POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

New General Framework for Ferrotitanium Process Control

MAHAN BALAL POUR

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie industriel

Décembre 2019

© Mahan Balal Pour, 2019.

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

New General Framework For Ferrotitanium Process Control

présenté par **Mahan BALAL POUR**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Bruno AGARD, président
Robert PELLERIN, membre et directeur de recherche
Vahid PARTOVI NIA, membre et codirecteur de recherche
Samuel-Jean BASSETTO, membre

DEDICATION

This thesis dedicates to my dear mother Nasrin Nazari, my dear father Abdalrasool Balalpour, and my dear wife Samira Namdary.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my advisor Prof. Robert Pellerin for his constructive support and commitment, I would like to thank him for his significant impact on my work and for the productive lessons that he taught me during my education.

I would like to thank Prof. Vahid Partovi Nia, who gave me influential support in the whole stage of this work with his supportive aid.

Additionally, I would like to thank all of my colleagues for their help during the academic years in the Metalliage company.

Finally, I want to express my deepest gratitude to my parents and my wife for their endless supports during this work.

RÉSUMÉ

Les producteurs d'acier sont connues pour être à l'origine d'autres industries manufacturières. Les sidérurgistes sont toujours en concurrence pour être en avance sur les autres en matière de qualité. Pour améliorer leurs produits, ces entreprises tentent toujours d'améliorer la qualité de leurs produits en appliquant des additifs chimiques. L'un de ces additifs est appelé ferroalliage. Le ferroalliage est une combinaison du métal avec une grande quantité d'alliages chimiques différents tels que le vanadium, le titane, le silicium, l'aluminium, etc. Chacun de ces ferroalliages a un effet constructif sur les performances de l'acier produit.

Le contrôle statistique des processus (SPC) est la méthode la plus utilisée par les producteurs de ferrotitane pour contrôler leurs processus. Le SPC a commencé à être utilisée dans les industries à partir des années 1920 et son utilisation a augmenté significativement pendant la Seconde Guerre mondiale. Les entreprises qui utilisent le contrôle statistique des processus gardent leurs principaux processus sous contrôle pour surveiller la stabilité, la capacité et les performances du processus. Cette méthode statistique est la façon la plus compréhensible et la plus simple de mettre en œuvre des méthodes de contrôle statistique univariées sur les processus. Mais dans la plupart des processus réels, il existe toujours plusieurs variables dans lesquelles les entreprises tentent de les contrôler. Par conséquent, l'utilisation de la méthode statistique univariée pour surveiller toutes leurs variables leur pose certaines difficultés. Il ne peut pas considérer les corrélations possibles entre les variables et la surveillance de plusieurs cartes de contrôle en même temps est impossible pour les opérateurs. Chacune des cartes de contrôle a sa propre signification et les données de contrôle ne peuvent pas être collectées dans une carte unique. Par conséquent, lorsqu'il y a un produit hors contrôle, il faut beaucoup de temps et d'énergie pour déterminer quelle variable était responsable de l'erreur.

L'objectif principal de cette recherche est de développer un cadre général pour déterminer les variables principales à contrôler dans le processus ferrotitanium et pour déterminer l'importance de chacune de ces variables. Dans ce modèle, après la collecte de l'ensemble de données historiques et après la préparation des données, nous analysons l'ensemble de données général, à l'aide de méthodes d'apprentissage supervisées et non supervisées, pour trouver la corrélation linéaire et non linéaire entre les variables et pour trouver les éléments les plus importants à surveiller. Les

méthodes d'apprentissage supervisées et non supervisées proposées ont été appliquées sur un sous-ensemble de données historiquespour valider l’approche proposée. Nos résultats montrent qu'en utilisant des méthodes d'apprentissage supervisé telles que la régression linéaire multiple et la méthode de forêt aléatoire, nous pouvons déterminer les principaux éléments de chaque variable de réponse et la priorité de chaque prédicteur par rapport à l'ensemble de données générales et au sous-ensemble de données de produits sélectionné. Dans les méthodes d'apprentissage non supervisées, en utilisant l'analyse des composants principaux comme méthode principale de contrôle des processus multivariés, toutes les variables cachées sont ainsi étudiées et analysées.

ABSTRACT

Steel producer companies are known as the root of other manufacturing industries. Steelmakers are always competing with each other to be ahead of the others in quality matters. To improve their products, these companies always are trying to improve the quality of their products by applying some chemical additives. One of these main additives is called ferroalloy chemical. Ferroalloy is a combination of the metal with a high amount of some different chemical alloys such as vanadium, titanium, silicon, aluminum, etc. Each of these ferroalloys has a constructive effect on the performance of produced steel (Holappa, 2010).

Statistical process control (SPC) is the most popular method between ferrotitanium producers to control their processes. SPC started to be used over the industries from early 1920s and its use has grown during the World War II. Companies that use the statistical process control keep their main processes under control to monitor the stability, capacity, and process performance. This statistical method is the most understandable and the easiest way to implement univariate statistical control methods over the processes. However, in most of the real processes, there are always several variables in which companies try to control them. Therefore, using the univariate statistical method to monitor all their variables generate some difficulties. They cannot consider the possible correlations between the variables and monitoring several control charts at the same time is impossible for operators. Each of the control charts has its significance level and they cannot be collected in a unique chart. Therefore, when there is any out of range product in the ferrotitanium process, it takes a lot of time and energy to find out which variable was responsible of the error.

The main goal of this research is to develop a general framework to determine the main variables to be controlled in the ferrotitanium process, and to determine the importance of each of these variables. In this model, after collecting the historical dataset and after data preparation, we analyze using supervised and unsupervised learning methods the general dataset to find out the linear and non-linear correlation between variables and to find out the most important elements of each response variable and to analyze whether the production results are predictable. Based on this result, the supervised and unsupervised learning methods was reapplied over a selected product data subset to see if the results match the general dataset results. Our results show that using supervised learning methods such as multiple linear regression and random forest method, we can

determine the main elements of each response variable and the priority of each predictor over both general dataset and selected product data subset. In unsupervised learning methods, using the principal component analysis as the main method of multivariate process control, all the hidden variables can be studied and analysed.

TABLE OF CONTENTS

DEDICATION III

ACKNOWLEDGEMENTS IV

RÉSUMÉ V

ABSTRACT VII

TABLE OF CONTENTS IX

LIST OF TABLES XII

LIST OF FIGURES XIV

LIST OF SYMBOLS AND ABBREVIATIONS XV

LIST OF APPENDICES XVII

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 LITERATURE REVIEW 3

2.1 Statistical methods for quality control 3

2.1.1 Univariate statistical process control (SPC) 3

2.1.2 Multivariate statistical process control 4

2.1.3 Design of experiments 6

2.2 Machine learning in manufacturing 7

2.2.1 Supervised machine learning 9

2.2.2 Unsupervised machine learning 12

2.3 Critical analysis 15

2.4 Conclusion 16

CHAPTER 3 RESEARCH METHODOLOGY 18

3.1 Research objectives and research questions 18

3.2 Research methodology 18
3.2.1 Phase 1: Define the objectives 20
3.2.2 Phase 2: Choose response variables 20
3.2.3 Phase 3: Identify important factors 21
3.2.4 Phase 4: Propose the model 22
3.2.5 Phase 5: Validate and analyze the results 22
3.2.6 Phase 6: Act 22
3.2.7 Phase 7: Report 23
3.3 Case study 23
3.4 Conclusion 24
CHAPTER 4 MODEL PROPOSAL 25
4.1 Proposed model 25
4.2 Collecting historical dataset 28
4.3 Data preparation and sorting out 28
4.4 Multiple linear regression over the dataset 30
4.5 Random forest 31
4.6 Result comparison between two methods 31
4.7 Principal component analysis 32
4.8 Consistency check and reapply the methods over a data subset 33
4.9 Conclusion 34
CHAPTER 5 MODEL VALIDATION AND NUMERICAL EXAMPLE 35
5.1 Collecting historical dataset 35
5.2 Data preparation and sorting out 35

5.3 Multiple linear regression over the dataset 39

5.4 Random forest 45

5.5 Result comparison between two methods 48

5.6 Principal components analysis 50

5.7 Consistency check and reapply the methods over a data subset 53

5.7.1 Multiple linear regression over data subset.....53

5.7.2 Random Forest over the data subset.....57

5.7.3 Linear regression and random forest results comparison over the data subset 61

5.7.4 PCA over data subset 62

5.8 Discussion 64

5.9 Conclusion.....68

CHAPTER 6 CONCLUSION AND RECOMMENDATIONS..... 69

REFERENCES..... 71

APPENDICES..... 75

LIST OF TABLES

Table 2.1 Pros and cons of SPC and MSPC....6

Table 2.2 Summary of reviewed statistical methods....15

Table 3.1 Dependent and independent variables....21

Table 5.1 Description of datasets' columns....35

Table 5.2 Multiple linear regression results of the equation (5.1) 39

Table 5.3 Multiple linear regression results of the equation (5.2) 40

Table 5.4 Multiple linear regression results of the equation (5.3) 41

Table 5.5 Multiple linear regression results of the equation (5.4) 41

Table 5.6 Multiple linear regression results of the equation (5.5) 42

Table 5.7 Multiple linear regression results of the equation (5.6) 43

Table 5.8 Multiple linear regression results of the equation (5.7) 43

Table 5.9 Multiple linear regression results of the equation (5.8) 44

Table 5.10 Summary of regression result over the general dataset....44

Table 5.11 Random forest results of equation (5.10)....45

Table 5.12 Random forest results of equation (5.11)....46

Table 5.13 Random forest results of equation (5.12)....47

Table 5.14 Random forest results of equation (5.13)....47

Table 5.15 Random forest results of equation (5.14)....48

Table 5.16 Comparison of LR and RF results over the general dataset....49

Table 5.17 General dataset main chemical components based on the random forest test 51

Table 5.18 Principal component results of the general dataset....51

Table 5.19 Multiple linear regression results of the equation (5.16) 53

Table 5.20 Multiple linear regression results of the equation (5.17) 54

Table 5.21 Multiple linear regression results of the equation (5.18) 54

Table 5.22 Multiple linear regression results of the equation (5.19) 55

Table 5.23 Multiple linear regression results of the equation (5.20) 55

Table 5.24 Multiple linear regression results of the equation (5.21) 56

Table 5.25 Summary of regression result over the selected product data subset (S) and the general dataset(G) 57

Table 5.26 Random forest results of equation (5.22).....58

Table 5.27 Random forest results of equation 5.23.....58

Table 5.28 Random forest results of equation (5.24).....59

Table 5.29 Random forest results of equation (5.25).....60

Table 5.30 Random forest results of equation (5.26).....60

Table 5.31 Comparison of linear regression(LR) and random forest(RF) results over the data subset 61

Table 5.32 Data subset main chemical components based on the random forest test.....62

Table 5.33 Principal component results over the selected data subset(S) compared with the general dataset(G) 63

Table 5.34 Regression results between general dataset and selected product data subset.....65

Table 5.35 Random forest results between general dataset and selected product data subset.....66

LIST OF FIGURES

Figure 3.1 DOE visualized definition 19

Figure 3.2 Steps involved in the design of experiments (Durakovic 2017).....20

Figure 3.3 Combination of Phases 4 and 5.....22

Figure 4.1 Proposed general model for ferrotitanium process control.....27

Figure 4.2 Control chart for vanadium.....29

Figure 4.3 Dataset structure 29

Figure 4.4 Response, predictor, and potential hidden variables.....30

Figure 4.5 Potential correlation between two response variables 33

Figure 4.6 Visual explanation for deducting direct effect of each predictor variable from response variables 33

Figure 5.1 Scatter plot to show the relationship between predictors 36

Figure 5.2 Correlation plot of predictors.....37

Figure 5.3 Five principal components plot of general dataset showing the observation record of data 52

Figure 5.4 Five principal components plot of data subset showing observation record of data 64

LIST OF SYMBOLS AND ABBREVIATIONS

Al	Aluminum
ARL	Applied Research Laboratories
C	Carbon
Cr	Chromium
CSV	Comma Separated Values
Cu	Copper
DOE	Design of Experiments
Fe	Iron
FeTi	Ferrotitanium
KIB	Keep In Bin
LASSO	Least Absolute Shrinkage and Selection Operator
LR	Linear Regression
LCL	Lower Control Limit
LECO	Laboratory Equipment Corporation
Mn	Manganese
Mo	Molybdenum
MSPC	Multivariate Statistical Process Control
N	Nitrogen
Ni	Nickle
O	Oxygen
PCA	Principal Component Analysis
RF	Random Forest

SPC	Statistical Process Control
SVM	Support vector machines
Si	Silicon
Sn	Tin
Ti	Titanium
UCL	Upper Control Limit
V	Vanadium
Zr	Zirconium

LIST OF APPENDICES

Appendix A r-studio codes.....75

CHAPTER 1 INTRODUCTION

Steel producers are considered as the foundation of other manufacturing industries. Steelmakers always compete to excel in quality and technical advancement. To improve the quality of their products, these companies apply chemical additives in their molten steel. One of these main inputs is a range of ferroalloys. A ferroalloy is a metal combined with a blend of alloys such as vanadium, titanium, silicon, aluminum, etc. Each of these ferroalloys has a constructive effect on the performance of produced steel in different applications such as aerospace, automotive, medical part production, construction and other industries (Holappa, 2010). Leading ferrotitanium producing countries include Brazil, China, India, Japan, Russia, Ukraine, the United Kingdom, the United States and Canada.

Ferrotitanium is used by steelmakers as a stabilizer to prevent chromium carbide forming at grain boundaries and in the production of low-carbon steels. During steelmaking, titanium is usually introduced as ferrotitanium because of its relatively low melting temperature and high density. Steels with relatively high titanium content include interstitial-free, stainless and high-strength low-alloy steels. Titanium reacts very fast with nitrogen, oxygen, carbon, and sulphur, and forms unsolvable composites. It is lighter, stronger and has higher resistance of corrosion compared with iron. Ferrotitanium is usually produced by induction melting of titanium scrap with iron or steel. However, it is also produced directly from titanium mineral concentrates. The standard grades of ferrotitanium are 30% and 70% titanium.

The company that initiated this research has become the leading manufacturer of ferrotitanium in Canada over the last twenty years. This company specializes in the crushing of FeTi of different grades and sizes ranging from standard, low carbon, low aluminum, low vanadium grades and sizes from 10 to 50 mm, 10 to 30 mm, 6 to 12 mm, until 0 to 2 mm FeTi powder based on the feeding requirements at steel mills. In this company, univariate statistical process control charts are used to control the client's required chemical components.

By applying some standard recipes, they produce high-quality products, except when they notice instabilities in chemical components that are not predictable. This causes the production of a low-quality product, even though the SPC method is in place. These off-grade products generate too much cost for this company because the only way to recover these off-grade products is to re-melt

and reproduce the batch, which is very expensive and time-consuming. Production of out of range material currently stands at 2.2% of total production, which equals 229,280 pounds per year.

Despite the importance of Ferroalloys in steel manufacturing, there is no formal guide to determine which variables are responsible for out of range products and there is no guide to recognize which variable is more important than the others. In this industry, identifying and ranking the element(s) that affect the essential and targeted chemical alloys can contribute to achieve stable quality and to reduce the cost of production. As such, the ***main objective of this research project is to develop a general framework to determine the main variables to be controlled in the ferrotitanium production process***. More specifically, this work determines whether or not there is any relation between dependent and independent variables in the ferrotitanium dataset, and if so, what are the most important elements of each dependent variable.

This report is organized as follows. Chapter 2 presents the related researches to reach an insight into statistical process control methods, machine learning algorithms, and the related subjects in supervised learning models such as multiple regression, random forests, and principal component analysis, and unsupervised learning methods. Chapter 3 then describes the proposed research methodology, including research objectives, research strategy and the case study. Our research methodology is based on the design of experiments approach. Chapter 4 includes the proposed model and an explanation of each step. Chapter 5 presents the implementation of the proposed model using the R-studio program. Data cleansing, training the data using supervised learning methods, and results obtained with the selected product dataset are presented. Principal component analysis as one of the main methods of multivariate statistical process control is presented, and results are compared with other methods in Chapter 5. Chapter 6 concludes this report with a discussion of the main contribution of this project as well as its limitations and future directions to continue this research.

CHAPTER 2 LITERATURE REVIEW

Chapter 1 emphasizes the importance of generating a statistical framework to determine the main variables of each target alloy in the ferrotitanium industry and to prioritize them, which can contribute to the stable quality and lower production costs in the ferrotitanium industry. This chapter intends to analyze the relevant works on quality prediction approaches that can be applied to the ferrotitanium process. In Section 2.1 we review the popular statistical control methods in this industry, which can improve the quality of the products. The popular machine learning methods that can develop a pattern from the existing data sets are presented in Section 2.2. A critical analysis of all this literature and the limitations and gaps in existing models is presented in Section 2.3.

2.1 Statistical methods for quality control

The statistical process control is known as one of the main quality control methods over the manufacturing companies (Montgomery, 2009). Univariate and multivariate statistical process control and design of experiments method are explained in the following sections as the statistical models which help to control the essential and targeted variables in manufacturing companies to improve the quality.

2.1.1 Univariate statistical process control (SPC)

Over the years, Statistical Process Control (SPC) becomes one of the most popular process control tools in industries as it can control critical manufacturing processes using easy-to-understand formulas and graphs. SPC is based on comprehensive fundamental principles, is easy to use, has substantial power, and can be applied to any process. Its seven major tools are histograms or stem-and-leaf plots, check sheets, pareto charts, cause-and-effect diagrams, defect concentration diagrams, scatter diagrams, and control charts (Montgomery, 2009). Historically, the 1920s was the start of Statistical Process Control (SPC) in industrial processes. Later, with the introduction of the six sigma methods, the application of statistical control methods increased due to the systematic usage in its procedure. Besides all technical aspects needed to implement statistical control methods, organizational and behavioural consideration by management is required to run this method successfully. In the 1980s, the demand for the implementation of SPC shifted to the control of critical processes instead of inspecting all final products. Such an application increases

productivity and decreases the variability of the products as well as making the consumption of energy more efficient (Toledo, Lizarelli, Junior, & Bispo, 2017).

2.1.2 Multivariate statistical process control

SPC is a statistical method that controls the stability and capability of a single variable of a process. The most popular tools in SPC are control charts, especially Shewhart control charts and capability indices. However, in most processes, several variables should be under control and supervised simultaneously. The equivalent of SPC in the multivariate domain is Multivariate Statistical Process Control (MSPC). It consists of monitoring multiple variables simultaneously in a defined process. The main tools of MSPC are T2 Hotelling charts and multivariate indices to control the capability. Principal component analysis (PCA) and clustering methods, which are classified under unsupervised machine learning methods, are also popular statistical methods used for MSPC implementation.

Among research in that domain, Kharbach et al. (2017) conducted a case study and applied MSPC in the pharmaceutical industry. Their study shows the difference between SPC and MSPC results and explains how the system could be out of control using MSPC control charts while it is presented under control using the SPC method.

The study of (Zou & Qiu, 2009) discusses the application of the LASSO statistical model in the MSPC quality control process over the parametric datasets. The authors advise conducting more research on the nonparametric datasets.

The use of latent structure-based multivariate statistical process control methods is also recommended as efficient quality improvement tools in a massive data context (Ferrer, 2014).

Using a different approach, Liu et al. (2017) discuss the application of the PCA method in the antibody production processes. They developed a multivariate statistical process control (MSPC) model to monitor cell cultures batch-wise for antibody production using in-line Raman spectroscopy. In a similar manner, Chen et al. (2000) applied PCA using Kernel Density Estimation (KDE) method. According to this study, applying KDE results in a better procedure for the selection of the right bandwidth, particularly in the case of highly clustered data with small variability.

In line with other researchers in that field (Rogalewicz & Poznańska, 2013), deal with the problem of controlling manufacturing processes with the use of statistical methods. They compared traditional (univariate) and multivariate (considering many variables and relationships between them) statistical process controls. The advantages and disadvantages of both approaches were pointed out. According to the author, the complicated mathematical calculations behind the MSPC make the implementation of MSPC difficult for engineers. This study emphasizes the essential need for an easy-to-understand way of applying this method. In addition, Rogalewicz (2012) explains the lack of guidelines for multivariate statistical control and its control charts. According to the author, there is a demand for having a step-by-step guideline for implementing the MSPC and its control charts. Similarly, Ferrer (2014) emphasizes the need to have a clear procedure and software to solve the lack of understandable statistical guidelines for the implementation and interpretation of MSPC control charts.

Table 2.1 presents a comparison of univariate and multivariate statistical process control methods. In general, the implementation of univariate SPC is much easier than multivariate SPC. The interpretation of control charts of MSPC needs some statistical expertise and much more complicated than univariate SPC control charts. The statistical algorithms which are used for MSPC are also more complex than SPC statistical methods. The main weakness of the control charts of multivariate SPC is that there is no possibility to easily determine which variable is out-of-range while the system is an alarming process is out-of-control. In this case, we should check all the univariate control charts to see which variable is not working in the range (Rogalewicz & Poznańska, 2013).

Table 2.1 Pros and cons of SPC and MSPC

Measure	Univariate control charts	Multivariate control charts
Acceptance by industry	very popular	Rarely used
Implementation	There are many guides and manual – Easy to apply	Lack of instruction and clear methodology – Very difficult to apply
Software exigence	Not a requirement but advisable	Necessary
Relationship between variable	Not consider	Consider
Multiple variable measurements.	It is hard to monitor several univariate control charts – Very complicated for the quality operator	Multiple variables could be controlled at the same control chart
Significance level	Every control chart has its significant level – very difficult to calculate	For all principal components, there is the same significant level
Out-of-control signal interpretation	Very easy and straightforward	Interpretation is practically impossible without the use of special methods
Unit of controlled statistic	The controlled statistic is in a unit of the controlled variable	The controlled statistic has no unit
Understanding the mechanism of functioning	Very easy	Complicate

2.1.3 Design of experiments

Design of Experiments (DOE) is a multipurpose tool that can be used in various situations such as design for comparisons, variable screening, transfer function identification, optimization, and robust design. The usage of DOE has been growing rapidly in manufacturing as well as non-manufacturing industries in the past 20 years. It was the most popular tool in scientific areas of biochemistry, engineering, medicine, physics, computer science and counts about 50% of its applications compared to all other scientific fields (Durakovic, 2017).

The design of experiments can be used in various positions for the identification of important input factors (predictor variable) and how they are related to the outputs (response variable). Therefore, DOE is mainly a regression analysis that can be used in different situations. It has some commonly used design types, such as *comparison* to select the best option and uses t-test, Z–test, or F–test.

Variable screening is another design type using a two-level factorial design to select important factors (variables) among several factors that affect performances of a system, process, or product. *Transfer function identification* is another commonly used design type in DOE that, if important input variables are identified, the relationship between the input variables and output variable can be used for further performance exploration of the system, process or product via transfer function. *System Optimization* is another design type of DOE where the transfer function can be used for optimization by moving the experiment to an optimum setting of the variables. Finally, the ***robust*** design such as Taguchi design intends to decrease variation in the system, process or product without elimination of its causes which can be considered in three main groups (Durakovic, 2017)

- external/environmental (temperature, humidity, and dust),
- internal (parts corrosion and ageing of materials), and
- unit to unit variation (variations in material, processes, and equipment).

2.2 Machine learning in manufacturing

Machine learning can be applied to determine patterns from the existing data sets or to predict the future behaviour of manufacturing systems. This generated knowledge can help the related process managers in their decision-making process and to improve the system automatically. Detecting certain patterns to describe data relations is the main purpose of machine learning (Alpaydin, 2009). The manufacturing industry today is facing a never seen increase in the presented data. These data include a variety of different structures, semantics, quality, e.g. sensor data from the production line, environmental data, machine tool parameters, etc. (Davis et al., 2015).

Many studies presented the use of machine learning for production or process control of the steel industry. For example, Umeshini and PSumathi (2017) explain the methods of data mining and describes its application in the industrial atmosphere and especially, in the steel industry, using matured theory like data mining and knowledge discovery to acquire new knowledge. The authors introduced clustering algorithms, then a clustering analysis of blast furnace operation parameters was carried out by K-means clustering. Analysis and comparison of practical data were conducted to determine the optimal number of clusters for blast furnace parameters analysis. Finally, the factor

analysis method to reduce the dimension of parameters was implemented, and the mining test of Tangshan iron and steel shows that the method is effective in practical application.

Laha et al. (2015) consider random forests, neural networks, dynamic evolving neuro-fuzzy inference system and support vector regression as competitive learning tools to verify the suitability of applications of these approaches and investigate their comparative predictive ability.

Prediction of wear loss quantities of ferroalloy coating using different machine learning algorithms can also be achieved Altay et al. (2019). Experimental wear losses under different loads and sliding distances of AISI 1020 steel surfaces coated with (wt.%) 50FeCrC-20FeW-30FeB and 70FeCrC-30FeB powder mixtures by plasma transfer arc welding were determined. The dataset comprised 99 different wear amount measurements obtained experimentally in the laboratory. The linear regression, support vector machine, and Gaussian process regression algorithms are used for predicting wear quantities.

Hansson et al. (2016) investigate characteristics of the most popular classifiers used at present industry using Support Vector Machines (SVM), Multilayer Perceptron, Decision Trees, Random Forests, and the meta-algorithm Bagging and Boosting. In another paper, an optimal method for prediction and adjustment on by-product gas holders in the steel industry is proposed (Zhang, Zhao, Wang, Cong, & Feng, 2011). Both single and multiple gasholders level prediction models are established by machine learning methodology. Furthermore, a hybrid parameter optimization algorithm is developed to optimize the model for high prediction accuracy. Then, based on the predicted gasholder level, the optimal adjustment amount is calculated by a novel reasoning method to sustain the gasholder within a safety zone.

A fault diagnosis method based on modified SVM for steel plates fault diagnosis (Tian, Fu, & Wu, 2015). With this method, the dimension of samples is effectively reduced by recursive feature elimination algorithm, and computing time is saved as well. Classification accuracy is improved by parameter optimizing and sample size balancing strategy. A faults dataset of steel plates is taken as a practical case, and SVM that are modified by different algorithms are utilized to complete fault diagnosis. This merged measure shows its dominance in sorting common faults of steel plates over the original SVM. Some essential procedures in model development, such as normalization and cross-validation, are also discussed by the authors.

Despite all the related research in steel manufacturing, how to determine the accurate reaction of each feedstock chemical alloys on the final products remains a challenge to all metallurgists in the ferrotitanium industry. Based on the reviewed research, there are very rare studies in this field to determine the main independent alloys and their importance on each dependant chemical element. As ferrotitanium is produced by numerous chemical reactions, the result of the added elements depends on many factors, which measurements are rough and tough. The decision on the exact number of feedstocks to be added often relies on a combination of knowledge-based models and expert perception of the engineers and operators, leading to many inaccurate results. Machine learning is an advanced approach that allows us to consistently increase the quality of decision-making, resulting in a decrease in the overall ferrotitanium out-of-range products. In the following sections, the main machine learning methods under two general categories of supervised and unsupervised learning methods are described.

2.2.1 Supervised machine learning

In the last two decades, machine learning algorithms have been progressed as an important part of information technology. This knowledge is tremendously useful in modelling complex logic. As a subfield of computer science, machine learning has generated new knowledge in statistical science. Machine learning is a method of data analysis that systematizes analytical model building. It is a division of artificial intelligence created on the idea that systems can learn from data, identify patterns and make decisions with minimal human interference. Supervised statistical learning contains structuring a statistical model for predicting or estimating an output based on one or more inputs. These types of problems will be more in the field, like a business, medicine, astrophysics, and public policy. In the supervised learning setting, we typically have access to a set of p features $x_1, x_2, \ldots x_p,$ measured on n observations, and a response y also measured on those same n observations. The goal is then to predict y using $x_1, x_2, \ldots x_p$ (James, Witten, Hastie, & Tibshirani, 2013). Statistical methods such as multiple regression, naïve Bayes classification, logistic regression, support vector machines, KNN and random forests are known as supervised machine learning methods.

2.2.1.1 Multiple regression

When we are dealing with a single predictor variable, simple linear regression is a suitable method for predicting response. However, in real life, there are often several predictor variables that have effects on the dependent variable. In multiple linear regression methods, by determining a separate slope coefficient, we try to find a single model to predict the relationship between the dependent and independent variables by finding a separate slope coefficient in a single model. As such, suppose that we have p distinct predictors. Then the multiple linear regression model will be as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_p x_p + \varepsilon \qquad (2.1)$$

where x_j represents the j th predictor, and β_j quantifies the association between that variable and the response, and ε is the measurement error. We interpret β_j as the average effect on y of a one-unit increase in x_j, holding all other predictors fixed (James et al., 2013).

2.2.1.2 Naive bayes classification

The naive Bayes classifier technique is established on Bayes theorem, and it is applied when the dimensionality of the inputs is high (Jadhav & Channe, 2016). It assumes that all variables contribute to classification and are reciprocally correlated. This concept is called class conditional independence. It is also called simple Bayes, idiot's Bayes, and independent Bayes. As the probability that a presented data item belongs to a class label, they can predict class membership probabilities. It considers that the presence or absence of a feature of a class is irrelevant to the presence or absence of any other feature when the class variable is given.

2.2.1.3 Logistic regression

Logistic regression is usually implemented where the border between the classes exists, also declares the class probabilities depend on distance from the boundary. Consider again the Default data set where the response default falls into one of two categories, Yes or No. Rather than modelling this response Y directly, logistic regression models the probability that Y belongs to a category. This moves towards the boundaries (0 and 1) more rapidly when the data set is larger. These statements about probabilities make logistic regression more than just a

classifier. It makes predictions stronger, more detailed and can be fit differently, but those strong predictions could be wrong. Logistic regression is a prediction approach. However, with logistic regression, prediction results in a branched outcome. Logistic regression is one of the most frequently used tools for applied statistics and discrete data analysis (Osisanwo et al., 2017).

2.2.1.4 Support vector machines

SVM have become increasingly widespread after its launch in the late 1990s, particularly within the Machine Learning community (Cortes & Vapnik, 1995). SVM functions have been successfully developed in numerous areas, such as bioinformatics, which is possibly the most rapidly developing discipline in terms of new methodologies due to the recent explosion of data volumes, econometrics, and biometrics. SVM have been proposed for the study of chemical data and have attracted the attention of the chemometrics community, both as a classification technique and because their use has been successfully extended to solve calibration problems. There is a growing number of articles focussing on the comparison of SVM with more traditional chemometrics approaches. Most applications of SVM are applied to datasets with small numbers of variables to those typically obtained in analytical chemistry. However, there is no reason why they cannot be expanded to highly multivariable datasets and therefore requiring a prior variable reduction step such as Principal Component Analysis first (Brereton & Lloyd, 2010).

2.2.1.5 KNN (k Nearest Neighbors)

As a nonparametric algorithm, KNN avoids preceding assumptions about the shape of the class boundary and can thus adapt more closely to nonlinear models when the amount of training data increases. SVM has lower variance than linear KNN, but KNN has the benefit of generating classification fits that adapt to any boundary. Even though the true class boundary is undetermined in most real-world applications, KNN has been shown to approach the theoretically optimal classification boundary as the training set increases to massive data (Bzdok, Krzywinski, & Altman, 2018).

2.2.1.6 Random forests

Random forests or random decision forests are strong supervised machine learning algorithms that can carry out both regression and classification tasks. A random forest is a group of classification, regression and other tasks that functions by generating several decision trees with the training data. A random forest is a practical nonlinear supervised machine learning method that applies large numbers of random decision trees to examine important sets of variables (Breiman, 2001). Based on the interaction with other variables, variable importance in the random forest is defined. Random forest approximates the importance of variable based on how much the prediction error increases when data for a variable is transposed while the other variables are left unaffected. The calculation for variable importance is taken away from each tree at a time as the random forest is created. Currently, the random forest is used in various applications such as the stock market, banking, medicine, retail, gene selection, and image analysis. In the random forest, the same algorithm could be used for both classification and regression problems. It can handle large datasets efficiently without variable deletion. There is no issue of overfitting when this algorithm is used either for classification or regression. The random forest can also be used for identifying important variables in the data while building the models (Amruthnath & Gupta, 2019).

2.2.2 Unsupervised machine learning

In unsupervised statistical learning, there are some inputs but no supervised output. In these methods, precise dependent variables available to train our algorithm is not known. For every observation $i = 1, \ldots, n$, we observe a vector of measurements x_i but no related response y_i. Predicting a response variable is not possible to use a linear regression model. In this setting, we are completely blind. We call this method as unsupervised because of the absence of a response variable that can supervise our analysis. We can seek to understand the relationships between the variables or between the observations. There are a couple of unsupervised learning methods, such as principal component analysis, topic models, and clustering that are explained as follows (Jolliffe, 2011).

2.2.2.1 Principal component analysis

Principal component analysis (PCA) is a famous method to derive a low-dimensional set of features from a large set of variables. PCA is considered as one of the main tools for unsupervised learning.

The principal component analysis is a method to decrease the variables when we have data on numerous variables, and we are sure there is some excess in some of these variables. This means they have some correlations that probably overlap the results of each other. Therefore, we think we can have fewer variables in a smaller number of principal components that will explain the maximum of the variance in the observed variables (Jolliffe, 2011).

It is possible to calculate a score for each subject on a given principal component by performing a principal component analysis. The formula to calculate scores on the first component extracted in a principal component analysis (2.2).

$$C_1 = b_{11} x_1 + b_{12} x_2 + \dots b_{1p} x_p, \qquad (2.2)$$

where

C_1 = the subject's score on principal component 1 (the first component extracted),

b_{1j} = loading (or weight) for observed variable j,

x_j= the subject's score on observed variable j.

The first component in a principal component analysis reflects the highest amount of total variance in the observed variables. Formally the loadings b_{1j} are determined by maximizing the variance of C_1. This means the first component is correlated to most of the observed variables. The characteristics of the second component are that this component will cover a maximal amount of variance in the dataset that was not considered by the first component. This means the second component will cover the variables which are not considered at the first component variables. Also, the second component C_2 is uncorrelated with the first component C_1. This means that the correlation between the two components is zero. The remaining components respect the same two characteristics. Every component reflects a maximal amount of variance in the observed variables that were not reflected by the previous components (James et al., 2013).

2.2.2.2 Topic models

Topic models are computer algorithms that identify hidden patterns of word occurrence using the distribution of words in a collection of documents. The result is a set of topics consisting of clusters of words that take place simultaneously in these documents according to certain patterns (Jacobi, Van Atteveldt, & Welbers, 2016). The power of topic models is that they are unsupervised. They do not require any preceding interpretations. The only requirement of a topic model is the text should be divided into documents, and the number of topics we want it to discover should be determined, and there are also models and processes for selecting the number of topics automatically (Teh, 2006).

2.2.2.3 Clustering

Clustering belongs to a very broad set of techniques to detect subgroups, or clusters, in a dataset. During the clustering of the observations of a dataset, we try to split them into distinct groups so that the observations within each group are quite like each other, while observations in different groups are quite different from each other. Indeed, this is often a consideration that must be made based on knowledge of the data being studied. In practice, we must determine the meaning of two or more observations to be similar or different (James et al., 2013). In the following table, the summary of each explained supervised and unsupervised method, including the scope of each method, main limitation, and supervised or unsupervised category of each method is explained.

Amongst all the supervised learning methods, linear regression, support vector machines, and random forest are placed in the regression category. Between the unsupervised learning methods, there is only factor analysis that is suitable for continuous hidden variables.

Table 2.2 Summary of reviewed statistical methods

Method	**Scope**	**Main limitation**	**Supervised/ Unsupervised**
KNN (k Nearest Neighbors)	Classification	When the dataset grows, the speed of the algorithm declines very fast. It does not perform well on imbalanced data	Supervised
Linear Regression	Regression	Only works for the linear relationship	Supervised
Naïve Bayes	Classification	Assumption of independent predictors.	Supervised
Logistic Regression	Classification	Logistic regression pushes the decision boundary towards the outlier	Supervised
Support Vector Machines	Classification/Regression	Does not provide importance for variables	Supervised
Random Forest	Classification/Regression	Only works with trees	Supervised
Clustering	Discrete hidden variable	Complexity, unable to estimate the correct number of clusters, unable to detect special patterns	Unsupervised
Factor Analysis	Continuous hidden variable	Complex, not easy to interpret.	Unsupervised
Topic Models	Discrete hidden variable	Does not do well for the non-linear dataset It cannot decrease the dimension of the dataset.	Unsupervised

2.3 Critical analysis

SPC is a statistical base method that controls the stability and capability of a single variable of a process. Control charts are the main tools of SPC to check for process stability. It is a well-founded

technique, which has demonstrated to be effective in manufacturing processes. However, in industry, there are many variables in which the simultaneous monitoring of two or more related quality process characteristics is required. Separately monitoring these quality characteristics can be very misleading. Product quality cannot be justified only by one product characteristic. Therefore, multivariate statistical process control (MSPC) is required to be applied in order to monitor many product quality variables at the same time.

Based on the related reviewed literature, the implementation of MSPC needs some statistical expertise and is much more complicated than univariate SPC control charts. Also, in MSPC, there is no possibility to easily determine which variable is out-of-range while the system is alarming that the process is out-of-control. As a result, operators should check all the univariate control charts to see which variable is not working in the range.

Moreover, progress in process automation and computerized equipment has changed the quality control process, from product quality monitoring to process monitoring. Automatic data acquisition and real-time process monitoring have provided massive data flow in a large and complex way. This trend has reinforced the application of machine learning and data mining techniques for process monitoring. However, there are a lot of different statistical methods that should be customized to exploit the specific data structure of every single industry. It is not possible to prescribe a unique statistical model for all industries. Indeed, many studies present the use of machine learning for production or process control in the steel industry. Nevertheless, it shows that analyzing the ferrotitanium producers' datasets and presenting a generic model for this industry has not been proposed yet. Combining different statistical methods to determine and prioritize the important predictor variables in the ferrotitanium industry is a potential solution. Such an approach would be very beneficial for the ferrotitanium producers, but they need to be guided in identifying critical alloys that have the main influence on the out of range results. Knowing the importance of every effective variable can save time and money by having a faster corrective reaction.

2.4 Conclusion

In this chapter, we studied the related literature of statistical methods for quality control, including univariate and multivariate statistical process control and design of experiment methods. Also, we

reviewed the stated researches in machine learning in production quality control and the related supervised and unsupervised learning methods. As shown in our analysis, there exist several approaches to support quality control processes. However, none is able to determine the main variables to be controlled in the ferrotitanium process and to determine the importance of each of these variables at the same time. Before presenting our proposed framework to address this issue, we first expose our research methodology in the next chapter.

CHAPTER 3 RESEARCH METHODOLOGY

This chapter presents the research approach used to carry out this study. The main research objective will be presented first and followed by the associated research questions. Next, we will describe the research strategy used as well as the case study carried out within the partner company, which served as the basis for the development and validation of the proposed model.

3.1 Research objectives and research questions

Usually, once there is an out of range result in the ferrotitanium production process, quality technicians halt the production to finalize their inquiries which are controlling the input material one by one to be matched with the issued recipe items which are a very time-consuming process. Hence, the research problem is the lack of a guide to determine which variables are responsible for out of range products in the ferrotitanium production process. Therefore, ***our main objective is to develop a general framework to determine the main variables to be controlled in the ferrotitanium process***. To meet our primary objective, we need to answer the following research questions (RQ):

RQ1: Is there any correlation between the dependent and independent variables?

RQ2: What are the most important elements of each dependent variable?

3.2 Research methodology

Considering the novelty of the phenomenon studied and the research objective to develop a general framework to determine the main variables to be controlled in the ferrotitanium process, a statistical methodology that can establish a statistical correlation between a set of input variables with a chosen outcome of the process under certain uncertainties, called uncontrolled inputs, seems appropriate to meet the objectives of this research project. By studying the collected dataset structure, determining dependent and independent variables by analyzing the potential correlation between these two groups of variables, we aim to develop a general framework that does not require any technical knowledge of internal chemical reactions affecting the ferrotitanium process. After reviewing the related literature, we found that such an approach, known in particular under the

name of DOE, is proposed by Fisher in his innovative book Design of Experiments and is illustrated in Figure 3.1 (Fisher, 1949).

The DOE structure is presented as a black box, like a device, system or object which can be viewed uniquely in terms of input, output, and transferred (correlation) characteristics without any knowledge of the internal mechanisms. Therefore, any established model using DOE is not a formal mathematical model but instead represents a proper statistical or correlation model that does not try to replicate the intrinsic physical properties of the system under study. Also, there is no need to generate some physical conclusions from the gained model as statistical correlation can be recognized between an input and the output which are not physically related (Astakhov, 2012).

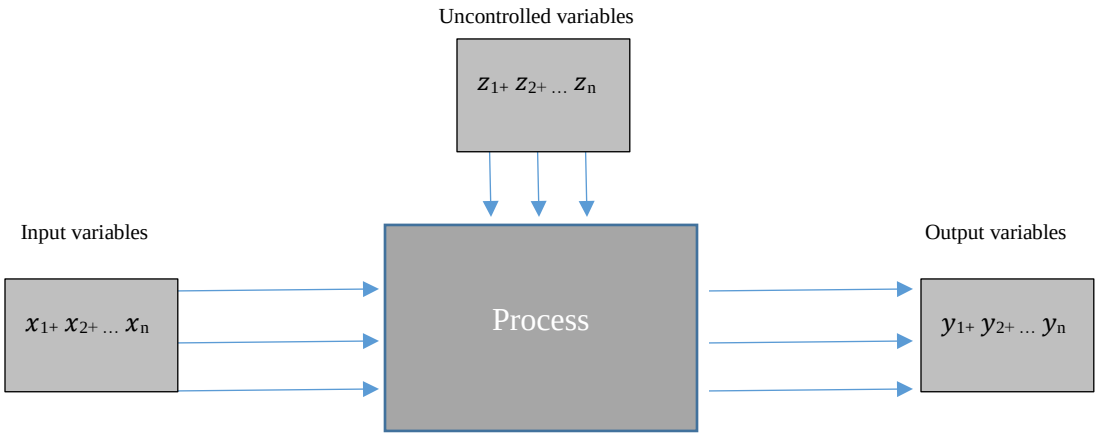

Figure 3.1 DOE visualized definition

As we want to find out their correlations and the importance of each on the response variables, including some hidden and uncontrolled variables, the DOE is an appropriate methodology to study the collected dataset and to generate the general model of our research objective. The practical steps and guidelines for planning and conducting DOE are listed below in Figure 3.2.

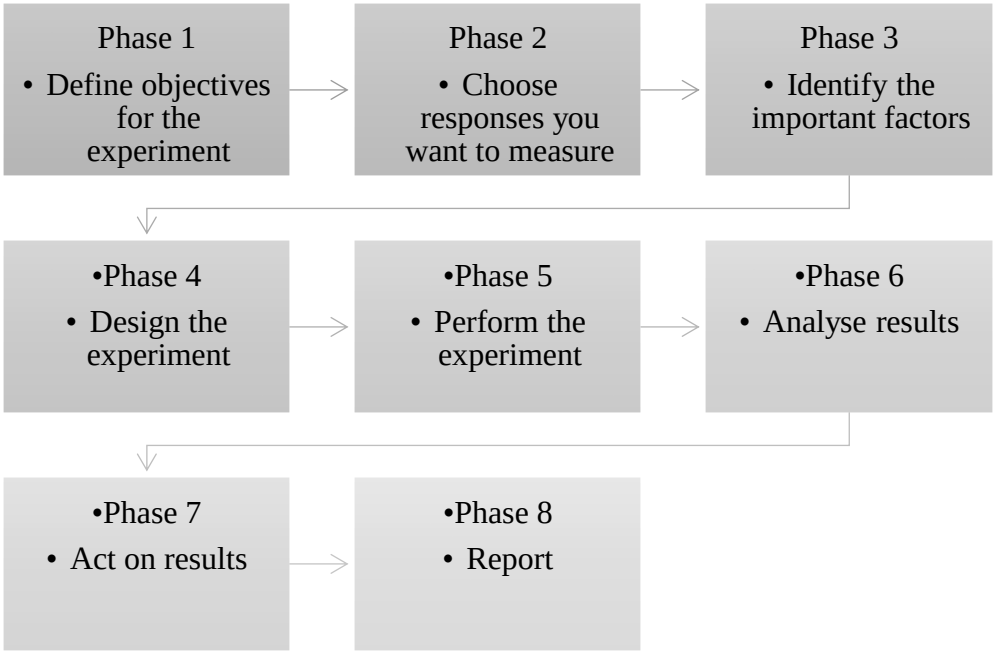

Figure 3.2 Steps involved in the design of experiments (Durakovic 2017)

We plan our research phases based on the DOE approach, as described in the following subsections.

3.2.1 Phase 1: Define the objectives

In our specific case, this first step is essentially based on the review of literature on the ferrotitanium industry. This made it possible to note the state of progress of the field studied and to determine our potential for contributions, which is the proposal of a generic model to determine the main variables to be controlled in the ferrotitanium process.

3.2.2 Phase 2: Choose response variables

In our case, after talking to quality experts of the studied company, we find out that over 15 ferrotitanium elements, five main elements are controlled using SPC control charts in the studied company. Therefore, we consider these five elements as our response variable (dependent) and the others as the predictor variables (independent).

The following table presents the segregation of dependent and independent variables, which we call them as response and predictor variables interchangeably over this research.

Table 3.1 Dependent and independent variables

Variables	Response	Predictor
al	√	
mo		√
si		√
sn		√
zr		√
mn		√
cr		√
v	√	
fe		√
c	√	
ni		√
cu		√
n		√
o	√	
ti	√	

3.2.3 Phase 3: Identify important factors

After reviewing all sale contracts, we found among all elements of ferrotitanium products that there are five main elements that all the steel producers as main clients require to meet the determined control limits. In our case, these five main elements are the response variables that, by studying the predictor variables, we try to find which predictors are the most important ones.

3.2.4 Phase 4: Propose the model

DOE is used for conducting systematic studies of a system, process or product in which input variables are manipulated to explore its effects on the measured response variable. We intentionally change one or more process variables in order to see the impact the changes have on one or more response variables (Durakovic, 2017). However, in our studied case, we are working with observed data, and there is no need and capability to change the variables or factors deliberately. So, we combine the two phases of *design the experiment* and *perform the experiment* together and replace them with a new one called *Propose the model*, as illustrated in Figure 3.3. This new activity is the main contribution of this work. As such, we recognize that a single experiment design would be insufficient in our case. Consequently, we suggest combining supervised and unsupervised learning methods. The proposed approach will be discussed in Chapter 4.

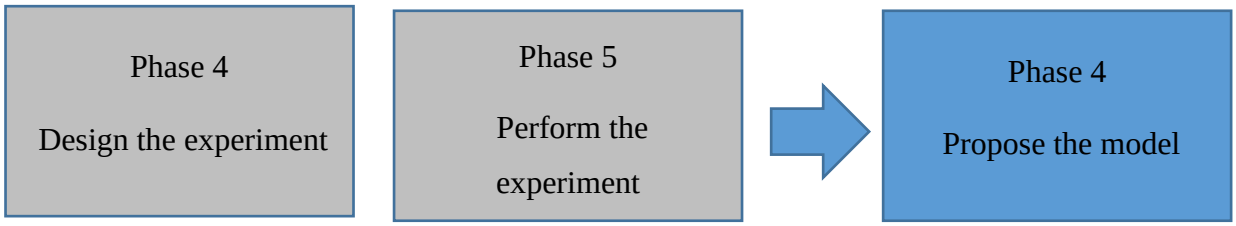

Figure 3.3 Combination of Phases 4 and 5

3.2.5 Phase 5: Validate and analyze the results

After proposing a general framework to determine the main variables to be controlled in the ferrotitanium process, we will validate it by running the included models using the R program. In this step, we will compare the result of the multiple linear regression method with a random forest to find out whether both results are the same or not and to control if the data fits these models.

3.2.6 Phase 6: Act

There are different types of products in the ferrotitanium industry, such as Low Aluminum, Low Carbon, Low Vanadium, Low Oxygen, etc. After applying multiple linear regression, random forest and principal component analysis across the main data set, to measure whether the results

are matched for all products, a special product called Low-Carbon is selected to repeat the three statistical methods and to make a comparison between results to see if the results are matched and fitting each other between dataset and the data subset.

3.2.7 Phase 7: Report

If the results of the data subset were matched with the general dataset results, we could apply the recommended model to other ferrotitanium producers. Furthermore, if the results were not equalled, we should continue our study to see what the main effective elements of each unmatched response variable are. The final research steps inspired by the DOE methodology are presented in Figure 3.4.

3.3 Case study

This research is applied in nature and requires the use of real data. As such, we will use the dataset of a ferrotitanium company of North America to investigate the quality data of this industry. Data can be obtained easily from the industrial partner as the main researcher of this study is an employee of the company. This project was requested by the management of the studied company from the researcher but was not funded by them. Over the last twenty years, this has become the leading manufacturer of ferrotitanium in Canada. The company is a globally recognized leader as a producer of the highest quality ferrotitanium (FeTi 70).

This company specializes in the crushing of FeTi of different grades and sizes ranging from standard, low carbon, low aluminum, low vanadium grades and sizes from 10 to 50 mm, 10 to 30 mm, 6 to 12 mm, until 0 to 2 mm FeTi powder. The products are intended primarily for the manufacturer of steel for use in construction, aviation, automotive, oil &gas and medical market.

Currently, in this company, they are using statistical process control charts to control the client's required chemical components. By using some standard recipes, they produce high-quality products, except when they see instabilities in chemical results that are not predictable and cause the production of low-quality products, even though the SPC method is in place. These off-grade products generate lots of costs for this company because the only way to recover these off-grade products is to remelt them, which is very expensive and time-consuming. For this company,

developing a general framework to determine the main variables to be controlled, leads to stable quality, and dramatically improves the production cost of this company and this industry.

3.4 Conclusion

This research is accomplished based on the observed data approach, in the sense that we do not touch the design part of DOE and we only focus on its modelling part. A new statistical framework by applying the observed data will be developed and presented in the next chapter since developing a general framework capable of determining the main variables to be controlled in the ferrotitanium process for a given observed dataset has not been covered by the scientific literature. We expect that combining multiple techniques would allow us to determine whether there is any relation between dependent and independent variables in the ferrotitanium dataset and, if so, what are the most important elements of each dependent variable.

CHAPTER 4 MODEL PROPOSAL

This chapter presents the development of the proposed model with a description of each of its components. It includes a description of the initial activities, collecting the historical dataset and preparing data in sections 4.2 and 4.3, respectively. Then, sections 4.4 and 4.5 discussed the use of multiple linear regression and random forest as supervised learning methods. In Section 4.6, the result comparison of the multiple linear regression and random forest methods is explained, followed by the principal component analysis activity in Section 4.7. As the last step of the proposed model, consistency check, and reapply the methods over a data subset is explained in Section 4.8.

4.1 Proposed model

In every company, massive data is generating every day. To exploit this data, we first need to collect and sort out the appropriate data after talking with the experts in the industry.

Once data is ready to be exploited, the right analysis techniques must be selected. As underlined in the literature review, different statistical methods can be used, and none can achieve all our objectives. Consequently, our strategy is first to study the potential correlation between the response and predictor variables and study the importance of each predictor variables using the supervised learning methods, as presented in Figure 4.1. Then, using unsupervised learning methods, we try to find the hidden variables that may influence the results.

More precisely, this general model for ferrotitanium process control uses multiple linear regression and random forest as supervised learning methods to determine the correlation and importance of each predictor variable. Multiple linear regression is applied to analyze the potential dependency between dependent and independent variables. Using multiple linear regression, we control the linearity of the correlation between the response and predictor variables. Multiple linear regression is calculated by giving each predictor a separate slope coefficient in a single model as Form 4.1:

$$y = \beta_{0+} \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_p x_p + \varepsilon, \qquad (4.1)$$

where xj represents the jth predictor, and βj quantifies the association between that variable and the response. We interpret βj as the average effect on y of a one-unit increase in xj, holding all other predictors fixed (Neter, Kutner, Nachtsheim, & Wasserman, 1996).

The result of multiple linear regression is verified, and the first two main chemical components which have the most effect on dependent variables are determined and compared with the multiple linear regression results to answer the second research question, using a random forest test,

Thereafter, principal component analysis is used to study hidden and unknown variables that affect the ferrotitanium results such as temperature, employees' experience, etc. We apply PCA over the residuals of random forest results to find out the hidden elements affecting the correlations between responses.

PCA is one of the main methods of multivariate statistical process control approach. When faced with a large set of correlated variables, principal components allow summarizing this set with a smaller number of representative variables that collectively explain most of the variability in the original set. PCA refers to the process by which principal components are computed, and the subsequent use of these components in understanding the data. PCA is an unsupervised approach since it involves only a set of features $x_1, x_2 \ldots x_p$, and no associated response y. Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization -visualization of the observations or visualization of the variables (Jolliffe, 2011).

The following sections describe each step of the proposed framework.

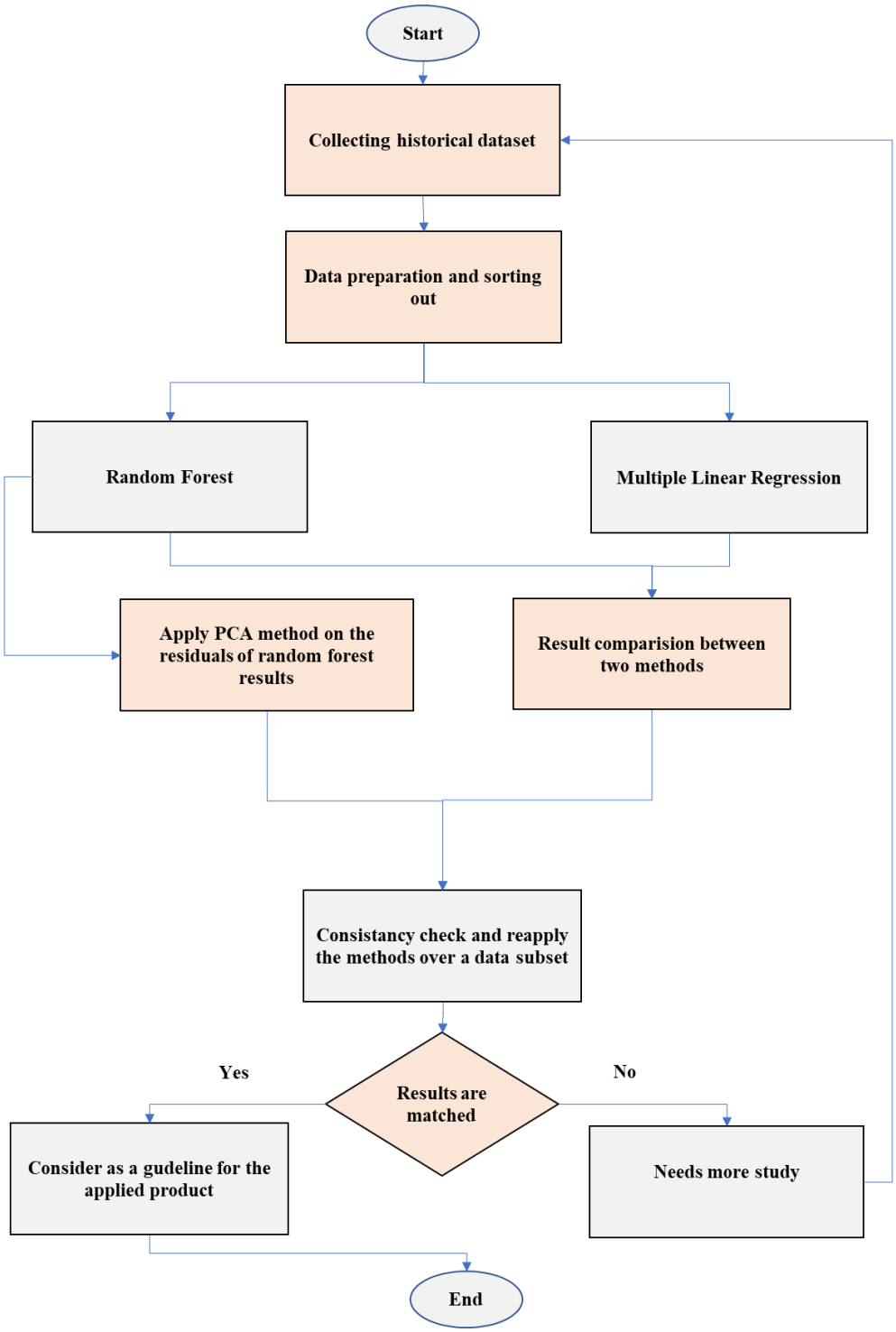

Figure 4.1 Proposed general model for ferrotitanium process control

4.2 Collecting historical dataset

One of the most important parts of every applied research is data collection. Using inaccurate data causes a great impact on the result of a study and finally leads to invalid results, which must be avoided. There are two main data classifications as primary and secondary data. Primary data is the data that will be collected by researchers at the source or the data originally collected by focus groups, individuals, and a panel of respondents specifically set up by the researcher whose opinions may be required on specific issues from time to time. Primary data is more accurate as compared with secondary data. It is less expensive and current. Required time and energy are the main weaknesses of primary data, as well as its difficulty in collecting.

Secondary data is generated by gathering and use of existing data for which they were originally collected, for example, computerized database, company records or archives, government publications, industry analysis offered by the media, and so on. Secondary data is less expensive and easier to collect than primary data. It is cost-effective, but the error rate is high.

There is another general classification of data as quantitative data and qualitative. Quantitative data is data that is mainly numbers. It refers to the information that is collected as or can be translated into numbers, which can then be displayed and analyzed mathematically (Osang, Udoimuk, Etta, Ushie, & Offiong, 2013). Therefore, the collected data we are using during this research is placed in the secondary and quantitative data because we are collecting them from the quality department of the partner company.

4.3 Data preparation and sorting out

Before starting data analysis, the data must be organized in an appropriate form. The data preparation is the process of processing and organizing data before analysis. Data preparation is the processing of raw data, which is often formless and messy, to change it to a more structured and useful form that is ready for analysis. The whole preparation process consists of a series of major activities, including profiling, cleansing, integration, and transformation of data (Abdallah, Du, & Webb, 2017). After investigating all sales contracts of 2018, it was considered that five main alloys are requested by clients such as titanium, aluminum, vanadium, carbon, and oxygen, and quality officers are controlling them at every sample result using the SPC control charts to

meet the customer requirements. The following figure shows the related control chart, which is used in the studied company to control vanadium as one of these response variables.

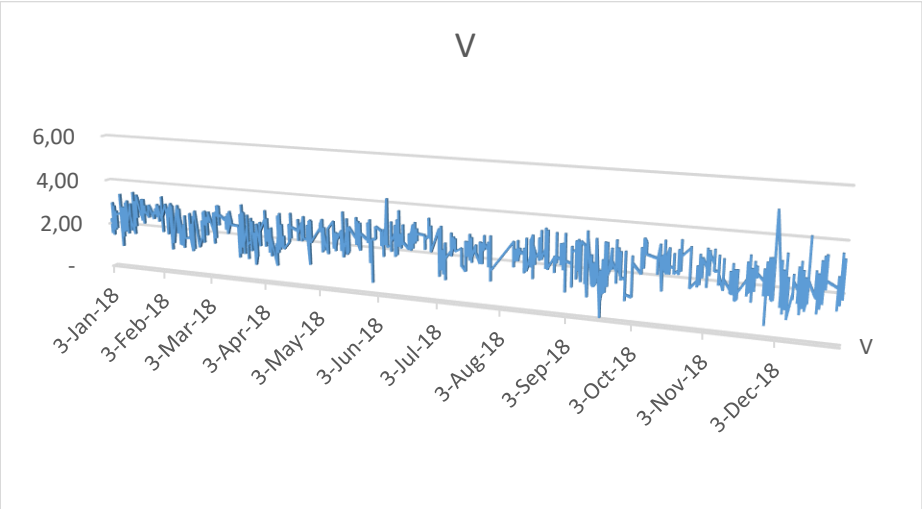

Figure 4.2 Control chart for vanadium

As depicted in Figure 4.2, the requested lower control limit is 1.5, and the upper control limit is 3.5 for vanadium. The X-axis shows the time of taken samples, which is separated in the monthly base, and Y-axis shows the percentage of vanadium in every taken sample. We consider these five main required alloys as our dependent variables to run the data analysis and to find out which alloys have more effects on these dependent variables. Figure 4.3 depicts the structure of the collected dataset.

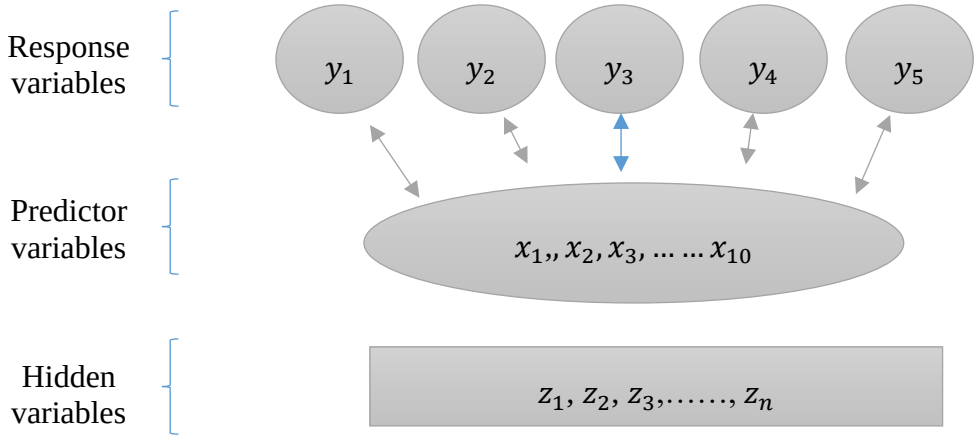

Figure 4.3 Dataset structure

As it is explained in Figure 4.3, there are five response variables in the collected dataset which are shown as $y_1, y_{2,} \dots, y_5$, and ten predictor variables which are shown as $x_1, x_{2,} \dots, x_{10}$. We are looking to find which of these x variables influence each of y variables, and to find out the importance of each x variable on every y. Also, some hidden variables have some potential effects on the results that are not controlled by the studied company, but we want to find out their potential influences on the results. We show these hidden variables as $z_1, z_{2,} \dots, z_n$ in Figure 4.3. Figure 4.4 shows the status of real variables we are analyzing using the proposed model.

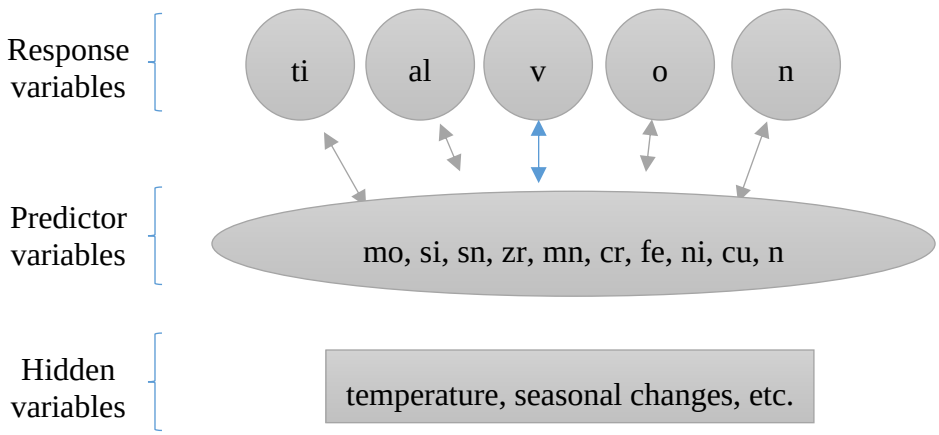

Figure 4.4 Response, predictor, and potential hidden variables

As a result, we use supervised learning methods to find potential correlations between response and predictor variables in the proposed model because the response variables and their control ranges are determined in our case. In supervised learning, we typically have access to a set of p features $x_1, x_2, \dots x_p$, measured on n observations and a response y also measured on those same n observations. The goal is then to predict y using $x_1, x_2, \dots x_p$ (James et al., 2013).

We use multiple linear regression as a fit model to analyze the linearity of the data correlation, and we use the random forest as a fit model to analyze non-linearity and to prioritizing the predictor variables in the proposed model.

4.4 Multiple linear regression over the dataset

Simple linear regression is a useful approach for predicting a response based on a single predictor variable. However, in practice, we often have more than one predictor.

There is an option to run some simple separate linear regressions for every predictor. However, there is a better option to develop simple linear regression by calculating the multiple predictors simultaneously. It will be calculated by giving each predictor a separate slope coefficient in a single model as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon, \quad (4.2)$$

where xj represents the jth predictor, and βj quantifies the association between that variable and the response. We interpret βj as the average effect on y of a one-unit increase in xj , holding all other predictors fixed (Neter et al., 1996).

Multiple linear regression is applied to analyze the potential linear dependency between dependent and independent variables because there is more than one independent variable over the collected dataset. In this way, we answer the first research question.

4.5 Random forest

A random forest is a group of classification, regression and other tasks that functions by generating plenty of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It becomes an important part of modern data scientists to purify the predictive model. There are very few assumptions attached to this model, so data preparation is pretty convenient and time-saving (Breiman, 2001).

The result of multiple linear regression is verified using a random forest test, and the first two main chemical components which have the most effect on dependent variables are determined and compared with the multiple linear regression results. In this way, we answer the second research question.

4.6 Result comparison between two methods

In this step, we compare the result of multiple linear regression methods with random forest to find out whether both results are the same or not and if the data fits these models. The model fit is tested with the dataset using an R-squared number for linear regression and P-value for the Random forest method. R-squared shows the portion of the variation in a dependent variable that is accounted for or predicted by independent variables. The P-value presents how confident we can be that each

variable has some correlation with the dependent variable. A predictor that has a low p-value (< 0.05) is expected to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Contrariwise, a larger (insignificant) p-value proposes that changes in the predictor are not related to changes in the response. In the case of T-value, a larger *t-value* indicates that it is less likely that the coefficient is not equal to zero purely by chance. Therefore, higher the t-value, the better.

4.7 Principal component analysis

In the unsupervised learning methods, we do not have precise dependent variables to train our algorithm. For every observation $i = 1, \ldots, n,$ we observe a vector of measurements xi but no related response y_i . We call this method as unsupervised because of the absence of a response variable that can supervise our analysis. We can seek to understand the relationships between the variables or between the observations (Jolliffe, 2011). Therefore, using unsupervised learning methods, we analyze these potential variables in the proposed model. The unsupervised method we are using in the proposed model is principal component analysis because we are projecting to hidden continuous variables, and the other unsupervised methods such as clustering and topic models are fitting for discrete hidden variables.

In the proposed model, using principal component analysis, we try to study the other hidden and unknown variables such as temperature, employees' experience, etc. that affect the ferrotitanium results. In this way, we can decrease the dataset dimension to one, and it gives us the possibility of studying and controlling each of these hidden variables using SPC control charts.

As we can see in the following figure, to study the relation between two response variables as y_1 and y_2, if they are in relation to the same predictor variable x_1, we cannot precisely say that this correlation is because of predictor $x_1$or the two response variables y_1 and y_2.

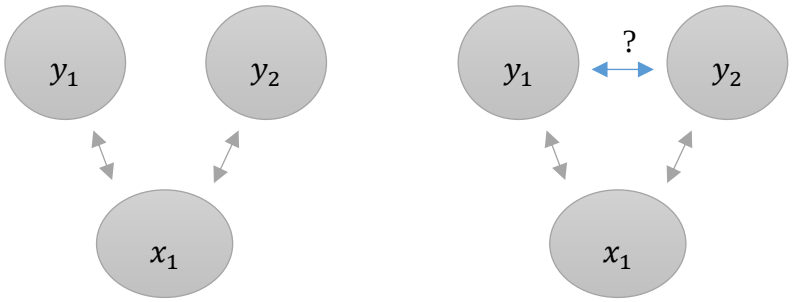

Figure 4.5 Potential correlation between two response variables

For this reason, we deduct the direct effect of each of these predictors in response variables which gives us the out of range variables, and we run PCA over these out of range variables, to have the possibility of controlling the errors. If we could control the errors of the system, the total system will be under control.

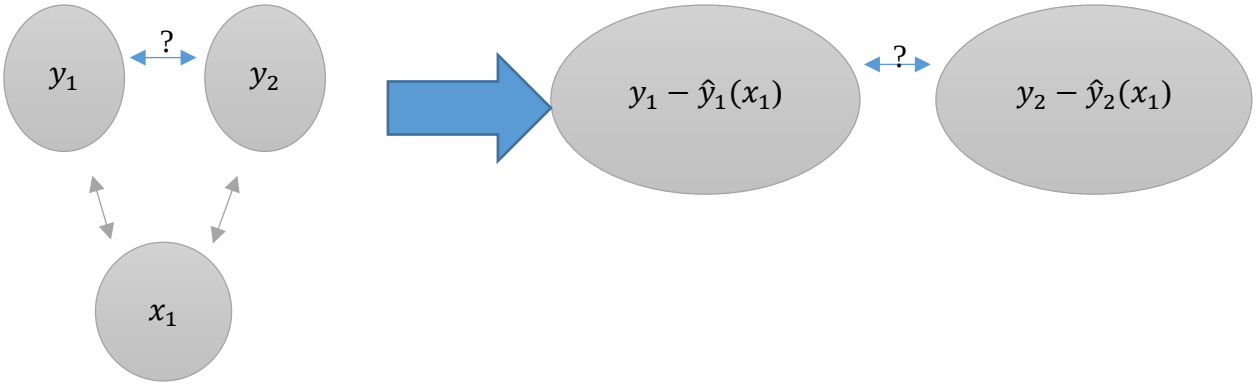

Figure 4.6 Deducting the direct effect of each predictor variable from response variables

As we see in the right panel of Figure 4.6, we remove the influence of predictor variable x_1 by applying PCA over the residuals of random forest results to find out the hidden elements influencing results.

4.8 Consistency check and reapply the methods over a data subset

Because there are different types of products in the ferrotitanium industry, such as Low Aluminum, Low Carbon, Low Vanadium, Low Oxygen, etc., a specific product will be selected to replicate the three statistical methods. If the results of the data subset match with the general dataset results,

we can use the proposed model as a guide for the selected product. Furthermore, if the results were not equalled, we would need to continue our study using the experts' feedback.

4.9 Conclusion

Our proposed framework is composed of multiple activities and techniques, including data collection and preparation, applying supervised and unsupervised learning methods, comparing results between linear and non-linear methods, and finally, performing consistency checks on other products. In the next chapter, we run the proposed model over a collected dataset in order to validate its coherence.

CHAPTER 5 MODEL VALIDATION AND NUMERICAL EXAMPLE

The preceding chapters made it possible to define the concepts and the functioning of our model as a new general framework for ferrotitanium process control. This chapter is dedicated to the practical demonstration of the proposed model by applying the model across a real dataset received by the researcher from the partner company.

5.1 Collecting historical dataset

The dataset, which was generated during 2018 to control 15 different alloys of daily production, is the most reliable data as the company has fully expanded its laboratory machinery from the beginning of 2018, and the quality control team has analyzed every melt sample after daily calibration. Overall, 2018 was the most consistent year to make data analysis.

5.2 Data preparation and sorting out

The collected dataset contains 15 different alloys and four more columns, including Date, SID1, Assignments, and Notes, which are explained in Table 5.1.

Table 5.1 Description of data sets' columns

Column Name	Description	Data type
Date	Date of sample production	Numeric
SID1	Sample number	Numeric
al	Aluminum	Numeric
mo	Molybdenum	Numeric
si	Silicon	Numeric
sn	Tin	Numeric
zr	Zirconium	Numeric
mn	Manganese	Numeric
cr	Chromium	Numeric
v	Vanadium	Numeric
fe	Iron	Numeric
c	Carbon	Numeric
ni	Nickle	Numeric
cu	Copper	Numeric
n	Nitrogen	Numeric
o	Oxygen	Numeric
ti	Titanium	Numeric
Assignment	Product name subgroup	Character
Notes	Product name	Character

The primary dataset format was changed to CSV (comma-separated values) file, which contained 2321 rows with 19 columns. Data exploration clarified that some observations were out of the range, and based on the discussion with laboratory technicians, we concluded to treat them as outliers. Therefore, observations with the "KIB" tag had calibration issues of ARL machines, so they are dismissed too. The final and cleaned dataset included 2084 rows and 17 columns, regardless of Date and SID1 columns. From these 17 columns, “type” and “product” are discrete, and others are continuous. After depicting the dataset variables in the scatter plot in Figure 5.1, it shows that there is a linear regression between some of the variables of the general dataset.

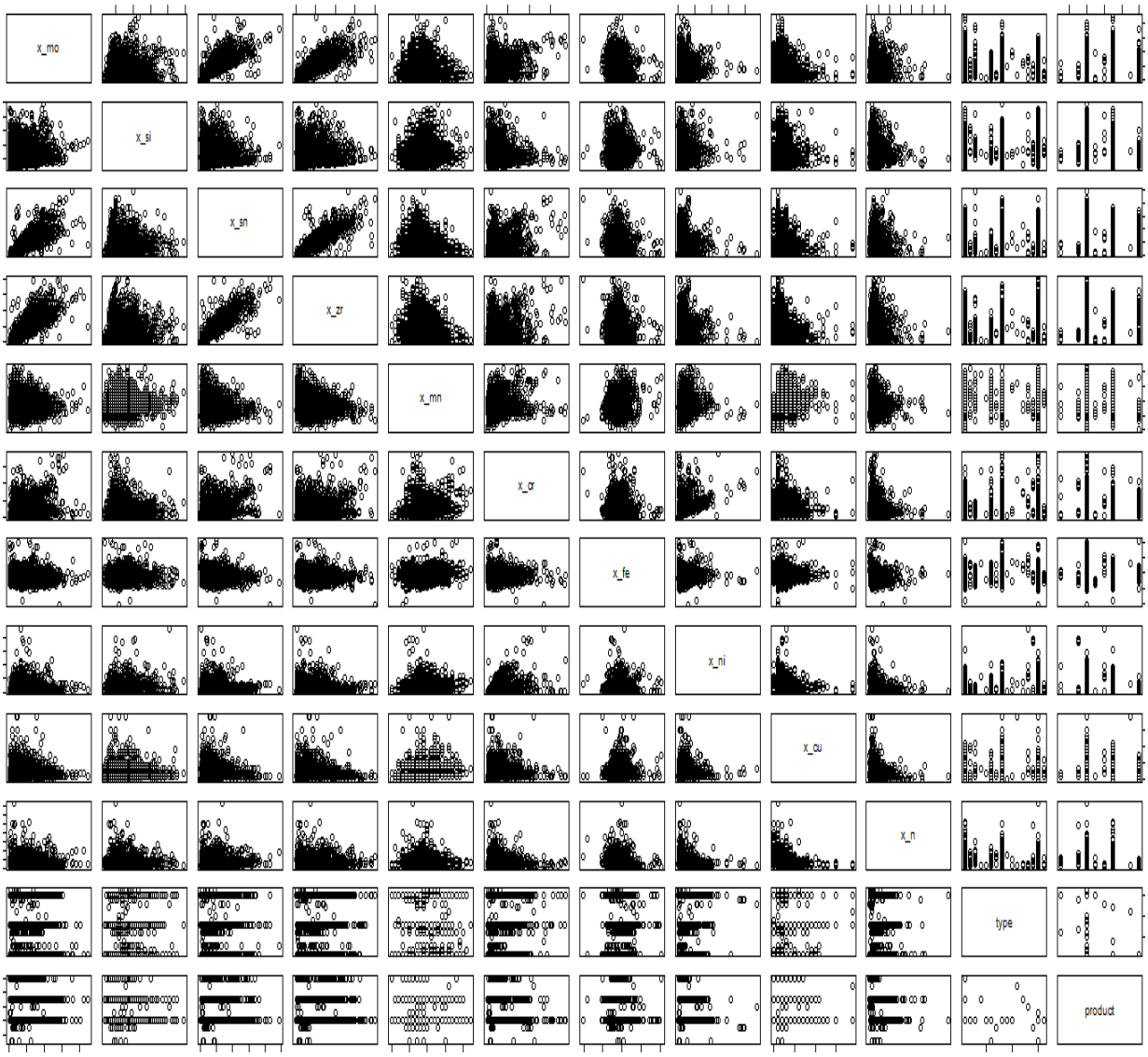

Figure 5.1 Scatter plot to show the relationship between predictors

As it is depicted in Figure 5.1, we can see a linear track between some variables like molybdenum (mo) and tin (sn), or tin (sn) and zirconium (zr). In Figure 5.2, and to clarify the observed correlation in Figure 5.1, the correlation plot is presented in Figure 5.2.

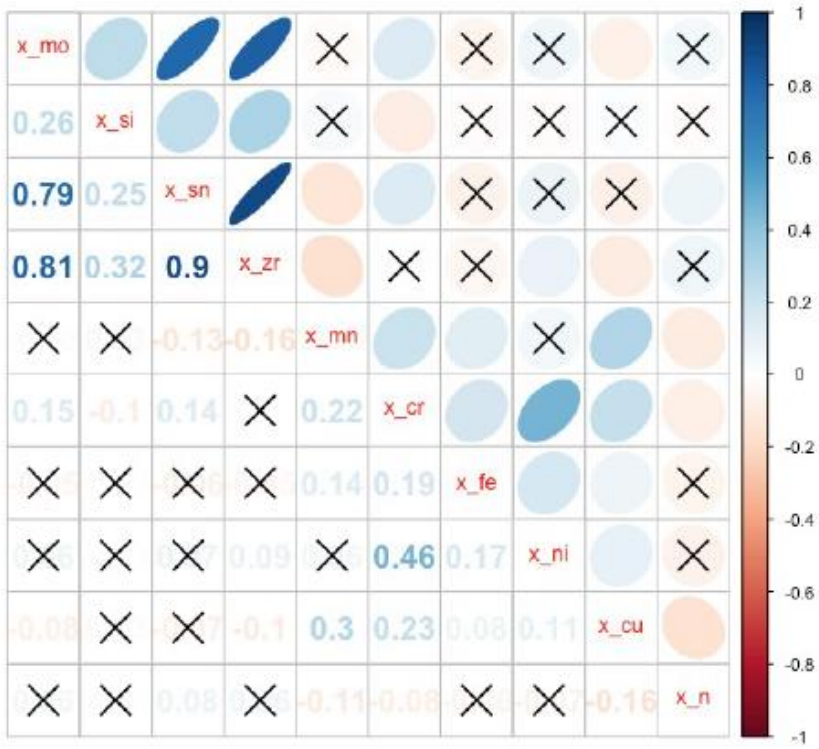

Figure 5.2 Correlation plot of predictors

As it is presented in Figure 5.2, there are some significant and considerable correlations between some predictors. The digits show the correlation amount between the two connected variables. The cross sign shows that there is no correlation between the two related variables. To visual identification of the correlations, as colours get darker than the others, means the correlation is higher in the associated variables, and lighter colours show a low correlation between linked variables. For example, there is a high correlation of 0.90 between tin (sn) and zirconium (zr) in the ferrotitanium production process. Also, we can observe the correlation between zirconium (zr) and molybdenum(mo) with 0.81 and finally the correlation between tin (sn) and molybdenum (mo) which is 0.79. In the rest of this research, using supervised and unsupervised machine learning methods, these correlations will be analyzed in more detail.

Statistical learning methods are divided into two categories supervised or unsupervised. For supervised learning, every observation of the predictor measurement(s) $x_i, i = 1, 2, \dots n$ there is an associated answer as y_i sometimes called the response. In supervised learning, we fit a model to predict responses. A simple statistical model may help to understand the relationship between predictors and response. There are some statistical measures we apply for the statistical analysis as below, to interpret our results.

The size of the coefficient for each independent variable gives the size of the effect that variable has on the dependent variable in simple or multiple linear regression, and positive or negative sign on the coefficient clarify the direction of the effect. In a single independent variable regression, the coefficient tells us how much the dependent variable is expected to increase or decrease (depends on the coefficient's sign positive or negative) when that independent variable increases by one unit. In multiple independent variable regression, the coefficient tells us how much the dependent variable is expected to increase when that independent variable increases by one, while all the other independent variables are constant.

The R-squared of the regression is the portion of the variation in a dependent variable that is accounted for or predicted by the independent variables. The R-squared is generally of secondary importance unless our main concern is using the regression equation to make accurate predictions.

The p-value presents how confident we can be that each variable has some correlation with the dependent variable, which is the important thing. A predictor that has a low p-value (< 0.05) is expected to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. Contrariwise, a larger (insignificant) p-value proposes that changes in the predictor are not related to changes in the response.

In the case of t-value, a larger t-*value* indicates that it is less likely that the coefficient is not equal to zero purely by chance. Therefore, the higher t-value is better.

In Section 5.3, supervised learning models are built to predict the relation between the response and predictors using multiple regression as a linear method, random forests and as a non-linear method. Afterward, the principal component analysis will be used as an unsupervised projection tool to improve multivariate statistical process control (James et al., 2013).

5.3 Multiple linear regression over the dataset

While dealing with a single predictor variable, simple linear regression is a suitable method for predicting response. In real life, there are often several predictors, which may affect each other. Multiple linear regression method determines a separate slope coefficient for each predictor to quantify the relationship between the response and predictors. (James et al., 2013)

We use the p-value to determine the statistical significance of the relationship, with a threshold level of 0.05. The obtained fitted model takes the following form for titanium chemical components.

$$\mathrm{ti_i} = \beta_0 + \beta_1 \mathrm{mo_i} + \beta_2 \mathrm{si_i} + \beta_3 \mathrm{sn_i} + \beta_4 \mathrm{zr_i} + \beta_5 \mathrm{mn_i} + \beta_6 \mathrm{cr_i} + \beta_7 \mathrm{fe_i} + \beta_8 \mathrm{ni_i} + \beta_9 \mathrm{cu_i} + \beta_{10} \mathrm{n_i} + \varepsilon_i, \quad (5.1)$$

where $i = 1, \ldots, 2084$ refers to the data samples.

The results of the coefficients from the fitted model on the training set are shown below.

Table 5.2 Multiple linear regression results of the equation (5.1)

	Estimate	Std-error	T-Value	P-value	Signif.Code
mo	0.36	0.12	2.83	0.00464	**
si	-2.09	0.34	-6.10	1.23e-09	***
sn	-0.87	0.37	-2.33	0.01981	*
zr	-0.67	0.16	-4.16	3.27e-05	***
mn	-11.72	0.83	-13.98	< 2e-16	***
cr	-0.51	0.12	-4.20	2.76e-05	***
fe	-0.59	0.01	-42.87	< 2e-16	***
ni	-0.98	0.07	-14.46	< 2e-16	***
cu	-5.92	1.30	-4.55	5.68e-06	***
n	-2.95	0.22	-13.23	< 2e-16	***

The results show that all predictors are statistically significant. This indicates that all predictors affect the response. The reported R-squared by the model is around 62%. It means this model fits the data well. Therefore, the trained model is used to predict the response variable on the dataset for titanium as one of our dependent variables. The obtained fitted model for aluminum regression takes the following format.

$$al_i = \beta_0 + \beta_1 mo_i + \beta_2 si_i + \beta_3 sn_i + \beta_4 zr_i + \beta_5 mn_i + \beta_6 cr_i + \beta_7 fe_i + \beta_8 ni_i + \beta_9 cu_i + \beta_{10} n_i + \varepsilon_i, \quad (5.2)$$

The results of the coefficient from the fitted model on the training set are shown below.

Table 5.3 Multiple linear regression results of the equation (5.2) [1]

	Estimate	Std-error	T-Value	P-value	Signif.Code
mo	-0.54	0.07	-7.07	2.02e-12	***
si	-0.90	0.21	-4.30	1.78e-05	***
sn	0.77	0.23	3.37	0.000745	***
zr	-0.28	0.09	-2.88	0.003947	**
mn	4.25	0.51	8.26	2.46e-16	***
cr	-0.30	0.07	-4.06	5.04e-05	***
fe	-0.13	0.01	-16.19	< 2e-16	***
ni	0.12	0.04	2.91	0.003561	**
cu	5.88	0.79	7.36	2.51e-13	***
n	-0.53	0.13	-3.89	0.000102	***

Results of Table 5.3 show all predictors are statistically significant and mean there is a linear regression between aluminum and other chemicals. The reported R-squared by the model is low 23%. It means the model fits the data poorly in terms of prediction power. Therefore, the trained model should be considered not as a strong model to predict the response. To obtain a fitted model for vanadium regression, it takes the following form.

$$v_i = \beta_0 + \beta_1 mo_i + \beta_2 si_i + \beta_3 sn_i + \beta_4 zr_i + \beta_5 mn_i + \beta_6 cr_i + \beta_7 fe_i + \beta_8 ni_i + \beta_9 cu_i + \beta_{10} n_i + \varepsilon_i, \quad (5.3)$$

which gives the following results as Table 5.4.

[1] Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Multiple R-squared: 0.2312, Adjusted R-squared: 0.2275

Table 5.4 Multiple linear regression results of the equation (5.3)

	Estimate	Std-error	T-Value	P-value	Signif.Code
mo	-0.53	0.04	-11.95	< 2e-16	***
si	-0.02	0.12	-0.16	0.868205	
sn	-0.43	0.13	-3.20	0.001372	**
zr	-0.20	0.05	-3.53	0.000412	***
mn	5.04	0.30	16.74	< 2e-16	***
cr	-0.26	0.04	-6.10	1.26e-09	***
fe	-0.14	0.01	-29.36	< 2e-16	***
ni	0.06	0.02	2.47	0.013476	*
cu	-0.78	0.46	-1.68	0.091619	.
n	0.50	0.07	6.33	2.98e-10	***

These results indicate there are two chemical components as *si* and *cu*, which are not significant. This means the model can be simplified, and *si* and *cu* can be removed from the model. We continue to regenerate the regression model by removing these two alloys, respectively. The removal part starts by taking away the *si* as it has a larger p-value compared to *cu* as the following form.

$$v_i = \beta_0 + \beta_1 mo_i + \beta_3 sn_i + \beta_4 zr_i + \beta_5 mn_i + \beta_6 cr_i + \beta_7 fe_i + \beta_8 ni_i + \beta_9 cu_i + \beta_{10} n_i + \varepsilon_i, \quad (5.4)$$

The results are shown in Table 5.5.

Table 5.5 Multiple linear regression results of the equation (5.4)

	Estimate	Std-error	T-Value	P-value	Signif.Code
mo	-0.54	0.04	-11.99	< 2e-16	***
sn	-0.43	0.13	-3.20	0.001372	**
zr	-0.20	0.05	-3.54	0.000405	***
mn	5.03	0.29	16.90	< 2e-16	***
cr	-0.26	0.04	-6.17	8.02e-10	***
fe	-0.14	0.01	-29.37	< 2e-16	***
ni	0.05	0.02	2.47	0.013388	*
cu	-0.78	0.46	-1.68	0.092079	.
n	0.50	0.07	6.33	2.96e-10	***

Results show that *cu* still has a p-value > 0.05, and we can remove this predictor for the vanadium prediction. So, we continue to redevelop the regression model by removing *cu* in addition to *si* as the following form.

$$v_i = \beta_0 + \beta_1 mo_i + \beta_3 sn_i + \beta_4 zr_i + \beta_5 mn_i + \beta_6 cr_i + \beta_7 fe_i + \beta_8 ni_i + \beta_{10} n_i + \varepsilon_i, \quad (5.5)$$

which gives the following fitting results, as illustrated in Table 5.6.

Table 5.6 Multiple linear regression results of the equation (5.5)

	Estimate	**Std-error**	**T-Value**	**P-value**	**Signif.Code**
mo	-0.53	0.04	-11.91	< 2e-16	***
sn	-0.44	0.13	-3.27	0.001073	**
zr	-0.20	0.05	-3.48	0.000501	***
mn	4.95	0.29	16.84	< 2e-16	***
cr	-0.27	0.04	-6.30	3.61e-10	***
fe	-0.14	0.00	-29.52	< 2e-16	***
ni	0.06	0.02	2.49	0.012749	*
n	0.52	0.07	6.64	3.93e-11	***

By removing *si* and *cu* from the trained model, all predictors are statistically significant. This indicates that all predictors affect the response except *si* and *cu*. The reported R-squared by the model is around 55%, so the model fits the data adequately. Therefore, this trained model can be used to predict the response variable on the dataset for vanadium content. To obtain a fitted model for carbon content regression, it takes the following form, and the statistical results will be explained in table 5.7

$$c_i = \beta_0 + \beta_1 mo_i + \beta_2 si_i + \beta_3 sn_i + \beta_4 zr_i + \beta_5 mn_i + \beta_6 cr_i + \beta_7 fe_i + \beta_8 ni_i + \beta_9 cu_i + \beta_{10} n_i + \varepsilon_i, \quad (5.6)$$

It gives the following fitting results.

Table 5.7 Multiple linear regression results of the equation (5.6)

	Estimate	Std-error	T-Value	P-value	Signif.Code
mo	-0.02	0.01	-2.541	0.011138	*
si	0.22	0.02	10.568	< 2e-16	***
sn	0.04	0.02	1.708	0.087747	.
zr	-0.01	0.01	-1.229	0.219231	
mn	0.07	0.05	1.450	0.147318	
cr	0.02	0.01	3.879	0.000108	***
fe	0.01	0.00	5.298	1.30e-07	***
ni	0.0106	0.00427	2.495	0.012670	*
cu	0.4814	0.08170	5.892	4.44e-09	***
n	0.0808	0.01399	5.781	8.58e-09	***

Following a similar removal procedure, we remove *sn*, *zr*, and *mn*. The final model is as follows

$$c_i = \beta_0 + \beta_1 mo_i + \beta_2 si_i + \beta_6 cr_i + \beta_7 fe_i + \beta_8 ni_i + \beta_9 cu_i + \beta_{10} n_i + \varepsilon_i, \quad (5.7)$$

which gives the following fitting results.

Table 5.8 Multiple linear regression results of the equation (5.7)

	Estimate	Std-error	T-Value	P-value	Signif.Code
mo	-0.01	0.01	-4.152	3.44e-05	***
si	0.23	0.02	10.867	< 2e-16	***
cr	0.03	0.01	4.955	7.82e-07	***
fe	0.00	0.00	5.529	3.63e-08	***
ni	0.01	0.00	2.304	0.0213	*
cu	0.50	0.08	6.255	4.82e-10	***
n	0.08	0.01	5.877	4.84e-09	***

It is indicated that by removing the *sn*, *zr*, and mn from the trained model, all predictors are statistically significant. This indicates that all predictors affect the response except *sn*, *zr*, and *cu*.

The reported R-squared by the model is around 12%. It means the model does not fit the data effectively, and multiple linear regression has difficulties predicting carbon content. Therefore, this trained model cannot be used individually to predict the response variable on the dataset. To obtain a fitted model for oxygen regression, it takes the following form.

$$o_i = \beta_0 + \beta_1 mo_i + \beta_2 si_i + \beta_3 sn_i + \beta_4 zr_i + \beta_5 mn_i + \beta_6 cr_i + \beta_7 fe_i + \beta_8 ni_i + \beta_9 cu_i + \beta_{10} n_i + \varepsilon_i, \quad (5.8)$$

It gives the following fitting results.

Table 5.9 Multiple linear regression results of the equation (5.8)

	Estimate	Std-error	T-Value	P-value	Signif.Code
mo	-0.14	0.02	-6.62	4.37e-11	***
si	1.70	0.10	16.20	< 2e-16	***
mn	2.06	0.24	8.30	< 2e-16	***
cr	-0.09	0.03	-2.79	0.00517	**
fe	-0.04	0.00	-9.52	< 2e-16	***
ni	0.05	0.02	2.48	0.01301	*
n	1.8534	0.06751	27.45	< 2e-16	***

It is indicated that by removing the *sn*, *zr*, and *cu* from the trained model, all predictors are statistically significant. This indicates that all predictors affect the response except *sn*, *zr*, and *cu*.

The reported R-squared by the model is around 38%. It means the model does not fit the data adequately, and multiple linear regression has the challenge to predict oxygen content. Therefore, this trained model cannot be used uniquely to predict the response variable on the dataset. All the results of multiple linear regression are shown in Table 5.10.

Table 5.10 Summary of regression result over the general dataset

Predictors		R-squared	mo	si	sn	zr	mn	cr	fe	ni	cu	n
Responses	ti	62%	√	√	√	√	√	√	√	√	√	√
	al	23%	√	√	√	√	√	√	√	√	√	√
	v	55%	√		√	√	√	√	√	√		√
	c	12%	√	√				√	√	√	√	√
	o	38%	√	√			√	√	√	√		√

As we can see in Table 5.10, there is a checkmark in every predictor cell, which shows that its regression coefficient is statistically significant (at 0.05 significance level) for the related response variable. In the next section, the responses will be prioritized regarding each predictor.

5.4 Random forest

A random forest is a practical nonlinear supervised machine learning method because it measures the non-linearity and importance of each predictor. Random forest applies big numbers of random decision trees to examine important sets of variables. In this section, using the random forest method, we verify the results of multiple linear regression using a nonlinear extension of multiple regression and to find out the most important elements to predict the dependent variables. The random forest algorithm can be used for both classification and the regression of problems (Breiman, 2001). The following algorithm will be developed over the variables, to apply a random forest test across the dataset in statistical software "R"

$$y \sim RF(x_1, x_2, \ldots, x_p) \tag{5.9}$$

To apply a random forest test to determine the main predictors of titanium content as one of the dependent variables, the following model will apply.

$$ti \sim RF(mo, si, sn, zr, mn, cr, fe, ni, cu, n). \tag{5.10}$$

where ~ means the model holds with some additive statistical error.

Table 5.11 Random forest results of equation (5.10)

	%IncMSE	IncNodePurity
mo	23.3	282.7
si	21.3	321.8
sn	27.4	242.3
zr	28.7	299.4
mn	45.0	546.6
cr	29.3	370.7
fe	139.8	2559.4
ni	50.5	752.0
cu	21.7	168.0
n	39.2	390.6

As illustrated in Table 5.11, Mean Decrease Accuracy (%IncMSE) shows how much our model accuracy decreases if we leave out that variable. Mean Decrease Gini (IncNodePurity) is a measure of variable importance based on the Gini impurity index used for calculating the splits in trees. The higher the value of mean decrease accuracy or mean decrease Gini score, the higher the importance of the variable to our model.

The results indicate that the main chemical components to be as the predictors for the titanium content are *fe* and *ni*, *mn*, and *n*, respectively. To apply a random forest test to determine the main predictors of Aluminum content as a dependent variable, the following model will apply

$$al \sim RF(mo, si, sn, zr, mn, cr, fe, ni, cu, n). \quad (5.11)$$

Table 5.12 Random forest results of equation (5.11)

	%IncMSE	**IncNodePurity**
mo	31.0	131.2
si	18.2	91.0
sn	25.0	85.2
zr	32.7	120.7
mn	30.4	79.2
cr	26.5	91.0
fe	50.4	203.9
ni	30.0	95.2
cu	33.3	64.4
n	13.7	84.9

Results, shown in Table 5.12, show that the first main chemical components to predict the aluminum content are *fe*, *cu*, *zr*, and *mo*, respectively.

To apply a random forest test to determine the main predictors of vanadium content as the other dependent variable, the following model will apply.

$$y \sim RF(mo, si, sn, zr, mn, cr, fe, ni, cu, n). \quad (5.12)$$

The results are shown in the following table.

Table 5.13 Random forest results of equation (5.12)

	%IncMSE	IncNodePurity
mo	32.8	102.5
si	15.7	30.8
sn	34.4	85.2
zr	34.9	107.5
mn	48.9	53.9
cr	25.8	43.5
fe	73.1	122.2
ni	17.8	31.2
cu	12.1	15.3
n	18.7	37.3

Results determine that the first main chemical components to predict the vanadium content are *fe*, *mn*, *zr*, and *sn*, respectively. To apply a random forest test on carbon content as the other dependent variable, the following model will apply

y ~ RF(mo, si, sn, zr, mn, cr, fe, ni, cu, n).
(5.13)

Table 5.14 Random forest results of equation (5.13)

	%IncMSE	IncNodePurity
mo	25.2	0.9
si	46.3	1.2
sn	24.3	0.7
zr	29.2	0.9
mn	21.6	0.6
cr	30.9	1.0
fe	24.5	1.2
ni	25.8	1.0
cu	28.5	0.6
n	24.6	0.9

It shows that the first main chemical components to predict the carbon content are *si*, *cr*, *zr*, and *cu*, respectively.

Finally, to apply a random forest test on oxygen content as the last dependent variable, the following model will apply.

o ~ RF(mo, si, sn, zr, mn, cr, fe, ni, cu, n).
(5.14)

The results, summarized in Table 5.15, shows that the main chemical components to predict the oxygen content are *n*, *si*, *mn*, and *cr*.

Table 5.15 Random forest results of equation (5.14)

	%IncMSE	**IncNodePurity**
mo	18.3	21.7
si	53.2	45.8
sn	17.6	18.1
zr	21.3	23.6
mn	24.8	23.8
cr	20.6	21.1
fe	18.6	29.9
ni	17.4	21.6
cu	15.5	12.3
n	111.8	109.1

5.5 Result comparison between two methods

To compare the results of both linear and nonlinear tools, we use the absolute t-value of each predictor in multiple linear regression to prioritize the first four main elements to see if the results of both methods are matched are not. The results are depicted in Table 5.16.

Table 5.16 Comparison of LR and RF results over the general dataset

Response	**Method**	**Predictors**										**Priority results**
		mo	si	sn	zr	mn	cr	fe	ni	cu	n	
ti	RF					3		1	2		4	Same priorities
	LR					3		1	2		4	
al	RF	4			3			1		2		Not the same for second and third items
	LR	4				2		1		3		
v	RF			4	3	2		1				Just the first two items are the same
	LR	3				2		1			4	
c	RF		1		3		2			4		Just first item has the same priority in both methods
	LR		1					4		2	3	
o	RF		2			3	4				1	Just the first two items are the same
	LR		2			4		3			1	

Each number is representing the importance rank of each predictor of related response. As is depicted in Table 5.16, the priority of the first four important elements across the general dataset is the same for the titanium using linear regression and random forest methods, which are iron, nickel manganese, and nitrogen. For vanadium, the priority of importance is the same for the first two elements as iron and manganese. For oxygen, the priority of importance is the same for the first two elements as nitrogen and silicon. For carbon, it results that just for the first item, the linear regression and random forest priority is the same as silicon. For aluminum, the first and fourth priority is the same over the two statistical methods linear regression and random forest as iron and molybdenum, respectively.

5.6 Principal components analysis

The basis for multivariate data analysis is the principal component analysis technique. The principal component analysis is a method to reduce the variables when there is redundancy in the linear relationship between variables. In other words, principal components are applied when some of the variables are correlated to one another, and probably they are computing a similar quantity. One may prefer to reduce observed variables into a smaller number of uncorrelated principal components that will approximate the data to a high extent in terms of the variance-covariance matrix.

A principal component is an optimal linear combination of weighted observed variables. A score for each subject should be calculated over a given principal component as below, to perform a principal component analysis.

$$C_1 = b_{11}\, x_1 + b_{12}\, x_2 + \ldots\, b_{1p}\, x_p \qquad (5.15)$$

where

C_1= the subject's score on principal component 1 (the first component extracted).

b_{1p} = the regression coefficient for observed variable p of the first principal component.

x_p = the subject's score on observed variable p.

The first component of a principal component analysis covers the highest amount of total variance in the observed variables. This means that the first component correlates with most of the observed variables. The second component has two important features. First, this component covers the maximum variance of other observed variables of the dataset except the variables included in the first component. This means that the observed variables that did not have strong correlations with the first component are correlated to the second principal component. Also, the second component does not have any correlation with the first component, and the correlation between the first and second components is zero. The remaining components have the same characters versus the other components (Jolliffe, 2011).

All the predictor variables should have identical units of comparison to have the same unit for measurement, to run the principal component analysis. In this way, to standardize the different

units of dependent variables and to have a comparable unit for every dependent variable, we run the random forest between response and predictor variables to determine the main chemical component of each response based on the random forest results. In this way, we control the response errors as one of the main items we need to decrease them over the process control phase. We believe the most important part of a statistical process is to control the out-of-range observations, which lead to keep all the systems under control.

The most important predictors of the random forest for every response variable are presented in the next table.

Table 5.17 General dataset main chemical components based on the random forest test

Dependent variable(response)	The most important predictor
ti	fe
al	fe
v	fe
c	si
o	n

After running principal component analysis on the residuals of random forest, results are as below.

Table 5.18 Principal component results of the general dataset

	comp.1	comp.2	comp.3	comp.4	comp.5
resid_ti	0.869	0.318	0.308	0.222	
resid_al	-0.443	0.783	0.437		
resid_v	-0.14	-0.535	0.812	0.186	
resid_c					0.999
resid_o	-0.17		-0.232	0.957	

It shows that the first principal component covers the maximum variance of around 87% of titanium residuals. It means the errors which happened for titanium and because of falling fe as the main

predictor of titanium in the out-of-range zone are controlled by component number one. Also, it displays that the second principal component covers the maximum variance around 78% of aluminum residuals. It means the errors which happened for aluminum and because of falling Fe as its main predictor in the out-of-range zone are controlled by component number two.

The third principal component covers the maximum variance of around 81% of vanadium residuals. It means the errors which happened for vanadium and because of falling fe as its main predictor in the out-of-range zone are controlled by component number three. The fourth principle component covers the maximum variance of around 96% of oxygen residuals. It means the errors which happened for oxygen and because of falling nitrogen as its main predictor in the out-of-range zone are controlled by component number four. The fifth principle component covers the maximum variance of around 99% of carbon residuals. It means the errors which happened for carbon and because of falling silicon as its main predictor in the out-of-range zone are controlled by component number five. In the following graph, the biplot of the main components is depicted.

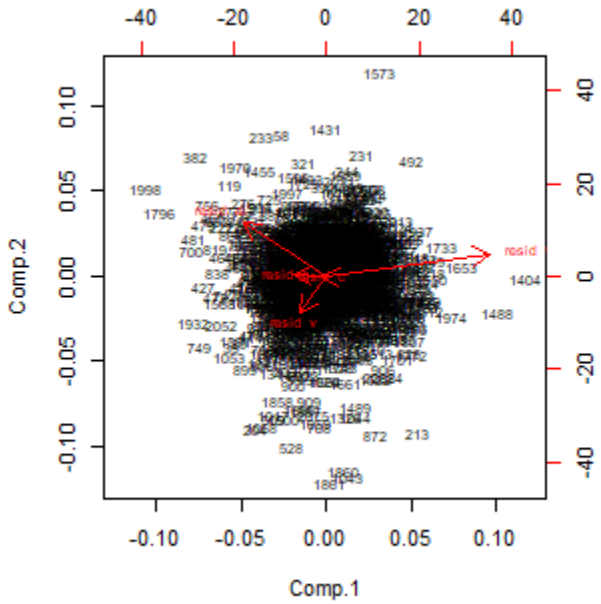

Figure 5.3 Five principal components plot of general dataset showing the observation record of data

5.7 Consistency check and reapply the methods over a data subset

After applying the multiple linear regression and the random forest test on the general dataset of 2018 and the analysis of the results, we check the consistency of the models presented for each of the dependent variables by reapplying the supervised learning methods on the dataset of a special product. In this regard, we chose a product that has been sending to one of the largest steel producers in the Middle East to compare its results with the general models presented and to see if these predicting models match each other.

The data set of this special product includes 17 variables and 675 observations.

5.7.1 Multiple linear regression over data subset

The fitted model to be checked for titanium regression for the selected product is as follows.

$$\mathrm{ti_i} = \beta_0 + \beta_1 \mathrm{mo_i} + \beta_2 \mathrm{si_i} + \beta_3 \mathrm{sn_i} + \beta_4 \mathrm{zr_i} + \beta_5 \mathrm{mn_i} + \beta_6 \mathrm{cr_i} + \beta_7 \mathrm{fe_i} + \beta_8 \mathrm{ni_i} + \beta_9 \mathrm{cu_i} + \beta_{10} \mathrm{n_i} + \varepsilon_i, \quad (5.16)$$

The relevant results are presented in Table 5.19.

Table 5.19 Multiple linear regression results of the equation (5.16)

	Estimate	Std-error	T-Value	P-value	Signif.Code
mo	0.49	0.18	2.709	0.00691	**
si	-0.02	0.60	-0.035	0.97186	
sn	-1.00	0.71	-1.406	0.16016	
zr	-0.87	0.30	-2.896	0.00391	**
mn	-7.37	1.31	-5.599	3.16e-08	***
cr	-0.48	0.23	-2.058	0.04002	*
fe	-0.65	0.02	27.939	< 2e-16	***
ni	-1.13	0.13	-8.510	< 2e-16	***
cu	-0.54	2.74	-0.197	0.84358	
n	-1.98	0.33	-6.006	3.13e-09	***

The results show that despite the general dataset results, which all predictors were statistically significant, here over the chosen product, *si*, *sn*, and *cu* are not significant for the titanium model.

So, by removing these chemical components from the model, the final titanium regression model of this selected product is as follows:

$$ti_i = \beta_0 + \beta_1 mo_i + \beta_4 zr_i + \beta_5 mn_i + \beta_6 cr_i + \beta_7 fe_i + \beta_8 ni_i + \beta_{10} n_i + \varepsilon_i, \quad (5.17)$$

Table 5.20 Multiple linear regression results of the equation (5.17)

	Estimate	Std-error	T-Value	P-value	Signif.Code
mo	0.46	0.17	2.578	0.01014	*
si	-1.19	0.19	-6.030	2.71e-09	***
sn	-7.34	1.26	-5.794	1.06e-08	***
zr	-0.59	0.21	-2.732	0.00647	**
mn	-0.64	0.02	27.940	< 2e-16	***
cr	-1.10	0.13	-8.401	2.66e-16	***
fe	-2.01	0.32	-6.123	1.57e-09	***

By removing the *si*, *sn*, and *cu* from the trained model, predictors become statistically significant. This indicates all predictors affect the response except *si*, *sn*, and *cu* for titanium at the selected product. The reported R-squared by the model is around 64%. It means the model fits the data adequately. The final fitted model for aluminum regression for the selected product takes the following form.

$$al_i = \beta_0 + \beta_1 mo_i + \beta_2 si_i + \beta_5 mn_i + \beta_7 fe_i + \varepsilon_i, \quad (5.18)$$

The related results are as follows:

Table 5.21 Multiple linear regression results of the equation (5.18)

	Estimate	Std-error	T-Value	P-value	Signif.Code
mo	-0.57	0.07	-7.890	1.24e-14	***
si	-2.22	0.40	-5.457	6.82e-08	***
mn	2.72	0.85	3.191	0.00148	**
fe	-0.13	0.01	-8.656	< 2e-16	***

It is indicated that by removing the *sn*, *zr*, *cr*, *ni*, *cu*, and *n* from the trained model, all predictors are statistically significant. This indicates all predictors affect the response *al* except these mentioned elements. The reported R-squared by the model is around 22%. It means the model fits

the data poorly in terms of prediction power. Therefore, the trained model should be considered not as a strong model to the response variable on the data subset. The final fitted model for the vanadium regression of the selected product data subset takes the following form.

$$v_i = \beta_0 + \beta_1 mo_i + \beta_2 si_i + \beta_4 zr_i + \beta_5 mn_i + \beta_6 cr_i + \beta_7 fe_i + \beta_8 ni_i + \varepsilon_i, \quad (5.19)$$

The related results are as follows:

Table 5.22 Multiple linear regression results of the equation (5.19)

	Estimate	Std-error	T-Value	P-value	Signif.Code
mo	-0.42	0.07	-5.358	1.16e-07	***
si	-0.64	0.26	-2.439	0.0150	*
zr	-0.63	0.08	-7.112	2.95e-12	***
mn	3.61	0.56	6.394	3.03e-10	***
cr	-0.50	0.09	-5.200	2.65e-07	***
fe	-0.13	0.01	13.325	< 2e-16	***
ni	0.12	0.05	2.127	0.0338	*

It is indicated that by removing the *sn*, *cu*, and *n* from the trained model, all predictors are significant. This means these three chemical components are not predictable variables for vanadium at the selected product. It is indicated that by removing the *sn*, *cu*, and *n* from the trained model, all predictors are statistically significant. This indicates that all predictors affect the response *v* except *sn*, *cu*, and *n*. The reported R-squared by the model is around 53%. It means the model fits the trained data adequately. The final fit model for the carbon regression of the selected product data subset takes the following form.

$$c_i = \beta_0 + \beta_1 mo_i + \beta_2 si_i + \beta_7 fe_i + \beta_8 ni_i + \beta_9 cu_i + \beta_{10} n_i + \varepsilon_i, \quad (5.20)$$

Related results are as follows:

Table 5.23 Multiple linear regression results of the equation (5.20)

	Estimate	Std-error	T-Value	P-value	Signif.Code
si	0.09	0.01	8.425	< 2e-16	***
fe	0.00	0.00	2.880	0.00411	**
cu	0.20	0.05	3.936	9.17e-05	***
n	0.01	0.01	2.680	0.00755	**

It is indicated that *mo*, *sn*, *zr*, *mn*, *cr*, and *ni* do not affect the carbon over the data subset. This means these six chemical components are not statistically significant predictors for carbon for the selected product. The reported R-squared by the model is around 13%. It means the model poorly fit the trained data and linear regression cannot be an efficient method to predict the carbon content. The final model for the oxygen regression of the selected product data subset is as the following form.

$$o_i = \beta_0 + \beta_1 mo_i + \beta_2 si_i + \beta_4 zr_i + \beta_7 fe_i + \beta_{10} n_i + \varepsilon_i, \tag{5.21}$$

Related results are presented in the next table.

Table 5.24 Multiple linear regression results of the equation (5.21)

	Estimate	**Std-error**	**T-Value**	**P-value**	**Signif.Code**
mo	-0.19	0.04	-4.87	1.36e-06	***
si	1.30	0.14	9.18	< 2e-16	***
zr	0.17	0.04	3.76	0.000182	***
fe	-0.03	0.01	-6.96	7.87e-12	***
n	1.26	0.07	16.16	< 2e-16	***

It shows *sn*, *mn*, *cr*, *ni*, and *cu* are not statistically significant predictors for oxygen over the data subset. The reported R-squared is around 39%. It means the model does not fit the trained data effectively, and we cannot rely on the result of linear regression as a sole method to predict the oxygen content over the data subset.

All the results of multiple linear regression over the selected product data subset and the general dataset are shown in Table 5.25.

As it is shown, the results are not the same between the general data set and the selected product data subset.

Table 5.25 Summary of regression result over the selected product data subset (S) and the general dataset(G)

Predictors		R-squared	mo	si	sn	zr	mn	cr	fe	ni	cu	n
ti	G	62%	√	√	√	√	√	√	√	√	√	√
	S	64%	√			√	√	√	√	√		√
al	G	23%	√	√	√	√	√	√	√	√	√	√
	S	22%	√	√			√		√			
v	G	55%	√		√	√	√	√	√	√		√
	S	53%	√	√		√	√	√	√	√		
c	G	12%	√	√				√	√	√	√	√
	S	13%		√					√		√	√
o	G	38%	√	√			√	√	√	√		√
	S	39%	√	√	√	√			√			√

Table 5.25 shows that for every predictor, if there is a checkmark in the associated cell, its regression coefficient is statistically significant for the related response variable.

5.7.2 Random Forest over the data subset

To determine the main predictors of titanium content as one of the dependent variables, we apply random forest, and the following model is applied

$$ti \sim RF(mo, si, sn, zr, mn, cr, fe, ni, cu, n). \quad (5.22)$$

The results are shown in Table 5.26.

Table 5.26 Random forest results of equation (5.22)

	%IncMSE	IncNodePurity
mo	13.8	77.1
si	10.7	65.9
sn	21.1	73.2
zr	19.6	87.3
mn	13.1	80.5
cr	17.6	115.0
fe	92.3	686.3
ni	28.7	177.0
cu	10.1	37.5
n	14.6	99.8

Results indicate that the main predictors for the titanium content are *fe*, *ni*, *sn*, and *zr*. However, the main predictors over the general dataset are *fe*, *ni*, *mn*, and *n*. To determine the main predictors of aluminum content as a dependent variable, we apply random forest as the following model.

$$al \sim RF(mo, si, sn, zr, mn, cr, fe, ni, cu, n). \quad (5.23)$$

Table 5.27 Random forest results of equation 5.23

	%IncMSE	IncNodePurity
mo	28.1	49.0
si	22.8	37.7
sn	17.8	26.4
zr	22.3	35.5
mn	11.3	21.6
cr	13.7	29.2
fe	25.7	56.0
ni	18.0	30.1
cu	10.7	14.5
n	10.4	26.2

Results presented in Table 5.27 illustrate that the first main chemical components to predict the aluminum content are *mo*, *fe*, *si*, and *zr*. However, the main predictors over the general dataset are

fe, *cu*, *zr*, and *mo*, respectively. The following model is applied to determine the main predictors of vanadium over the selected product dataset

v~RF(mo, si, sn, zr, mn, cr, fe, ni, cu, n).

(5.24)

Table 5.28 Random forest results of equation (5.24)

	%IncMSE	**IncNodePurity**
mo	24.9	44.6
si	8.8	10.4
sn	20.8	31.6
zr	27.4	39.1
mn	18.3	14.5
cr	13.3	13.7
fe	39.7	36.5
ni	11.5	12.2
cu	7.9	4.9
n	7.6	12.1

Results in Table 5.28 determine that the first main chemical components to predict the vanadium content are *fe*, *zr*, *mo*, and *sn*. The main predictors over the general dataset are *fe*, *mn*, *zr*, and *sn*, respectively. The following model is applied as random forest test on carbon content over the subset data

c ~ RF(mo, si, sn, zr, mn, cr, fe, ni, cu, n).

(5.25)

Table 5.29 indicates that the first main chemical components to predict the carbon content are *si*, *mo*, *zr*, and *n*. Though, the main predictors over the general dataset are *si*, *cr*, *zr*, and *cu*, respectively.

Table 5.29 Random forest results of equation (5.25)

	%IncMSE	IncNodePurity
mo	14.7	0.022
si	46.4	0.049
sn	11.4	0.019
zr	14.1	0.023
mn	10.7	0.016
cr	10.1	0.022
fe	9.5	0.028
ni	7.6	0.020
cu	9.8	0.011
n	12.4	0.025

And finally, to apply a random forest test on oxygen content as the last dependent variable over the data subset, the following model is applied

o ~ RF(mo, si, sn, zr, mn, cr, fe, ni, cu, n).

(5.26)

Table 5.30 Random forest results of equation (5.26)

	%IncMSE	IncNodePurity
mo	14.1	3.3
si	30.1	6.6
sn	13.0	2.3
zr	9.8	3.1
mn	8.4	2.7
cr	12.7	3.6
fe	11.4	5.8
ni	12.3	3.9
cu	10.3	1.6
n	50.0	15.0

Results presented in Table 5.30 shows that the first main chemical components to predict the oxygen content are *n*, *si*, *mo*, and *sn*. However, the main predictors over the general dataset are *n*, *si*, *mn*, and *cr*, respectively.

5.7.3 Linear regression and random forest results comparison over the data subset

To compare the results of both linear and nonlinear tools over the data subset, we use the absolute t-value of each predictor in multiple linear regression to prioritize the first four main elements to see if the results of both methods are matched are not. The results are depicted in Table 5.31.

Table 5.31 Comparison of linear regression (LR) and random forest (RF) results over the data subset

Response	**Method**	**Predictors**										**Priority results**
		mo	si	sn	zr	mn	cr	fe	ni	cu	n	
ti	RF			3	4			1	2			Not the same for third and fourth items
	LR				4			1	2		3	
al	RF	1	3		4			2				Just the first items are the same
	LR	2	3			4		1				
v	RF	3	4		2			1				Just the first two items are the same
	LR	4			2	3		1				
c	RF	2	1		3						4	First and fourth items have the same priority in both methods
	LR		1					3		2	4	
o	RF	3	2	4							1	Just the first two items are the same
	LR	4	2					3			1	

Each number is representing the importance rank of each predictor of related response over the random forest and linear regression methods.

5.7.4 PCA over data subset

All the predictor variables should have identical units of comparison to have the same unit for measurement, and to run principal component analysis over the data subset of the selected product. In this way, to standardize the different units of dependent variables and to have a comparable unit for every dependent variable, we run the random forest between response and predictor variables to determine the main chemical component of each response. Whereas the most important part of a statistical process control method is to control the out-of-range observations which lead to keep all the system under control, we control the response errors as one of the main items we need to decrease them over the process control phase by running the principal component analysis over the out-of-range observations of the main predictor variables of each response.

The most important predictors of random forest tests for every response variable are as below.

Table 5.32 Data subset main chemical components based on the random forest test

Response	The most important predictor
ti	fe
al	mo
v	fe
c	si
o	n

The results of the principal component analysis, which applied over the data subset, are shown below.

Table 5.33 Principal component results over the selected data subset(S) compared with the general dataset(G)

Items	Dataset	comp.1	comp.2	comp.3	comp.4	comp.5
resid_ti	G	0.869	0.318	0.308	0.222	
	S	0.865	0.325	0.379		
resid_al	G	-0.443	0.783	0.437		
	S	-0.457	0.800	0.373	-0.110	
resid_v	G	-0.140	-0.535	0.812	0.186	
	S	-0.179	-0.502	0.845		
resid_c	G					0.999
	S					1.000
resid_o	G	-0.170		-0.232	0.957	
	S	-0.110			0.991	

It shows that the first principal component of the data subset again covers the maximum variance around 87% of titanium residuals, which is very close to the result of the general dataset. It means the errors which happened for titanium and because of falling fe as the main predictor of titanium in the out-of-range zone are controlled by component number one.

Also, it displays that the second principal component of the data subset covers the maximum variance around 80% of aluminum residuals, which is close to the general dataset result 78%. It means the errors which happen for aluminum and because of falling molybdenum as its main predictor in the out-of-range zone are controlled by component number two. The third principal component covers the maximum variance of around 85% of vanadium residuals over the data subset, which is closed with the result of the general dataset 81%. It means the errors which happen for vanadium and because of falling fe as its main predictor in the out-of-range zone are controlled by component number three. The fourth principle component covers the maximum variance of around 99% of oxygen residuals, which is very close to the general dataset result 96%. It means the errors which happen for oxygen and because of falling nitrogen as its main predictor in the out-of-range zone are controlled by component number four.

The fifth principle component covers the maximum variance of around 100% of carbon residuals, which is the same for the general dataset. It means the errors which happened for carbon and because of falling silicon as its main predictor in the out-of-range zone are controlled by component number five. In the following graph, the biplot of the main components is depicted.

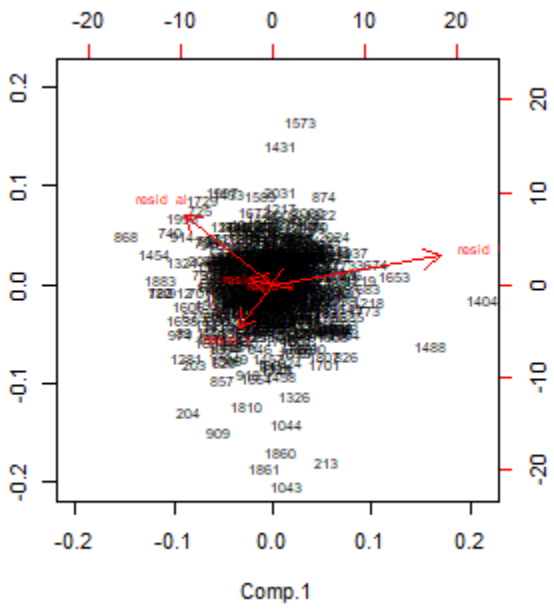

Figure 5.4 Five principal components plot of data subset showing observation record of data

5.8 Discussion

In this section, the results of the supervised and unsupervised statistical method over the general dataset and the subset of the selected product will be compared to see if we can present a unique algorithm to guide the titanium producers' production and quality processes. For this purpose, a low-carbon product dataset is selected to see if the results meet the general dataset results.

After running the multiple linear regression over both the general dataset and selected product data subset, it was noticed that the chemical reaction of components is not matched when products are different. Table 5.34 shows the difference of results between the general dataset and the selected product dataset.

Table 5.34 Regression results between general dataset and selected product data subset

Dataset	**Dependent variable**	**mo**	**si**	**sn**	**zr**	**mn**	**cr**	**fe**	**ni**	**cu**	**n**
General	ti	√	√	√	√	√	√	√	√	√	√
Selected Product	ti	√			√	√	√	√	√		√
General	al	√	√	√	√	√	√	√	√	√	√
Selected Product	al	√	√			√		√			
General	v	√		√	√	√	√	√	√		√
Selected Product	v	√	√		√	√	√	√	√		
General	c	√	√				√	√	√	√	√
Selected Product	c		√					√		√	√
General	o	√	√			√	√	√	√		√
Selected Product	o	√	√		√			√			√

As is depicted in Table 5.34, there is a checkmark in every predictor cell in which its regression coefficient is statistically significant for the related response variable. The predictors are different when we choose a specific product of the general dataset as the selected dataset. In this test, a low-carbon product dataset is selected which is called selected product in Table 5.34, and titanium (ti) predictors are molybdenum (mo), zirconium (zr), manganese (mn), chromium (cr), iron (fe), nickel (ni), and nitrogen (n) for this product. For the general dataset, the predictors are molybdenum (mo), silicon (si), tin (sn), zirconium (zr), manganese (mn), chromium (cr), iron (fe), nickel (ni), cupper (cu), and nitrogen (n). It shows that we cannot expand the results of the general dataset to other specific products. As shown in the table, the predictors are different for the other responses as aluminum, vanadium, carbon, and oxygen. The priority of the first four predictor variables based

on the random forest results across the general dataset and selected product data subset are shown below.

Table 5.35 Random forest results between general dataset and selected product data subset

Dataset	**Dependent variable**	**mo**	**si**	**sn**	**zr**	**mn**	**cr**	**fe**	**ni**	**cu**	**n**
General	ti					3		1	2		4
Selected Product	ti			3	4			1	2		
General	al	4			3			1		2	
Selected Product	al	1	3		4			2			
General	v			4	3	2		1			
Selected Product	v	3	4		2			1			
General	c		1		3		2			4	
Selected Product	c	2	1		3						4
General	o		2			3	4				1
Selected Product	o	3	2	4							1

The mentioned numbers in Table 5.35 show the importance of each predictor of every response. As it is presented in Table 5.35, the significance of the first four predictors is different for the general dataset and selected product dataset. In the case of titanium (ti), the first more important predictors over the general dataset are iron (fe), nickel (ni), manganese (mn), and nitrogen (n), respectively. Over the selected product data subset, iron (fe), nickel (ni), tin (sn), zirconium (zr) are the most important predictors. By looking at this table, it shows that for titanium and oxygen, the first two predictors have the same level of significance. The number of applied trees in the random forest method is 500, and the line selection rate is 2321. Also, the column selection rate in our analysis using R is 10/3 = 3.33. The tree type is regression, and there is not vote type because it is a regression tree and not a classification tree.

Regarding the applying of principal component analysis over the general dataset and selected product data subset explained in Sections 5.6 and 5.7, the most important predictor of each response is determined based on the result of random forest and out-of-range observations of this main predictor that have been extracted to run principal component analysis across these residuals.

The result shows that the first principal component of the data subset covers the maximum variance of titanium residuals, which meets the result of the general dataset. It means the errors that happened for titanium, and because of falling fe as the main predictor of titanium in the out-of-range zone, are controlled by component number one.

Also, it displays that the second principal component of the data subset covers the maximum variance of aluminum residuals, which is close to the general dataset result. It means the errors that happen for aluminum, and because of falling molybdenum as its main predictor in the out-of-range zone, are controlled by component number two.

The third principal component covers the maximum variance of vanadium residuals over the data subset, which is close to the result of the general dataset. It means the errors which happen for vanadium and because of falling fe as its main predictor in the out-of-range zone are controlled by component number three.

The fourth principle component covers the maximum variance of oxygen residuals, which is very close to the general dataset result. It means the errors that happen for oxygens and because of falling nitrogen as its main predictor in the out-of-range zone are controlled by component number four.

The fifth principle component covers the maximum variance of carbon residuals, which is the same for the general dataset. It means the errors which happened for carbons and because of falling silicon as its main predictor in the out-of-range zone are controlled by component number five.

5.9 Conclusion

This chapter made it possible to validate the proposed model using a real dataset of a titanium producer. This practical demonstration was performed using the statistical software “R”. Quality technicians, and metallurgists of the partner company confirmed the feasibility and accuracy of the proposed model based on their experience.

However, our results show that the results are not identical when we choose a specific product of the general dataset as the selected dataset. Therefore, it is needed to run this model across the data subset of every product and analyze the results. The framework limitations are discussed in the following chapter as well as future research opportunities.

CHAPTER 6 CONCLUSION AND RECOMMENDATIONS

The main objective of this project was to propose a general framework to determine the most crucial variables to be controlled in the ferrotitanium process and to prioritize them. During this research, using the design of experiments (DOE) approach, a new statistical model was developed to analyze the potential correlation between the variables and to determine the important elements of the ferrotitanium process.

We ran multiple linear regressions over the collected dataset to find out if there is any linear correlation between variables. We ran the random forest across the collected dataset to study any nonlinear regression between variables and to prioritize the most important predictors of every response. Then, we ran a principal component analysis over the residuals of random forest results to study other unknown effective elements of the process. Finally, we reapplied the recommended statistical methods over a data subset of a specific product to compare the results and to find out whether the predictor variables are the same in the response variables over both datasets and data subsets. We can then validate the consistency of the proposed model. If the results were not the same for a response variable, it would mean that more research would be necessary using the experts' viewpoints to find the linear or non-linear nature of the trained data.

Using the proposed model, we discovered a correlation between the dependent and independent variables in the ferrotitanium dataset and identified the most critical elements of each dependent variable. The results show that there is not any correlation between vanadium (v) with copper (cu), no correlation between oxygen (o) with tin (sn), and copper (cu), and no correlation between carbon (c) with tin (sn), zirconium (zr), and manganese (mn) in both general data set and selected product data subset. Therefore, if one of these mentioned response variables falls out of the control limits, these answers help to avoid wasting time analyzing the predictor variables which do not correlate with the response.

Also, the main predictors of titanium (ti), oxygen (o), and vanadium (v) are iron, nitrogen, and iron, respectively; and they are the same between two general datasets and selected product data subset. For aluminum and carbon, the results are different between the general dataset and selected data subset. Therefore, in the case of falling the result of these response variables out of the quality

limits, by starting to analyze the most important predictor variables, we can better and faster find the reason for the unpredicted reaction and decrease the production process stop time.

These practical results can help ferrotitanium producers to understand better the link between out of range results and the main reasons for each error. Indeed, this study complements the current scientific literature in statistical process control. The strength of the proposed model lies in using different statistical models to analyze the linearity and non-linearity of the collected dataset and to study the hidden and unknown variables of the ferrotitanium process.

However, we acknowledge that the proposed model has some limitations. First, the model was validated using the dataset of a ferrotitanium producer in Canada and needs to be applied to other ferrotitanium producers' datasets to test its reliability. The model could also be tested in other business units or companies to test its performance and robustness for other products. From a practical standpoint, another limit to implement this proposed model is the requirement of vast data and the relevant specialized knowledge in statistical and data analysis. Such expertise is rare in this industry currently. On a technical level, we also need more research on the output of the principal component analysis method to analyze and find out the other hidden and unknown variables affecting the results.

This study, however, represents the first step towards better process control in the ferrotitanium industry. The structure of this research can be a guide for future researchers attempting to understand better material reactions based on historical data observations.

REFERENCES

Abdallah, Z. S., Du, L., & Webb, G. I. (2017). Data Preparation. In *Encyclopedia of Machine Learning and Data Mining* (pp. 318–327). Melbourne: Springer.

Alpaydin, E. (2009). *Introduction to Machine Learning* (2th ed.). Cambridge: MIT Press.

Altay, O., Gurgenc, T., Ulas, M., & Özel, C. (2019). Prediction of wear loss quantities of ferro-alloy coating using different machine learning algorithms. *Friction*, 1–8.

Amruthnath, N., & Gupta, T. (2019). Factor Analysis in Fault Diagnostics Using Random Forest. *Industrial Engineering & Management*, *8*(1), 1–10.

Astakhov, V. P. (2012). Design of experiment methods in manufacturing: basics and practical applications. In *Statistical and computational techniques in manufacturing* (pp. 1–54). Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. *Analyst*, *135*(2), 230–267.

Bzdok, D., Krzywinski, M., & Altman, N. (2018). *Points of significance: Machine learning: supervised methods*. Berline: Nature Publishing Group.

Chen, Q., Wynne, R. J., Goulding, P., & Sandoz, D. (2000). The application of principal component analysis and kernel density estimation to enhance process monitoring. *Control Engineering Practice*, *8*(5), 531–543.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Davis, J., Edgar, T., Graybill, R., Korambath, P., Schott, B., Swink, D., … Wetzel, J. (2015). Smart manufacturing. *Annual Review of Chemical and Biomolecular Engineering*, *6*, 141–160.

Durakovic, B. (2017). Design of experiments application, concepts, examples: State of the art. *Periodicals of Engineering and Natural Sciences*, *5*(3), 421–439.

Ferrer, A. (2014). Latent structures-based multivariate statistical process control: A paradigm shift. *Quality Engineering*, *26*(1), 72–91.

Fisher, R. A. (1949). The design of experiments (5th ed.). Edinburgh: Oliver & Boyd.

Holappa, L. (2010). Towards sustainability in ferroalloy production. *Journal of the Southern African Institute of Mining and Metallurgy*, *110*(12), 703–710.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89–106.

Jadhav, S. D., & Channe, H. P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR)*, *5*(1), 1842–1845.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (1st ed.). New York: Springer.

Jolliffe, I. (2011). *Principal component analysis* (2th ed.). New York: Springer.

Kharbach, M., Cherrah, Y., Vander Heyden, Y., & Bouklouze, A. (2017). Multivariate statistical process control in product quality review assessment – A case study. *Annales Pharmaceutiques Françaises*,*75*(6), 446–454.

Laha, D., Ren, Y., & Suganthan, P. N. (2015). Modeling of steelmaking process with effective machine learning techniques. *Expert Systems with Applications*, *42*(10), 4687–4696.

Liu, Y.-J., André, S., Saint Cristau, L., Lagresle, S., Hannas, Z., Calvosa, É., … Duponchel, L. (2017). Multivariate statistical process control (MSPC) using Raman spectroscopy for in-line culture cell monitoring considering time-varying batches synchronized with correlation optimized warping (COW). *Analytica Chimica Acta*, *952*, 9–17.

Montgomery, D. C. (2009). *Statistical quality control* (7th ed.). Arizona: Wiley New York.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago: Irwin Chicago.

Osang, Je., Udoimuk, A. B., Etta, E. B., Ushie, F. O., & Offiong, N. E. (2013). Methods of gathering data for research purpose and applications using ijser acceptance rate of monthly paper publication (march 2012 edition-may 2013 edition). *IOSR Journal Of ComputerEngineering (IOSR-JCE)*, *15*(2), 59–65.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128–138.

Rogalewicz, M. (2012). Some notes on multivariate statistical process control. *Management and Production Engineering Review*, *3*(4), 80–86.

Rogalewicz, M., & Poznańska, P. (2013). The Methodology of Controlling Manufacturing Processes with the Use of Multivariate Statistical Process Control Tools. *Journal of Trends in the Development of Machinery and Associated Technology*, *17*(1), 89–93.

Teh, Y. W. (2006). A Bayesian Interpretation of Interpolated Kneser-Ney NUS School of Computing Technical Report TRA2/06. *National University of Singapore*, 1–21.

Tian, Y., Fu, M., & Wu, F. (2015). Steel plates fault diagnosis on the basis of support vector machines. *Neurocomputing*, *151*, 296–303.

Toledo, J. C. de, Lizarelli, F. L., Junior, S., & Bispo, M. (2017). Success factors in the implementation of statistical process control: action research in a chemical plant. *Production*, *27*.

Umeshini, S., & PSumathi, C. (n.d.). *A Survey ON DATA MINING IN STEEL INDUSTRIES*.

Zhang, X., Zhao, J., Wang, W., Cong, L., & Feng, W. (2011). An optimal method for prediction and adjustment on byproduct gas holder in steel industry. *Expert Systems with Applications*, *38*(4), 4588–4599.

Zou, C., & Qiu, P. (2009). Multivariate statistical process control using LASSO. *Journal of the American Statistical Association*, *104*(488), 1586–1596.

APPENDIX A R-STUDIO CODES

```
original_data<-as.data.frame(read.csv(paste(path,"2018.txt", sep=""), sep="\t"))


# take Al, C, O

original_data <- original_data[ ,3:ncol(original_data)]

# remove date

colnames(original_data) <- c("y_al","x_mo", "x_si", "x_sn", "x_zr", "x_mn", "x_cr",

                "y_v", "x_fe", "y_c", "x_ni", "x_cu", "x_n", "y_o", "y_ti",

                "type", "product")


X<- as.data.frame(original_data[,c(2,3,4,5,6,7,9,11,12,13,16,17)])


# change discrete variable to continuous variable

X$x_cu <- as.numeric(as.character(X$x_cu))

Y <- original_data[,c(1,8,10,14,15)]

# plot scatter plot

# change discrete variable to continuous variable

Y$y_o <- as.numeric(as.character(Y$y_o))

x <- cbind(X,Y)

# remove NAs

x <- x[ which( complete.cases(x)) , ]

# remove outlier data

# step 1
```

```
x <- x[!(x$y_ti<65),]
x <- x[!(x$y_ti>75),]
# step 2
x <- x[!(x$x_zr>2),]
x <- x[!(x$y_o>3),]
# step 3
x <- x[!(x$x_mn>0.3),]
x <- x[!(x$x_cu>0.2),]
x <- x[!(x$x_cr>2.5),]

# step 4
x <- x[!(x$x_sn>1),]
x <- x[!(x$x_si>0.5),]
# remove KIB
x <- x[!(x$product=='KIB'),]
# remove PCA outlier
x <- x[-c(1165,1045),]
rownames(x)=NULL
pairs(x[,1:6], oma=c(0,0,0,0))
pairs(x[,7:12], oma=c(0,0,0,0))
pairs(x[,1:12], oma=c(0,0,0,0))

##set all x
```

```
summary(lm(y_ti~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
##Replacing dependent variable y by al
summary(lm(y_al~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
##Replacing dependent variable y by v
summary(lm(y_v~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
##Replacing dependent variable y by c
summary(lm(y_c~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
##Replacing dependent variable y by o
summary(lm(y_o~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
# remove x one by one until all x are significant
summary(lm(y_ti~x_si+x_sn, data=x))
### Test of significant for Ti
summary(lm(y_ti~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
### All are significant for ti, so stop-------
### Test of significant for al
summary(lm(y_al~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
### All are significant for al, so stop-------
##Test of significant for v
summary(lm(y_v~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
##Test of significant for v without si
summary(lm(y_v~x_mo+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
##Test of significant for v without si, and Cu
summary(lm(y_v~x_mo+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_n, data=x))
```

```
### All are significant for v, so stop-------

##Test of significant for c

summary(lm(y_c~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))

##Test of significant for c without sn

summary(lm(y_c~x_mo+x_si+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))

##Test of significant for c without sn, zr

summary(lm(y_c~x_mo+x_si+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))

##Test of significant for c without sn, zr,mn

summary(lm(y_c~x_mo+x_si+x_cr+x_fe+x_ni+x_cu+x_n, data=x))

### All are significant for c, so stop-------

##Test of significant for o

summary(lm(y_o~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))

##Test of significant for o without sn

summary(lm(y_o~x_mo+x_si+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))

##Test of significant for o without sn, zr

summary(lm(y_o~x_mo+x_si+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))

##Test of significant for o without sn, zr, cu

summary(lm(y_o~x_mo+x_si+x_mn+x_cr+x_fe+x_ni+x_n, data=x))

### All are significant for o, so stop-------

# this is called backward model selection.

# then doing the Random Forest test

library(randomForest)

###RandomForest test for ti
```

```
bigmodel_y_ti <- randomForest(y_ti~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x, importance=TRUE)

importance(bigmodel_y_ti)

# the most important random forest predictor

smallmodel_y_ti <- randomForest(y_ti~x_fe, data=x, importance=TRUE)

yhat <- predict(smallmodel_y_ti)

xhat <- x$x_fe[order(x$x_fe)]

yhat <- yhat[order(x$x_fe)]

par(mar=c(4,4,0.1,0.1))

#Plot ti versus fe

plot(x$x_fe, x$y_ti, ylab="Ti", xlab="Fe", col='gray')

points(smooth.spline(x$y_ti~x$x_fe, spar = 0.5), type='l', lty=3, col='blue', lwd=5)

abline(lm(x$y_ti~x$x_fe), col='green', lwd=3)

resid_ti <- x$y_ti-predict(bigmodel_y_ti)

PATH="C:/"

###RandomForest test for al

bigmodel_y_al <- randomForest(y_al~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x, importance=TRUE)

importance(bigmodel_y_al)

# the most important random forest predictor

smallmodel_y_al <- randomForest(y_al~x_fe, data=x, importance=TRUE)

yhat <- predict(smallmodel_y_al)

xhat <- x$x_fe[order(x$x_fe)]
```

```
yhat <- yhat[order(x$x_fe)]


resid_al <- x$y_al-predict(bigmodel_y_al)

PATH="C:/"

###RandomForest test for v

bigmodel_y_v <- randomForest(y_v~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n,
data=x, importance=TRUE)

importance(bigmodel_y_v)

# the most important random forest predictor

smallmodel_y_v <- randomForest(y_v~x_fe, data=x, importance=TRUE)

yhat <- predict(smallmodel_y_v)

xhat <- x$x_fe[order(x$x_fe)]

yhat <- yhat[order(x$x_fe)]

resid_v <- x$y_v-predict(bigmodel_y_v)

PATH="C:/"

###RandomForest test for c

bigmodel_y_c <- randomForest(y_c~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n,
data=x, importance=TRUE)

importance(bigmodel_y_c)

# the most important random forest predictor

smallmodel_y_c <- randomForest(y_c~x_si, data=x, importance=TRUE)


yhat <- predict(smallmodel_y_c)

xhat <- x$x_si[order(x$x_si)]
```

```
yhat <- yhat[order(x$x_si)]

resid_c <- x$y_c-predict(bigmodel_y_c)

PATH="C:/"

#RandomForest test for O

bigmodel_y_o <- randomForest(y_o~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n,
data=x, importance=TRUE)

importance(bigmodel_y_o)

# the most important random forest predictor

smallmodel_y_o <- randomForest(y_o~x_n, data=x, importance=TRUE)

yhat <- predict(smallmodel_y_o)

xhat <- x$x_n[order(x$x_n)]

yhat <- yhat[order(x$x_n)]

resid_o <- x$y_o-predict(bigmodel_y_o)

### Principal components on residuals of random forest

resid_matrix <- cbind(resid_ti, resid_al, resid_v, resid_c, resid_o)

pca <- princomp(resid_matrix)

pca$loadings

biplot(pca, pch='o', expand=0.5)

##########################################

summary(pca)

#PC1

pc1 <- pca$scores[,1]

pc2 <- pca$scores[,2]

library('qicharts')
```

```
par(mfrow=c(2,1))
qic(pc1, chart = 'i')
qic(resid_al, chart = 'i')
PATH="C:/"
# create path+filename
FILE = paste(PATH, "yo_si.png", sep='')
# start file
par(mar=c(4,4,0.1,0.1))


#Plot ti versus fe
plot(x$x_fe, x$y_ti, ylab="Ti", xlab="Fe", col='gray')
points(xhat, yhat, col='red', type='l', lwd=3)
points(smooth.spline(x$y_ti~x$x_fe, spar = 0.3), type='l', lty=3, col='blue', lwd=5)

#dev.off()
# end file
#Plot al versus fe
plot(x$x_fe, x$y_al, ylab="Al", xlab="Fe", col='gray')
points(xhat, yhat, col='red', type='l', lwd=3)
points(smooth.spline(x$y_al~x$x_fe, spar = 0.3), type='l', lty=3, col='blue', lwd=5)

#Plot v versus fe
```

```
plot(x$x_fe, x$y_v, ylab="v", xlab="Fe", col='gray')
points(xhat, yhat, col='red', type='l', lwd=3)
points(smooth.spline(x$y_v~x$x_fe, spar = 0.3), type='l', lty=3, col='blue', lwd=5)


#Plot c versus si
plot(x$x_si, x$y_c, ylab="C", xlab="Si", col='gray')
points(xhat, yhat, col='red', type='l', lwd=3)
points(smooth.spline(x$y_c~x$x_si, spar = 0.3), type='l', lty=3, col='blue', lwd=5)

#Plot O versus si
plot(x$x_si, x$y_o, ylab="Oxigen", xlab="Silice", col='gray')
points(xhat, yhat, col='red', type='l', lwd=3)
points(smooth.spline(x$y_o~x$x_si, spar = 0.3), type='l', lty=3, col='blue', lwd=5)
################
# select JSW as the product (data subset)
x <- x[x$product=='JSW',]
#par(mar=c(0,0,0,0))
pairs(x[,1:6], oma=c(0,0,0,0))
pairs(x[,7:12], oma=c(0,0,0,0))
pairs(x[,1:12], oma=c(0,0,0,0))
#####Linear Regression for one product(JSW)
##set all x for Ti
```

```
summary(lm(y_ti~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
##Replacing dependent variable y by al
summary(lm(y_al~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu, data=x))
##Replacing dependent variable y by v
summary(lm(y_v~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu, data=x))
##Replacing dependent variable y by c
summary(lm(y_c~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu, data=x))
##Replacing dependent variable y by o (it gives error)
summary(lm(y_o~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu, data=x))

# remove x one by one until all x are significant
summary(lm(y_ti~x_si+x_sn, data=x))
### Test of significant for Ti
summary(lm(y_ti~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
## si, sn, cu are not significant
#remove si
summary(lm(y_ti~x_mo+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
#remove si and sn
summary(lm(y_ti~x_mo+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
#remove si and sn and cu
summary(lm(y_ti~x_mo+x_zr+x_mn+x_cr+x_fe+x_ni+x_n, data=x))
### All are significant for ti, so stop-------
### Test of significant for al
```

```
summary(lm(y_al~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
##remove sn
#summary(lm(y_al~x_mo+x_si+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu, data=x))
##remove sn , cr
#summary(lm(y_al~x_mo+x_si+x_zr+x_mn+x_fe+x_ni+x_cu, data=x))
##remove sn , cr, ni
#summary(lm(y_al~x_mo+x_si+x_zr+x_mn+x_fe+x_cu, data=x))
##remove sn , cr, ni, cu, zr,n
summary(lm(y_al~x_mo+x_si+x_mn+x_fe, data=x))

### All are significant for al, so stop-------

##Test of significant for v
summary(lm(y_v~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
##remove sn
##summary(lm(y_v~x_mo+x_si+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu, data=x))
##remove sn,cu
##summary(lm(y_v~x_mo+x_si+x_zr+x_mn+x_cr+x_fe+x_ni, data=x))
##remove sn,cu,n
summary(lm(y_v~x_mo+x_si+x_zr+x_mn+x_cr+x_fe+x_ni, data=x))

### All are significant for v, so stop-------
```

```
##Test of significant for c
summary(lm(y_c~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
##remove sn
#summary(lm(y_c~x_mo+x_si+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu, data=x))
##remove sn, zr
#summary(lm(y_c~x_mo+x_si+x_mn+x_cr+x_fe+x_ni+x_cu, data=x))
##remove sn, zr,cr
#summary(lm(y_c~x_mo+x_si+x_mn+x_fe+x_ni+x_cu, data=x))
##remove sn, zr,cr,mn
#summary(lm(y_c~x_mo+x_si+x_fe+x_ni+x_cu, data=x))
##remove sn, zr,cr,mn,mo, ni
summary(lm(y_c~x_si+x_fe+x_cu+x_n, data=x))
### All are significant for c, so stop-------
##Test of significant for o
summary(lm(y_o~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x))
##remove sn
#summary(lm(y_o~x_mo+x_si+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu, data=x))
##remove sn,mn
#summary(lm(y_o~x_mo+x_si+x_zr+x_cr+x_fe+x_ni+x_cu, data=x))
##remove sn,mn,cr
#summary(lm(y_o~x_mo+x_si+x_zr+x_fe+x_ni+x_cu, data=x))
##remove sn,mn,cr,ni,cu
summary(lm(y_o~x_mo+x_si+x_zr+x_fe+x_cu+x_n, data=x))
```

```
### All are significant for o, so stop-------
par(mfrow=c(2,5), mar=c(1,2,4,2))
boxplot(x[, 1], main = 'x_al')
boxplot(x[, 2], main = 'x_')
boxplot(x[, 3], main = 'x_')
boxplot(x[, 4], main = 'x_')
boxplot(x[, 5], main = 'x_')
boxplot(x[, 6], main = 'x_')
boxplot(x[, 7], main = 'x_')
boxplot(x[, 8], main = 'x_')
boxplot(x[, 9], main = 'x_')
boxplot(x[, 10], main = 'x_')
par(mfrow=c(1,5), mar=c(2,2,4,2))
boxplot(x[, 13], main = 'y_', col='gray')
boxplot(x[, 14], main = 'y_', col = 'yellow')
boxplot(x[, 15], main = 'y_')
boxplot(x[, 16], main = 'y_')
boxplot(x[, 17], main = 'y_')
# this is called backward model selection.
###########Random Forest
library(randomForest)
###RandomForest test for ti
bigmodel_y_ti <- randomForest(y_ti~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n,
data=x, importance=TRUE)
```

```
importance(bigmodel_y_ti)
# the most important random forest predictor
smallmodel_y_ti <- randomForest(y_ti~x_fe, data=x, importance=TRUE)
xhat <- x$x_fe[order(x$x_fe)]
yhat <- yhat[order(x$x_fe)]
resid_ti <- x$y_ti-predict(bigmodel_y_ti)



#####


##Plot Ti versus Fe
par(mfrow=c(1,1),mar=c(4,4,0.1,0.1))


plot(x$x_fe, x$y_ti, ylab="Ti", xlab="Fe", col='gray')
points(xhat, yhat, col='red', type='l', lwd=3)
points(smooth.spline(x$y_ti~x$x_fe, spar = 0.3), type='l', lty=3, col='blue', lwd=5)
###RandomForest test for al
bigmodel_y_al <-
randomForest(y_al~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n, data=x,
importance=TRUE)
importance(bigmodel_y_al)
# the most important random forest predictor
smallmodel_y_al <- randomForest(y_al~x_mo, data=x, importance=TRUE)
```

```
xhat <- x$x_mo[order(x$x_mo)]

yhat <- yhat[order(x$x_mo)]

resid_al <- x$y_al-predict(bigmodel_y_al)


##Plot al versus mo

par(mar=c(4,4,0.1,0.1))

plot(x$x_mo, x$y_al, ylab="Al", xlab="Mo", col='gray')

points(xhat, yhat, col='red', type='l', lwd=3)

points(smooth.spline(x$y_al~x$x_mo, spar = 0.3), type='l', lty=3, col='blue', lwd=5)


###RandomForest test for v

bigmodel_y_v <- randomForest(y_v~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n,
data=x, importance=TRUE)

importance(bigmodel_y_v)

# the most important random forest predictor


smallmodel_y_v <- randomForest(y_v~x_fe, data=x, importance=TRUE)


xhat <- x$x_fe[order(x$x_fe)]

yhat <- yhat[order(x$x_fe)]

resid_v <- x$y_v-predict(bigmodel_y_v)


##Plot V versus Fe

par(mar=c(4,4,0.1,0.1))
```

```
plot(x$x_fe, x$y_v, ylab="V", xlab="Fe", col='gray')

points(xhat, yhat, col='red', type='l', lwd=3)

points(smooth.spline(x$y_v~x$x_fe, spar = 0.3), type='l', lty=3, col='blue', lwd=5)

###RandomForest test for c

bigmodel_y_c <- randomForest(y_c~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n,
data=x, importance=TRUE)

importance(bigmodel_y_c)

# the most important random forest predictor

smallmodel_y_c <- randomForest(y_c~x_si, data=x, importance=TRUE)


xhat <- x$x_si[order(x$x_si)]

yhat <- yhat[order(x$x_si)]



resid_c <- x$y_c-predict(bigmodel_y_c)



##Plot C versus si

par(mar=c(4,4,0.1,0.1))


plot(x$x_si, x$y_c, ylab="C", xlab="Si", col='gray')

points(xhat, yhat, col='red', type='l', lwd=3)

points(smooth.spline(x$y_c~x$x_si, spar = 0.3), type='l', lty=3, col='blue', lwd=5)
```

```
#RandomForest test for O
bigmodel_y_o <- randomForest(y_o~x_mo+x_si+x_sn+x_zr+x_mn+x_cr+x_fe+x_ni+x_cu+x_n,
data=x, importance=TRUE)
importance(bigmodel_y_o)

# the most important random forest predictor
smallmodel_y_o <- randomForest(y_o~x_si, data=x, importance=TRUE)
yhat <- predict(smallmodel_y_o)
xhat <- x$x_si[order(x$x_si)]
yhat <- yhat[order(x$x_si)]
resid_o <- x$y_o-predict(bigmodel_y_o)

PATH="C:/"

### Principal components on residuals of random forest
resid_matrix <- cbind(resid_ti, resid_al, resid_v, resid_c, resid_o)
pca <- princomp(resid_matrix)
pca$loadings
biplot1<-biplot(pca)


# start file
par(mar=c(4,4,0.1,0.1))
```

```
plot(x$x_si, x$y_o, ylab="Oxigen", xlab="Silice", col='gray')

points(xhat, yhat, col='red', type='l', lwd=3)

points(smooth.spline(x$y_o~x$x_si, spar = 0.3), type='l', lty=3, col='blue', lwd=5)
```