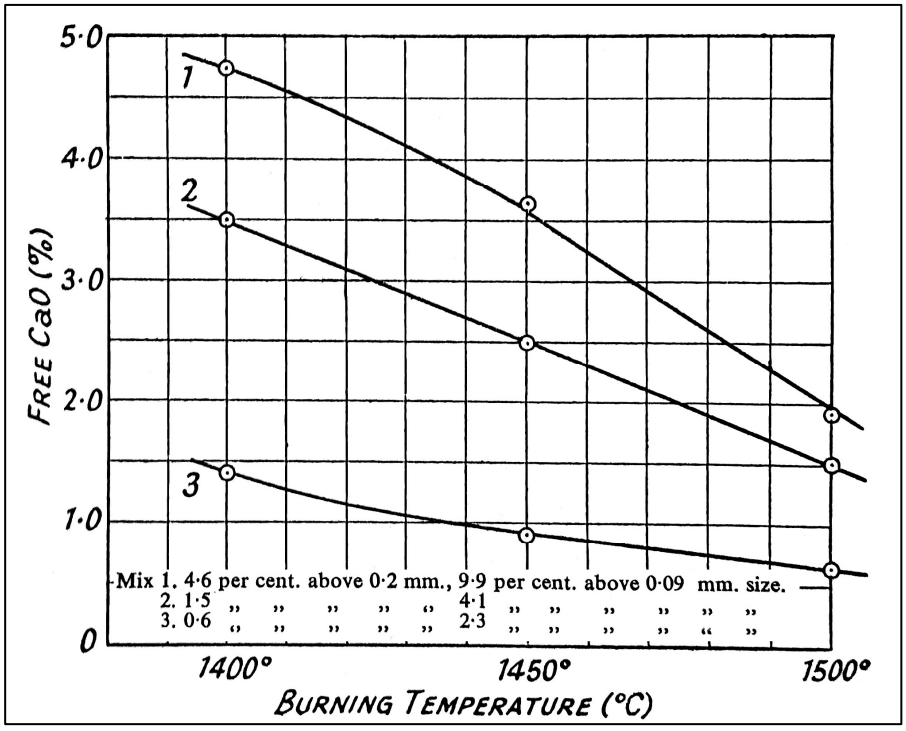
Source: (PCA-Portland Cement Association, 2004).

3.4.4 Raw meal granulometry

For chemical reactions with solid material as a reagent, the contact surface is a crucial parameter in the system. It is not different for clinkerization reactions. In the cement industry, the contact surface is controlled by raw meal granulometry. Fine raw meal favours the chemical reaction. Unfortunately, too fine material has to be avoided because it increases the risk to form a coating, mainly in the preheater (Holcim Group Support Ltd, 2008).

The standard values for raw meal granulometry are between 12 and 18% residue in 90 microns and between 1 and 3% residue in 200 microns. The influence of granulometry in the raw meal burning process is exemplified in Figure 11. There are 3 curves with distinct granulometry for a limestone-clay mix. Curve 1 is the coarse material and has a higher value for *f*-CaO (~4.75%) at 1400°C. In this same temperature, curve 3 from the finest material has a lower value for *f*-CaO (<1.5%) (HEWLETT, 2006).

Figure 11 - Burning of pure hard limestone-clay mix.



Source: (HEWLETT, 2006).

3.4.5 Burnability

Burnability is a concept for 'how difficult it is to burn a specific raw meal'. In other words, a high burnability value means more energy to achieve the same *f*-CaO content in clinker compared with a low burnability value. In the kiln system, high burnability means a higher temperature profile, for the same *f*-CaO content in the clinker. Consequently, the specific heat consumption is also higher for high burnability values.

The raw meal chemical composition and granulometry are the factors that define the raw meal burnability. There are many chemical tests in the literature to define raw meal burnability. Burnability Index (BI), *cf.* Equation (7), and Burnability Factor (BF), *cf.* Equation (8), are two of them (PCA-Portland Cement Association, 2004).

$$BI = \frac{C_3S}{C_4AF + C_3A} \tag{7}$$

$$BF = \text{LSF} + 10 \times SM - 3 \times \left(MgO + (Na_2O + 0{,}658 \times K_2O)\right) \tag{8}$$

Where:
BI = burnability index, [-];
BF = burnability factor, [-];
C_3S, C_4AF and C_3A = clinker crystals obtained by x-ray diffraction analysis for clinker, [%mass];
LSF = lime saturation factor, [-];
SM = silica modulus, [-];
MgO, Na_2O, K_2O = oxides values obtained from clinker x-ray fluorescence analysis, [%mass];

As mentioned before, in section *3.4.2 Chemical modules*, the LSF and SM modules have a direct impact on the raw meal burnability, which is evidenced by Equations (7) and (8).The Burnability Index dependence with raw meal granulometry (% residue in 90 micros) is not taken into account in Equations (7) and (8).

3.4.6 Retention time in the kiln system and distribution of residence times

Retention time is another important factor that cannot be ignored in the kiln system dynamic. The total retention time in the kiln system is the sum of three steps: retention time in the preheater tower + retention time in the rotary kiln + retention time in the cooler.

As can be observed in Figure 4, the retention time in the preheater is less than 1 minute, so, it can be ignored when the entire system is analysed. For the rotary kiln, the analysis is more complicated, since the total retention time is influenced by the viscosity of the liquid phase that is complex to incorporate in a mathematical model. A lot of equations for the rotary kiln retention time were proposed, with most of them based on empirical models and a good review for it is carried out by Ndiaye et al (2010). For one example, there is the kiln retention time in Equation (9) that is commonly used in the cement industry, it is also an empirical model (PCA-Portland Cement Association, 2004).

$$t_{kiln} = \frac{0{,}19 \times L}{Di \times n \times \emptyset} \tag{9}$$

Where:

t_{kiln} = retention time in the rotary kiln, [min];

L = rotary kiln length, [m];

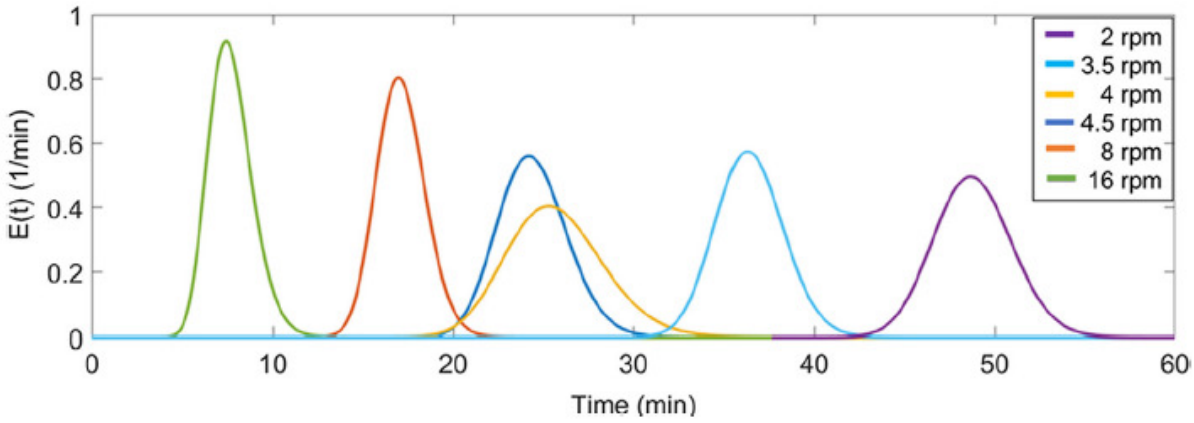
Di = rotary kiln internal diameter, [m];

n = kiln speed, [rpm];

∅= kiln slope, [%m/m];

Moreover, the empirical models for retention time in the rotary kiln consider that it is an ideal reactor and can be classified as PFR (continuous tubular reactor) since this value is constant (FOGLER, 2016). In fact, for real reactors, there is a residence time distribution (RTD), and the same is applied to cement rotary kilns. Experimental measurement of the residence time distribution for rotary kiln was conducted by Paredes et al. (2018), one of the results is the residence time distribution at various kiln speeds, as can be seen in Figure 12. The current study was carried out at ambient temperature.

Figure 12 - Residence time distributions at various rotation rates.



Source: (PAREDES et al., 2018).

For grate coolers, the retention time calculation, *cf.* Equation (10), is simpler than for rotary kilns and is based on common reactors (FOGLER, 2016). Again, it is considered an ideal reactor, the value obtained by Equation (10) must be interpreted as an average value.

$$t_{cooler} = \frac{60 \times A \times H \times \rho_{cl}}{Pr_{cl}} \quad (10)$$

Where:

t_{cooler} = cooler retention time, [min];

A= total grate area, [m^2];

H= clinker height at grate, [m];

ρ_{cl}= clinker apparently density, [t/m^3];

Pr_{cl}= clinker production, [tph];

The total retention time average in the kiln system (rotary kiln retention time + cooler retention time) is ≈ 40 to 60 minutes for most modern systems.

3.4.7 Kiln filling

Kiln filling is the volumetric percentage that material occupies in the rotary kiln, and it influences the material mix during the burning process, how well and efficient the mix is between phases (solid and liquid). Another impact of kiln filling is in heat exchange between combustion gas and material surface exposed to it. Finally, it has

a role in the clinker nodulation process. The kiln filling percentage is calculated by Equation (11) (Holcim Group Support Ltd, 2008).

$$\%F = \frac{3{,}2 \times Pr_{cl}}{Di^3 \times \emptyset \times n} \quad (11)$$

Where:

%F= kiln filling, [%volume];

Pr_{cl}= clinker production, [tph];

Di = rotary kiln internal diameter, [m];

n = kiln speed, [rpm];

∅= kiln slope, [%m/m];

3.5 Fluid and heat dynamics considerations about chemical reactions to form clinker

As mentioned before, the heat consumed in the kiln system is supplied by fuels when they are burned. Hot gases are the product of the combustion reaction, and the heat is transferred from the hot gases to the system (kiln walls, cyclone tower, calciner, raw material, etc.) by mechanisms as radiation, conduction, and convection.

The relationship between gas flow, temperature and heat is highlighted in Equation (12). The heat is the key parameter, but, the day by day, it is not the value available in real-time to evaluate the kiln system. Moreover, it is not the parameter that guides make decisions at operation. A case study energy audit for pyro-processing is executed and discussed by Ghalandari, Majd and Golestanian (2019).

$$Q = \dot{m}.C_{pgas} \cdot (T_{gas} - T_0) \quad (12)$$

Where:

Q = sensible heat from hot gases, [kJ/hour];

$\dot{m}$ = hot gas flow, [Nm^3/hour];

C_{pgas} = specific heat capacity at constant pressure, [kJ/ Nm^3.ºC];

T_{gas} = gas temperature, [ºC].

T_0 = gas at reference temperature, 20 ºC, [ºC];

There are 3 variables related to heat at Equation (12). Specific heat capacity at constant pressure (C_{pgas}), is a thermodynamic property, and it is calculated when the heat balance of the system is carried out. The other two variables (temperature and gas flow) are present in the guides for kiln system operation. The temperature profile at the system is measured directly by thermocouples as PT100 or pyrometers. Gas flow and airflow (in the case of cooling fans) are a little more complicated. There are direct measurements at flows in some strategic points, an example is the piezometric rings for flow measurements at cooling fans, but flow measurements do not occur for all fans. Additionally, the confidence of flow measured by piezometric rings is a weak point in the cement industry, the calibration is complex, sometimes hard to be performed. In general, piezometric rings indication is good to follow tendencies in the kiln operation, but it is not ideal to exact values of flow. The alternative to engineers and operators to evaluate the flow profile is the indirect indication, with other parameters correlated with it, such as pressure, fan power and fan speed. Fan laws make the connection between these variables, as can be seen in Equations (13) to (15) (FOX et al., 2015) (Axair Fans UK Ltd, 2022).

The First Fan law is related to volume of air. Volumetric flow rate is directly proportional to the ratio of the rotational speed of the impeller, *cf.* Equation (13). The Second Fan Law is about the pressure, it describes the relationship between the pressure developed by the fan and the impeller rotational speed. Pressure is proportional to the square to the ratio of the rotational speed of the impeller, *cf.* Equation (14). Finally, the third law provides the required power when the impeller speed changes. The cubic nature of this relationship between power and the impeller rotational speed shows how even for small speed increase, large amounts of additional power are needed, *cf.* Equation (15) (FOX et al., 2015) (Axair Fans UK Ltd, 2022).

$$\dot{m}_2 = \left(\frac{U_2}{U_1}\right) \times \dot{m}_1 \tag{13}$$

$$P_2 = \left(\frac{U_2}{U_1}\right)^2 \times P_1 \tag{14}$$

$$kW_2 = \left(\frac{U_2}{U_1}\right)^3 \times kW_1 \tag{15}$$

Where:

$\dot{m}_1$= initial volume of air or gas, [m³/hour];

$\dot{m}_2$ = new volume of air or gas, [m³/hour];

U_1 = initial fan impeller rotation, [rpm];

U_2 = new fan impeller rotation, [rpm];

P_1 = initial pressure, [mbar];

P_2 = new pressure, [mbar];

kW_1 = initial power required by the fan, [kW];

kW_2 = new power required by the fan, [kW];

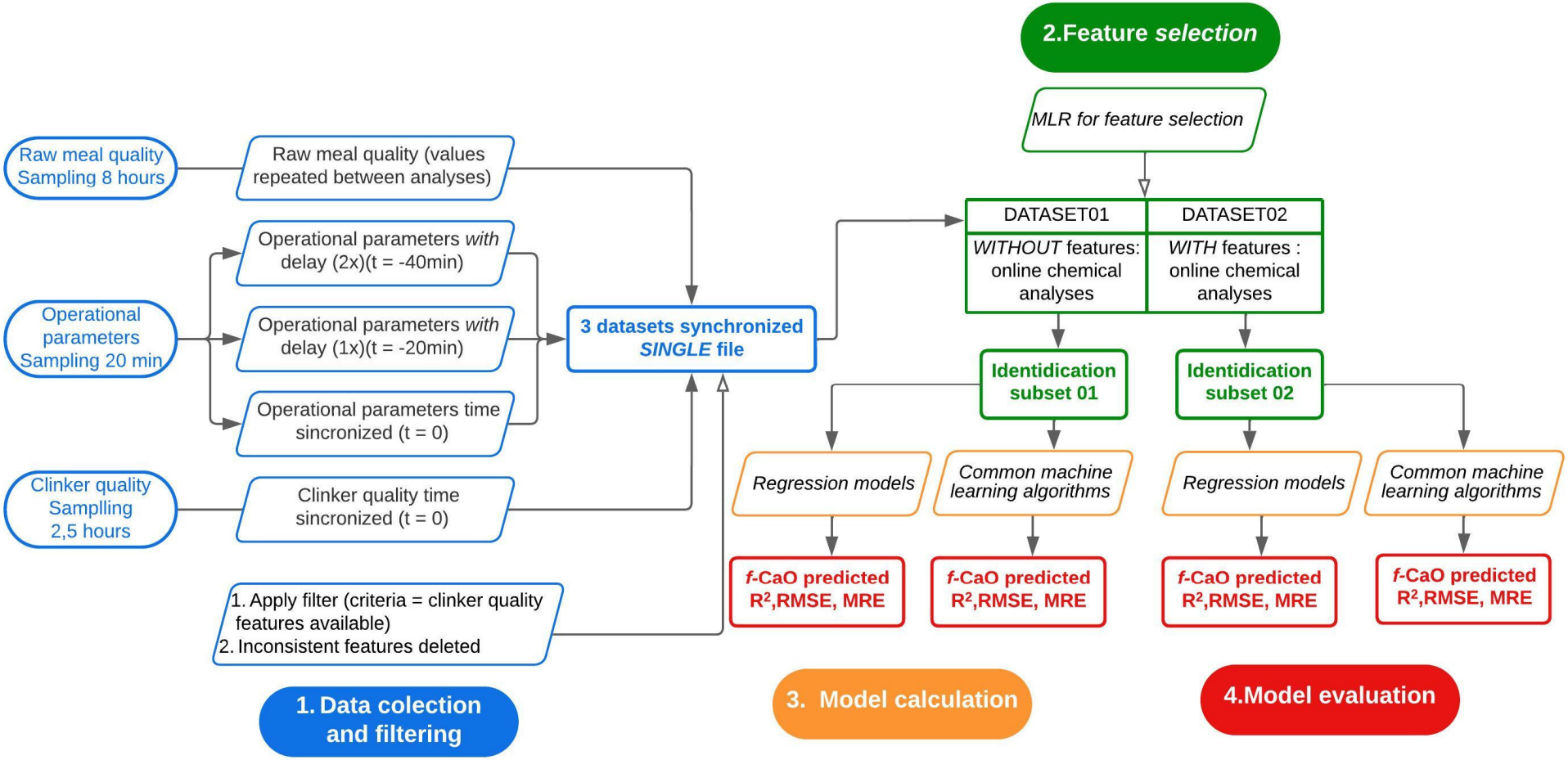
In summary, temperatures, pressures together with fan power and fan speed provide the pieces of information about the heat transfer process in the kiln system. As Equation (12) makes clear, it is not possible only to consider the temperatures, since the total heat involved in the system is directly dependent on both, the temperature and flow, so, any heat analysis of the system must evaluate operational parameters related to both (temperatures, pressures, fan power and fan speed). Temperatures and pressures are measured directly by thermocouples/pyrometers and manometers respectively. Fan power and fan speed values are also available at common cement plants.

4 METODOLOGY

Industrial data were investigated with a combination of deep knowledge of system and statical tools to determine the optimal feature set between operational and quality parameters that impact the *f*-CaO at clinker; with the features sub-set defined, the second step was applying the machine learning techniques in the data to develop a soft sensor to predict *f*-CaO in the clinker.

MATLAB, Microsoft Excel® software, codes using C++ language, libraries written for the Python language as pandas, NumPy, and statsmodels were used to implement the data analysis, which includes the pre-processing data, MLR, and most used machine learning algorithms to prediction models (RAY, 2019) (SARKER, 2021). The methodology is summarized in the flowchart presented in Figure 13; description of each step is in the coming topics, 4.1 to 4.3.1.

Figure 13 - Flowchart with the main steps in the development of a soft sensor to predict *f*-CaO in the clinker.

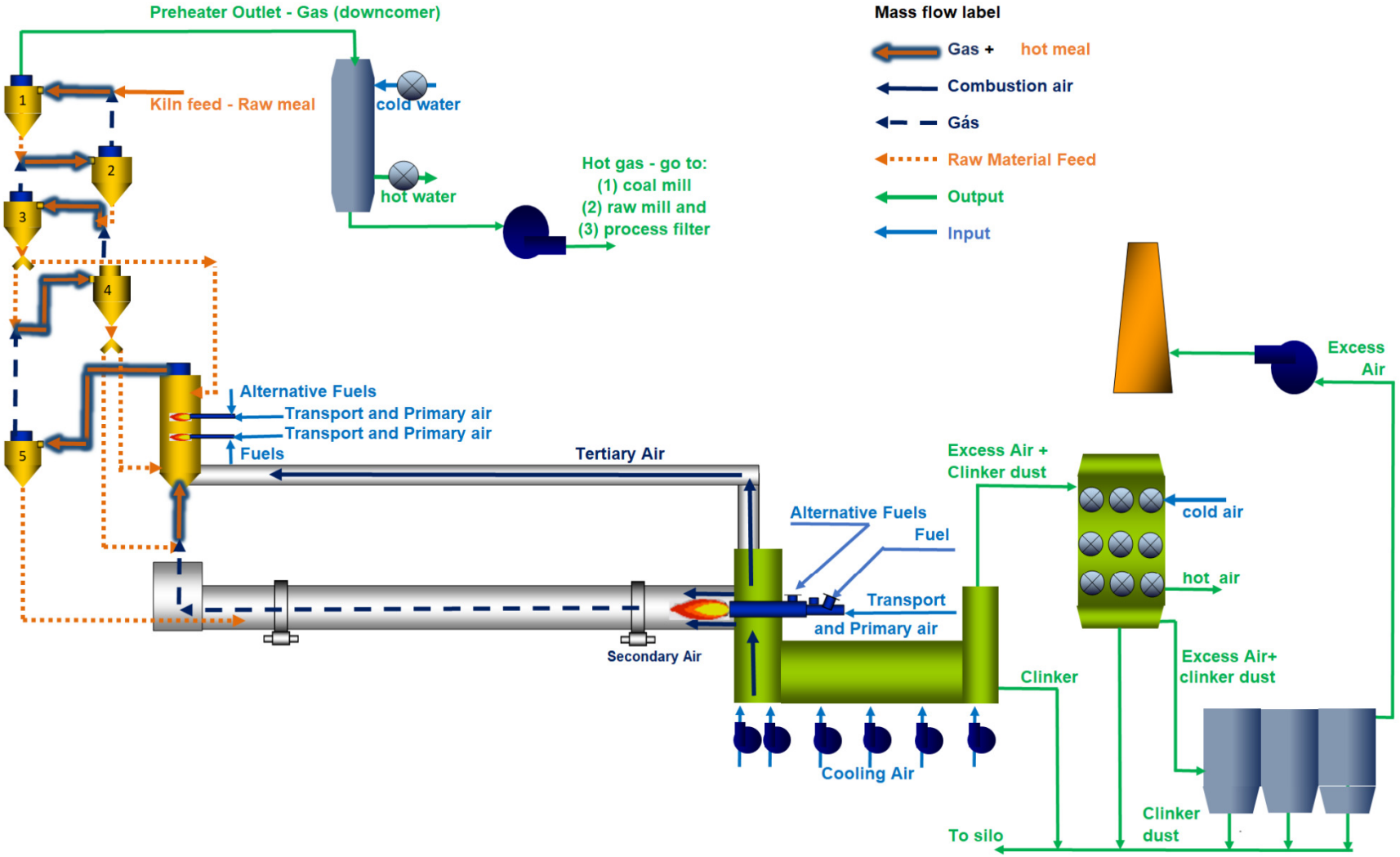


Source: author

4.1 Industrial data description and pre - processing data

The data used in this work came from a Portland Cement Plant with a modern pyro-process line supplied by FLSmidth which began operation in 2015 and is located in the south of Brazil. The kiln system is designed for 3.000 tpd of clinker, guarantee value, and consists of a preheater with 5 cyclones, an ILC calciner with 6,0 m diameter x 30,0 m length, a ROTAX-2 rotary kiln with 4,35m diameter x 51,0 m length. The gas comes from the preheater outlet (downcomer) is cooled at the gas conditioning tower before the ID fan. Hot clinker is colling at Cross-Bar 10x40 cooler with 67,8 m^2 aerated grate area. The excess air from the cooler is conditioned at the air-air heat exchanger before going to the bag filter. The fuel matrix is diverse, with coke as the main fuel and distinct alternative fuels for the main burner and calciner burners. A schematic mass flowsheet with inputs and outputs is in Figure 14 .

Figure 14 - Schematic flowsheet for kiln system considered in this work with mass flow highlighted.



Source: author.

The raw data consists of 6 months of data (1st July to 27th December/2021) in 3 distinct data sets, one for raw meal quality control, the other for kiln operational parameters and the last one for clinker quality control. Continuous values are available only for operational features.

The data sets were joined in a unique file using a code implemented in MATLAB. The two criteria for joining the data sets were frequency and residence time at the system, calculated by Equations (9) and (10).

For the raw meal, the frequency for analysis is one by shift (approximately every 8 hours) with a compound sample. It was not necessary to add any delay and the values repeated between analyses. Operational parameters were sampled as the mean value every 20 minutes. An interval smaller than 20 minutes for operational features is not a reasonable choice since the frequency of the clinker sampling is ≈2.5 hours with a punctual sample. The clinker quality values were synchronized with operational parameters considering retention time (kiln + cooler) for each sample.

Additionally, operational parameters *with* 20 and *with* 40 minutes of delay were considered 'new variables' and synchronized with clinker quality features accordingly.

After joining the three datasets in a single file, inconsistent features were deleted, and knowledge about kiln operation was the basis for decision-making. The filter for the final dataset was the clinker quality features since there is no clinker sampling if the kiln system fails. This approach based on process knowledge is not usual in the literature, which normally applies the filter using statistical tools. (LIU et al., 2019) (ZHAO, 2021) (LI; WANG; CHAI, 2015). Then, considering the timeline, previous clinker quality features, were input together with the time series sample.

In the last step of pre-processing, two datasets were generated, due to around 50% of the data having the continuous values of raw meal quality available (online chemical analyses). The first dataset, DATASET01, is with all data *WITHOUT* the features of the online chemical analyses. The second dataset, DATASET02, is the data *WITH* features of the online chemical analyses available.

4.2 Multiple linear regression for feature selection

For each dataset various simulations were carried out with MLR combined with the forward stepwise methodology to select the feature set. The first simulation considered the initial features set without mathematical manipulation of the dataset before applying the MLR. The second simulation considered the initial features set *plus* operational parameters with 20 minutes of delay, and the next simulation considered the previous one *plus* operational parameters with 40 minutes delay. This approach is based on the concept of residence time distribution, it is an attempt to measure this effect. Moreover, it carried out simulations considering at least all statistically significant features found in the previous simulations *plus* selected features.

The statistical relevance of individual features was decided by the null-hypothesis significance testing, where if the p-value is lower than 0.05, the feature influences the *f-CaO* in the clinker. To evaluate if the variable is statistically relevant the p-value for each variable, in each simulation, can be obtained from the results calculated by Equations (16) and (17). The correlation coefficient is calculated by Equation (16) and it is an input for test statistic (t_i^*) in the t-test given by Equation (17); the t_i^* value has t-Student distribution with ns–2 degrees of freedom; r_i is the metric responsible measuring the linear dependence between y_i (*f*-CaO) and x_i (process variables); p-value of t_i^* is then compared to the p-value established for the test, the feature influences the *f*-CaO in the clinker if p-value is lower than 0.05 (HASTIE; TIBSHIRANI; FRIEDMAN, 2017) (NIQUINI et al., 2019).

$$r_i = \frac{\sum_{j=1}^{ns}\left(x_{i,j} - \bar{x_i}\right)\cdot\left(y_j - \bar{y}\right)}{\sqrt{\sum_{j=1}^{ns}\left(x_{i,j} - \bar{x_i}\right)^2}\cdot\sqrt{\sum_{j=1}^{ns}\left(y_j - \bar{y}\right)^2}} \tag{16}$$

$$t_i^* = \frac{r_i\cdot\sqrt{ns-2}}{\sqrt{1-r_i^2}} \tag{17}$$

Where:

r_i = correlation coefficient for each variable with index *I*, [-];

t_i^* = test statistic for the variable with index *i* , [-];

ns = number of sampling points, [-];

$x_{i,j}$, y_j = measured value of variables x_i and y with index *j*;

$\bar{x_i}$, $\bar{y}$ = average values for the variables.

4.3 Machine learning algorithms for prediction *f*-CaO in clinker

Initially, data were split randomly into two sets: 80% for the algorithm's training, and the remaining 20% used for the model´s test; timeline was not a criterion. The same algorithms were applied for DATASET01 and DATASET02. An additional simulation, with data split in 85% for training and 15% for testing, was realised in the DATASET02, with the fourth-order polynomial model.

Three distinct metrics were applied to evaluate the models: two metrics are about the errors, root mean squared error (RMSE) and mean relative error (MRE), given by Equations (18) and (19). The third metric is the coefficient of determination (R^2) given by Equation (20), the closer R^2 is to 1, the better the outcome predicted by the model.

$$RMSE = \sqrt{\frac{1}{ns} \cdot \sum_{i=1}^{ns} (y_i - \hat{y}_i)^2} \tag{18}$$

$$MRE = \frac{1}{ns} \cdot \sum_{i=1}^{ns} \frac{|y_i - \hat{y}_i|}{y_i} \tag{19}$$

Where:

$RMSE$ = root mean squared error, [-];

MRE = mean relative error, [%];

ns = number of sampling points, [-];

y_i = real value of model (*f*-CaO) with index *i*, [%mass];

$\overline{y}$ = average for real values (*f*-CaO), [%mass].

The overview of the main steps to develop the model to predict *f-CaO* is in Figure 13.

4.3.1 Regression models

The forward stepwise regression approach, which starts with a null model and adds variables one by one, was used for multivariate polynomial algorithms (first to fourth order), where MLR is a special case of a multivariate polynomial *with* a degree equal to 1. In the stepwise (forward-backward) method, a selected variable can be discarded. The metric for stopping or adding a new variable is the same, coefficient of determination (R^2), *cf.* Equation (20). For each new variable added, between all possibilities, was chosen to the one with the higher R^2 value for training and testing data. The stopping criterion was the value for R^2 in the test data. The algorithm stops adding variables when the R^2 of test data is not increasing anymore with an additional one; this criterion avoids the overfitting phenomenon. After that, the null-hypothesis significance testing, based on p-value , *cf.* Equations (16) and (17), was applied to define the statistical relevance of each variable, where the final number of features is represented by (*nf*).

$$R^2 = 1 - \frac{\sum_{i=1}^{ns}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{ns}(y_i - \bar{y})^2} \tag{20}$$

Where:

R^2 = coefficient of determination, [-];

$\hat{y}_i$ = predict value of model (*f*-CaO) with index *i,* [%mass];

ns = number of sampling points, [-];

y_i = real value of variable y_i (*f*-CaO) with index *i,* [%mass];

$\bar{y}$ = average for real values (*f*-CaO), [%mass];

Multivariate polynomials consider the multiplication of process variables (x_i) or their inverses to model the correlation between them and *f-CaO* (y). The general expression for polynomials is in Equation (21). It has to be highlighted that a polynomial of degree *m* can contain all possible functions w_i related to the degree *m* and the polynomials with a lower degree than *m*, *cf.* Table 9. Generic examples for the third and fourth order are in Equations (22) and (23). For all polynomials, the models are linear in the coefficients (b_i) and were calculated by MLR with the optimal line that fits the data; with low computational cost, the MLR methodology to estimate the coefficients does not have the disadvantages from other methods as errors associated

to the truncation of Taylor series or interactive algorithms (RASMUSON et al., 2014) (HASTIE; TIBSHIRANI; FRIEDMAN, 2017).

$$y = b_0 + b_1 \cdot w_1(\boldsymbol{x}) + b_2 \cdot w_2(\boldsymbol{x}) + \cdots + b_{ni} \cdot w_{ni}(\boldsymbol{x}) \tag{21}$$

Where:

y = content of *f*-CaO predict by multivariate polynomial model, [%mass];

b_{ni}= linear coefficients where *ni* is the number of regression variables *before* applying the null-hypothesis significance testing, [-];

w_i = functions of x_i and depend on the degree of adjusted polynomial, *cf.* Table 9.

Table 9 - w_i functions for each degree of polynomial.

Polynomial degree (m)	Possible w_i functions for each degree of polynomial
1	x_i, $(x_i)^{-1}$ for $\forall\ i \leq nv$
2	$x_i.x_j$, $x_i.(x_j)^{-1}$, $(x_i.x_j)^{-1}$ for $\forall\ i \leq nv$, $\forall\ j \leq nv$
3	$x_i.x_j.x_k$, $x_i.x_j.(x_k)^{-1}$, $x_i.(x_j.x_k)^{-1}$, $(x_i.x_j.x_k)^{-1}$ for $\forall\ i \leq nv$, $\forall\ j \leq nv$, $\forall\ k \leq nv$
4	$x_i.x_j.x_k.x_v$, $x_i.x_j.x_k.(x_v)^{-1}$, $x_i.x_j.(x_k.x_v)^{-1}$, $x_i.(x_j.x_k.x_v)^{-1}$,$(x_i.x_j.x_k.x_v)^{-1}$ for $\forall\ j \leq nv$, $\forall\ k \leq nv$, $\forall\ v \leq nv$

Source: elaborated by Esly Ferreira da Costa Junior.

$$y = b_0 + b_1 \cdot x_1 \cdot x_{10} + b_2 \cdot x_1 \cdot x_{20} \cdot x_{35} + \cdots + b_{20} \cdot x_7 \cdot x_{22} \cdot {x_{35}}^{-1} + \cdots + b_{70} \cdot (x_{33} \cdot x_{30})^{-1} \tag{22}$$

$$y = b_0 + b_1 \cdot x_3 + b_2 \cdot x_3 \cdot x_{22} \cdot x_{30} \cdot x_{31} + \cdots + b_{30} \cdot x_5 \cdot x_{15} \cdot (x_{31} \cdot x_{27})^{-1} + \cdots + b_{63} \cdot {x_{33}}^{-2} \cdot {x_{30}}^{-2} \tag{23}$$

4.3.2 Common machine learning algorithms for prediction models

The MLP, XGBoost, CatBoost, RDF and SVM are robust algorithms for non-linear problems. The codes were written in Python for each algorithm and the tuning process was manual. Three distinct metrics were applied to evaluate the models: coefficient of determination (R^2) given by Equation (20), the closer R^2 is to 1, the better the outcome predicted by the model. The other two metrics are about the errors, root mean squared error (RMSE) and mean relative error (MRE), Equations 24 and 25.

$$RMSE = \sqrt{\frac{1}{ns} \cdot \sum_{j=1}^{ns} \left(y_j - \hat{y}_j\right)^2} \quad (24)$$

$$MRE = \frac{1}{ns} \cdot \sum_{j=1}^{ns} \frac{\left|y_j - \hat{y}_j\right|}{y_j} \quad (25)$$

Where:

RMSE = root mean squared error, [-];

MRE = mean relative error, [%];

ns = number of sampling points, [-];

y_j = real value of model (*f*-CaO) with index j, [%mass];

$\hat{y}_j$ = predict value of model (*f*-CaO) with index j, [%mass];

The overview of the main steps to develop the model to predict *f*-CaO is in Figure 13.

5 RESULTS AND DISCUSSION

This chapter is split into four topics following the flowchart's steps, *cf* Figure 13: 5.1 presents the results related to the data collection and filtering; 5.2 discusses the results from feature selection step; 5.3 and 5.4 evaluate the results from regression and other models, respectively.

5.1 Pre – processing data

After cleaning the three original databases for the initial period analysed, 1st July to 27th December/2021, the useful data found on each database before joining data into a single file are:

1. Raw meal quality: 276 samples.
2. Kiln operational parameters: 12945 samples.
3. Clinker quality: 1453 samples.

A summary of raw data features before and after pre-processing data is in Table 10 by sub-systems and quality control. A complete list with individual features description (operational and quality) after pre-processing data is available in APPENDIX A, page 91. As a result, 04 features for raw meal quality (online analysis), 59 features for kiln operational, *without* delay, and 11 features for clinker quality control, including *f*-CaO, totalizing 70 and 74 features selected for the DATASET01 and DATASET02 respectively.

Table 10 - Features description by area before (raw databases) and *after* pre-processing.

Area Description	Features / Parameters	Unit	nº before	nº after	Note
Raw meal at kiln feed	Modulus (LSF, SM, AR)	-	3	0	Calculated. by Equations (3) to (5)
	Residue at 90 microns	%mass	1	0	Measured
	Chemical components	%mass	9	0	Measured
Quality control	Modulus (LSF, SM, AR) MgO. Online Analysis *FOR DATASET02	-	4*	4*	Calculated by Equations (3) to (5)
	Total - Kiln Feed Quality		**17**	**4**	
Preheater Tower	Temperatures	ºC	13	13	Measured
	Pressures	mbar	6	6	Measured
	Dampers - Open Position	%open	2	0	Measured
	Gas composition (O_2%; CO; SO_2; NO_2)	% or ppm	4	0	Measured
	Kiln feed	tph	1	0	Measured
	Fuel feed (coke + alternative fuels)	tph	2	0	Measured
Conditioning Tower + ID Fan	Temperatures	ºC	1	1	Measured
	Pressures	mbar	2	2	Measured
	Power (fans)	kW	1	1	Measured
	Speed (fans)	rpm	1	1	Measured
	Water flow	m^3/h	1	1	Measured
Rotary Kiln	Temperatures	ºC	1	1	Measured
	Pressures	mbar	2	2	Measured
	Gas composition (O_2%; CO; SO_2; NO_2)	% or ppm	4	0	Measured
	Fuel feed (coke + alternative fuels)	tph	2	0	Measured
	Electrical Current	A	2	1	Measured
	Kiln Speed	rpm	1	1	Measured
	Main Drive – Power	kW	2	1	Measured
	Filling Degree – Kiln	%	1	1	Calculated by Equation (11)
	Residence time – Kiln	min	1	1	Calculated by Equation (9)
Cooler + Excess Air Fan	Temperatures	ºC	2	2	Measured
	Pressures	mbar	7	7	Measured
	Power (fans)	kW	7	7	Measured
	Speed (fans)	rpm	7	7	Measured
	Electrical Current (fan)	%A	1	1	Measured
	Cooler Speed	strocks/min	1	1	Measured
	Residence time	min	-	1	Calculated by Equation (10)
	Dampers - Open Position	%open	1	0	Measured
	Air flow	m^3/s	6	0	Measured
	Total - Operational Parameters		**82**	**59**	
Clinker Quality control - Parameters	Modulus (LSF, SM, AR)	-	3	3	Calculated by Equations (3) to (5)
	Chemical components	%mass	10	0	Measured
	Total alkalis as Na_2O	-	1	1	Measured
	Liquid phase at 1450ºC	%mass	1	1	Calculated by Equation (2)
	Calcium crystals	%mass	4	4	Measured by X-ray crystallography
	Burnability Index	-	-	1	Calculated by Equation (7)
	Calcium Oxide	%mass	2	1	Measured
	Total - Clinker Quality		21	11	
Total features			**120**	**74**	

The reason for deleted features is listed below:

Modulus (LSF, SM, AR) and Residue at 90 microns: by previous simulations, it is noted that no variable for raw meal quality by laboratory appears at the feature selection. The explanation is about the sampling time being very long, approximately every 8 hours, against a retention time of ≈45 minutes at the whole system (kiln + cooler), so, it is not possible to capture the effect of granulometry or Modulus at *f*-CaO.

Chemical components : as explained on *3.4.2 Chemical modules*, page 42, raw meal and clinker composition are described by modules and not for each chemical component, so, it is not reasonable to include the oxides ($CaCO_3$, Fe_2O_3, Al_2O_3 and SiO_2) and minor components concentration at analysis.

Dampers: damper position (%open) features deleted did not change the position during the time considered in this work. The %open for these dampers had fixed values.

Gas composition (O_2%; CO; SO_2; NO_2): It is normal for standard gas analysers in the cement industry to have cleaning cycles. During these cycles, the gas composition values are equal to air, so, averages values for these features are not reliable. The frequency of these cycles is not fixed and is often related to other parameters, the pressure drop at the dust filters is one example. Moreover, it is relatively common for the online analyser to stay out of service for maintenance.

Kiln feed: the values are not available for most of the 6 months database analysed for feature selection. On the other hand, the kiln filling degree feature is available, and it is an indirect indication of the total amount of material at the system, *cf.* Equation (11).

Fuel feed flow (coke + alternative fuels): in the literature, the choice of fuel feed flow is common for empirical models to predict *f*-CaO (LIU et al., 2020). However, the choice of this feature is not indicated because the parameter which matters is the total energy supplied to the system with the combustion of the fuels, and, the energy value does not depend only on the fuel flow, the low heating value (LHV) is also an essential property, so, as was explained in *3.3 Thermodynamics considerations about chemical reactions to form clinker*, page 37, the LHV value is a quality control feature and is not

available as the other operational parameters. The features which are indirect indications about the amount of energy in the system and heat transfer phenomena are temperatures, pressures together with fan power and fan speed, as explained in *3.5 Fluid and heat dynamics considerations about chemical reactions to form clinker*, page 51.

Main drive: there are two drives responsible for the rotary kiln speed. For one of them, the electrical current value is not available, the instrumentation is out of service. The power at final selection is the two drives power sum.

Air flows at cooler : they are measured by piezometric rings, and, as mentioned in the topic *3.5 Fluid and heat dynamics considerations about chemical reactions to form clinker*, page 51, they are not reliable instruments, they are good to follow tendencies in the kiln operation, but they are not ideal to exact values of airflow, other features as fan speed, power and pressure are better choices.

5.2 Feature selection by multiple linear regression

The metrics for the relevant simulations, including the number of features before and after the selection by forward stepwise methodology, are in Table 11.

Table 11 - Number of features selection for each simulation - Initial Input and Final Input after selection by forward setwise as aproach.

Data set	**DATASET01 WITHOUT Online Quality**				**DATASET02 WITH Online Quality**			
Identification	**sim01**	**sim02**	**sim03**	**sim04**	**sim1**	**sim2**	**sim3**	**sim18**
Delay time	No Delay	20 min	20 and 40 min	20 and 40 min	No Delay	20 min	20 and 40 min	20 and 40 min
Sample size	1450	1450	1450	**1450**	671	671	671	**671**
Inputs (features)	70	131	191	**32**	74	139	203	**36**
Factor observation for each feature	20:1	11:1	7:1	**40:1**	9:1	5:1	3:1	**18:1**
Output (features)	24	25	32	**32**	20	27	-	**36**
Multiple R	0.74	0.75	0.76	0.76	0.79	0.81	-	0.82
Multiple R^2	0.54	0.56	0.57	**0.57**	0.62	0.66	-	**0.67**
Remark	-	-	-	**Selected**	-	-	-	**Selected**

The ratio between sample size and features is not unanimous in the statistic field. The widely cited two general rules suggest 5 or 10 minimum observations for each independent variable. The recommendation is a ratio of 50:1 if the stepwise methodology is applied (University of Cambridge, 2023). It was not possible to perform

the complete simulation with the delay times for the DATASET02, since the sample size was small, with a ratio of 3:1, *cf.* Table 11. Selected for the next steps, the **sim04** and **sim18** have 40 and 18 observations for each feature respectively. The sample size is not too small for them, but a larger sample is a considerable improvement for future research.

There are many papers about soft sensor models to predict *f*-CaO at clinker, but, in all of them, the focus is on complex mathematical models for machine learning, such as novel support vector machine ensemble (ESVM) (LIU et al., 2020) or multivariate time series analysis (ZHAO et.al, 2021) among other models used along the last years by different research lines, so the feature selection itself had a secondary role in these works. As mentioned in *5.1 Pre* – processing data , page 64, there are more than one hundred variables in the system, it is difficult to deal with all of them with complex mathematical models, because of this limitation, in general, the researchers opted for keeping a small number of features, less than fifteen, and the feature selection was based on other criteria, as operator knowledge.

In the present work, the initial 120 features as input decreased for 32 features with statistical relevance as output, in sim04, with $\boldsymbol{R^2 = 0.57}$ and 36 features in sim18, with $\boldsymbol{R^2 = 0.67}$, *cf.* Table 11. The system cannot be described satisfactorily with a small number of variables. Due to the correct feature set-up, the MLR model had best results than other complex models proposed in literature like novel support vector machine ensemble (ESVM), with $\boldsymbol{R^2=0.48}$, support vector machines (SVM), with $\boldsymbol{R^2=0.42}$, convolutional neural network (MVTS–CNN), with $\boldsymbol{R^2=0.62}$, among others (LIU et al., 2020) (ZHAO et al., 2021). Consequently, it is expected that a complex model with the correct feature selection as input has a better result than the MLR. The description of each feature selected by sim04 and sim18 is in APPENDIX B, page 93.

The following discussions are related to feature selection presented in APPENDIX B, Table 16**,** and interpreted from the process point of view.

Firstly, by quality features, it is noted that only 2 raw meal quality features (Alumina Ratio and Mg content - online analysis) in the SIM18 make a reasonable improvement when compared with SIM04, $\boldsymbol{R^2}$ increases from **0.57** to **0.67**. Regarding the clinker quality control, with 4 features in both datasets, the chemical and physical clinker´s properties are dependent on raw meal quality which does not change abruptly. Consequently, the previous clinker quality values carry essential information

for the mathematical modelling to predict the *f*-CaO. Moreover, ashes from alternative fuels are incorporated into the clinker, which impacts quality control parameters.

The influence of the cooler operation in the *f*-CaO is an intriguing result. There are 15 and 13 features statistically relevant from cooler for SIM 04 and SIM18, up than 40% of the operational parameters for both datasets. Most of the time, the grate cooler is classified in the cement industry only as a heat machine, with correct operation connected to save heat energy in the system with the analyses involved to aim heat consumption, decreases (kcal/kg clinker). However, the feature selection results show a high dependence between *f*-CaO and cooler operation, which is evidenced in Figure 9, optimal cooler operation means a quick interruption of the C_3S decomposition process. In the cement industry, control of calciner temperature by the operation is standard, when there is an *f*- CaO out of quality, the operator acts by increasing the fuel flow at the calciner to increase temperature (Holcim Group Support Ltd, 2008). However, the results show that the strategy needs to be improved, and the cooler must be executed correctly to avoid high *f*-CaO in the clinker. The optimal cooler operation increases the temperature of tertiary air, consequentially increasing the calciner temperature, correlated variables that are also selected, *cf.* Table 17. The operator and process engineers must not pay attention only to calciner operation but should look at the entire system, the cooler operation philosophy cannot be neglected.

A distinct approach discussing the results from feature selection by fluid, heat dynamics and chemical kinetic concepts. Pressures, temperatures, fan power/speed are properties that cannot be analysed separately, because, together, they have information about the heat transfer process in the kiln system, as explained on *3.5 Fluid and heat dynamics considerations about chemical reactions to form clinker*, page 51. Variables with pressure appear more than temperatures, and fan powers/speed appear but less than the other two. The exact connection among them, for this entire and complex system, is not clear, what characterizes the empirical models. It is possible to understand the concept between them but proposing a mechanistic model with all variables and resolving it is complex and out of the scope of this work.

There are retention time and the delay time from chemical kinetic concepts, *cf. 3.4.6 Retention time in the kiln system and distribution of residence times,* page *49,* for the kiln is represented by the filling degree and/or kiln speed, for cooler, cooler grate speed and residence time. The features with delays can be explained by the residence time distribution; *f*-CaO dependence with the feature dedusting fan is unexpected,

since the dedusting bag filter is small, only to avoid excess dust in the field. Moreover, it is not mandatory for the cooler system, and this dependence appears in many simulations. It is possible to infer that the dependence is real and not random. The hypothesis is that the dedusting bag filter is installed at the wrong position, which affects the air balances at the cooler (secondary, tertiary, and excess air). To end this section, the information that features describes are summarized at Table 12.

Table 12 - Summarize for process information that features describes.

Concept	Area Description	Features	Unit	Notes
Chemical composition (quality control)	Raw meal at kiln feed	AR / Mg content		online analysis / SIM18
	Clinker	x_{29} to x_{32} for SIM04 x_{33} to x_{36} for SIM18	%mass %mass	*previous* values for Alite, Belite, Alumina, Ferrite, Alkalis, *f*-CaO
Air flow	Kiln Preheater Tower Cooler + Excess Air Fan	Fan power Fan speed Pressures	kW rpm mbar	connected by fan laws (13) to (15)
Gas flow	Preheater Tower Conditioning Tower + ID Fan	Fan power Fan speed Pressures	kW rpm mbar	connected by fan laws (13) to (15)
Heat exchange gas - material	Preheater Tower	Raw meal temperatures	℃	Quality of heat exchange, topic *4.3* and (12)
Sensible heat from hot gases/air	Kiln Preheater Tower Conditioning Tower + ID Fan Cooler + Excess Air Fan	Temperatures Pressures Fan power Fan speed	℃ mbar kW rpm	Thermal energy available at system (flow + temperatures) (12) connected by fan laws (13) to (15)
Retention time	Kiln	Kiln Speed Filling Degree - Kiln	rpm %	(9) (11)
	Cooler	Cooler Grate Speed Residence Time	strocks/min min	(10)
Distribution of residence times	Kiln Preheater Tower Conditioning Tower + ID Fan Cooler + Excess Air Fan	Variables with distinct delays	min	no additional delay 20min 40 min

5.3 Prediction *f*-CaO in clinker by regression models

The algorithm which finds the regression variables with optimal coefficient values (b_i) for the polynomial models was written in C++. The calculation for p-value, RMSE, MRE and a cross-check for the coefficients (b_i) for each model was carried out in Python, using libraries such as pandas, NumPy, and statsmodels; Microsoft Excel® was used for additional analyses. Considering the training data, the metric results for regression models are in Table 13.

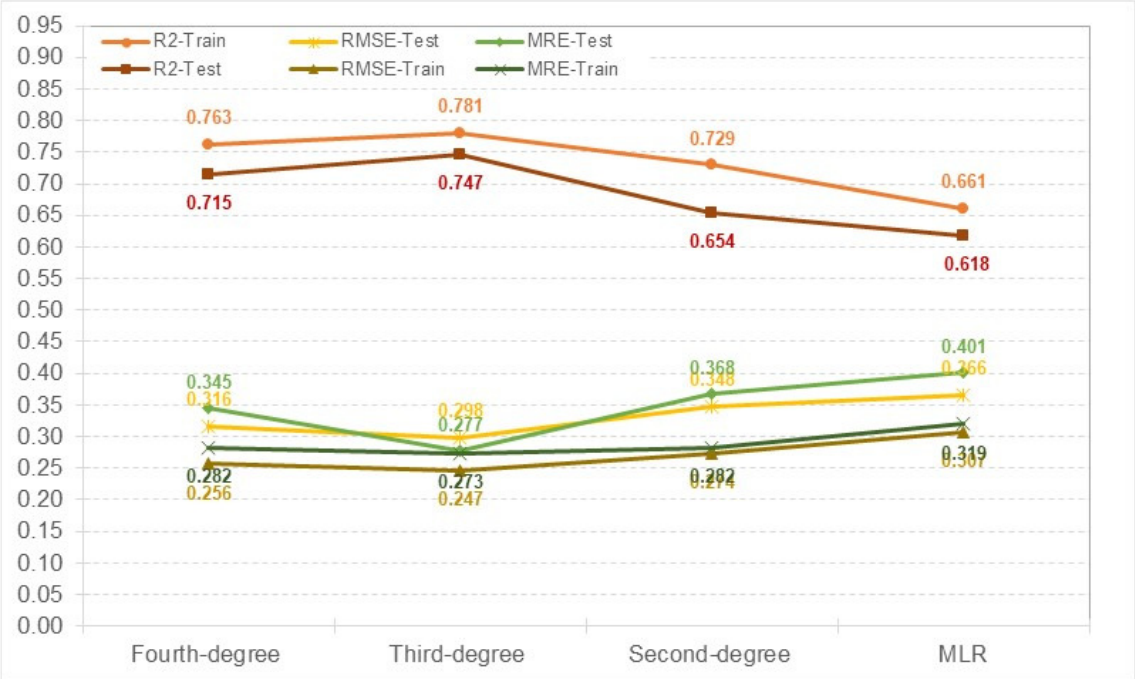
Reasonable second and fourth-degree polynomials for Sim04 were not found. After applying the statistical significance test, p-value lower than 0.05, the number of regression variables (*n*) decreased by 18% for Sim04 and between 25 and 30% for Sim18 with minimum impact in the metrics; the R^2 decreased by less than 5%, errors metrics increased by less than 8%, *cf.* Table13.

Table 13 - Results for the polynomial models before and after applying the null-hypothesis significance testing for training data.

Model		Sim04			Sim18		
	Metrics	**Initial (*ni*)**	**Final (*nf*)**	**%difference**	**Initial (*ni*)**	**Final (*nf*)**	**%difference**
Fourth -degree Polynomial	Inputs				60	45	-25.0%
(Quartic function)	p-value (max)				0.465	0.039	-91.6%
	R^2				0.794	0.763	-3.9%
	RMSE				0.239	0.256	7.2%
	MRE				0.510	0.531	4.0%
Third-degree Polynomial	Inputs	50	41	-18.0%	79	57	-27.8%
(Cubic function)	p-value (max)	0.819	0.024	-97.0%	0.987	0.041	-95.9%
	R^2	0.637	0.635	-0.4%	**0.782**	**0.781**	**-0.1%**
	RMSE	0.307	0.308	0.3%	0.246	0.247	0.2%
	MRE	0.315	0.322	2.1%	0.275	0.273	-0.8%
Second-degree Polynomial	Inputs				53	37	-30.2%
(Quadratic function)	p-value (max)				0.919	0.048	-94.8%
	R^2				0.751	0.729	-2.8%
	RMSE				0.263	0.274	4.1%
	MRE				0.268	0.282	5.0%
First-degree Polynomial (MLP - multiple linear regression)	Inputs	32	26	-18.8%	36	25	-30.6%
	p-value (max)	0.232	0.021	-91.0%	0.812	0.008	-99.0%
	R^2	0.569	0.563	-1.0%	0.672	0.661	-1.7%
	RMSE	0.334	0.336	0.7%	0.302	0.307	1.7%
	MRE	0.340	0.343	1.0%	0.316	0.319	1.2%

Sim18 reached better results than Sim04, with higher R^2, lower RMSE and MRE values for all regression models. The metrics for polynomial models considering Sim18 are in Table 13, where the third-degree polynomial has the best performance, with **R^2= 0.747**, RMSE and MRE lower than 0.300 for the testing dataset.

Figure 15 - Sim18: Comparison of performance for polynomial´s models using R^2, RMSE and MRE as a metrics.



The metrics for the additional fourth-order polynomial model for Sim18 are in Table 14. The dataset was split into 85% for training and 15% for testing for this modelling.

The results show that the stopping criterion, based on R^2 for the testing data combined with the p-values lower than 0.05 to maintain the variables in the final model, is efficient, and overfitting phenomena did not occur. Furthermore, increasing the variable's number and model complexity does not mean performance improvement, *cf.* Figure 15. The better metrics for the additional fourth-order polynomial model, *cf.* Table 14, is due to the larger dataset size for training, 85% against 80% for previous models, which is an indicative that the same methodology applied to a bigger dataset could improve the model's performance. The final regression variables (w_i ($\boldsymbol{x}$)) and coefficients (b_i) for the multivariate polynomials are in Table 18 and Table 19.

Table 14 - Additional fourth-order polynomial model for Sim18 (dataset:85% for training and 15% for testing).

Metrics	Training	Testing
Inputs (*nf*)	55	55
p-value (max)	0.046	-
R^2	**0.801**	**0.779**
RMSE	0.251	0.321
MRE	0.238	0.269

In the literature, a value for R^2= 0.958 was obtained by Li, Wang and Chai (2015). Nevertheless, the model used features from the flame image, which increased the complexity of the model substantially and became complicated the application in the real world when compared with the polynomial models, which use only continuous numeric features as input; another point is the sample size, with 157 *f*-CaO values associated with the flame image, the data was split in 50% for training and 50% for testing, the dataset size was small. An R^2= 0.89 was achieved by Yao et al. (2021), but the *f*-CaO range used (0 to 0.7%) was too short and does not reflect the reality in a cement plant where the *f*-CaO range is between 0 to 3.5% (LIU et al.,2020) (ZHAO et al., 2021) (LI; WANG; CHAI, 2015).

5.4 Prediction *f*-CaO in clinker by machine learning algorithms

Additionally, five common machine learning algorithms were used for modelling *f*-CaO content in clinker. The metric results for MLP, XGBoost, CatBoost, RDF, and SVM are in Table 15 with MLR and third-degree polynomials from section 5.3, totalizing seven models.

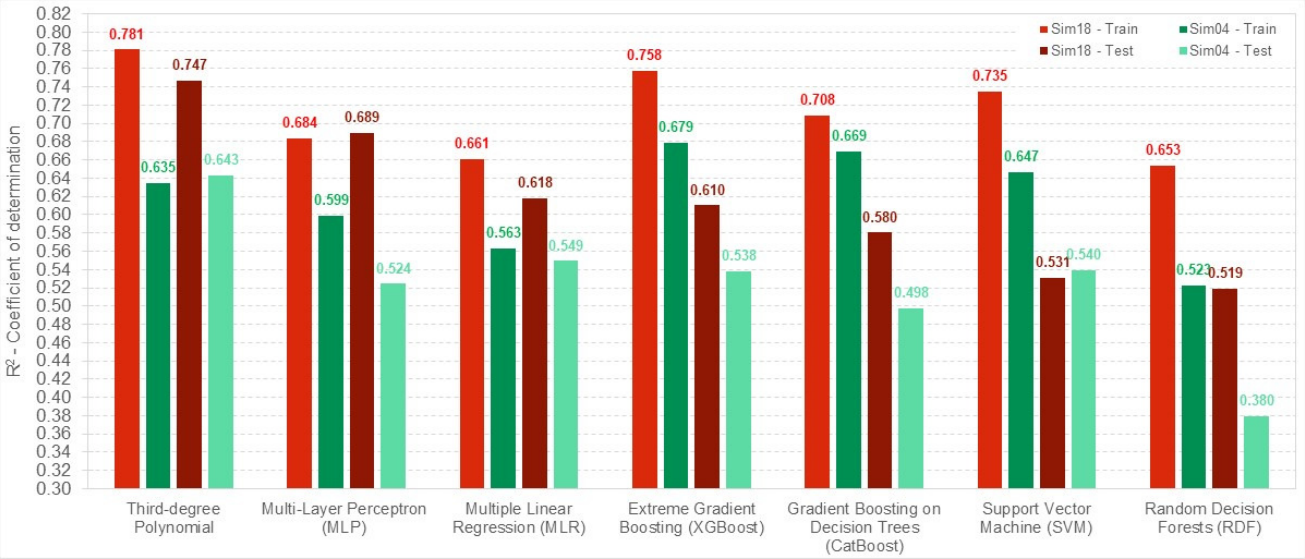
A bar chart with R^2 values for training and testing data for both sub-datasets (Sim04 and Sim18) is in Figure 16.

Table 15 - Comparison of performance for different models using R^2, RMSE and MRE as a metrics.

Model		Train			Test			%difference Test/Train	
	Metrics	Sim04	Sim18	%difference*	Sim4	Sim18	%difference*	Sim4	Sim18
Third-degree Polynomial	R^2	0.635	0.781	23.0%	0.643	0.747	16.1%	1.3%	-4.6%
	RMSE	0.308	0.247	-19.9%	0.287	0.298	3.7%	-7.1%	17.3%
	MRE	0.322	0.273	-15.1%	0.328	0.277	-15.5%	1.9%	1.5%
Multi-Layer Perceptron (MLP)	R^2	0.599	0.684	14.2%	0.524	0.689	31.4%	-14.1%	0.8%
	RMSE	0.317	0.304	-4.2%	0.354	0.300	-15.2%	10.2%	-1.4%
	MRE	0.351	0.303	-13.6%	0.326	0.363	11.3%	-7.5%	16.5%
Multiple Linear Regression (MLR)	R^2	0.563	0.661	17.4%	0.549	0.618	12.5%	-2.5%	-6.9%
	RMSE	0.337	0.307	-8.9%	0.323	0.366	13.3%	-4.2%	16.2%
	MRE	0.343	0.319	-6.9%	0.377	0.401	6.5%	9.0%	20.4%
Extreme Gradient Boosting (XGBoost)	R^2	0.679	0.758	11.6%	0.538	0.610	13.4%	-26.1%	-24.2%
	RMSE	0.288	0.259	-10.1%	0.327	0.370	13.3%	11.8%	30.0%
	MRE	0.300	0.278	-7.3%	0.372	0.392	5.5%	19.4%	29.1%
Gradient Boosting on Decision Trees (CatBoost)	R^2	0.669	0.708	5.9%	0.498	0.580	16.6%	-34.4%	-22.1%
	RMSE	0.293	0.284	-2.9%	0.341	0.383	12.5%	14.1%	25.9%
	MRE	0.321	0.313	-2.7%	0.396	0.391	-1.3%	18.9%	20.1%
Support Vector Machine (SVM)	R^2	0.647	0.735	13.7%	0.540	0.531	-1.7%	-19.8%	-38.5%
	RMSE	0.303	0.280	-7.5%	0.326	0.359	10.1%	7.2%	22.1%
	MRE	0.319	0.195	-38.8%	0.375	0.377	0.6%	15.0%	48.3%
Random Decision Forests (RDF)	R^2	0.523	0.653	25.0%	0.380	0.519	36.5%	-37.6%	-26.0%
	RMSE	0.352	0.312	-11.2%	0.379	0.411	8.4%	7.1%	24.0%
	MRE	0.406	0.337	-16.9%	0.487	0.421	-13.6%	16.7%	19.9%

*%Difference between Sim18 and Sim04.

Figure 16 - Comparison of performance for different models using R^2 as a metric.



Considering the R^2 for testing dataset as the principal performance metric, the results presented in Figure 15 and Figure 16 are interpreted and clustered into two topics:

Comparison of performance of the two sub-datasets (Sim04 and Sim18):

1. Sim18 has superior performance than Sim04 for all models, in the training and testing data, as is evidenced in Figure 16;

2. Sim18 has minor RMSE and MRE values for all models in the training dataset, some discrepancies occurred in the testing data, one reason is the dataset sizes are not equal. *cf. 5.1 Pre – processing data* ,page 64, the size is input in RMSE and MRE calculation, *cf.* Equations (18) and (19) ;

3. The discrepancies between training and testing data metrics also evidence the importance of multiple parameters to evaluate the performance of the models;

4. The most significant difference between Sim04 and Sim18 is the process variables (x_{23} and x_{24}) related to the raw meal chemical analyses online in Sim18, *cf.* Table 17, what had a considerable impact on the models' performance;

5. In the literature, some works considered the raw meal quality as inputs, but values come from the quality laboratory that the frequency is not continuous as the online analysis (ZHAO et al., 2021) (LI; WANG; CHAI, 2015).

6. On the other hand, due to the difference in the dataset size, it is necessary to repeat the current work with both datasets of the same size, using equal data for both datasets, with the only difference will be the features related to the raw meal chemical analyses online in one of them. In this way, the impact of the raw meal quality variables in the models can investigate deeply.

Comparison of the distinct models (MLP, XGBoost, CatBoost, RDF, SVM, MLR and third-degree polynomials):

1. Performance did not change with the datasets, from the highest to the lower values;

2. The third-degree polynomial had the best metrics followed by MLP, with MLR in the third position. These models had a lower metrics difference between training and testing data;

3. Oppositely, the remaining models, XGBoost, CatBoost, SVM and RDF, had the worst achievements, with the highest difference between training and testing data;

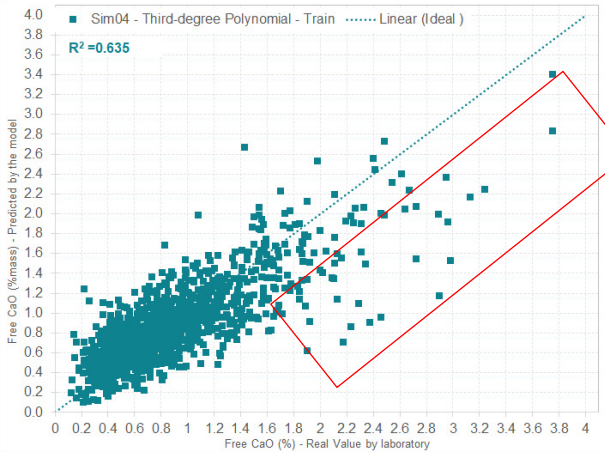
4. The algorithms with poor performance have in common the higher mathematical complexity and the facility of occur overfitting if the hyperparameters are not tuning properly;

5. As mentioned by Bashir et al. (2020): “overfitting occurs when an algorithm reduces error through memorization of training examples, with noisy or irrelevant features, rather than learning the true general relationship between X and Y”, in other words, *the algorithm starts modelling errors*;

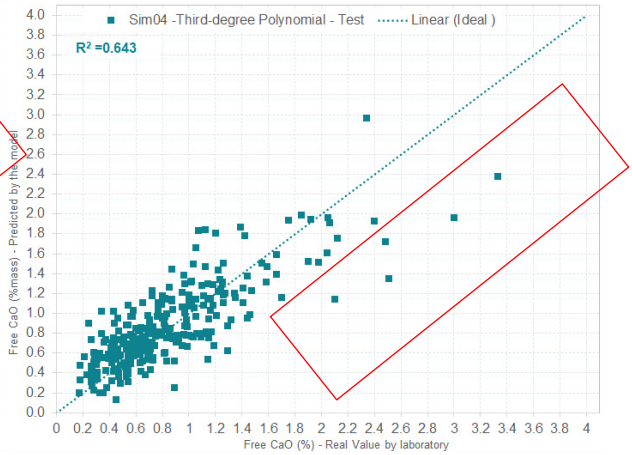
6. The optimization of hyperparameters combined with the methodology present in the current work is suggested for future research.

The results from the three best models, with *f*-CaO prediction versus real values, are in Figures 17 ,18 for Sim04 and Figures 19, 20 for Sim18.

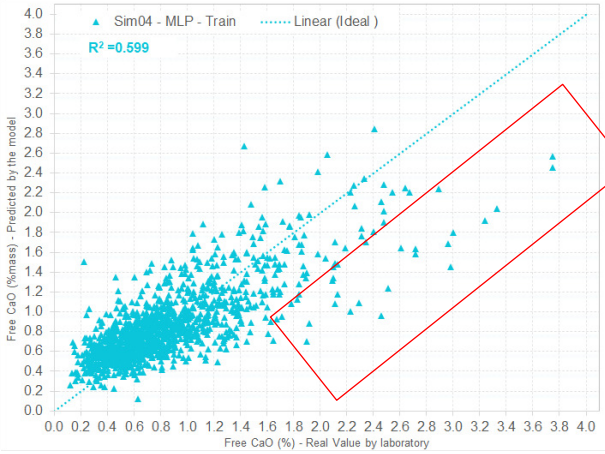
Figure 17 - Sim04 : Prediction *versus* real values for (a) training and (b) testing of the third-degree polynomial model with the highest testing R^2; (c) training and (d) testing of the MLP model with the second highest testing R^2 and (e) training and (f) testing of the MLR model with the third highest testing R^2.



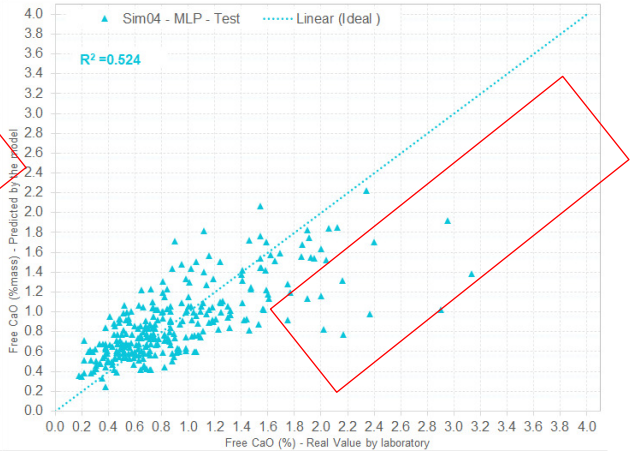
(a)**Sim04:** Prediction *versus* real value for cubic model – Training set.



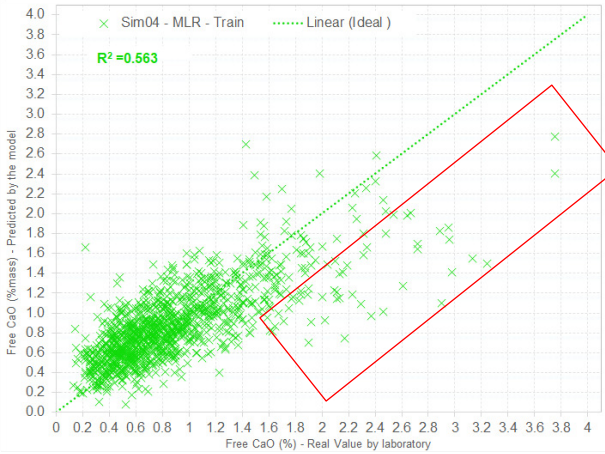
(b)**Sim04:** Prediction *versus* real value for cubic model – Testing set.



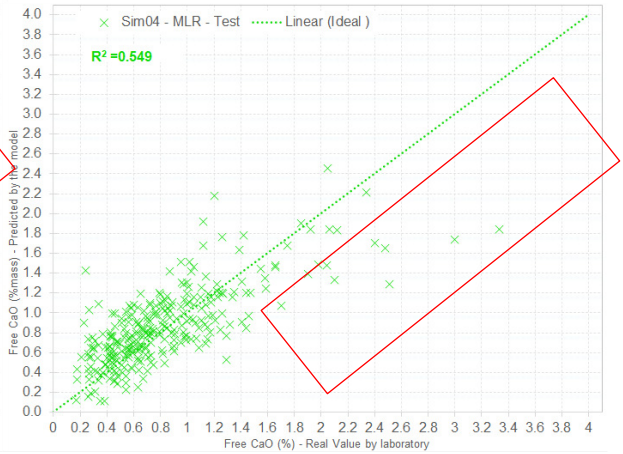
(c)**Sim04:** Prediction *versus* real value for MLP model – Training set.



(d)**Sim04:** Prediction *versus* real value for MLP model – Testing set.

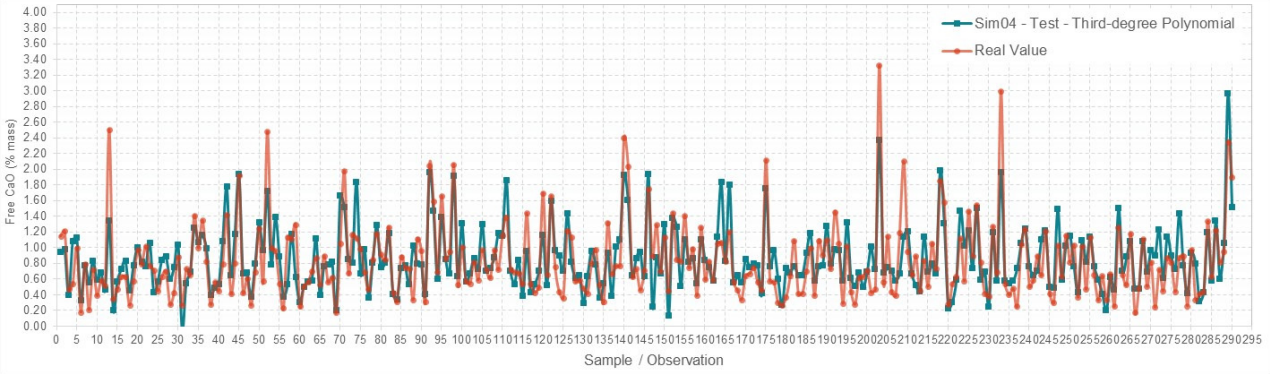


(e)**Sim04:** Prediction *versus* real value for MLR model – Training set.

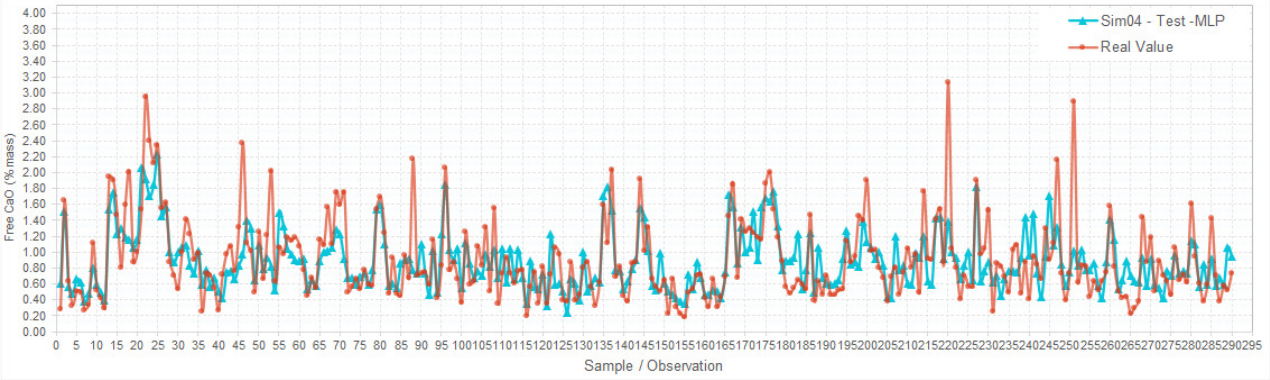


(f)**Sim04:** Prediction *versus* real value for MLR model – Testing set.

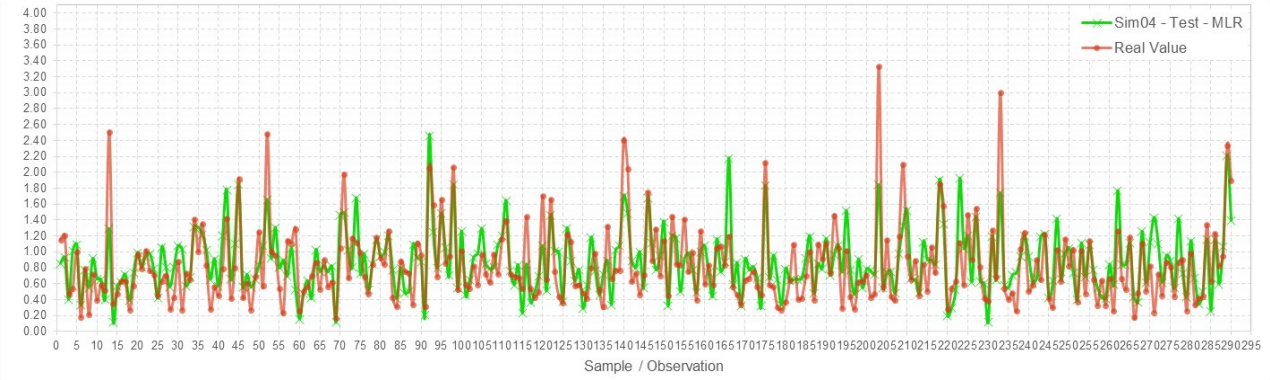
Figure 18 - Sim04: Comparison of *f*-Cao prediction by three different models with the highest R^2 value for testing data (a) third-degree polynomial model; (b)MLP model; (c) MLR model.



(a)**Sim04**: Comparison between prediction and real value for cubic model – Testing set.

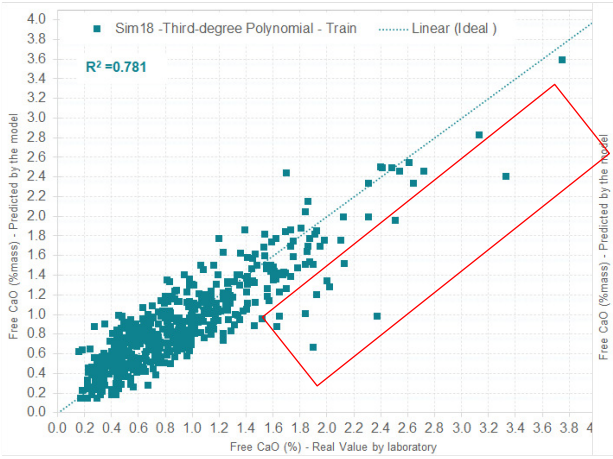


(b)**Sim04**: Comparison between prediction and real value for MLP model – Testing set.

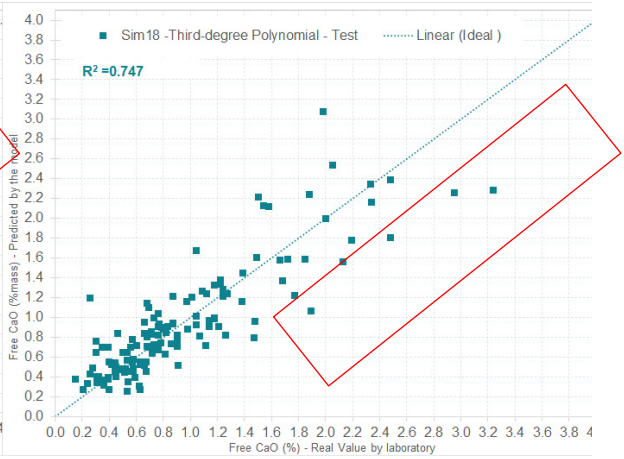


(c)**Sim04**: Comparison between prediction and real value for MLR model – Testing set.

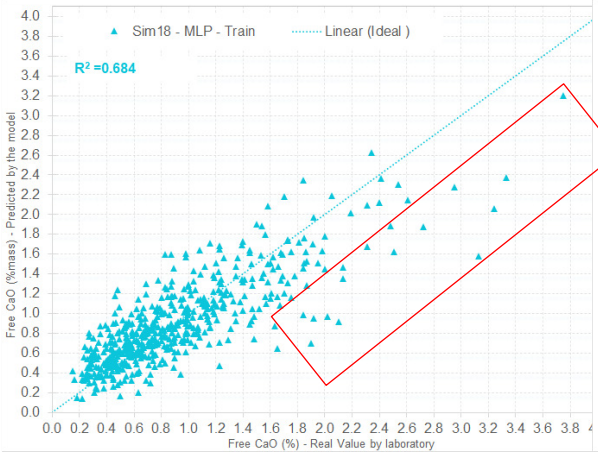
Figure 19 - Sim18 : Prediction *versus* real values for (a) training and (b) testing of the third-degree polynomial model with the highest testing R^2; (c) training and (d) testing of the MLP model with the second highest testing R^2 and (e) training and (f) testing of the MLR model with the third highest testing R^2.



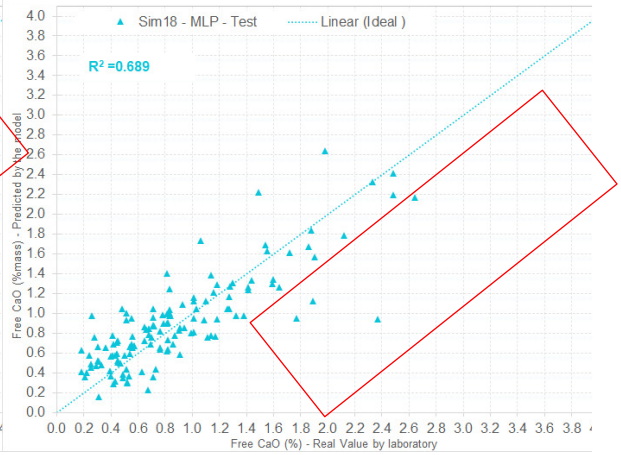
(a)**Sim18:** Prediction *versus* real value for cubic model – Training set.



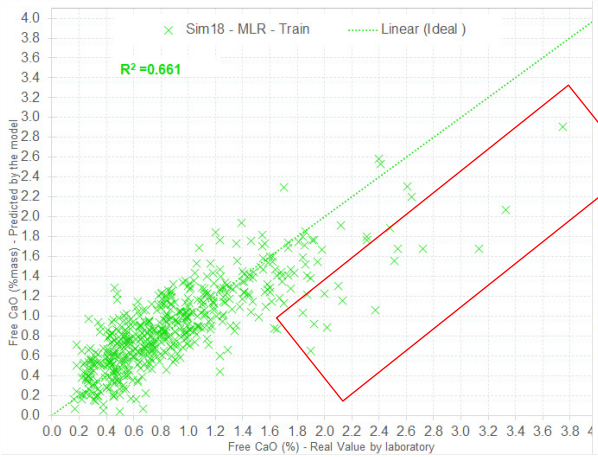
(b) **Sim18:** Prediction *versus* real value for cubic model – Testing set.



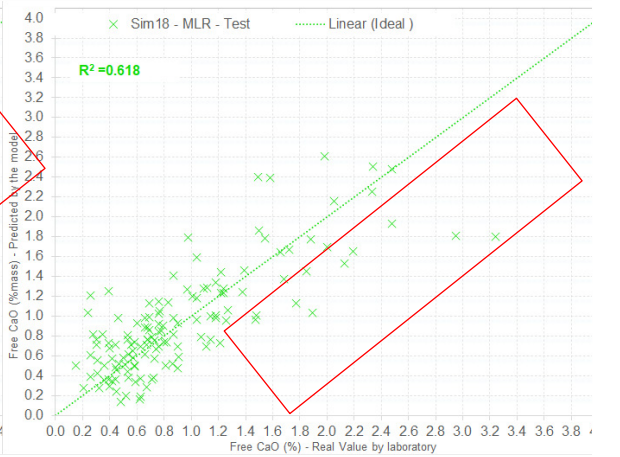
(c)**Sim18:** Prediction *versus* real value for MLP model – Training set.



(d) **Sim18:** Prediction *versus* real value for MLP model – Testing set.

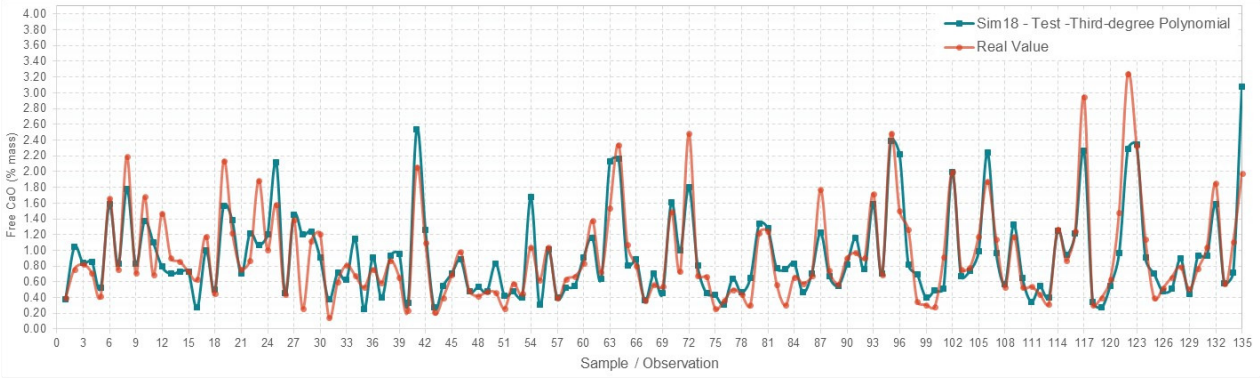


(e)**Sim18:** Prediction *versus* real value for MLR model – Training set.

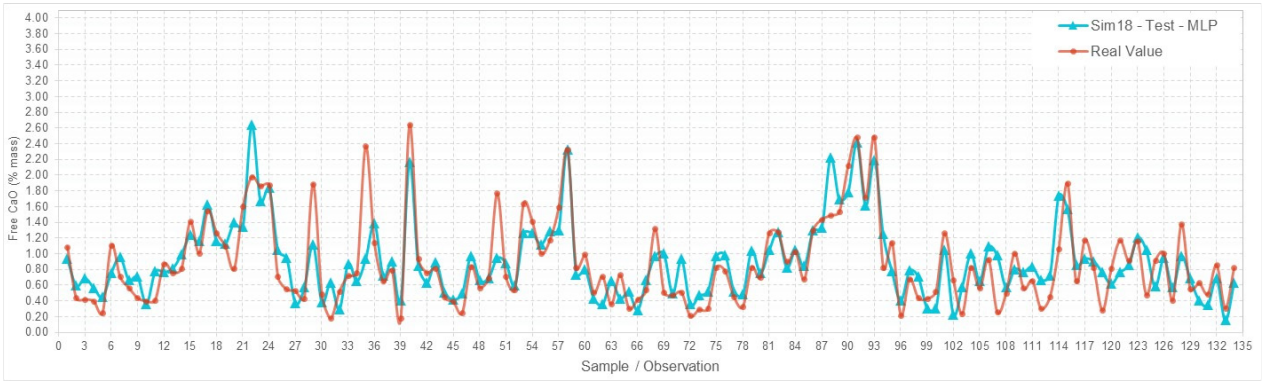


(f) **Sim18:** Prediction *versus* real value for MLR model – Testing set.

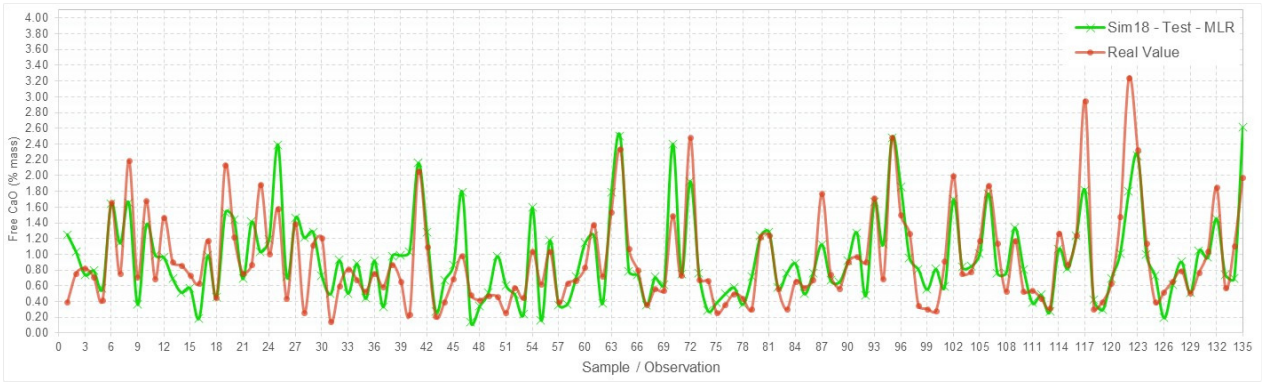
Figure 20 - Sim18: Comparison of *f*-Cao prediction by three different models with the highest R^2 value for testing data (a) third-degree polynomial model; (b)MLP model; (c) MLR model.



(a)**Sim18**: Comparison between prediction and real value for cubic model – Testing set.



(b)**Sim18**: Comparison between prediction and real value for MLP model – Testing set.



(c)**Sim18**: Comparison between prediction and real value for MLR model – Testing set.

The furthest points from the ideal line are shown in a red rectangle in the graphs presented in Figures 17 and 19. Prediction values with higher errors occur for *f*-CaO above 1.8% for both datasets, and the higher density of outliers above 1.8% is identified in the graphics. Just 8% of raw data has *f*-CaO values higher than 1.8%, so this sample size is insufficient to train the models satisfactorily in this range, with values higher than 1.8%, and is an improvement for future modelling.

6 CONCLUSIONS

The results obtained using industrial data combined with process knowledge for decision making, including the filter for the final dataset, which was the clinker quality features and not a statistical criterion, shows chemical and physical laws should be accounted for modelling a chemical engineering system in the real world and it is reflected in features selection. It doesn't matter if the models are empirical, robust algorithms are not able to compensate for the failure in the incorrect variable setup, so feature selection is a step as important as applying the algorithm itself. The results in the MLR model, the more straightforward and most widely used algorithm in engineering, demonstrated the importance of the feature selection step. The initial 120 features as input decreased for 32 with statistical relevance as output, in sim04, with **R^2= 0.57** and 36 features in sim18, with **R^2= 0.67**. Due to the correct feature set-up, the MLR model had better results than other complex models proposed in literature like novel support vector machine ensemble (ESVM), with **R^2=0.48**, support vector machines (SVM), with **R^2=0.42**, convolutional neural network (MVTS–CNN), with **R^2=0.62**, among others (LIU et al., 2020) (ZHAO et al., 2021).

Additionally, the variables related to the chemical composition and cooler operation have a substantial influence on the prediction models for *f*-CaO content. Pressures, temperatures, and fan power/speed are properties that cannot be analysed separately because these variables together have information about the heat transfer process in the kiln system. Variables with pressure appear more than temperatures, and fan powers/speed appear but less than the other two. The exact connection among them, for this entire and complex system, is unclear, which characterizes the empirical models. About the soft sensor, an effective model that saves thermal energy in the cement industry, the robust and complex statistical models applied (XGBoost, CatBoost, SVM and RDF) had poor performance, with **R^2** lower than **0.55** for the testing dataset and optimization of hyperparameters combined with the methodology present in the current work is suggested for future research. Finally, the multivariate polynomial models had satisfactory results for predicting the *f*-CaO content in clinker, with **R^2=0.78** in the fourth-degree model and **R^2=0.75** in the three-degree model, for the testing dataset from sim18, which included online chemical analysis for raw meal that feeds the kiln system. There are opportunities to improve the polynomials model's performance for a bigger dataset size with raw meal chemical analysis online available.

7 SUGGESTIONS FOR FUTURE RESEARCH

The current study can be interpreted as the first step in the methodology that combined chemical engineering concepts and MLR for feature selection and this step can be considered as fundamental as the modelling itself. There are a lot of opportunities to improve the results found as listed:

1. Repeat the modelling with two equal datasets, one with the features related to the raw meal online chemical analyses and the other without;

2. Use a larger dataset, around 3000 samples or more of *f*-CaO. The sampling size used in this research was around 1500 for Sim04;

3. Consider a dataset from a new cement site, with a modern system to obtain the data and reliable instrumentation;

4. Consider the same data with a shorter time to *f*-CaO analyses, around 1 hour or less. The *f*-CaO analyses were carried out around every 2,5 hours in the data used in this research;

5. Optimize the hyperparameters for robust machine learning algorithms (XGBoost, CatBoost, SVM and RDF).

8 BIBLIOGRAPHY

ALI, M.B.; SAIDUR, R.; HOSSAIN, M.S..***A review on emission analysis in cement industries***. Renewable and Sustainable Energy Reviews, volume 15, issue 5, pages 2252-2261, 2011, ISSN 1364-0321,
https://doi.org/10.1016/j.rser.2011.02.014

ALTMAN, Naomi; KRZYWINSKI, Martin. ***Points of Significance: Association, correlation and causation***. Nature Methods, pages 899–900, September 2015, https://doi.org/10.1038/nmeth.3587

AMERICAN SOCIETY FOR TESTING AND MATERIALS. ASTM C150 / C150M-21: ***Standard specification for Portland cement***. West Conshohocken, Pennsylvania, USA: ASTM, 2021,
https://doi.org/10.1520/C0150_C0150M-22

AMERICAN SOCIETY FOR TESTING AND MATERIALS. ASTM D5865/D5865M-19: ***Standard Test Method for Gross Calorific Value of Coal and Coke***. West Conshohocken, Pennsylvania, USA: ASTM, 2019,
https://doi.org/ 10.1520/D5865_D5865M-19

Axair Fans UK Ltd. ***www.axair-fans.co.uk***., available https://www.axair-fans.co.uk/news/understanding-basic-fan-laws/
Accessed: 03 of March 2022

BASHIR, Daniel; MONTAÑEZ, George D.; SEHRA, Sonia; SEGURA, Pedro Sandoval; LAUW, Julius. ***An Information-Theoretic Perspective on Overfitting and Underfitting***. Australasian Joint Conference on Artificial Intelligence, AI 2020: Advances in Artificial Intelligence, volume 12576, pages 347–358, November 2020,
https://doi.org/10.1007/978-3-030-64984-5_27

BASU, Prabir. ***Biomass Gasification, Pyrolysis and Torrefaction - Practical Design and Theory***. Third Edition. Academic Press, 2018, ISBN 9780128129920,
https://doi.org/10.1016/C2016-0-04056-1

BOMMERT, Andrea; SUN, Xudong; BISCHL, Bernd; RAHNENFÜHRER, Jörg; LANG, Michel. ***Benchmark for filter methods for feature selection in high-dimensional classification data***. Computational Statistics & Data Analysis, volume 143, 106839, 2020, ISSN 0167-9473,
https://doi.org/10.1016/j.csda.2019.106839

EUROPEAN COMMITTEE FOR STANDARDIZATION - CEN. ***Cement - Part 1: Composition, specifications, and conformity criteria for common cements***. Brussels: CEN - EUROPEAN COMMITTEE FOR STANDARDIZATION, 2011.

FLSmidth. ***FLSmidth / Discover***. Available in FLSmidth: https://www.flsmidth.com/en-gb/discover/mining-2020/wind-will-power-clean-energy-and-the-demand-for-crucial-minerals
Accessed: 29 of December 2020

FOGLER, H. Scott. ***Elements of Chemical Reaction Engineering***. 5th Edition. Boston: Prentice Hall, 2016, ISBN 9780133888096, ISBN 0133888096

FOX, Robert W.; MCDONALD, Alan T.; PRITCHARD, Philip J.; MITCHELL, John W. ***Introduction to Fluid Mechanics***. 9th ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2015, ISBN 9781119034582

GIELEN, Dolf; BOSHELL, Francisco; SAYGIN, Deger; BAZILIAN, Morgan D; WAGNER, Nicholas; GORINI, Ricardo. ***The role of renewable energy in the global energy transformation***. Energy Strategy Reviews, volume 24, pages 38-50, 2019, ISSN 2211-467X,
https://doi.org/10.1016/j.esr.2019.01.006

GHALANDARI, Vahab; MAJD, Mahdieh Mozaffari; GOLESTANIAN, Amir. ***Energy audit for pyro-processing unit of a new generation cement plant and feasibility study for recovering waste heat: A case study***. Energy, volume 173, pages 833-843,2019, ISSN 0360-5442,
https://doi.org/10.1016/j.energy.2019.02.102.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. ***The elements of statistical learning: data mining, inference, and prediction***. Second Edition. Switzerland: Springer Nature Switzerland AG, 2017,
https://doi.org/10.1007/978-0-387-84858-7

HEWLETT, P. C. ***Lea's Chemistry of Cement and Concrete***. 4thEdition. Oxford: Butterworth-Heinemann, 2006, ISBN:9780080535418

Holcim Group Support Ltd. ***Cement Manufacturing Course - Technical Documentation***. *Materials Technology I,* volume 1. Holcim Group Support Ltd, 2008.

Holcim Group Support Ltd. ***Cement Manufacturing Course - Technical Documentation***. *Process Technology I,* volume *3*. Holcim Group Support Ltd, 2008.

Holcim Group Support Ltd. ***Cement Manufacturing Course - Technical Documentation***. *Process Technology II,* volume *4*. Holcim Group Support Ltd, 2008.

KHAIRE, Utkarsh Mahadeo; DHANALAKSHMI, R. ***Stability of feature selection algorithm: A review***. Journal of King Saud University - Computer and Information Sciences, volume 34, issue 4, pages 1060-1073, 2022, ISSN 1319-1578,
https://doi.org/10.1016/j.jksuci.2019.06.012.

LIU, Xiaoyan; JIN, Jiao; WU, Weining; HERZ, Fabian. ***A novel support vector machine ensemble model for estimation of free lime content in cement clinkers***. ISA Transactions, volume 99, pages 479-487,2020, ISSN 0019-0578,
https://doi.org/10.1016/j.isatra.2019.09.003

LI, Weitao; WANG, Dianhui; CHAI, Tianyou. ***Multisource Data Ensemble Modelling for Clinker Free Lime Content Estimate in Rotary Kiln Sintering Processes***. IEEE Transactions on Systems, Man, and Cybernetics: Systems, volume 45, no. 2, pages 303-314, February 2015,
https://doi.org/10.1109/TSMC.2014.2332305

MAULUD, Dastan; ABDULAZEEZ, Adnan M. ***A Review on Linear Regression Comprehensive in Machine Learning***. Journal of Applied Science and Technology Trends, volume 1, number 4, pages 140-147, 31 of December 2020.
https://doi.org/10.38094/jastt1457

MOHAMED, Yasir A.; KASIF, A. Elhameed MO; ALLA, Elrafie AA; ELMAHADI, Muaz Musa. ***Calculation of the formation process of clinker inside the rotary cement kiln***. Vestnik VGUIT, volume 80, no. 1, pages 233-239, Mar 2018,
https://doi.org/10.20914/2310-1202-2018-1-233-239

NDIAYE, L.G; CAILLAT, S.; CHINNAYYA, A.; GAMBIER, D.; BAUDOIN, B. ***Application of the dynamic model of Saeman to an industrial rotary kiln incinerator*: *Numerical and experimental results***. Waste Management, volume 30, pages 1188–1195, July 2010,
https://doi.org/10.1016/j.wasman.2009.09.023

NEWMAN, J.; CHOO, B. S. ***Advanced Concrete Technology***. Butterworth-Heinemann, 2003, ISBN: 9780080490014.

NIQUINI, Gabriela R; SILVA, Suzimara R.; JUNIOR, Esly Ferreira da Costa; COSTA, Andréa Oliveira Souza. ***Feedstock and inoculum characteristics and process parameters as predictors for methane yield in mesophilic solid-state anaerobic digestion***. Anais da Academia Brasileira de Ciências, volume. 91, n. 4, page e20181181, 2019,
https://doi.org/10.1590/0001-3765201920181181

OZGUR, Ceyhun; COLLIAU, Taylor; ROGERS, Grace; HUGHES, Zachariah. ***MatLab vs. Python vs. R*. Journal of Data Science**, volume 15, issue 3, pages 355-372, 2017,
https://doi.org/10.6339/JDS.201707_15(3).0001

PAREDES, Ingrid J.; YOHANNES, Bereket; EMADY, Heather N.; MUZZIO, Fernando J.; MAGLIO, Al; BORGHARD, William G.; GLASSER, Benjamin J.; Cuitiño, Alberto M. ***Measurement of the residence time distribution of a cohesive powder in a flighted rotary kiln***. Chemical Engineering Science, volume 191, pages 56-66,2018, ISSN 0009-2509,
https://doi.org/10.1016/j.ces.2018.06.044.

PCA-Portland Cement Association. ***Innovations in Portland Cement Manufacturing***. Illinois: Portland Cement Association, 2004, ISBN 0893122343, 9780893122348

RASMUSON, Anders; ANDERSSON, Bengt; OLSSON, Louise; ANDERSSON, Ronnie. ***Mathematical Modelling in Chemical Engineering***. Cambridge: Cambridge University Press,2014,
https://doi.org/10.1017/CBO9781107279124

RAY, Susmita. ***A Quick Review of Machine Learning Algorithms***. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, pages. 35-39, 2019, https://doi.org/10.1109/COMITCon.2019.8862451

RODRIGUES, Diulia C.Q.; JUNIOR, Atílio P. Soares; JUNIOR, Esly Ferreira da Costa; COSTA, Andréa Oliveira Souza. ***Dynamic Analysis of the Temperature and the Concentration Profiles of an Industrial Rotary Kiln Used in Clinker Production***. Anais da Academia Brasileira de Ciências, Engineering Sciencies, volume 89, n.4, pages 3123-3136, October 2017,
https://doi.org/10.1590/0001-3765201720160661

SARKER, Iqbal H. ***Machine Learning: Algorithms, Real-World Applications and Research Directions***. SN Computer Science, volume 2, article number :160, 2021, https://doi.org/10.1007/s42979-021-00592-x

TELSCHOW, Samira. ***Clinker Burning Kinetics and Mechanism***.2012.179 pages. Ph.D. thesis - Department of Chemical and Biochemical Engineering, Technical University of Denmark (DTU), Kongens Lyngby, Denmark, 2012. Available in:
https://orbit.dtu.dk/en/publications/clinker-burning-kinetics-and-mechanism

TENÓRIO, J. A.S; PEREIRA, S. S.R; FERREIRA, A. V.; ESPINOSA, D. C.R; BARROS, A.M; ARAUJO, F. G. ***CCT diagrams of tricalcium silicate decomposition***. Advances in Cement Research, volume 20, issue 1, pages 31-33, January 2008,
https://doi.org/10.1680/adcr.2008.20.1.31

THEISEN, Kirsten. ***Burnability: From raw meal to clinker***. Lecture 06-12. FLSmidth Institute, 2010.

University of Cambridge. ***MRC Cognition and Brain Sciences Unit***. Available in
https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/RatCaseVar
Accessed:18 of May 2023.

UWASU, Michinori; HARA, Keishiro; YABAR, Helmut. ***World cement production and environmental implications***. Environmental Development, volume 10, pages 36-47, 2014, ISSN 2211-4645,
https://doi.org/10.1016/j.envdev.2014.02.005

YAO, Le; JIANG, Xiaoyu; HUANG, Gaopan; QIAN, Jinchuan; SHEN, Bingbing; XU, Lu; GE, Zhiqiang. ***Virtual Sensing f-CaO Content of Cement Clinker Based on Incremental Deep Dynamic Features Extracting and Transferring Model***. IEEE Transactions on Instrumentation and Measurement, volume 70, pages 1-10, article number 2500610, 2021,
https://doi.org/10.1109/TIM.2020.3011251

ZHAO, Yantao; DING, Bochuan; ZHANG, Yuling; YANG, Liming; HAO, Xiaochen. ***Online cement clinker quality monitoring: A soft sensor model based on multivariate time series analysis and CNN***. ISA Transactions, volume 117, pages 180-195,2021, ISSN 0019-0578, https://doi.org/10.1016/j.isatra.2021.01.058.

ZIERI, Wilfred; ISMAIL, Ibrahim. ***Alternative Fuels from Waste Products in Cement Industry***. Handbook of Ecomaterials. Springer, Cham, pages 1183-1206, 2019, https://doi.org/10.1007/978-3-319-68255-6_142

9 APPENDIX

APPENDIX A - Individual features description after pre-processing

Table 16 - Individual features description after pre-processing (to be continued).

nº	Area	Feature / Parameter	Unit
1	**Raw meal**	Lime saturation factor -on line analysis	-
2	**at kiln feed**	Silica Modulus - on line analysis	-
3	**Quality**	Alumina Ratio - on line analysis	-
4	**Control**	Mg content - on line analysis	%mass
5	**Preheater**	Gas Temperature at Cyclone 1 Outlet	℃
6	**Tower**	Gas Pressure at Cyclone 1 Outlet	mbar
7		Gas Temperature at Cyclone 2 Outlet	℃
8		Gas Pressure at Cyclone 2 Outlet	mbar
9		Gas Temperature at Cyclone 3 Outlet	℃
10		Gas Pressure at Cyclone 3 Outlet	mbar
11		Gas Temperature at Cyclone 4 Outlet	℃
12		Gas Pressure at Cyclone 4 Outlet	mbar
13		Gas Temperature at Cyclone 5 Outlet	℃
14		Gas Pressure at Cyclone 5 Outlet	mbar
15		Gas Temperature at Calciner Outlet	℃
16		Gas Pressure at Calciner Outlet	mbar
17		Gas Pressure at Kiln Inlet	mbar
18		Raw Meal Temperature at Cyclone 1	℃
19		Raw Meal Temperature at Cyclone 2	℃
20		Raw Meal Temperature at Cyclone 3	℃
21		Raw Meal Temperature at Cyclone 4	℃
22		Raw Meal Temperature at Cyclone 5	℃
23		Gas Temperature at Downcomer Inlet	℃
24	**Conditioning**	Gas Temperature at Downcomer Outlet	℃
25	**Tower + ID**	Water Flow at Downcomer (gas conditioning tower)	m^3/h
26	**Fan**	Gas Pressure at ID Fan Inlet	mbar
27		Gas Pressure at ID Fan Outlet	mbar
28		ID Fan – Power	kW
29		ID Fan – Speed	rpm
30	**Kiln**	Air Temperature at Tertiary Air	℃
31		Air Pressure at Tertiary Air	mbar
32		Main Drive - Electrical Current - Kiln	A
33		Kiln Speed	rpm
34		Main Drive - Power - (sum of both drives)	kW
35		Filling Degree - Kiln	%
36		Residence time - Kiln	min
37		Air Pressure at Kiln Hood	mbar

Table 16 - Individual features description after pre-processing (conclusion).

nº	Area	Feature / Parameter	Unit
38	**Cooler + Excess Air Fan**	FN300 Cooling Fan Speed	rpm
39		FN300 Pressure at Cooling Fan Outlet	mbar
40		FN300 Cooling Fan Power	kW
41		FN305 Cooling Fan Speed	rpm
42		FN305 Pressure at Cooling Fan Outlet	mbar
43		FN305 Cooling Fan Power	kW
44		FN310 Cooling Fan Speed	rpm
45		FN310 Pressure at Cooling Fan Outlet	mbar
46		FN310 Cooling Fan Power	kW
47		FN315 Cooling Fan Speed	rpm
48		FN315 Pressure at Cooling Fan Outlet	mbar
49		FN315 Cooling Fan Power	kW
50		FN320 Cooling Fan Speed	rpm
51		FN320 Pressure at Cooling Fan Outlet	mbar
52		FN320 Cooling Fan Power	kW
53		FN325 Cooling Fan Speed	rpm
54		FN325 Pressure at Cooling Fan Outlet	mbar
55		FN325 Cooling Fan Power	kW
56		Excess Air Temperature before Air-Air Heat Exchanger	℃
57		Excess Air Temperature after Air-Air Heat Exchanger	℃
58		Excess Air Fan Speed	rpm
59		Excess Air Fan Power	kW
60		Pressure Drop through to Bag Filter before Excess Air Fan	mbar
61		Dedusting Fan (located above clinker transport) - Electrical Current	%A
62		Cooler Grate Speed	strocks/ min
63		Residence time - Cooler	min
64	**Clinker Quality Control**	Lime saturation factor	-
65		Silica Modulus	-
66		Alumina Ratio	-
67		Total Alkalis as Na_2O	-
68		Liquid phase at 1450℃	%mass
69		Tricalcium Silicate (Alite)	%mass
70		Dicalcium Silicate (Belite)	%mass
71		Tricalcium Aluminate (Aluminate)	%mass
72		Tetracalcium Aluminoferrite (Ferrite)	%mass
73		Burnability Index	-
74		Calcium Oxide	%mass

APPENDIX B – Description of each feature selection for SIM04 and SIM18.

Table 17 - Description of each feature selection for SIM04 and SIM18.

	SIM04				SIM18			
Area Description	**where (x_i)**	**Process variables (*nv*)**	**Delay**	**Unit**	**where (x_i)**	**Process variables (*nv*)**	**Delay**	**Unit**
Raw meal at kiln feed Quality Control					x_{23}	Alumina Ratio - on line analysis	20 min	-
					x_{24}	Mg content - on line analysis	20 min	%mass
Preheater	x_1	Gas Temperature at Cyclone 1 Outlet	-	℃	x_{30}	Gas Temperature at Calciner Outlet	40 min	℃
Tower	x_3	Gas Temperature at Cyclone 5 Outlet	-	℃	x_5	Gas Temperature at Downcomer Inlet	-	℃
					x_{16}	Gas Temperature at Downcomer Inlet	20 min	℃
	x_5	Raw Meal Temperature at Cyclone 2	-	℃	x_{18}	Raw Meal Temperature at Cyclone 2	20 min	℃
					x_{19}	Raw Meal Temperature at Cyclone 3	20 min	℃
	x_{18}	Gas Pressure at Cyclone 4 Outlet	20 min	mbar	x_1	Gas Pressure at Cyclone 4 Outlet	-	mbar
	x_2	Gas Pressure at Cyclone 3 Outlet	-	mbar	x_2	Gas Pressure at Cyclone 5 Outlet	-	mbar
					x_3	Gas Pressure at Calciner Outlet	-	mbar
					x_{15}	Gas Pressure at Cyclone 5 Outlet	20 min	mbar
Conditioning Tower + ID Fan					x_6	ID Fan – Speed	-	rpm
	x_4	ID Fan – Power	-	kW	x_{17}	ID Fan – Power	20 min	kW
Kiln	x_6	Air Temperature at Tertiary Air	-	℃				
	x_{25}	Air Temperature at Tertiary Air	40 min	℃				
	x_8	Air Pressure at Kiln Hood	-	mbar	x_4	Air Pressure at Kiln Hood	-	mbar
	x_{19}	Air Pressure at Kiln Hood	20 min	mbar	x_{22}	Air Pressure at Kiln Hood	20 min	mbar
	x_{20}	Air Pressure at Tertiary Air	20 min	mbar	x_{31}	Air Pressure at Kiln Hood	40 min	mbar
	x_7	Filling Degree - Kiln	-	%	x_7	Filling Degree - Kiln	-	%
	x_{26}	Filling Degree - Kiln	40 min	%	x_{20}	Kiln Speed	20 min	rpm
					x_{21}	Main Drive - Power - (sum of both drives)	20 min	kW
Cooler + Excess Air Fan	x_9	FN305 Pressure at Cooling Fan Outlet	-	mbar	x_9	FN310 Pressure at Cooling Fan Outlet	-	mbar
	x_{10}	FN310 Pressure at Cooling Fan Outlet	-	mbar	x_{12}	FN325 Pressure at Cooling Fan Outlet	-	mbar
	x_{22}	FN315 Pressure at Cooling Fan Outlet	20 min	mbar				
	x_{12}	FN320 Pressure at Cooling Fan Outlet	-	mbar				
	x_{13}	FN325 Pressure at Cooling Fan Outlet	-	mbar				
	x_{14}	FN325 Cooling Fan Power	-	kW	x_{32}	FN305 Cooling Fan Power	40 min	kW
	x_{21}	FN310 Cooling Fan Power	20 min	kW	x_{25}	FN300 Cooling Fan Power	20 min	kW
	x_{27}	FN320 Cooling Fan Power	40 min	kW	x_8	FN300 Cooling Fan Speed		rpm
	x_{11}	FN315 Cooling Fan Speed	-	rpm	x_{26}	FN305 Cooling Fan Speed	20 min	rpm
					x_{10}	FN320 Cooling Fan Speed	-	rpm
					x_{11}	FN325 Cooling Fan Speed	-	rpm
	x_{15}	Excess Air Temperature before Air-Air Heat Exchanger	-	℃	x_{13}	Excess Air Temperature before Air-Air Heat Exchanger	-	℃
	x_{23}	Excess Air Temperature after Air-Air Heat Exchanger	20 min	℃	x_{14}	Excess Air Temperature after Air-Air Heat Exchanger	20 min	℃
	x_{17}	Cooler Grate Speed	-	strocks/ min	x_{28}	Cooler Grate Speed	20 min	strocks/ min
	x_{24}	Residence time - Cooler	20 min	min	x_{29}	Residence time - Cooler	20 min	min
	x_{28}	Residence time - Cooler	40 min	min				
	x_{16}	Dedusting Fan (located above clinker transport) - Electrical Current	-	%A	x_{27}	Dedusting Fan (located above clinker transport) - Electrical Current		%A
Clinker Quality Control	x_{29}	Tricalcium Silicate (Alite)	**previous*	%mass	x_{33}	Dicalcium Silicate (Belite)	**previous*	%mass
	x_{30}	Alumina Ratio	**previous*	%mass	x_{34}	Tetracalcium Aluminoferrite (Ferrite)	**previous*	%mass
	x_{31}	Total Alkalis as Na_2O	**previous*	-	x_{35}	Total Alkalis as Na_2O	**previous*	-
	x_{32}	Calcium Oxide	**previous*	%mass	x_{36}	Calcium Oxide	**previous*	%mass

**Previous value from quality control laboratory.*

APPENDIX C - Results for the final variables (wi (x)) and coefficients (bi) for multivariate polynomial models

Table 18 - SIM04 – Results for the final variables (w_i (x)) and coefficients (b_i) for multivariate polynomial models (MLR and degree 3).

Degree 1 (Multiple Linear Regression)			Degree 3 (Cubic function)		
Function (w_i)	w_i (x)	Coefficients (b_i)	Function (w_i)	w_i (x)	Coefficients (b_i)
	Interception (b_0)	0.00E+00			-4.94E+02
w_1	x_1	-1.87E-02	w_1	$(x_{10}.x_{32})/x_{15}$	5.40E+00
w_2	x_4	-2.39E-04	w_2	$x_6/(x_{10}.x_{25})$	-8.27E+01
w_3	x_5	1.37E-02	w_3	$x_4/(x_{14}.x_{23})$	1.14E+00
w_4	x_6	-1.66E-03	w_4	$x_{14}/(x_{12}.x_{30})$	1.49E+00
w_5	x_7	3.62E-02	w_5	$x_{26}.x_{26}.x_{32}$	-2.06E-03
w_6	x_8	-2.36E-01	w_6	$x_3/(x_{16}.x_{24})$	-3.33E+00
w_7	x_9	-3.20E-02	w_7	$x_{32}/(x_{15}.x_{30})$	-9.02E+02
w_8	x_{10}	5.78E-02	w_8	$x_4.x_{32}.x_{32}$	-9.53E-05
w_9	x_{11}	-9.07E-04	w_9	$x_{24}.x_{28}$	3.21E-03
w_{10}	x_{12}	4.39E-02	w_{10}	$(x_{24}.x_{24})/x_{28}$	7.80E-03
w_{11}	x_{13}	-8.76E-02	w_{11}	$x_{14}.x_{27}.x_{29}$	4.59E-06
w_{12}	x_{14}	1.39E-02	w_{12}	$(x_1.x_9)/x_{10}$	-3.44E-03
w_{13}	x_{15}	-1.53E-03	w_{13}	$1/(x_{15}.x_{16}.x_{24})$	8.59E+05
w_{14}	x_{16}	-1.25E-02	w_{14}	$(x_{26}.x_{31})/x_{27}$	3.52E+00
w_{15}	x_{18}	1.84E-02	w_{15}	$1/(x_6.x_{16}.x_{17})$	1.37E+04
w_{16}	x_{20}	-7.13E-02	w_{16}	$(x_{15}.x_{25})/x_{23}$	-1.20E-04
w_{17}	x_{21}	-5.37E-03	w_{17}	$x_5.x_{25}.x_{27}$	1.23E-08
w_{18}	x_{22}	9.22E-03	w_{18}	$x_7.x_{28}.x_{32}$	7.48E-03
w_{19}	x_{23}	-6.89E-03	w_{19}	$(x_{20}.x_{20})/x_{22}$	1.51E-01
w_{20}	x_{24}	-3.12E-02	w_{20}	$(x_1.x_8)/x_{27}$	1.87E-01
w_{21}	x_{25}	1.29E-03	w_{21}	$(x_8.x_{32})/x_{20}$	1.68E+00
w_{22}	x_{28}	1.13E-01	w_{22}	$x_{24}/(x_6.x_{14})$	-1.13E+02
w_{23}	x_{29}	2.80E-02	w_{23}	$(x_2.x_{19})/x_{32}$	2.70E-04
w_{24}	x_{30}	1.22E-01	w_{24}	$x_{10}.x_{27}.x_{30}$	-1.48E-04
w_{25}	x_{31}	1.33E+00	w_{25}	$1/(x_{11}.x_{12}.x_{16})$	2.75E+06
w_{26}	x_{32}	4.37E-01	w_{26}	$(x_8.x_8)/x_{32}$	8.49E-02
			w_{27}	$(x_7.x_8)/x_{13}$	-3.47E+00
			w_{28}	$x_{24}/(x_3.x_{11})$	-2.09E+05
			w_{29}	$(x_4.x_{14})/x_{31}$	5.52E-06
			w_{30}	$(x_{11}.x_{13})/x_{12}$	-1.72E-03
			w_{31}	$(x_1.x_{16})/x_{14}$	-1.82E-03
			w_{32}	$(x_2.x_{18})/x_{15}$	-5.77E-02
			w_{33}	$x_{27}.x_{29}.x_{32}$	7.91E-05
			w_{34}	$(x_6.x_6)/x_{15}$	-2.78E-04
			w_{35}	$x_{16}.x_{25}.x_{30}$	3.33E-06
			w_{36}	$x_{26}/(x_{21}.x_{28})$	-1.40E+01
			w_{37}	$x_{25}/(x_6.x_{31})$	-7.57E-01
			w_{38}	$x_{22}.x_{23}.x_{30}$	5.90E-05
			w_{39}	$(x_{21}.x_{23})/x_{15}$	-5.51E-02
			w_{40}	$(x_{14}.x_{14})/x_{30}$	-1.13E-03
			w_{41}	x_4/x_{10}	-5.44E-02

Table 19 - SIM18 – Results for the final variables (w_i ($\boldsymbol{x}$)) and coefficients (b_i) for multivariate polynomial models (MLR to degree 4).

Function (w_i)	Degree 1 (MLR) w_i ($\boldsymbol{x}$)	Degree 1 (MLR) Coefficients (b_i)	Degree 2 w_i ($\boldsymbol{x}$)	Degree 2 Coefficients (b_i)	Degree 3 w_i ($\boldsymbol{x}$)	Degree 3 Coefficients (b_i)	*Degree 4 w_i ($\boldsymbol{x}$)	*Degree 4 Coefficients (b_i)
Interception	b_0	-6.08E+00	b_0	0.00E+00	b_0	1.17E+01	b_0	-7.52E+01
w_1	x_1	3.35E-02	x_{36}/x_{27}	5.41E+01	$(x_4.x_{20})/x_{15}$	7.82E-02	$(x_9.x_{32}.x_{36})/x_{13}$	9.35E-02
w_2	x_3	-7.65E-02	$x_4.x_4$	3.40E-03	$(x_3.x_3)/x_9$	8.12E-01	$(x_{12}.x_{13}.x_{17})/x_{36}$	6.70E-09
w_3	x_4	-2.51E-02	$x_7.x_{14}$	-1.42E-03	$x_8/(x_6.x_{13})$	-3.41E+02	$x_{15}/(x_2.x_{20}.x_{26})$	6.48E+03
w_4	x_7	1.52E-01	x_9/x_{12}	3.21E+00	$(x_9.x_{33})/x_{12}$	2.06E-01	$(x_7.x_{20})/x_{19}$	1.03E+02
w_5	x_8	-3.20E-03	$x_{12}.x_{28}$	4.02E-03	x_{33}/x_{13}	3.82E+02	$x_3.x_{22}.x_{25}.x_{33}$	-3.57E-05
w_6	x_9	4.02E-02	$1/(x_{11}.x_{35})$	-1.90E+04	$x_{10}/(x_5.x_{13})$	4.68E+02	$(x_{13}.x_{35})/(x_6.x_8)$	-1.61E+04
w_7	x_{11}	2.41E-03	$x_9.x_{20}$	-8.33E-03	$x_3.x_3.x_{31}$	2.00E-03	$(x_{13}.x_{29})/(x_{12}.x_{25})$	4.38E-01
w_8	x_{12}	-5.38E-02	x_{22}/x_9	6.73E+01	$(x_9.x_{24})/x_{28}$	6.94E-02	$(x_{34}.x_{36}.x_{36})/x_{25}$	2.16E+00
w_9	x_{13}	-1.28E-03	x_3/x_1	7.13E+00	$x_{29}/(x_{21}.x_{24})$	1.23E+02	$(x_{13}.x_{23}.x_{36})/x_{34}$	4.49E-02
w_{10}	x_{14}	-1.51E-02	$x_6.x_8$	2.33E-06	$(x_{13}.x_{27})/x_9$	-3.96E-03	$(x_5.x_9)/(x_8.x_{18})$	1.38E+03
w_{11}	x_{16}	-4.89E-02	$x_8.x_{11}$	-2.70E-06	$(x_7.x_7)/x_{23}$	9.99E-03	x_{18}	1.64E-01
w_{12}	x_{17}	-7.26E-04	$x_{23}.x_{36}$	1.26E-01	$x_{33}/(x_{13}.x_{19})$	-2.74E+05	$x_{23}/(x_{13}.x_{29})$	6.63E+03
w_{13}	x_{18}	4.18E-02	$1/(x_{26}.x_{30})$	5.35E+06	$x_5.x_{15}.x_{15}$	-4.62E-06	$(x_{15}.x_{22}.x_{26})/x_2$	-1.82E-03
w_{14}	x_{20}	3.95E-01	x_1/x_3	6.72E-01	$(x_{17}.x_{20})/x_{13}$	-5.23E-02	$(x_{12}.x_{21}.x_{33})/x_{36}$	8.36E-07
w_{15}	x_{21}	-2.85E-03	$x_{13}.x_{24}$	6.05E-04	$x_{13}.x_{17}.x_{25}$	1.50E-08	$(x_{12}.x_{21}.x_{33})/x_{25}$	-4.24E-04
w_{16}	x_{22}	-3.04E-01	$x_{16}.x_{35}$	-6.46E-02	$(x_1.x_3)/x_{35}$	-3.80E-03	$(x_{17}.x_{22})/(x_3.x_{32})$	-7.56E-01
w_{17}	x_{24}	1.89E-01	$x_{18}.x_{35}$	2.53E-02	$(x_7.x_{36})/x_{26}$	9.18E+01	$x_3.x_4.x_5.x_7$	-3.39E-06
w_{18}	x_{25}	1.49E-02	$1/(x_8.x_{13})$	2.28E+06	$x_{15}/(x_2.x_{34})$	1.16E+01	$(x_6.x_{14}.x_{14})/x_{17}$	3.22E-04
w_{19}	x_{27}	-9.16E-03	$x_3.x_{31}$	-7.19E-02	$(x_{21}.x_{32})/x_8$	-2.60E-01	$(x_{13}.x_{25})/(x_7.x_{26})$	6.60E-01
w_{20}	x_{28}	4.00E-01	x_{36}/x_6	1.06E+03	$x_7.x_{33}$	-1.44E-02	$x_{12}.x_{14}.x_{22}.x_{24}$	1.23E-04
w_{21}	x_{29}	8.39E-02	$x_{22}.x_{26}$	-3.33E-03	$x_{20}.x_{31}.x_{33}$	6.90E-03	$x_{20}.x_{23}.x_{29}.x_{33}$	-2.32E-04
w_{22}	x_{30}	-4.26E-03	x_{18}/x_{11}	3.72E+01	x_{22}	-1.09E+01	$x_{34}/(x_7.x_{18}.x_{27})$	-4.14E+04
w_{23}	x_{33}	-3.70E-02	$x_6.x_{31}$	-1.01E-03	$(x_{11}.x_{21})/x_{30}$	1.56E-02	$(x_1.x_2.x_5)/x_{24}$	3.19E-05
w_{24}	x_{35}	1.61E+00	$x_8.x_{33}$	2.66E-04	$x_{28}/(x_{21}.x_{36})$	-1.04E+01	$x_7.x_{10}.x_{13}.x_{20}$	-1.36E-07
w_{25}	x_{36}	4.83E-01	$x_{10}.x_{21}$	5.50E-06	$(x_9.x_{22})/x_{19}$	-2.69E+02	$(x_9.x_{23}.x_{36})/x_{35}$	-6.65E-03
w_{26}			$x_{16}.x_{33}$	1.55E-03	$x_{13}.x_{14}.x_{27}$	1.21E-06	$(x_{12}.x_{24}.x_{27})/x_6$	3.94E-02
w_{27}			x_{36}/x_{32}	-1.64E+02	$(x_9.x_{10})/x_{16}$	-1.54E-02	$(x_{12}.x_{20}.x_{36})/x_{13}$	-2.08E+00
w_{28}			$x_{36}.x_{33}$	-1.79E-01	$(x_{16}.x_{36})/x_{27}$	-1.37E+00	$x_3.x_3.x_3.x_3$	2.84E-05
w_{29}			x_{11}/x_{-13}	1.59E+00	$x_8.x_{35}$	3.47E-03	$x_3.x_{15}.x_{17}$	-4.24E-06
w_{30}			x_5/x_{13}	-6.29E+00	$x_{18}/(x_{28}.x_{29})$	7.58E+00	$x_1/(x_3.x_{35})$	-4.38E-01
w_{31}			x_{21}/x_{20}	-4.61E-02	$x_{21}.x_{23}.x_{33}$	-1.84E-04	$x_5.x_{10}$	4.35E-05
w_{32}			x_{33}	-1.04E+00	$x_5/(x_{12}.x_{20})$	-1.03E+00	$x_5.x_5$	-2.21E-04
w_{33}			x_{29}/x_{36}	4.59E-03	$x_{34}.x_{36}.x_{36}$	-2.33E-02	$x_1.x_3.x_9.x_{18}$	-2.35E-07
w_{34}			$x_{17}.x_{19}$	-1.41E-06	$x_7.x_{13}$	-3.37E-04	$x_{10}.x_{16}$	-3.11E-05
w_{35}			$1/(x_{35}.x_{35})$	2.14E+00	$x_8.x_{11}.x_{35}$	-2.65E-06	$(x_1.x_4.x_{26})/x_{27}$	3.72E-04
w_{36}			$x_8.x_{22}$	2.40E-03	$(x_2.x_{28})/x_3$	1.01E-01	$x_7.x_{23}.x_{27}.x_{36}$	-5.32E-04
w_{37}			$1/(x_7.x_{13})$	-1.18E+04	$x_{12}/(x_{13}.x_{23})$	-2.33E+01	$(x_9.x_{23}.x_{28})/x_{34}$	-3.40E-02
w_{38}					$x_{34}/(x_{20}.x_{32})$	4.93E+01	$(x_{13}.x_{36}.x_{36})/x_{34}$	-1.36E-02
w_{39}					$x_{36}/(x_{24}.x_{34})$	2.63E+01	$(x_1.x_3.x_{24})/x_{32}$	1.20E-01
w_{40}					$x_{24}/(x_7.x_{36})$	-1.55E+00	$(x_1.x_1.x_{27})/x_{13}$	-3.90E-03
w_{41}					$x_{12}.x_{30}.x_{32}$	8.70E-07	$x_9.x_{10}.x_{33}.x_{33}$	-2.45E-07
w_{42}					$1/(x_{14}.x_{36})$	5.50E+01	$x_9.x_{16}.x_{33}.x_{33}$	1.56E-06
w_{43}					$(x_{22}.x_{26})/x_{16}$	1.95E+00	$x_{13}.x_{24}.x_{33}$	5.71E-05
w_{44}					$x_9.x_{22}$	3.59E-01	$1/(x_1.x_3.x_3.x_4)$	1.45E+02
w_{45}					x_5/x_{12}	4.41E-01	$(x_9.x_{18})/x_8$	-1.40E+00
w_{46}					$(x_{17}.x_{36})/x_{23}$	-1.01E-03	$(x_{14}.x_{30})/x_{23}$	-4.85E-05
w_{47}					$x_{23}.x_{36}$	-3.20E-01	$x_{12}.x_{21}.x_{28}$	2.49E-05
w_{48}					$(x_{26}.x_{36})/x_{27}$	3.60E-01	$(x_5.x_8.x_{36})/x_{20}$	-2.54E-06
w_{49}					$x_{14}.x_{36}$	2.39E-01	$(x_{23}.x_{36})/(x_{10}.x_{10})$	-6.08E+05
w_{50}					$(x_{12}.x_{22})/x_{32}$	7.50E+00	$x_{33}.x_{33}$	-1.46E-02
w_{51}					$x_8.x_{29}$	-1.66E-04	$x_{13}/(x_{23}.x_{30})$	5.21E+00
w_{52}					$x_{14}.x_{23}.x_{31}$	-4.76E-03	$(x_{13}.x_{35})/x_{12}$	4.14E-01
w_{53}					x_{16}	-4.70E-02	$x_5/(x_{23}.x_{36})$	-1.08E-03
w_{54}					x_{19}	-4.69E-02	$(x_4.x_{23}.x_{35})/x_3$	-4.86E-01
w_{55}					$(x_{16}.x_{22})/x_{15}$	-2.90E-02	$(x_5.x_8.x_{21})/x_{10}$	-1.89E-05
w_{56}					$(x_{14}.x_{26})/x_{24}$	-7.27E-05		
w_{57}					$(x_{14}.x_{36})/x_{19}$	-1.66E+02		

Model with higher R^2 (85% for training and 15% for testing).