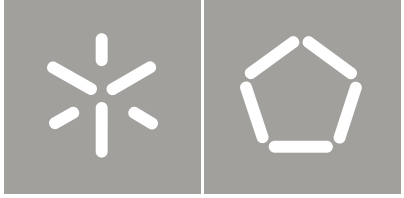
Universidade do Minho
Escola de Engenharia

Joaquim Agostinho Barbosa Tinoco

Application of Data Mining Techniques to Jet Grouting Columns Design

November 2012



Universidade do Minho
Escola de Engenharia

Joaquim Agostinho Barbosa Tinoco

Application of Data Mining Techniques to Jet Grouting Columns Design

Doctoral Thesis
Civil Engineering

Work performed under the supervision of
Professor António Gomes Correia

And co-supervision of
Professor Paulo Alexandre Ribeiro Cortez

November 2012

DECLARATION

Name:	Joaquim Agostinho Barbosa Tinoco
E-mail:	jabtinoco@civil.uminho.pt jabtinoco@hotmail.com
Phone number:	(+351) 916 518 646

Doctoral thesis title:	Application of Data Mining Techniques to Jet Grouting Columns Design
Supervisor:	Professor António Gomes Correia
Co-supervisor:	Professor Paulo Alexandre Ribeiro Cortez
Year of conclusion:	2012
Area of knowledge:	Civil Engineering

University of Minho, November 2012

Acknowledgements

The present research work was developed at the Environment and Construction Centre (C-TAC) of the University of Minho in Portugal under the supervision of Professor António Gomes Correia and Paulo Cortez. Accordingly, the author would like to thank all institutions and persons who made this project possible, namely:

- *Fundação Para a Ciência e a Tecnologia* (FCT), through the doctoral grant SFRH/BD/45781/2008;
- University of Minho through the Territory, Environment and Construction Centre (C-TAC) and the Department of Civil Engineering (DEC), for the logistical and human conditions;
- Prof. António Gomes Correia, for his encouragement, stimulation, supervision and insights into various aspects of this work. His broad view of geotechnics and critical spirit provided an important contribution to this research work;
- Paulo Cortez, for his explanations and support concerning the Data Mining field and insights into various aspects of this work, as well as of academic and scientific life;
- Tiago Valente, for the dataset gathered from the laboratory formulations that was a key element in this research work and for all help and friendship;
- Tecnasol-FGE company, particularly to engineer João Falcão, for the interest and for making available all the data concerning to jet grouting projects, without which this work would not have been possible;
- Sandra Coelho, for her patience and support during the database compilation process and for her advices and insights;
- Finally, I would like to thank to everyone that, directly or indirectly, have contributed to this work's development. Particularly, my greatest gratitude goes to my parents and friends for their constant support, belief and encouragement.

This page was intentionally left blank.

Abstract

Jet Grouting (*JG*) is a reference method on soil improvement technologies, allowing improvements to the strength, stiffness and permeability of soft soils. However, even after several years of practice and notable technological advances, there are still some limitations to overcome. In particular, the main limitation is related to the actual approaches for *JG*'s design. The actual approaches are scarce and have important applicability limitations, either in terms of jet systems or soil types. Indeed, the actual design approaches are often too conservative. As a result, the economy and quality of the soil improvement can be affected. Therefore, it is fundamental to develop new approaches that are able to accurately predict a *JG* column's mechanical properties as well as its diameter. However, due to the high number of variables involved in the *JG* process and the heterogeneity of the soils improved, the accomplishment of such complex tasks represents a major challenge. This challenge relies on the fact that a *JG* model design should be able to incorporate simultaneously the effects of different variables (e.g., soil and cement slurry properties and jet system).

Thus far, the traditional statistical approaches were unable to address the complexity of *JG* data. However, in the past few years, powerful tools have emerged for extracting useful information from large and complex data sets. These tools are currently known as *Data Mining* (*DM*) techniques and have been successfully applied in different application domains. In the present work, some of the most well-known *DM* algorithms were applied in the prediction of the mechanical properties of *JG* mixtures, as well as the respective column diameters. Therefore, as a first step, a multiple regression, artificial neural network, support vector machine and functional network algorithms were trained to predict the uniaxial compressive strength and stiffness of *JG* laboratory formulations. Moreover, the analytical expressions proposed by Eurocode 2 and CEB-FIP Model Code 1990 for strength and stiffness prediction of concrete were adapted to *JG* mixtures. After that, the same methodologies were applied in the prediction of the same properties of *JG* mixtures, as well as the diameter of the respective column.

As the main outcomes of this work, high-quality predictive models were achieved, as well as a better understanding of the *JG* mixtures' behaviour (given by a global sensitivity analysis). Such results are quite useful for *JG* design, which can expect economic and technical improvements through better optimisation of the available resources.

Keywords: Soft soils, soil improvement, jet grouting, artificial intelligence, data mining, support vector machines, artificial neural networks, functional networks, sensitivity analysis.

This page was intentionally left blank.

Sumário

Jet Grouting (*JG*) surge atualmente como um método de referência entre as tecnologias de melhoramento de solos, permitindo o aumento da resistência e rigidez bem como a diminuição da permeabilidade de solos moles. No entanto, mesmo após vários anos de prática e de notáveis avanços tecnológicos, existem ainda algumas limitações a vencer. Uma das mais relevantes prende-se com as actuais abordagens de dimensionamento, as quais são escassas e com importantes limitações de aplicabilidade, quer em termos de tipo de jet ou tipo de solo. De facto, as atuais abordagens de cálculo são por vezes demasiado conservativas, condicionando assim a eficiência técnica e económica do melhoramento. Neste sentido, é fundamental desenvolver novas abordagens capazes de prever com maior precisão as propriedades mecânicas do material *JG* e respectivo diâmetro das colunas. Contudo, devido ao elevado número de variáveis envolvidas e à heterogeneidade dos solos tratados, tal tarefa representa um enorme desafio. Este desafio prende-se com o facto de um modelo de dimensionamento da tecnologia de *JG* dever ser capaz de incorporar simultaneamente o efeito de diferente variáveis (propriedades do solo e da calda injetada, tipo de jet, etc.).

Até aos dias de hoje, as ferramentas estatísticas tradicionais foram incapazes de lidar com a complexidade caracteristica de dados de *JG*. No entanto, nos últimos anos têm emergido ferramentas com enorme potencial, capazes de analisar e extrair informação útil de grandes volumes de dados complexos. Estas ferramentas são correntemente conhecidas como técnicas de *Data Mining* (*DM*) e têm sido aplicadas com sucesso em diferentes áreas do conhecimento. No presente trabalho de investigação, alguns dos mais conhecidos algoritmos de *DM* foram aplicados na previsão das propriedades mecânicas de misturas de *JG* bem como na previsão do diâmetro das respetivas colunas. Assim, numa primeira fase, os algoritmos de regressão múltipla, redes neuronais artificiais, máquina de vetores de suporte e redes funcionais foram treinados para prever a resistência à compressão uniaxial e o módulo de deformabilidade de formulações laboratoriais de *JG*. Além disso, as expressões analíticas propostas pelo Eurocódigo 2 e pelo CEB-FIP Model Code 1990 usadas na previsão destas propriedades do betão, foram também adaptadas a misturas de *JG*. Posteriormente, as mesmas metodologias foram aplicadas na previsão da resistência e módulo de deformabilidade de misturas de *JG*, bem como do diâmetro das respectivas colunas.

Como principais contribuições do presente trabalho, destaca-se a elevada qualidade previsional dos modelos obtidos, bem como uma melhor compreensão do comportamento de misturas de *JG* (conseguida através da aplicação de análises de sensibilidade globais). Estes resultados são um claro contributo para o dimensionamento de colunas de *JG*,

antevendo-se uma maior eficiência técnica e económica, através de uma melhor otimização dos recursos disponíveis.

Palavras-chave: Solos moles, tratamento de solos, jet grouting, inteligência artificial, mineração de dados, máquina de vetores de suporte, redes neuronais artificiais, redes funcionais, analises de sensibilidade.

Résumé

Jet Grouting (*JG*) se pose actuellement comme une méthode de référence entre les technologies d'amélioration des sols, en permettant l'augmentation de la résistance et de la rigidité et également la diminution de la perméabilité des sols mous. Cependant, même après des années de pratique et de notables avancées technologiques, il existe encore quelques limitations à vaincre. Une des plus pertinentes concerne les approches de dimensionnement, lesquelles sont limitées dans le domaine d'application, notamment pour la prise en compte des différents types de *JG* et de sols. En effet, les approches actuelles de calcul sont essentiellement supportées par des méthodes empiriques et parfois même trop conservatives. Par conséquence, l'efficacité technique et économique du traitement peut être compromise. Il est donc fondamental de développer des nouvelles approches, plus précis et capable de prévoir les propriétés mécaniques du matériau *JG*, ainsi que les diamètres des respectives colonnes. Cependant, dû aux nombres élevés des variables impliquées et à l'hétérogénéité des sols traités, cette tâche est un énorme défi. Ce défis réside dans le fait qu'un modèle de dimensionnement de la technologie de *JG* doit être capable d'incorporer simultanément l'effet des différents variables (propriétés du sol, propriétés du coulis, type de *JG*, entre autres).

Jusqu'á aujourd'hui, les outils statistiques traditionnels n'étaient pas en mesure de faire face la complexité des données caractéristique du *JG*. Cependant, dans les dernières années des outils avec énorme potentiel ont émergés, capable d'analyser et d'extraire de l'information utile de grands volumes de données complexes. Ces outils sont habituellement connus comme techniques de *Data Mining* (*DM*) et sont appliqués avec succès dans différents domaines de la connaissance. Dans ce travail de recherche quelques un des plus connus algorithmes de *DM* ont été appliqués à la prévision des propriétés mécaniques de mélanges de *JG* comme dans la prévision du diamètre des respectives colonnes. Ainsi, dans une première étape, les algorithmes de régression multiples, réseaux neuronales artificielles, machine à vecteurs de support et réseaux fonctionnels ont été formés pour prévoir la résistance à la compression unidimensionnelle et le module de déformabilité des formulations de laboratoire de *JG*. En outre, les expressions analytiques proposés par l'Eurocode 2 et par le CEB-FIP Model Code 1990, utilisées dans la prévision de ces propriétés pour le béton, ont été adaptées aux mélanges de *JG*. Ensuite, les mêmes méthodologies ont été aussi appliquées pour les matériaux des vraies colonnes de *JG*, ainsi que pour la prévision du respectif diamètre.

Comme principaux contributions de ce travail on peut soulever l'haute qualité des modèles de prévision et une meilleure connaissance du comportement des mélanges de *JG* (donné par une analyse de sensibilité globale). Ces résultats sont très utiles pour le

dimensionnement du *JG* avec des importants avantages économiques et techniques au moyen d'une meilleure optimisation des ressources disponibles.

Mots-clés : Sols mous, l'amélioration des sols, jet grouting, l'intelligence artificielle, data mining, machine à vecteurs de support, réseaux neuronales artificielles et réseaux fonctionnels, analyses de sensibilité.

Contents

This page was intentionally left blank.

List of Figures

This page was intentionally left blank.

List of Tables

Acronyms

D	*jet grouting column diameter.*
E_0	*elastic Young's modulus.*
E_{max}	*maximum secant deformability modulus.*
$E_{sec50\%}$	*secant deformability modulus at 50% of the maximum applied stress.*
$E_{tg50\%}$	*tangent deformability modulus at 50% of the maximum applied stress.*
UCS	*Uniaxial Compressive Strength.*
AI	*Artificial Intelligence.*
ANN	*Artificial Neural Network.*
$CRISP\text{-}DM$	*Cross-Industry Standard Process for Data Mining.*
DM	*Data Mining.*
$EC2$	*Eurocode 2.*
FN	*Functional Network.*
FS	*Feature Selection.*
$GAMS$	*General Algebraic Modelling System.*
GSA	*Global Sensitivity Analysis.*
$JGLF$	*Jet Grouting Laboratory Formulations.*
JG	*Jet Grouting.*
KDD	*Knowledge Discovery from Databases.*
MAD	*Mean Absolute Deviation.*
$MC90$	*Model Code 1990.*
MR	*Multiple Regression.*
REC	*Regression Error Characteristic.*
$RMSE$	*Root Mean Square Error.*
SA	*Sensitivity Analysis.*
$SEMMA$	*Sample, Explore, Modify, Model, and Assess.*
SVM	*Support Vector Machine.*
UD	*Uniform Design.*
VEC	*Variable Effect Characteristic.*

This page was intentionally left blank.

Chapter 1

Introduction

1.1 Motivation

Currently, due to strong urbanisation and industrialisation, any piece of soil may be required for construction purposes, even soft soils usually characterised by high porosity, plasticity, compressibility and low strength (Liu et al., 2008). Good examples of this situation are harbour areas, where there is an increasing need for reclaimed land (Van Impe et al., 2005). Unfortunately, the soil foundation at such places does not always have the appropriate characteristics for construction purposes. Some situations[1] arise in which some undesirable behaviour of a soil foundation needs rectifying with minimal impact on neighbouring construction. In these situations, the solution considers the improvement of the mechanical and physical properties of the soil to increase its strength and stiffness and to decrease its permeability. Moreover, the soil improvement method should respect the growing concerns about environment issues. This consideration means that, for instance, the in situ soil should be reused instead of being replaced by another one with better proprieties.

To satisfy such needs, several soil improvement methods have been developed in recent decades. In this field, *Jet Grouting* (*JG*) technology plays an important role as one of the most used soft soil improvement methods worldwide. This technology has been applied in different situations, such as ground water control, settlement or excavation control and tunnelling support. Important advances have also been observed in injection systems, improving energy efficiency and increasing the area treated. However, despite being widely applied worldwide, namely in important geotechnical projects, the existing methods for *JG* technology design are scarce and have important limitations in terms of jet systems and soil types, namely for *soilcrete's*[2] mechanical properties and column diameter

[1] For example due to changes in functionality of the building.

[2] *Soilcrete* – practical designation for soil-cement mixture resulting from *JG* technology.

prediction. Indeed, even in large-scale works, *JG* design is essentially based on empirical methods and strongly supported by *JG* companies' experience. As a result, considering the subjectivity of such approaches and the conservative values of safety factors used in empirical design methods, the economic and technical efficiency of the soil improvement can be compromised. Therefore, keeping in mind the high versatility of *JG* technology and its role in important geotechnical works, there is a need to develop new approaches for accurately predicting *JG* column diameter and *Soilcrete* mechanical properties.

One of the main reasons for the scarcity of *JG* columns design (with a considerable applicability in terms of jet systems and soil types) is related to the high number of variables involved in the entire construction process and to the heterogeneity of the treated soils. Furthermore, there are also situations in which, due to budget limitations, the available information (e.g., soil characterisation or test columns) to feed the empirical approaches is limited. On the other hand, particularly in large-scale *JG* works, much information has been produced that could be used in future projects after being properly analysed and interpreted. Therefore, the question that arises is how to explore all of the available information related to past *JG* projects to support decisions in the preliminary stages of future designs, mainly in small-scale *JG* works where information is scarce, while keeping in mind the high dimensionality and nonlinearity of the problem.

An interesting solution can be the use of *Artificial Intelligence* (*AI*) tools that has shown successful results in different knowledge domains (Liao et al., 2012), namely in Geotechnics field (Miranda et al., 2011; Goh and Goh, 2007; Narendra et al., 2006). Indeed, the application of *Data Mining* (*DM*) techniques to data gathered from large geotechnical works can provide a strong framework to the development of models that can be very useful in future projects. These tools are supported in the idea that there are usefully information behind the data, aiming the extraction of patterns and rules from the data through specific algorithms. The main advantage of *DM* techniques over traditional statistical analysis is in its ability/superiority to deal with big amount of data, characterized by high-dimensionality and complexity. Furthermore, the developed models based on these tools can be easily updated when new data are available.

1.2 Scope of the work

The main goal of this research is to develop a new reliable approach for *JG* design that predicts the *Uniaxial Compressive Strength* (*UCS*) and stiffness of both laboratory and field mixtures. Moreover, we also intend to develop analytical models for real *JG* column diameter predictions. The proposed methods aim to overcome some of the most relevant

limitations of the current approaches, namely in terms of jet systems and soil types. This need for such models arises from the high versatility and potential of *JG* technology as a soft soil improvement method and from the increasing number of *JG* projects around the world, namely in Portugal (ICOG, 2012).

Due to the high number of variables involved in soft soil improvement by *JG* technology and the complex relationships between *soilcrete's* mechanical properties and their contributing factors, the development of the *JG* design approaches proposed in this research was supported on new and powerful tools currently known as *DM* techniques. The use of these tools, apart from their high learning capabilities, also gives an important reliability to the models when compared to those supported on traditional statistical analysis methods.

A better understanding of *JG* mixtures' behaviour supported by a detailed *Sensitivity Analysis* (*SA*) will certainly contribute to an improvement of the *JG* technology efficiency and will lead to technical and economic benefits. Therefore, a *Global Sensitivity Analysis* (*GSA*) was applied over each one of the proposed models for the studied properties (strength, stiffness and diameter) of both laboratory and field mixtures. These analyses allowed identification of the key variable on each studied property as well as its average effect on the output variable.

1.3 Outline of the thesis

This thesis intends to highlight the benefits resulting from the implementation of *AI* tools to solve complex geotechnical problems, particularly *JG* technology design, and for this purpose, it is divided into seven chapters and two appendices, organised as schematised in Figure 1.1. In this section, the content of each chapter is described in detail.

- Chapter 1, entitled **Introduction**, describes the initial considerations and motivations of the thesis. It also presents the description of the performed work in each chapter.

- Chapter 2, entitled **Artificial intelligence tools**, presents a global overview of *AI* tools. Here, the reader can find the main concepts behind knowledge discovery database processes. It also notes all *DM* algorithms and methodologies implemented in the present research, namely the applied approaches for feature selection and model selection, as well as for model assessment and interpretability. Furthermore, the main aspects related to the software used in the performed experiments are also introduced.

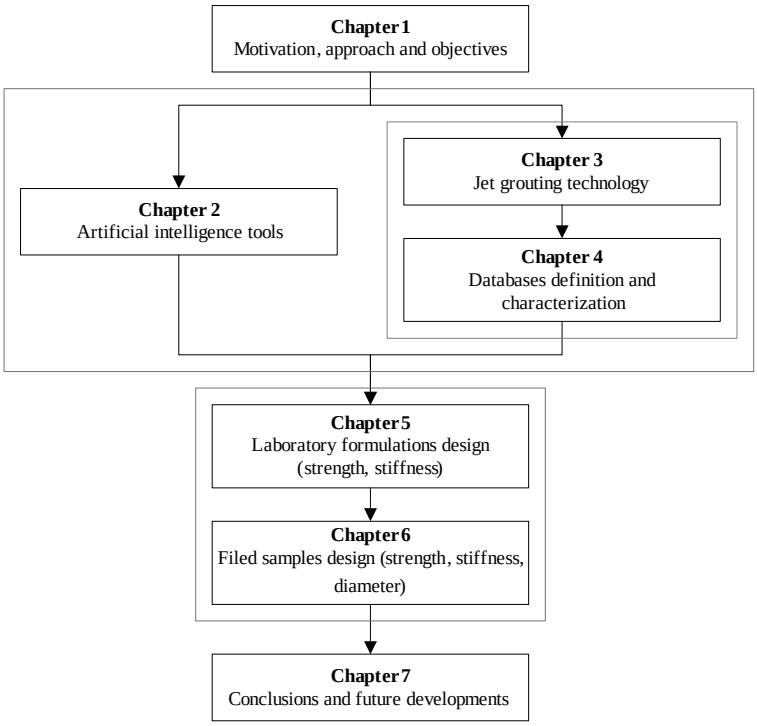


Figure 1.1: Outline and organisation of the thesis

– Chapter 3, entitled **Jet grouting technology**, starts by addressing the main aspects of *JG* technology, highlighting the importance of this soft soil improvement method, as well as the complexity of its design. Then, a description of the main equipment and construction process, as well as the different *JG* systems, is given. After that, the main approaches currently used for the mechanical properties of *JG* mixtures and column diameter design are summarised.

– Chapter 4, entitled **Jet grouting database characterisation**, enumerates the information sources for each database (laboratory and field) used in this research. It also highlights the methodology followed during the database compilation process. The input variables considered in both laboratory and field studies are also enumerated and presented as a correlation matrix for each studied property that shows the relationship level between all input and output variables.

– Chapter 5, entitled ***DM* techniques applied to laboratory data**, presents the main results of the application of *DM* techniques toward the development of analytical models for UCS and stiffness prediction of *Jet Grouting Laboratory Formulations* (*JGLF*). The high learning capabilities of *DM* techniques, particularly the *Support Vector Machine* (*SVM*) algorithm, are highlighted and compared with *Eurocode 2* (*EC2*) and *Model Code 1990* (*MC90*) approaches currently used for

concrete strength and stiffness predictions. Moreover, the results of the application of a *GSA* are presented and discussed, emphasising the key variables on mechanical behaviour of *JGLF*, as well as its average effect on the target variable.

– CHAPTER 6, entitled ***DM* techniques applied to field data**, describes the data-driven predictive models for mechanical properties of *JG* field mixtures collected directly from real *JG* columns, as well as for their diameters. Moreover, the key variables in strength, stiffness and diameter prediction of real *JG* columns, as well as their average effect on the target variables, are enumerated. It also presents a relationship between the strength of laboratory and field samples and a correlation between strength and stiffness of *soilcrete* mixtures. At the end, a proposal for *JG* column diameter design is presented.

– CHAPTER 7, entitled **Main summary**, summarises the main important conclusion of the present work, pointing out some advice for a better economic and technical efficiency of soft soil improvement performed by *JG* technology. It also presents some research possibilities for future developments.

– **Appendix A** summarises the main statistics and histograms of all input and output variables considered in the present research for both laboratory and field studies.

– **Appendix B** details the mathematical expressions applied to calculate some input variables used during the entire study.

This page was intentionally left blank.

Chapter 2

Artificial intelligence tools

2.1 Background

In the middle of 50's a new branch of computer science started to attract the attention of specialists in this field. This new branch, termed as *AI*, can be defined as the study of how to make computers do things at which, at the moment, people are better (Rich, 1983). In the begin, the goal was to develop a computer that could mimic human behaviour. In the 70s, *AI* was more focused on developing expert systems that would acquire knowledge from experts and support decision making. Later, in the 90s, there was a shift in *AI* to learn useful knowledge directly from the data. Currently, *AI* encompasses several methods and solutions. For demonstration purposes, Figure 2.1 shows an historic view of some the main *AI* methods.

With the boom of Information Technology, the generation and collection of data grew rapidly. At the current stage, vast datasets are becoming commonplace. All this data hold valuable knowledge (e.g. trends, patterns) that can be used to support decision making and optimize success.

Classical statistics may fail to analyse vast amounts of data and/or when there are complex relationships between the data variables. Also, the number of experts is limited and they may overlook important details. Hence, to overcome these limitations, it is desirable to have more automated processes for data analysis, based on computers.

Given the interest in (semi-)automatic knowledge extraction from data, in the last few decades there has been an increase in a new research area that intersects several scientific domains, such as Artificial Intelligence, Statistics and Information Systems. Formally, this area was defined as *Knowledge Discovery from Databases* (*KDD*) (Fayyad et al., 1996b) but through the years the term Data Mining become more popular. As such, in this thesis, *DM* terminology will be often used as a synonym of *KDD*.

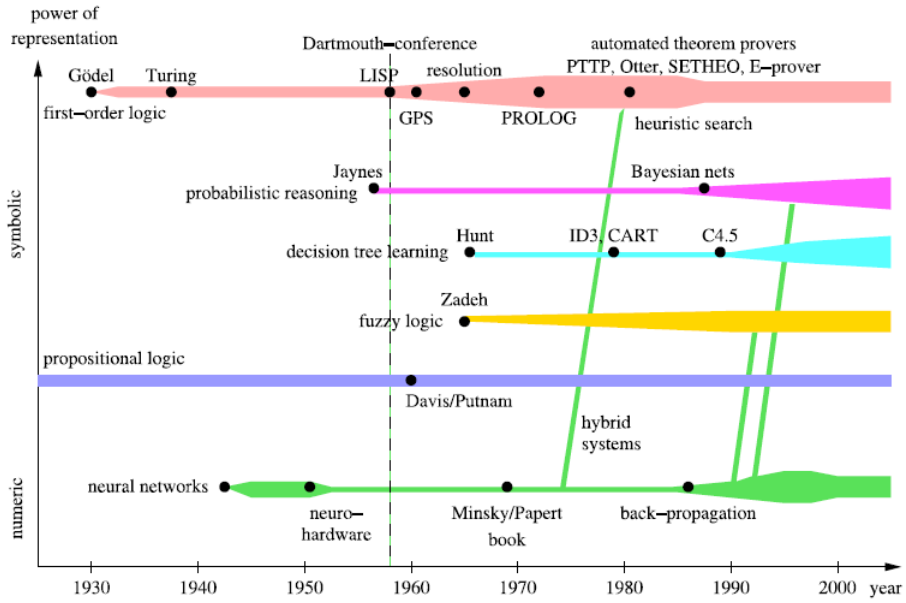


Figure 2.1: History of the various *AI* areas. The width of the bars indicates prevalence of the methods use (Ertel, 2009)

DM is receiving widespread attention in the academic and public literature. Many case studies suggest that companies are increasingly investigating the potential of *DM* technology to deliver competitive advantage. Nowadays, there are inclusively many successfully applications of *DM* techniques in different knowledges domains. For instance, these techniques are widely used in business fields, such as direct target marketing campaigns, fraud detection, and development of models to aid in financial predictions (Miranda, 2007). Liao et al. (2012) carried out a deep literature review, showing the developments and applications of *DM* techniques during the past decade. The survey focussed on the period from 2000 to 2011, having found 216 articles concerning to *DM* techniques applications on different research and practical domains of knowledge. Additionally, Liao and his collaborators presented some perspectives about expected future developments in *DM* techniques, methodologies and applications. In particular, Liao et al. (2012) present important application in the civil engineering domain, namely:

- Lai and Serra (1997) applied *Artificial Neural Network*s (*ANN*s) to predict compressive strength of cement conglomerates;
- Prasad et al. (2009) propose an *ANN* to predict a 28-day compressive strength of a normal and high strength self compacting concrete and high performance concrete with high volume fly ash;
- Chou et al. (2011) aimed to optimize the prediction accuracy of the compressive strength of high-performance concrete by comparing data-mining methods;

Within the more specific Geotechnical field, there are also some relevant studies where *DM* tools were applied to solve different geotechnical problems:

- Miranda et al. (2011) proposed new alternative regression models using *ANN*s for the analytical calculation of strength and deformability parameters of rock masses;
- Goh and Goh (2007) used *SVM*s to assess seismic liquefaction;
- Erzin (2007) studied the relationship between the swell pressure and soil suction behaviour in specimens of Bentonite-Kaolinite clay mixtures with varying soil properties using *ANN*s;
- Narendra et al. (2006) applied computational intelligence techniques for *UCS* prediction of soft grounds.

As further highlighted in Chapter 3, geometric and mechanical properties design of *JG* mixtures is a complex task involving a high number of variables that have shown nonlinear relationships between input and output variables. Hence, the use of *DM* tools can be seen as a interesting alternative to the development of more reliable and accurate methods for *JG* design.

For a reliable design of any *JG* work, the first step is to carry out a soil characterization as detailed as possible. Unfortunately, such characterization is scarcely performed due to schedule and budget limitations. In addition, for quality control purposes, some test columns should be built near to the improvement spot, from which some samples are extracted and tested at different ages. Once again, and particularly in small *JG* works, these test columns consist of a very reduced number due to the inherent costs with materials and the time demands. On the other hand, particularly in important and big scale *JG* works, is usual to perform a detailed soil characterization and built several test columns, from which a significant number of samples are collected and tested. This scenario leads to the following question: *Is there a way to optimize this useful information?* That is, how can the information produced, particularly in the big scale *JG* works, be efficiently used in new *JG* works, particularly in the smallest ones. It is precisely here that the application of automated process for data analysis, such as *DM* techniques, can give a valuable contribution, helping to overcome the limitations of the actual *JG* approaches design, namely in terms of jet systems and soil types.

A simple compilation of all available data related with *JG* works in an adequate structure could be seen as a first step to help the development of new and more accurate methodologies for *JG* design. Database theories and tools provide the necessary infrastructure to store, access, and manipulate data. Data warehousing, a recently popularized

term, refers to the current business trend of collecting and cleaning transactional data to make them available for online analysis and decision support. A popular approach for analysis of data warehouses is called online analytical processing (OLAP). However, a simple organization and manipulation of the data is not enough when dealing with *JG* data, because of its high nonlinear characteristics/complexity and dimensionality. In such situations, there is a potential contribution that can be provided by *DM* techniques, since these can be used to explore all this data and automatically extract valuable rules and patterns. The obtained knowledge/models can be further applied in the project stage, helping the definition of the parameters for *JG* columns construction, as well as during the soil improvement (in real time), advising eventual adjustments to overcome unexpected conditions.

2.2 Knowledge discovery in databases

The main purpose of *AI* domain is to develop machines that mimic real persons, thus showing intelligent behaviour (Ertel, 2009). To achieve this goal, there is one essential ability: to learn from experience. Here is where *KDD* plays a key role.

KDD process can be defined as *the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* (Fayyad et al., 1996b). The term *nontrivial* means that some search or inference is involved. This means that *KDD* process is not a straightforward computation of predefined quantities, such as computing the average value of a set of numbers. *KDD* is interactive and iterative, involving numerous steps with many decisions made by the user.

The first *KDD* works were motivated by fields concerned with inferring models from data, including statistical pattern recognition, applied statistics, machine learning, databases, visualization, and neural networks (see Figure 2.2). *KDD* largely relies on methods from these fields to find patterns from data in the *DM* step of the *KDD* process. According to Fayyad et al. (1996a) a large degree of the current interest in *KDD* is the result of the media interest surrounding successful *KDD* applications. For example, the focus articles in Business Week, Newsweek, Byte, PC Week, and other large-circulation periodicals. In science, one of the primary application areas is astronomy. *KDD* focuses on the overall process of knowledge discovery from data, including how the data is stored and accessed, how algorithms can be scaled to massive datasets and still run efficiently, how results can be interpreted and visualized, and how the overall human-machine interaction can be modelled and supported. *KDD* places a special emphasis on finding understandable patterns that can be interpreted as useful or interesting knowledge. Scal-

ing and robustness properties of modelling algorithms for large noisy datasets are also of fundamental interest. Statistics has much in common with *KDD*.

A *DM* model and the resulting knowledge should satisfy some important requests such as:

- be valid when applied to new data;
- bring something new (at least to the system and preferably to the user);
- be useful to the knowledge domain or user;
- be understandable.

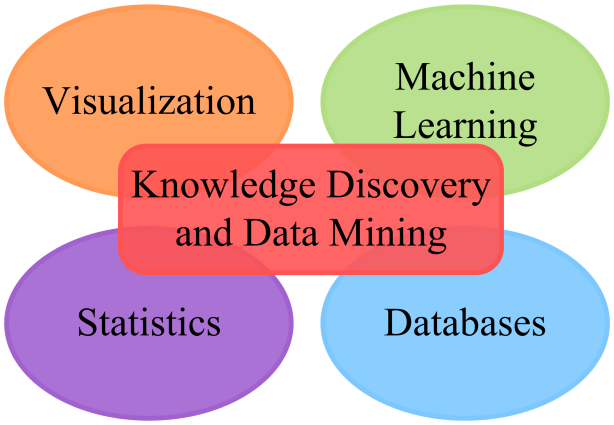


Figure 2.2: Scientific fields related with *KDD* and *DM*

According to Ertel (2009) the processing of knowledge follows the structure shown in Figure 2.3. According to this schema the knowledge is stored in a knowledge base. This knowledge is provided by those who are called *Knowledge Engineering* that is supported on several knowledge sources such as humans experts, the knowledge engineer and databases. It is also possible obtain knowledge through an active exploration of the world. Here, the agent learns from a database and from the interaction with the world. The knowledge stored in the *knowledge base* is processed allowing the final user to apply such knowledge.

An important notion, called *interestingness*, is usually taken as an overall measure of model value, combining validity, novelty, usefulness, and simplicity. In order to achieve such request, the process of developing a data-driven model should evolve several steps. Figure 2.4 depicts the main steps of the interactive and iterative (with many decisions made by the user) *KDD* process, which are following summarized:

- Selection: based on problem domain, a target dataset with the relevant information is compiled from the database, on which the discovery will be performed;

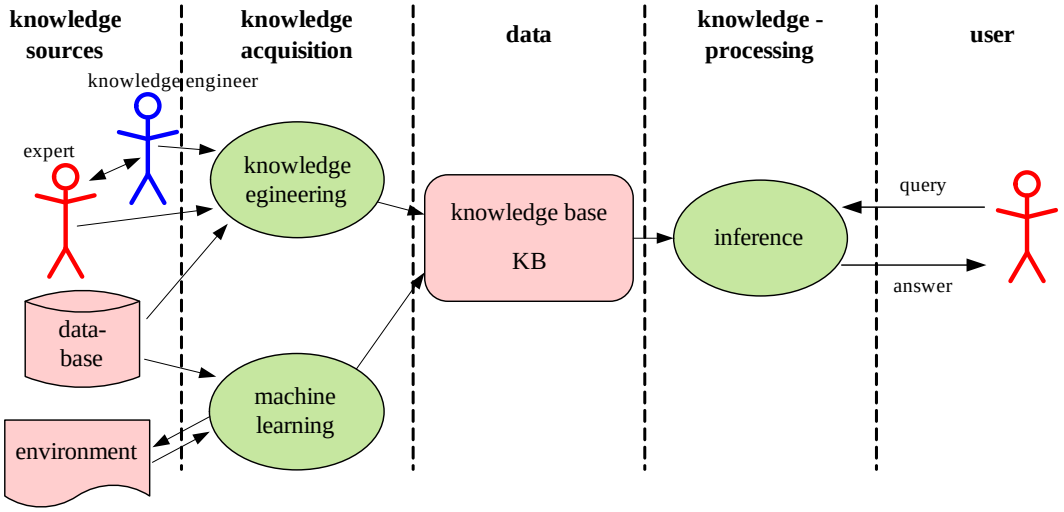


Figure 2.3: Structure of a classic knowledge-processing system (Ertel, 2009)

- Pre-processing: this stage consists on the target data cleaning (outliers and noise removal), handling missing data and other pre-processing in order to obtain consistent data;
- Transformation: data are transformed using dimensionality reduction or transformation methods in order to present the adequate form for *DM* stage;
- Data Mining: application of *DM* algorithms for searching for patterns of interest;
- Interpretation: this stage consists on the interpretation and evaluation of the mined patterns, in order to obtain understandable and useful knowledge.

The *KDD* process should start with an understanding of the problem domain and the collection/compilation of all available and interesting information in a database. After that, a subset of the main database, only with the relevant attributes is extracted. For this step it is very useful a multidisciplinary team of specialists, which are fundamental to support the variable selection task. It is also at this stage that the main goals of study are established. The target dataset is carefully and rigorously analysed and important operations are made. Tasks related with removing noise or outliers are performed in order to improve the data quality. In addition, decisions about how to handle missing values are taken. The main approaches for dealing with missing values can be classified into four categories (Brown and Kros, 2003):

- use of complete data only;
- deleting cases or attributes with missing data;

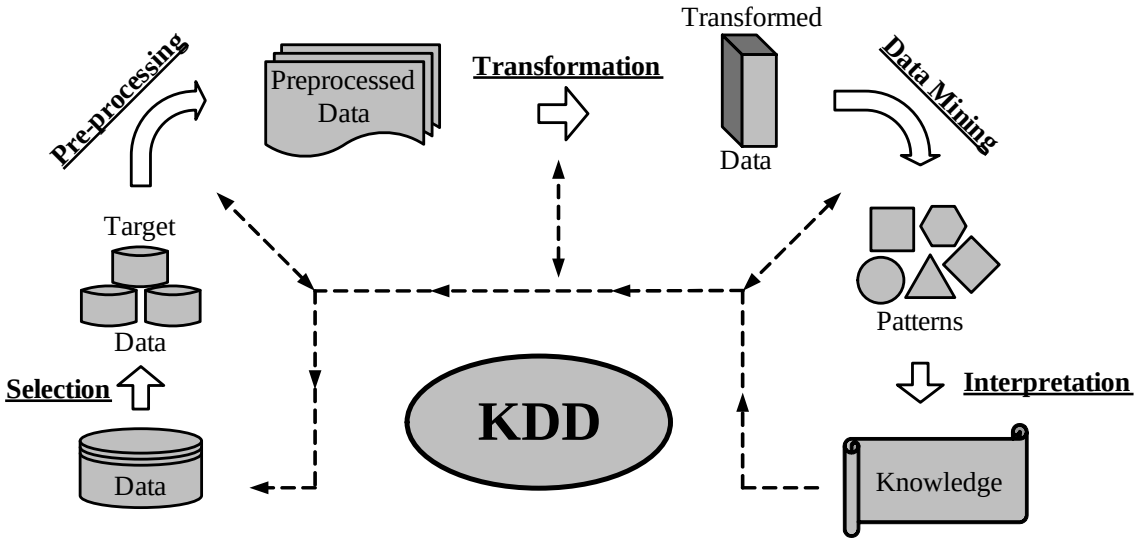


Figure 2.4: Steps of *KDD* process (adapted from Fayyad et al. (1996b))

- data imputation; and
- model-based approaches.

The first two approaches are very simple and direct but are only best suited for situations where the amount of missing data is scarce. An imputation method replaces missing data by estimated values, under distinct approaches, such as:

- Case substitution: use domain experts to replace missing values;
- Mean substitution: use the mean value of the data variable;
- Cold-deck imputation: use values from other sources of data;
- Hot-deck imputation: missing values are replaced with values drawn from the most similar case. The hard part of the application of this method is the difficulty in defining what is "similar". The conceptual simplicity maintenance and proper measurement level of variables are its main advantages;
- Regression imputation: regression analysis is used to predict missing values based on the variable's relationship to other variables in the data set. Single and/or multiple regression can be used to impute missing values. The first step consists of identifying the independent variables and the dependent variables. In turn, the dependent variable is regressed on the independent variables. The resulting regression equation is then used to predict the missing values. An advantage of this method is that it preserves the variance and covariance structures of variables with missing data;

- Multiple imputation: combination of a number of imputation methods into a single procedure.

The main purpose of the transformation step is to transform the data in order to take the correct form, to apply the different *DM* algorithms available. For instance, it may be advantageous to normalize the inputs and/or outputs to a zero mean and a one standard deviation.

After these steps, the *DM* algorithms are applied to the data. Section 2.3 focuses on this particular *KDD* step. As the last *KDD* step, the obtained patterns are interpreted and analysed in order to obtain knowledge, possibly using visualizations tools and applying *SA* procedures or removing redundant patterns.

It can be necessary to return to any of the previous steps in an iterative procedure of correcting options and errors in order to improve the final results. The understandable knowledge can be used in a decision support process or be incorporated in other intelligent systems, such as the expert or knowledge based systems. Furthermore, the new knowledge is checked by domain experts, in order to find possible conflicts, stressing the importance of the user in the *KDD* process.

The previously described *KDD* steps are strongly connected. For instance, the procedures applied on "Transformation" step are conditioned by the *DM* algorithms chosen on *DM* step. Furthermore, it should be mentioned that the quality of the results is dependent of a good interaction between all *KDD* steps and the user, and should not bet viewed as independent steps.

2.3 Data mining

This section focuses on the main issues related with *DM* stage, namely *DM* tasks, methodologies and algorithms.

In this *KDD* step, a *DM* algorithm is fitted to the dataset used during the learning phase, leading to a data-driven model. Such model can be described as the relationship between the inputs and the output, which can represent useful knowledge. Depending on the type of patterns that can be found, *DM* tasks are normally classified into two categories: **predictive** and **descriptive**. Predictive tasks perform inference on data in order to predict unknown values of the output variable, given known values of the input variables. Descriptive tasks try to characterize and summarize the general properties of the data in order improve its understanding. Figure 2.5 summarizes the main *DM* tasks currently used.

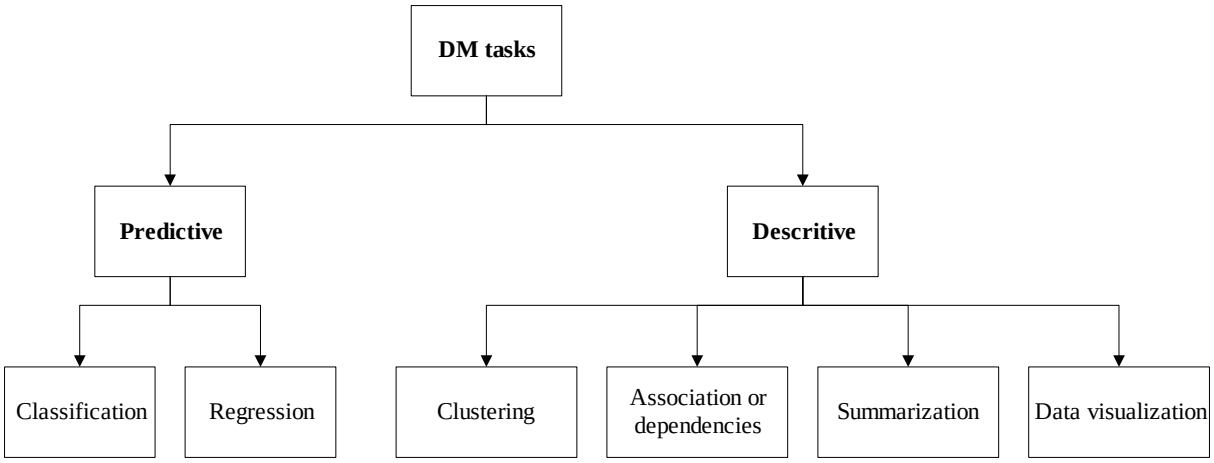


Figure 2.5: Hierarchy of the main *DM* tasks

2.3.1 DM tasks

For *DM* purposes there are several tasks that can be applied. This section briefly presents the most popular *DM* tasks (Figure 2.5).

Classification is one of most used *DM* tasks and has the purpose to find a model that classifies an example into a class within a predefined set of classes. The trained model should be able to correctly classify a new example based on its attributes. The model used to carry out such classification is normally built using a set of labelled examples (supervised learning). Some of the most used *DM* algorithms in classification tasks are decision tree, neural networks and support vector machines. The performance of a classification model is normally accessed by classification metrics, such as the percentage of correct predictions.

Regression is a *DM* task very similar to classification, sometimes also termed prediction, used to estimate unknown values of the dependent variable based on a set of independent variables. The main difference between classification and regression is related with the output variable. In classification the output is discrete while in regression the target is continuous. Classification can be considered as a particular case of regression. For model accuracy, distinct metrics are used (when compared to classification). Examples of such metrics are: mean absolute deviation and root mean squared error. The present study adopts a regression approach, where mechanical and geometric properties of *JG* columns will be predicted based on a set of selected attributes.

Clustering consists of grouping similar objects into classes (clusters). In contrast with classification, there are no class labels and the clusters (or groups) are determined by an unsupervised learning from the data. Ideally, all objects of each group should be close to each other and the distance between groups should be as high as possible. Normally,

clustering is a *DM* task used in early analysis with the purpose of finding clusters in the data and then applying to each cluster the most adequate *DM* algorithm.

Association rules, which try to find a model that describes significant dependences between variables through the identification of groups of highly associated data. These dependencies can exist at two levels:

- Structural: the model presents locally dependent variables in a graphical way;
- Quantitative: the model specifies the strength of the dependencies using a numerical scale.

Summarization makes use of methods that find a succinct description for the dataset. The most sophisticate summarization methods involve rules, visualization techniques and functional relationships between variables. Summarization functions are often used in data exploratory analysis and automatic generation of reports. A very simple example of summarization could be a histogram or a statistical measure of a certain attribute of data.

Data visualization deals with displaying final or intermediate *DM* results through a visual way. Its purpose is to describe complex relationships in a easily understandable way, normally through graphics or other visual representations. Visualizing the results in different forms together with interestingness measures can be very usefully to enhance comprehension of the domain, selection of the patterns which represent useful knowledge and provides guidelines for further discovery..

2.3.2 DM methodologies

Due to the increased interest in the field of *DM*, particularly due to the rising of vast databases in an increasing and differentiate number of organizations, there was a need to develop standard methodologies that can guide the implementation of *DM* projects. The main efforts were developed by academics and people in the industry field. Nowadays, the most two popular approaches are: *Cross-Industry Standard Process for Data Mining* (*CRISP-DM*) and *Sample, Explore, Modify, Model, and Assess* (*SEMMA*). These two approaches were developed in different environments, but with the same purpose, i.e., define a standard methodology to increase the success of the implementation of *DM* projects. The former methodology was developed by the means of the efforts of a consortium of companies from different activities: NCR, Daimler Chrysler AH, SPSS Inc. and OHRA (Chapman et al., 2000). The latter methodology was proposed by an organization that delivers services in the areas of *DM* and decision support.

CRISP-DM methodology

CRISP-DM methodology was developed at the end of the 90's and is supported by strong theoretical principles, as well as by the experience of those who develop *DM* projects. This methodology can be described by an hierarchical, iterative and interactive process, which sets six phases (Figure 2.6):

- Business understanding: identification and understanding of the project objectives and requirements from a business perspective. This knowledge is converted into a *DM* problem definition and a preliminary plan is proposed to achieve the goals;
- Data understanding: collection and analysis of the data in order to access its quality, discover first insights and detect subsets or trends. With this first data analysis, some hypotheses are formulated for hidden information;
- Data preparation: compilation of the final dataset that will be used during the learning phase (modelling) to build the *DM* model. Include the selection of the records and attributes from the initial raw data as well as its cleaning and transformation;
- Modelling: selection of the *DM* algorithms and optimization of its parameters in order to find patterns within the data;
- Evaluation: deep assessment of all fitted models and revision of all previous steps in order to verify if the business objectives were achieved;
- Deployment: organization of the obtained knowledge and its implementations in a way that the customer can use it.

SEMMA methodology

The *SEMMA* methodology, can be viewed as a guideline for a *DM* project, from its initial specification until its implementation, allowing an organized and adequate development and maintenance. This methodology is composed by a cycle with five main stages (see Figure 2.7) that start with the selection of the data and finish with the assessment of model obtained during the learning phases (Bulkley et al., 1999).

- Sample: selection of a representative sample from the studied universe, which should have an adequate dimension in order to optimize the costs, profitability and performance of the methodology. That is, the data sample extracted should contain the significant information and at the same time be easily manipulated;

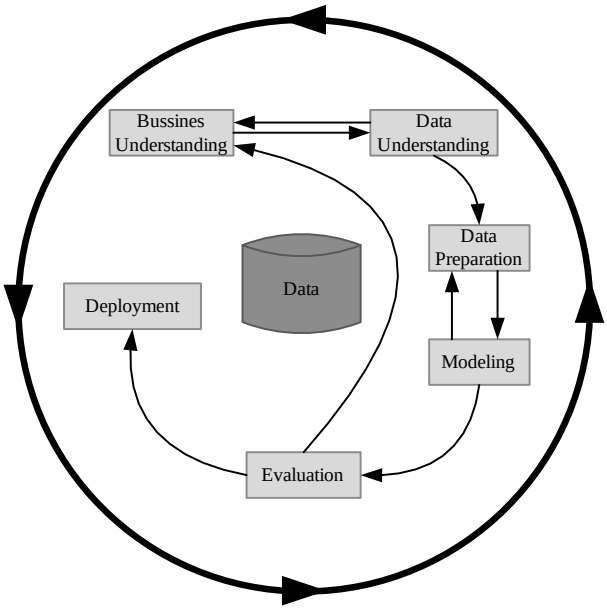


Figure 2.6: Phases of *CRISP-DM* methodology (adapted from Chapman et al. (2000))

- Explore: application of statistical and visual techniques to get an insight on the data, in order to identify tendencies and/or anomalies and gain understanding and ideas;

- Modify: based on the results of previous stages some transformations can be applied. For instance, new attributes can be included or modified;

- Model: after preparing and exploring the data, in this stage the appropriate *DM* algorithms are chosen and applied, in order to achieve the fitted models;

- Assess: finally, the obtained models are evaluated in order to infer about its performance, reliability and usefulness. For this purpose the model is applied to a new dataset (not used during the training phase) and its response is assessed.

When comparing *CRISP-DM* and *SEMMA* methodologies, we can conclude that they are very similar/equivalent and that there is a strong correlation with the five stages of *KDD* process (Figure 2.4). Azevedo and Santos (2008) establish an correspondence between these two *DM* methodologies and the *KDD* process (Section 2.2). The five stages of *SEMMA* methodology can be seen as a practical implementation of the five stages of *KDD* process. On the other hand, on the *CRISP-DM* methodology the first and the last stages can represent a *pre KDD* and *post KDD* stages respectively, while *data understanding* stage can be identified as the combination of the *selection* and *pre-processing*. The remaining stages are directly correlated.

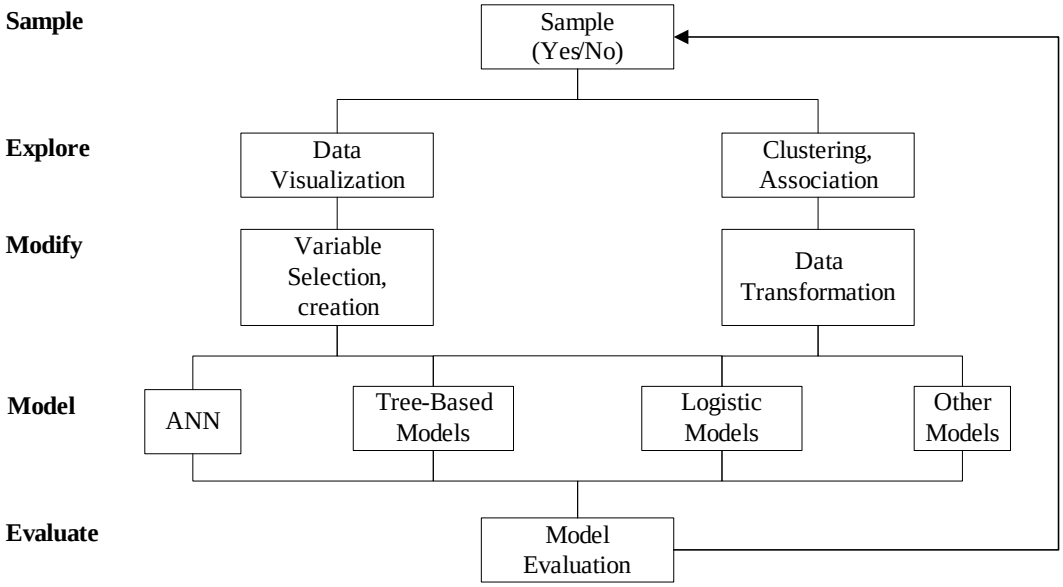


Figure 2.7: Phases of *SEMMA* methodology (Bulkley et al., 1999)

Given that the *CRISP-DM* methodology is more complete (and also neutral in terms of the *DM* tool explored), we adopt this methodology in this work.

2.3.3 DM algorithms

For each *DM* task (regression, classification, etc.) there are several algorithms that can be used, each one with its own advantages and limitations. Therefore, the first step is to choose the most suited algorithm to solve the problem at hands, viewed as skill of the analyst (Fayyad et al., 1996b). In this work, we explore four *DM* methods, which are described in the next subsections.

Multiple regression

Multiple Regression (*MR*) is a statistical technique used in different domains, ranging from engineering to social sciences. This linear approach is defined as:

$$\hat{y} = \beta_0 + \sum_{i=1}^{I} \beta_i \cdot x_i \tag{2.1}$$

where $\hat{y}$ is the predicted value, $x_1, ..., x_i$ are input variables and $\beta_0, \beta_1 ..., \beta_i$ are coefficients to be adjusted, normally using the least squares technique. Due to its additive nature, this model is easy to interpret and it is widely used in regression tasks. In the present research work, *MR* was essentially used as a baseline comparison.

Artificial neural networks

ANN is a computational technique inspired by nervous system structure of the human brain (Kenig et al., 2001). This technique has shown high performance in modelling complex nonlinear mappings and is robust when dealing with noisy data. It is particularly useful for problems that do not have an analytical formulation or where explicit knowledge does not exist. *ANN*s can be defined as a network of neurons connected in a simplified structure very similar to the neurons of living beings. This structure is able to learn with its own experience, store such knowledge and apply it to new examples not used during the learning process. This generalization capacity allows its application to solve complex problems, recognize patterns and predict future events.

Biologically, neurons are composed by a nucleus and are connected with millions of other neurons as schematically represent in Figure 2.8. They receive electrochemical inputs from their neighbours through connections called synapses. The synapses are formed by axons and dendrites. This simple structure allow perform three basic functions: input, processing and output of signals. Throughout the dendrite, the input signals reach the neuron, which process such information. The output signal flows throughout the axon, which is connected to other neurons throughout the synapses. Neurons form complex, nonlinear and highly parallel structures.

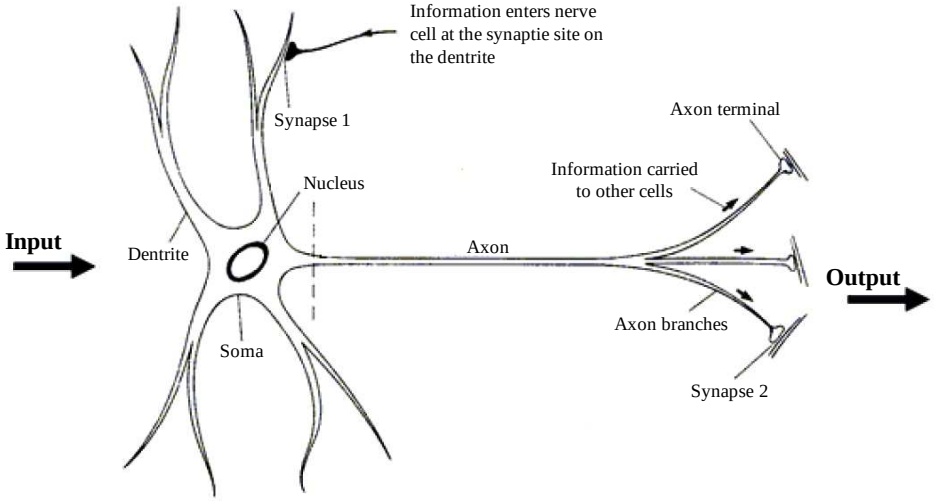


Figure 2.8: Schematic of the human neuron constitution

The first mathematical models of neural networks were designed by McCulloch, Pitts and Hebb in the 1940s, based on results from neuroscience. Imitating the human brain structure and its neurons, *ANN*s are complex parallel computational structures based on connected processing units (neurons) organized in layers. Figure 2.9 shows the configuration of an artificial neuron, which is composed by three key elements:

- A set of connexions ($w_{i,j}$): each input is weighted by a real or binary number. May also exist an extra connection, called *bias* that takes the value +1 and introduces some tendency to the computational process;
- The integrator ($\sum$): all inputs are converted to a single value by weighting each one through a linear combination;
- An activation function: this function convert the input to the output (response), which is passed on to the neighbouring neurons as output over the synaptic weights. Here can be introduced a nonlinear effect by adding a nonlinear component to the computational process.

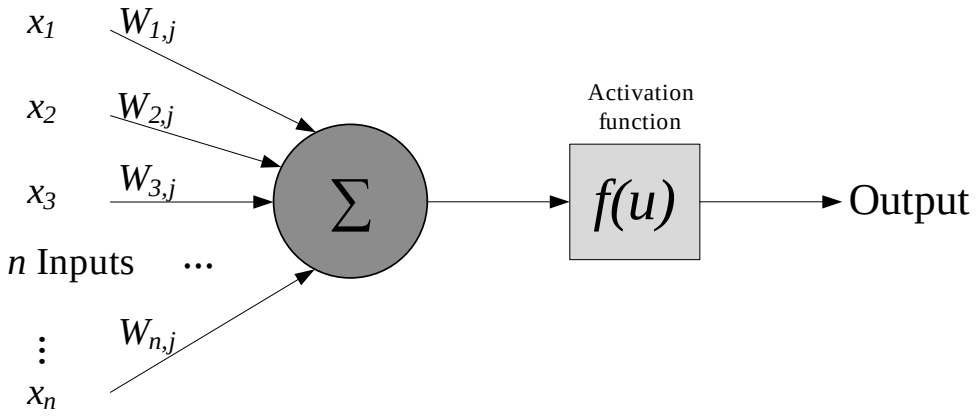


Figure 2.9: Scheme of an artificial neuron configuration

For the activation function there are a number of possibilities. The simplest is the *identity* where the neuron just calculates the weighted *sum* of the input values and passes this on. However, this frequently leads to convergence problems with the neural dynamics because the function is unbounded and the function values can grow beyond all limits over time (Ertel, 2009). In the present work was adopted the sigmoid function which is frequently used when adopting *ANN*s. Figure 2.10 depicts such function, which is translated by the following equation:

$$f(x) = \frac{1}{1 + e^{-\alpha \cdot x}} \tag{2.2}$$

Beyond these activation functions, the *threshold*, *linear* and *ramp* functions are also very common (Ertel, 2009).

The way how the neurons are organized and connected define the network architecture or topologies. Depending on the number of layers and how information flows throughout the network, they can be grouped in three main network topologies:

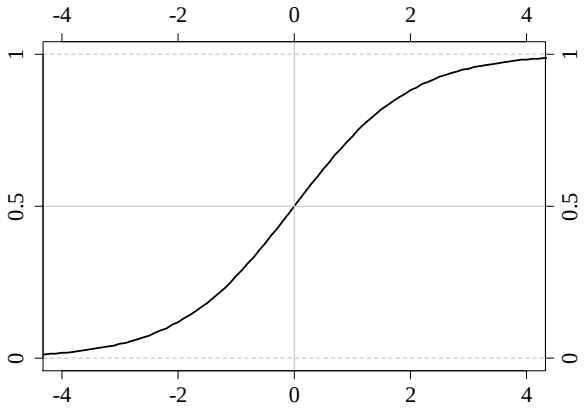


Figure 2.10: Sigmoid activation function

- One layer feedforward networks: this is the simplest network type, only composed by two layers - input and output (see Figure 2.11). The input layer is not considered because it does not perform any calculations. In this type of topology, connections are unidirectional (from input to output) and there are no connections between neurons in the same layer forming an acyclic network;

- Multilayer feedforward networks: this architecture is composed by at least two intermediate parallel layers called hidden layers (see Figure 2.12). The first is the input and the last the output layer. This is the most common type of network. By increasing the number of hidden layers, it is possible to develop more complex functions. However, the time for learning also increases, under an exponential rate;

- Recurrent: the output neurons can be connected with input ones forming cycles, conferring a spatial and/or temporal nonlinear behaviour to the network (see Figure 2.13). This type of ANN can lead to arbitrary topologies.

Optimizing a network topology is a trial and error process for there is no rule to define *a priori* the best topology. Furthermore, before start the learning process of the *ANN*, the initial values of the weights need to be defined by the user, which should be small and randomly generated. These initial values may affect the results accuracy. Therefore, if the network accuracy is not acceptable, it is common to define a different topology and to initialize the weights with a different set of values. In addiction, it is also need to define the learning ratio. Adopting small values for learning ratio the training convergence is slow but the obtained error values are also low. Only after this considerations, the learning process begin.

The learning process of an *ANN* is based on specific algorithms with very well defined rules. In this context, there are three main methods, normally called paradigms, used for

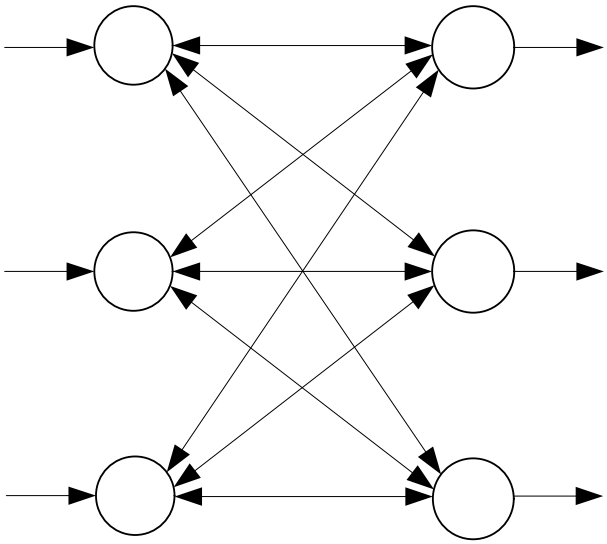


Figure 2.11: Example of a one layer feedforward network

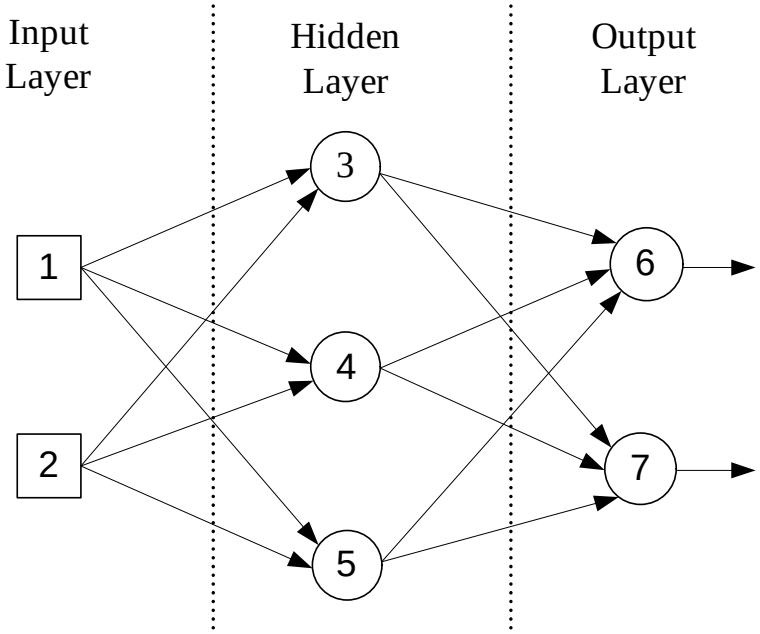


Figure 2.12: Example of a multilayer feedforward network

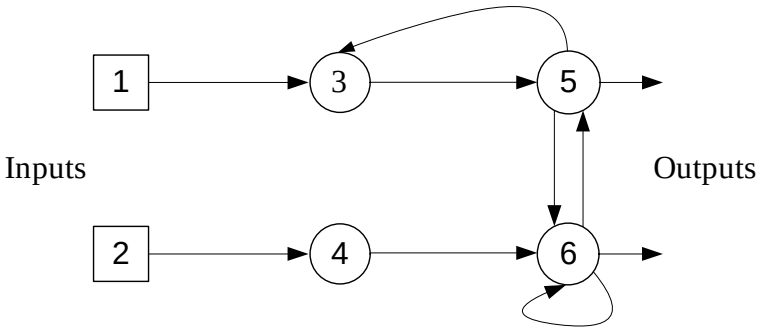


Figure 2.13: Example of a recurrent network

ANN learning process:

- Supervised learning;
- Reinforcement learning;
- Unsupervised learning.

The **supervised learning** is very popular and requires the presence of a "teacher", which gives the right answer to the network (output target). The network learns with each individual example and the proposed answer (the output) is then compared with the real value resulting in an error measure. This error is used to adjust the weights of the connection in order to minimize it in an iterative process. This type of learning is typically used for modelling dynamic systems, classification and prediction problems, and was the learning type used throughout the present research work.

In **reinforcement learning** there are also a *teacher* during the learning process. However, in this case it is just given to the network if its output is right or wrong. The right answer is not provided. Based on this information, the learning algorithm tries to improve its accuracy.

In **unsupervised** learning follows a different approach, since there is no output target. The learning process is performed through the identification of certain characteristics within in the input data, such as statistical regularities and clusters.

The first type of *ANN*s were developed in 50's. The *perceptrons* are one layer feedforward networks with several inputs and outputs and are characterized by its simplicity, since there are just a few parameters to fit. However, due to its simplicity, they are limited to solve problems with low complexity. In 60's decade, Minsky and Papert (1969) show that multilayer feedforward networks can overcome most of the limitations of *perceptron*. In 1986, Rumelhart, Hinton and Williams (Rumelhart et al., 1986), presented an algorithm for the adjustment of the weight of hidden layers called *backpropagation*.

Backpropagation algorithm performs learning in multilayer feedforward networks, which are characterized by high learning capabilities supported by its nonlinearity, existence of intermediate neurons and high connectivity degree. They are the most widely used paradigm in supervised learning. *Backpropagation* can be seen as a nonlinear extension of *perceptrons*. It is based on the selection of an error function whose value is determined by the difference between the outputs of the network and the real values. This function is minimized through the correction of the weights in an iterative process normally using the gradient descent method (Witten and Frank, 2005). The learning is ended when the stopping criterion is met. This may occur when a sufficiently low er-

ror is reached or when there are insignificant variations of weights or error function in consecutive iterations.

The *backpropagation* algorithm can be summarized in two steps (Cortez, 2002):

- Forward phase: the input vector is given to the network following forward layer by layer with fixed weights;
- Backward phase: the weights are adjusted in accordance with the error, which is propagated in a backward fashion from the output until the input layer.

Backpropagation networks are powerful learning tools and have been used with success in several applications. They are able to learn from noisy and highly nonlinear data and can recognize different sets of data within a broader dataset. Moreover, they do not require any pre-existing knowledge and statistical models.

Despite of all capabilities of *ANN*, there are also some important limitations, mainly those that use *backpropagation* algorithm:

- Absence of explanatory knowledge: models induced by *ANN*s are not comprehensible to the user. They are frequently called as "black-box" models since they give the answer but not explain it. As a result, there is a lack of theoretical basis for validation of the outcomes produced by the networks. In order to overcome such drawback, research is ongoing for the development of algorithms for the extraction of rules from trained neural networks. In this work, we adopted a sensitivity analysis procedure (*GSA* method) in order to open the "black-box";
- Computational time: the computational time during training process can be very high due to a slow convergence of the learning procedure;
- Overfitting and generalization: there are no reliable methods to define the ideal number of hidden layers as well as the correspondent number of neurons. Networks with many hidden nodes have the ability to "memorize" the desired output instead of learning the patterns. This phenomenon is classified as overfitting. When this happen the induced model can perform poorly outside its range of training. On the other hand, a too low number of hidden neurons can induce models with low learning capabilities, loosing prediction accuracy.

Support vector machines

*SVM*s, developed by Vapnik (Vapnik, 1998), have received a large attention due to their promising abilities in terms of achieving optimum supervised learning models. *SVM*s

have shown high learning capabilities even when working with complex data and can be used for either classification or regression analysis (Chen and Councill, 2003). For a given dataset, the *SVM* algorithm fits an unique and globally optimal solution. The underlying concept of *SVM*s is to map the original data into a higher dimensional feature space and to fit optimally a linear function in this feature space.

*SVM*s are a very specific class of algorithms, which are characterized by the use of kernels, absence of local minima during the learning phase, sparseness of the solution and capacity control obtained by acting on the margin, or on the number of support vectors. When compared with other types of base learners, such as the well known multilayer perceptron (also known as backpropagation neural network), *SVM* represents a significant enhancement in functionality. The supremacy of *SVM* lies in their use of nonlinear kernel functions that implicitly map inputs into high dimensional feature spaces, as schematically represented in Figure 2.14. In this feature space, linear operations may be possible that try to find the best linear separating hyperplane ($y_i = \omega_o + \sum_{i=1}^{m} \omega_i \phi(x)$), related to a set of support vector points. It is interesting that the optimal dividing hyperplane is determined by a few parameters, namely by the support vectors. Optimal separation of the support vectors is equivalent to optimal separation the entire data.

As a result of the transformation of the real space into the feature space, the number of dimensions of the new vector space grows exponentially with the number of dimensions of the original vector space. However, the large number of new dimensions is not so problematic because, when using support vectors, the dividing plane, as mentioned above, is determined by only a few parameters. This new method of representing decision functions is especially useful for a high dimensional input space: the number of free parameters in this representation is equal to the number of support vectors but does not depend on the dimensionality of the space (Vapnik et al., 1997). Although *SVM*s are linear learning machines with respect to the feature space, they are in effect nonlinear in the original input space. This means that *SVM* can learn nonlinear behaviors without the drawbacks of nonlinear approaches, i.e., occurrence of local minima, convergence problems and overfitting. *SVM*s are indeed currently very popular. This is mainly due to their capacity to combine the advantages of linear and nonlinear models, as well as their predictive results that were achieved in several domains.

Let $XY = \{(x, y) \,|\, (x_1, y_1), ..., (x_N, y_N)\}$ denote the training dataset, where N is the number of training samples. In linear *SVM*s, the relation between input variable x_k (where k represent the k^{th} model attribute) and the predicted variable $\hat{y}_k$ can be described by

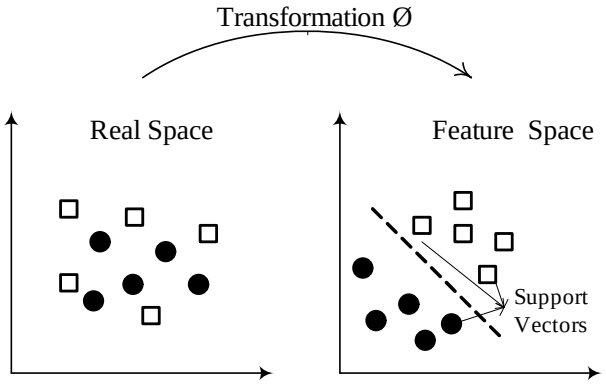


Figure 2.14: Example of a *SVM* transformation (adapted from Cortez (2010)).

the linear function $f(x)$ taking the form of:

$$\hat{y}_k = f(x_k) = \langle w, x_k \rangle + b \tag{2.3}$$

where$\langle ., . \rangle$ denotes the dot product, w and b are the weight vector and bias parameter, respectively. For *SVM* regression, the aim is to find a pair of unknown vectors of (w, b) that minimize the prediction error for training samples and has at most an ϵ deviation from actual target y_k. This implies that there is no penalty during optimization for the pairs when $|y_k - f(x_k)| \leq \epsilon$ and is defined by the ϵ-insensitive loss function, L_ϵ, which can be expressed as follows (Gunn, 1998):

$$L_\epsilon(y) = \begin{cases} 0 & \text{for} \quad |f(x) - y| < \epsilon \\ |f(x) - y| - \epsilon & \text{otherwise} \end{cases} \tag{2.4}$$

To ensure that the minimal complexity risk would be obtained, in order to have optimal structural risk minimization, one can minimize the norm of w, $\|w\|^2 = \langle w, w \rangle$. Hence, the constrained regression problem can be mathematically written as a convex optimization problem according to the following equations:

$$min_{w,b,\xi_k,\xi_k^*} \frac{1}{2} \|w\|^2 + C \cdot \sum_{k=1}^{N} (\xi_k + \xi_k^*) , \tag{2.5}$$

$$\text{subject to} \begin{cases} y_k - \langle w, x_k \rangle - b \leq \epsilon + \xi_k \\ \langle w, x_k \rangle + b - y_k \leq \epsilon + \xi_k^* \\ \xi_k, \xi_k^* \geq 0 \end{cases} \tag{2.6}$$

where, ξ_k and ξ_k^* are slack variables. The constant regularization parameter $C \geq 0$ in Equation 2.5 determines the trade of between the complexity of the function and the deviation from the tolerable error ϵ. The problem represented in Equations 2.5 and 2.6 is a convex quadratic programing optimization which can be converted to a Lagrange function by introducing a dual set of positive Lagrange multipliers variables. This Lagrange function could be solved by maximizing its dual optimization problem and has a saddle point regarding its primary and dual variables. The final solution of the optimization problem is given by:

$$w = \sum_{k=1}^{N} (\alpha_k - \alpha_k^*) \cdot x_k \xrightarrow{\text{yields}} \hat{y}_{new} = f(x_{new}) = \sum_{k=1}^{N} (\alpha_k - \alpha_k^*) \cdot \langle x_k, x_{new} \rangle + b \tag{2.7}$$

where $\alpha_k \geq 0$ and $\alpha_k^* \geq 0$ are Lagrange multipliers. As seen in Equation 2.7, w can be completely described as a linear combination of the training vectors and the Lagrange multipliers. The samples that are inside the ϵ-insensitive tube make both Lagrange multipliers zero, and w actually is represented by only some training vectors, called *support vectors* (*SVs*), which lie outside the ϵ-insensitive tube. Thus, complexity of the solution is not dependent of the dimensionality of the problem, whereas SVs define the complexity of the function.

For enriching *SVM* algorithm to deal complex nonlinear relationships, some preprocessing procedures of training patterns can be implemented (Vapnik, 1998; Gunn, 1998). This can be done by mapping input vectors into a higher-dimensional feature space by the means of kernel functions, which yields the nonlinear *SVM* for the kernel function of $K \langle ., . \rangle$. Its solution is given by:

$$\hat{y}_{new} = f(x_{new}) = \sum_{k=1}^{N} (\alpha_k - \alpha_k^*) \cdot K \langle x_k, x_{new} \rangle + b \tag{2.8}$$

Kernel functions are important to control the complexity of final solution. One may choose any arbitrary kernel functions (Hamel, 2009), such as:

$$\text{Linear: } k \langle x, x' \rangle = \langle x, x' \rangle \tag{2.9}$$

$$\text{Polynomial: } k \langle x, x' \rangle = \langle x, x' \rangle^d, \quad d > 0 \quad \text{or}$$

$$k \langle x, x' \rangle = (\langle x, x' \rangle + 1)^d, \quad d > 0 \tag{2.10}$$

$$\text{Gaussian radial basis function: } k \langle x, x' \rangle = exp\left(-\gamma \cdot \|x - x'\|^2\right), \quad \gamma > 0 \tag{2.11}$$

$$\text{Exponential radial basis function: } k \langle x, x' \rangle = exp\left(-\gamma \cdot \|x - x'\|\right), \quad \gamma > 0 \tag{2.12}$$

In highly nonlinear spaces, radial basis function kernel usually yields more promising results in comparison with other mentioned kernels and present less parameters than other kernels (Gunn, 1998). Hence, in this work we adopt the popular Gaussian kernel throughout all experiments.

SVM was initially proposed for classification problems by Vladimir Vapnik and his co-workers (Cortes and Vapnik, 1995). Later, after the introduction of an alternative loss function proposed by Vapnik (Smola, 1996), called ϵ-insensitive loss function, was possible to apply *SVM* to a regression problems (Smola and Schölkopf, 2004). When working with *SVM*, it is well known that its generalization performance (estimation accuracy) depends on a good setting of meta-parameters C (regularization parameter), ϵ (width of a ϵ-insensitive zone) and the kernel parameters (Gilan et al., 2012). The problem of choosing a good parameter setting in a learning task is the so-called model selection. This task is further complicated by the fact that *SVM* model complexity (and hence its generalization performance) depends on all three parameters. Parameter C controls the trade-off between complexity of the machine (flatness) and the number of non-separable data points and may be viewed as a "regularization" parameter (Goh and Goh, 2007). For example, if C is too large (infinity), then the objective is to minimize the empirical risk only, without regard to model complexity part in the optimization formulation. This parameter is usually determined experimentally (trial and error) via the use of a training and test (validation) set. Parameter ϵ controls the width of the ϵ-insensitive zone, used to fit the training data. The value of ϵ can affect the number of support vectors used to construct the regression function. The bigger ϵ, the fewer support vectors are selected. On the other hand, bigger ϵ-values results in more "flat" estimates. Hence, both C and ϵ-values affect model complexity, but in a different way. Selecting a particular kernel type and kernel function parameters is usually based on application-domain knowledge and also should reflect distribution of input (x) values of the training data.

The problem of finding the best combination of hyper-parameters (model selection) is often troublesome due to the highly nonlinear space of the model performance with respect to these parameters (Gilan et al., 2012). Although an exhaustive search method could be used to tune these hyper-parameters, it suffers from the main drawbacks of being very time-consuming and lacking of a guarantee of convergence to the globally optimal solution. Hence, several approaches have been proposed in order to find the best set of parameters with less effort (time and computing consuming) (Huang et al., 2007; Cherkassky and Ma, 2004; Frohlich and Zell, 2005; Gilan et al., 2012).

Huang et al. (2007) propose a nested *Uniform Design* (*UD*) methodology for efficient, robust and automatic model selection for *SVM*. In contrast to conventional exhaustive grid search, this method can be treated as a deterministic analogue of random search.

The key theoretic advantage of the *UD* model selection over the grid search is that the *UD* points are "far more uniform" and "far more space filling" than lattice grid points. The better uniformity and space-filling phenomena make the *UD* selection scheme more efficient by avoiding wasteful function evaluations of close-by patterns. Furthermore, this model selection scheme is robust and efficient and can be carried out fully automatically. In addition, *UD* approach provides the flexibility to adjust the candidate size under computational cost constraint. In practice, it can be combined with variants of *SVM* implementations easily. Following *UD* approach, a heuristic for setting up a two-dimensional search box in the parameter space, which is able to automatically scale the distance factor in the Gaussian kernel, is given. Regardless of the search scheme, it is always important to set up a proper search region. Once the search region is determined, it is applied the 2-stage *UD* methodology to select the candidate set of parameter combinations and perform a k-fold cross validation to evaluate the generalization performance of each parameter combination. The 2-stage *UD* procedure first sets out a crude search for a highly likely candidate region of global optimum and then confines a finer second-stage search therein. In the present research work, this model selection approach was implemented in the feature selection step, taking advantage of its the flexibility (the three hyperparameters are automatically defined).

During the learning phase of all *SVM* models were adopted the recommendations proposed by Cherkassky and Ma (2004). Following this approach, the parameter C is analytically selected from the training data. Therefore, a "good" value for C can be chosen equal to the range of output (response) values of training data but considering the presence of outliers. So, the following expression is proposed to calculate the regularization parameter:

$$C = max\left(\left|\bar{y} + 3\sigma_y\right|, \left|\bar{y} - 3\sigma_y\right|\right) \tag{2.13}$$

where $\bar{y}$ and σ_y are the mean and the standard deviation of the y values of training data.

For ϵ parameter is proposed an analytical selection based on the input noise level in the training data (assuming that the standard deviation of the noise σ is known or estimated from the data) and on the number of training samples. Thus, the following expression is suggested: $\epsilon = \hat{\sigma}/\sqrt{N}$, where $\hat{\sigma} = 1.5/N \cdot \sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2$, y_i is the measured value, $\hat{y}_i$ is the value predicted by a 3-nearest neighbour algorithm and N the number of examples.

Functional networks

In a first look to *Functional Network* (*FN*), we can found some similarities with *ANN*. However, there are important differences that should be stressed. Unlike *ANN*, in a *FN*

the goal is allowing the neuron functions to be learned and suppressing the weights between connexions (Castillo et al., 1998). This new type of networks is a general framework useful for solving a wide range of problems such as statistics and engineering applications (Castillo et al., 2001; El-Sebakhy et al., 2006; Li et al., 2001) and it has been successfully applied in both prediction (Alonso-Betanzos et al., 2004) and classification (Zhou et al., 2005) problems. Its neural functions can be multivariate, multi-argument and it is also possible to use different learnable functions, instead of fixed functions. Moreover, there is no need to associate weights to the connections between nodes, since the learning is achieved by the neural functions. These features represent a remarkable difference between *FN* and *ANN* networks. It should be noted that *FN*s are not arbitrary but subject to strong constraints to satisfy the compatibility conditions imposed by the existence of multiple links going from the last input layer to the same output units. When compared with *ANN*s, there are inclusively some advantages that deserve be highlighted (Zhou et al., 2005). Unlike *ANN*, *FN* can reproduce certain physical characteristics that lead to the corresponding network in a natural way. However, such reproduction only takes place if one use an mathematical expression with a physical meaning inside the function database. Moreover, the estimation of the network parameters can be obtained by resolving a linear system of equations, which returns a fast and unique solution, i.e. the global minimum is always achieved.

While presenting a similar structure, *ANN* and *FN* also have important differences. For example, the selection of the initial topology of the *FN* is normally based on the problem domain, instead of several topologies and choosing one using an optimal criterion, such as happen in *ANN*. The initial topology in a *FN* can be further simplified using functional equations and its neural functions can be multidimensional and set during the learning phase. Moreover, *FN* incorporates different neural functions, normally functions from a given family, such as polynomial or exponential, and they are not restricted to be a linear combination of inputs. Finally, the neurons outputs can be connected to each others, which is not the case of the standard *ANN*. In Figure 2.15, it is shown the *FN* associations. The structure of a *FN* consists in (see Figure 2.15):

- a layer of input storing units;
- a layer of output storing units;
- one or several layers of processing units, which evaluate a set of input values, coming from the previous layer and delivers a set of output values to the next layer;
- none, one or several layers of intermediate storing units;

- and a set directed links, that connect units in the input or intermediate layers to neuron units, and neuron units to intermediate or output units.

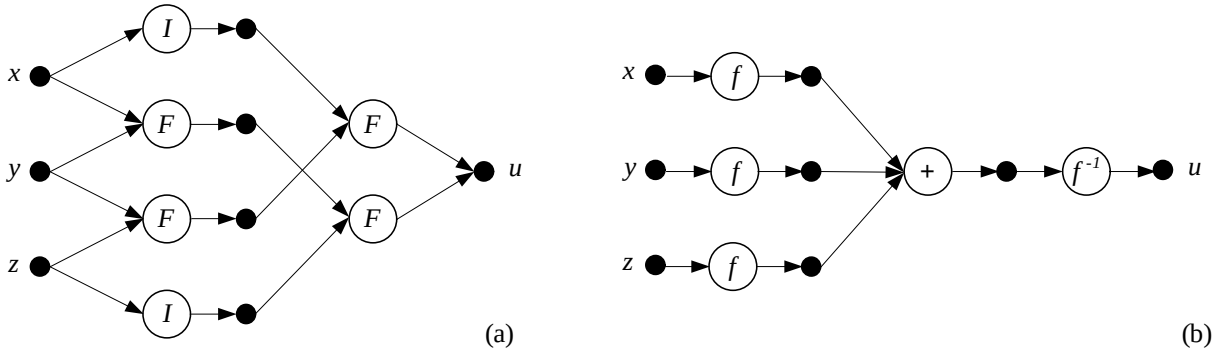


Figure 2.15: Example of the *FN* associations: a) initial network, b) equivalent simplified network (Castillo et al., 1998).

When working with *FN*, several steps are necessary to be set. The first one is to define the initial topology of the network, based on problem to be solved. Next, the architecture using functional equations and the equivalence concept needs to be initialized, and then checked the uniqueness condition of the desired architecture. Third, using the available data, the learning procedure (i.e. training algorithm) is realized by considering the combinations of linear independent functions, $\psi = \{\psi_{s1}, ..., \psi_{sm_s}\}$, for all s to approximate the neuron functions, that is:

$$g_s(x) = \sum_{i=1}^{m_s} (\alpha_{si} \cdot \psi_{si}(x)) \quad \text{for all s} \tag{2.14}$$

where the coefficients α_i are the parameters in *FN*. The most common linearly independent functions are:

$$\begin{aligned} \psi &= \{1, X, ..., X^m\} \\ \psi &= \{1, e^x, e^{-x}, ..., e^{mx}, e^{-mx}\} \\ \psi &= \{1, cos(x), ..., cos^l(x), sin^l(x)\} \end{aligned} \tag{2.15}$$

where m is the number of elements in the combination of sets of linearly independent function. To learn the parameters in Equation 2.14, different optimization techniques can be used, such as the least squares algorithm, conjugate gradient, iterative least squares, minimax or maximum, such as likelihood estimation. The last step in implementation process is to select the best model and validate it.

2.4 Feature selection

Feature Selection (*FS*) is a process of selecting a subset of original features according to a given criterion. It is an important and frequently used technique in *DM* for dimension reduction and an essential step in successful *DM* applications. *FS* has been an active research field in the last decades in *DM*, having been widely applied to many fields. This research area is of great practical significance and has been developed and evolved to answer the challenges due to data of increasingly high dimensionality. There are many potential benefits of *FS*: facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance. Furthermore, it reduces the number of features, removing irrelevant, redundant, or noisy features, and brings about palpable effects for applications, by improving *DM* performance, providing faster and more cost-effective predictors, allowing a better understanding of the underlying process that generated the data, and helping prepare, improving learning accuracy, and leading to better model performance (Liu et al., 2010).

The key point on *FS* is: *what variable are redundant?* A presumably redundant variable could be useful when taken with another set of variables. Guyon and Elisseeff (2003) pointed out some important observations related to redundant variable:

- Noise reduction and consequently better class separation may be obtained by adding variables that are presumably redundant. Variables that are independently and identically distributed are not truly redundant;
- Perfectly correlated variables are truly redundant in the sense that no additional information is gained by adding them;
- Very high variable correlation (or anti-correlation) does not mean absence of variable complementarity;
- A variable that is completely useless by itself can provide a significant performance improvement when taken with others;
- Two variables that are useless by themselves can be useful together.

Based on several studies has been shown that some features can be removed without performance deterioration (Liu et al., 2010). On the other hand, it is known that including too many input variables to a model are often harmful, since it can lead to overfitting phenomenon, especially for small databases. Likewise, including only few variables are not

always beneficial due to underfitting problem. Therefore, some trade-off between these extremes is highly important.

The best subset of variables contains the least number of dimensions that most contribute to accuracy, being the unimportant dimensions removed. This is an important stage of preprocessing and is one of two ways of avoiding the curse of dimensionality (Swell, 2007). To perform *FS*, there are two main approaches with practical application:

- Forward selection: the process start with no variables and add them one by one. At each step it is add the one that most decrease the error, until any further addition does not significantly decrease the error (or improve the model performance);
- Backward selection: here, the process star with all variable and remove them one by one. At each step is removed the one that most decreases the error (or increases it only slightly), until any further removal increases the error significantly.

Swell (2007) summarize at his paper a list of different algorithms for *FS*, given an overview of different approaches that can be used (see Figure 2.16).

Figure 2.17 shows a unified view for a *FS* process. This process comprise two phases:

- Feature selection
- Model fitting and performance evaluation

In few words, a subset of the original features is selected via certain research strategies, which is evaluated in order to analyse the utility of the candidate set. Some features can be add or discard to the candidate set. If the set of selected features is good enough using certain stopping criterion, then the selected data are used to train a particular learning model and test it with the test dataset. The decision whether proceed to a new iteration is normally supported on the test error, which is calculated with the validation dataset in order to reduce overfitting problems.

In order to improve the chances to select the best set of variables, the following list enumerate 10 questions that should be answered to help to solve a *FS* problem (Guyon and Elisseeff, 2003):

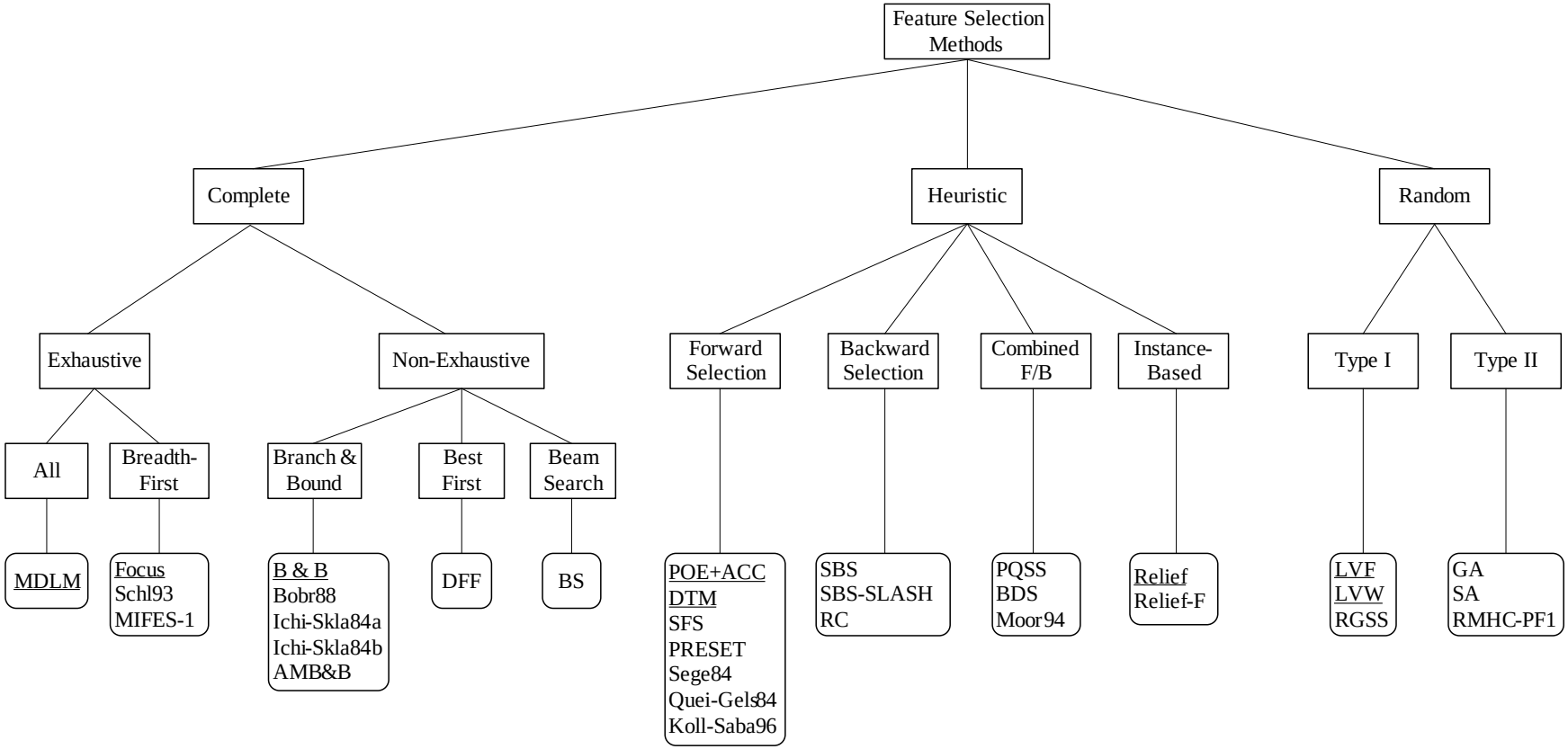


Figure 2.16: Overview of *FS* methods (adapted from Dash and Liu (1997))

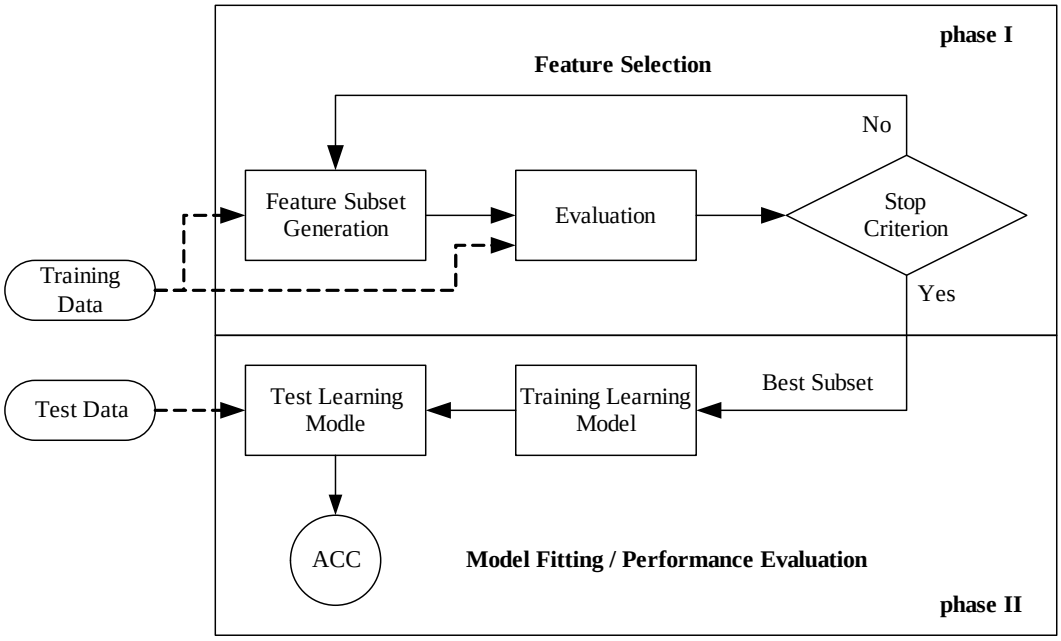


Figure 2.17: A unified view of a *FS* process (Liu et al., 2010)

1. Do you have domain knowledge?
2. Are your features commensurate?
3. Do you suspect interdependence of features?
4. Do you need to prune the input variables (e.g. for cost, speed or data understanding reasons)?
5. Do you need to assess features individually (e.g. to understand their influence on the system or because their number is so large that you need to do a first filtering)?
6. Do you need a predictor?
7. Do you suspect your data is "dirty" (has a few meaningless input patterns and/or noisy outputs or wrong class labels)?
8. Do you know what to try first?
9. Do you have new ideas, time, computational resources, and enough examples?
10. Do you want a stable solution (to improve performance and/or understanding)?

2.5 Model assessment and interpretation

As previously mentioned, it is fundamental to perform a rigorous assessment of the *DM* model when applied to on unseen data, in order to measure its generalization capacity. For model assessment two different approaches can be taken:

- Objective: when model performance is evaluated based on statistics and structures of patterns (e.g, support, confidence, etc);
- Subjective: the model is subjectively accessed when the user's belief in the data are applied (e.g. unexpected, novelty, etc).

Furthermore, and due to the high mathematics complexity of some data-driven models, particularly those resulting from *SVM* and *ANN* algorithms, some procedures need to be applied in order to extract understandable information from them. In this section, we describe the approaches used to perform model assessment, as well as its interpretability.

2.5.1 Evaluation measures

Depending if the problem at hands is a classification or a regression task, different evaluation measures can be applied. In regression, evaluation metrics are computed based on the difference between observed and predicted values (the errors). Typically, the lower the error, the better is the predictive model, being a value of zero the ideal goal to be achieved.

In this work, we adopt three common metrics: *Mean Absolute Deviation* (*MAD*) (Equation 2.16); *Root Mean Square Error* (*RMSE*) (Equation 2.17) and the Squared Correlation Coefficient (R^2) (Equation 2.18). Low values of *MAD* and *RMSE*; and R^2 close to the unit value should be interpreted as high model predictive capacity. The main difference between *MAD* and *RMSE* is that the latter one is more sensitive to extreme values since it uses the square of the distance between the real and predicted values. When compared with *MAD*, *RMSE* penalizes more heavily a model that in a few cases produces high errors. Thus, these two error measurements give different and complementary perspectives about the behaviour of the induced models, allowing its comparison.

These three metrics can be calculated by the following way. Let y_k be the actual value and $\hat{y}_k$ be the predicted value of the k^{th} observation and N be the number of observations, then *MAD*, *RMSE* and R^2 could be defined, respectively, as follows:

$$MAD = \frac{\sum_{i=1}^{N} |y_k - \hat{y_k}|}{N} \tag{2.16}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (y_k - \hat{y_k})^2}{N}} \tag{2.17}$$

$$R^2 = \left(\frac{\sum_{i=1}^{N} (y_k - \bar{y}) \cdot (\hat{y_k} - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N} (y_k - \bar{y})^2 \cdot \sum_{i=1}^{N} (\hat{y_k} - \bar{\hat{y}})^2}} \right)^2 \tag{2.18}$$

Furthermore, different regressions *DM* models can be easily compared by plotting the *Regression Error Characteristic* (*REC*) curve proposed by Bi and Bennett (2003), which plots the error tolerance on the x-axis *versus* the percentage of points predicted within the tolerance on the y-axis. In this work, we also adopt this representation for the model performance analysis.

2.5.2 Generalization capacity

Another important issue in a model evaluation is its generalization capacity. That is, how a *DM* model is able to accurately predict unseen values. The most common methods to infer about generalization capacity of a predictive model are *holdout*, *cross-validation* and *leave-one-out*.

Following an *Holdout* approach the dataset is randomly partitioned into two independents sets, one for training and the other for test. The training set, used to induce the model, allocates typically 2/3 of the records and the remaining 1/3 are used for model accuracy measurement. The main advantage of this approach is its simplicity and speed. However, this method is not much robust, tending to produce different results for different data random splits.

The *Cross-Validation*, schematically presented in Figure 2.18, is an improvement of holdout approach, allowing to use all data available for training and testing. According to this approach, the data (P) are randomly sampled into k mutually exclusive subsets ($P_1, P_2, ..., P_k$), with the same length. Training and testing is performed k times and the overall error of the model is taken as the average of the errors obtained in each iteration. The values of k can range from 2 to N, where N is the number of data sample. The typical value for k are 5, 10 or 20, depending of the dimension of the dataset. This method is more robust than the holdout but requires more computation.

The *Leave-One-Out* (Hastie et al., 2009) approach can be seen as a special case of Cross-Validation. This method is especially suited when the dataset is small (e.g. lower

than 100 examples). Under leave-one-out, sequentially one example is used to test the model and the remaining data is used to fit the model. Under this scheme, all data is used for training and testing. Yet, this method requires around N times more computation, since N models are fitted. The final generalization estimate is evaluated by computing evaluation metrics for all N test samples.

In order to improve model reliability, each one of the above approaches described can be performed T times (executions, also known as runs).

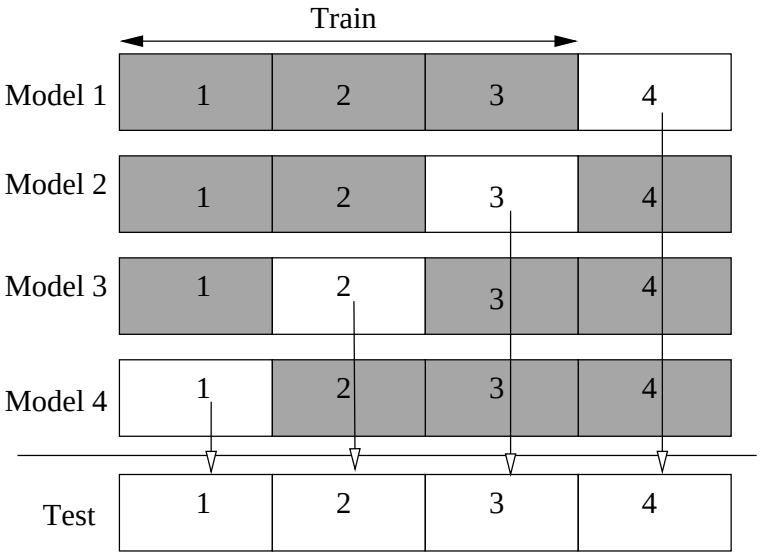


Figure 2.18: Cross-Validation approach

2.5.3 Sensitivity analysis

Basically, there are two fundamental requirements that a data-driven model should satisfy. On one hand, it is required a high prediction quality. On the other hand, namely within the engineering domain, the predictive model should be understandable and easy to interpret. However, this is precisely one of the main drawbacks related with black-box data-driven models, such as *ANN* and *SVM*. In order to solve this issue, Cortez and Embrechts (2011) proposed a novel visualization approach based on a *SA* method, which is used in this work. *SA* is a simple method that is applied after the training phase and measures the model responses when a given input is changed, allowing to quantify the relative importance of each attribute, as well as its average effect on the target variable.

In particular, we applied the *GSA* method (Cortez and Embrechts, 2011), which is able to detect interactions among input attributes. This is achieved by performing a simultaneous variation of F inputs (that can range from 1, one dimensional *SA*, denoted as 1-D, to I, I-D *SA*). Each input is varied through is range with l levels and the remaining

inputs are kept fixed to a b baseline value. In this research work, it was set: $l = 12$, which allows an interesting detail level under a reasonable amount of computational effort; and b is set to the average input variable value.

First, the *DM* model is fitted to the whole dataset. Then, the *GSA* algorithm (Algorithm 1) is applied to the fitted *DM* model, being the respective sensitivity responses stored. In the Algorithm 1 the S_D jagged array is built using the Algorithm 2, while the $predict(M, X)$ is a function that returns the responses of model M given the input matrix X (of $N \times I$ size). The **REP** procedure is equivalent to the R rep function (R Development Core Team, 2009) (e.g. **REP**((1,2),2,2)=(1,1,2,2,1,1,2,2)).

Using the sensitivity responses, two important visualization techniques can be computed. The input importance barplot shows the relative influence of each input in the model (from 0% to 100%). The rational of *SA* is that the higher the changes produced in the output, the more important is the input. To measure this effect, following the suggestion of Cortez and Embrechts (2011), it was adopted the gradient metric:

$$g_a = \sum_{j=2}^{l} |\hat{y}_{a,j} - \hat{y}_{a,j-1}| \,/\, (l-1) \tag{2.19}$$

where a denotes the input variable under analysis, $\hat{y}_{a,j}$ is the sensitivity response for $x_{a,j}$. Having computed the gradient for all inputs, then the relative importance (R_a) is calculated using:

$$R_a = g_a / \sum_{i=1}^{I} g_i \cdot 100(\%) \tag{2.20}$$

To analyse the average impact of a given input x_a in the fitted model, the *Variable Effect Characteristic* (*VEC*) curve can be used, which plots the attribute l level values (x-axis) versus the *SA* responses (y-axis). Between two consecutive $x_{a,j}$ values, the *VEC* plot performs a linear interpolation. To enhance the visualization analysis, several *VEC* curves can be plotted in the same graph. In such case, the x-axis is scaled (e.g. within [0,1]) for all x_a values. Similarly, when a pair of inputs (x_{a1},x_{a2}) is simultaneously varied ($F > 2$), the *VEC* surface can be plotted, showing the average responses to changes in the pair.

Algorithm 1 Global Sensitivity Analysis (Cortez and Embrechts, 2011)

procedure GSA($M, S_D, F, b, N_{\hat{y}}$)
  $m_x \leftarrow 1$
  for $a \in F$ **do** ▷ compute m_x length
    $m_x \leftarrow m_x \times length(S_D[a, *])$
  end for
  $X \leftarrow$ matrix $m_x \times (I + Y_{col})$ ▷ rows × columns
  for $a \in \{1, ..., I\}/F$ **do**
    $X[*, a] \leftarrow b(a)$ ▷ set $/F$ columns to baseline
  end for
  $e \leftarrow 1$
  for $a \in F$ **do** ▷ set SA inputs
    $x'_a \leftarrow S_D[a, *]$
    $t \leftarrow m_x/(e \cdot length(x'_a))$
    $X[*, a] \leftarrow$ **REP**(x'_a, e, t) ▷ replicate x'_a
    $e \leftarrow e \cdot length(x'_a)$
  end for
  $y_{col} \leftarrow \{I + 1, ..., I + N_{\hat{y}}\}$ ▷ output columns
  $X[*, y_{col}] \leftarrow predict(M, X[*, \{1, ..., I\}])$
  Output: X ▷ matrix with SA inputs and responses
end procedure
procedure REP($x, each, times$) ▷ auxiliary function
  $x_r = \emptyset$ ▷ empty vector
  for $j \in \{1, ..., times\}$ **do**
    $x_e = \emptyset$ ▷ empty vector
    for $i \in x$ **do**
      $x'_e \leftarrow$ vector with $each \times length(x)$ elements
      $x'_e[*] \leftarrow i$ ▷ all x'_e elements are set to i
      $x'_e \leftarrow c(x_e, x'_e)$ ▷ concatenate operator
    end for
    $x_r \leftarrow c(x_r, x_e)$ ▷ concatenate operator
  end for
  Output: X_r ▷ vector with replicates from x
end procedure

Algorithm 2 Scanning data method (Cortez and Embrechts, 2011)

procedure SCAN DATA(D, F, l)
  for $a \in F$ **do**
    $S_D[a, *] \leftarrow scan(D[*, a], l)$
  end for
  Output: S_D ▷ jagged array with scanned inputs
end procedure

2.6 Data mining tools

Nowadays, there are several data analysts, such as *RapidMiner*, *R*, *Excel*, *Weka*, *SAS* and *Matlab* between others. In this work, we adopted the *R* environment, which has gained attention of the *DM* community in the past few years (see Figure 2.19). The *R* environment is a multiple platform (e.g. Windows, Mac OS) and free open-source tool that is based on a high-level matrix programming language, broadly used for statistical and data analysis. *R* environment is based on objects and on a high-level language, being its functionalities easily extended by installing new packages, which are continuously being developed by an very active *R* community. In addition, an extensive help system is included and available from the prompt (*help.start()* calls the full tutorial in an HTML browser). Furthermore, there is also a large documentation freely available on the *R* Web site (http://www.r-project.org/) as well on books (Muenchen and Hilbe, 2010). While not specifically oriented for Business Intelligent / *DM*, the *R* environment includes a large variety of Business Intelligent / *DM* algorithms (e.g. Neural Networks, Support Vector Machines, Bayesian Networks or Decisions Trees). Furthermore, *R* is currently used by a large number of Business Intelligent / *DM* analysts. As a drawback, *R* requires some effort for non expert users to initially learn the tool, due to the lack of an easy to use graphical user interface (GUI), as well as the absence of technical support. Usually, almost usage of *R* is under a console command interface as shown in Figure 2.20, where all commands are typed. Yet, after some experience and training, the user achieves a better control and understanding of what is being executed (in contrast with several "black-box" *DM* GUI products).

The R environment was not specifically developed for conducting *DM* projects. Thus, some packages were developed to improve this issue. Two of the most interesting interfaces, are *Rattle* and *rminer* packages. The main advantage of *Rattle* is its graphical interface (Figure 2.21), while *rminer* is easier to install and requires much less *R* packages. Moreover, *rminer* presents more *ANN* and *SVM* capabilities (e.g. in *Rattle* version 2.6.18, *SVM* cannot be used for regression tasks and the *ANN* algorithm is unable to automatically search for the best number of neurons on the hidden layer).

In this work, we adopted the *rminer* library (available at http://www3.dsi.uminho.pt/pcortez/rminer.html or R CRAN packages). This library is an integrated framework that uses a console based approach and facilitates the use of *DM* algorithms in classification and regression tasks (Cortez, 2010). Moreover, *rminer* is particularly suited for *ANN* and *SVM* (two of the main *DM* algorithms used in the present work), making use of a short and coherent set of functions:

(a)

What data mining tools have you used for a real project (not just for evaluation) in the past 6 months? (364 voters)
SPSS PASW Modeler (formerly Clementine) (68 alone, 52 with other tools, 120 total)
RapidMiner (36 alone, 41 w. other tools, 77 total)
SAS (39 alone or with SAS EM; 36 with other tools, 75 total)
Excel (1 alone, 68 total)
SAS Enterprise Miner (39 alone or with SAS; 28 w/ other tools; 67 total)
R (2 alone, 51 total)
Your own code (3 alone, 44 total)
KXEN (25 alone, 31 total)
Weka (now Pentaho) (0 alone, 31 total)
MATLAB (0 alone, 26 total)
Other commercial tools (0 alone, 19 total)
KNIME (1 alone, 18 total)
Other free tools (0 alone, 15 total)
Microsoft SQL Server (1 alone, 15 total)
Zementis (5 alone, 13 total)
Oracle DM (0 alone, 9 total)
Statsoft Statistica (0 alone, 8 total)
Orange (0 alone, 5 total)
Salford CART, Mars, other (1 alone, 5 total)
C4.5/C5.0 (0 alone, 4 total)
Angoss (0 alone, 4 total)
Inference for R (0 alone, 3 total)
Viscovery (0 alone, 2 total)
Megaputer (0 alone, 2 total)
Insightful Miner/S-Plus (now TIBCO) (0 alone, 2 total)
Bayesia (1 alone, 2 total)
Thinkanalytics (1 total)
Miner3D (1 total)
Clario Analytics (1 total)

(b)

Which data mining/analytic tools you used in the past 12 months for a real project (not just evaluation) [912 voters]	
RapidMiner (345)	37.8%
R (272)	29.8%
Excel (222)	24.3%
KNIME (175)	19.2%
Your own code (168)	18.4%
Pentaho/Weka (131)	14.3%
SAS (110)	12.0%
MATLAB (84)	9.2%
IBM SPSS Statistics (72)	7.9%
Other free tools (67)	7.3%
IBM SPSS Modeler (former Clementine) (67)	7.3%
Microsoft SQL Server (63)	6.9%
Statsoft Statistica (57)	6.2%
Other commercial tools (56)	6.1%
SAS Enterprise Miner (50)	5.5%
Zementis (34)	3.7%
Orange (25)	2.7%
Oracle DM (19)	2.1%
KXEN (19)	2.1%
Salford CART Mars other (15)	1.6%
VisuaLinks (12)	1.3%
Viscovery (10)	1.1%
Angoss (8)	0.9%
TIBCO Insightful Miner (7)	0.8%
Miner3D (7)	0.8%
REvolution Computing (4)	0.4%
Megaputer Polyanalyst/TextAnalyst (3)	0.3%
Portrait Software (2)	0.2%
Data Applied (2)	0.2%
Centrifuge (2)	0.2%
PRSD Studio (1)	0.1%
Clario Analytics (1)	0.1%
Bayesia (1)	0.1%

Figure 2.19: *DM*/analytic tools used poll: a) May 2009, b) May 2010. Source: `http://www.kdnuggets.com/polls` (kdnuggets web page)

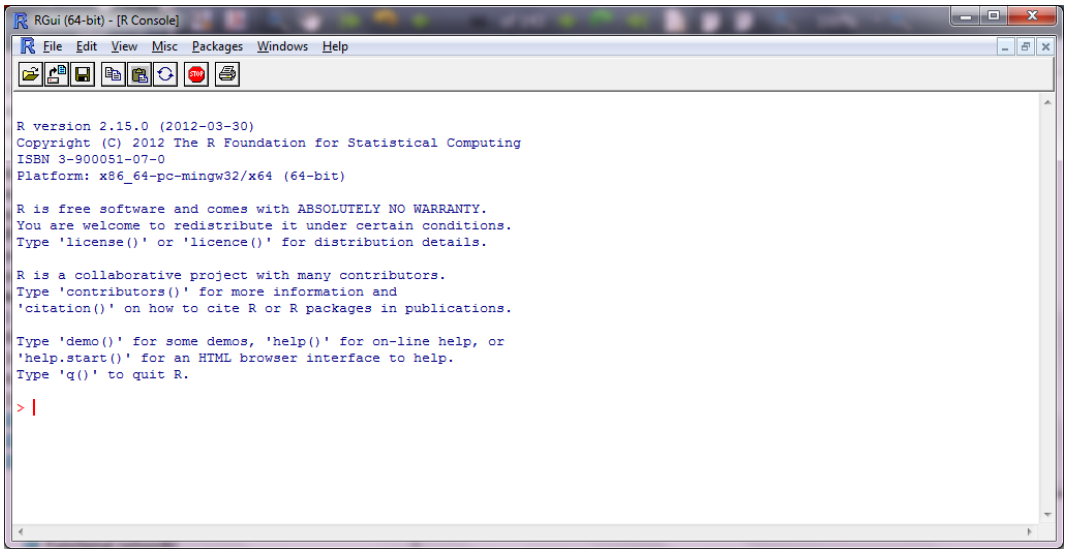


Figure 2.20: Snapshot of *R* console

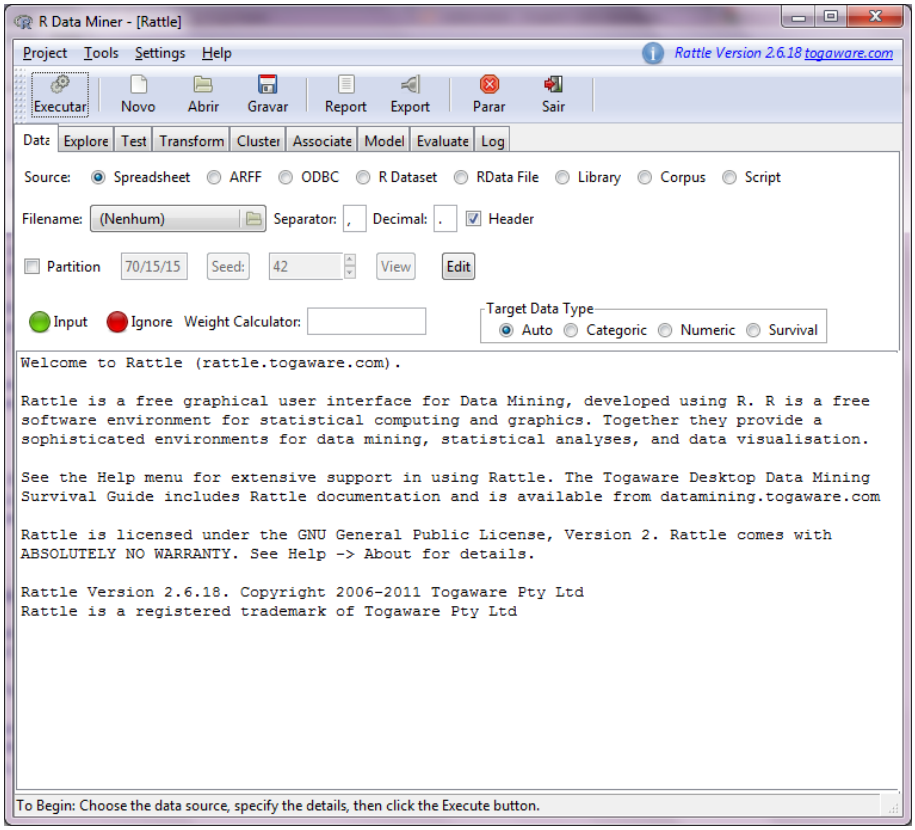


Figure 2.21: Snapshot of the *Rattle* graphical interface for *DM* in *R*

- fit: create and adjust a given *DM* model using a dataset;
- predict: returns the predictions for new data;
- mining: a powerful function that trains and tests a particular model under several runs;
- mgraph: returns several graphs;
- metrics and mmetric: compute classification or regression error metrics.

For regression tasks, *rminer* package allows implement the following *DM* algorithms: *naive*- most common class; *dt* - decision tree; *rm* - multiple regression; *bruto* - additive spline model; *mars* - multivariate adaptive regression splines; *knn* - k-nearest neighbour; *mlp* - multilayer perceptron with one hidden layer; *mlpe* - multilayer perceptron ensemble; *svm* - support vector machine; and *randomforest* - random forest algorithm.

Additionally to the statistical *R* environment and *rminer* packages, we also used the free version of *General Algebraic Modelling System* (*GAMS*) (GAMS Development Corporation, 2012) for the implementation of the *FN*s. *GAMS* is an high-level modelling

system for mathematical programming and optimization. It consists of a language compiler and a stable of integrated high-performance solvers. *GAMS* is tailored for complex, large scale modelling applications, allowing to build large maintainable models that can be adapted quickly to new situations.

Moreover, particularly for initial data observation, we used the powerful visualisation characteristics of *GGobi* software (Cook and Swayne, 2007) as well its ability to connect with *R*. *GGobi* is an open source visualization program for exploring high-dimensional data. It provides highly dynamic and interactive graphics such as tours, as well as familiar graphics such as the scatterplot, barchart and parallel coordinates plots. Plots are interactive and linked with brushing and identification.

2.7 Conclusions

There are a large number of successful *DM* projects in different domains, including the geotechnical field. Such success motivates this work, which aims to use *DM* techniques for enhancing *JG* column design. However, there are some important issues that should be taken into account for achieving a valuable impact. On one hand, it is fundamental that sufficient data with significant attributes are available for the discovery task. On the other hand, quality and reliability of the data are also relevant issues in a *DM* problem. Moreover, for a successful implementation of a *DM* project, several steps should be taken. For these issues, the application of *SEMMA* or *CRISP-DM* methodologies can give a valuable contribution.

Currently, powerful *DM* algorithms are available to explore high-dimensionality data and extract useful rules and patterns. Two of the most well-known and implemented in this research are the *ANN*s, which are inspired by the neurons system structure of the human brain, and the *SVM*s supported in statistical theory.

Another issue related to a *DM* problem is the selection of the model attributes, particularly in problems with high dimensionality. To help in the task, several approaches have been proposed. In the present work, the forward and backward *FS* approaches were applied to guide the process of selecting the input variables.

For model assessment, particularly in regression problems, different metrics, such as *MAD*, *RMSE* and R^2, can be calculated to measure the deviation between prediction and experimental values. The model's interpretability is as important as its performance. This is a relevant issue because data-driven models are normally characterised by high mathematical complexity. Accordingly, the application of a *GSA* can give a valuable contribution. Particularly, this analysis is able to measure the relative importance of the

input variables as well as their average effects on the target variable.

As a final note, it should be emphasised that actually there are several data analysis methods, each with its advantages and limitations. The *R* environment has gained attention within the *DM* community in the past few years and was adopted in the present work. One of the most attractive features of the *R* environment is the possibility of installing new packages, which extend its functionalities.

Chapter 3

Jet grouting technology

3.1 Background and definitions

The main goal of any ground improvement method is to improve those soil characteristics that match the desired results of a project. For example, an increase in density and shear strength to overcome stability problems; reduction of soil compressibility; influencing permeability to reduce and control ground water flow; increase the rate of consolidation; or improve soil homogeneity.

Ground improvement techniques are continually in progress, both quantitatively and qualitatively, as a result of not only technology developments but also of an increasing awareness of the environmental and economic advantages of modern ground improvement methods. Moreover, the last decade has seen an increasing demand for *in situ* deep soil mixing work in Europe and North America (Moseley and Kirsch, 2004).

Within ground improvement techniques, there is distinction between methods of compaction or densification (e.g. deep vibro techniques, or dynamic compaction) and methods of soil reinforcement through the introduction of additional material into the ground (e.g. cement grouting, compaction grouting or jet grouting). Following, a brief summary of some of the most relevant soil improvement techniques is presented, emphasizing *JG* technology.

According to the fundamental concepts of soil mechanics, the placement of an external load on a low-permeable soil layer will induce excess pore water pressure, causing a consolidation process in which pore water is pushed out of the soil. As a result, the effective stress increases gradually and the excess pore water pressure decreases. This process is termed as consolidation, and will continue until the excess pore water pressure has dissipated. The duration of this process is mainly related with the drained path. Therefore, the idea behind the installation of vertical *band drains* is to reduce the length

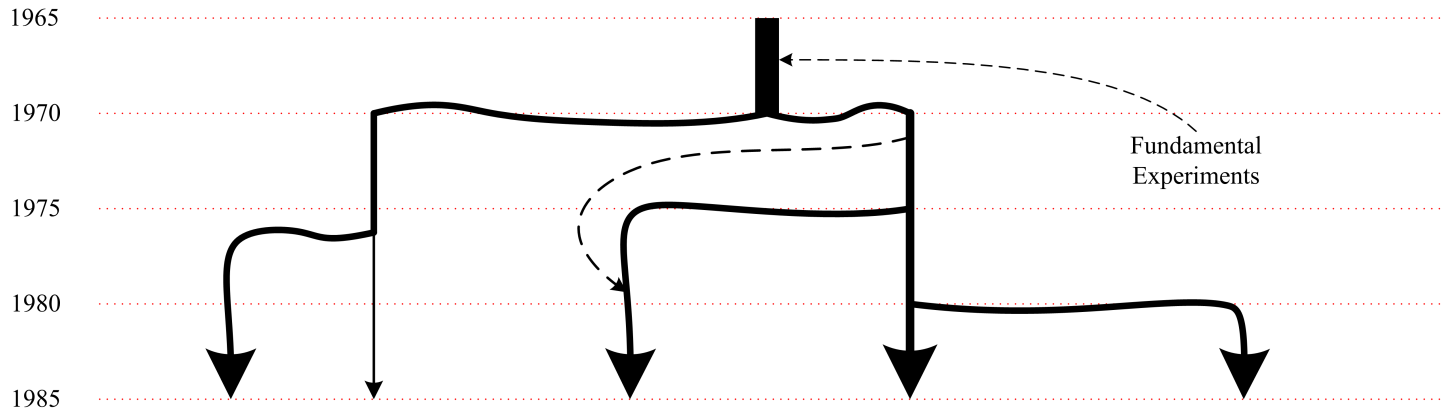
of the drainage paths and thereby reduce the time of consolidation. For this purpose, different strategies can be adopted, such as the used of vertical sand drains, cardboard wicks or geodrains.

Another well known soil improvement method is *cement grouting*. Normally, grouting is used to fill voids in the ground, aiming to increase resistance against deformation, to supply cohesion, shear-strength and UCS or to reduce conductivity and interconnected porosity in an aquifer. Grouting uses liquids which are injected under pressure into the pores and fissures of the ground. Liquid grout mixes consist of mortar, particulate suspensions, aqueous solutions and chemical products, such as polyurethane, acrylate or epoxy. By displacing gas or groundwater, these fluids fill pores and fissures in the ground, conferring new properties (after setting and hardening) to the subsoil.

The concept behind JG, i.e. the use of high pressure water for disrupt the ground, dating from middle of 60's decade and was proposed by Japanese specialists (Xanthakos et al., 1994). In 1965, Yamakada brothers (Miki and Nakanishi, 1984) applied this concept not only for cutting purposes but also to mixture soil with cement. These developments gave rise to the first two forms of JG, which date by early 1970.

Since then, several JG forms had been developed, improved and merged leading to the three main system currently applied (Xanthakos et al., 1994). The major categories of JG applied in Japan in 1985 are summarized and described in Figure 3.1. The strong improvements on equipment development, providing significantly higher flow rate at higher pressures, allowed, since the early 1990, improve volume of soils 20 times as large as the conventional systems. This technology progress enabled to obtain JG columns with around 5 meters in diameter or even up to 9 meters in softer ground.

By the late 1970's, JG technology was initially applied in Japan, Germany, UK, Italy, France, Singapore and Brazil, by groups of geotechnical contractors, and then throughout the world. Despite of all potential of JG technology as a soil improvement method, its acceptance found some obstacles. The risk/legal concerns, inherent to any novel method, appear on the top of the list. Moreover, inappropriate applications and initial technical problems leading to poor performance, are responsible for its slow acceptance, particularly in North America. In Portugal, JG was introduced in the middle of 90's decades, mainly on Lisbon underground extension works. Nowadays, JG solutions have become competitive and advisable in several and more usual geotechnical scenarios (Falcão et al., 2000).



Method	**CJG ≡ JET3**	**JSG ≡ JET2**	**CCP ≡ JET1**	**SSS - MAN**
Definition	Half-replacement method using a high speed water jet enveloped by air jet.	Mixing method using high speed gorut jet enveloped by air.	Mixing method using high speed gorut jet.	Rverse type replacement method using a high speed water jet enveloped by air jet.
Outline of method	Uplift of a rotating high pressure water jet enveloped by air jet cut in situ soil, which is halfly removed by the uplift flow of the air, then a space formed by that removed is mixed with the grout continuously supplied from a road to produce large size cylindrical solidified body.	Uplift of a rotating high speed horizontal grout jet enveloped by air jet mixes in situ soil with the grout and produce a large size cylindrical solidified body.	Uplift of a rotating horizontal grout jet with a super high pressure (20 MPa) mixes in situ soil with the grout and produces a cylindrically solidified body.	A pilot hole is drilled by reverse circulation, then a rotating superhigh pressure (60 MPa) water jet enveloped by air jet lowered removing the cut soil through the reverse road to produce a large size space, which is filled up with grout after the confirmation of the sapce shape.
Fit soil	Cohesive soil, sandy soil.	Cohesive soil, sandy soil, gravelly soil.	Cohesive soil (N<5), sandy soil (N<15).	Cohesive soil, sandy soil, gravelly soil.
Improvement efficiency	A large size improved diameter of 1500-3000 mm and high strength with uniformity are attainable.	A large size improved diameter of 800-2000 mm and a high strength with unifirmity are attainable.	Na improvement diameter of 300-500 mm and high strength with uniformity are attainable.	A large size cylindrical solidified body having the diameter of 2000-4000 mm is attainable per one road. As grout concrete, clay, cement mortar, etc can be used. The diameter of the body can be confirmed on the ground, so a serve management can be done.

Figure 3.1: Development of *JG* methods in Japan from 1965 to 1985 (adapted from Miki and Nakanishi (1984))

JG technology is classified as a grouting on ground improvement methods and is defined as placement of a pumpable material (normally a cementitious material) directly into the subsoil, without previous excavation. The cinematic energy of the drilling fluid cut the soil, allowing its mixture with the injected grout. At the end, a new material, also known as *soilcrete*, with an controlled geometry structure is obtained, presenting better physical and mechanical proprieties when compared with natural soil.

JG is actually a viable solution for a wide range of problems when conventional injection methods are unsuitable, unsafe or too expensive. Although be a recent technology on ground improvement, it is notable its fast growing worldwide (Terashi and Juran, 2000). Its growth has been in response to the need to treat fine and/or grain soils that can not be treated with permeation grouting, to produce very high strengths and to comply with major environmental controls that chemical grouts may not meet.

Nowadays, *JG* is one of the most used deep mixing improvement methods worldwide (Nikbakhtan et al., 2010), where slurry cement is injected into the natural soil, obtaining a new material characterized by an enhancement in terms of resistance, stiffness and permeability.

JG technology has aroused interest within the geotechnical community due to it great versatility, enabling to improve mechanical and physical properties of different soil types, obtaining different geometries shapes (columns, panels, etc.) with different orientations (vertically, horizontally or inclined). As shown in Figure 3.2, that compares the applicability of different soil improvements methods, *JG* technology can be economically used from coarse to fine-grained soils (GmbH, 2002). Moreover, it requires just few equipments, can be applied from confined places, such as from inside of buildings, allows to treat a specific zone (e.g. a confined stratum) and is economically attractive when compared with other soil improvement methods (Falcão et al., 2000). The bearing capacity of *JG* columns can still be improved by introducing steel profiles inside them.

It is this high versatility of *JG* technology, namely in terms of soil type and geometry, that give it the ability to solve a large diversity of geotechnical problems. Proof of this is the high diversity of *JG* applications scattered throughout the world. These different applications can be grouped under the following headings (Essler and Yoshida, 2004):

- groundwater control;
- movement control;
- support; and
- environmental.

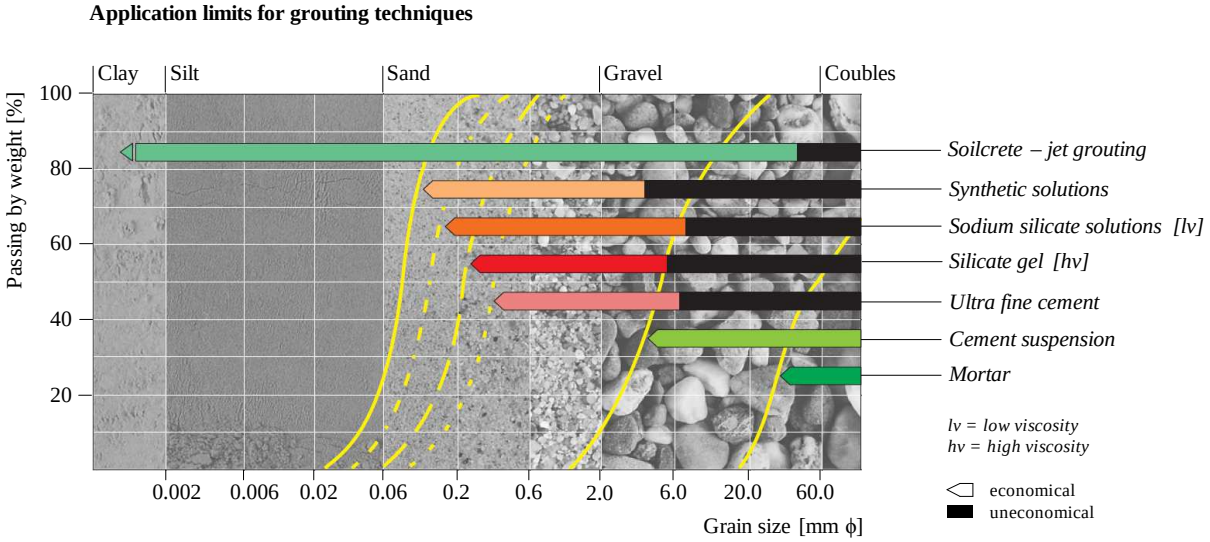


Figure 3.2: Comparison of the applicability of different soil improvement methods (adapted from GmbH (2002))

JG technology is also frequently required for groundwater control works, such as to create waterproof barriers or to perform sealing works. In fact, low permeability values, normally around to 10^{-9} to 10^{-10}, are a key characteristic of *soilcrete*. Moreover, it can be used within environmental issues for preventing or reducing contamination flow through the ground or encapsulating contaminants in the ground or into sensitive water systems (Gazaway and Jasperse, 1992).

Taken advantage of the improved mechanical properties of *soilcrete*, *JG* technology is often applied on tunnel protection, underpinning buildings during excavation or transferring foundation load through weak material to a competent strata. Furthermore, can also be used in embankments or cuttings by increasing the safety factor (Welsh and Burke, 1991; Padura et al., 2009; Gazzarrini et al., 2008; Shibazaki and Yoshida, 1997; Gazzarrini et al., 2005).

Despite of all particular characteristics of *JG* technology, there are also some less positives aspects that should be enumerated. The high cement and water consuming is one of the main less attractive points of *JG* technology. Furthermore, the bearing capacity of the soil immediately after the soil improvement is very low. This means that can occur undesirable settlements (Wang et al., 1998). In addiction, the high pressure used during the soil improvement can damage neighbour structures induced by uncontrolled soil movements (Wang et al., 1999), particularly if for some reason the excess material can not achieve the surface. For this reason it very important to check if the excess material, that result from the soil improvement, can freely ascend to the surface throughout the free space between the open borehole and the rod. This spoil that ascends to the surface

can also represent a environment threat if were not taken the appropriate measures.

Another important issue related with *JG* technology is its design. As previously pointed out in Chapter 1, *JG* columns diameter and *soilcrete* mechanical properties prediction are a complex task, mainly due to the high dimensionality of the problem. This subject is of particular importance for *JG* technical and economic efficiency, and represents the scope of the present research work.

3.2 Function and effects of the JG technology equipment on soil improvement

Conceptually, soil improvement by *JG* technology can be described in two main steps: drilling phase followed by the mixture process. In the first step, a *JG* string with simple, double or triple inner conduit, which convey the *JG* fluid(s) to the monitor, is drilled into the soil until the intended depth and with the orientation of the column that will be built. During this stage, a water jet flow can be used to facilitate the penetration process and clean the space between the borehole walls and the rods, which is an important aspect to successfully carry out the soil improvement in terms of security and technical requirements. In the second phase, the improved mass of soil is obtained by jetting of the disaggregating and cementing fluid(s) through small nozzles (2 mm to 4 mm of diameter) screwed to the monitor. At the same time, a jet grouting rig apply a pre-established withdrawal and rotation speed to the rods while the fluids are pumped with a pressure (until 550 times of atmosphere pressure) and flow rate pre-specified. The excess water soil-cement mixture, currently termed *spoil*, is removed to the surface through the annular space between drill rod and borehole wall. Figure 3.3 schematically represents the *JG* process as well as the main equipment and materials involved on the entire process.

As previously underlined, one of the aspects that make *JG* technology a remarkable soft soil improvement method is the few amount of equipment necessary for its application. Following are enumerated the main equipment used on soil improvement by *JG* technology, where most of them can be identified in Figure 3.3.

- Drilling rigs;
- Jet grouting string;
- Monitor;
- Nozzle(s);
- Cement;

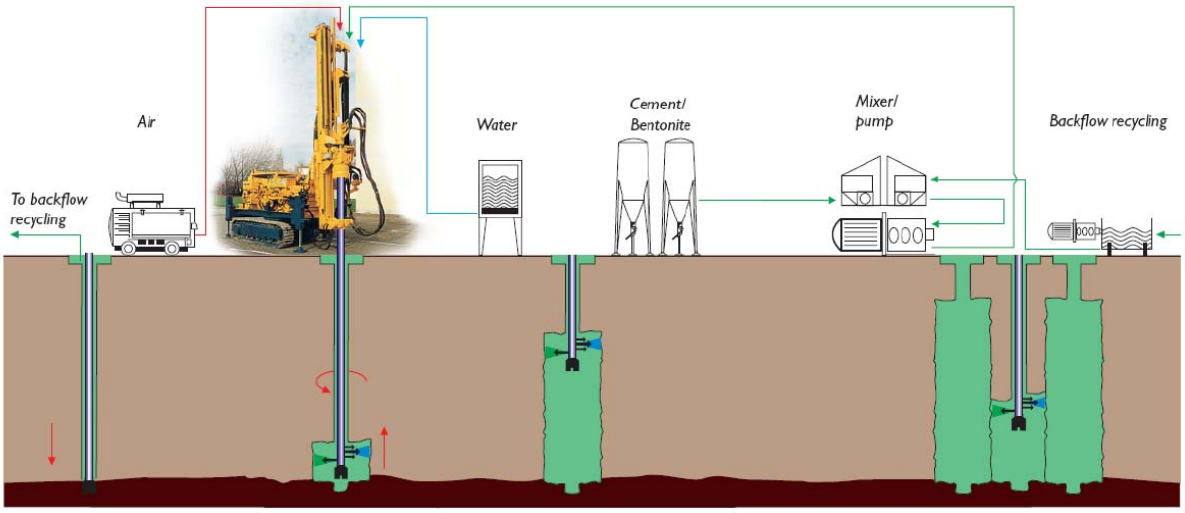


Figure 3.3: *JG* process (adapted from GmbH (2002))

- Water reservoir;
- Grout slurry mixture station;
- Soilcrete pump and control station;
- Air compressor.

Being the *JG* technology a soil improvement method, the soil properties are a key element in the final characteristics of the new material resulting, in terms of both mechanical behaviour and column diameter.

A practical way to assess the influence of the soil, is to separate it between granular and cohesive soils. Accordingly, and based on several studies, it was observed that unconfined compression tests results (a standard test for quality control) follow the distribution shown in the histograms plotted in Figure 3.4 for cohesive and granular ground. Essler and Yoshida (2004) also propose some reference values for UCS, cohesive strength, bond strength and bending tensile strength for granular and cohesive soils (see Table 3.1). Moreover, there are also some reference values, proposed by several authors, for different soil types, which are summarized in Table 3.2.

Relating to the *JG* column diameter, the influence of the soil it is also assessed in terms of its structure, i.e. considering whether the soil is granular or cohesive. Figure 3.5 shows a relation between the N_{SPT} of the soil and the *JG* column diameter, as a function of the jet system applied.

[1]JET1 - Single fluid system; JET2 - double fluid system; JET3 - triple fluid systems (see Section 3.3 for more details.

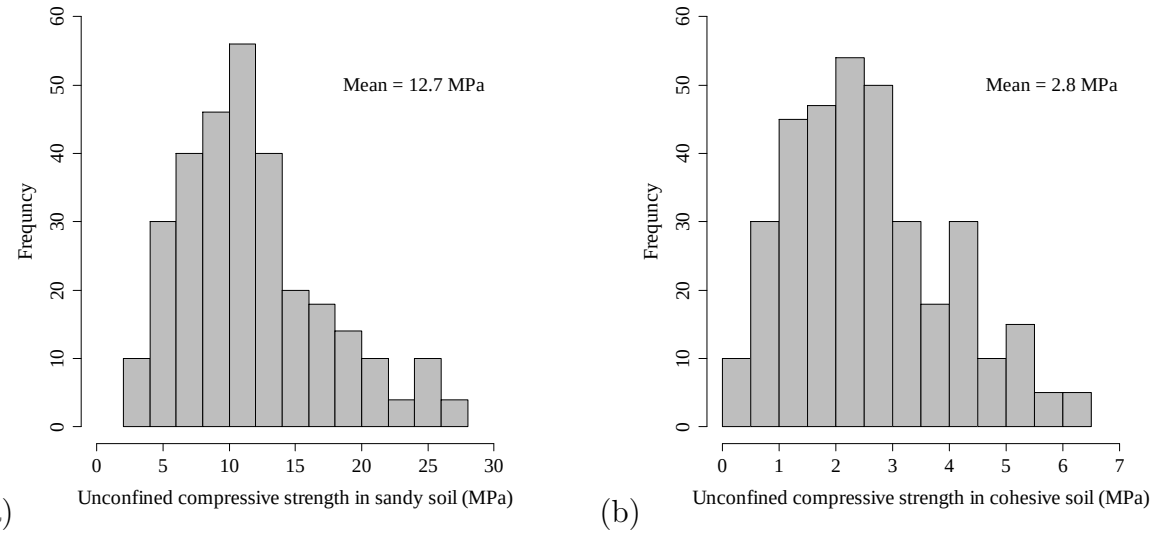


Figure 3.4: Histogram of experimental UCS of *soilcrete* for: a) sandy soil, b) cohesive soil. (adapted from Essler and Yoshida (2004))

Table 3.1: Standard strengths in design (adapted from Essler and Yoshida (2004)).

Soil type	Unconfined compressive strength ($\mathrm{MN\,m^{-2}}$)	Cohesive strength ($\mathrm{MN\,m^{-2}}$)	Bond strength ($\mathrm{MN\,m^{-2}}$)	Bending tensile strength ($\mathrm{MN\,m^{-2}}$)
Cohesive	1	0.3	0.1	0.2
Granular	3	0.5	0.17	0.33

Table 3.2: UCS of materials treated by JG technology (adapted from Carreto (2000)).

Author/Data	W/C	Soil Type - UCS (MPa)				
		Organic clay	Clay	Silt	Sand	Gravel
Welsh and Burke (1991)	-	-	1 to 5	1 to 5	5 to 11	5 to 11
Baumann et al. (1984)*	1:1,5	-	-	6 to 10	10 to 14	12 to 18
	1:1,0	-	-	3 to 5	5 to 7	6 to 10
Paviani (1989)*	-	-	1 to 5	1 to 5	8 to 10	20 to 40
Teixeira et al. (1987)*	-	0,5 to 2,5	1.5 to 3.5	2 to 4.5	2.5 to 8	-
JJGA (1995)*	-	0.3	1	1 to 3	-	-
Guatteri et al. (1994)*	-	-	0.5 to 4	1.5 to 5	3 to 8	-

* In Carreto (2000)

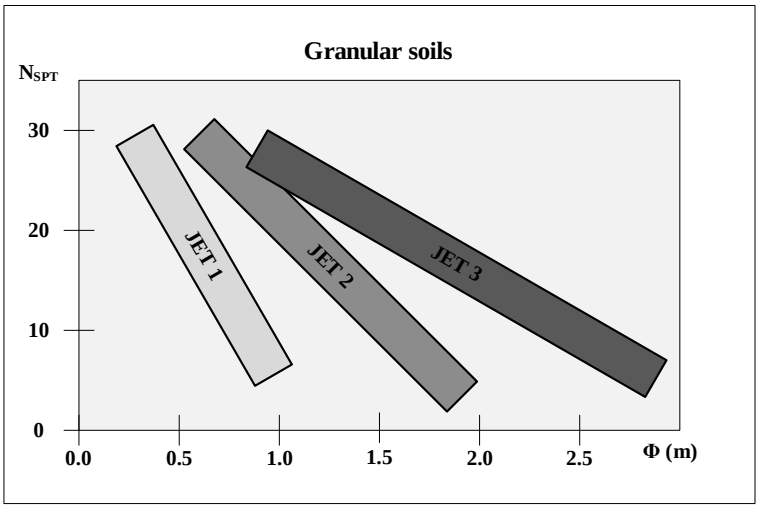


Figure 3.5: Relation between *JG* column diameter and N_{SPT} for different jet systems[1]

There are also some empirical abacus similar to those plotted in Figure 3.6 that depicts the column diameter as a function of N_{SPT} for different soil types and *JG* systems. It should be stressed that *JG* column diameter is one of the most important parameters used for quality control purposes. Therefore, *JG* column diameter quantification is of particular importance for the economy of the soil improvement.

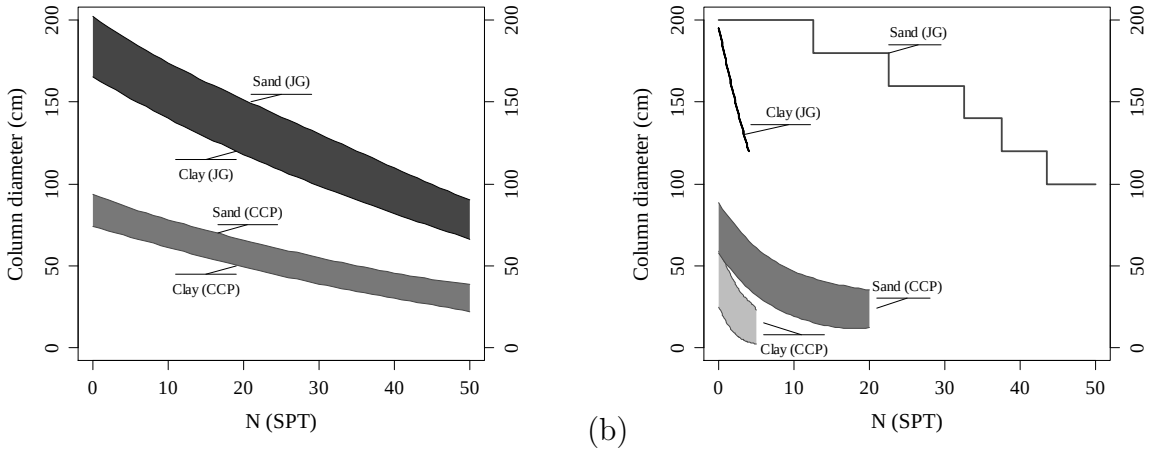


Figure 3.6: *JG* column diameter in function of SPT number for different soil types and *JG* systems. a) according to Brazilian practice (NOVATECNICA, 2003), b) proposed by Miki and Nakanishi (1984) and Abramento et al. (1998) (CCP - single fluid system, JG - double fluid system)

Concerning to the equipment used to perform the soil improvement, the drilling rigs is placed as close as possible to the improvement spot and is linked to the cement slurry pump station and, if necessary, to the air compressor throughout high pressure hoses, through which the different fluids are conducted. This machine is normally coupled to

an truck and is provided by a drill pipe that range just from few meters (allowing to access confined places) to several meters high (allowing to increase productivity). The main function of the drilling rigs is to control both rotating and withdrawal speeds as well as column orientation. The first two *JG* parameters allow control the cement content of the mixture as well as the cut effect. Higher withdrawal and rotation speeds means lower cement content and cut effect (keeping the remaining parameters constant). As a result, monitoring these two parameters, can be controlled the mechanical properties of *soilcrete* as well as *JG* column diameter. Moreover, this machine also disposes of an operational panel and a recording station (see Figure 3.7) where are displayed and recorded several *JG* parameters (e.g withdrawal and rotation speeds) for supervision and control in real time. This informations can then be analysed and interpreted.

Figure 3.7: *JG* record station

Jet grouting string is coupled to the drilling rigs and is formed by jointed rods provided by one monitor (see Figure 3.8) coupled at its end. This monitor enables jetting of the fluids into the ground and is provided by a drill bit, which enables/facilitates the drilling process. Figure 3.9 shows some details of the nozzles used on *JG* technology. The nozzle is a specially manufactured device screwed to the monitor and designed to transform the high pressure fluid flow within the *JG* strings into a high speed jet directed against the soil. They are normally placed perpendicularly to the monitor. However, its orientation is part of *JG* design as well as its dimension (diameter) and number. The influence of this important element, is more noticeable in *JG* column diameter than in *soilcrete* mechanical properties. Its number and diameter, as well as orientation will affect the jet energy and therefore the ability to cut and reach highest distances. The importance of this element is reflected in the strict control that is targeted during soil improvement.

Upstream of the drilling rigs, are the remains equipments listed above and shown in Figure 3.3. Among them, it should be stressed the importance of cement silo and water

Figure 3.8: *JG* monitor details, showing nozzles and drill bit position

Figure 3.9: *JG* nozzle details

reservoir that ensure a continuous supply of cement and water respectively. Concerning to cement type, it should be stressed that its influence is particularly noticeable in the strength development ratio (Limprasert, 1995). Finally, the cement slurry obtained from mixing cement with water is pressurized in the pump station (see Figure 3.10). This equipment is responsible to create all necessary energy to disrupt the soil and mix them with the cement slurry. Combining the injection pressure of the fluids with the diameter of the nozzles, it is developed the sufficient energy to perform the soil improvement.

Figure 3.10: Pump station (equipment used in Multiusos - Viana do Castelo)

3.3 Jet grouting systems

Since its first application until nowadays, *JG* technology has undergone several developments and refinements. One of the main developments is related with the number of fluids injected, which define the three main systems currently in use, i.e., single fluid system, double fluid system and triple fluid system[2]. More recently, other systems has been proposed, where Xjet system is highlighted. Following are emphasized the main characteristics related to the different *JG* systems, as well as the influence of each one in the mechanical properties of *soilcrete* and *JG* column diameter.

3.3.1 Single fluid system

Single fluid system is the simplest form of *JG*, where it is just injected cement slurry, at high pressure and velocity. This fluid is responsible to erode the soil and mix with it. A schematic representation of this system is presented in Figure 3.11. Single fluid system is predominantly used in horizontal *JG* works, namely in tunnel protection. Furthermore,

[2]Single, double and triple fluid systems are also currently known as JET 1, JET 2 and JET 3 respectively.

it is normally the best alternative when there are concerns about the air usage and loss of strength. This was the first system to be developed and the small columns diameter produced, usually up to 1 m in diameter, is one of its main limitations. Moreover, the borehole opened to introduce the rods has a tendency to become blocked, often resulting in ground heave. Under single fluid system, six jetting parameters must be specified: grout pressure, flow rate, number and diameter of the nozzles and withdrawal time and rotation speed of the drill rod. There are also some other parameters, related with cement slurry properties, that also need to be defined, such water/cement ratio, cement or water type (e.g. drinking or *in situ* water).

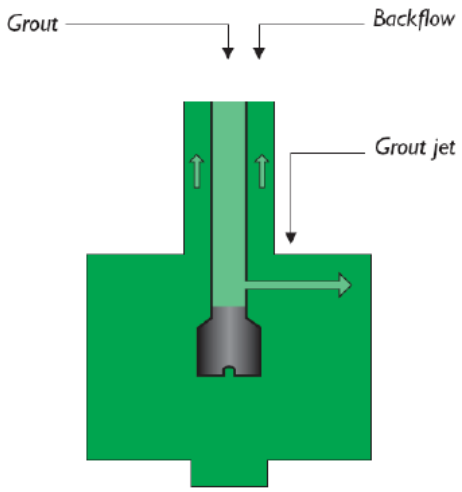


Figure 3.11: Single fluid system schema (adapted from GmbH (2002))

3.3.2 Double fluid system

Double fluid system, schematically represented in Figure 3.12, is very similar to the single system but with the addition of an air shroud the cement grout jet. Adding air to the grout jet the cutting energy increases, allowing higher eroding distance, mainly above water table. Beyond its benefits related with the erosion energy, the compressed air is very importance for conveying spoil up to the ground surface. However, due to the injection of air during the mix process, the final mixture present highest porosities, which normally leads to lower strength values. Moreover, on double fluid system a lot of grout may be lost to the surface due to the airlift, decreasing soil improvement efficiency. In this system, additionally to all parameters related with single fluid system, it is also need to control the pressure and flow rate of air jet.

The development of double fluid system was strongly supported on the observation of jet behaviour on different media. The experience has shown that a water jet is very

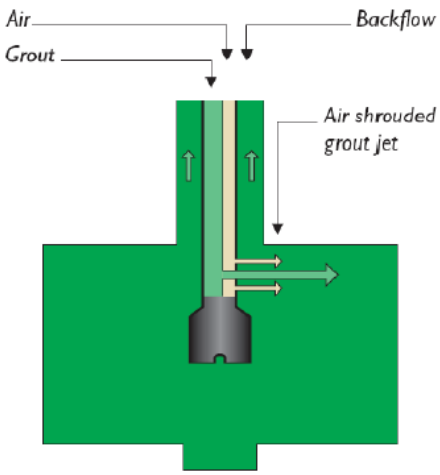


Figure 3.12: Double fluid system schema (adapted from GmbH (2002))

effective in air (for instance as a fire extinguisher), and that its effectiveness is significantly decreased in water. This observation is shown in Figure 3.13 that sketches the eroding distance of a jet in air, in water, and in water with an air shroud. Therefore, and taking into account that *JG* technology is frequently performed beneath the water table, the efficiency of an alone grout jet will be low. So, shrouding the liquid jets with compressed air, is created a atmosphere effect by eliminating ground water around the jets. However, to increase the effectiveness of the air shrouding liquid jet, its velocity should be higher than half the sonic velocity, ensure a thickness of one millimetre and provide sufficient air flow. A compressed air may be generated by a low-pressure compressor rated at 0.7 MPa for work up to 20 m deep, but is dependent of the ground water pressure. For deeper works high-pressure compressor is required.

3.3.3 Triple fluid system

Triple fluid system, schematically represent in Figure 3.14, is slightly different and more complex than single and double fluid systems. In this system the erosion of the ground is carried out by a high pressure water jet shrouded by air and the mixture process is performed by an additional low pressure grout line. Typically, grouting nozzles are placed half a meter below the water jetting nozzle in order to convey as much excavated soil particles as possible to the surface while limiting the grout ejected. By controlling independently the erosion and grout ejection, this system is superior to the other two systems from the point of view of control quality. Moreover, a higher column diameter can be obtained. The triple system is usually less viscous and hence offers less risk for blockage and potential structural or ground movement. However, and similar to double fluid system, the strength of the final mixture is lower due to the injection of air during the

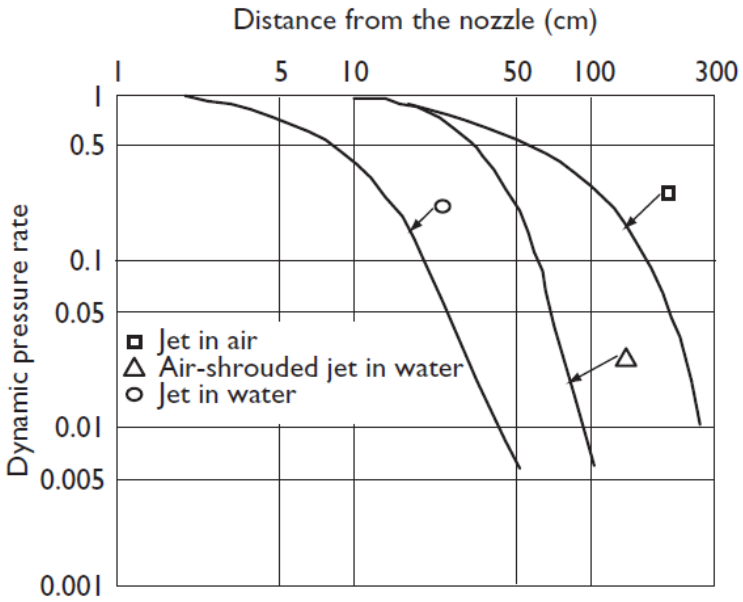


Figure 3.13: Relationships of dynamic pressure rates and distance from nozzle in various media. (adapted from Essler and Yoshida (2004))

process. Furthermore, since the achieved diameter is higher, the cement content is lower, contributing for a strength decreasing. On this system, beyond of the all parameters related to double system, it is also need to define the number and diameter of water nozzles as well as the pressure and flow rate of water.

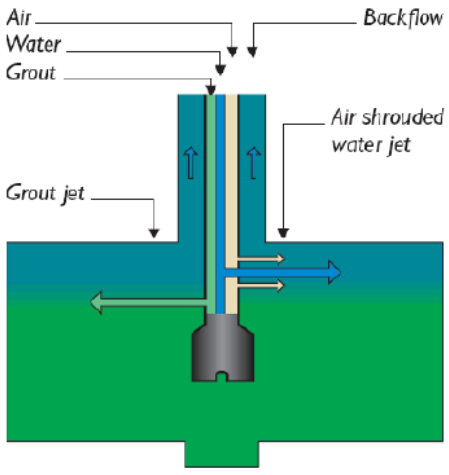


Figure 3.14: Triple fluid system schema (adapted from GmbH (2002))

3.3.4 Xjet system

Additionally to single, double and triple fluid systems, there is another concept, proposed in the late of 1980, providing an innovative progress for *JG* systems. This novel system,

termed Xjet system, also known as Cross-jet or collided jet, consists of a pair of intersecting air-shrouded water jets (see Figure 3.15) and is designed to cut a nominal 2 m to 2.5 m column diameter in any ground (Shibazaki et al., 1996). Figure 3.16 compares conceptually the profiles of conventional jetting and Xjet systems. Cementitious grout is injected below the erosion nozzles to displace and mix with the soil to create a high quality *soilcrete* column. When compared to the other systems, this concept allows control the eroding capability and thus achieve a better control of the column diameter regardless to the soil conditions (Welsh and Burke, 1997). Furthermore, the enhancement in this *in situ* mixing system results in more than 4 times the treated volume using the same equipment (Essler and Yoshida, 2004). Xjet is mostly applicable in variable weak ground such as soft clays and peat where overcutting of the design diameter can be a problem. This method is becoming popular in Japan and Europe due to its considerable technical and cost advantages. Xjet substantially replaced the *in situ* material, rather than mixing it with cement, thereby producing a very high quality *soilcrete*, can reduce up to 25% the spoil production and allow reduce the project schedule around 50%. The main drawback of this new concept is that requires sophisticated, more costly equipment and speciality contractors experienced in *JG* technology.

Figure 3.15: Xjet system

3.3.5 Jet grouting system selection

One of the first steps on a *JG* project is to choose the *JG* system to implement, which represents an important step in *JG* design. In this task, soil properties are within the

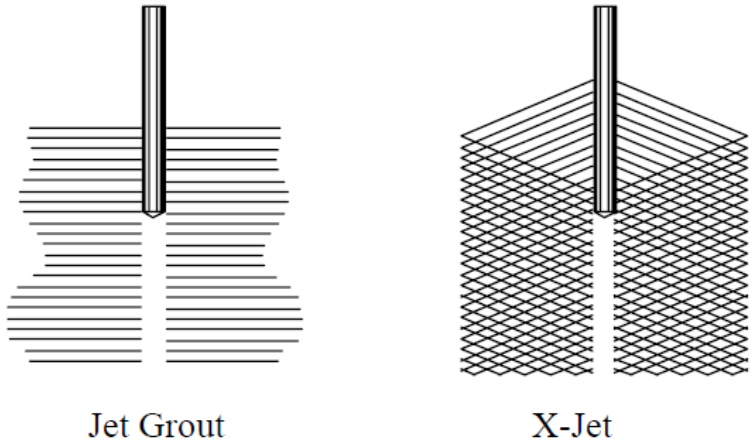


Figure 3.16: Comparison between the conceptual profiles of the conventional jetting and Xjet system

main factors to take into account. Moreover, and considering the cut energy associated to each jet system, it is expected that the highest diameters are achieved for triple fluid system. However, there are also other aspects that need to be considered, such as the economy and the project requirements. To help to accomplish such task, Figure 3.5 and particularly Figure 3.17 give an idea of the applicability of the three main *JG* systems in cohesive and granular soils.

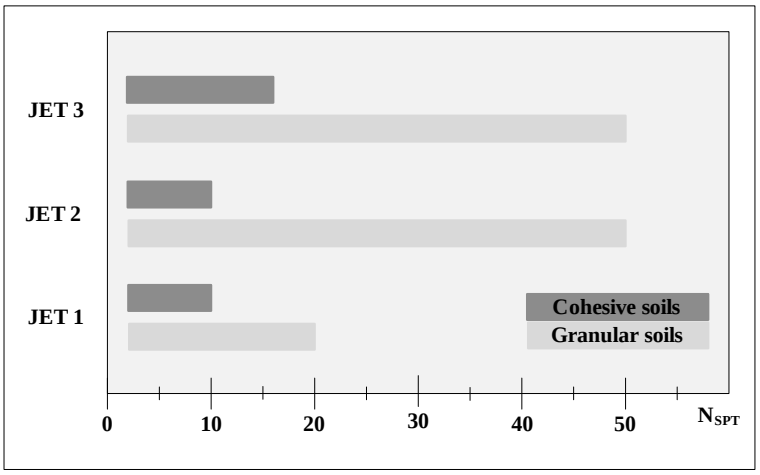


Figure 3.17: Applicability of the three main *JG* systems for cohesive and granular soils

Despite of the strong influence of the final column diameter on *JG* system choice (Figure 3.5), there are other factors that should also be taken into account. For instance, if there are concerns about air usage and loss of strength, the single system is the available alternative. Otherwise, the choice normally rests between the use of the double or triple system. Triple system generally offers less risk for blockage and potential structural or ground movement. Another issue that can affect the *JG* system choice is related with *JG*

equipment limitations. As previously mentioned Xjet requires sophisticated equipment and speciality contractors with high experience in *JG* technology that may not always be available.

3.4 Quality control and empirical approaches

3.4.1 Quality control

JG design is a task involving several steps. The choice of the most adequate *JG* system is one of the first tasks, followed by the definition of all parameters related with *JG* process (pressures, velocities, flow rate, etc.), as well as the definition of the cement slurry properties (water/cement ratio, cement type, etc.). Additionally, it is very important to perform a detailed soil site investigation in order to characterize it correctly. These aspects evidence the complexity behind *JG* design, where are involved several parameters. Moreover, it should be remainder that the soil is a very heterogeneous material, increasing the complexity of such task. Therefore, and keeping in mind that the actual approaches for *JG* design have important applicability limitations, it is fundamental to perform a rigorous quality control procedure throughout the entire process. Figure 3.18 summarize the main steps that should be followed during a *JG* work in order to ensure that the project requirements will be achieved.

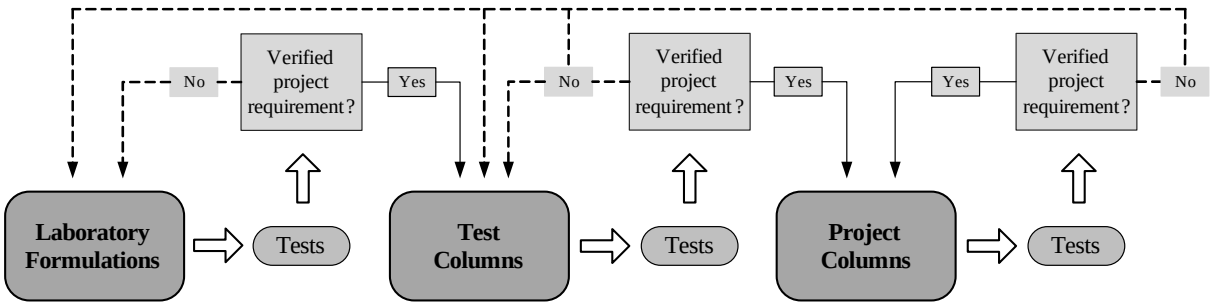


Figure 3.18: *JG* quality control procedure

In few words, this procedure start with the preparation of a set of laboratory formulations using the same materials that will be used during the soil improvement (e.g. the same soil, water and cement). This formulations allow the designer define some parameters related with soil-cement mixture, such as the water/cement ratio or the better choice for cement type. In addition, it is also assessed whether the *in situ* water can be used to prepare the cement slurry. Moreover, these formulations will give the first idea of the mechanical properties of the *soilcrete*.

The next step is to build some test columns in a representative place, normally close to the site where the project columns will be build. These test columns are built with the "same" parameters of the project columns (e.g. cement type, water, ejection pressure, withdrawal and rotation speed, etc.). Based on the tests results of some samples collected from this columns, normally measuring its strength and stiffness, is assessed whether some adjustments are necessary or not. Moreover, it is based on the column diameter of the these test columns that the decision about the construction of the project columns is supported. This means that the test column diameter is a key element on *JG* quality control assessment, giving indication whether to proceed or not for the project columns. Therefore, it is expected that the project columns will achieve the same diameter of test columns.

Finally, the project columns are constructed with all parameters previously defined. During the works, some samples are extracted periodically from this columns, in order to verify the project requirements, particularly in terms of strength and stiffness of *soilcrete*, and eventually procedure to some parameter refinements. More recently, additionally to core samples collected from *JG* columns after some days of curing time, some samples of *fresh material*[3] are also collected immediately after the columns construction, which are saved in a controlled environment and tested (unconfined compression tests) at different days time of cure. The diameter of the project columns, usually, is not verified, since it is assumed that the expected diameter is accomplished, considering the measurements performed over the test columns. This assumption is supported on the idea that for the same conditions (i.e. soil and jet parameters), the same results are always achieved, and it is contemplated by Eurocode 7 (CEN, 2004b).

Additionally to these main steps, during the jet grouting process some procedures are followed in order to guaranty that everything is in accordance with the design specifications. For example, the nozzles diameter are rigorously inspected before the soil improvement and periodically during the works, because this element has a preponderant influence of the jet energy, and consequently in the column diameter. Moreover, the specific gravity and viscosity of the injected cement slurry are also periodically checked. Furthermore, spoil is continuously observed in order to analysis its aspect and flow rate, avoiding underground overpressures that can damage neighbouring structures.

Another important aspect is drilling tolerance, particularly when overlapping of columns is crucial, namely on groundwater control works, base slabs or tunnel break-in or break-out. In these situations, omission or misplacement of a column can have the most serious effect on performance or safety. For this reason column position and the

[3] *fresh material* is the designation currently used to the material collected from the *JG* columns immediately after its construction.

expected diameter should be rigorously controlled.

All procedures and aspects above described are very important to guarantee the expected results (*soilcrete* physical and mechanical properties and *JG* column diameter). However, it is also fundamental be able to correctly define all *JG* parameters, according to the *JG* system chosen and soil properties, in order to achieve the project requirements. Moreover, understanding the effect of changing a given parameter, it is also crucial to efficiently correct undesired results, and so far there are almost no information.

The current state of knowledge about *JG* technology has shown that *JG* efficiency and effectiveness are strongly dependent of all *JG* parameters previously enumerated. Furthermore, it has been observed that such parameters present complex relationships between them, which has hindered the development of analytical models for *JG* design. Indeed, so far there just few mathematical expression, supported on traditional statistics analysis and using data from some *JG* works carried out in the last decades. As a result, they are very limited to the conditions under which were developed. Some of the most relevant analytical expressions that perform a relationship between *JG* parameters and *soilcrete* mechanical properties and column diameter are summarized on Sections 3.4.2 and 3.4.3 respectively. Moreover, many other authors have proposed some reference values and recommendations that can be seen as useful tips for *JG* design. According to Gazaway and Jasperse (1992) experience, grout pressure and flow rate, jet nozzle diameters, rotation and lift rate are some of the most important parameters that are involved in *JG* soil improvement. Van Impe et al. (2005) highlight the influence of the depth on *soilcrete* strength. Essler and Yoshida (2004) suggest that for lift speed should be adopted a 5 cm lift for up to 2 m of column diameter and a 10 cm lift for more than to 4 m of column diameter.

The core of the *JG* design is essentially supported in the know-how of each *JG* companies, which developed their own design tables. These tables perform a direct correlation between the expected results (normally the column diameter) and the *JG* parameters values that should be applied. However, although practical and simple these tables are very conservatives, compromising sometimes the economy of the soil improvement. Moreover, they also not explain the influence of each parameter in the final mixture. As above mentioned, these tables represent the know-how of each company and, for this reason are confidential. However, these tables are similar to that presented in Table 3.3, which summarizes the range of some of the most influential *JG* parameters currently used, according to the three main *JG* systems.

Table 3.3: *JG* parameters range (adapted from Carreto (2000)).

JG parameter		Single System	Double System	Triple System
Pressure	Grout (MPa)	20 to 60	20 to 55	0.5 to 27.6
	Air (MPa)	-	0.7 to 1.7	0.5 to 1.7
	Water (MPa)	PJ	PJ	0.5 to 27.6
Flow rate	Grout (l/min)	30 to 180	60 to 150	60 to 250
	Air (m^3/min)	-	1 to 9.8	0.33 to 6
	Water (l/min)	PJ	PJ	30 to 150
Nozzles diameter	Grout (mm)	1.2 to 5	2.4 to 3.4	2 to 8
	Water (mm)	PJ	PJ	1 to 3
Nozzles number	Grout	1 to 6	1 to 2	1
	Water	PJ	PJ	1 to 2
Water/Cement ratio		1:0.5 to 1:1.25	1:0.5 to 1:1.25	1:0.5 to 1:1.25
lift speed (m/min)		0.1 to 0.8	0.07 to 0.3	0.04 to 0.5
rotation rate		6 to 30	6 to 30	3 to 20

PJ - prejetting

3.4.2 Empirical approaches for mechanical properties prediction

For quality control purposes, the UCS of *soilcrete* is the mechanical properties currently used. In some situations, where structure's serviceability are required, deformability properties of the improved soils are also needed. For this reason, the use of reliable approaches for early predict the final mechanical properties of *soilcrete* is useful. Accordingly, several approaches (analytical models) have been proposed for its prediction. These expressions, normally supported on experimental studies, aims to establish a relationship between UCS and some of the most relevant *JG* parameter. Followed are summarized some of the most widely known empirical expressions with this purpose. It should be stressed that all mathematical expressions bellow presented are limited to its own development conditions. Therefore, it is recommended to consult the author works, for its full description and applicability.

Following the experiences of Nikbakhtan and Osanloo (2009), it was observed a good relationship between grout flow rate (FR, l/min) or grout pressure (P_{grout}, bar) with UCS (MPa) for *soilcrete* material. These two relationships are mathematically expressed by Equations 3.1 and 3.2 respectively and are graphically depicted on Figures 3.19 and 3.20. It should be underlined that, among other conditions, these two expression were adjusted to data collected from *JG* columns built with triple fluid system to improve low-strength

clay and fine soils.

$$UCS = 0.4376 \cdot e^{0.0079 \cdot FR} \tag{3.1}$$

$$UCS = 0.6334 \cdot e^{0.0937 \cdot P_{grout}} \tag{3.2}$$

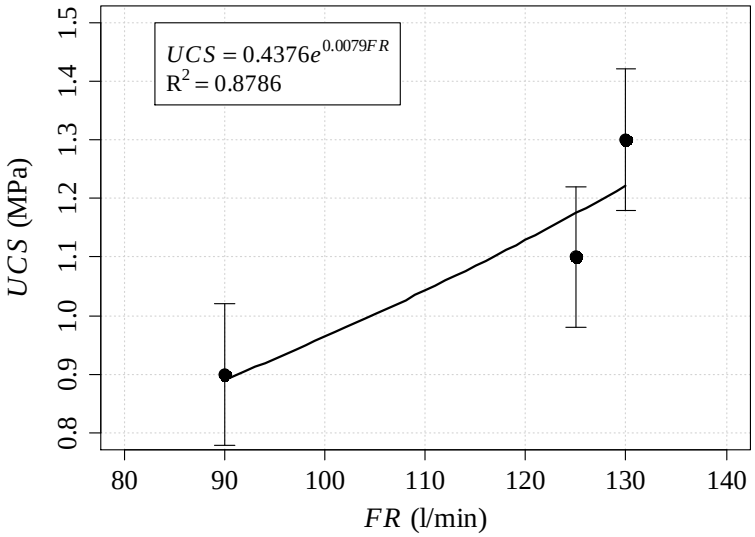


Figure 3.19: Relationship between UCS and FR for triple fluid system JG columns (adapted from Nikbakhtan and Osanloo (2009))

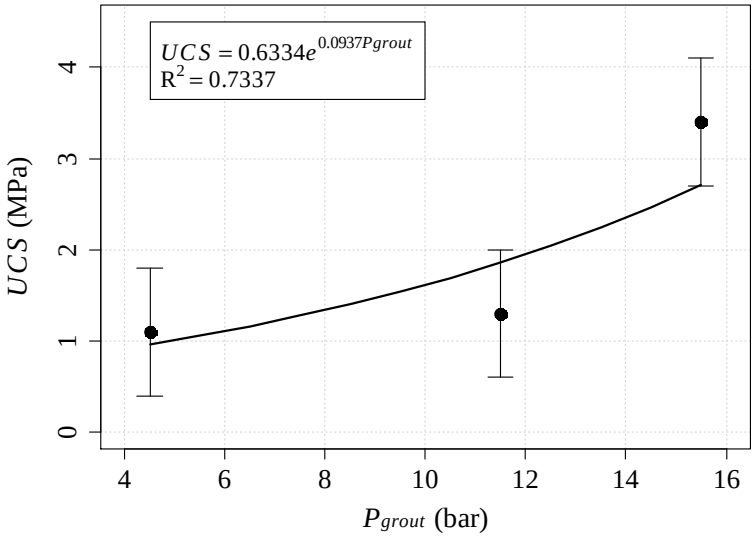


Figure 3.20: Relationship between UCS and P_{grout} for triple fluid system JG columns (adapted from Nikbakhtan and Osanloo (2009))

Later, Nikbakhtan in cooperation with Ahangari (Nikbakhtan and Ahangari, 2010) proposed a new expression to correlate UCS (MPa) with grout pressure (P_{grout}, bar) (see

Figure 3.21) as well as three others expressions that correlate Cement/Water ratio (C/W) (see Figure 3.22), lift speed (WS, cm/min) and rotation speed (rpm) with UCS (MPa) (see Figure 3.23):

$$UCS = 0.7131 \cdot e^{0.0523 \cdot P_{grout}} \tag{3.3}$$

$$UCS = 1.6141 \cdot e^{-0.0784 \cdot WS} \tag{3.4}$$

$$UCS = 1.6141 \cdot e^{-0.0784 \cdot rpm} \tag{3.5}$$

$$UCS = 2.4507 \cdot e^{0.2296 \cdot C/W} \tag{3.6}$$

Again, these expressions were developed based on data collected from triple fluid system

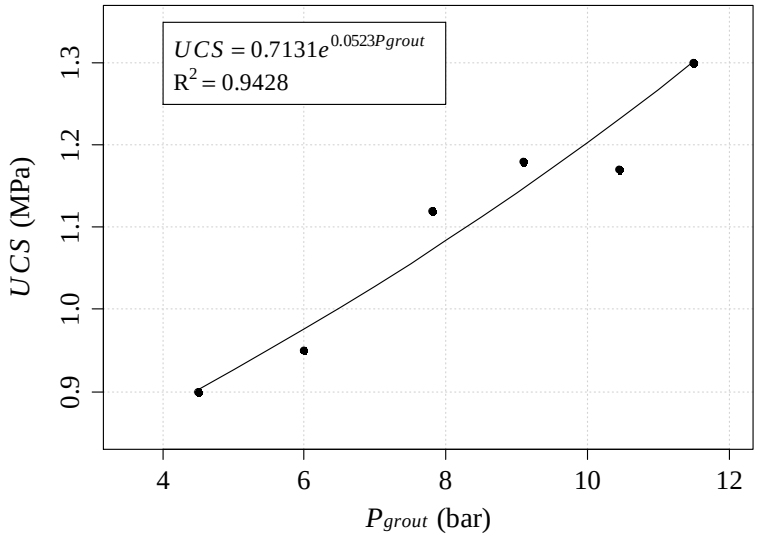


Figure 3.21: Relationship between P_{grout} and UCS (adapted from Nikbakhtan and Ahangari (2010))

JG columns built on fine grain soils, mainly from clay with low plastic property or plastic sediment.

Croce and Flora (1998) showed that UCS of JG mixtures can be successfully correlated ($R^2 = 0.70$), within a set of restrictions, with its dry unit weight (γ_d, $\mathrm{kg\,m^{-3}}$), following an linear law:

$$UCS = 2933 \cdot \gamma_d - 32427 \tag{3.7}$$

This relationship was obtained from a case study where pyroclastic soils were treated with single fluid system injecting a slurry of cement with a water/Cement ratio equal to 1 at 45 MPa and applying a lifting step of 40 mm.

Shen et al. (2010) proposed inferring the UCS of *soilcrete* based on the degree of

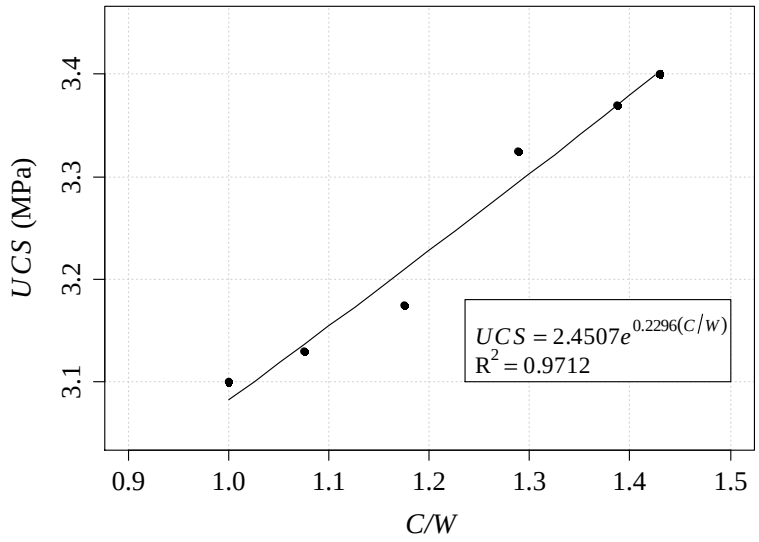


Figure 3.22: Relationship between C/W ratio and UCS (adapted from Nikbakhtan and Ahangari (2010))

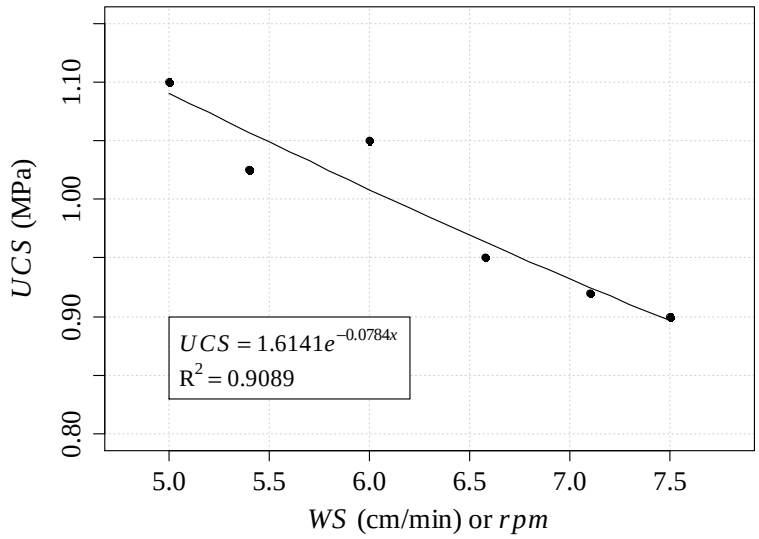


Figure 3.23: Relationship between WS and rpm and UCS (adapted from Nikbakhtan and Ahangari (2010))

mixing uniformity (D_u), which is determined using samples collected immediately after mixing. This coefficient is defined as:

$$D_u = \frac{N_1}{N_2} \times 100\% \tag{3.8}$$

Where N_1 is the number of collected samples and N_2 the number of samples with an pH value higher than the critical value. The average strength of *soilcrete* can be obtained by multiplying the degree of mixing uniformity with the strength from a standard laboratory mixing test.

There are also some other expressions that can give an idea of the strength values of soil-cement mixture, particularly for laboratory formulations. Narendra et al. (2006) proposed the following equation:

$$UCS = \frac{A}{B^{W_c/C}} \tag{3.9}$$

where A is a coefficient related to the type of clay, liquidity index and age of the mixture; W_c/C is the soil-water/cement ratio and B is an empirical constant that range from 1.22 to 1.24 and is independent of the type of clay.

Lee et al. (2005), based on previous works, particularly those developed by Gallavresi (1992), Kaushinger et al. (1992), Nagaraj et al. (1996), observed that for a given type of cement and cohesive soil, the UCS (kPa) can be correlated with water/cement ratio (W/C) and soil/cement ratio (S/C). Thus, after some experiments proposed the following relationship:

$$UCS = UCS_0 \cdot \frac{e^{m\cdot(S/C)}}{(W/C)^n} \tag{3.10}$$

where UCS_0 (kPa), m and n are experimentally fitted values.

Liu et al. (2008) introduced a simple index, the total water/Cement ratio (R_m), that present a good correlation with UCS of marine clay stabilized with cement. This index is defined as follows:

$$R_m = m_w/m_c \tag{3.11}$$

where m_w represents the weight of water in the mixed soil-cement, including the water in the original soil and the water in slurry cement; and m_c represents the weight of cement in dry state. Figure 3.24 shows the relation between the UCS (MPa) and the inverse of proposed index ($1/R_m$), presenting a good adjustment for a given age.

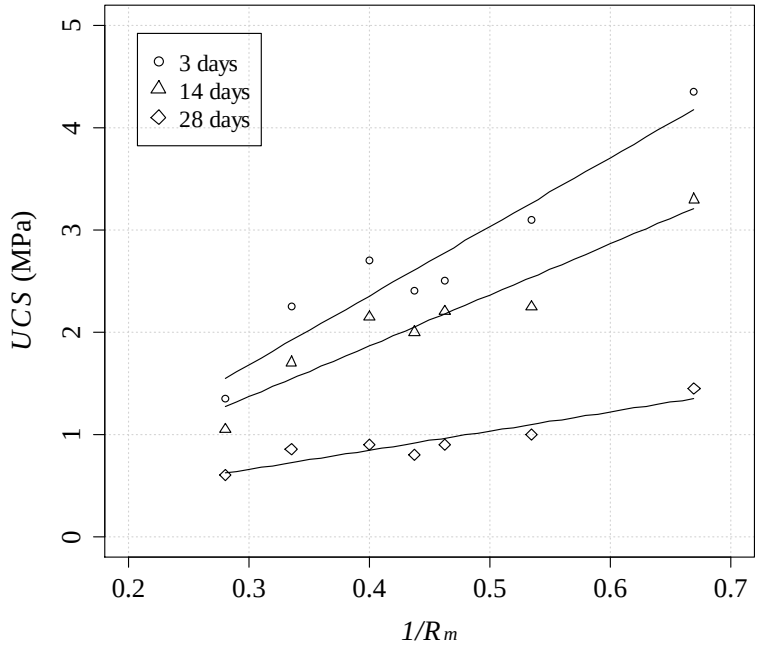


Figure 3.24: Relationship between UCS and total water-cement ratio (Liu et al. (2008))

Liu et al. (2008) also summarize some others mathematics expressions proposed by many others authors to predict UCS of soil-cement mixtures. According to Mitchell et al. (1974) there are the following relationship between UCS and curing time:

$$UCS_t = UCS_{t0} + K \cdot \log\left(t/t_0\right) \tag{3.12}$$

where UCS_t (kPa) is UCS at t days; UCS_{t0} is UCS (kPa) at t_0 days; $K = 480 \cdot C$ for granular soils and $K = 70 \cdot C$ for fine grain soil; C is cement content (% by mass.)

Nagaraj and Miura (1996) carried out unconfined compressive tests on four inland clays that had different liquid limits, and obtained the generalized relationship as follows:

$$UCS_t/UCS_{14} = a + b \cdot \ln\left(t\right) \tag{3.13}$$

where UCS_t is the UCS at age t (days); UCS_{14} is UCS the 14 days time of cure with initial water content as much as liquid limit of soil. It is reported that $a = -0.20$ and $b = 0.458$ for inland clays. Yamadera et al. (1997) further investigated the strength development with time of three different marine Ariake clays at their liquid limit. They found that $a = 0.190$ and $b = 0.299$.

Tan et al. (2002) have established an empirical relationship to predict the strength development based on cement content, water content and curing period. The strength developed at specific cement, water content and curing period is used as the reference

compressive strength for a given soil and the strength developed under other conditions for the same soil is normalized using the reference strength.

$$\frac{UCS_{soil1}}{UCS_{soil1}\left(a_w,\omega,t\right)} = \frac{UCS_{soil2}}{UCS_{u,soil2}\left(a_w,\omega,t\right)} = \frac{UCS_{soil3}}{UCS_{soil3}\left(a_w,\omega,t\right)} \tag{3.14}$$

where UCS_{soil1}, UCS_{soil2}, and UCS_{soil3} are the UCS of soil 1, soil 2, and soil 3, respectively; a_w is the ratio of cement to clay by weight both in their dry states (%); ω is the water content of soil; and t is the curing time.

Miura et al. (2001) and Horpibulsuk et al. (2003) used Abram's law as the basis for model development. With the concept explained in the literature of Horpibulsuk et al. (2003), the empirical model is developed as follows:

$$\frac{UCS_{(W_c/C)_1,t}}{UCS_{(W_c/C)_2,28}} = \begin{cases} 1.24^{\left[(W_c/C)_2-(W_c/C)_1\right]}\left(0.038+0.281\cdot\ln\left(t\right)\right) & \text{if LI} = 1.0 \sim 2.5 \\ 1.24^{\left[(W_c/C)_2-(W_c/C)_1\right]}\left(-0.216+0.342\cdot\ln\left(t\right)\right) & \text{if LI} > 2.5 \end{cases} \tag{3.15}$$

where t is the curing period in days; $UCS_{(W_c/C)_1,t}$ is the UCS at $(W_c/C)_1$ for the curing period of t days; W_c is the water content; C is the cement content; $UCS_{(W_c/C)_2,28}$ is the UCS at $(W_c/C)_2$ for the reference curing period of 28 days; LI is the liquidity index.

Lorenzo and Bergado (2004) found that the ratio between after-curing void ratio (e_{ot}) and cement content (C) is sufficient to characterize the strength of cement-admixed clay at high water contents. The following relationship has been derived to describe the UCS of any cement-admixed clay:

$$UCS = A\cdot p_a\cdot e^{B\cdot(e_{ot}/C)} \tag{3.16}$$

where A and B are dimensionless constants and p_a is atmospheric pressure. Based on the results presented, for soft Bangkok clay mixed with Type I Portland cement, the constants are $A = 10.33$ and $B = -0.046$. The constant A is affected by the type of admixture (or type of cement), while the constant B is affected by the type and mineralogy of the original clay. Thus, the empirical relationship of after-curing void ratio, e_{ot}, which is related to clay water content, cement content, and curing time, is put forward.

Additionally to all empirically expression previously enumerated, there are other approaches used on different areas that can be adapted to predict soil-cement mixtures mechanical properties, namely of JG material. Two of these approaches are those contemplated on *EC2* (CEN, 2004a) and *MC90* (CEB-FIP, 1991) regulations, currently applied to predict mechanical properties (strength and stiffness) of concrete. These to ap-

proaches will be adapted and tested in the present research work to predict both strength and stiffness of *JG* material, particularly for *JGLF*.

EC2 proposes the following mathematical expression to estimate concrete strength over time:

$$f_{cm}(t) = e^{\left(s\cdot\left[1-\left(\frac{28}{t}\right)^{a}\right]\right)} \cdot f_{cm} \tag{3.17}$$

where $f_{cm}(t)$ is the strength at age t; f_{cm} is 28 days strength of the mixture; s is a coefficient related with cement type and t is the age of the mixture. The coefficient a, taken equal to $a = 1/2$ for concrete will be adapted to *JG* mixtures.

For stiffness estimation, *EC2* proposes a similar expression, defined as follows:

$$E_{cm}(t) = \left(e^{\left(s\cdot\left[1-\left(\frac{28}{t}\right)^{a}\right]\right)}\right)^{b} \cdot E_{cm} \tag{3.18}$$

where $E_{cm}(t)$ is the stiffness at age t; E_{cm} is 28 days stiffness of the mixture; s is a coefficient related with cement type, t is the age of the mixture and a and b are coefficients to be adjusted using *JG* data.

Based on *MC90* regulation, concrete stiffness can be estimated according to the following equation:

$$E_{ci}(t) = \left(e^{\left(s\cdot\left[1-\left(\frac{28}{t}\right)^{a}\right]\right)}\right)^{b} \cdot \alpha_E \cdot E_{c0} \cdot (f_{cm}/f_{cm0})^{c} \tag{3.19}$$

where $E_{ci}(t)$ is the stiffness at age t; E_{cm} is 28 days stiffness of the mixture; s is a coefficient related with cement type, t is the age of the mixture; α_E is a coefficient that depends on the type of aggregate (for soil clay, a 0.99 value can be adopted); $f_{cm0} = 10\,\text{MPa}$; f_{cm} is 28 days stiffness of the mixture; E_{c0} was determined for each formulation based on 28 days stiffness and a, b and c are coefficients to be adjusted. For strength development through the time, the proposed model by *MC90* is equal to those present by *EC2* (see Equation 3.17).

3.4.3 Empirical approaches for diameter prediction

JG column diameter prediction is one of the most important issues in *JG* technology design. Particularly on groundwater control works, It is fundamental that there is no free space between columns, i.e., that all columns intersect with the adjacent. Once again, in order to guarantee such conditions it is necessary to dispose of design tools able to accurately predict *JG* column diameter. Since the begin of *JG* technology, several attempts were made in order to develop a mathematical model able to predict *JG* column

diameter as accurate as possible, under different soil conditions and *JG* systems.

One of the most interesting approaches so far developed is the proposed by Modoni et al. (2006). However, in spite of its strong theoretical support and applicability to different soil types, also presents important limitation, namely its restriction to *JG* single fluid system. Conceptually, the proposed model, approach the problem of column diameter as the distance achieved by the jet grout. This mean that the column diameter will be equal to distance travelled by jet until its energy is null, keeping in mind that the jet energy is maximum immediately after the nozzle and decrease during its travel throughout the soil.

Following the proposed approach, the jet propagation is performed in two steps. The first one correspond to the jet propagation across the space included between the injection nozzles and the intact soil, which is modelled based on the theory of submerged flows. The second step coincide with the jet propagation within the soil. Here, different interactions are assumed for gravels, sands and clays. In the case of gravels, grout seepage is considered to be the most relevant mechanism. For sandy soils, the injected fluid is assumed to penetrate, for a limited extent, into the soil skeleton, producing a considerable increment of the pore pressures and a corresponding reduction of the grain-to-grain contact forces. The removal of the soil particles is then triggered by the dragging action of the fluid threads, and the analysis is developed under drained conditions. For clayey soils, the jet action is considered as a load imposed on the jet-soil interface, and the erosion process is modelled as an evolving sequence of undrained failures.

For granular (gravels and sands) and cohesive (clayey) soils Modoni et al. (2006) proposed the Equations 3.20a and 3.20b respectively, to predict the maximum radius (theoretical, i.e., for high jetting time) of single fluid *JG* columns.

$$\text{Granular:} \quad R = \frac{2 \cdot \nu_0 \cdot \Lambda \cdot C \cdot D_{grout}}{\sqrt{\frac{\Omega_s \cdot g \cdot N}{\gamma_f} \cdot \frac{c' + \sigma_z \cdot \tan(\phi')}{1 + \Omega_s \cdot [\tan(\phi')/2]}}} \tag{3.20a}$$

$$\text{Cohesive:} \quad R = \frac{2 \cdot \Lambda \cdot C \cdot D_{grout} \cdot \nu_0}{\sqrt{\frac{\Omega_c \cdot g \cdot N \cdot c_u}{\gamma_f}}} \tag{3.20b}$$

In the above equations ν_0 is the initial speed of the jet threads (immediately after the nozzle); Λ is a coefficient (experimentally quantified) related with the nozzle shape that affect the attenuation of the fluid velocity along the jet axis (x); $C = \sqrt{\xi}/2$, where $\xi = \nu_x/\nu_{xmax}$ which represent a fraction of the maximum velocity of the jet at distance x from the nozzle (ν_x is the mean velocity of the jet at distance x and ν_{xmax} represent the respective maximum velocity); D_{grout} is the nozzles diameter; Ω_s and Ω_c are dimensionless

parameter accounting for energy dissipation of the injected fluid on granular and cohesive soils respectively; g is the gravitational acceleration; N represent the turbulent kinematic viscosity ration of injected fluid and water ($N = \epsilon_f/\epsilon_w$); γ_f represent the unit weight of the injected fluid; c' and ϕ' are respectively the effective cohesion and friction angle of the soil; c_u is the undrained soil cohesion; σ_z is the initial vertical overburden stress.

It should be stressed that Equations 3.20a and 3.20b allow predict the maximum theoretical column radius for JG single fluid system and for a reference time of jetting (t^*) that allows obtain such radius. One of the main contributions of the works developed by Modoni et al. (2006) was to show the dependency of JG column radius on the fluid velocity, number and diameter of the nozzles, as well as monitor lifting speed. Particularly, for clayey soils, the proposed approach shows that JG it is only effective if applied high flow rates and low withdrawal speeds.

Three years latter, Carletto (2009) proposed a simplification to Modoni et al. (2006) method. After observe that the JG column should consider both the effect of jet energy and soil resistance, he try to simplify the two equations proposed by Modoni et al. (2006) for granular and cohesive soils (Equations 3.20a and 3.20b). One of the first guidelines for its development is related with the fact that the soil resistance should be considered by its shear strength (under drained conditions for granular soils and under undrained conditions for clayey soils). Therefore, one of the main tasks is to quantify the shear strength for the different soil types and conditions. On the other hand, it was expected that the entire effect of the jet action could ever be considered by a single parameter (J). This parameter is then defined as the product between the proportionality relationship observed on Modoni et al. (2006) equations, i.e., between the maximum theoretical JG radius diameter (R) and the reference time that allow obtain such theoretical diameter (t^*) (for a detailed description of these considerations is recommended to consult Carletto (2009)). The jet effect can so be mathematically expressed as follows:

$$J = \frac{\nu_0 \cdot D_{grout} \cdot \sqrt{\gamma_f}}{\sqrt{N}} \cdot \left(\frac{M \cdot \sqrt{N}}{WS}\right)^{\chi} = \nu_0 \cdot D_{grout} \cdot \left(\frac{M}{WS}\right)^{\chi} \cdot N^{0.5\cdot(\chi-1)} \cdot \gamma_f \qquad (3.21)$$

where M is the number of nozzles; WS represent the lifting speed of the rods; and χ is a calibration parameter that depends of the soil type (granular or clayey), being quantified through numerical simulations. This expression can then be simplified by replacing χ by the values obtained by Carletto (2009) for granular and cohesive soils as well as taken into account that both N and γ_f are dependent of water cement ratio (W/C). Hence, the

following expressions are obtained:

$$\text{Granular:} \quad J_s = \nu_0 \cdot D_{grout} \cdot \left(\frac{M}{WS}\right)^{0.50} \cdot \left(1.16(W/C)^2 - 2.06(W/C) + 3.55\right) \tag{3.22a}$$

$$\text{Cohesive:} \quad J_c = \nu_0 \cdot D_{grout} \cdot \left(\frac{M}{WS}\right)^{0.77} \cdot \left(0.72(W/C)^2 - 1.52(W/C) + 4.07\right) \tag{3.22b}$$

The last step on Carletto (2009) approach is to combine the effect of all JG parameters, which is represented by J parameter, with shear strength of the soil. From this interaction, it was observed that the column diameter follows an power law ($S \cdot J^{\beta}$), where S is the function of shear strength and β is the power coefficient. After quantified S function and β coefficient, the following equations are proposed by Carletto (2009) to predict single JG column diameter for granular and clayey soils, using a simplified approach:

$$\text{Granular:} \quad D = 0.58 \cdot s^{-0.40} \cdot J_s^{0.67} \tag{3.23a}$$

$$\text{Cohesive:} \quad D = 0.11 \cdot s_u^{-0.26} \cdot J_c^{0.55} \tag{3.23b}$$

where s and s_u are respectively the drained shear strength for granular soils and undrained shear strength for clayey soils.

Additionally to these two main approaches (Modoni et al., 2006; Carletto, 2009) characterized by a strong theoretical explanation, there are other simplest approximations used to predict JG column diameter and to support JG system selection. Some of these approximations are described herewith.

Kanematsu (1980) proposed that JG column diameter should be around 300 times the diameter nozzle (both in meters), without consider any soil proprieties or JG parameter. In turn, Langbehn (1986), considering the grout pressure, proposed a range for JG diameter in function on soil type (soft clays or soft compact sands), which are depicted on Figure 3.25.

Nikbakhtan and Ahangari (2010), based on their works where three JG columns were constructed with triple fluid system under fine grained soil (clay with low plastic property or plastic sediment) with different JG parameters, also correlated JG column diameter with grout pressure according to the law depicted in Figure 3.26. Furthermore, a good relationship was observed between JG column and lifting or rotating speed as shown in Figure 3.27, as well as with Cement/Water ratio plotted Figure 3.28.

These four relationships show that eroding distance increase with grout pressure and Cement/Water ratio, and decrease with lifting and rotating speed, according an exponential law.

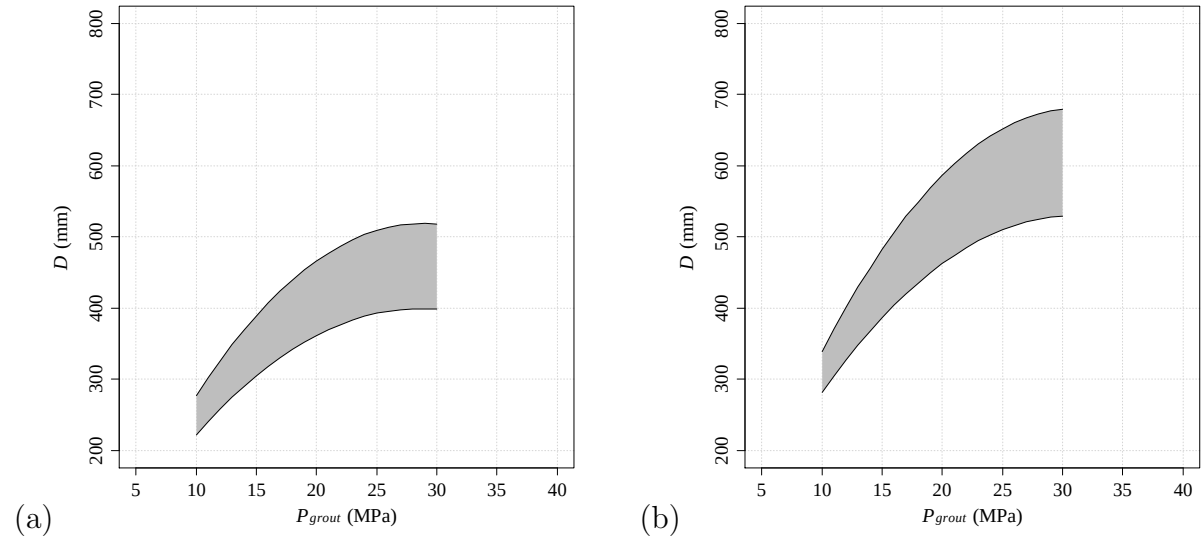


Figure 3.25: Relationship between P_{grout} and D: a) soft clay, b) sandy soil medium dense (adapted from Langbehn (1986))

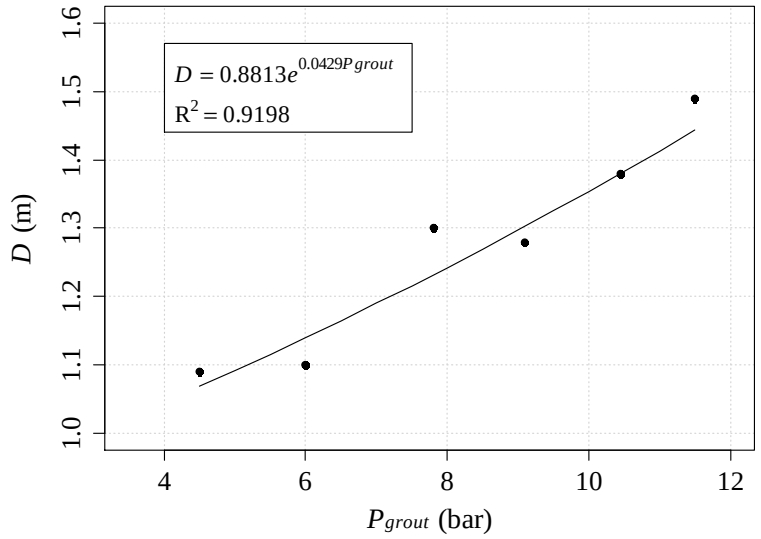


Figure 3.26: Relationship between P_{grout} and D (adapted from Nikbakhtan and Ahangari (2010))

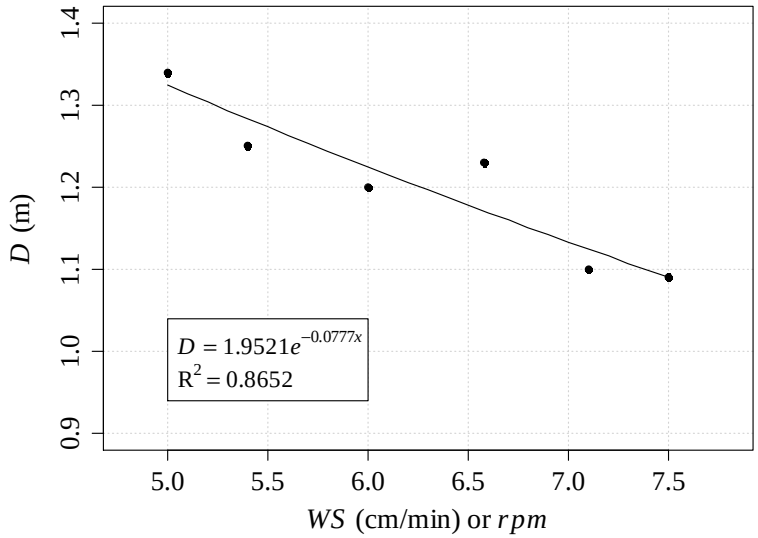


Figure 3.27: Relationship between WS and rpm and D (adapted from Nikbakhtan and Ahangari (2010))

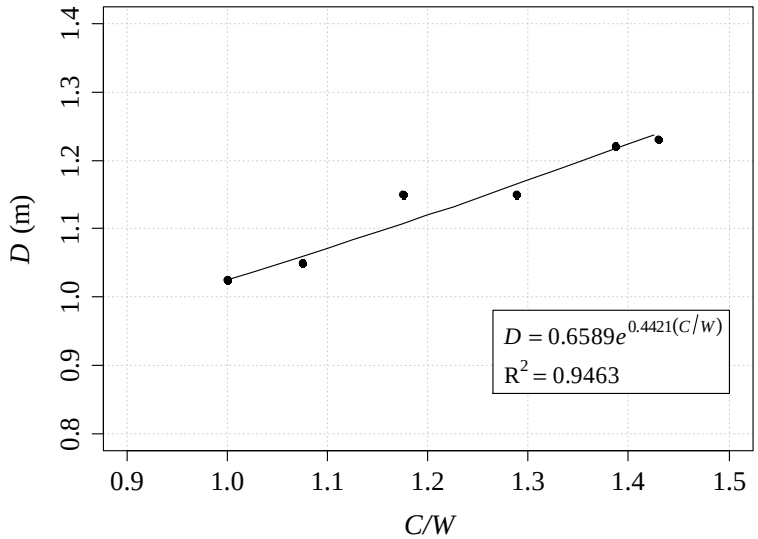


Figure 3.28: Relationship between C/W ratio and D (adapted from Nikbakhtan and Ahangari (2010))

As Modoni et al. (2006), also Wang et al. (2012) followed the turbulent kinematic flow theory to support his approach for *JG* column diameter prediction. Thus, based on such theory, the following expression was proposed, which can be applied to most soil types:

$$R = \frac{d_0}{2} + b \cdot \frac{4FR}{M\pi D_{grout}\sqrt{UCS/p_a}} \tag{3.24}$$

where R is the radius of the *JG* column; d_0 is the rod diameter; FR is the flow rate of the fluid injected; M the number of nozzle of the rod; D_{grout} the nozzle diameter; UCS is the unconfined compressive strength of the *soilcrete*; p_a the atmospheric pressure; and b is a parameter related to the soil characteristics, which can represent the eroding ability of jet fluid on different soils. Following the results of Wang et al. (2012), b should range between 1.2 to 2.0 for very soft clay, 0.75 to 1.4 for clayed silt and 0.25 to 0.75 for sand. Observing Equation 3.24, it is evident the *JG* column diameter dependency of the UCS of the improved mixture. This means that such approach can only be used after the soil improvement and after perform unconfined compression tests in order to quantify *soilcrete* strength.

Motivated by the need of obtain *JG* columns up to 5 meters in diameter, and after carried out an experimental program where *JG* triple fluid system was applied to built the columns, Shibazaki and Yoshida (1997) proposed an empirical formula to predict the cutting distance, defined as:

$$R = \left(4.95 \cdot K \cdot P_{grout}^{-1.4} \cdot FR^{-1.6} \cdot N^{-0.2} \cdot v_n^{-0.3}\right) - 0.7 \tag{3.25}$$

where R is the column radius (m); K is a constant related with jetting liquid (2.5 for cement slurry and 1.0 for water); P_{grout} correspond to the discharge pressure ($\mathrm{kg\,cm^{-2}}$); FR is the flow rate (l/min); N represent the number of passes; v_n is the tangential velocity at a nozzle outlet ($\mathrm{m\,s^{-1}}$). Since this expression was developed based on a small experimental program, a special carefully should be taken to the range of each parameter. Thus, P_{grout} is limited to 200∼500 kgf/cm^2, FR to 70∼300 l/min, N to 1∼20 and v_n to 0.1∼0.2 m/s.

Another proposed approach to estimate *JG* column diameter is applying the mathematical expression developed by Croce and Flora (1998). Based on his works, where single fluid system was applied to treat pyroclastic soils, the following equation was proposed to predict *JG* column diameter (D):

$$D = 2 \cdot \left\{ \frac{\alpha \cdot V_j}{\pi \cdot [1 - (1 - \beta) \cdot (1 - n)]} \right\}^{0.5} \tag{3.26}$$

In this equation V_j represent the injected grout volume per unit length and n is the initial soil porosity. The coefficient α and β are related with the percentage of mortar retained by the subsoil and the percentage of soil removed by jet action respectively.

3.5 Conclusions

The literature review presented in the current chapter focuses on two main aspects related to *JG* technology. The first emphasises the high versatility of such technology and its importance in geotechnical works as a soft soil improvement method. It illustrates the diversity of applications under different soil characteristics and logistical conditions of *JG* technology, as well as its economic advantages when compared with other soil improvement methods. On the other hand, the main drawback of *JG* technology is related to the actual approaches for *JG* design. As presented, the actual approaches for such purposes are scarce and have important applicability limitations. In some cases, such approaches are only valid for particular soil conditions and for a given jet system. In the case of *JG* column diameter, there are some theoretical approaches, but they are also limited to a particular jet system (single fluid system). Indeed, JG companies' experience remains the principal source of knowledge for *JG* design, which is then validated through the construction of some test columns and laboratory tests over extracted samples.

Based on the performed literature review, it was observed that the grout flow rate, grout pressure, water/cement ratio, withdrawal and rotation speeds and dry unit weight are some of the most commonly used variables for the prediction of mechanical properties. For *JG* column diameter, the number and diameter of the nozzles are also usually considered. Moreover, the importance of a detailed soil characterisation (or at least a distinction between granular and cohesive soils) for a reliable *JG* technology design was stressed. In addition, the complexity of *JG* column design caused by the high number of variables involved and nonlinear relationships between *JG* mechanical properties or column diameter and its contributing factors was also underlined.

So far, several attempts were performed toward the development of more reliable approaches for *JG* design, which were almost supported by traditional statistical analysis. Until now, however, no proposed approaches were completely successful. Therefore, this long path needs to be continued to encourage the use of new and advanced tools to solve this complex problem. This work addresses this step and aims to develop new approaches for *JG* mechanical properties and column diameter design while contemplating different soil types and jet systems.

This page was intentionally left blank.

Chapter 4

Jet grouting database characterisation

4.1 Introduction

Information can be seen as a synonym of knowledge, representing a key issue on any business. This is what happen on *JG* technology, where knowledge is fundamental for a successful *JG* column design because so far, reliable methods for such task are scarce. Particularly in small *JG* works, such knowledge is still more preponderant due to the higher budget limitations in these situations. As a result the number of field and laboratory experimental tests for soil characterization are reduced to only a few number.

For this reason it is very important to collect and store all information related to each *JG* work. Such information is normally related with three main aspects, as shown in Figure 4.1: soil and materials characterization and *JG* parameters. Moreover, information concerning to mechanical properties of both laboratory formulations and *soilcrete* are also collected through laboratory tests, as well as columns geometry (diameter). *JGLF* are soil-cement mixtures prepared in the laboratory, using the same materials (e.g. soil and cement) of the *JG* columns, with the purpose to guide early stages of *JG* process. This formulations, almost not performed on small *JG* works due to budgets limitations, are tested at different ages giving an idea of the mechanical properties of the *soilcrete*, and allowing to define some *JG* parameters such as cement type, cement content or Water/Cement ratio.

Soilcrete mechanical properties measurement is a key aspect on *JG* technology quality control. Through a simple and not so expensive procedure, it is possible to quantify *soilcrete* strength and stiffness, by performing laboratory tests over some samples directly collected from the *JG* columns. Additionally to the mechanical properties of *soilcrete*, the measurement of the *JG* column diameter at different depths, particularly in the test columns, it is also fundamental. Indeed, the diameter of the test columns represents the

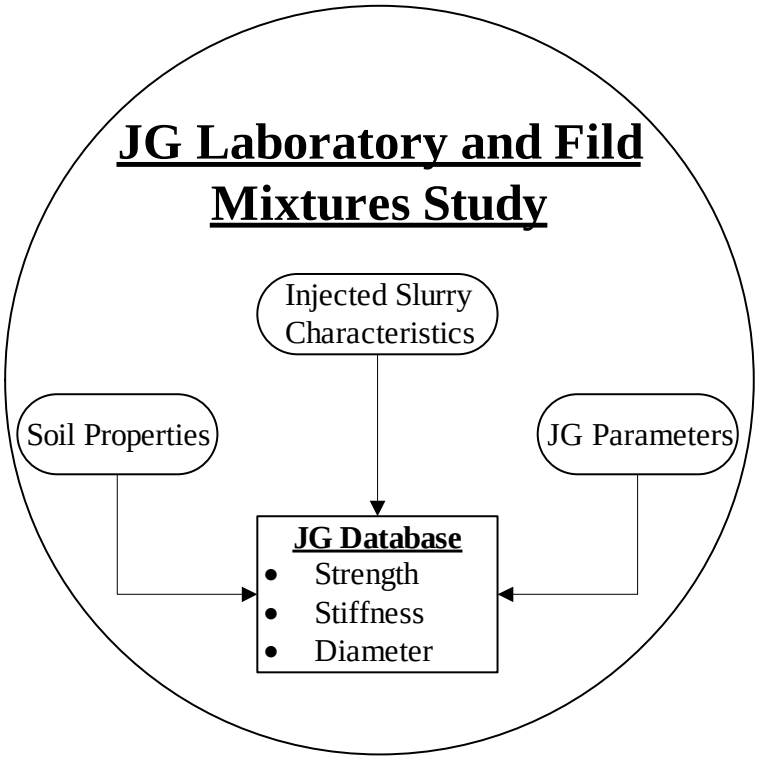


Figure 4.1: Structure of the compiled database

main decision criterion used to assess the soil improvement quality, allowing to start the construction of the project columns. These are the three key element (strength, stiffness and column diameter) usually taken for a quantitatively assessment of *JG* soil improvement quality. Moreover, particularly in big-scale *JG* projects, are also prepared and tested some laboratory formulations that supply important informations related with the materials used for its preparation, since usually a detailed characterization is performed for all used materials.

Another fundamental aspect for any *JG* project is a geotechnical characterization of the soil to be improved, although sometimes this characterization is minimal. Moreover, several *JG* parameters related with the soil improvement process (e.g. grout pressure, rotation speed) are continually monitored in the record station (see Figure 3.7). At the end, for each *JG* project, it is stored information related with *JG* column diameter and *soilcrete* mechanical properties, the materials used in the soil improvement (particularly soil, cement and water characterization), as well as about the *JG* parameters applied during the soil improvement. Now, the challenge is to cross and deeply analyse all this information in order to find patterns and useful tendencies for future *JG* projects.

So far, all these informations are essentially used for quality control procedures and to guide the designer to make decisions, i.e., to verify if the project requirements are being

satisfied, and to help the engineering to choice the best solution able to correct undesired behaviours. However, it's known that all these information/data handle useful knowledge that can be very useful in future *JG* projects. Therefore, the first step is organise them in a structured database and then explored them, particularly through the application of *DM* techniques and guided by a panel of expertises. Moreover, and keeping in mind that due to economic constrains, not always a convenient soil and materials characterization is performed, such analysis is even more important toward a better efficiency of *JG* technology.

In the following sections, the two main databases used in the present research work for *JGLF* and *soilcrete* samples study will be presented and characterized.

4.2 Laboratory data

The study of *JGLF* mechanical properties was supported on a database compiled with data taken from a large experimental program carried out at University of Minho. This program aimed to analyse the influence of different parameters in mechanical properties of *JG* laboratory mixtures (Gomes Correia et al., 2009). Hence, during the preparation of the laboratory formulations, a special care was taken to record all information potentially usefully, such as those related to the soil properties, cement type, water quality, cement dosages, soil and water content of each formulation, etc. A full list of all variables considered for the study of *JGLF* is further presented. These particular circumstances, i.e., the fact that all information used in this study came from a singular source, represent a key factor on the quality and confidence of the research results, since this provides greater uniformity in the procedures adopted for *JGLF* preparation during the entire experimental program.

This experimental program contemplated the preparations of *JGLF* for seven different *JG* projects. This means that seven different ground types (soil types) were concerned, which will be further characterized. After mixing and prepare several samples for each formulation, each one was tested in order to measure either its UCS, stiffness or both. Figure 4.2 shows the different moduli that can be defined in a nonlinear stress strain relationship, which are determined through a unconfined compression test with a sample strain instrumentation (Gomes Correia et al., 2009), measuring the local deformations of the tested sample with LDTs (local deformation transducers) and LVDTs (linear variable differential transformers), as shown in Figure 4.3.

Table 4.1 summarizes the number of records and formulations used in the study of strength and stiffness of *JGLF*. This table shows that for the study of the *tangent de-*

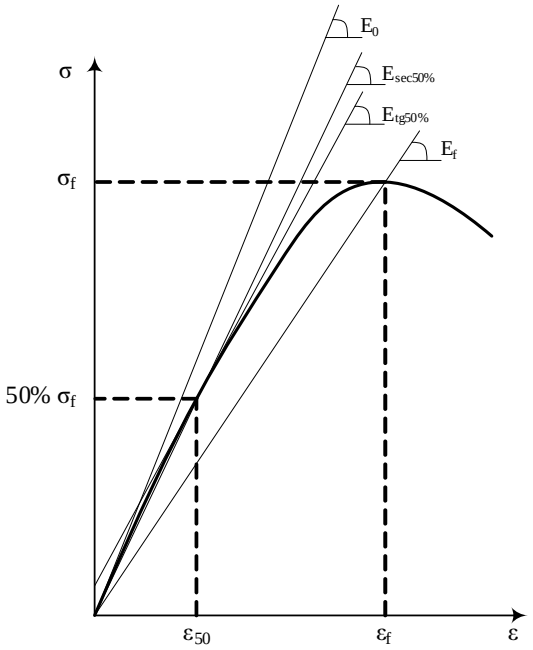


Figure 4.2: Illustration of the different deformability properties (i.e. moduli) that can be defined in a unconfined compressed test(x-axis denotes the strain ε and y-axis the stress σ)

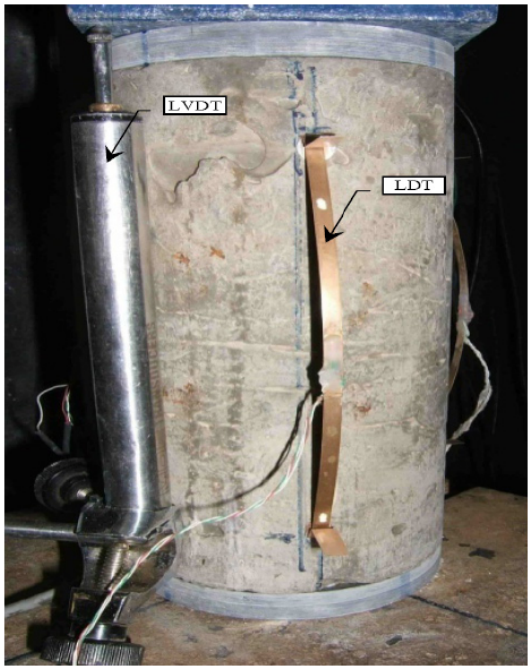


Figure 4.3: Specimen of the laboratory mixture instrumented with LDT and LVDT (Gomes Correia et al., 2009)

formability modulus at 50% of the maximum applied stress ($E_{tg50\%}$), *secant deformability modulus at 50% of the maximum applied stress* ($E_{sec50\%}$) and *maximum secant deformability modulus* (E_{max}) there are only 48 records available, which can be seen as a rather small number for *DM* purposes. Yet, it should be stressed that the acquisition of each data example requires considerable costs and amount of time, as well as demanding laboratory work.

Table 4.1: Number of records and formulations used in both mechanical properties study of *JGLF*

	UCS	E_0	$E_{tg50\%}$	$E_{sec50\%}$	E_{max}
Number of records	175	188	48	48	48
Number of formulations	35	9	8	8	8

For *JGLF* mechanical proprieties study a total of 24 variables were considered, for which the histograms are presented in Appendix A.1, following listed:

- W/C - Water/Cement ratio
- CT - cement type
- SCC - strength cement class
- s - coefficient related with cement type
- kg/m^3 - kilograms of cement by cubic meter of soil
- t (days) - age of the mixture
- ρ ($\mathrm{kg\,m^{-3}}$) - natural density of the mixture
- ω (%) - water content of the mixture
- ρ_d ($\mathrm{kg\,m^{-3}}$) - dry density of the mixture
- $1/\rho_d$ ($\mathrm{m^3\,kg^{-1}}$) - inverse of the dry density of the mixture
- $\%Soil$ - soil content in the mixture
- $\%Cement$ - cement content in the mixture
- $\gamma_{s.mixt}$ ($\mathrm{kg\,m^{-3}}$) - unit weight of the mixture

- e - void ratio of the mixture
- n - mixture porosity
- $1/n$ - inverse of the mixture porosity
- ω_{sat} (%) - saturated water content
- S_ω - degree of saturation
- C_{iv} - volumetric content of cement
- $n/(C_{iv})^d$ - relation between mixture porosity and volumetric content of cement
- $\%Sand$ - percentage of sand in the natural soil
- $\%Silt$ - percentage of silt in the natural soil
- $\%Clay$ - percentage of clay in the natural soil
- $\%OM$ - percentage of organic matter in the natural soil
- UCS (MPa) - uniaxial compressive strength
- E_0 (GPa) - elastic Young's modulus
- $Etg50\%$ (GPa) - tangent deformability modulus at 50% of the maximum applied stress
- $Esec50\%$ (GPa) - secant deformability modulus at 50% of the maximum applied stress
- $Emax$ (GPa) - maximum secant deformability modulus

Among all considered variables, just three of them are discrete: CT (1, 2 and 4), SCC (32.5R and 42,5R) and s (0.2 and 0.25). Moreover, since, not all of them were directly measured from the samples, the mathematical expressions used for its calculation are presented in Appendix B.

One of the first steps on data analysis is to describe the data with a simple parameter, which can be provided by statistics. On Tables A.1, A.2 and A.3 of Appendix A.1 are presented the main statistics, i.e., *maximum*, *minimum*, *mean* and *standard deviation*, of each input and output variables considered in the study of $JGLF$ mechanical properties. Figures 4.4 and 4.5 show the histograms of UCS and *elastic Young's modulus* (E_0), $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} respectively for $JGLF$. It is interesting to observe that the shape of

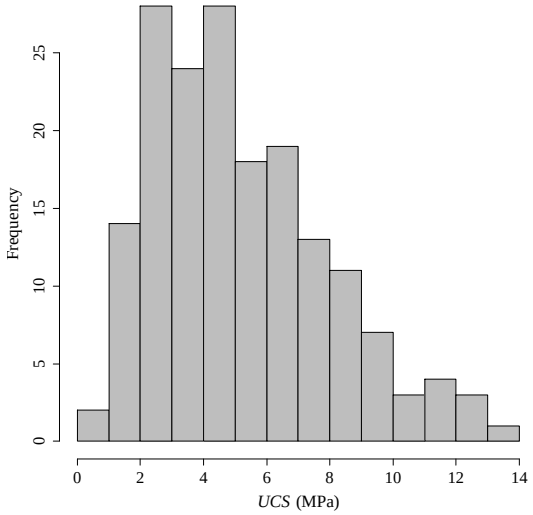


Figure 4.4: Histogram of UCS in the study of $JGLF$

the UCS histogram shown in Figures 4.4 is similar to that found in the literature (see Figure 3.4b), although they are related to different types of materials (i.e. laboratory and field JG mixtures respectively).

For the analysis of multidimensional data, it is also important to verify if two variables x_i and x_j are statistically dependent. For example, the covariance (defined in Equation 4.1) gives information about this issue. In this sum, the summand returns a positive entry for the pth data vector exactly when the deviations of the ith and jth components from the average both have the same sign. If they have different signs, then the entry is negative.

$$\sigma_{ij} = \frac{1}{N-1} \cdot \sum_{p=1}^{N} \left(x_i^p - \bar{x_i}\right) \left(x_j^p - \bar{x_j}\right) \tag{4.1}$$

However, the covariance also depends on the absolute value of the variables, which makes comparison of the values difficult. To compare the degree of dependence in the case of multiple variables, it is preferable to calculate the correlation coefficient (Ertel, 2009):

$$K_{ij} = \frac{\sigma_{ij}}{S_i \cdot S_j} \tag{4.2}$$

for two values x_i and x_j, which is nothing but a normalized covariance. The matrix K of all correlation coefficients contains values between –1 and 1, is symmetric, and all of its diagonal elements have the value 1. In order to facilitate the interpretation of K matrix, it can be represented as a density plot. Hence, instead of the numerical values, the matrix elements are filled with grey values. Figure 4.6 represents the correlation matrix for all 24

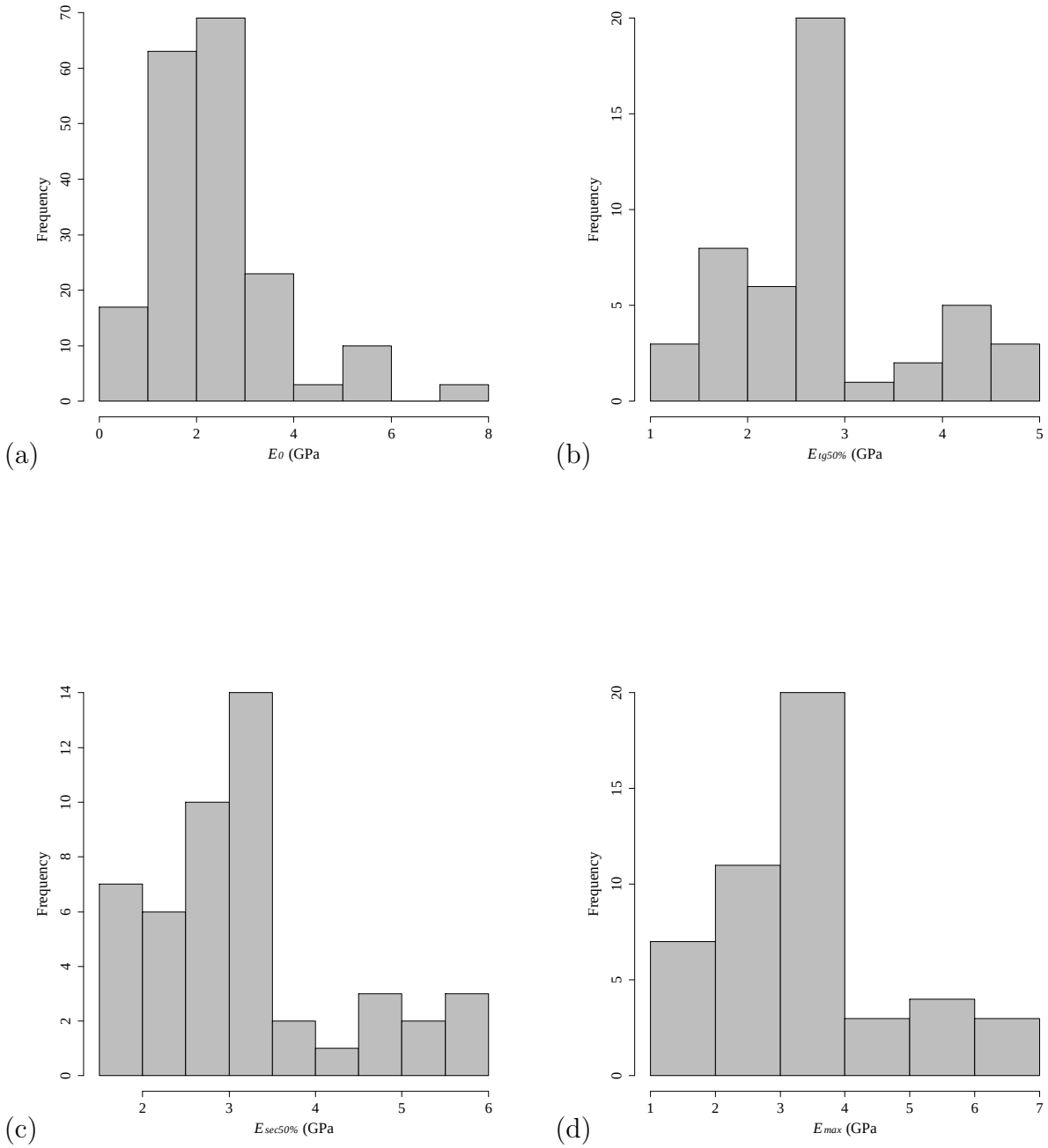


Figure 4.5: Histograms of: a) E_0, b) $E_{tg50\%}$, c) $E_{sec50\%}$ and d) E_{max} in the study of *JGLF*

attributes and target variables in the *JGLF* strength study. Taken into account that the number of records of the database used in *JGLF* deformability modulus study is not the same than in *UCS* study, the equivalent representation for E_0 is presented in Figure 4.7 and, for E_{max}, $E_{sec50\%}$ and for $E_{tg50\%}$ (here the database is the same for all three moduli) is shown in Figure 4.8.

As previously mentioned, in the study of *JGLF* mechanical properties seven different *JG* project were considered. This means that were prepared *JGLF* using seven different soil types. Thus, soil samples were collected from the different test sites of the present research work and submitted to laboratory tests to obtain a physical characterisation of the natural soils used in the *JGLF* preparation. Although all the soils are of a clayey nature, they contain different percentages of sand, silt, clay and organic matter. Considering the information from the literature review (see Section 3.2), where the soil influence is defined only for cohesive and granular soils, it is expected just a slight influence of the soil properties in the present research work. A detailed classification of the natural soils is provided in Table 4.2, where the first column denotes the construction site and the third column shows the number of records that contain that soil. The soil classification was based on the Unified Soil Classification System – ASTM D2487–83 (ASTM, 1985). This system is based on identifying soils according to their textural and plasticity qualities and on their grouping with respect to behaviour. Soils seldom exist in nature separately as sand, gravel, or any other single component. They are usually found as mixtures with varying proportions of particles of different sizes, which independently contribute for the global characteristics of the soil mixture. Based on such characteristics the soil is evaluated as an engineering construction material. For soil classification, the following properties, which can be determined by simple tests, are considered: percentages of gravel, sand, and fines (fraction passing the #200 sieve); shape of the grain-size-distribution curve; and plasticity and compressibility characteristics. Combining all this information, the Unified Soil Classification System, label the soil with a letter symbol and a descriptive name indicating its principal characteristics.

All laboratory formulations used in the study of *JGLF* deformability were prepared with cement type CEM I 42.5R (Portland cement with 100% clinquer) and CEM II 42.5R (composed Portland cement with ≥65% clinquer). For *UCS* study, additionally to this two cement types were also prepared some samples with pozzolanic cement (CEM IV/A 35.5R with ≥20% clinquer).

It should be remarked that in the current stage of knowledge within *JG* technology domain, there are still no specific procedures to prepare *JG* laboratory mixtures. However, some specifications/standards currently applied to similar materials, such as concrete, can be adopted and used to guide the preparation of *JG* laboratory mixtures (Magalhães,

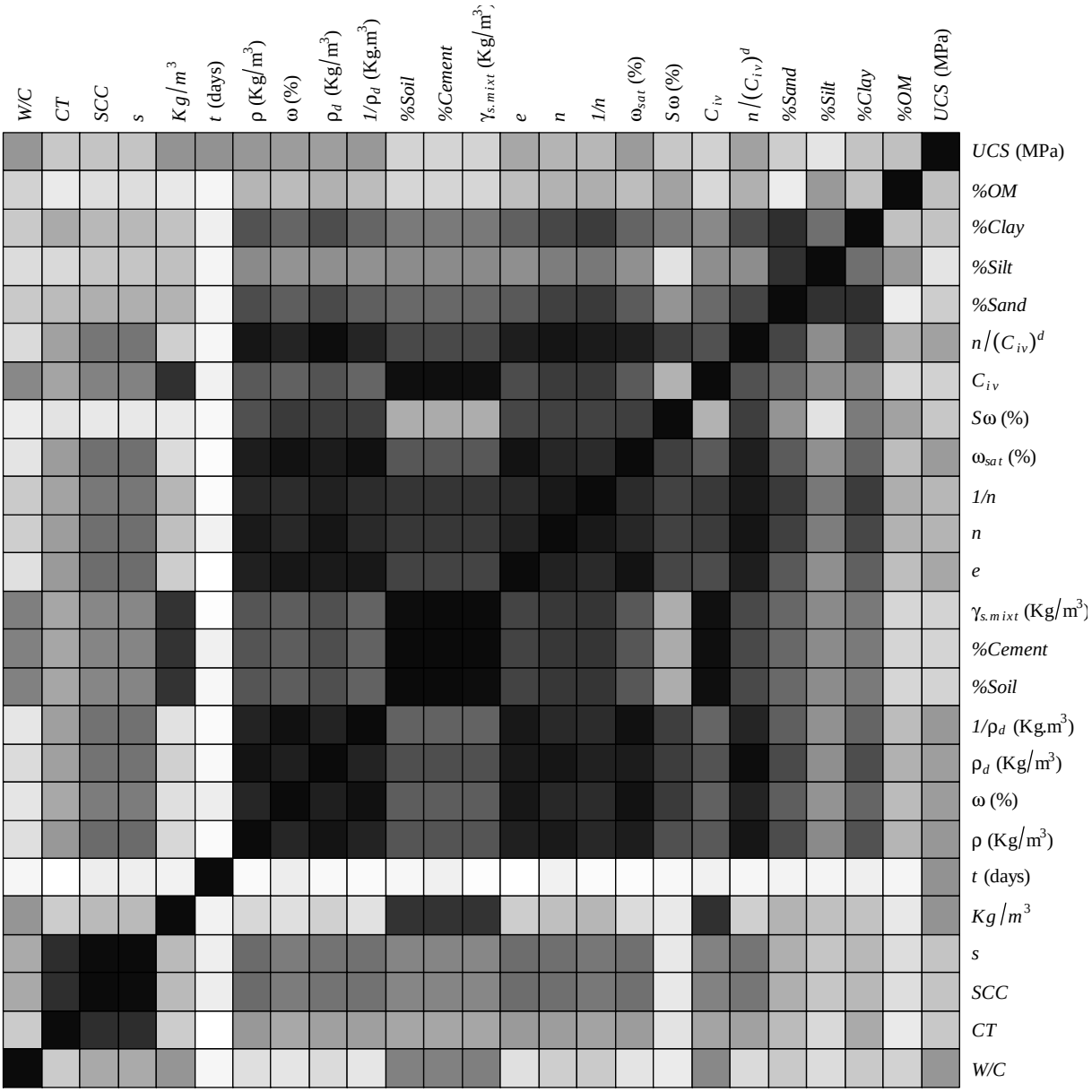


Figure 4.6: Correlation matrix as a frequency graph for all 24 variables considered in UCS prediction of $JGLF$. In this representation the absolute values were considered, here *white* means $k_{ij} \approx 0$ (uncorrelated) and *black* $|k_{ij}| \approx 1$ (strongly correlated)

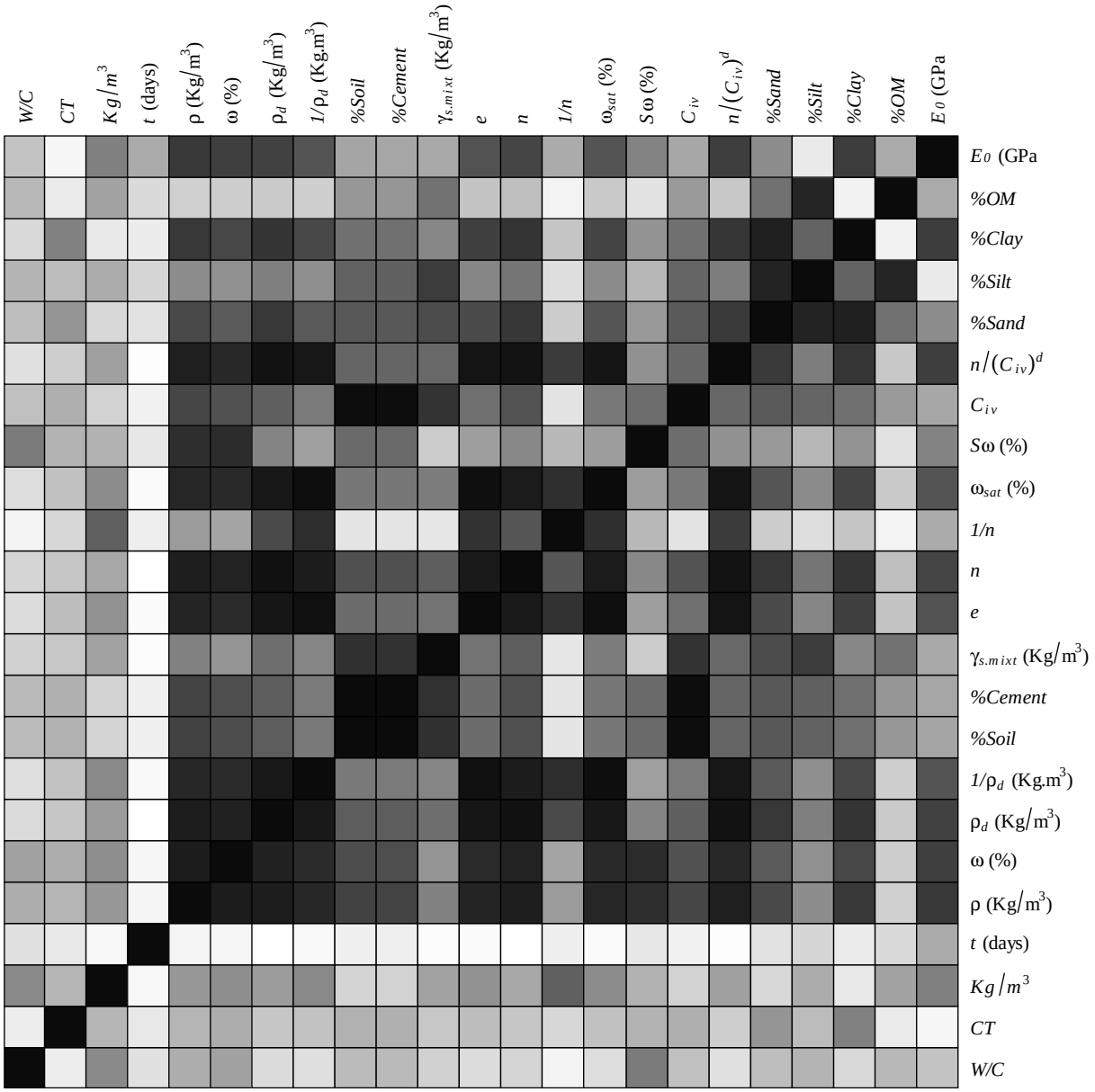


Figure 4.7: Correlation matrix as a frequency graph for all 22 variables considered in E_0 prediction of *JGLF*. In this representation the absolute values were considered, here *white* means $k_{ij} \approx 0$ (uncorrelated) and *black* $|k_{ij}| \approx 1$ (strongly correlated)

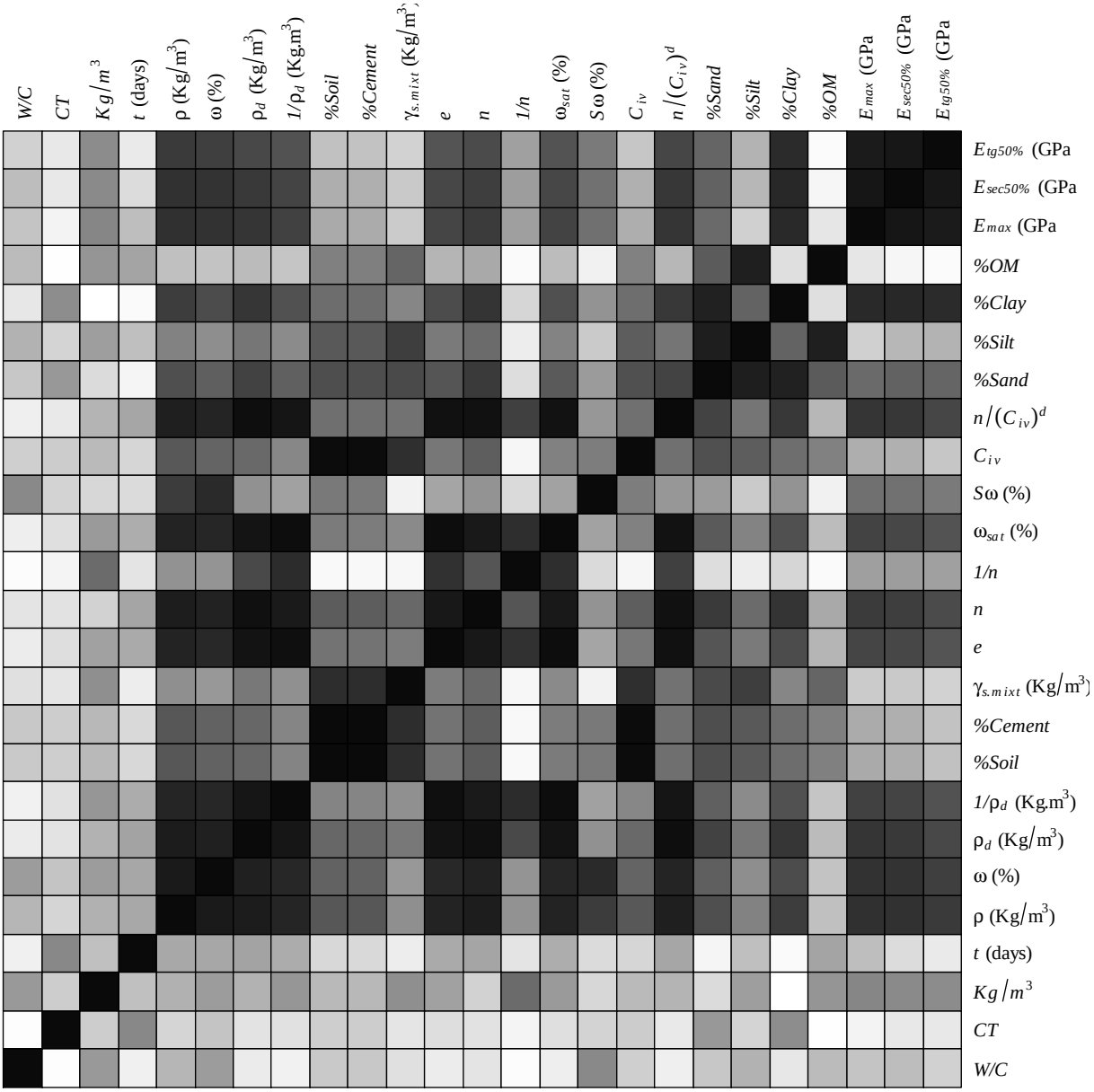


Figure 4.8: Correlation matrix as a frequency graph for all 22 variables considered in $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} prediction of *JGLF*. In this representation the absolute values were considered, here *white* means $k_{ij} \approx 0$ (uncorrelated) and *black* $|k_{ij}| \approx 1$ (strongly correlated)

Table 4.2: Soil types present in the collected data for *JGLF* mechanical properties study and its classification according to ASTM D2487–00 (ASTM, 1985)

Site	Soil Type	Frequency			$\%Sand$	$\%Silt$	$\%Clay$	$\%MO$
		UCS	E_0	E^*				
A	Lean clay (CL)	10	28	9	39.0	33.0	27.0	8.3
B	Organic lean clay (OL)	5	18	6	6.0	57.0	37.0	1.8
C	Fat clay (CH)	85	93	22	7.0	53.0	40.0	3.2
D	Silty clay (CL-ML)	20	27	6	25.0	52.5	22.5	0.4
E	Lean clay (CL)	15	22	5	0.0	55.0	45.0	3.9
F	Silty clay (CL-ML)	20	-	-	32.5	43.5	24.0	1.2
G	Lean clay (CL)	20	-	-	10.5	48.5	41.0	1.0

E^* - E_{max} or $E_{sec50\%}$ or $E_{tg50\%}$

2006). For instance, the *JGLF* were mixed in laboratory using an electrical machine, allowing a better homogeneous mixture. A special care was also given during samples manipulation in order to not introduce vibrations. Another important issue is related with the cure conditions, that should be as similar as possible to the *in situ* conditions. Hence, after prepared, each *JGLF* sample was coated with a film waterproofing and stored under the adequate temperature and humidity conditions.

4.3 Field data

The capability of accurately predict the mechanical properties of *JGLF* is just the first step on *JG* technology design. Therefore, after overcome this issue, the next and most important step is to develop reliable methods to predict *soilcrete* mechanical properties (strength and stiffness) and *JG* column diameter.

Once again, as in the study of *JGLF* mechanical proprieties, the first step is to compile a dataset with all available and potential useful information. In this respect it should be noted that, this simple and apparently vulgar task is more complex than looks like and consumed a lot of time. The main reason for this observation is related with the absence of systematic process for information organization during a *JG* project. This means that, although most of the information exist, it is spread in different "places" within the *JG* company, which represents a huge obstacle on the database compilation process for *DM* purposes.

In order to guarantee the highest reliability as possible of the present research work results, the entire database compilation process was guided by a rigorous verification

procedure, contemplating the following steps:

1. Collection of all information available in spreadsheets, old archives and many other sources;

2. Compilation of the collected information in a structured and organized database;

3. Performing a first attempt in order to fill the missing data by researching the collected information, looking for into other information sources, talking with the engineers responsible for each project as well as with experts;

4. Deep and careful revision of the entire database with the collaboration of an engineer (employer of the company that supplied all information) who was involved in a significant number of *JG* projects included in the database;

5. Sending the database for a detailed revision. In this step, the database was split by project and sent to the responsible engineer of such project;

6. Revision of the entire database, considering the comments introduced in the previous step, and compilation of the final database.

It should be noted that the compiled database, had two main purposes. The first and foremost was to support the present research work. The second one was to boost an important process for the company related with the development of a structured database, representing the framework for storing information for future *JG* projects.

An overview of the compiled database showed that the most complete records, i.e., containing information for almost all attributes, belong just to five *JG* works, within a total of 107 *JG* projects. Therefore, and in order to have available the highest number of records, independently of the set of input variables selected, for all experiments performed aiming the development of predictive models for strength, stiffness and column diameter, it was only considered the data related with these five *JG* projects, which are related to works carried out in Portugal and Spanish.

During the compilation process of the field database, and keeping in mind the second purpose of the database above underlined, it was made an additional effort toward to fill the database with all information available related with each *JG* project. Thus, additionally to the variables listed above in the scope of the study of *JGLF* (see Section 4.2), some other variables/information were introduced, in order to consider all parameters related with *JG* process as well as to describe in detail each work. The following list enumerates all variables available and considered in the study of the *soilcrete* mechanical properties and *JG* column diameter:

- W/C - Water/Cement ratio
- CT - cement type
- SCC - strength cement class
- s - coefficient related with cement type
- kg/m^3 - kilograms of cement by cubic meter of soil
- kg/ml - kilograms of cement by linear meter of column
- t (days) - age of the mixture
- ρ ($\mathrm{kg\,m^{-3}}$) - natural density of the mixture
- ω (%) - water content of the mixture
- ρ_d ($\mathrm{kg\,m^{-3}}$) - dry density of the mixture
- $1/\rho_d$ ($\mathrm{m^3\,kg^{-1}}$) - inverse of dry density of the mixture
- $\%Soil$ - soil content in the mixture
- $\%Cement$ - cement content in the mixture
- $\gamma_{s.mixt}$ ($\mathrm{kg\,m^{-3}}$) - unit weight of the mixture
- e - void ratio of the mixture
- n - mixture porosity
- $1/n$ - inverse of the mixture porosity
- ω_{sat} (%) - saturated water content
- S_ω - degree of saturation
- C_{iv} - volumetric content of cement
- $n/(C_{iv})^d$ - relation between mixture porosity and volumetric content of cement
- W_c/C - soil water/cement ratio: ratio between water content of soil and cement content
- S/C - soil/Cement ratio: ratio between weight of soil and weight of cement

- OM/C - Organic Matter/Cement ratio: ratio between organic matter content and cement content
- $OM/C^{W_c/C}$ - relation between organic matter, cement content and soil water content
- ρ_{grout} - grout density
- $\%Sand$ - percentage of sand in the natural soil
- $\%Silt$ - percentage of silt in the natural soil
- $\%Clay$ - percentage of clay in the natural soil
- $\%OM$ - percentage of organic matter in the natural soil
- H (m) - depth where sample was collected
- JS - Jet system
- WS (cm/min) - withdrawal speed of the rod
- rpm - rotation speed of the rod
- WT (s) - withdrawal time of the rod
- $Step$ (cm) - withdrawal step
- FR (l/min) - flow rate of grout slurry
- D_{grout} (mm) - mean diameter of grout nozzles
- N_{Dgrout} - number of grout nozzles
- D_{water} (mm) - diameter of water nozzle
- P_{grout} (bar) - grout pressure
- P_{air} (bar) - air pressure
- P_{water} (bar) - water pressure
- Imp_{grout} (kg) - grout impact
- UCS (MPa) - uniaxial compressive strength;
- E_0 (GPa) - Young's modulus;

- D (mm) - column diameter

Relating to the list of variables above presented, particularly for the *impact*, it was balanced its empirical relevance and the number of records available if such variable was considered or not. Accordingly, we considered only the impact of the grout, since there is a lot of missing data related to the pressure and nozzle diameter for air and water jets (required for the calculation of the total impact). This way, it was possible to include the impact variable and maximize the number of records available.

Similarity to the study of *JGLF*, not all variables above enumerated were directly measured from the *JG* columns (e.g. Imp_{grout} is calculated from other variables). In these cases, the mathematical expressions used for its calculation are presented in the Appendix B. The main statistics, i.e., *maximum*, *minimum*, *mean* and *standard deviation*, of each input and output variables considered in the study of *soilcrete* mechanical properties and *JG* column diameter are present on Tables A.4, A.5 and A.6 of Appendix A.2. Figure 4.9 shows the histograms of *soilcrete* mechanical properties, where one can observe that the shape of UCS histogram (Figure 4.9a) is similar to that found in the literature and shown in Figure 3.4b. Figure 4.10 plots the histogram of *JG* column diameter. The histograms of each attribute considered in the study of *soilcrete* mechanical properties and *JG* column diameter are presented in Appendix A.2.

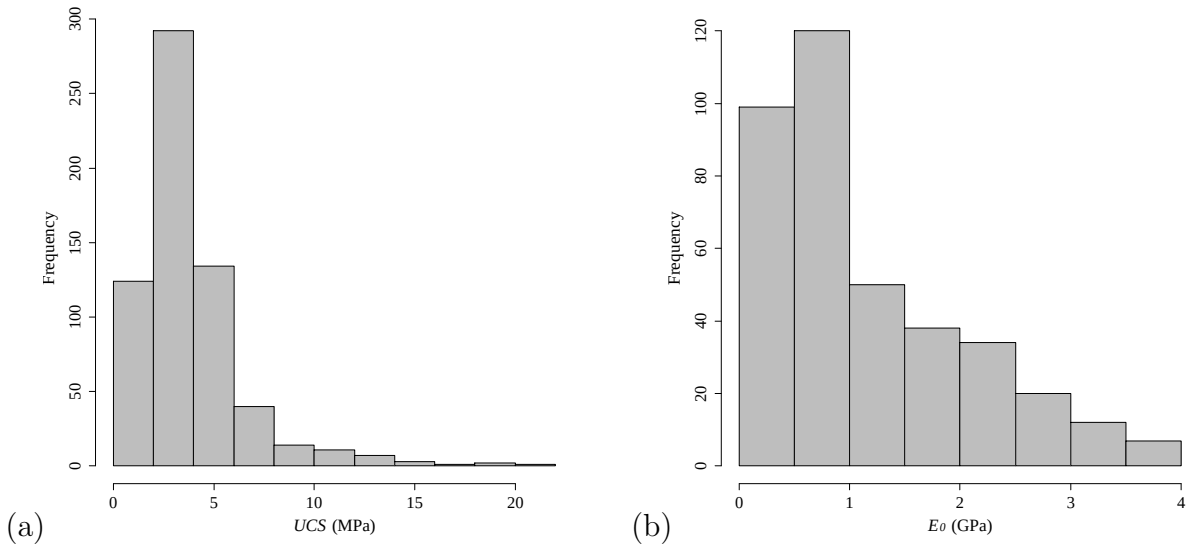


Figure 4.9: Histogram of: a) UCS and b) E_0, in the study of *soilcrete* mechanical properties

Also here, as has been done in the study of *JGLF*, the correlation matrix for all variables considered in the study of both *soilcrete* mechanical properties and *JG* column diameter were calculated. Thus, Figure 4.11 shows the correlation matrix for all variables used in the study of UCS of *soilcrete* samples. The equivalent matrix for E_0 and *JG*

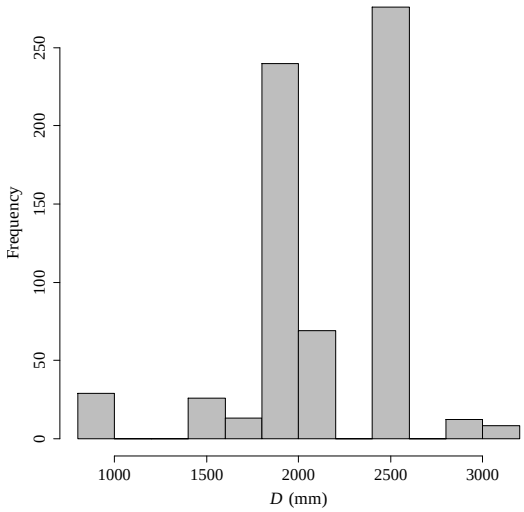


Figure 4.10: Histogram of D study

column diameter are shown in Figure 4.12 and Figure 4.13 respectively. It should be stressed that for the calculation of the correlation values, it was only considered the complete and not constant ($\sigma \neq 0$) records. Moreover, and contrary to the study of *JGLF* where were studied the different moduli that can be defined in an unconfined compressed test (see Figure 4.2), here it was only analysed the E_0, since there are no information related to the remains moduli.

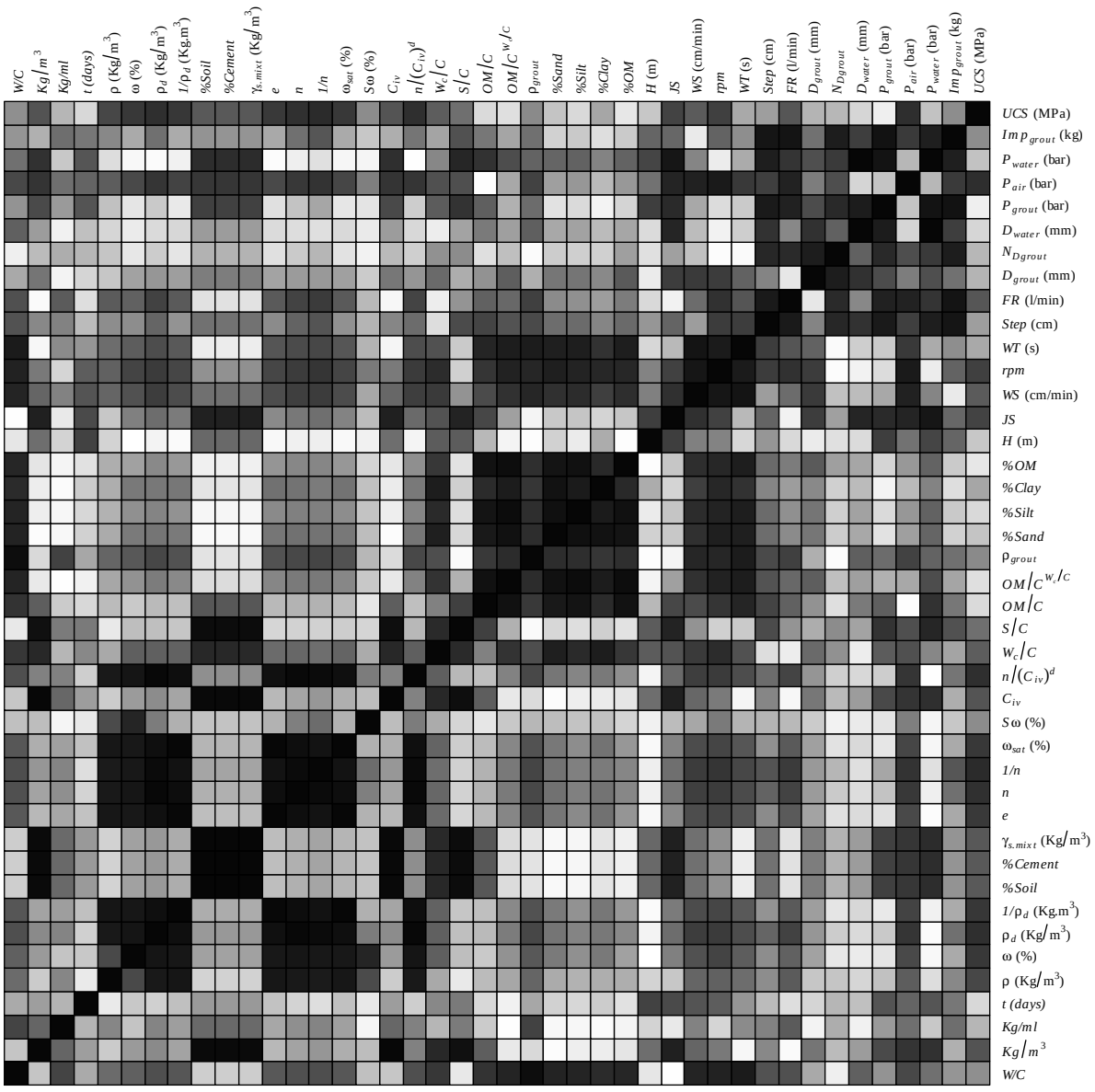


Figure 4.11: Correlation matrix as a frequency graph for all 41 variables considered in UCS prediction of JG field samples. In this representation the absolute values were considered, here *white* means $k_{ij} \approx 0$ (uncorrelated) and *black* $|k_{ij}| \approx 1$ (strongly correlated)

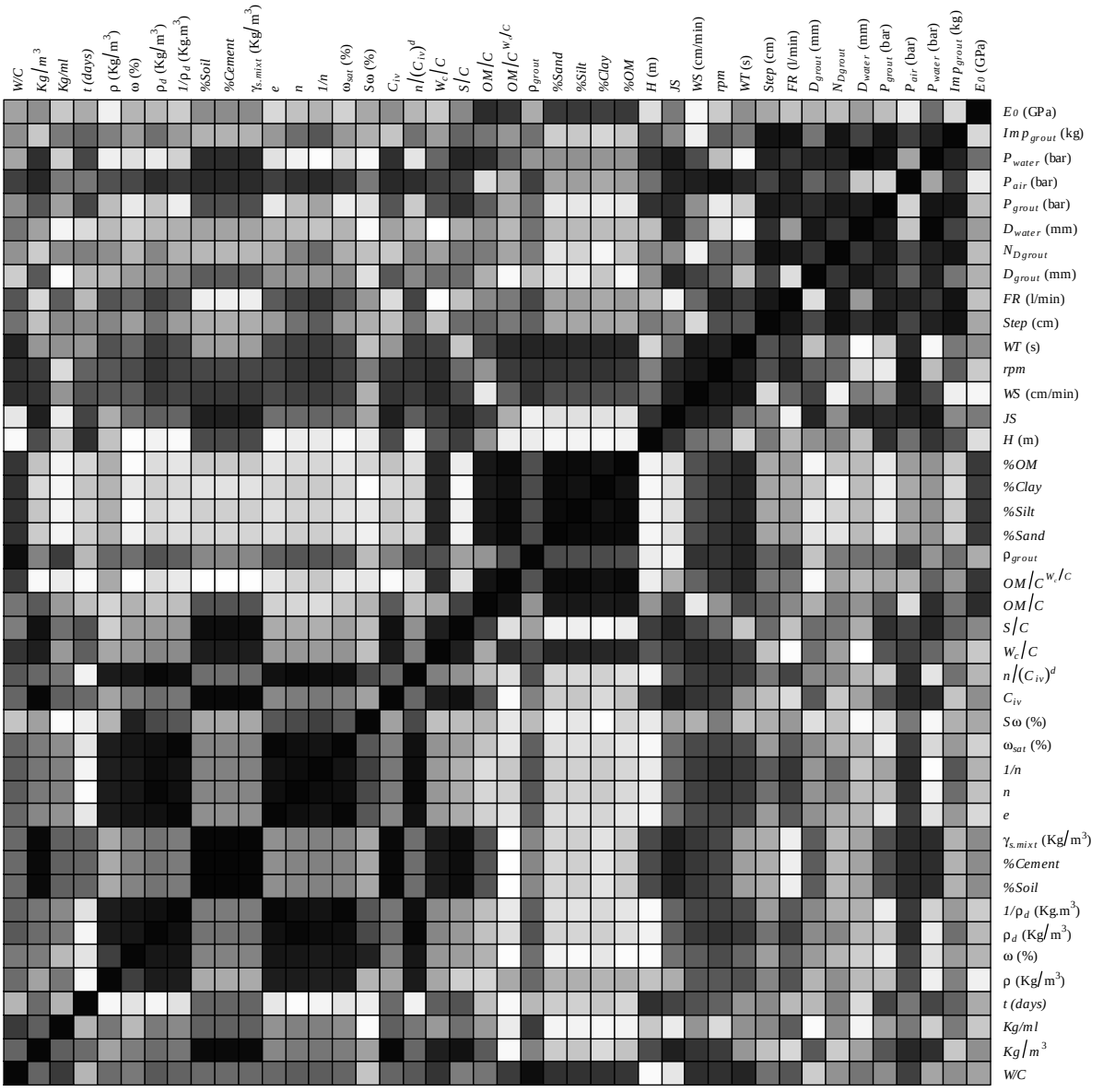


Figure 4.12: Correlation matrix as a frequency graph for all 41 variables considered in E_0 prediction of *JG* field samples. In this representation the absolute values were considered, here *white* means $k_{ij} \approx 0$ (uncorrelated) and *black* $|k_{ij}| \approx 1$ (strongly correlated)

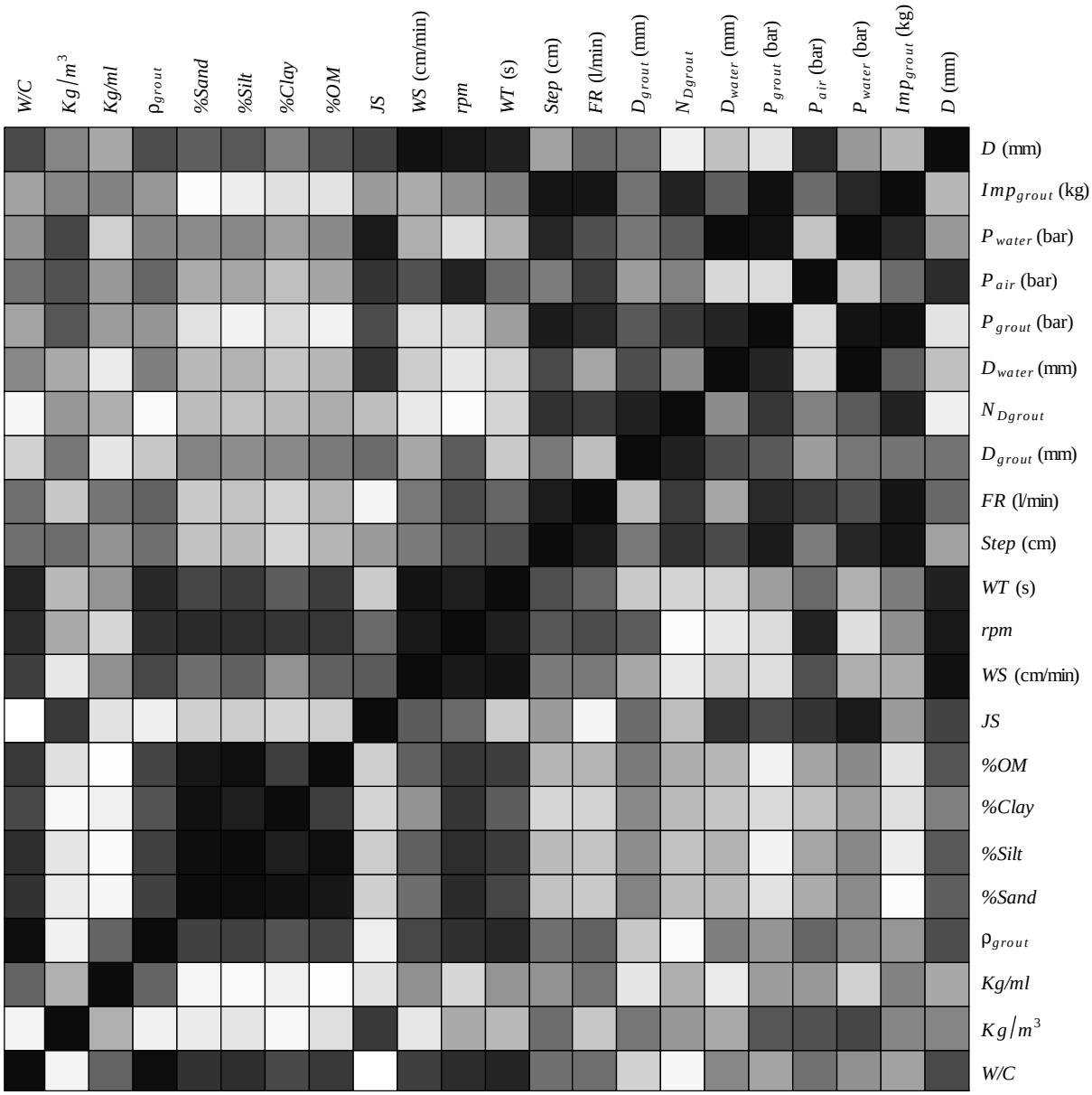


Figure 4.13: Correlation matrix as a frequency graph for all 21 variables considered in D prediction. In this representation the absolute values were considered, here *white* means $k_{ij} \approx 0$ (uncorrelated) and *black* $|k_{ij}| \approx 1$ (strongly correlated)

4.4 Conclusions

The success of the present research, similar to that of any study that involves the application of *DM* techniques, is strongly dependent on the database quality used during the experiments. Hence, throughout the entire process of the database compilation a rigorous methodology was followed to guarantee the highest reliability possible. Moreover, an extra effort was put forth to use as many of the number of records as possible and to include all potentially useful variables.

Despite all the difficulties and obstacles found during the database compilation process, in the end two main databases were prepared with the most relevant variables within the *JG* technology domain and with a significant number of records that represent a particular issue for *DM* application studies purposes. However, it should also be noted that in the case of the laboratory database for deformability study, namely for $E_{tg50\%}$, $E_{sec50\%}$ and E_{max}, the number of records is particularly small (only 48 records are available). However, due to the interest of these moduli for practical purposes, some experiments will still be performed.

Due to the large amount of missing data in the field database, special care was required. Although there are approaches to deal with missing data, in the present work, and after some experiments, only the complete records were considered because the implementation of such approaches could compromise the results' reliability.

Concerning the soil characterisation, its effect was considered based on the %Sand, %Silt, %Clay and %OM of each soil type (seven in the study of laboratory mixtures and five in the study of field mixtures), all of which are of a clayed nature.

Chapter 5

DM techniques applied to laboratory data

5.1 Introduction

This section presents the explored data-driven models for *JGLF* mechanical properties prediction through the application of *DM* techniques.

For the *FS* task, we applied two different approaches, namely a forward sequential *FS* and backward selection scheme. For the last one, we opted for the procedure implemented in the *rminer* package that is guided by a *SA* procedure, as previously explained in Section 2.5.3. Adopting a *SA* to guide the variable deletion, the computational effort can be reduce by a factor of I (when compared to the standard backward procedure) (Cortez et al., 2009). A schematic representation, contemplating the main steps, of these two *FS*s schemas is depicted in Figure 5.1. After run a *FS* method, based on the achieved results and on the statistic analysis information, particularly those related with the correlate coefficient shown in Figures 4.6, 4.7 and 4.8 and considering the empirical knowledge related with the soil-cement laboratory mixtures behaviour, a manual *FS* was performed. In other words, several models with different sets of variables were trained (using the *SVM* algorithm) and compared its performance, considering *MAD*, *RMSE*, and R^2 as a performance criteria. At the end, a set of nine input variables was selected (eight in the case of stiffness prediction) and trained each one of the four *DM* algorithms (i.e. *MR*, *ANN*, *SVM* and *FN*). A full search of all possible combination between all variables was not considered because this study contemplates 24 variables and such combinatorial exploration is not practically possible (in the conditions, there are around 16777215 combinations between all 24 variables).

For model selection purposes, particularly during the *FS* step, where only *SVM* was applied, we adopted the methodology proposed by Huang et al. (2007). The main advantage of this approach lies in the fact that the three *SVM* parameters {C, γ, ϵ} can be

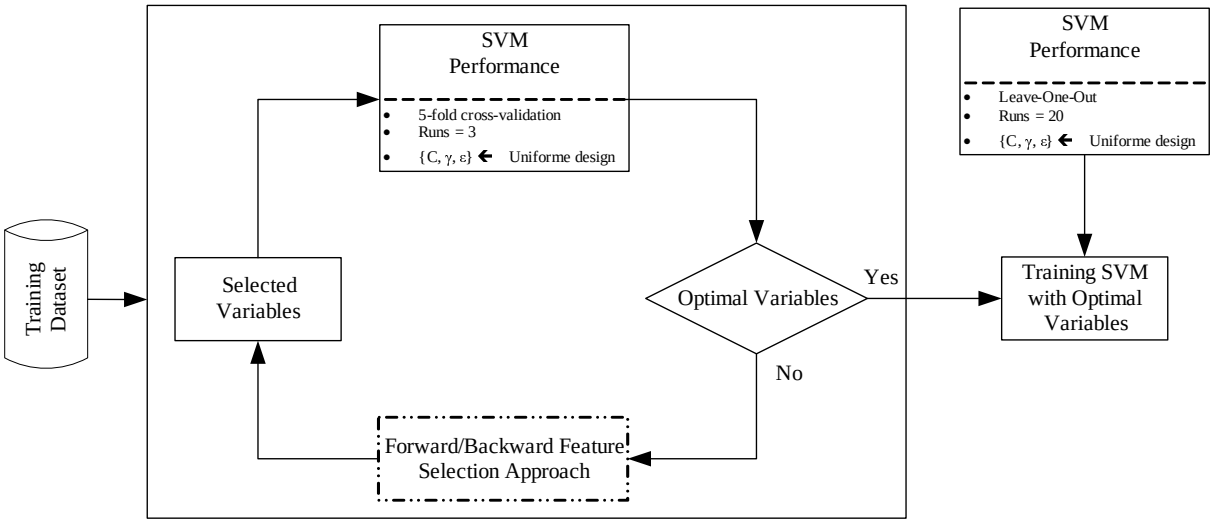


Figure 5.1: Schematic of the *FS* approaches

automatically defined, which is very useful during the *FS* process.

During the learning phase (after choosing the input variables), for *ANN* we adopted a fully connected multilayer perceptron, with one hidden layer with H processing units, bias connections and logistic activation functions $1/(1+e^{(-x)})$. To find the best value for H, a grid search within the range $\{2, 4, ..., 10\}$, under an internal (i.e. applied over training data) 5-fold cross validation (Hastie et al., 2009) was executed. Under this grid search, the H value that produced the lowest *MAD* was selected and then the *ANN* was retrained with all training data. For *SVM* algorithm, in order to reduce the search space, we adopted the popular gaussian kernel and considered the heuristics proposed by Cherkassky and Ma (2004) to set the complexity penalty parameter, C=3, and the size of the insensitive tube, $\epsilon = \hat{\sigma}/\sqrt{N}$, where $\hat{\sigma} = 1.5/N \cdot \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$, y_i is the measured value, $\hat{y}_i$ is the value predicted by a 3-nearest neighbour algorithm and N the number of examples. The most important *SVM* parameter, the kernel parameter γ, was set using a grid search within $\{2^{-15}, 2^{-13}, ..., 2^{3}\}$, under the same internal 5-fold cross validation scheme (Hastie et al., 2009).

Additionally to *ANN* and *SVM* algorithms, we also tested a *MR*, as a baseline comparison. All these three *DM* algorithms (*ANN*, *SVM* and *MR*) were implemented in the *R* tool (R Development Core Team, 2009) and *rminer* library. Furthermore, before fitting the *ANN*, *SVM* and *MR* models, the data attributes were standardized to a zero mean and one standard deviation and before analysing the predictions, the outputs post-processed with the inverse transformation (Hastie et al., 2009).

In this research, we also explored a *FN* to solve the following generic expression,

aiming to predict the mechanical properties of *JGLF*:

$$\hat{y} = \beta_0 \cdot \prod_{i=1}^{n} x_i^{\alpha_i} \tag{5.1}$$

where, $\{x_1, ..., x_i\}$ are the input parameters, $\{\beta_0, \alpha_1 ..., \alpha_i\}$ are coefficients to be adjusted. To learn the coefficients in Equation 5.1 the following minimization problem was used:

$$Minimize \quad Q = \sum_{S=1}^{S} \delta_S^2 = \sum_{S=1}^{S} \left(y_s - \beta_0 \cdot \prod_{i=1}^{I} x_i^{\alpha_i} \right)^2 \tag{5.2}$$

The formulation and resolution of this *FN* was implemented in the free version of the GAMS software (GAMS Development Corporation, 2012).

Additionally to the four *DM* algorithms above enumerated, the analytical models proposed by *EC2* (CEN, 2004a) and *MC90* (CEB-FIP, 1991) for strength and stiffness prediction of concrete were also adapted to *JGLF* to predict such properties.

5.2 Uniaxial compressive strength prediction

5.2.1 Model performance

Table 5.1 compares the models performance of the two *FS* approaches implemented and the two models where the set of variables were manually selected (Tinoco et al., 2009, 2011b), using *MAD*, *RMSE* and R^2 as a performance criteria. This table shows that both forward and backward *FS* approaches were unable to define the best set of variables to predict *JGLF* mechanical properties. However, the information given by these two approaches represent an important information source in the definition of the nine input variables (termed as MS_{ql1}) that lead to the best predictive models, which were used during the *UCS* study of *JGLF* (Tinoco et al., 2009).

After optimizing the coefficients of *EC2* analytical expression to *UCS* data of *JGLF*, coefficient a of Equation 3.17 took the value of $a = 0.5$, leading to the following model (in this work, this model is termed *EC2-UCS.Lab*):

$$UCS = e^{\left(s \cdot \left[1 - \left(\frac{28}{t}\right)^{0.5}\right]\right)} \cdot UCS_{28days} \tag{5.3}$$

The mathematical expression resulting from the optimization of the coefficients in Equation 5.1 using the *FN* algorithm and *UCS* data is written as (this model will be termed

as *FN-UCS.Lab*):

$$\begin{aligned} UCS = 12023149 \cdot W/C^{-1.052} \cdot s^{-2.090} \cdot t^{0.239} \cdot \\ \cdot (n/(C_{iv})^{d})^{-3.064} \cdot C^{1.473} \cdot \%Sand^{-0.028} \cdot \\ \cdot \%Silt^{-1.594} \cdot \%Clay^{0.397} \cdot \%OM^{-0.028} \end{aligned} \tag{5.4}$$

The above equation, trained using the Leave-One-Out estimation method and the minimization problem according to Equation 5.2, achieved a high performance with small values of $MAD = 0.58$ MPa and $RMSE = 0.75$ MPa and an R^2 close to the unit vale ($R^2 = 0.92$). Moreover, as shown in Figure 5.2, the number of predictions above of diagonal line is approximately equal to the number of predictions below the same line, which is a sign of reliability, since the model predictions are not either underestimated or overestimated.

The average hyperparameters and fitting time values (and respective 95% level confidence intervals according to a t-student distribution) of all *DM* models trained using the set of nine input variables assigned in Table 5.1 as MS_{ql1} are shown in Table 5.2. These models (trained to predict *UCS* of *JGLF*) will be termed as *MR-UCS.Lab*, *ANN-UCS.Lab* and *SVM-UCS.Lab*, and are respectively the result of the training of *MR*, *ANN* and *SVM* algorithms with *UCS* data of *JGLF*.

Table 5.3 shows the predictive capacity of all trained models, comparing its performance in *UCS* prediction of *JGLF* based on the *MAD*, *RMSE* and R^2 metrics, computed for the test data under a leave-one-out approach (mean value and 95% confidence intervals). This table shows that *UCS* of *JGLF* can be accurately predicted by each one of the four *DM* models, particularly by *ANN-UCS.Lab* and *SVM-UCS.Lab* models. Moreover, it is shown that *EC2-UCS.Lab* model also represents a good alternative to predict *UCS* of *JGLF* over time, which is characterized by its simplicity. However, it should be noted that such model has limitations, being impossible its application during the project level, since it requires the 28 days strength of each formulation, which implies waiting 28 days before performing experimental tests for its quantification.

Scatterplots of *ANN-UCS.Lab* and *SVM-UCS.Lab* models illustrated in Figures 5.3 and 5.4 respectively, corroborate the high predictive performances shown in Table 5.3. As shown, in both models the predictions are very close to the experimental values (diagonal line). In Figure 5.5 it's compared the predictive performance of all models trained for *UCS* prediction of *JGLF* (*EC2-UCS.Lab*, *FN-UCS.Lab*, *MR-UCS.Lab*, *ANN-UCS.Lab* and *SVM-UCS.Lab*), depicting the model accuracy as a function of the absolute deviation (*REC* curves, (Bi and Bennett, 2003)). The shape of these curves evidence once more the high performance of the models, namely of *ANN-UCS.Lab* and *SVM-UCS.Lab* models.

For instance, the *REC* curve of *SVM-UCS.Lab* model shows that if an absolute deviation around 1.0 MPa is tolerated, then 80% of the records can be accurately predicted by the model. It is also appealing to observe that *EC2-UCS.Lab* model predicts accurately 20% of the records (absolute deviation equal to zero). However, it should be stressed that this is just a consequence of the model structure. This means that implicitly the *EC2-UCS.Lab* model is able to predict correctly the 28 days strength of each formulation since this is a model input.

Table 5.1: Model performance comparison of the two *FS* approaches implemented and those where the attributes were manually selected, in *UCS* prediction of *JGLF*

Var	FFS	BFS	MS_{ql0}	MS_{ql1}
s	×	×	×	✓
$\%Sand$	✓	✓	✓	✓
$\%Silt$	×	✓	✓	✓
$\%Clay$	×	×	✓	✓
$\%OM$	✓	✓	✓	✓
C_{iv}	×	✓	×	✓
$n/(C_{iv})^d$	✓	×	×	✓
t	✓	✓	✓	✓
W/C	×	✓	✓	✓
CT	×	✓	✓	×
SCC	×	×	✓	×
ρ	✓	✓	✓	×
$1/\rho_d$	×	✓	×	×
S_r	✓	✓	×	×
ω	✓	✓	✓	×
kg/m^3	×	✓	✓	×
$1/n$	×	✓	×	×
MAD	0.62 ± 0.01	0.56 ± 0.02	0.54 ± 0.01	0.55 ± 0.00
RMSE	0.82 ± 0.02	0.74 ± 0.02	0.73 ± 0.01	0.73 ± 0.00
R^2	0.91 ± 0.01	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00

5.2.2 Model interpretability

Besides achieving a high predictive performance, it is also important to consider the explanatory power of the data-driven model. This is particularly relevant in the engineering domain. When "black-box" *DM* models are applied (e.g. *ANN* or *SVM* algorithms), involving complex mathematical expressions, the application of a given procedure able

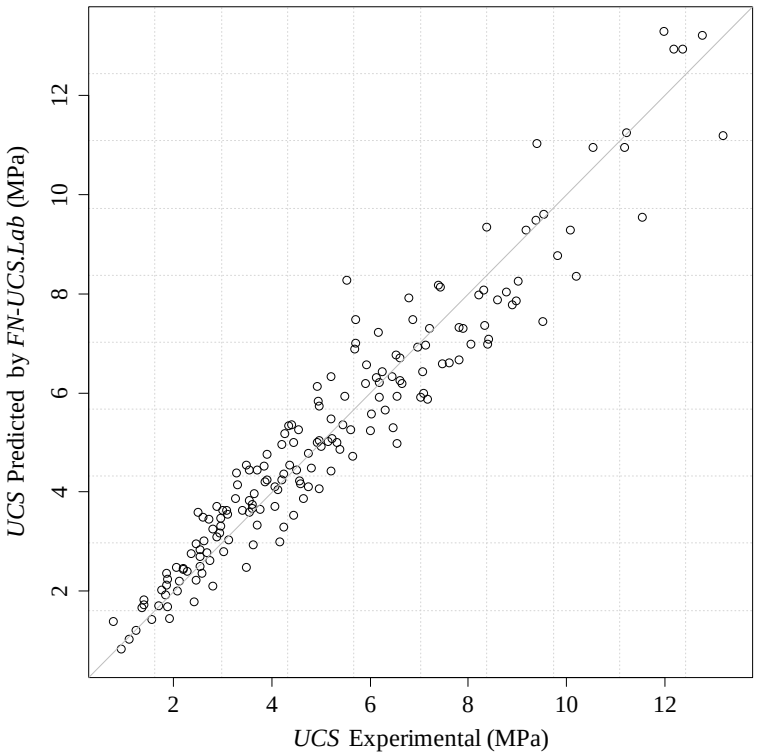


Figure 5.2: Relationship between UCS experimental versus predicted values by *FN-UCS.Lab* model

Table 5.2: Hyperparameters and computation time of each *DM* model for UCS prediction of *JGLF*

Model	Hyperparameters	time (s)
FN-UCS.Lab	-	35.5 ± 00.00
MR-UCS.Lab	-	11.08 ± 00.03
ANN-UCS.Lab	$H = 6 \pm 0$	346.63 ± 02.47
SVM-UCS.Lab	$\gamma = 0.12 \pm 0.00$, $\epsilon = 0.11 \pm 0.00$	1087.42 ± 12.82

Table 5.3: Error metrics of all *DM* models for UCS prediction of *JGLF* (test set values, best values in **bold**)

Model	MAD	RMSE	R^2
EC2-UCS.Lab	0.60 ± 0.00	0.88 ± 0.00	0.90 ± 0.00
FN-UCS.Lab	0.58 ± 0.00	0.75 ± 0.00	0.92 ± 0.00
MR-UCS.Lab	0.86 ± 0.00	1.13 ± 0.00	0.83 ± 0.00
ANN-UCS.Lab	0.61 ± 0.02	0.82 ± 0.02	0.91 ± 0.01
SVM-UCS.Lab	$\mathbf{0.55} \pm 0.00$	$\mathbf{0.73} \pm 0.00$	$\mathbf{0.93} \pm 0.00$

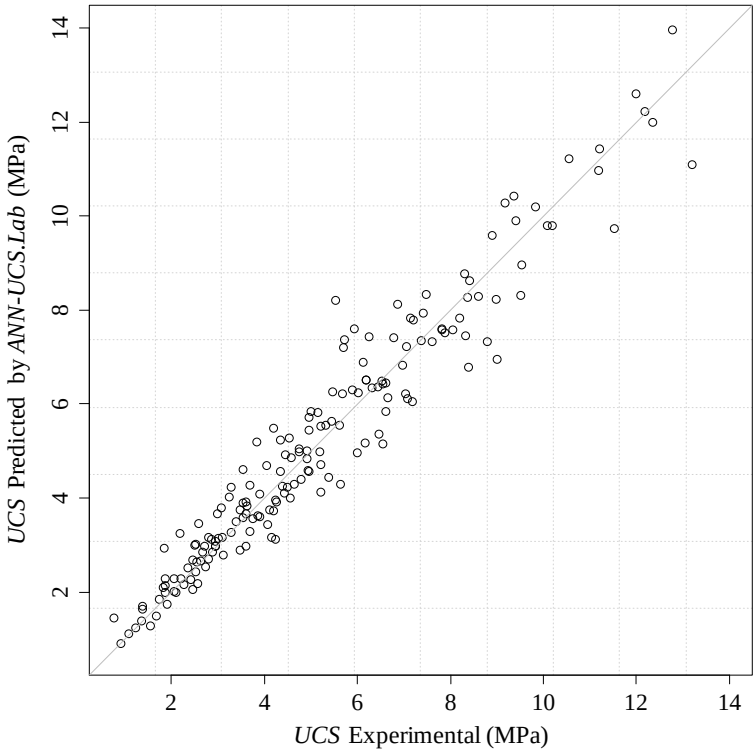


Figure 5.3: Relationship between UCS experimental versus predicted values by $ANN\text{-}UCS.Lab$ model

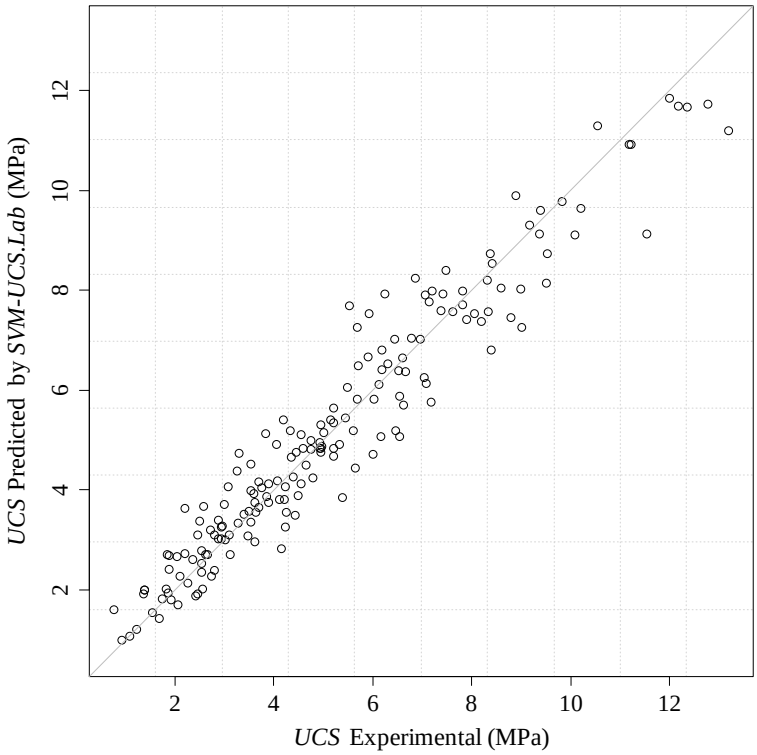


Figure 5.4: Relationship between UCS experimental versus predicted values by $SVM\text{-}UCS.Lab$ model

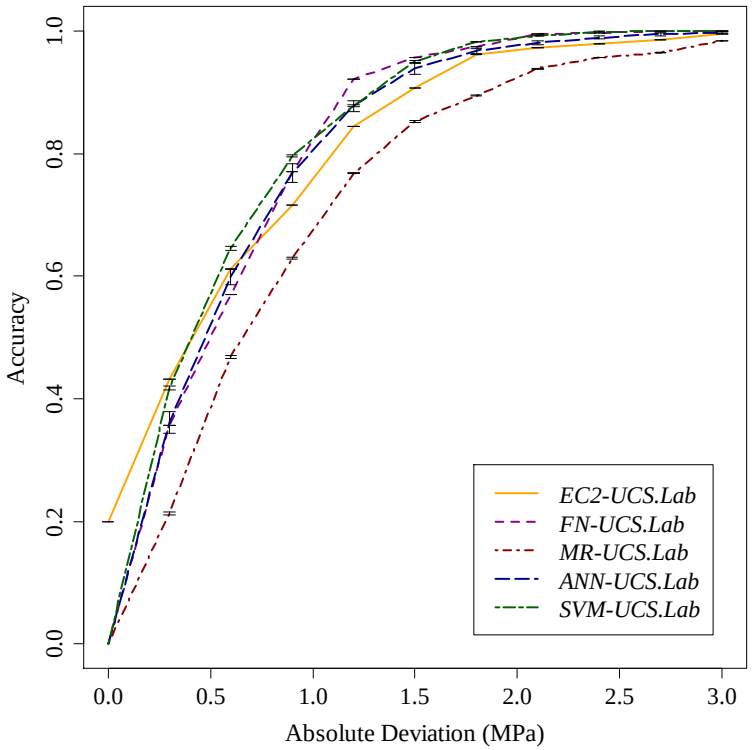


Figure 5.5: *REC* curves of *EC2-UCS.Lab*, *FN-UCS.Lab*, *MR-UCS.Lab*, *ANN-UCS.Lab* and *SVM-UCS.Lab* models, comparing the *UCS* predictive performances of *JGLF*

to "open" the model plays an important role. In this work, model interpretability was measured by quantifying what are the key input variables in *UCS* prediction of *JGLF* and their average effects in such prediction. For such purpose, the *GSA* algorithm (Cortez and Embrechts, 2011) was applied.

Figure 5.6 shows and compares the relative importance of each variable in *UCS* prediction of *JGLF* according to *MR-UCS.Lab*, *ANN-UCS.Lab*, *SVM-UCS.Lab* and *FN-UCS.Lab* models, as measured by the 1-D *SA*, with the correspondent t-student 95% confidence intervals for all 20 runs performed.

A first analysis to Figure 5.6 shows that the *UCS* behaviour of *JGLF* should not be guided only by a linear model. This observation is supported on the relative importance of each variable according to *MR-UCS.Lab* model that consider the soil properties, namely its sand, clay and silt content, as the only variables that control the *UCS* of *JGLF*. On the other hand, and according to *FN-UCS.Lab* model, the C_{iv} and the relation $n/(c_{iv})^d$ play the major role in strength prediction of *JGLF*. Moreover, the soil properties, mainly the $\%Silt$ and $\%OM$ also control the *UCS* of *JGLF* prediction. When compared with the empirical knowledge, *FN-UCS.Lab* model seems to underestimate the effect of the age of the mixture (4.25%) in the development of *JGLF* strength.

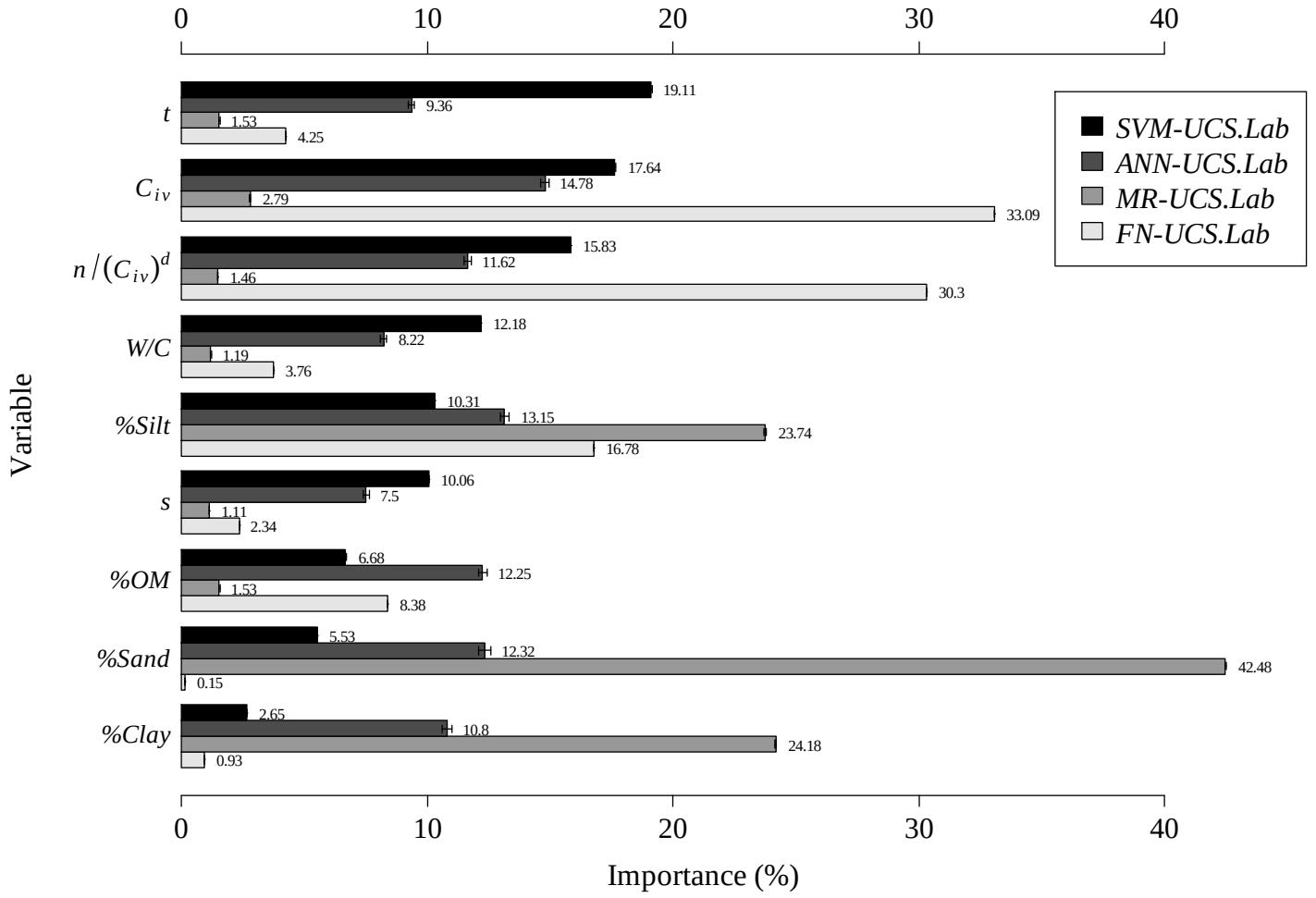


Figure 5.6: Relative importance of each input variable quantified by 1-D *SA*, comparing *MR-UCS.Lab*, *ANN-UCS.Lab*, *SVM-UCS.Lab* and *FN-UCS.Lab* models

The relative importances of each variable according to *ANN-UCS.Lab* and *SVM-UCS.Lab* are similar and in agreement with what is known empirically. Following *SVM-UCS.Lab* model, the three key variable for UCS prediction of *JGLF* are t (19%), C_{iv} (18%) and $n/(C_{iv})^d$ (16%). The soil properties, mainly according to *ANN-UCS.Lab* model, also have an important influence on UCS behaviour of *JGLF*. The W/C ratio and the s have a smaller impact in the strength behaviour of *JGLF*, with an relative importance around 12% and 10%, respectively. It is well known, from the experience with soil-cement mixtures (Coulter and Martin, 2006), that t has a strong influence in the behaviour of these kind of mixtures, mainly if one takes in account the range of t variable in the dataset used during the learning phase, i.e. $t \leq 56$ days time of cure (see Table A.1 on Appendix A). On the other hand, it makes sense that C_{iv} has a strong impact on cementitious materials (Horpibulsuk et al., 2003). The mixture porosity, also relevant in the strength behaviour of soil-cement mixtures, is indirectly considered in $n/(C_{IV})^d$ variable. Relatively to the influence of the soil properties and according to the *SVM-UCS.Lab*, it may seem strange its low influence. However, it should be stressed that all laboratory formulations were prepared using soils relatively similar, i.e., all soil were classified as clayed nature, just differing on its sand, silt, clay and organic matter content.

Making a global appreciation of all five models presented, and considering both metrics values and *SA* as performance criteria, the *SVM-UCS.Lab* model seems to be the most interesting. When compared with *ANN-UCS.Lab*, the metrics values are slightly better and the relative importance of the input attributes is more coherent in terms of what is known empirically in the *JG* domain. Furthermore, through the 20 runs performed, *SVM-UCS.Lab* model shows more consistency in the metrics values, as well as in the variables importance. Comparing *SVM-UCS.Lab* with *MR-UCS.Lab* and *EC2-UCS.Lab*, the advantages are more enhanced. On one hand, *MR-UCS.Lab* performance is lower (see Table 5.3) and unrealistic in terms of the relative influence of each input variable. On the other, comparing with *EC2-UCS.Lab*, besides of a higher performance, *SVM-UCS.Lab* has the important advantage of being applicable during the project level, where *EC2-UCS.Lab* is restricted because of its need for 28 days time of cure of each formulation. Finally, *FN-UCS.Lab* has a very similar performance in terms of the metrics values and relative importance distribution. However, the underestimation of the t effect by *FN-UCS.Lab*, as well as the strong asymmetry of the relative importance, make *SVM-UCS.Lab* more interesting for practical purposes. The main disadvantage of *SVM-UCS.Lab*, particularly when compared with *MR-UCS.Lab*, *EC2-UCS.Lab* and *FN-UCS.Lab*, is the high complexity of its mathematical expression that makes difficult its understanding by humans.

Given the previous explained analysis, we adopt the *SVM-UCS.Lab* as a reference model (Tinoco et al., 2012b). In order to achieve a better understanding of the modelled UCS behaviour of $JGLF$, we performed a more detailed SA analysis, under a 1-D and 2-D approaches, to measure the key input effects in the model.

Figure 5.7 depicts the effect of the four most relevant variables in UCS prediction of $JGLF$ according to *SVM-UCS.Lab* model. As a first observation, it can be pointed out that all four variables have a nonlinear effect on UCS behaviour of $JGLF$. Then, and as empirically expected, t and C_{iv} have a positive impact in strength prediction of $JGLF$ (Tinoco et al., 2011c). On the other hand, increasing $n/(C_{iv})^d$ (i.e. increase the porosity of the mixture or decrease the volumetric content of cement) or W/C leads to a decrease of mixture strength. This means that these both variables have a negative impact in UCS behaviour. Moreover, the VEC curve of t shows a concave shape, which

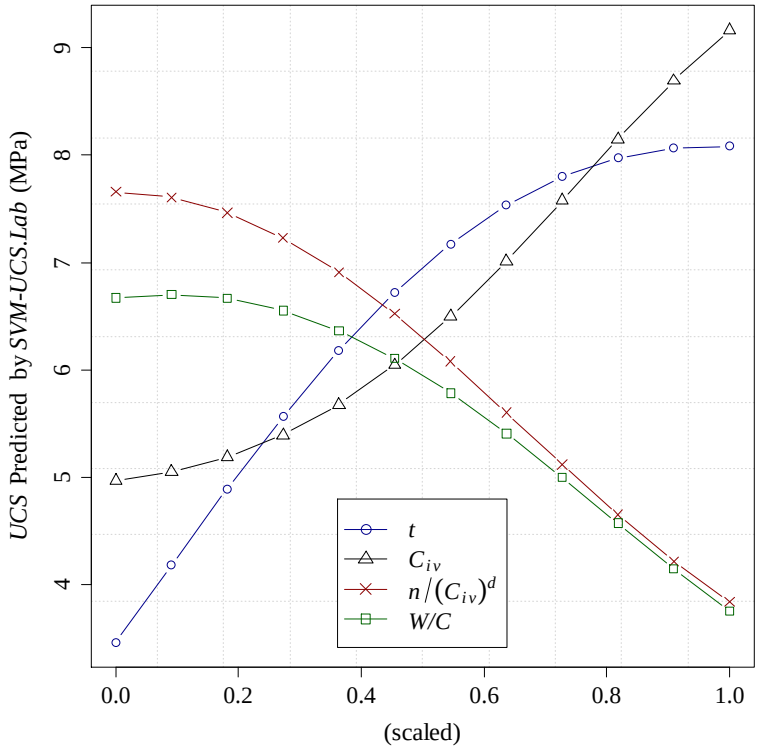


Figure 5.7: VEC curves for the four key input variables according to *SVM-UCS.Lab* model in UCS prediction of $JGLF$, quantified by 1-D SA

means that the mixture strength increases more quickly in early ages (up to 45 days time of cure) and then more slowly, until it stabilizes (Horpibulsuk et al., 2003; Van Impe et al., 2005).It is also interesting to observe the shape of C_{iv} VEC curve, showing that C_{iv} improves considerably UCS for values higher than 45%. Lastly, it is possible to observe that $n/(C_{iv})^d$ and W/C VEC curves have a very similar effect (concave shape) on UCS prediction of $JGLF$ (Lee et al., 2005).

All previous results, i.e., relative importance and averaged effect of the key input variables, were taken based on a 1-D *SA*, i.e., holding all variables at their mean values, ranging only one at each time. As known, such conditions rarely or even never happen in practical conditions. Therefore, and keeping in mind a more realistic and detailed analysis, we discuss some important observations taken from a 2-D *SA*, i.e., changing simultaneously two input variables, keeping the remaining ones at their mean values. This approach allows measuring the interaction level between variables and quantifying the average effect on UCS when two variables are changed simultaneously. Hence, we measured the interaction level of all variables with t and C_{iv} (the two most relevant variables in UCS prediction of *JGLF*) and plotted the *VEC* surfaces for: t and W/C; t and $n/(C_{iv})^d$; C_{iv} and W/C and C_{iv} and t (Tinoco et al., 2012b).

Figure 5.8 shows the relative importance of the interaction between all variables with t (Figure 5.8a) and C_{iv} (Figure 5.8b). In both situations the highest interaction is observed with W/C, presenting a relative importance around 14%. This observation shows that in spite of W/C being just the fourth variable with more impact in UCS prediction of *JGLF* (based on a 1-D *SA*), it should be taken into account in *JGLF* behaviour because it has a strong interaction with other variables, namely with t and C_{iv}. The highest interaction of

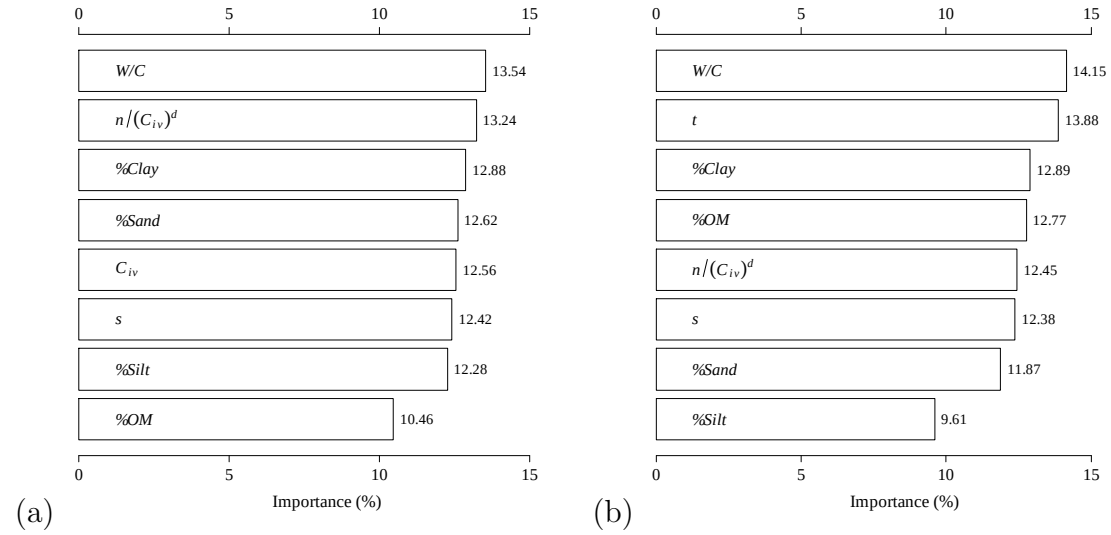


Figure 5.8: Interaction level between all variables with: a) t and b) C_{iv}, according to *SVM-UCS.Lab* model in UCS prediction of *JGLF*, measured by a 2-D *SA*

W/C with t can be explained if we take into account that the gain of strength is related to the decreasing of free water in the mixture (hardness process). This means that *JGLF* with high W/C ratio needs more time to obtain the same strength than for a lower W/C ratio. From this, it can be concluded that in order to obtain a faster hardness process, *JGLF* should be prepared with lower values of W/C ratio. On the other hand, the high

interaction between C_{iv} and W/C is related with mixture preparation. Normally, mixtures with high C_{iv} are prepared using grout slurry with lower W/C ratio. Therefore, is clear that C_{iv} and W/C has a strong interaction. Another interesting observation from these two figures is related with the soil properties. Once again, this input shows low impact on UCS prediction of $JGLF$ (within the database conditions).

Plotting the interaction effect between t and W/C in UCS prediction of $JGLF$, the VEC surface shown in Figure 5.9 is obtained. This surface shows precisely the high effect of the interaction between these two variables, evidenced by the high range of UCS values for different combinations of t and W/C (since 2 MPa to 9 MPa). Furthermore, it is also possible to observe that mixtures with high W/C ratios tend to stabilize for early ages. Based on VEC surface of t and $n/(C_{iv})^d$ plotted in Figure 5.10, we can see the high

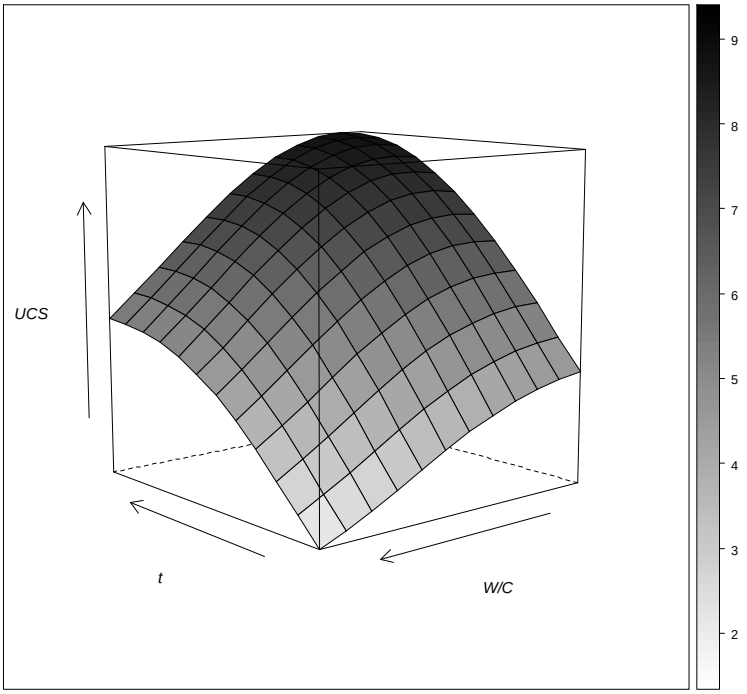


Figure 5.9: VEC surface for t and W/C interaction in UCS prediction of $JGLF$, according to *SVM-UCS.Lab* model, quantified by 2-D *SA*

impact interaction that these two variables also have in UCS prediction of $JGLF$ (UCS range from 2 MPa to 9 MPa). In addition, it is observed that the effect of t on UCS is more pronounced for lower values of $n/(C_{iv})^d$ than for higher values. This means that for mixture with high porosity (or lower cement content) the UCS will just slight increase over time.

Figure 5.11 shows the VEC surfaces for C_{iv} and W/C and Figure 5.12 the equivalent representation for C_{iv} and t. In these two graphs, we can see once again the high impact

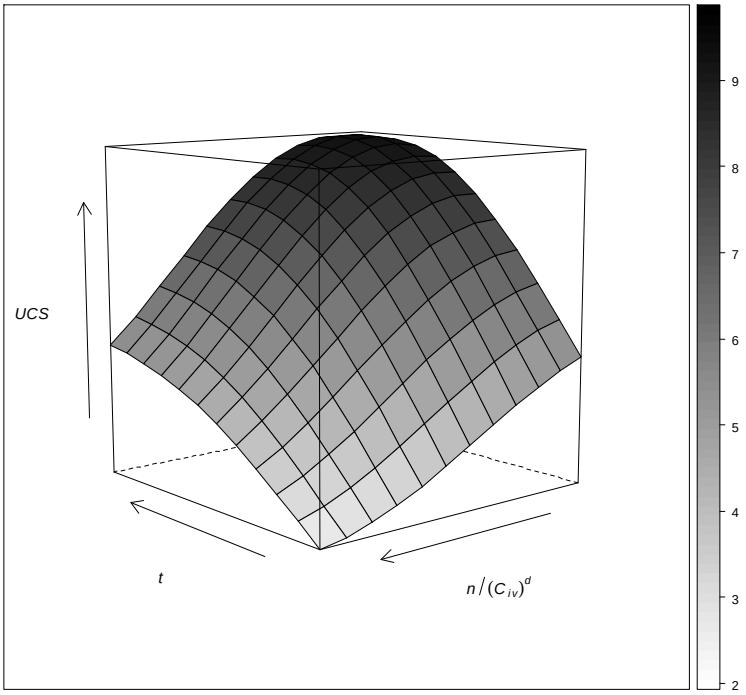


Figure 5.10: *VEC* surface for t and $n/(Civ)^d$ interaction in UCS prediction of $JGLF$, according to *SVM-UCS.Lab* model, quantified by 2-D *SA*

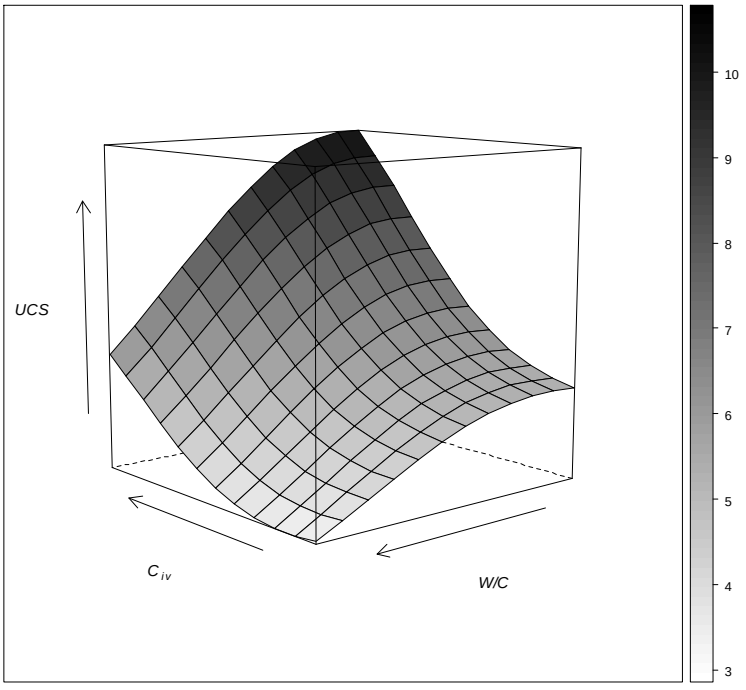


Figure 5.11: *VEC* surface for C_{iv} and W/C interaction in UCS prediction of $JGLF$, according to *SVM-UCS.Lab* model, quantified by 2-D *SA*

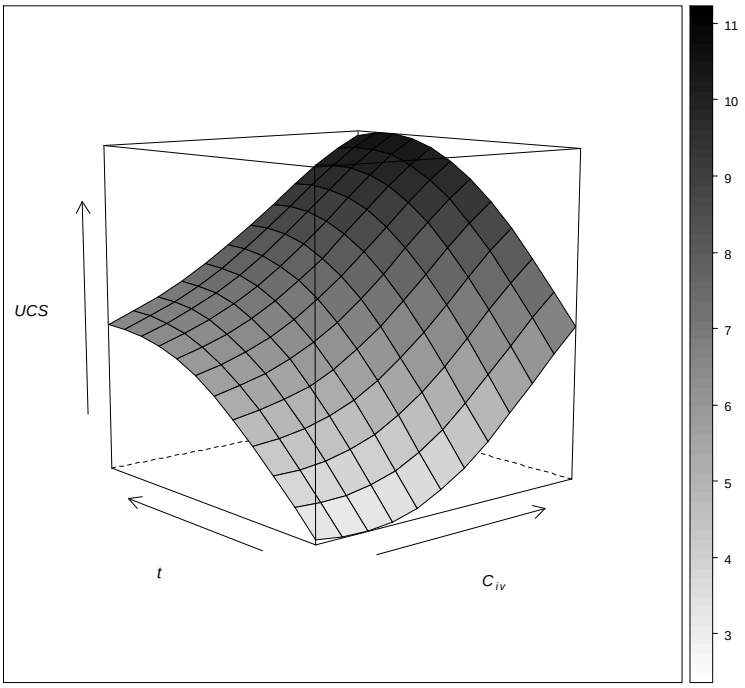


Figure 5.12: *VEC* surface for C_{iv} and t interaction in UCS prediction of *JGLF*, according to *SVM-UCS.Lab* model, quantified by 2-D *SA*

interaction of both W/C and t with C_{iv} (UCS range from 3 MPa to 11 MPa). The *VEC* surface of C_{iv} and W/C depicted in Figure 5.11 shows a fast increasing of UCS for higher values of C_{iv} when W/C decrease. This behaviour can be explained by the high amount of cement in such condition (high C_{iv} and low W/C ratio). Observing the *VEC* surface of C_{iv} and t plotted in Figure 5.12 we can see an almost linear effect of C_{iv} for advanced ages.

5.3 Deformability modulus prediction

5.3.1 Model performance

Similar to what was done in UCS study of *JGLF*, Table 5.4 compares (using *MAD*, *RMSE* and R^2 as a performance criteria) the model performance (*SVM* algorithm) achieved in E_0 prediction of *JGLF* by forward and backward *FS* approaches with a model where a set of eight variables were manually selected. However, in this case, the main purpose of this exercise is to validate/corroborate the set of variable chosen in UCS study. In other words, once the goal of the problem at hands is to predict the mechanical properties of a given material, in this case *JGLF*, it is more rational to consider the same set of

variables in both strength and stiffness studies. Therefore, the two *FS* approaches were just applied in E_0 study, since it is where the amount of data is higher and its purpose is just to compare the results with *UCS* study.

Analysing Table 5.4, we can observe that the achieved performance by *SVM* model in the three approaches is slightly higher than in *UCS* study (see Table 5.1). It is also observed that FFS leads to the most accurate model in E_0 prediction of *JGLF* using only three input variables, which seems to be unrealistic, although recognizing the importance of t and C_{iv} variables empirically known as relevant in soil-cement mixtures mechanical properties behaviour. However, and considering the above reasons, in the present research work the study of *JGLF* stiffness was performed using the same set of variables considered in *UCS* study, which is termed in Table 5.4 as MS_{E0l1} (Tinoco et al., 2011f).

Table 5.4: Model performance comparison of the two *FS* approaches implemented and that where the attributes were manually selected, in E_0 prediction of *JGLF*

Var	FFS	BFS	MS_{E0l1}
$\%Sand$	×	×	✓
$\%Silt$	×	×	✓
$\%Clay$	×	✓	✓
$\%MO$	×	✓	✓
C_{iv}	✓	✓	✓
$n/(C_{iv})^d$	×	✓	✓
t	✓	✓	✓
W/C	×	×	✓
ρ_d	×	✓	×
S_r	×	✓	×
CT	×	✓	×
e	×	✓	×
kg/m^3	×	✓	×
$\%Solo$	✓	✓	×
MAD	0.15 ± 0.00	0.19 ± 0.02	0.17 ± 0.00
RMSE	0.21 ± 0.00	0.26 ± 0.02	0.25 ± 0.01
R^2	0.97 ± 0.00	0.96 ± 0.01	0.96 ± 0.00

The average hyperparameters and fitting time values (and respective 95% level confidence intervals according to a t-student distribution) of the four *DM* algorithms trained for stiffness prediction of *JGLF* (i.e. *MR*, *ANN*, *SVM*and *FN*) are shown in Table 5.5. Table 5.6 shows and compares the performance of these algorithms trained for E_0, $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} prediction of *JGLF*. In order to facilitate the referencing to each trained

model, and following the same criterion used in *UCS* study, each one of the predictive models of stiffness of *JGLF* will be termed as shown in Table 5.7.

Table 5.5: Hyperparameters and computation time of each *DM* model for stiffness prediction of *JGLF*

Model	Hyperparameters	time (s)
FN-E_0.Lab	-	56.60 ± 0.00
FN-$E_{tg50\%}$.Lab	-	17.30 ± 0.00
FN-$E_{sec50\%}$.Lab	-	16.60 ± 0.00
FN-E_{max}.Lab	-	14.00 ± 0.00
MR-E_0.Lab	-	10.82 ± 0.02
MR-$E_{tg50\%}$.Lab	-	2.54 ± 0.01
MR-$E_{sec50\%}$.Lab	-	2.57 ± 0.02
MR-E_{max}.Lab	-	2.67 ± 0.01
ANN-E_0.Lab	$H = 7 \pm 1$	869.93 ± 0.95
ANN-$E_{tg50\%}$.Lab	$H = 3 \pm 1$	128.69 ± 0.67
ANN-$E_{sec50\%}$.Lab	$H = 5 \pm 1$	134.92 ± 1.06
ANN-E_{max}.Lab	$H = 3 \pm 1$	136.27 ± 0.25
SVM-E_0.Lab	$\gamma = 0.70 \pm 0.02$, $\epsilon = 0.06 \pm 0.00$	1168.92 ± 0.97
SVM-$E_{tg50\%}$.Lab	$\gamma = 0.74 \pm 0.06$, $\epsilon = 0.02 \pm 0.0$	190.07 ± 1.46
SVM-$E_{sec50\%}$.Lab	$\gamma = 0.36 \pm 0.05$, $\epsilon = 0.01 \pm 0.00$	202.39 ± 0.16
SVM-E_{max}.Lab	$\gamma = 0.39 \pm 0.02$, $\epsilon = 0.02 \pm 0.00$	201.79 ± 1.73

As done for *UCS* study, also here the mathematical expression proposed by *EC2* (CEN, 2004a) for deformability estimation of concrete was applied to stiffness prediction of *JGLF*. In addition, the analytical expression used by *MC90* (CEB-FIP, 1991) for concrete stiffness estimation, was also adapted to *JGLF* stiffness prediction. These two analytical models were only applied in E_0 study due to the following reasons. On one hand, for practical purposes, E_0 and $E_{tg50\%}$ are the two moduli currently used. The first one has demonstrated a good relationship with non-destructive tests with very small deformations, such as bender elements or sonic tests, while the second is a key geotechnical parameter that better defines the deformability properties of soil-cement mixtures and has important practical use. On the other hand, as depicted in Figure 5.13, a strong relationship is observed between E_0 and $E_{tg50\%}$ ($R^2 = 0.89$), after comparing these two moduli for all the tested samples for which such data are available. Moreover, both these expressions require some *a priori* information related with a given formulation. This means that to apply *EC2* expression it is need to know the 28 days deformability modulus of each formulation and in the case of *MC90* expression the 28 days strength of each

formulation. Therefore, and since the databases for $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} studies are rather small (see Table 4.1), not contemplating such information, these two approaches were only applied for E_0 prediction of *JGLF*.

Optimizing the two coefficients of Equation 3.18 using Young's modulus data of *JGLF*, the following model is achieved:

$$E(t) = \left(e^{\left(s \cdot \left[1 - \left(\frac{28}{t} \right)^{0.0011} \right] \right)} \right)^{959.56} \cdot E_{cm} \tag{5.5}$$

This model presents a performance equivalent to *DM* models, particularly *SVM-E_0.Lab* and *ANN-E_0.Lab*, as shown in Table 5.8 that compares the performance (based on *MAD*, *RMSE* and R^2 metrics) between *EC2-E_0.Lab*, *FN-E_0.Lab*, *MR-E_0.Lab*, *ANN-E_0.Lab* and *SVM-E_0.Lab* models in E_0 prediction of *JGLF*. Figure 5.14 corroborates such performance illustrating an excellent relationship between the experimental E_0 values and those predicted by the *EC2-E_0.Lab* model adapted to *JGLF*.

Table 5.6: Error metrics of all *DM* models for E_0, $E_{tg50\%}$, $E_{sec50\%}$, E_{max} prediction of *JGLF* (test set values, best values in **bold**)

Model	MAD	RMSE	R^2
FN-E_0.Lab	0.22 ± 0.00	0.30 ± 0.00	0.95 ± 0.00
FN-$E_{tg50\%}$.Lab	0.18 ± 0.00	0.24 ± 0.00	0.93 ± 0.00
FN-$E_{sec50\%}$.Lab	0.20 ± 0.00	0.25 ± 0.00	0.95 ± 0.00
FN-E_{max}.Lab	0.20 ± 0.00	0.27 ± 0.00	0.95 ± 0.00
MR-E_0.Lab	0.34 ± 0.00	0.48 ± 0.00	0.87 ± 0.00
MR-$E_{tg50\%}$.Lab	0.32 ± 0.00	0.40 ± 0.00	0.81 ± 0.00
MR-$E_{sec50\%}$.Lab	0.30 ± 0.00	0.39 ± 0.00	0.87 ± 0.00
MR-E_{max}.Lab	0.31 ± 0.01	0.42 ± 0.01	0.90 ± 0.00
ANN-E_0.Lab	$\mathbf{0.15} \pm 0.00$	$\mathbf{0.21} \pm 0.00$	$\mathbf{0.97} \pm 0.00$
ANN-$E_{tg50\%}$.Lab	0.20 ± 0.01	0.29 ± 0.01	0.90 ± 0.01
ANN-$E_{sec50\%}$.Lab	$\mathbf{0.12} \pm 0.01$	$\mathbf{0.16} \pm 0.01$	$\mathbf{0.98} \pm 0.00$
ANN-E_{max}.Lab	$\mathbf{0.18} \pm 0.01$	$\mathbf{0.26} \pm 0.02$	$\mathbf{0.96} \pm 0.01$
SVM-E_0.Lab	0.17 ± 0.00	0.25 ± 0.01	0.96 ± 0.00
SVM-$E_{tg50\%}$.Lab	$\mathbf{0.15} \pm 0.00$	$\mathbf{0.20} \pm 0.00$	$\mathbf{0.95} \pm 0.00$
SVM-$E_{sec50\%}$.Lab	0.15 ± 0.01	0.21 ± 0.03	0.96 ± 0.01
SVM-E_{max}.Lab	$\mathbf{0.18} \pm 0.00$	0.31 ± 0.01	0.94 ± 0.00

Both Table 5.8 and Figure 5.14 show that *EC2-E_0.Lab* model can be used to accurately predict Young's modulus of *JGLF*. However, looking to the arguments of Equation 5.5 (mathematical expression of *EC2-E_0.Lab* model), an important limitation is also

Table 5.7: Adopted nomenclature for model referencing in stiffness study of *JGLF*

Algorithm	Modulus	Designation
FN	E_0	$FN\text{-}E_0.Lab$
	$E_{tg50\%}$	$FN\text{-}E_{tg50\%}.Lab$
	$E_{sec50\%}$	$FN\text{-}E_{sec50\%}.Lab$
	E_{max}	$FN\text{-}E_{max}.Lab$
MR	E_0	$MR\text{-}E_0.Lab$
	$E_{tg50\%}$	$MR\text{-}E_{tg50\%}.Lab$
	$E_{sec50\%}$	$MR\text{-}E_{sec50\%}.Lab$
	E_{max}	$MR\text{-}E_{max}.Lab$
ANN	E_0	$ANN\text{-}E_0.Lab$
	$E_{tg50\%}$	$ANN\text{-}E_{tg50\%}.Lab$
	$E_{sec50\%}$	$ANN\text{-}E_{sec50\%}.Lab$
	E_{max}	$ANN\text{-}E_{max}.Lab$
SVM	E_0	$SVM\text{-}E_0.Lab$
	$E_{tg50\%}$	$SVM\text{-}E_{tg50\%}.Lab$
	$E_{sec50\%}$	$SVM\text{-}E_{sec50\%}.Lab$
	E_{max}	$SVM\text{-}E_{max}.Lab$

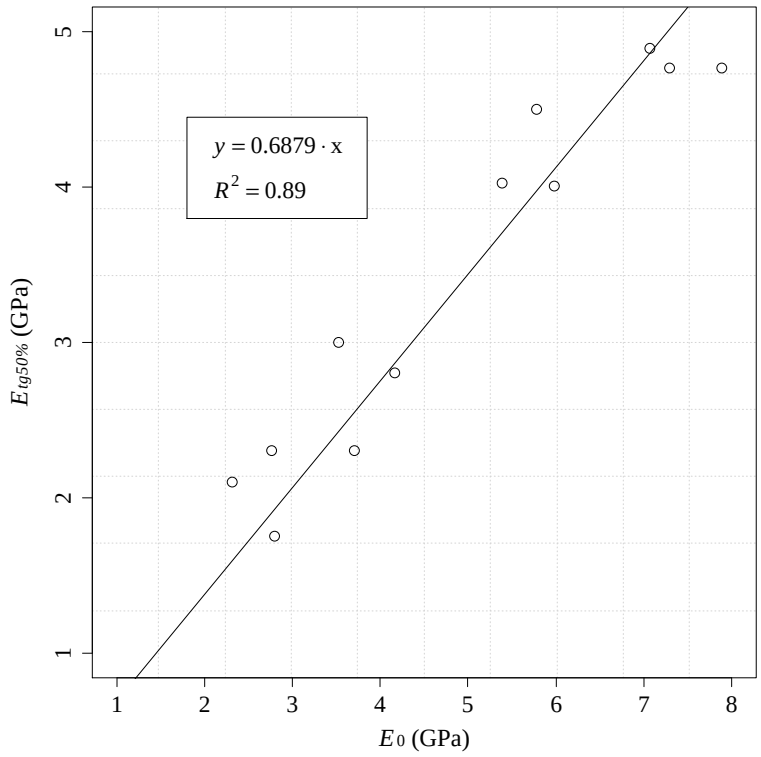


Figure 5.13: Relationship between E_0 and $E_{tg50\%}$ of *JGLF*, illustrating a strong correlation

Table 5.8: Comparison of the performance between the four *DM* models and the *EC2* analytical approach in E_0 prediction of *JGLF*, based on *MAD*, *RMSE* and R^2 metrics (mean values and 95% level confidence intervals according to a t-student distribution)

Metric	*EC2-E_0.Lab*	*FN-E_0.Lab*	*MR-E_0.Lab*	*ANN-E_0.Lab*	*SVM-E_0.Lab*
MAD (GPa)	0.16 ± 0.00	0.22 ± 0.00	0.34 ± 0.00	0.15 ± 0.00	0.17 ± 0.00
RMSE (GPa)	0.25 ± 0.00	0.30 ± 0.00	0.48 ± 0.00	0.21 ± 0.00	0.25 ± 0.01
R^2	0.96 ± 0.00	0.95 ± 0.00	0.87 ± 0.00	0.97 ± 0.00	0.96 ± 0.00

identified. The argument E_{cm} means that its application requires the knowledge of the deformability modulus at 28 days time of cure. Therefore, its application must be postponed for 28 days to perform stiffness tests on each formulation and quality control during construction. Nevertheless, it should be mentioned that ongoing research to predict stiffness based on earlier measurements (Azenha et al., 2011) will most likely eliminate this problem in the future.

As previously mentioned, the mathematical expression proposed by *MC90* was also applied to Young's modulus prediction of *JGLF*. Therefore, the three coefficients of Equation 3.19, were adapted to *JGLF* using E_0 data. However, contrary to the *EC2-E_0.Lab* model, in this case the performance achieved was very poor, as shown in Table 5.9 (Tinoco et al., 2010b). After analysing the *MC90* analytical expression (Equation 3.19), it was observed that the coefficient E_{co} should not be constant as initially considered and defined in CEB-FIP (1991). Thus, this coefficient was taken for each laboratory formulation (see Table 5.9), keeping the remaining coefficients at the following values: $a = 1/2$, $b = 1/2$ and $c = 1/3$. However, even when considering different values for E_{c0} coefficient, the performance achieved was worse than the *EC2-E_0.Lab* model. Moreover, and similarly to *EC2* analytical model, the *MC90* approach also requires laboratory tests to quantify UCS of each formulation at 28 days time of cure (f_{cm} argument). Therefore, it contains the same limitations previously explained related to its application during the project level.

Comparing the mathematical expression of both *EC2* and *MC90* approaches, it is observed that the main differences are related with E_{cm} and f_{cm} parameters, respectively. Hence, and as expected, we can conclude that E_0 prediction of *JGLF* based on E_{cm} is more reliable than on f_{cm} of each formulation. However, the prediction of *JGLF* stiffness based on the respective strength values (as considered by *MC90* approach), has a value in a practical application and therefore should be adopted. An attempt toward to this goal is present in Section 5.4.

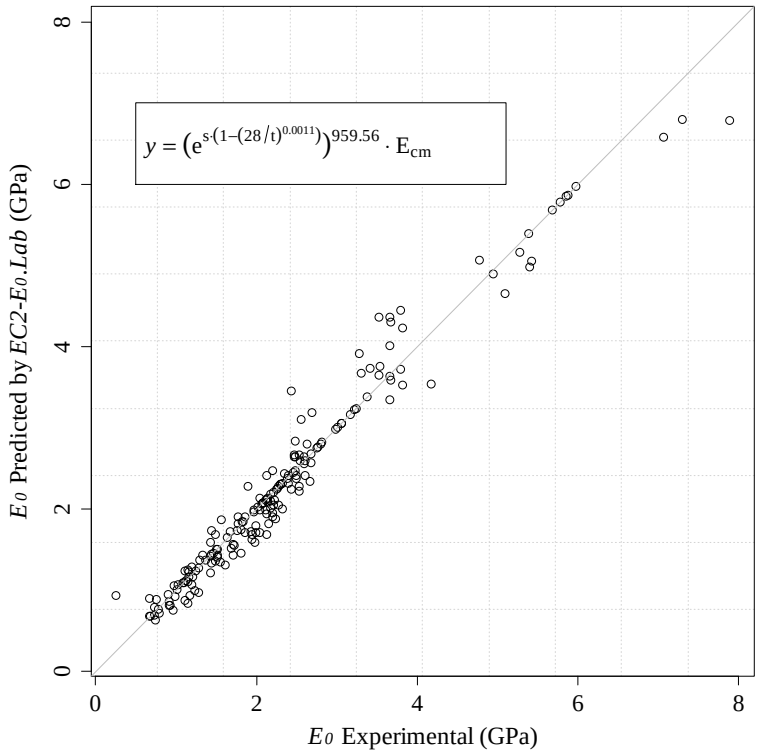


Figure 5.14: Relationship between E_0 experimental versus predicted values by $EC2\text{-}E_0.Lab$ model

The mathematical expression formulated in Equation 5.1 demonstrated a good performance in strength prediction of $JGLF$. Therefore, it was also applied in the study of $JGLF$ stiffness. Hence, it was applied the minimization problem formulated in Equation 5.2 and the FN algorithm to optimize the coefficients of Equation 5.1 in order to predict E_0, $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} of $JGLF$. The optimized coefficients (using the Leave-One-Out approach), related to each moduli predictive model are summarized in Table 5.10.

Table 5.9: Metrics values in $MC90\text{-}E_0.Lab_{adapted}$ and $MC90\text{-}E_0.Lab_{modified}$ models for E_0 prediction of $JGLF$

Metric	$MC90_{adapted}$	$MC90_{modified}$								
		LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9
MAD (GPa)	0.84	0.33	0.82	0.24	0.32	0.13	0.18	0.21	0.15	0.15
RMSE (GPa)	1.11	0.45	1.01	0.31	0.43	0.17	0.23	0.26	0.18	0.22
R^2	0.48	0.64	0.75	0.89	0.93	0.92	0.93	0.53	0.80	0.48
E_{co} (GPa)	3.54	4.06	6.64	2.59	4.03	3.08	2.08	3.17	2.88	1.80

LF - Laboratory Formulation

Table 5.10: Optimized coefficients of Equation 5.1 to the prediction of *JGLF* stiffness, i.e., E_0, $E_{tg50\%}$, $E_{sec50\%}$ and E_{max}

Model	β_0	$\alpha_{\%Sand}$	$\alpha_{\%Silt}$	$\alpha_{\%Clay}$	$\alpha_{\%OM}$	$\alpha_{W/C}$	α_t	α_C	$\alpha_{n/(C_{iv})^d}$
FN-E_0.Lab	10.0^{10}	-0.11	-9.80	4.60	-1.99	-1.03	0.23	1.11	-0.73
FN-$E_{tg50\%}$.Lab	2.47^{3}	-0.10	-1.63	-1.40	-0.26	-0.61	0.24	1.02	0.02
FN-$E_{sec50\%}$.Lab	6.91^{2}	-0.09	0.46	-2.65	0.12	-0.73	0.18	0.90	-0.49
FN-E_{max}.Lab	10.0^{10}	-0.03	-8.58	4.92	-1.72	-0.57	0.13	0.94	-1.96

As shown in Table 5.6, all *DM* algorithms (i.e., *MR*, *ANN*, *SVM* and *FN*) achieved once again a good performance on *JGLF* stiffness prediction, particularly *ANN* and *SVM* algorithms.

Figure 5.15 compares all data-driven models in E_0 prediction of *JGLF* based on the *REC* curves, underling the superiority of *ANN-E_0.Lab*, *SVM-E_0.Lab* models on such task. The Scatterplots of *MR-E_0.Lab*, *ANN-E_0.Lab*, *SVM-E_0.Lab* and *FN-E_0.Lab* models are shown in Figure 5.16.

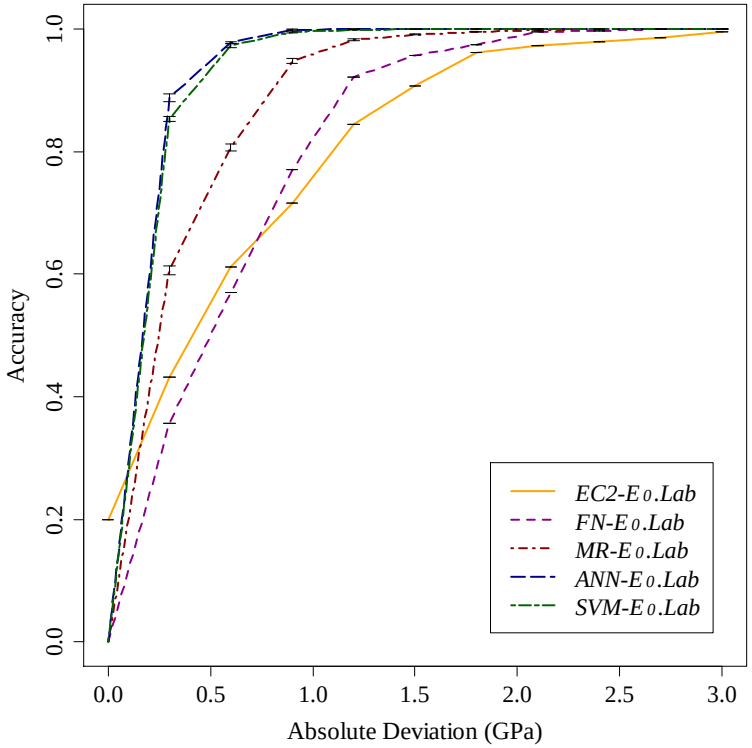


Figure 5.15: *REC* curves of *MR-E_0.Lab*, *ANN-E_0.Lab*, *SVM-E_0.Lab*, *FN-E_0.Lab* and *EC2-E_0.Lab* models, comparing its performance in E_0 prediction of *JGLF*

Considering the performance achieved by all *DM* models in stiffness prediction summarized in Table 5.6 using *MAD*, *RMSE* and R^2 as a performance criteria, as well as

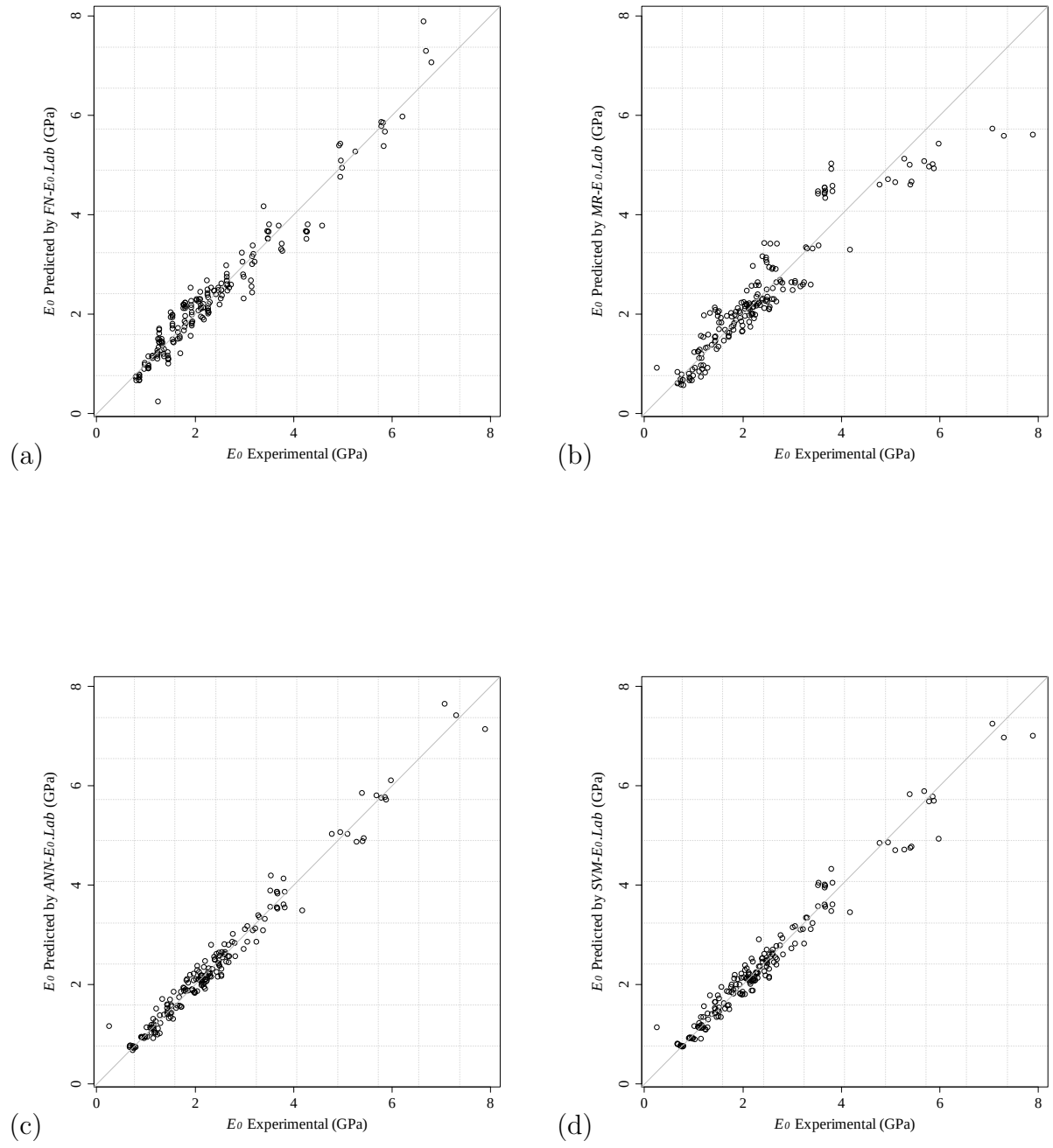


Figure 5.16: Relationship between E_0 experimental versus predicted values by: a) *FN-*E_0*.Lab*, b) *MR-*E_0*.Lab*, c) *ANN-*E_0*.Lab* and d) *SVM-*E_0*.Lab* models

the relationships depicted in Figure 5.16 for the particular case of E_0 study, *ANN* models seems to be a little more accurate than *SVM*. However, taken into account that *SVM* was used as a reference model in *UCS* study, this algorithm will be also used as reference in *JGLF* stiffness study, particularly on model interpretability, where this algorithm showed once again a good interpretation of *JGLF* behaviour, as discussed in Section 5.3.2.

Following this observation, Figure 5.17a compares *SVM* models performance in stiffness prediction of *JGLF*, showing that *SVM* algorithm is able to predict almost with the same performance either E_0, $E_{tg50\%}$, $E_{sec50\%}$ or E_{max} of *JGLF* (Tinoco et al., 2011a, 2010a, 2011f). The plots b), c) and d) of Figure 5.17 corroborate the high learning capabilities of *SVM* algorithm, even when the database that supports the learning is rather small.

5.3.2 Model interpretability

Keeping in mind a better understanding and interpretation of the developed *DM* models for *JGLF* stiffness prediction, a *GSA* method was applied over such models. Hence, a 1-D and 2-D *SA* was applied aiming to measure the relative importance of each variable, as well as its effect in stiffness behaviour of *JGLF*. Moreover, and following previous observations related to the practical importance of E_0 and the learning capabilities of *SVM* algorithm, a particular emphasize was given to *SVM-*E_0*.Lab* model.

Through the application of a 1-D *SA*, the relative importance of each variable, according to each one of the four *DM* models trained for E_0 prediction of *JGLF* (i.e., *FN-*E_0*.Lab*, *MR-*E_0*.Lab*, *ANN-*E_0*.Lab* and *SVM-*E_0*.Lab* models), is illustrated in Figure 5.18. Under a quick analysis, it is clear that *FN-*E_0*.Lab* model is unsuitable to predict E_0 of *JGLF* over time, despite of its good predictive performances based on the *MAD*, *RMSE* and R^2 metrics, as well as the results obtained for the *UCS* study. According to this model, E_0 of *JGLF* is only controlled by *%OM* and *%Silt* of the soil, which does not make sense and is unrealistic in soil-cement mixtures studies. These results stress the importance of involving domain experts in *DM* projects (as suggested by the *CRISP-DM* methodology).

According to *MR-*E_0*.Lab* model, used mainly for a baseline comparison, the Young's modulus of *JGLF* is almost only conditioned by soil properties, namely by its sand, silt and clay content. As expected, these results point out that the relationship between E_0 of *JGLF* and its contributing factors follows nonlinear laws and consequently cannot be described by a linear model. Analysing the relative importance of each variable according to *ANN-*E_0*.Lab* and *SVM-*E_0*.Lab* models, one can see that $n/(C_{iv})^d$ and t are key parameters in E_0 prediction of *JGLF*. Furthermore, it is also observed a strong influence of the soil properties, mainly according to *ANN-*E_0*.Lab* model. There results are in agreement

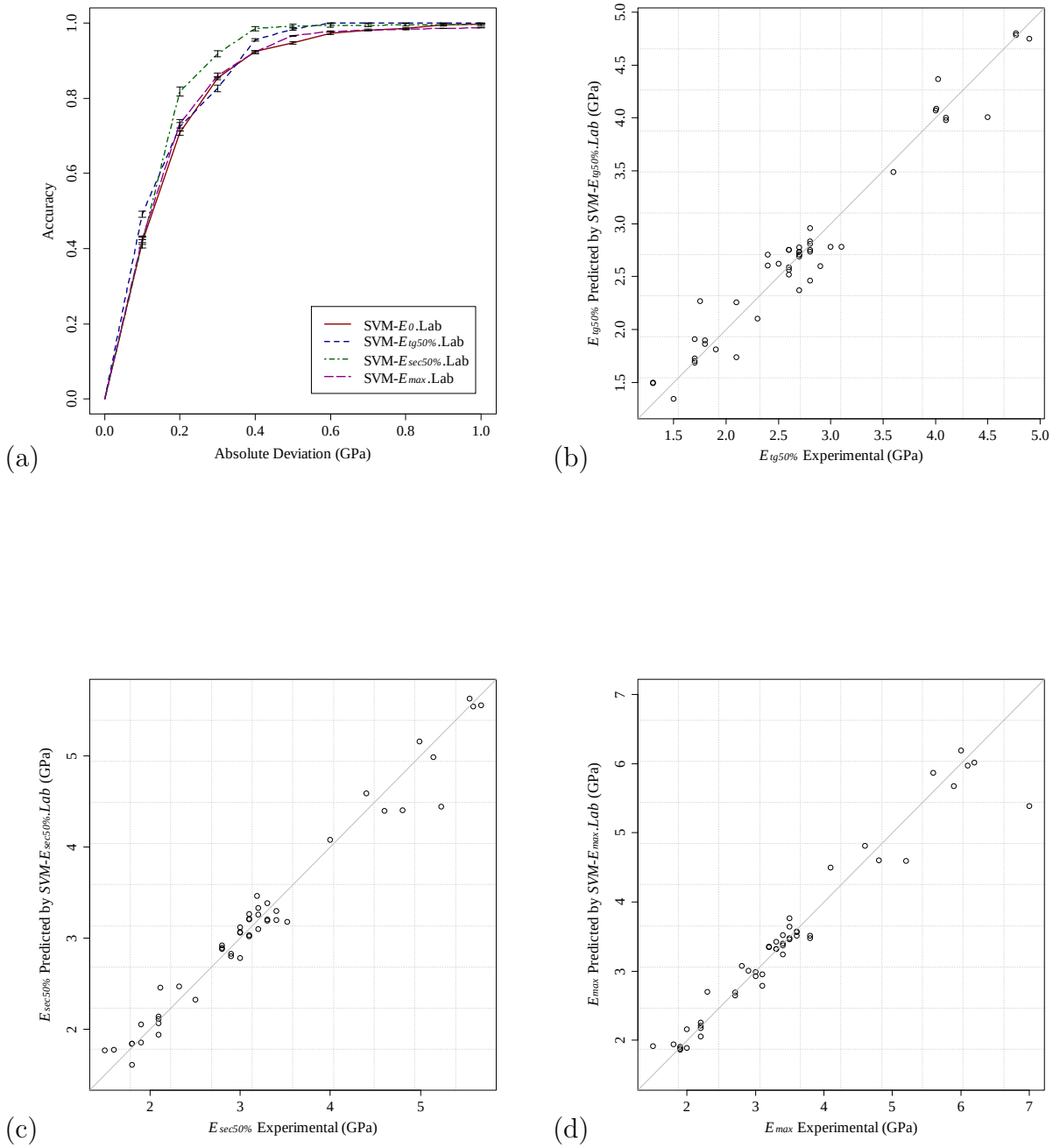


Figure 5.17: Stiffness prediction performance of *JGLF*: a) *REC* curves of *SVM* model for E_0, $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} prediction, b) scatterplot of $E_{tg50\%}$ according to *SVM-*$E_{tg50\%}$*.Lab*, c) scatterplot of $E_{sec50\%}$ according to *SVM-*$E_{sec50\%}$*.Lab* and d) scatterplot of E_{max} according to*SVM-*E_{max}*.Lab*

with the empirical knowledge related with soil-cement mixtures and also coincide with those observed in UCS study.

Figure 5.19 depicts the relative importance of each variable in $E_{tg50\%}$ prediction according to *FN-*$E_{tg50\%}$*.Lab*, *MR-*$E_{tg50\%}$*.Lab*, *ANN-*$E_{tg50\%}$*.Lab* and *SVM-*$E_{tg50\%}$*.Lab* (Tinoco et al., 2011a). Also here, *FN-*$E_{tg50\%}$*.Lab* model proves to be inappropriate for stiffness prediction. According to this model the soil properties have an influence around 77% in $E_{tg50\%}$ prediction that seems to be excessive. Also *MR-*$E_{tg50\%}$*.Lab* model shows the same behaviour, giving an excessive importance to soil properties in $E_{tg50\%}$ prediction, following the same approach than in E_0 study. Following the relative importance ranking according to *ANN-*$E_{tg50\%}$*.Lab* and *SVM-*$E_{tg50\%}$*.Lab* models, its shown that $n/(C_{iv})^d$ is the key variable in $E_{tg50\%}$ prediction. The soil properties, namely $\%Clay$, and C_{iv} also have a strong influence in $E_{tg50\%}$ prediction. It is still interesting to observe that the t effect on $E_{tg50\%}$ prediction is almost insignificant, even according to *ANN-*$E_{tg50\%}$*.Lab* and *SVM-*$E_{tg50\%}$*.Lab* models. This behaviour is understood and explained by the range of t in the dataset used during the training phase of the models (see Table A.3) of Appendix A. As shown in this table, this variable only ranges from 28 to 84 days. On the other hand, it is known that in cementitious mixtures (including soil-cement mixtures), t performs an important role (in both strength and stiffness behaviour), mainly for $t \leq 28$.

Figures 5.20 and 5.21 show the relative importance of each variable in $E_{sec50\%}$ and E_{max} prediction according to *SVM-*$E_{sec50\%}$*.Lab* and *SVM-*E_{max}*.Lab* respectively, and measured by a 1-D *SA* (Tinoco et al., 2011f). Again, similar conclusions are drawn. On one hand, both *MR* and *FN* models consider the soil properties as the most important factor in $E_{sec50\%}$ and E_{max} prediction of *JGLF*. On the other hand, and according to *ANN* and *SVM* models, the key variables in stiffness prediction of *JGLF* are the relation $n/(C_{iv})^d$ and the t.

Making a global appreciation of all data-driven models for stiffness prediction of *JGLF*, it should be stressed that both *MR* and *FN* models are unable to understand the stiffness behaviour of *JGLF*. On the other hand, it should be remarked the high learning capabilities of *ANN* and *SVM* algorithms in stiffness prediction of *JGLF*. Comparing these two algorithms, it can be concluded that *SVM* is more interesting, leading to better results. While achieving a similar predictive performance, the relative input importances according to *SVM* models is more coherent in terms of what is known empirically in the *JG* domain. Moreover, also in the study of *UCS* of *JGLF*, *SVM* achieved the most interesting results.

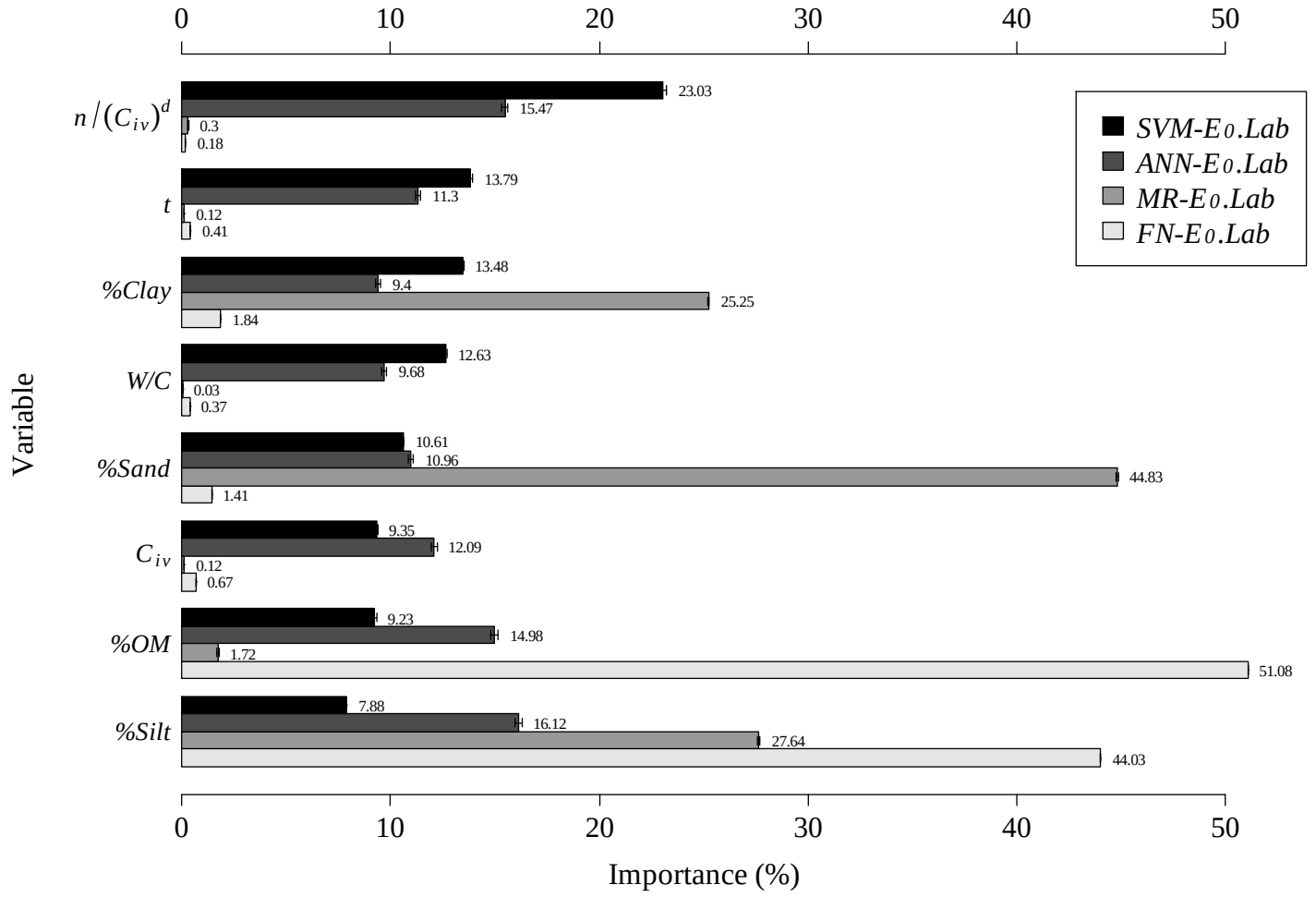


Figure 5.18: Relative importance of each input variable quantified by 1-D *SA*, comparing *MR-E*$_0$*.Lab*, *ANN-E*$_0$*.Lab*, *SVM-E*$_0$*.Lab* and *FN-E*$_0$*.Lab* models

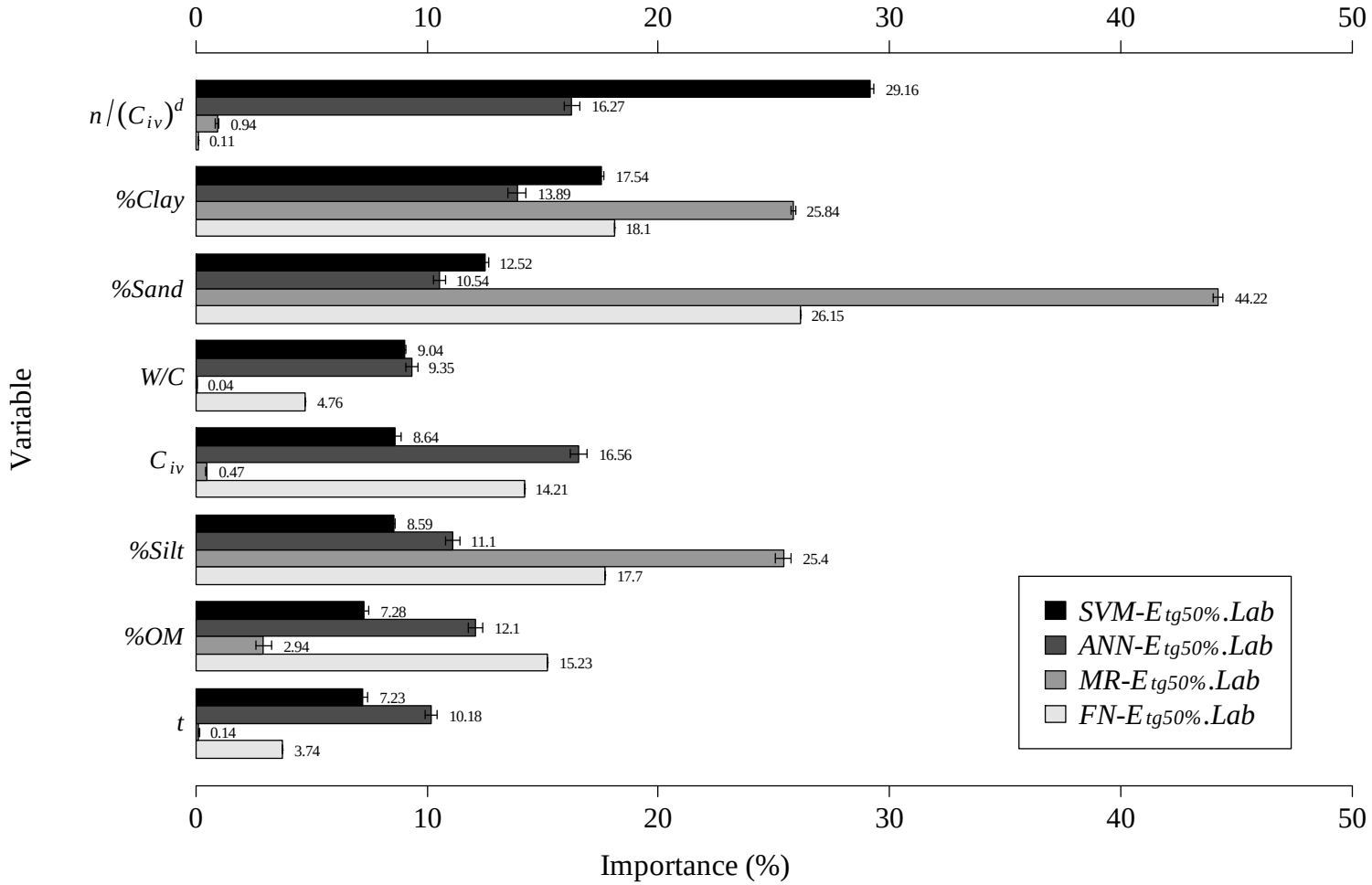


Figure 5.19: Relative importance of each input variable quantified by 1-D *SA*, comparing *MR-*$E_{tg50\%}$*.Lab*, *ANN-*$E_{tg50\%}$*.Lab*, *SVM-*$E_{tg50\%}$*.Lab* and *FN-*$E_{tg50\%}$*.Lab* models

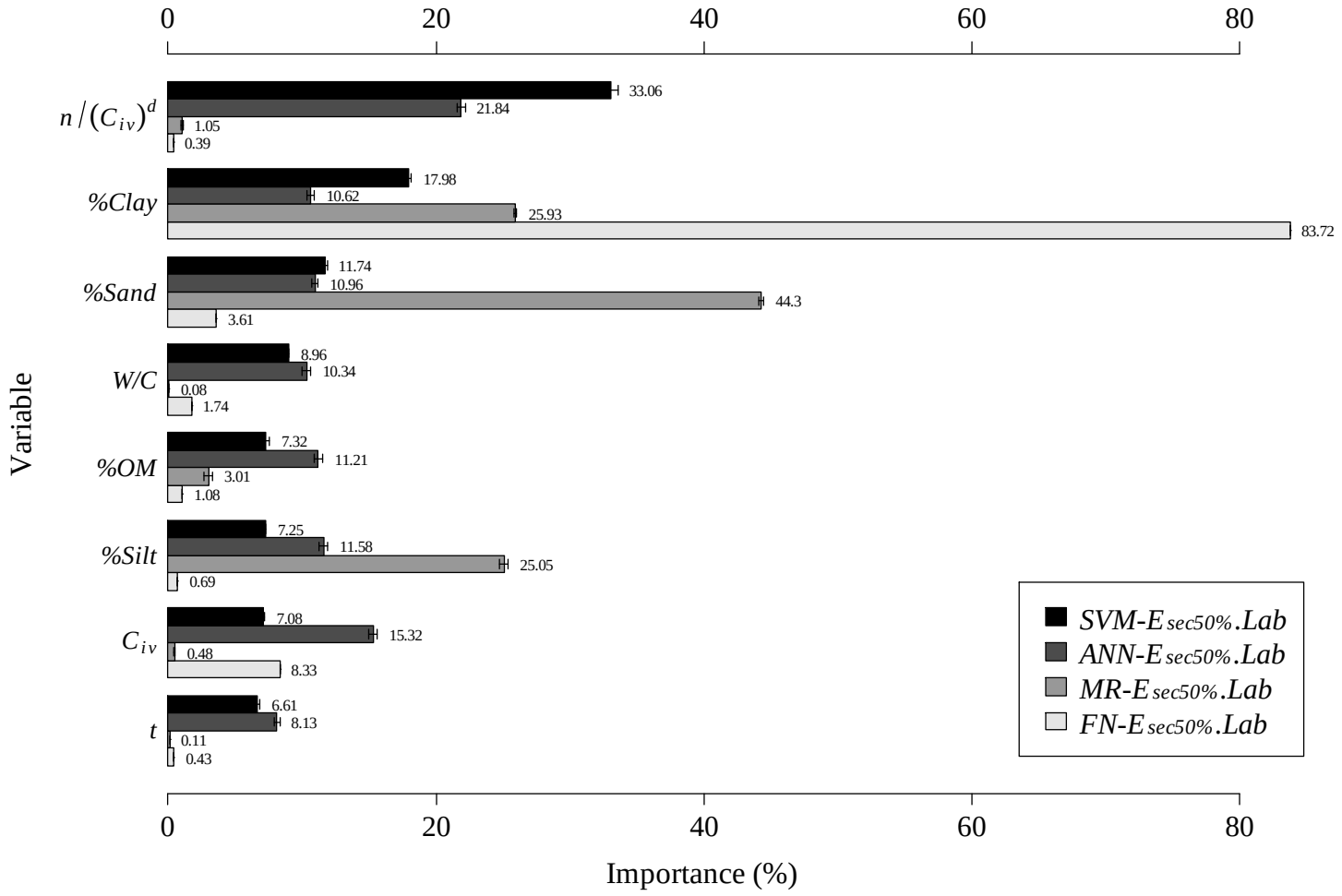


Figure 5.20: Relative importance of each input variable quantified by 1-D *SA*, comparing *MR-*$E_{sec50\%}$*.Lab*, *ANN-*$E_{sec50\%}$*.Lab*, *SVM-*$E_{sec50\%}$*.Lab* and *FN-*$E_{sec50\%}$*.Lab* models

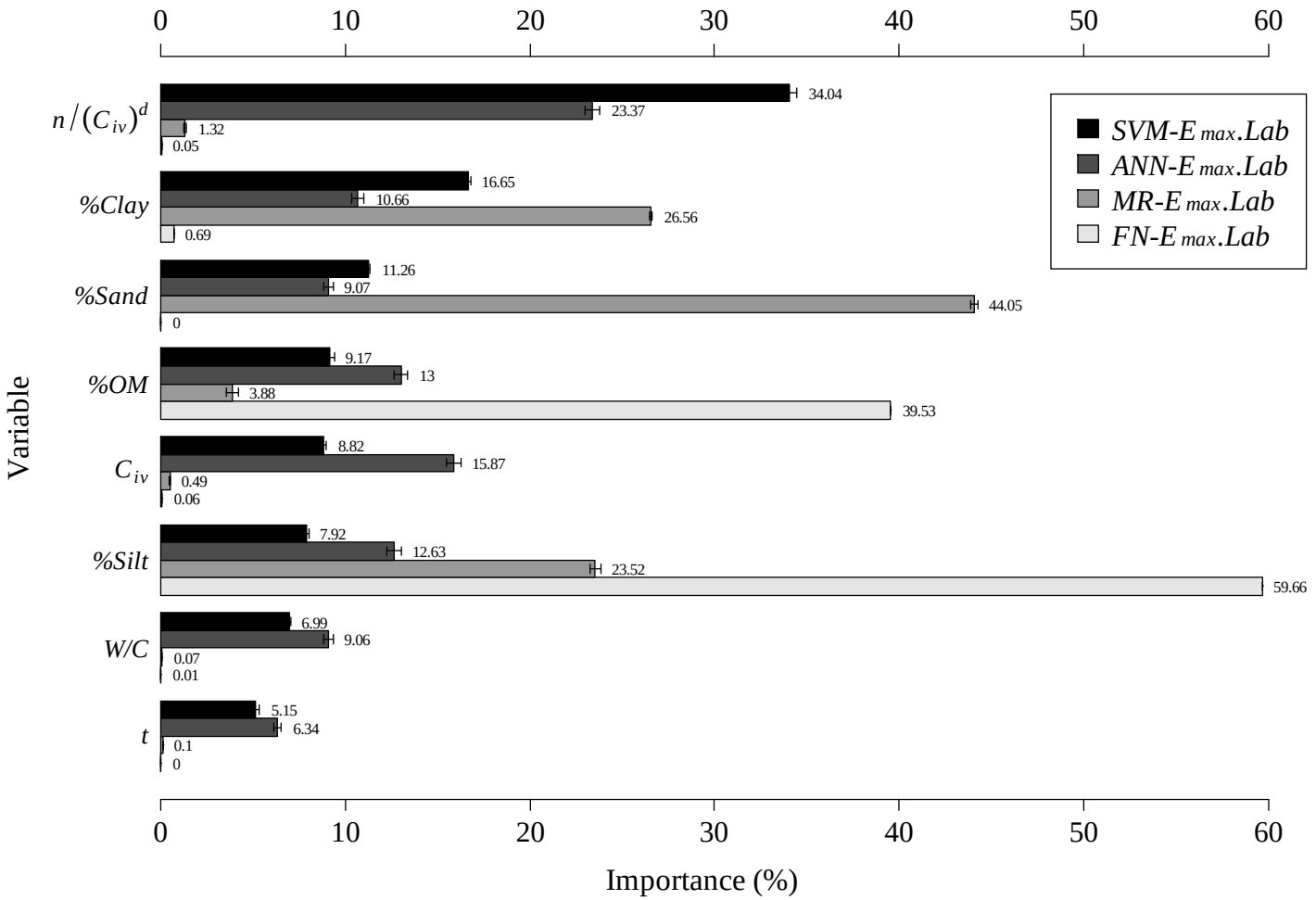


Figure 5.21: Relative importance of each input variable quantified by 1-D *SA*, comparing *MR-E_{max}.Lab*, *ANN-E_{max}.Lab*, *SVM-E_{max}.Lab* and *FN-E_{max}.Lab* models

Therefore, and using the *SVM* models as a reference (i.e. *SVM-E_0.Lab*, *SVM-$E_{tg50\%}$.Lab*, *SVM-$E_{sec50\%}$.Lab* and *SVM-E_{max}.Lab*), Figure 5.22 compares the relative importance of each variable in stiffness prediction of *JGLF*. This figure shows that the relation $n/(C_{iv})^d$ and the soil properties, mainly its clay and sand content, are the key variables in *JGLF* stiffness prediction. Particularly in the study of $E_{tg50\%}$, the W/C ratio also evidences a strong influence. Figure 5.22 also underlines previous observations related with t effect, showing that t is only preponderant (the second more relevant variable) in E_0 study. This is precisely the only situation where this variable take values lower than 28 days time of cure ($3 \leq t \geq 56$), as shown in Table A.2 of Appendix A. Thus, these results come to corroborate the empirically knowledge related to the effect of the age of the mixture in cementitious mixtures, i.e., that t is more preponderant for ages lower than 28 days time of cure in both strength and stiffness prediction.

Model interpretability given by a *SA* in terms of the relative importance of each variable can be improved measuring the effect of the key variables on the target variable. Therefore, using *SVM* models as reference, particularly *SVM-E_0.Lab* model, we analysed the effect of the key variable in *JGLF* stiffness prediction, based on a 1-D and 2-D *SA*.

Figures 5.23 and 5.24 plot respectively the *VEC* curves of $n/(C_{iv})^d$ and $\%Clay$ variables (the two most relevant variables in stiffness prediction of *JGLF*), according to SVM-E_0.Lab, *SVM-$E_{tg50\%}$.Lab*, *SVM-$E_{sec50\%}$.Lab* and *SVM-E_{max}.Lab* models respectively. In both situations it is observed a decreasing on *JGLF* stiffness when $n/(C_{iv})^d$ or $\%Clay$ increase. On $\%Clay$ *VEC* curves it is observed a slight increase on *JGLF* stiffness for higher values of clay content. This phenomena is probably a consequence of the high amount of cement added when the soil treated has a high content of clay. Thus, it is anticipated that the stiffness will increase despite of the high amount of clay in the soil.

As illustrated in Figure 5.22, t and W/C also have an important influence in *JGLF* stiffness prediction, particularly in E_0 and $E_{tg50\%}$ respectively. Accordingly, Figures 5.25 and 5.26 plot the *VEC* curve for these two variables according to *SVM-E_0.Lab* and *SVM-$E_{tg50\%}$.Lab* models. In these graphs, for a given input, each plot shows the histogram (frequency values are shown at the right of the y-axis) and the *VEC* curves (predicted values, shown at the left of the y-axis) when the analytical test values (x-axis) are changed through their domain values (with l=6 levels). Since several experiments were held, vertical averaging is performed (with the respective 95 % confidence intervals) of all *VEC* curves. The main advantage of this representation is to easily compare the *VEC* curve and the histogram for a given attribute. As empirically expected, the *VEC* curve of t illustrated in Figure 5.25 shows an exponential shape, evidencing the higher effect of t until 28 days time of cure.

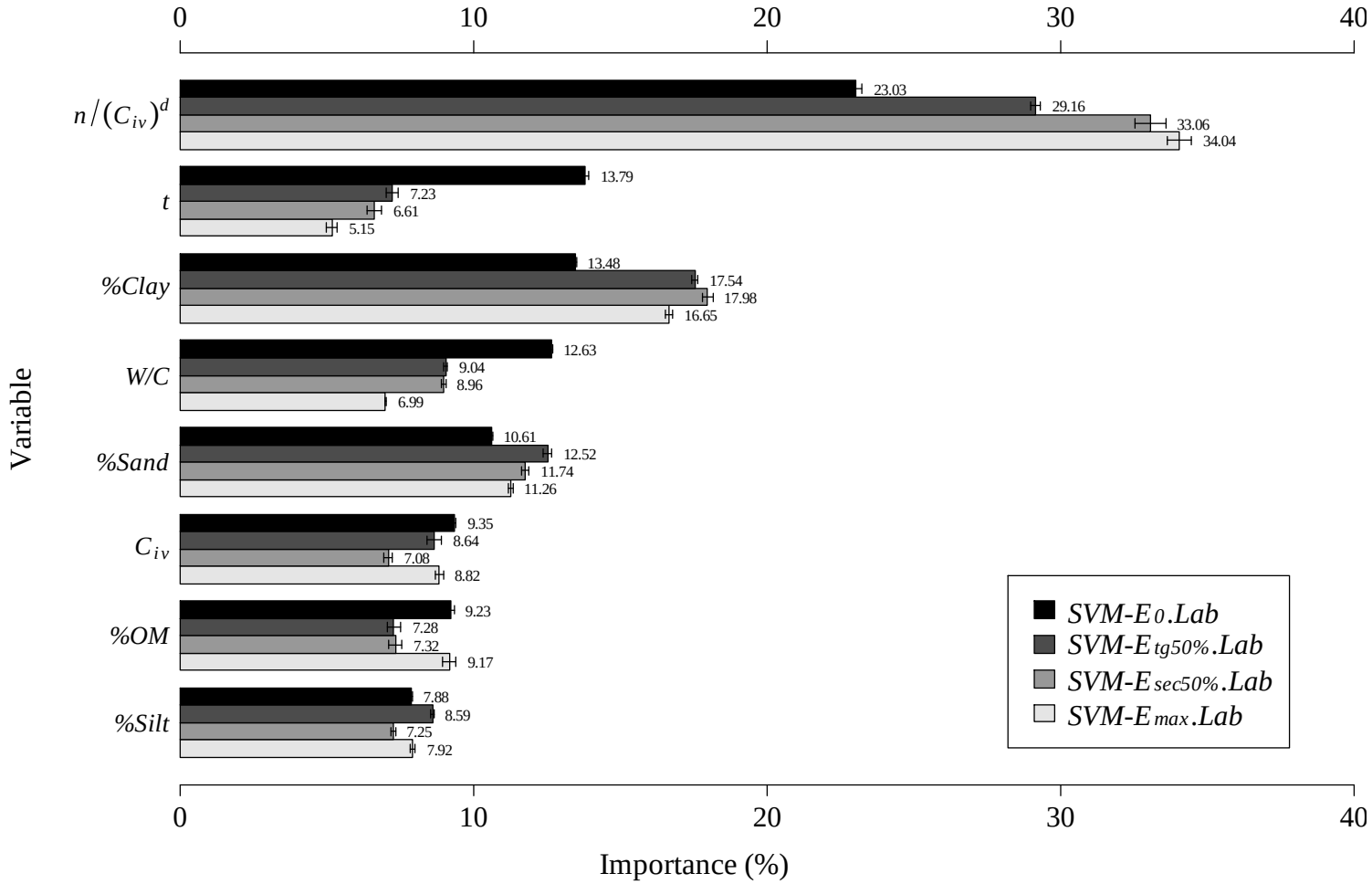


Figure 5.22: Relative importance of each variable in stiffness prediction according to *SVM-E_0.Lab*, *SVM-$E_{tg50\%}$.Lab*, *SVM-$E_{sec50\%}$.Lab* and *SVM-E_{max}.Lab* models, based on a 1-D *SA*

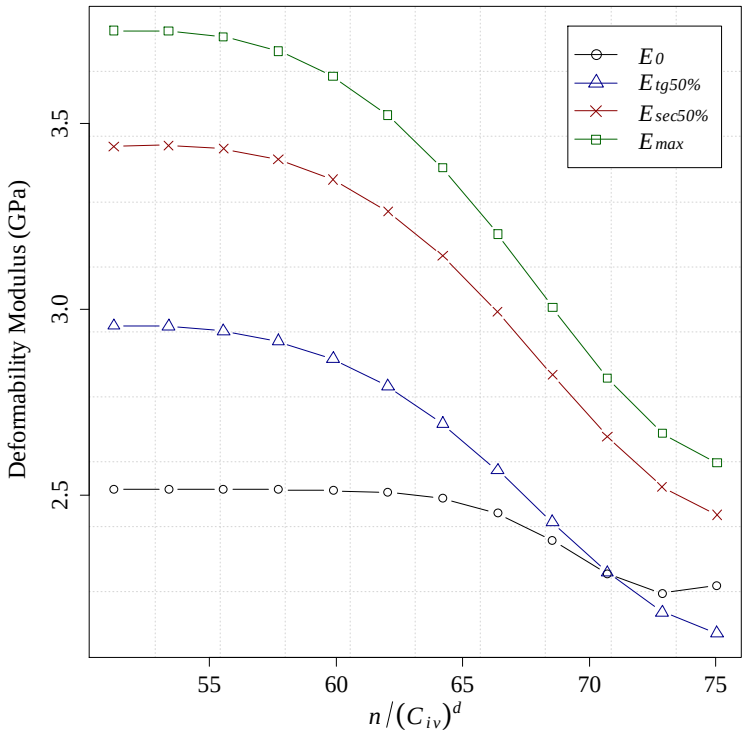


Figure 5.23: *VEC* curves of $n/(c_{iv})^d$ according to each *SVM* model in stiffness prediction of *JGLF*, quantified by 1-D *SA*

Relating to the W/C *VEC* curve, the slight increase of $E_{tg50\%}$ observed if Figure 5.26 for low values of W/C is probably related with the absence of data for such range of values, for what the model found some difficulties to learn. It is also observed an almost linear effect of W/C in $E_{tg50\%}$ prediction of *JGLF*.

Similar to what was executed in the *UCS* study, and in order to improve the interpretability of the models, a 2-D *SA* was performed, allowing to measure the interaction level between variables and quantify its average effect in stiffness prediction when two variables are changed simultaneously. Accordingly, Table 5.11 summarizes the interaction level between all variables with $n/(C_{iv})^d$ according to *SVM-*E_0*.Lab*, *SVM-*$E_{tg50\%}$*.Lab*, *SVM-*$E_{sec50\%}$*.Lab* and *SVM-*E_{max}*.Lab* models, after applying a 2-D *SA*. The $\%Clay$ and t are the two variables that have the higher overall interaction with $n/(C_{iv})^d$ in *JGLF* stiffness prediction. The strong interaction between $n/(C_{iv})^d$ and t helps to understand the less relative importance of t in $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} prediction (see Figure 5.22), and complement previous observations related with the range of t in this situations. The *VEC* contour plotted in Figure 5.27a, depicts the interaction effect between $n/(C_{iv})^d$ and $\%Clay$ in $E_{tg50\%}$ study according to *SVM-*$E_{tg50\%}$*.Lab* model, showing that the highest values of $E_{tg50\%}$ are achieved on samples with lower $n/(C_{iv})^d$ and prepared using soils with low clay content. In Figure 5.27b, it is plotted the effect of $n/(C_{iv})^d$ and t interaction in

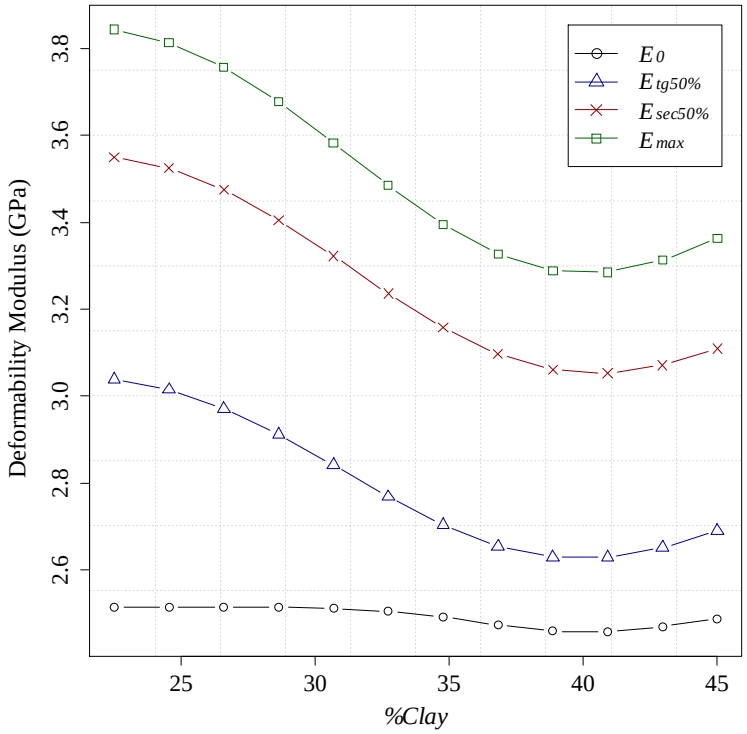


Figure 5.24: *VEC* curves of *%Clay* according to each *SVM* model in stiffness prediction of *JGLF*, quantified by 1-D *SA*

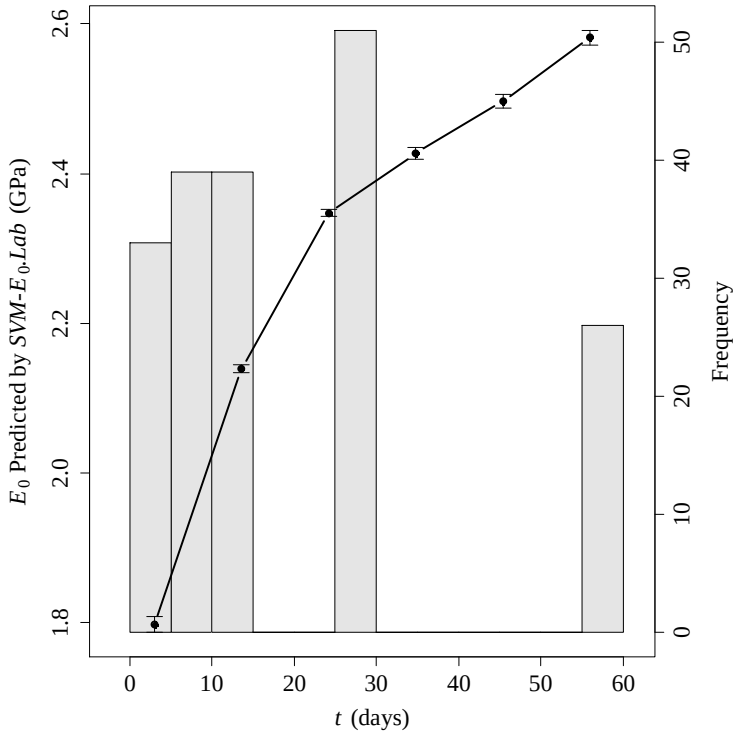


Figure 5.25: Vertical averaging of the *VEC* curves (points and whiskers) and histogram (in bars) according to *SVM-E_0.Lab* model for t variable in E_0 prediction of *JGLF*

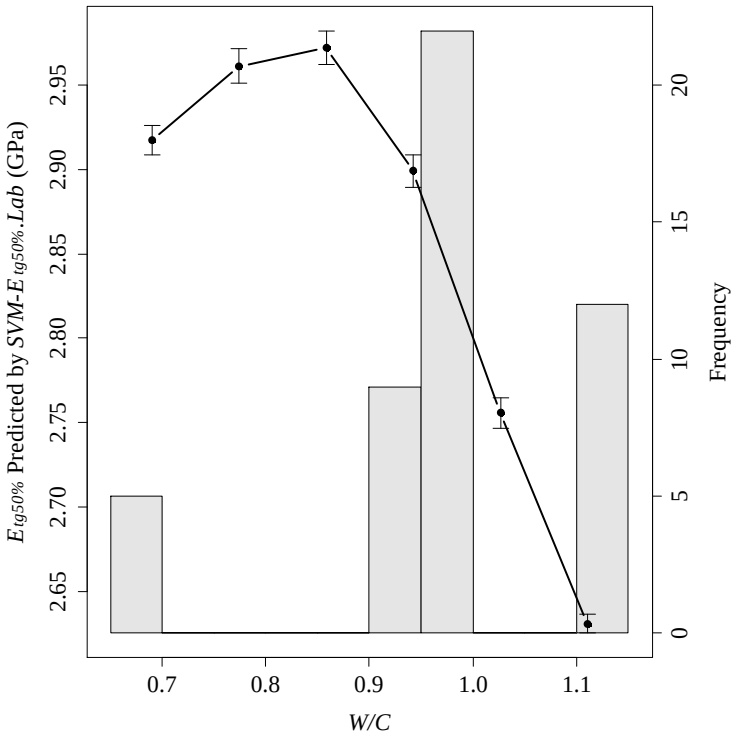
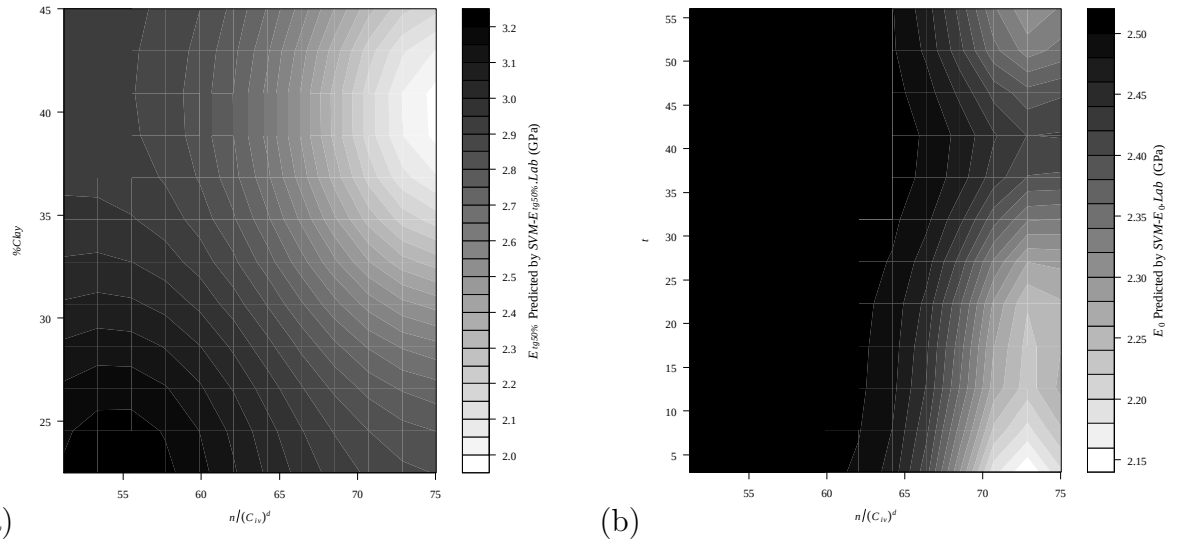


Figure 5.26: Vertical averaging of the *VEC* curves (points and whiskers) and histogram (in bars) according to *SVM-*$E_{tg50\%}$*.Lab* model for W/C variable in $E_{tg50\%}$ prediction of *JGLF*

E_0 prediction, where it is possible to observe a slightly influence of t in E_0 prediction of *JGLF* samples with low values of $n/(C_{iv})^d$.

Table 5.11: Interaction level between all variables with $n/(C_{iv})^d$ according to *SVM-*E_0*.Lab*, *SVM-*$E_{tg50\%}$*.Lab*, *SVM-*$E_{sec50\%}$*.Lab* and *SVM-*E_{max}*.Lab* models in stiffness prediction of *JGLF*, measured by a 2-D *SA*

Variable		t	c_{iv}	W/C	$\%Sand$	$\%Silt$	$\%Clay$	$\%OM$
$n/(C_{iv})^d$	E_0	**21.22**	10.60	14.48	15.86	10.45	17.94	9.45
	$E_{tg50\%}$	15.34	13.20	13.39	15.00	12.46	**16.48**	14.13
	$E_{sec50\%}$	15.57	13.49	13.18	14.92	12.44	**16.31**	14.10
	E_{max}	15.80	13.77	12.89	14.90	12.42	**16.16**	14.07



(a) (b)

Figure 5.27: *VEC* contour for: a) $n/(C_{iv})^d$ and $\%Clay$ interaction in $E_{tg50\%}$ prediction of *JGLF*, according to *SVM-*$E_{tg50\%}$*.Lab* model and b) $n/(C_{iv})^d$ and t interaction in E_0 prediction of *JGLF*, according to *SVM-*E_0*.Lab* model, quantified by a 2-D *SA*

5.4 Strength and stiffness relationship

As shown in Sections 5.2 and 5.3, an high performance was achieved, particularly by *SVM* algorithm, in both strength and stiffness prediction of *JGLF* (Tinoco et al., 2012e). Table 5.12 summarizes the metrics values (*MAD*, *RMSE* and R^2) of *SVM-UCS.Lab*, *SVM-*E_0*.Lab*, *SVM-*$E_{tg50\%}$*.Lab*, *SVM-*$E_{sec50\%}$*.Lab* and *SVM-*E_{max}*.Lab* models, comparing its performance. Using R^2 as performance criterion, this table underlines the high learning capabilities of *SVM* algorithm in the study of *JGLF* mechanical properties.

Table 5.12: Comparison of the performance of each *SVM* predictive model in UCS, E_0, $E_{sec50\%}$ and $E_{tg50\%}$ of *JGLF*, using *MAD*, *RMSE* and R^2 as performance criteria

Model	MAD	RMSE	R^2
SVM-UCS.Lab	0.55 ± 0.00	0.73 ± 0.00	0.93 ± 0.00
SVM-E_0*.Lab*	0.17 ± 0.00	0.25 ± 0.01	0.96 ± 0.00
SVM-$E_{tg50\%}$*.Lab*	0.15 ± 0.00	0.20 ± 0.00	0.95 ± 0.00
SVM-$E_{sec50\%}$*.Lab*	0.15 ± 0.01	0.21 ± 0.03	0.96 ± 0.01
SVM-E_{max}*.Lab*	0.18 ± 0.00	0.31 ± 0.01	0.94 ± 0.00

Figure 5.28, which compares the relative importance of each input variable according to *SVM* predictive models of UCS, E_0, $E_{tg50\%}$, $E_{sec50\%}$ and E_{max}, illustrates that the relation $n/(C_{iv})^d$ is the key variable in both mechanical properties prediction of *JGLF* (Tinoco

et al., 2012d). Moreover, in the UCS study the t and C_{iv} should be also taken into account. On the other hand, we can observe that the soil properties are apparently more relevant in stiffness prediction of *JGLF* than in strength study. This observation is explained if one takes into account that for low deformations, the grain size of the soil particles is responsible for the main resistance capacity of the material. After the grains broke, the cohesion is sustained by soil-cement matrix. So, after this time, the age of the mixture and the percentage of cement take the main role in the strength capacity of the soil-cement mixture. Furthermore, it should also be stressed that all conclusions herein pointed out underlie the characteristics of the database used during the learning of each mechanical properties studied. Indeed, it was shown that the range of some variables support some of the observations, such as the small influence of t in $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} prediction.

As underlined in Section 5.3.1, when *MC90* model results were analysed, the prediction of soil-cement mixtures stiffness (e.g. *JG* mixtures) based on strength values has an important practical application. It is known that the unconfined compression test is a standard, simple and relatively inexpensive way to assess *JG* soil improvement quality. However, the deformability properties of *JG* material are sometimes required for structure's serviceability evaluation (Gomes Correia, 2004). On the other hand, and keeping in mind that deformability tests are more expensive and require more time, it would be useful to predict *JGLF* stiffness based on an unconfined compression test in practice.

Therefore, a novel approach using *DM* techniques is proposed, aiming to predict *JGLF* stiffness based on the UCS of the respective mixture and considering elementary variables related to the mix properties. The proposed model is capable of predicting the E_0 of *JGLF* based on the $\%Clay$, C_{iv}, t and the UCS of the mix at the same age. The choice of these variables is supported, on one hand by the empirical knowledge related to soil-cement mixtures, particularly the variable t and, on the other hand by the comparison of the key variables in *JGLF* strength (Tinoco et al., 2012b) and stiffness prediction (see Figure 5.28). This comparison indicates that $\%Clay$ plays an important role in *JGLF* stiffness behaviour and is almost insignificant in UCS prediction, whereas C_{iv} is more important to *JGLF* strength prediction than deformability. This evidences that these two variables are key elements to determining the stiffness of a given *JGLF* based on its unconfined strength.

The proposed approach was developed based on a rather small database, containing only 11 samples, extracted from the main database of E_0 study. The database dimension can represent an important limitation in such circumstances because *DM* techniques are particularly designed to work with high amounts of data. However, the practical relevance of such approach justifies its use, even under such conditions.

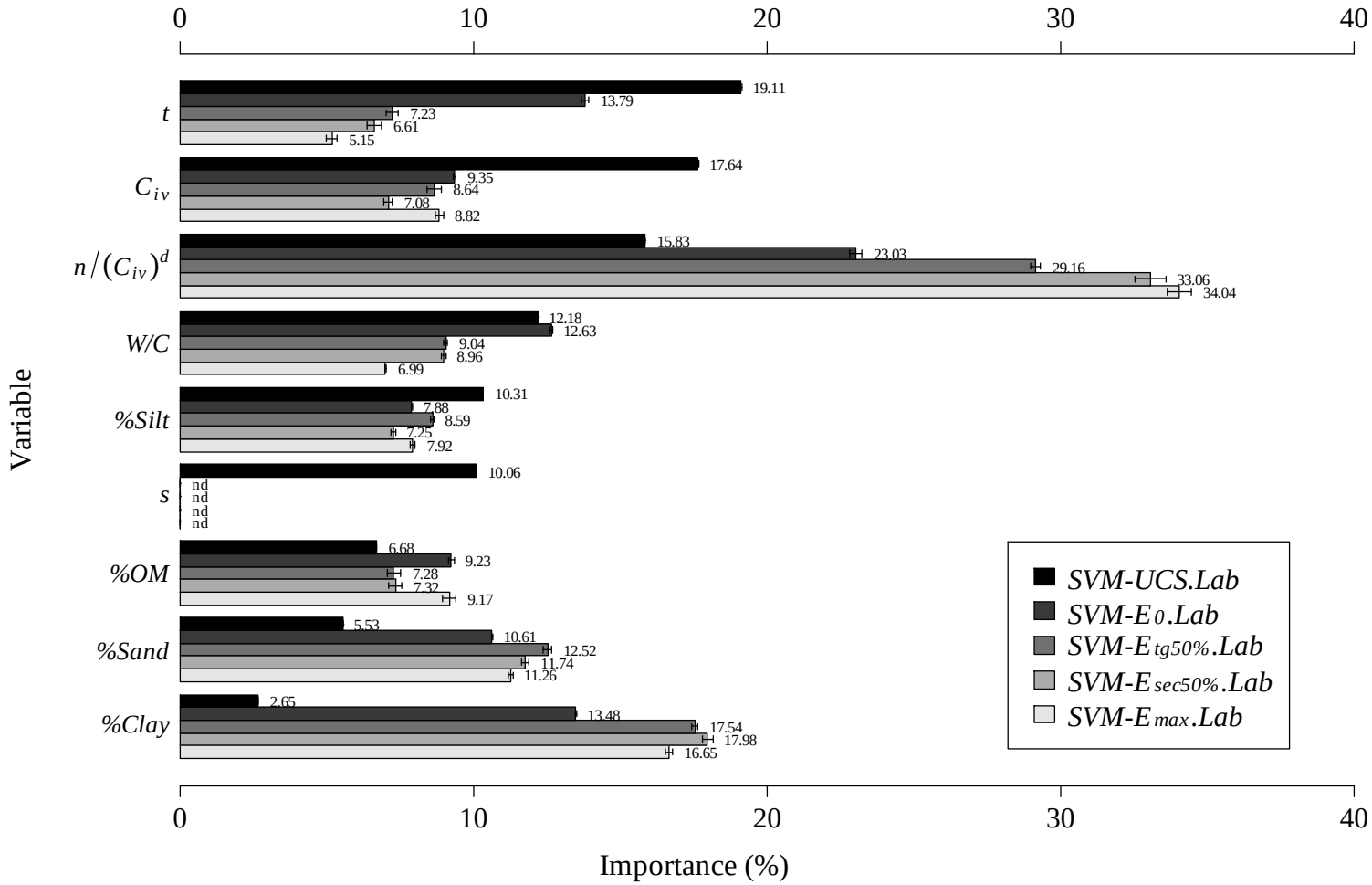


Figure 5.28: Comparison of the relative importance of each variable according to each *SVM* predictive model of UCS, E_0, $E_{tg50\%}$, $E_{sec50\%}$ and E_{max}

In order to optimize all available data, the leave-one-out procedure was applied to the model generalisation capacity assessment and during the learning phase (i.e., using only training data). Table 5.13 summarises the main statistics of the database's input and output variables used in this experiment.

Table 5.13: Summary of both input and output variable statistics of the database used in the experiments performed in Section 5.4, aiming to correlate strength and stiffness of *JGLF*

Variable	Minimum	Maximum	Mean	Standard Deviation
$\%Clay$	22.50	45.00	32.73	11.75
t	28.00	56.00	40.73	14.62
C_{iv}	35.85	36.83	36.30	0.51
UCS (MPa)	1.52	7.27	4.72	2.43
E_0 (GPa)	2.32	7.89	5.00	1.99

The results show an high performance by the *SVM-*E_0UCS*.Lab* model (termed in this way following the same nomenclature previously adopted), despite of the low number of records used during the training and test phases. Figure 5.29 depicts the relationship between the E_0 experimental values and those predicted by the SVM-E_0UCS.Lab model, showing a small deviation between them, which is corroborated by an R^2 value very close to the unit ($R^2 = 0.94$). These results indicate once again the advanced learning capabilities of such an algorithm, namely in *JGLF* data analysis. Among the 20 runs performed, the *SVM* hyperparameters (described in Section 2.3.3) that best fit the data are $\epsilon = 0.07 \pm 0.01$ and $\gamma = 0.05 \pm 0.00$ (mean values and 95% confidence intervals).

The relative importance of each input variable according to *SVM-*E_0UCS*.Lab* model was measured by performing the *GSA* described in Section 2.5.3. Figure 5.30 shows, as expected, that the UCS is strongly correlated with the E_0 of a given sample and that t, C_{iv}, and $\%Clay$ also play important roles in the relationship between these two mechanical properties of *JGLF*.

Moreover, the *GSA* analysis also confirms that these variables have an almost linear effect on E_0 prediction, as illustrated in the *VEC* curves of UCS and t depicted in Figure 5.31. This linear behaviour, indicates that all nonlinear components in E_0 prediction are incorporated through the UCS variable, confirming what was expected.

The proposed approach is compared with the *EC2-*E_0*.Lab* model adapted for *JGLF* presented in Section 5.3 for the purposes of a baseline comparison and practical application. Figure 5.29 shows the deviation between the values predicted by these two models

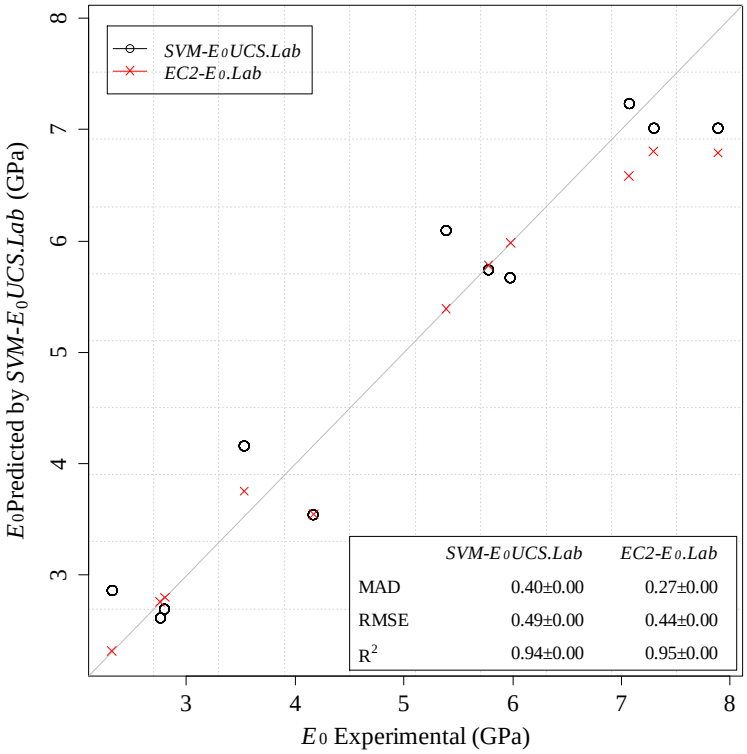


Figure 5.29: Relationship between E_0 experimental versus predicted values by $SVM\text{-}E_0UCS.Lab$ and $EC2\text{-}E_0.Lab$ models

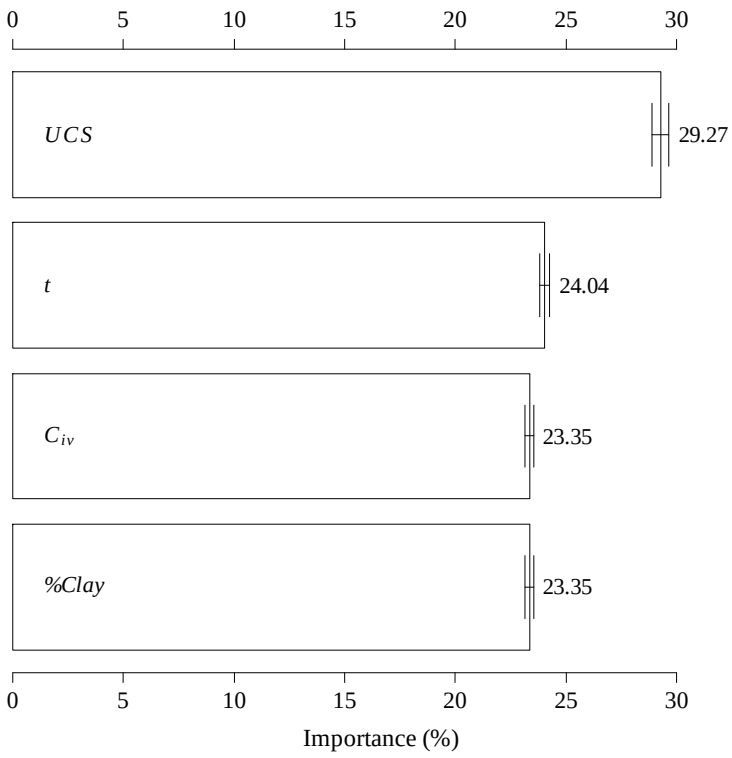


Figure 5.30: Relative influence of each input variable according to the $SVM\text{-}E_0UCS.Lab$ model, quantified by 1-D *SA*

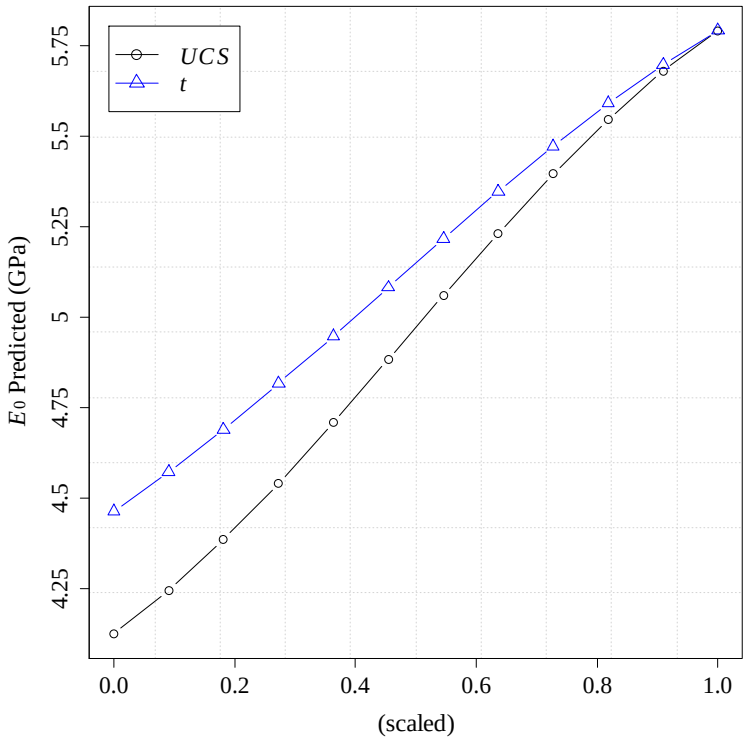


Figure 5.31: *VEC* curves for the *UCS* and t variables during the E_0 prediction of *JGLF* according to the *SVM-E_0UCS.Lab* model, quantified by 1-D *SA*

(*SVM-E_0UCS.Lab* and *EC2-E_0.Lab*) and their metric values. A simple comparison of the results evidences that the *EC2-E_0.Lab* model presents a slightly better performance than the proposed approach (*SVM-E_0UCS.Lab* model). However, it should be stressed that the *EC2-E_0.Lab* model is strongly dependent of the 28 day deformability modulus of each formulation, which requires more complex and expensive laboratory testing than what is required to get the *UCS*. Moreover, five of the eleven samples were tested at 28 days, meaning that the *EC2-E_0.Lab* model error is equal to zero for these records. The proposed approach (the *SVM-E_0UCS.Lab* model) presents a better accuracy for the remaining samples (see Figure 5.29), which were tested at 56 days. Furthermore, it should be stressed that the proposed approach (i.e. *SVM-E_0UCS.Lab* model) can further be updated, as new data becomes available, which will certainly improve the model's accuracy.

5.5 Conclusions

The study of laboratory formulations is an important task that can supply valuable information related to the future behaviour of *soilcrete*, which helps in the definition of

the *JG* technology parameters. This study is particularly important if the information related to the soil to be treated is scarce. In these particular situations, several laboratory formulations need to be prepared and tested to define the best solutions that meet the project requirements at lower environmental and economic costs. For such purposes, it is very useful to have a set of predictive models capable of accurately estimating mechanical properties of such mixtures. In this work, we attempted to develop new reliable models for *JGLF* mechanical properties prediction.

The results presented in this chapter show that both UCS and stiffness of *JGLF* can be accurately predicted by data-driven models over time for seven different ground types. It was shown that the *SVM* algorithm is able to learn the complex relationships between *JGLF* mechanical properties and their contributing factors. Moreover, based on the application of a *GSA*, the proposed models, particularly those obtained from the *SVM* algorithm that are characterised by a high mathematical complexity, were "opened", allowing extraction of useful information. This analysis suggested that t, $n/(C_{iv})^d$ and soil properties, particularly $\%Clay$, are key variables in both strength and stiffness prediction of *JGLF*. In addition, we also measured the effect of such key variables in *JGLF* mechanical behaviour, determining the exponential effect of t as well as the negative impact of %Clay in the development of *JGLF* mechanical properties. Moreover, it was shown that the analytical expressions proposed by *EC2* for strength and stiffness prediction of concrete can be successfully adapted to predict mechanical properties of *JGLF*. However, this approach presents an important limitation related to its application in the early stages of a *JG* project.

Additionally, an attempt to predict E_0 of *JGLF* based on UCS values was performed. Although supported by a rather small number of records, this experiment showed a good performance in E_0 prediction, and it showed an almost linear relationship between stiffness and strength for a given sample of *JGLF*. This approach has an important practical application because it allows the prediction of E_0 for a given sample based on unconfined compression tests that are less expensive.

The knowledge developed herein can potentially contribute a better understanding of the *JGLF* behaviour and improve the technical and economic efficiency of *JG* technology because the number of *JGLF* to prepare can be significantly reduced without a loss of information.

CHAPTER 6

DM techniques applied to field data

6.1 Introduction

In Chapter 5, we proposed several data-driven models for *JGLF* mechanical properties prediction. It was shown that *DM* algorithms, particularly *SVM*, are able to explore *JGLF* and learn its mechanical behaviour. Moreover, through the application of a *GSA*, useful information was extracted, representing a greet contribution for a better understanding of *JGLF* behaviour and model interpretability.

Accordingly, in this chapter the same framework was applied to *JG* data related with real *JG* columns. Therefore, we developed predictive models for UCS and Young's modulus of *soilcrete* material, as well as for *JG* columns diameter, using *ANN* and *SVM* algorithms, and *MR* as a baseline comparison. Moreover, the generic expression written in Equation 5.1 was also optimized to *soilcrete* mechanical properties, using the minimisation problem of Equation 5.2 and the *FN* algorithm. In addition, and considering the good performance of *EC2* approach in *JGLF* mechanical properties study, its analytical expressions were also adapted to strength and stiffness prediction of *soilcrete* mixtures.

Similar to what was executed in the study of *JGLF*, we also applied the two *FS* approaches described in Section 2.4, (i.e. forward and backward *FS* algorithms), aiming to help to define the best set of input variables. Also, during the *FS* task, we applied the *SVM* algorithm and the methodology proposed by Huang et al. (2007) for model selection (i.e. to select the best values of the hyperparameters C, ϵ and γ). Then, during the learning phase of the models, i.e. after choosing the final set of input variables, we used the same parameters for *ANN* and *SVM* algorithms, as described in Section 5.1 (e.g. the same activation function for *ANN* algorithm as well as the same grid search for H and γ hyperparameters). The only difference is related with the approaches applied during the search for the best value of H and γ, as well for generalizations purposes. Thus,

instead of an internal (i.e. applied over training data) 5-fold cross-validation,we applied a 3-fold cross validation, and instead of a leave-one-out scheme we applied a 20-fold cross-validation for model generalization assessment. The reason for these choices is basically related with the higher number of records of the databases used for strength, stiffness and diameter study of real *JG* columns, which requires more computational effort. Finally, we adopted *R* environment and *rminer* library for *MR*, *ANN* and *SVM* predictions, and for *FN* estimation we adopted the GAMS software.

It should be noted that in the study of the mechanical properties and diameter of *JG* columns, the choice of the input variables was somewhat conditioned by the availability of some variables. On one hand, aiming to maximize the number of records, some variables were not considered because there was just few records with such informations. On the other hand, there are some variables that despite of its empirical relevance, were not considered as inputs because either are constant in the compiled database or just are not used by the company that supplied the data (e.g. nozzles orientation).

6.2 Uniaxial compressive strength prediction

6.2.1 Model performance

On Table 6.1 it's compared the performance (using metrics *MAD*, *RMSE* and R^2 as performance criteria) of the *SVM* predictive models developed based on the forward and backward *FS* approaches, with that (termed as MS_{qf1}) where the input variables were manually selected considering the literature review, knowledge from *JGLF* study, as well as on the contribute given by the two *FS* approaches implemented. This table shows that the best performance, considering the metric values and its confidence interval as well as the empirical relevance of the chosen variables, was achieved by *SVM* model using the set of variables assigned as MS_{qf1}. Hence, this set of nine variables will be used during the entire study of *UCS* of *soilcrete* mixtures (Tinoco et al., 2012a). On Table 6.2 are summarized the main statistics of the database used during this study, i.e. the database that contemplates just the nine variables assigned in Table 6.1 as MS_{qf1} and is composed by 472 records.

Table 6.1: Comparison of the *SVM* models performance developed using the forward and backward *FS* approaches with a manual selection of attributes, aiming to predict UCS of *soilcrete* mixtures

Var	FFS	BFS	MS_{qf1}
JS	×	✓	✓
$n/(C_{iv})^d$	✓	✓	✓
t	×	✓	✓
C_{iv}	×	×	✓
$1/\rho_d$	×	×	✓
e	×	×	✓
ω	✓	✓	✓
W/C	×	×	✓
$\%Clay$	✓	✓	✓
kg/ml	✓	×	×
N_{Dgrout}	✓	×	×
$1/n$	×	✓	×
Imp_{grout}	×	✓	×
kg/m^3	×	✓	×
rpm	×	✓	×
ρ	×	✓	×
ρ_d	×	✓	×
n	×	✓	×
W_c/C	×	✓	×
S_r	×	✓	×
MAD	1.37 ± 0.02	1.37 ± 0.03	1.38 ± 0.01
RMSE	1.98 ± 0.03	1.97 ± 0.06	1.99 ± 0.01
R^2	0.52 ± 0.02	0.52 ± 0.03	0.51 ± 0.01

FFS - forward feature selection; BFS - backward feature selection

The mathematical expression proposed by *EC2* (see Equation 3.17), which had shown a good performance in strength prediction of *JGLF*, it was also adapted to *soilcrete* material. After optimize the coefficient a of Equation 3.17 to UCS data of *soilcrete* mixtures, the best value is $a = 0.5$, and the resulting model (further termed as *EC2*-

UCS.Field) is written by the following equation:

$$UCS = e^{\left(s\cdot\left[1-\left(\frac{28}{t}\right)^{0.5}\right]\right)} \cdot UCS_{28days} \tag{6.1}$$

However, *EC2-UCS.Field* model performs badly UCS of *soilcrete* as illustrated in Figure 6.1, achieving an R^2 value of just 0.13. This low performance is probably related with the higher heterogeneity of *soilcrete* material when compared with *JGLF*, as well as with the not consideration of many others variables that are important in *soilcrete* strength prediction over time (e.g. mixture porosity or cement content). This means that, even when knowing the UCS of each formulation at 28 days time of cure, the proposed approach by *EC2* for strength prediction of concrete is unable to accurately predict UCS of *soilcrete* mixtures over time.

Table 6.2: Summary statistics of both input and output variable of the database used during the study of UCS of *soilcrete* mixtures, which contemplates the nine input variables assigned in Table 6.1 as MS_{qf1}

Variable	Minimum	Maximum	Mean	Standard Deviation
JS	1.00	3.00	2.03	0.38
W/C	0.83	1.05	0.93	0.07
ω	2.50	96.80	38.89	12.12
$\%Clay$	22.50	45.00	30.84	6.87
t	9.00	181.00	46.12	32.80
$1/\rho_d$	$5.63E^{-4}$	$1.44E^{-3}$	$8.43E^{-4}$	$1.23E^{-4}$
C_{iv}	0.14	0.28	0.22	0.03
e	0.56	2.99	1.32	0.34
$n/(C_{iv})^d$	37.88	79.17	59.49	6.88
UCS	0.32	20.27	4.05	2.83

The coefficients of Equation 5.1, optimized to UCS data of *soilcrete* mixtures, using the *FN* algorithm and the minimization problem described in Equation 5.2 are present in Equation 6.2 (this model will be termed as *FN-UCS.Field*).

$$\begin{aligned} UCS = 1.000E^{+10} \cdot JS^{-0.730} \cdot W/C^{1.142} \cdot \omega^{0.328} \cdot \%Clay^{-0.550} \cdot \\ \cdot t^{0.133} \cdot 1/\rho_d^{2.892} \cdot C_{iv}^{0.891} \cdot e^{-3.193} \cdot (n/(C_{iv})^d)^{0.422} \end{aligned} \tag{6.2}$$

The *FN-UCS.Field* model, trained/assessed under the Leave-One-Out estimation approach, presents a slightly better performance in UCS prediction of *soilcrete* mixtures

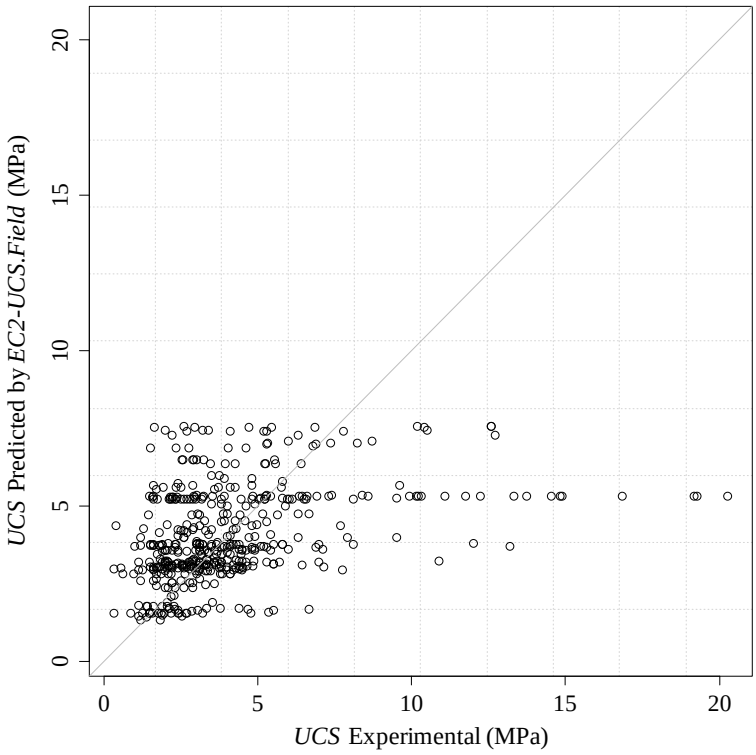


Figure 6.1: Relationship between UCS experimental versus predicted values by *EC2-UCS.Field* model

when compared to *EC2-UCS.Field* model, as shown in Table 6.4. However, such performance is still very poor, as illustrated in Figure 6.2 that depicts the relationship between UCS experimental and predicted by *FN-UCS.Field* model. This figure shows a significant dispersion, particularly when compared which *FN-UCS.Lab* model used for strength prediction of *JGLF*, but not unrealistic for field data analysis.

The averaged hyperparameters and fitting time values (and respective 95% level confidence intervals according to a t-student distribution) of all *DM* models trained using the set of nine input variables assigned in Table 6.1 as MS_{qf1} are summarized in Table 6.3. These models, developed to predict UCS of *soilcrete* will be further termed as *MR-UCS.Field*, *ANN-UCS.Field* and *SVM-UCS.Field*, and are respectively the result of the training of *MR*, *ANN* and *SVM* algorithms with UCS data of *soilcrete* mixtures.

Table 6.4 shows and compares the predictive capacity of all trained models for UCS prediction of *soilcrete* based on *MAD*, *RMSE* and R^2 metrics, computed for the test data under a 20-fold cross-validation approach (mean value and 95% confidence intervals). This table shows a considerable decrease in predictive performance when compared to *JGLF* study (see Table 5.3). However, keeping in mind that here we analyse *JG* field data, with all its with all its complexity and heterogeneity, an R^2 value around 0.5 could be considered satisfactory. The *MR-UCS.Field* model, used mainly for a baseline comparison achieved

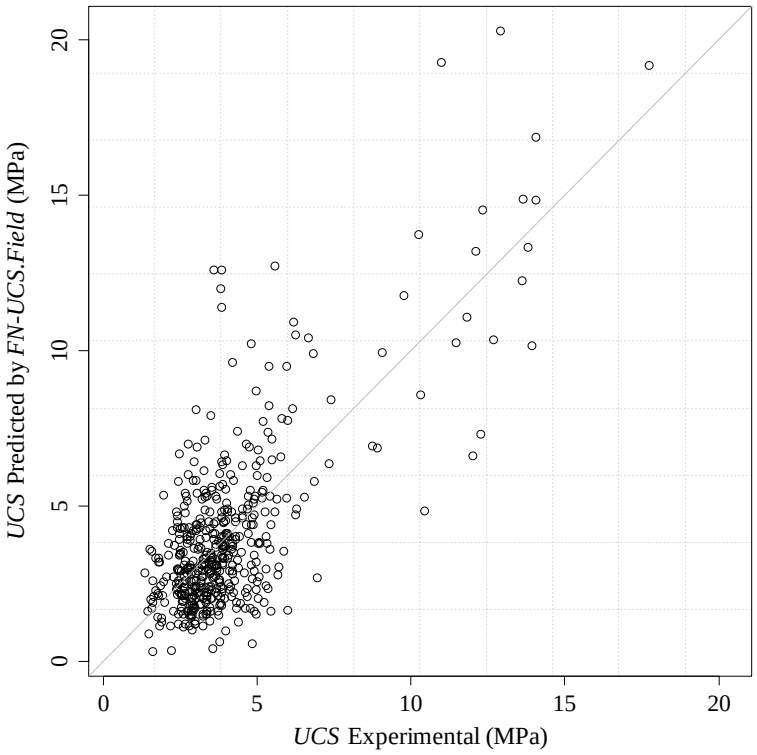


Figure 6.2: Relationship between UCS experimental versus predicted values by *FN-UCS.Field* model

the lowest performance. This observation can be interpreted as an indication that the UCS of *soilcrete* mixtures are not guided by linear laws.

Scatterplots of *ANN-UCS.Field* and *SVM-UCS.Field* models, illustrated in Figures 6.3 and 6.4 respectively, confirming the non ideal performance shown in Table 6.4. As depicted, in both models the predictions are not very close to the diagonal line that represent the experimental values. However, approximately 81% of the records are predicted with an absolute error lower than 2 MPa and just around 7% are unsafe predictions, which represents an acceptable performance for field data predictions.

Table 6.3: Hyperparameters and computation time of each DM model for UCS prediction of *soilcrete* material

Model	Hyperparameters	time (s)
FN-UCS.Field	-	121.20 ± 0.00
MR-UCS.Field	-	0.64 ± 0.01
ANN-UCS.Field	$H = 3 \pm 1$	46.21 ± 0.35
SVM-UCS.Field	$\gamma = 0.07 \pm 0.01$, $\epsilon = 0.10 \pm 0.00$	40.94 ± 0.13

Table 6.4: Error metrics of all *DM* models for *UCS* prediction of *soilcrete* (test set values, best values in **bold**)

Model	MAD	RMSE	R^2
EC2-UCS.Field	1.75 ± 0.00	2.65 ± 0.00	0.13 ± 0.00
FN-UCS.Field	1.40 ± 0.00	$\mathbf{1.95} \pm 0.00$	0.19 ± 0.00
MR-UCS.Field	1.53 ± 0.00	2.13 ± 0.01	0.43 ± 0.00
ANN-UCS.Field	1.41 ± 0.02	2.01 ± 0.06	0.49 ± 0.03
SVM-UCS.Field	$\mathbf{1.38} \pm 0.01$	$\mathbf{1.99} \pm 0.01$	$\mathbf{0.51} \pm 0.01$

Figure 6.5 compares the performance of all models trained for *UCS* prediction of *soilcrete* mixtures (i.e. *EC2-UCS.Field*, *FN-UCS.Field*, *MR-UCS.Field*, *ANN-UCS.Field* and *SVM-UCS.Field* models), depicting the model accuracy as a function of the absolute deviation (*REC* curves, (Bi and Bennett, 2003)). The shape of these curves evidence once more the non ideal performance of the developed models, and that *SVM-UCS.Field* and *ANN-UCS.Field* are the two more accurate models in *UCS* of soilcrete mixtures. Reading the *REC* curve of *SVM-UCS.Field* model (the most accurate), it is shown that if an absolute deviation around 2 MPa is tolerated, then around 81% of the records can be accurately predicted by the model, as above underlined.

Making a global appreciation of *EC2-UCS.Field*, *FN-UCS.Field*, *MR-UCS.Field*, *ANN-UCS.Field* and *SVM-UCS.Field* models, it can be concluded that even the last two models have difficulties to learn the complex relationships between *UCS* of *soilcrete* mixtures and its contributing factors. However, the achieved performance by *ANN-UCS.Field* and *SVM-UCS.Field* models, with an R^2 close to 0.5, can be considered acceptable within field data analysis.

6.2.2 Model interpretability

The obtained data-driven models, namely *SVM-UCS.Field* and *ANN-UCS.Field* models, perform *UCS* prediction of *soilcrete* mixtures with a considerable but acceptable dispersion. In this section, we apply a *GSA* over such models to extract useful information, helping to understand better the *UCS* behaviour of *soilcrete* mixtures. Accordingly, and based on a 1-D *SA*, Figure 6.6 shows and compares the relative importance of each input variable according to *MR-UCS.Field*, *ANN-UCS.Field*, *SVM-UCS.Field* and *FN-UCS.Field* models.

Interpreting Figure 6.6, one can observe that according to *SVM-UCS.Field* model (the most accurate in *UCS* prediction), the relation $n/(C_{iv})^d$, JS, t and $\%Clay$ are the

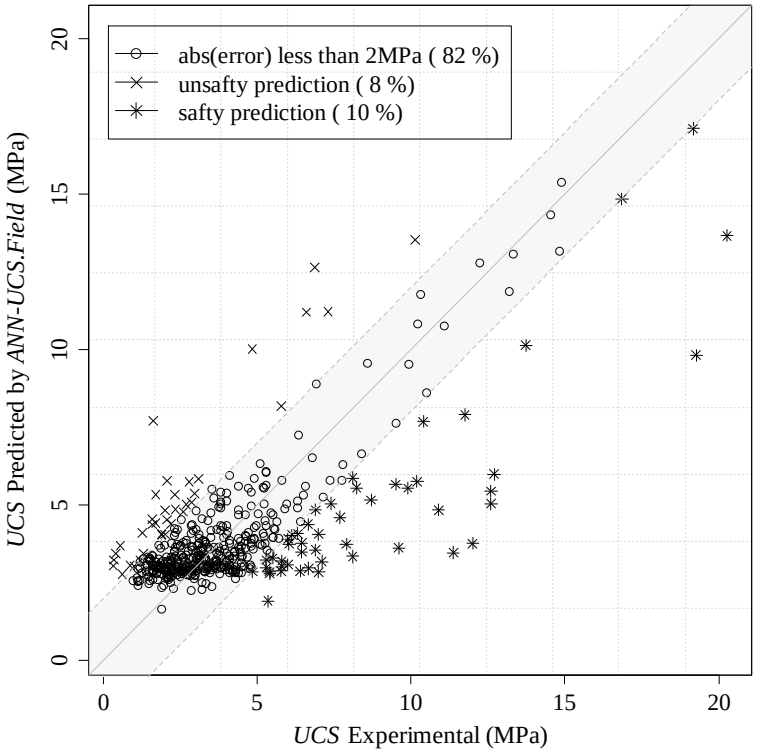


Figure 6.3: Relationship between UCS experimental versus predicted values by $ANN\text{-}UCS.Field$ model

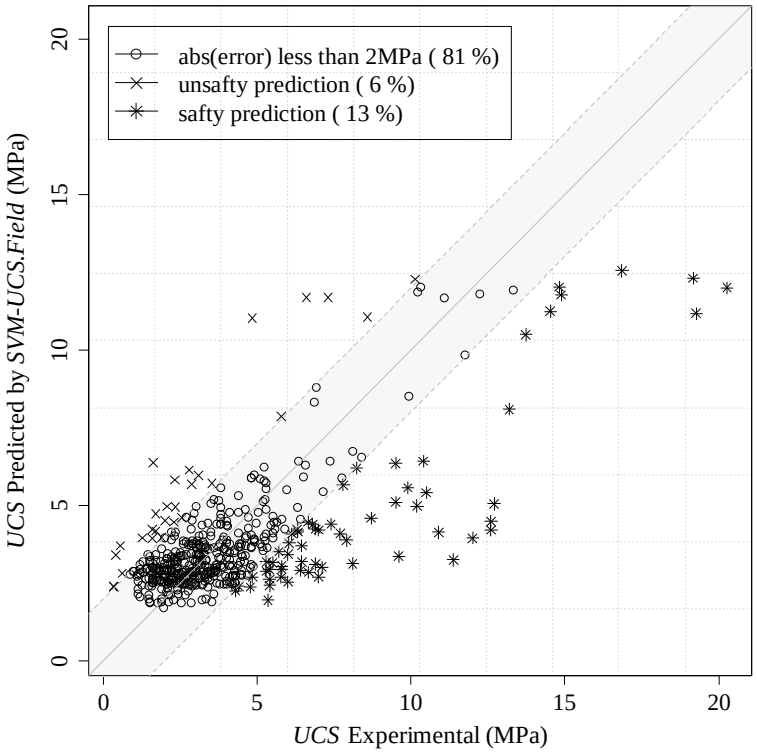


Figure 6.4: Relationship between UCS experimental versus predicted values by $SVM\text{-}UCS.Field$ model

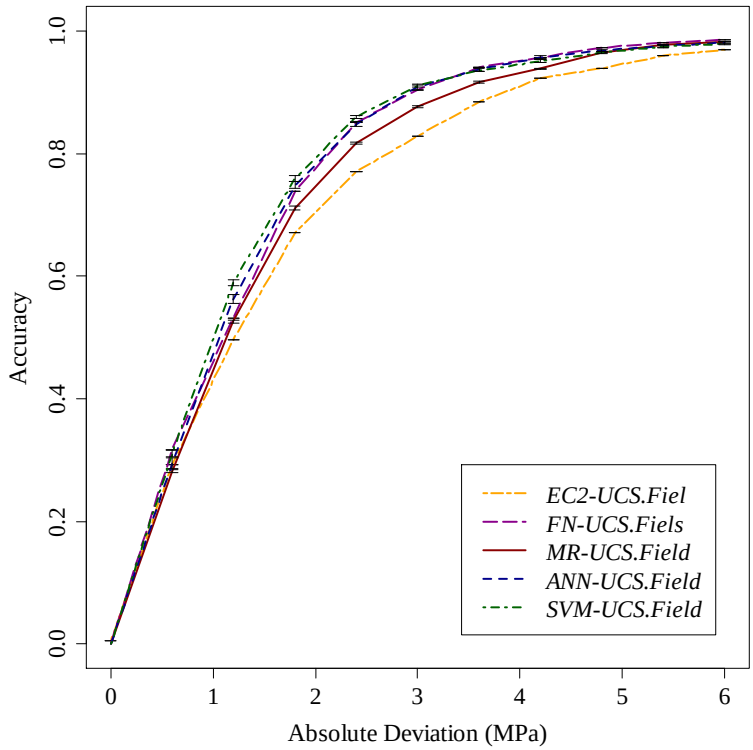


Figure 6.5: *REC* curves of *EC2-UCS.Field*, *FN-UCS.Field*, *MR-UCS.Field*, *ANN-UCS.Field* and *SVM-UCS.Field* models, comparing its performance in UCS prediction of *soilcrete* mixtures

four key variables in UCS prediction of *soilcrete* mixtures (Tinoco et al., 2011e). Among them it is identified one that is related to the soil type (%Clay), another related with the JG process (JS) and two others related to the JG mixture, namely its age and the relation $n/(C_{iv})^d$ that combines the porosity and cement content effect. In other words, to predict UCS of *soilcrete* mixtures, the models ask for information about the soil to be improved, how the improvement was performed and the actual conditions of the obtained mixture. Moreover, it is also interesting to observe that such variable ranking has a physical explanation and is empirically understandable. Experimental studies related with soil-cement mixtures have been shown that both soil properties and age of the mixture should be taken into account in its behaviour (Van Impe et al., 2005; Liu et al., 2008). Furthermore, concerning to soil improvement using JG technology, it makes sense that the JG system used should be considered, since it will determine the energy applied or the impact of the fluids against the soil. On the other hand, *FN-UCS.Field* model evidences an unrealistic behaviour by considering e (66%) and $1/\rho$ (19%) the two key variables in UCS prediction of *soilcrete* mixtures. Hence, based on these observations, and although of the dry density ($1/\rho_d$) of the *soilcrete* mixture as well as its void ratio (the key variables according to *ANN-UCS.Field* model), show a lower relevance in *SVM-UCS.Field* model (but not dismissed), *SVM-UCS.Field* model seems to be the most interesting one.

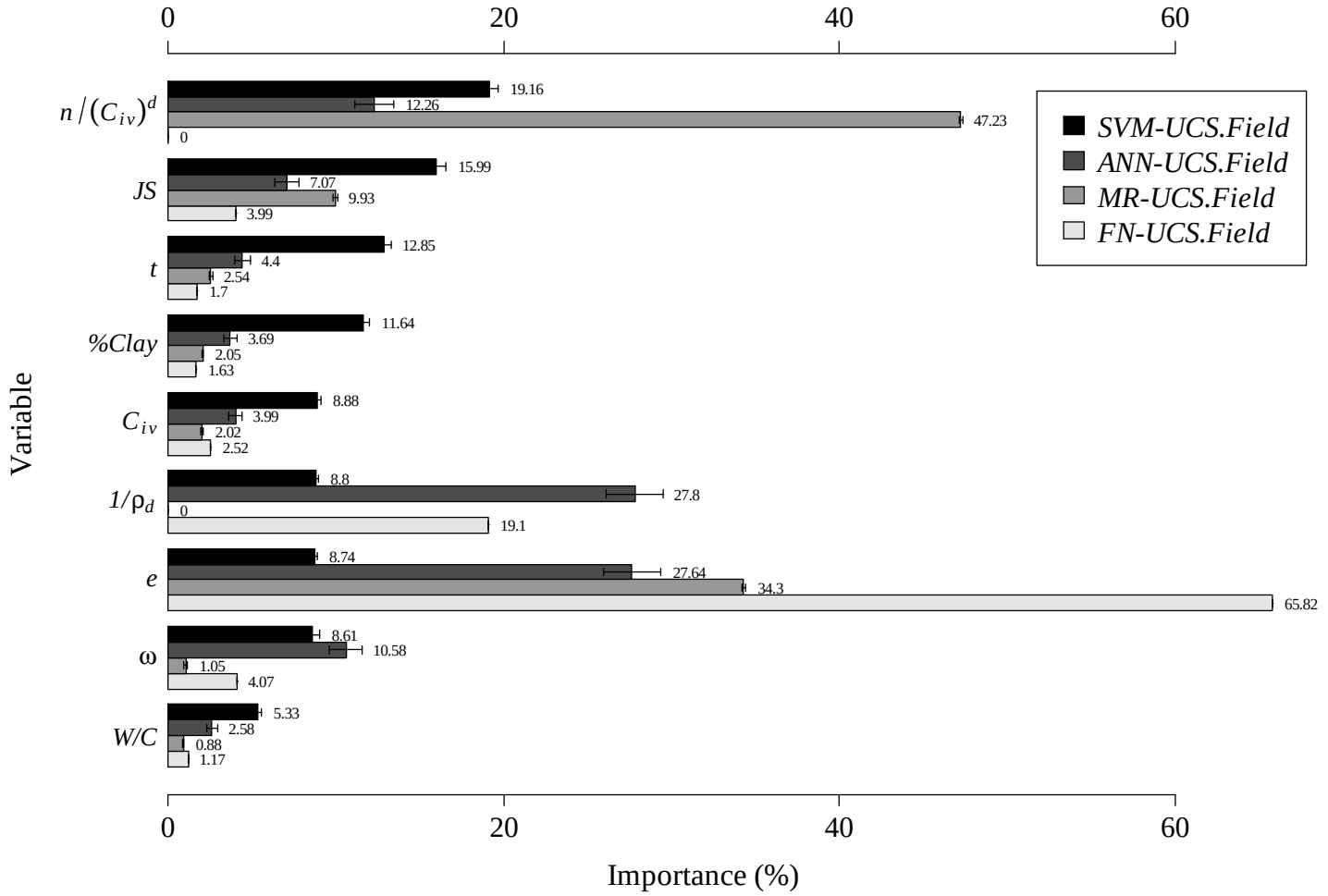


Figure 6.6: Relative importance of each input variable quantified by 1-D *SA*, comparing *MR-UCS.Field*, *ANN-UCS.Field*, *SVM-UCS.Field* and *FN-UCS.Field* models

Thus, in the next paragraphs, *SVM-UCS.Field* model will be used as the reference model in the performed analysis. Moreover, this algorithm was already used as reference in mechanical properties study of *JGLF*.

Toward to a better understanding what has been learned by *SVM-UCS.Field* model about UCS behaviour of *soilcrete* mixtures, the *VEC* curves of its four key input variables are plotted in Figure 6.7. Analysing such curves, it can be observed that the influences of the four key variables are in agreement with the empirical knowledge, showing a predominant nonlinear effect in UCS behaviour of *soilcrete* mixtures (Tinoco et al., 2011d). On one hand, UCS increases with the age of the mixture according to an exponential law (Van Impe et al., 2005; Coulter and Martin, 2006). This convex shape evidences that the first days of cure are responsible by the main gain of strength of the mixture. On the other hand, the relation $n/(C_{iv})^d$ and the $\%Clay$ have a similar and negative impact in UCS prediction of *soilcrete*. This means that when increasing the mixture porosity or clay content, or decreasing the cement content, the UCS of the mixture will decrease. In addition, the highest values of UCS are achieved for mixtures produced with single fluid system, decreasing almost linearly for double and triple fluid system. This outcome makes sense if we take into account that when increasing the energy of the jet (from single to triple fluid system), the achieved distance is higher. As a result, the content of cement by unit volume of soil is lower, leading to a decrease in UCS of the mixture.

Aiming a better understanding of *soilcrete* strength behaviour when two variables are changed simultaneously, a 2-D *SA* was applied over *SVM-UCS.Field*. Accordingly, it was measured the interaction level between all variables with t and $\%Clay$. Table 6.5 summarizes the relative importance of each variable in these interactions. In both situations, it is observed that W/C ratio presents a high interaction level with t and $\%Clay$ despite of its low relative importance in strength prediction of *soilcrete*, as shown in Figure 6.6. This observation is by itself the reason of such small influence in UCS behaviour of *soilcrete*, since t and $\%Clay$ are within the four more relevant variables.

Table 6.5: Interaction level (%) between all variables with t and $\%Clay$, according to *SVM-UCS.Field* model for UCS prediction of *soilcrete* mixtures, measured by a 2-D *SA*

Variables	JS	W/C	ω	$\%Clay$	t	$1/\rho_d$	C_{iv}	e	$n/(C_{iv})^d$
t	13.18	**13.77**	11.36	13.13	-	11.45	12.92	11.40	12.78
$\%Clay$	12.47	**13.82**	13.04	-	15.07	10.06	12.66	10.00	12.89

Plotting the effect in UCS of *soilcrete* mixtures when t and JS are changed simul-

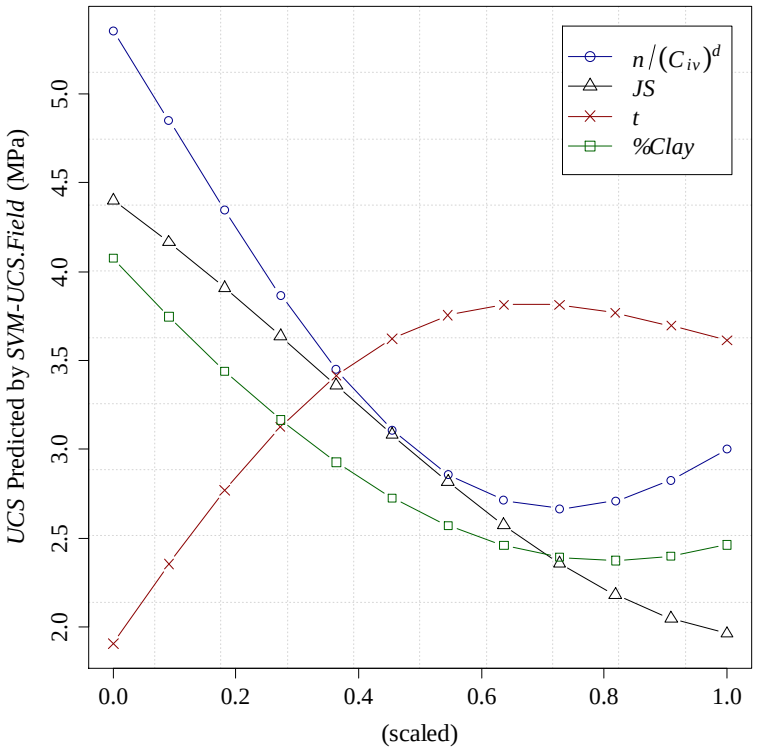
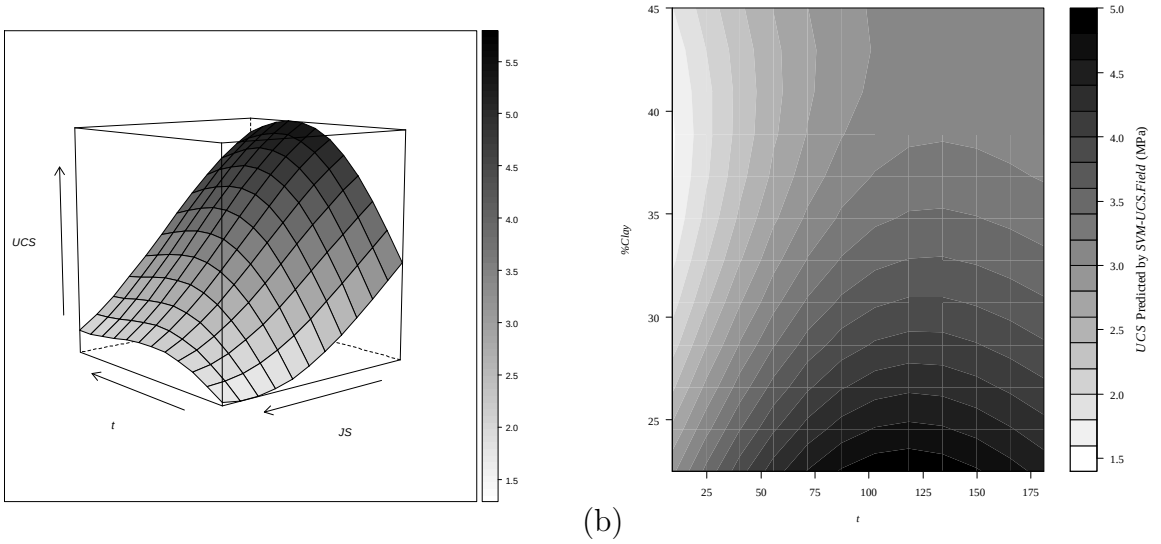


Figure 6.7: *VEC* curves of the four key input variables according to *SVM-UCS.Field* model in *UCS* prediction of *soilcrete*, quantified by 1-D *SA*

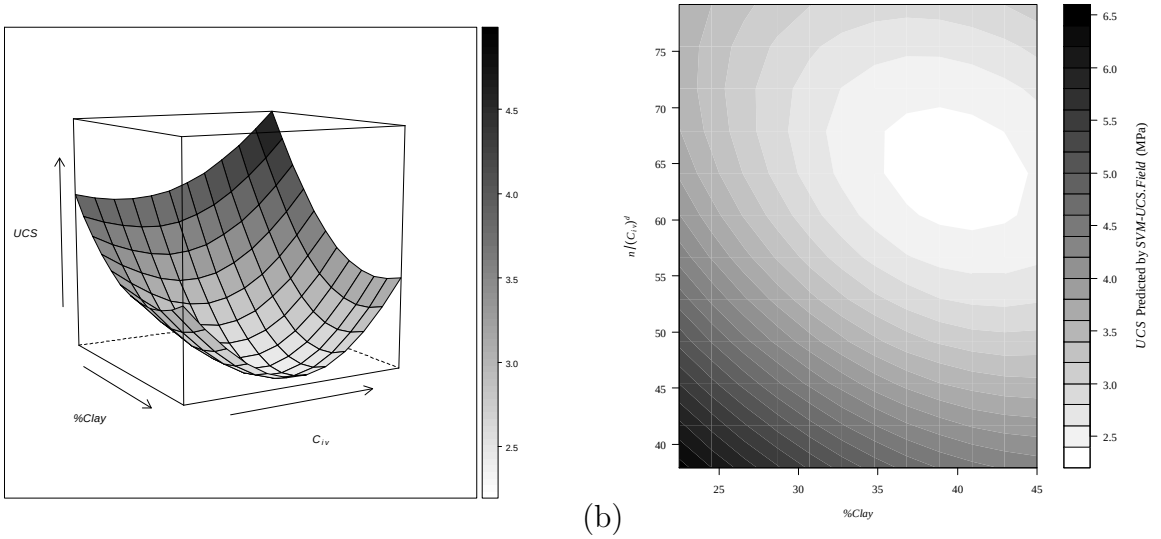
taneously, the *VEC* surface of Figure 6.8a is obtained. The *VEC* contour depicted in Figure 6.8b represents the effect in *UCS* of *soilcrete* mixtures for different combination between t and $\%Clay$. In the first situation, it is observed that the t effect in *UCS* is more pronounced when the soil improvement is performed with single fluid system and that this gain ratio, as well as the maximum strength values, decreases for double and triple fluid systems (Tinoco et al., 2011d). Indeed, for triple system, *UCS* of *soilcrete* mixtures just slightly increase over time. Moreover, it is also observed that the influence of the *JG* system, will be particularly noteworthy for advanced ages. In the second case, a similar behaviour is observed, i.e. that the gain of strength is more pronounced in soil with low clay content, noting also that for high $\%Clay$ the *UCS* of *soilcrete* mixtures just slightly increases over time.

Figure 6.9a shows that for soils with high clay content, the effect of C_{iv} variations on *UCS* of *soilcrete* mixtures is hardly noticeable. Additionally, it is also observed that for soils with low $\%Clay$, even for low C_{iv}, it is obtained a considerable strength. On Figure 6.9b it is observed a uniform variation of *UCS* of *soilcrete* when $\%Clay$ and $n/(C_{iv})^d$ are changed simultaneously, being the highest values of *UCS* achieving for *soilcrete* mixtures with a low $n/(C_{iv})^d$ ration and prepared in a soil with low clay content.



(a) (b)

Figure 6.8: 2-D *SA* according to *SVM-UCS.Field* model in UCS prediction of *soilcrete*: a) VEC surface for t and JS interaction and b) VEC contour for t and $\%Clay$ interaction



(a) (b)

Figure 6.9: 2-D *SA* according to *SVM-UCS.Field* model in UCS prediction of *soilcrete*: a) VEC surface for $\%Clay$ and C_{iv} interaction and b) VEC contour for $\%Clay$ and $n/(C_{iv})^d$ interaction

6.3 Comparison between laboratory and field strength predictions

As shown in Chapter 5, a new approach was developed to accurately predict UCS of $JGLF$, namely by *SVM-UCS.Lab* model. On the other hand, the same novel approach was applied for field samples from JG columns of geotechnical works. Moreover, in Section 6.5, it will be proposed a relationship between strength and stiffness of *soilcrete* mixtures. In order to set an integrated approach for JG mechanical properties design, a final step needs to be executed, allowing to make the bridge between laboratory formulations and field samples. Accordingly, a plausible, although tentative framework as a step toward the prediction of UCS of *soilcrete* from the laboratory database was attempted. In Figure 6.10, the mean values at 28 days of UCS experimental field samples by geotechnical works are compared with those predicted by *SVM-UCS.Lab* laboratory model (also by mean values at 28 days of geotechnical works). The analysis of these results show an acceptable relationship between UCS of laboratory formulations and field samples except for one geotechnical work (G). Indeed, if this case is excluded, a relationship between laboratory formulations and field samples with an $R^2 = 0.64$ is achieved. Moreover, it is observed that UCS of *soilcrete* is around 11% higher than the equivalent laboratory formulation, following reference values found in the literature (Van Impe et al., 2005).

For geotechnical work G, the mean value predicted by *SVM-UCS.Lab* model was considerably overestimated. Since the *SVM-UCS.Lab* model applicability is satisfied for all field records used in this experiment, it was performed an attempt to find a plausible justification for such situation.

Considering that the study of *soilcrete* mixtures is based on laboratory formulations, we agree that such deviation should be related with a given variable not contemplated by the laboratory model. However, since this deviation is observed just for one geotechnical work, this behaviour is probably related with a particular situation of this geotechnical work. Therefore, it was performed a deep analysis of all available information related with each of the geotechnical works, such as the amount of cement applied during the soil improvement, the water content of the mixture, the depth where the samples were collected (influence of environment effects), water table level, etc. Some experiments were also performed, using UCS of each sample normalized by the 28 days strength of the respective formulation. However, no significant differences were observed. The only relevant difference is related with the water table level of geotechnical work G. In this case, there are information that the columns were built bellow water table level, leading us to conclude that this is probably the reason for the low values of UCS observed for

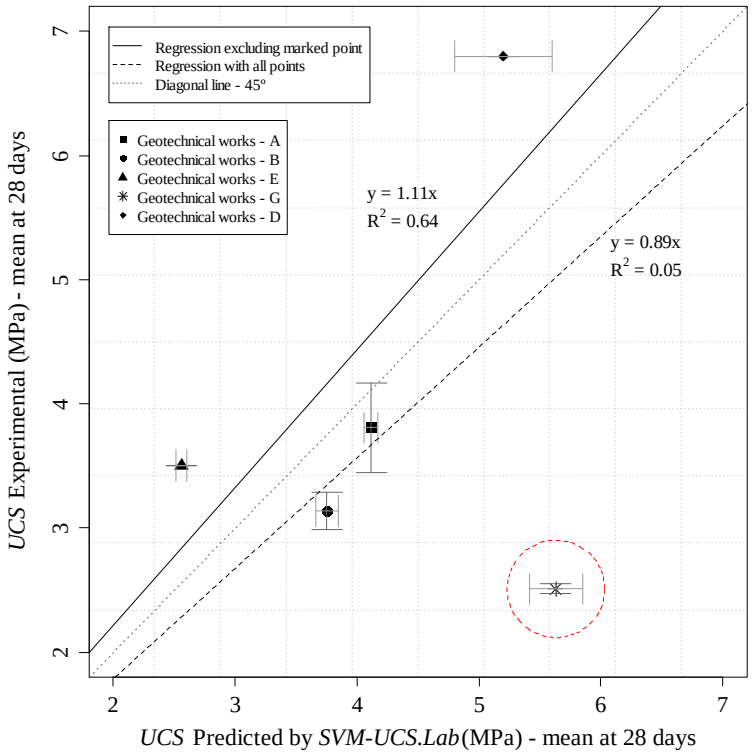


Figure 6.10: Deviation of the mean by geotechnical works of UCS experimental field samples against predicted by *SVM-UCS.Lab* model (mean values by geotechnical works)

field samples when compared to what was expected based on laboratory formulations behaviour. However, this issue requires a more detailed analysis in order to establish stronger evidence-based conclusions.

6.4 Deformability modulus prediction

6.4.1 Model performance

The prediction of *soilcrete* stiffness is often of high importance, namely in the evaluation of structure's serviceability. Therefore, and following the same framework of Section 6.2, some analytical models are herein proposed for E_0 prediction of *soilcrete* mixtures.

The forward and backward *FS* approaches were also applied to guide the selection of the best set of input variables. At the end, the selection for *soilcrete* stiffness study was based on the information given by the *FS* approaches and empirical knowledge, but also supported on the *soilcrete* strength study described in Section 6.2. Moreover, the experience obtained with the study of *JGLF* also gave an important contribution.

Using the metrics *MAD*, *RMSE* and R^2, Table 6.6 compares the performance of the *SVM* predictive models developed based on the forward and backward *FS* approaches,

with that (termed as MS_{qf1}) where the input variables were manually selected considering the literature review, knowledge from *JGLF* and *soilcrete* strength studies, as well as on the contribute of the two *FS* approaches implemented.

Based on the metric values and on the empirical importance of the input variables of each model, the set eight attributes assigned in Table 6.6 as MS_{Ef1} will be used during the entire study of E_0 of *soilcrete* mixtures, compiled in a database with 261 records. Table 6.7 summarizes the main statistics of E_0 and the eight input variables assigned in Table 6.6 as MS_{Ef1} that will be used during the study of *soilcrete* stiffness.

Table 6.6: Comparison of the *SVM* models performance developed using the forward and backward *FS* approaches and that where the attributes were manually selected, aiming to predict E_0 of *soilcrete* mixtures

Var	FFS	BFS	MS_{Efl}
JS	×	×	✓
$n/(C_{iv})^d$	✓	×	✓
t	✓	✓	✓
C_{iv}	×	×	✓
$1/\rho_d$	×	✓	✓
e	×	✓	✓
ω	×	✓	✓
W/C	×	×	✓
$\%Sand$	×	✓	×
$\%Silt$	×	✓	×
$\%Clay$	×	✓	×
$\%OM$	×	✓	×
P_{grout}	✓	×	×
$1/n$	×	✓	×
kg/m^3	✓	×	×
rpm	×	✓	×
ρ	×	✓	×
ρ_d	×	✓	×
W_c/C	✓	×	×
OM/C	×	✓	×
$OM/C^{W_c/C}$	×	✓	×
MAD	0.31 ± 0.01	0.32 ± 0.01	0.31 ± 0.00
RMSE	0.46 ± 0.01	0.47 ± 0.02	0.46 ± 0.01
R^2	0.54 ± 0.03	0.49 ± 0.05	0.53 ± 0.01

FFS - forward feature selection; BFS - backward feature selection

Table 6.7: Summary statistics of both input and output variable of the database used during the study of E_0 of *soilcrete* mixtures, which contemplates the eight input variables assigned in Table 6.1 as MS_{Ef1}

Variable	Minimum	Maximum	Mean	Standard Deviation
JS	1.00	3.00	2.04	0.51
W/C	0.83	1.00	0.89	0.07
ω	2.50	96.80	36.38	13.34
t	9.00	181.00	36.72	35.07
$1/\rho_d$	5.63E^{-4}	1.40E^{-3}	8.18E^{-4}	1.23E^{-4}
C_{iv}	0.14	0.28	0.21	0.04
e	0.56	2.85	1.25	0.33
$n/(C_{iv})^d$	37.88	78.61	58.00	7.50
E_0	0.06	3.63	0.89	0.68

Similar to what was performed in the study of *JGLF* and in strength prediction of *soilcrete* mixture, also for stiffness prediction of *soilcrete* mixture, the proposed expression by *EC2* for concrete deformability prediction was adapted to *soilcrete* mixtures. Relating to the approach proposed by *MC90* (CEB-FIP, 1991) for the same purpose, and taking into account its poor performance in *JGLF* study, it was not applied here. The model obtained from the optimization of coefficients a and b of Equation 3.18 to *soilcrete* stiffness data is written in Equation 6.3 (further termed as *EC2-E_0.Field*).

$$E(t) = \left(e^{\left(s\cdot\left[1-\left(\frac{28}{t}\right)^{0.5}\right]\right)}\right)^{0.3} \cdot E_{cm} \tag{6.3}$$

Again, the *EC2* analytical expression adapted to *soilcrete* mixtures (*EC2-E_0.Field* model) is unable to accurately predicts E_0 of *JG* mixtures. The weak performance achieved by *EC2-E_0.Field* is plotted in Figure 6.11 and corroborated by the low R^2 value achieved ($R^2 = 0.24$). These results illustrate the complexity of *soilcrete* stiffness prediction, even knowing E_0 of each formulation at 28 days time of cure.

The coefficients of Equation 5.1, optimized to *soilcrete* data for stiffness prediction, using the *FN* algorithm and the minimization problem according to Equation 5.2 are shown in Equation 6.4 (this model will be termed as *FN-E_0.Field*).

$$\begin{aligned} E_0 = 1.000\text{E}^{+10} \cdot JS^{0.238} \cdot W/C^{1.904} \cdot \omega^{-0.100} \cdot t^{0.625} \cdot \\ \cdot\, 1/\rho_d^{-18.517} \cdot C_{iv}^{2.194} \cdot e^{20.226} \cdot (n/(C_{iv})^d)^{-22.516} \end{aligned} \tag{6.4}$$

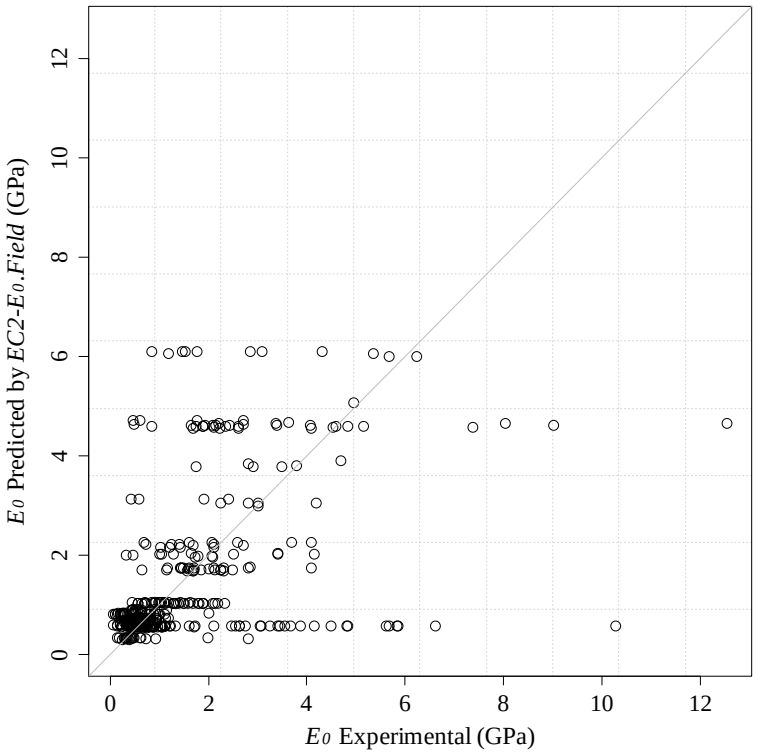


Figure 6.11: Relationship between E_0 experimental versus predicted values by *EC2-E_0.Field* model

FN-E_0.Field model written in Equation 6.4 and trained using the Leave-One-Out estimation method, performs E_0 prediction of *soilcrete* mixtures with some dispersion, as depicted in Figure 6.12, but considerably better than *EC2-E_0.Field* model, obtaining a $R^2 = 0.55$. However, keeping in mind that this model was trained with *JG* field data, such performance may be acceptable.

The average hyperparameters and fitting time values (and respective 95% level confidence intervals according to a t-student distribution) of all *DM* models trained using the set of eight input variables assigned in Table 6.6 as MS_{Ef1} are shown in Table 6.8. These models, developed to predict E_0 of *soilcrete* will be further termed as *MR-E_0.Field*, *ANN-E_0.Field* and *SVM-E_0.Field*, and are respectively the result of the training of *MR*, *ANN* and *SVM* algorithms with E_0 data of real *JG* columns.

Table 6.9 shows the predictive capacity of all trained models (i.e. *EC2-E_0.Field*, *FN-E_0.Field*, *MR-E_0.Field*, *ANN-E_0.Field* and *SVM-E_0.Field*), comparing its performance in E_0 prediction of *soilcrete* mixtures using *MAD*, *RMSE* and R^2 metrics as performance criteria (mean value and 95% confidence intervals), which were computed for the test data under a 20-fold cross-validation approach. Once again, the best performance was achieved by *ANN-E_0.Field* and *SVM-E_0.Field* models, together with *FN-E_0.Field* model. However, the last one is unrealistic in terms of the relative importance of the attributes

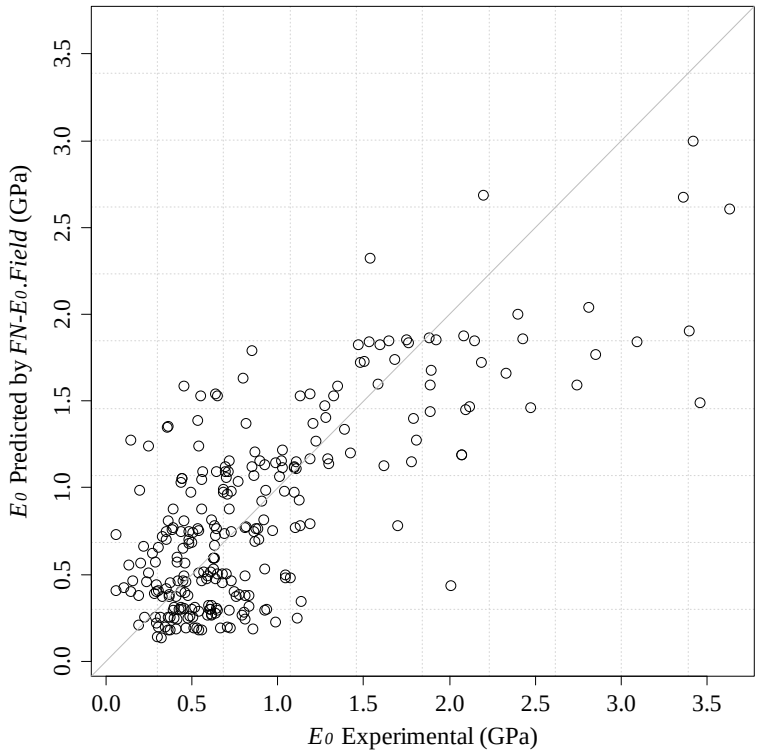


Figure 6.12: Relationship between E_0 experimental versus predicted values by *FN-E_0.Field* model

as further discussed. Comparing *ANN-E_0.Field* and *SVM-E_0.Field* models, the latter seems to be a better choice for E_0 prediction of *soilcrete* mixtures because it gets smaller confidence intervals along the 20 runs performed (Tinoco et al., 2012c). In addition, as further explained when model interpretability is discussed, this model is more coherent, namely in terms of what is empirically known.

Analysing Table 6.9, it is also observed a significant decrease of models accuracy, namely *ANN-E_0.Field* and *SVM-E_0.Field* models, when compared to the proposed models for stiffness prediction of *JGLF* (see Table 5.6). However, and as above underlined, since these models were trained using *JG* field data that normally are characterized by high complexity and heterogeneity, an value of 0.5 for R^2 could be considered satisfactory and acceptable. On the other hand, among the four *DM* algorithms trained, the lowest performance was achieved by *MR-UCS.Field* model, similarly to what occurred in the *UCS* study of *soilcrete* mixtures. This means that also *soilcrete* stiffness behaviour cannot be set by linear laws.

Scatterplots of *ANN-E_0.Field* and *SVM-E_0.Field* models are shown in Figures 6.13 and 6.14 respectively, corroborating that non ideal performance shown in Table 6.9. As observed, there are several predictions that are far from the diagonal line (i.e. higher predictive errors). However, these two Scatterplots also illustrate that the model predictions

Table 6.8: Hyperparameters and computation time of each *DM* model for E_0 prediction of *soilcrete* material

Model	Hyperparameters	time (s)
FN-E_0.Field	-	117.09 ± 0.00
MR-E_0.Field	-	1.05 ± 0.01
ANN-E_0.Field	$H = 3 \pm 1$	59.30 ± 0.40
SVM-E_0.Field	$\gamma = 0.19 \pm 0.02$, $\epsilon = 0.17 \pm 0.00$	39.68 ± 0.05

Table 6.9: Error metrics of all *DM* models for E_0 prediction of *soilcrete* (test set values, best values in **bold**)

Model	MAD	RMSE	R^2
EC2-E_0.Field	0.84 ± 0.00	1.52 ± 0.00	0.24 ± 0.00
FN-E_0.Field	0.34 ± 0.00	$\mathbf{0.45} \pm 0.00$	$\mathbf{0.55} \pm 0.00$
MR-E_0.Field	0.39 ± 0.00	0.55 ± 0.00	0.33 ± 0.01
ANN-E_0.Field	$\mathbf{0.31} \pm 0.01$	$\mathbf{0.46} \pm 0.01$	$\mathbf{0.54} \pm 0.02$
SVM-E_0.Field	$\mathbf{0.31} \pm 0.00$	$\mathbf{0.46} \pm 0.01$	$\mathbf{0.53} \pm 0.01$

tend to follow the diagonal line. Indeed, both models are able to predict approximately 85% of the records within an absolute error less than 0.5 GPa. Moreover, within the prediction with an absolute error higher than 0.5 GPa, 60% (around 9% of all predictions) are conservatives, i.e. the prediction is performed below the experimental value. These two observations give a considerable reliability to the model in spite of the R^2 value around 0.53.

Figure 6.15 compares the predictive performance of all models trained for E_0 prediction of *soilcrete* mixtures (i.e. *EC2-E_0.Field*, *FN-E_0.Field*, *MR-E_0.Field*, *ANN-E_0.Field* and *SVM-E_0.Field* models), depicting the model accuracy as a function of the absolute deviation (*REC* curves, (Bi and Bennett, 2003)). These curves confirm the poor performance of *EC2-E_0.Field* even for higher absolute deviations. Furthermore, it is shown that *ANN-E_0.Field* and *SVM-E_0.Field* models have the highest performance, which is very similar. It is still appealing to observe that *FN-E_0.Field* model performs better E_0 prediction of *soilcrete* mixtures for an absolute deviation higher than 0.9 GPa. Reading the *REC* curve of *SVM-E_0.Field* or *ANN-E_0.Field* models, it is concluded that these models are able to predict accurately more than 80% of the records within an absolute deviation less than 0.5 GPa, as above underlined.

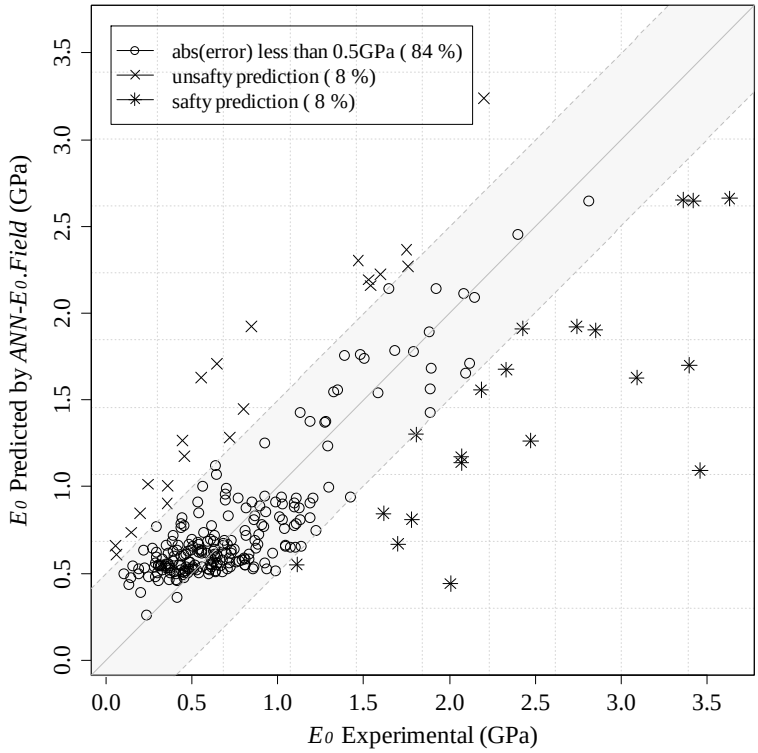


Figure 6.13: Relationship between E_0 experimental versus predicted values by *ANN-*E_0*.Field* model

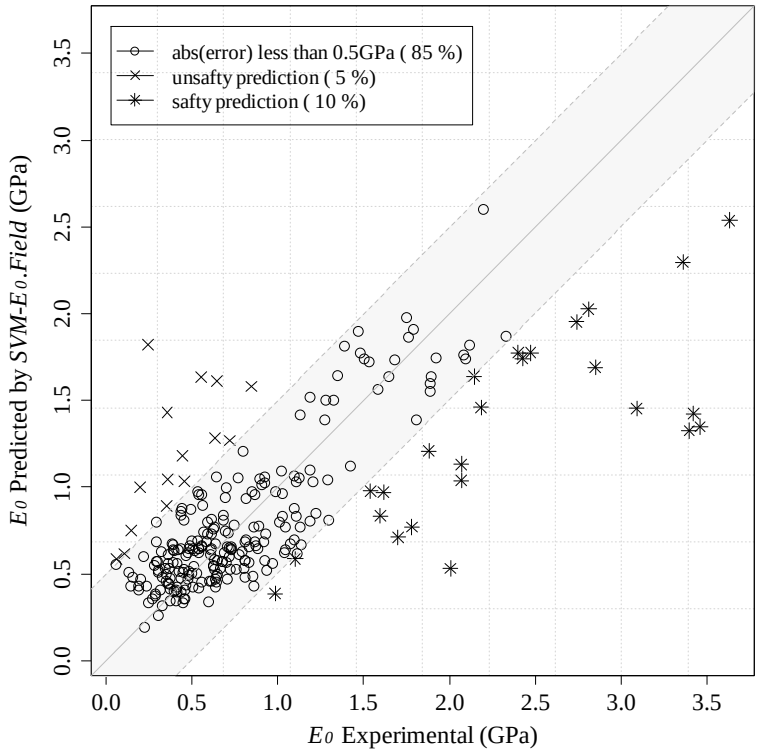


Figure 6.14: Relationship between E_0 experimental versus predicted values by *SVM-*E_0*.Field* model

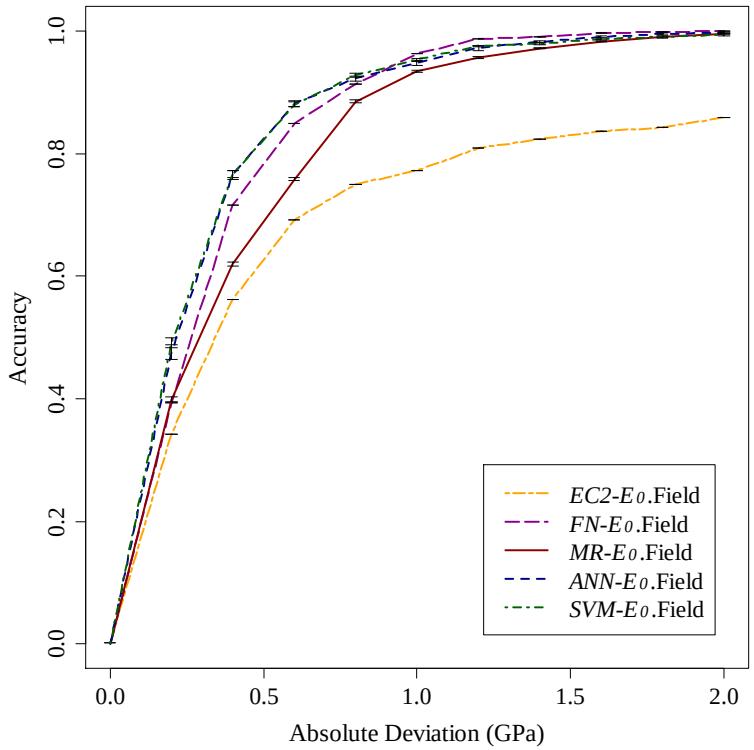


Figure 6.15: *REC* curves of *EC2-E_0.Field*, *FN-E_0.Field*, *MR-E_0.Field*, *ANN-E_0.Field* and *SVM-E_0.Field* models, comparing its performance in E_0 prediction of *soilcrete*

6.4.2 Model interpretability

As previously underlined, namely in Chapter 2, the main drawback of complex *DM* models is related with its interpretability, due to high mathematical complexity of the algorithms. Thus, although the difficulty found by the proposed models for *soilcrete* stiffness prediction, as above presented, the application of a *GSA* over such models, namely *ANN-E_0.Field* and *SVM-E_0.Field*, can give a valuable help in models interpretability and *soilcrete* stiffness behaviour. Accordingly, and based on a 1-D *SA*, Figure 6.16 shows and compares the relative importance of each variable according to *FN-E_0.Field*, *MR-E_0.Field*, *ANN-E_0.Field* and *SVM-E_0.Field* models.

Analysing Figure 6.16, and according to *SVM-E_0.Field* model, it is observed that the t and C_{iv} present the highest impact in E_0 prediction of *soilcrete* (Gomes Correia et al., 2011). Based on the *ANN-E_0.Field* model, e and ρ_d also have an important influence in E_0 behaviour. Although with a similar performance in terms of *MAD*, *RMSE* and R^2, *FN-E_0.Field* model does not have a physical meaning in terms of relative importance of the attributes, because considers that the e is the only variable that controls E_0 behaviour of *soilcrete* mixtures. Moreover, it is appealing to observe, particularly according to *SVM-E_0.Field* model, that the jet system applied shows just a slightly influence in

stiffness prediction of *soilcrete* mixtures. This behaviour may be related with the statistical distribution of such variable in the database. Indeed, a significant number of records are from double jet system and just few records are from single and triple fluid system.

Based on the interpretation of Figure 6.16, as well as in models performance shown in Table 6.9, the *SVM-E_0.Field* seems to be the most interesting one to predict *soilcrete* stiffness with the highest accuracy. Moreover, *SVM* algorithm shows good learning capabilities in *JGLF* study and in *soilcrete* strength prediction. Therefore, *SVM-E_0.Field* model will be used as reference in the following analysis, performed toward to a better understanding of *soilcrete* stiffness behaviour.

To improve model interpretability and better understand what has been learned by *SVM-E_0.Field* model, the *VEC* curves of its three key input variables, identified in Figure 6.16 are plotted in Figure 6.17. Both *VEC* curves of t and C_{iv} show a positive effect in deformability properties of *soilcrete* mixtures (Tinoco et al., 2012c). Particularly, the concave shape of t *VEC* curve corroborates once again the exponential effect of t in soil-cement mixtures behaviour (Coulter and Martin, 2006; Van Impe et al., 2005). On the other hand, the convex shape of C_{iv} *VEC* curve gives the idea that for lower cement contents, *soilcrete* stiffens just slight increases with C_{iv} and only after a given dosage (around $0.20 \Rightarrow 0.40$ according to the scaled x-axis of Figure 6.17), it increases quickly. The *VEC* curve for ω presents an unexpected shape, namely for high water contents of the mixture, where *soilcrete* stiffness increases with ω (Liu et al., 2008). This unexpected behaviour is probably related with the interaction between variables that forced mixtures with high ω to reach higher stiffness than other with low ω. The not so high *SVM-E_0.Field* model accuracy can also contribute for such behaviour. As previously shown, all models, even *SVM-E_0.Field* experienced some difficulties to learn the complex relationships between E_0 and its contributing factors.

Aiming a more realistic interpretation of the models and a more detailed understanding of *soilcrete* stiffness behaviour, a 2-D *SA* was performed over *SVM-E_0.Field*, allowing to measure the interaction level between variables, as well as its effect in E_0 prediction of *soilcrete* mixtures. Figure 6.18a plots the interaction level between all variables with t where W/C is in the top of the ranking with an relative importance around 17%. This observation shows that although W/C is considered the second variable with less impact in E_0 prediction of *soilcrete* mixtures (see Figure 6.16), it should also be considered in *soilcrete* stiffness behaviour. It is appealing to observe that this behaviour was also identified is strength study of *soilcrete* mixtures.

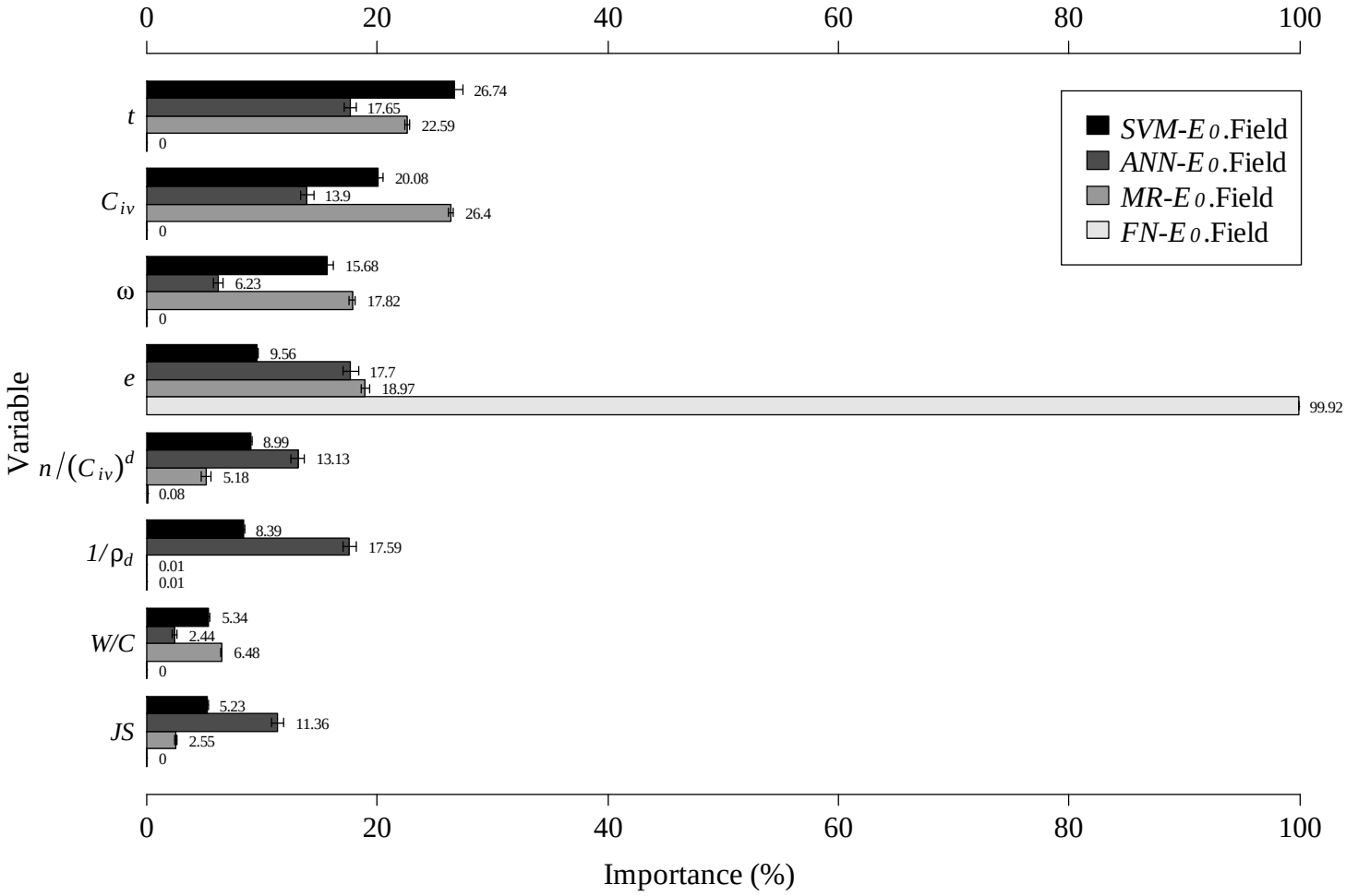


Figure 6.16: Relative importance of each input variable quantified by 1-D *SA*, comparing *FN-E_0.Field*, *MR-E_0.Field*, *ANN-E_0.Field* and *SVM-E_0.Field* models

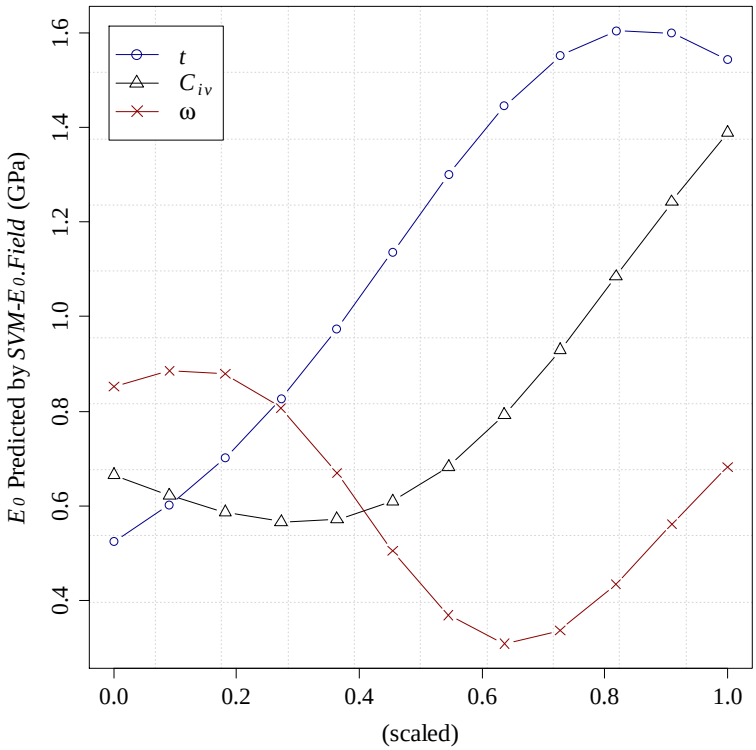
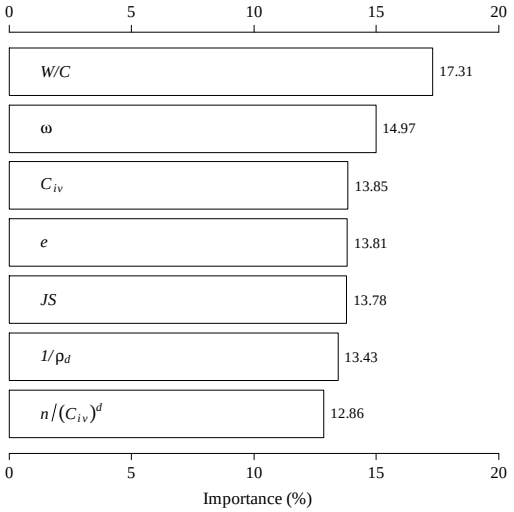
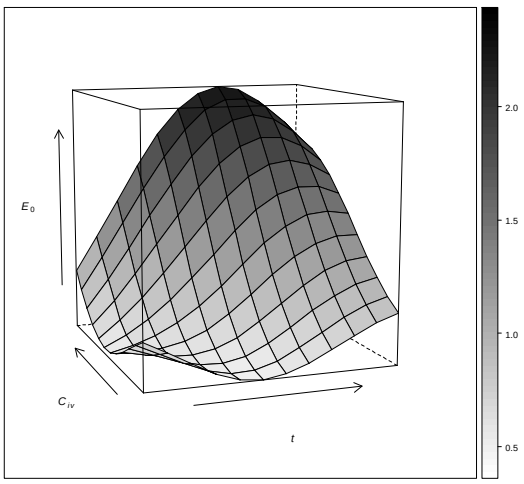


Figure 6.17: *VEC* curves for the three key input variables according to *SVM-E_0.Field* model on *soilcrete* stiffness prediction, quantified by 1-D *SA*

The *VEC* surface of t and C_{iv} interaction, depicted in Figure 6.18b, illustrated that the stiffness gain is proportional to t and C_{iv}. This means that, for instance, the gain of stiffness over time is higher in mixtures with higher cement content. Observing *VEC*



(a)



(b)

Figure 6.18: 2-D *SA* according to *SVM-E_0.Field* model in E_0 prediction of *soilcrete*: a) interaction level between all variables with t and b) *VEC* surface for t and C_{iv} interaction

contour plotted in Figure 6.19a, it is pointed out that the gain of E_0 through the time is faster for mixtures with low ω. On *VEC* contour of t and W/C interaction, depicted in Figure 6.19b, it is observed a slight increase of *soilcrete* stiffness when W/C increases, mainly for advanced ages (Lee et al., 2005). Although not expected, this phenomenon is probably related to the low relative importance of W/C (see Figure 6.16), as well as to the non ideal performance of *SVM-E_0.Field* model.

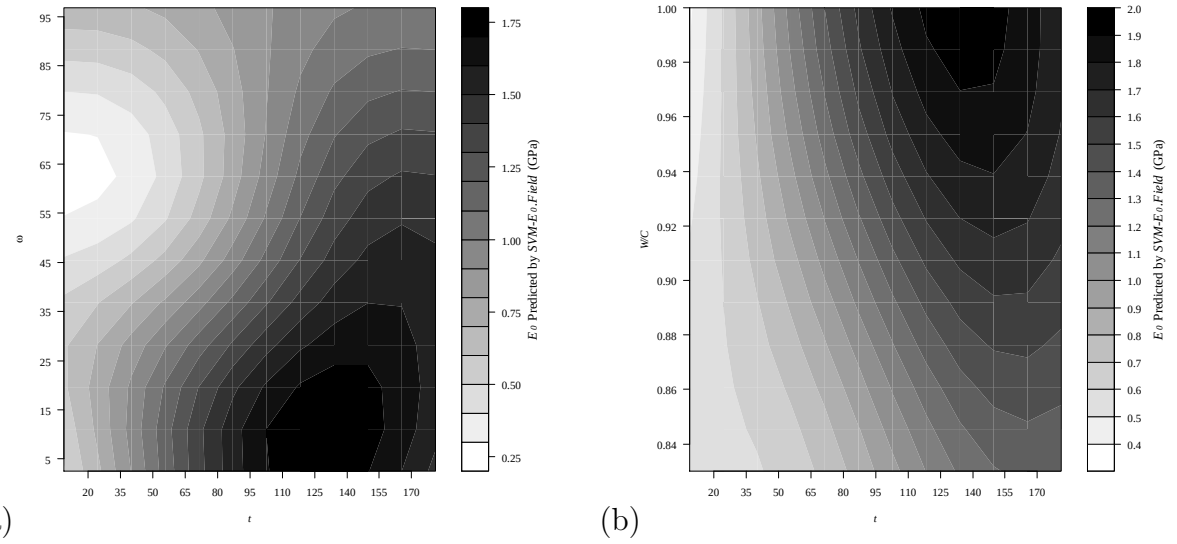


Figure 6.19: 2-D *SA* according to *SVM-E_0.Field* model in E_0 prediction of *soilcrete*: a) *VEC* contour for t and ω interaction and b) *VEC* contour for t and W/C interaction

6.5 Soilcrete strength and stiffness - comparison and relationship

Bearing capacity of *JG* columns is normally performed based on *soilcrete* mechanical properties, i.e. its strength and stiffness. Table 6.10 summarizes the metrics values (*MAD*, *RMSE* and R^2) of all models so far proposed for *UCS* and E_0 prediction of *soilcrete*, comparing its performance. A global overview of this table, and using R^2 as performance criterion, shows that for both mechanical properties prediction each algorithm achieved a similar performance. The only exception is for *FN* algorithm that performs better strength prediction then stiffness of *soilcrete* mixtures. Moreover, it is underlined the poor performance of *EC2* models in both mechanical properties prediction.

Figure 6.20 compares the relative importance of each input variable in *UCS* and E_0 prediction of *soilcrete* mixtures according to *SVM-UCS.Field* and *SVM-E_0.Field* models. This figure shows that there are some differences between the key variables in the strength

Table 6.10: Comparison of the performance of all predictive models in UCS and E_0 of *soilcrete* using *MAD*, *RMSE* and R^2 as performance criteria

Model	MAD	RMSE	R^2
EC2-UCS.Field	1.75 ± 0.00	2.65 ± 0.00	0.13 ± 0.00
EC2-E_0.Field	0.84 ± 0.00	1.52 ± 0.00	0.24 ± 0.00
FN-UCS.Field	1.40 ± 0.00	1.95 ± 0.00	0.19 ± 0.00
FN-E_0.Field	0.34 ± 0.00	0.45 ± 0.00	0.55 ± 0.00
MR-UCS.Field	1.53 ± 0.00	2.13 ± 0.01	0.43 ± 0.00
MR-E_0.Field	0.39 ± 0.00	0.55 ± 0.00	0.33 ± 0.01
ANN-UCS.Field	1.41 ± 0.02	2.01 ± 0.06	0.49 ± 0.03
ANN-E_0.Field	0.31 ± 0.01	0.46 ± 0.01	0.54 ± 0.02
SVM-UCS.Field	1.38 ± 0.01	1.99 ± 0.01	0.51 ± 0.01
SVM-E_0.Field	0.31 ± 0.00	0.46 ± 0.01	0.53 ± 0.01

and stiffness of *soilcrete* mixtures behaviour. While in UCS study the three most relevant variables are $n/(C_{iv})^d$, JS and t, in *soilcrete* stiffness study the key variables are t, C_{iv} and ω. Among the key variables in strength and stiffness study of *soilcrete* mixtures, t is the only one common to both mechanical properties, although with different relative importances. It is also observed that both models (i.e. *SVM-UCS.Field* and *SVM-E_0.Field* models) also include C_{iv} as a key variable. However, in the case of UCS prediction this variable is only considered indirectly through $n/(C_{iv})^d$ relation. For the remaining variables, significant differences are observed.

As previously highlighted, the prediction of *soilcrete* stiffness based on its strength values has an important practical application, particularly because the tests for measuring mixtures deformability are more expensive. Accordingly, and similar to what was done for *JGLF* presented in Section 5.4, we present a novel approach, aiming to predict *soilcrete* stiffness based on the UCS of the respective mixture, and considering some additional elementary variables. The proposed approach, developed using *DM* techniques, is intended to predict E_0 of *soilcrete* mixtures based on $n/(C_{iv})^d$, JS, t, C_{iv} and ω, as well as the UCS of the mix at the same age. The choice of these set of input variables is essentially supported on the observation of Figure 6.20, where it is found a significant difference in its relative importance depending on whether strength or stiffness is studied, which means that probably these variables make the bridge between these two mechanical properties. Moreover, some of these variables, namely t and C_{iv}, were also identified as relevant for this correlation in *JGLF* study (see Figure 5.30).

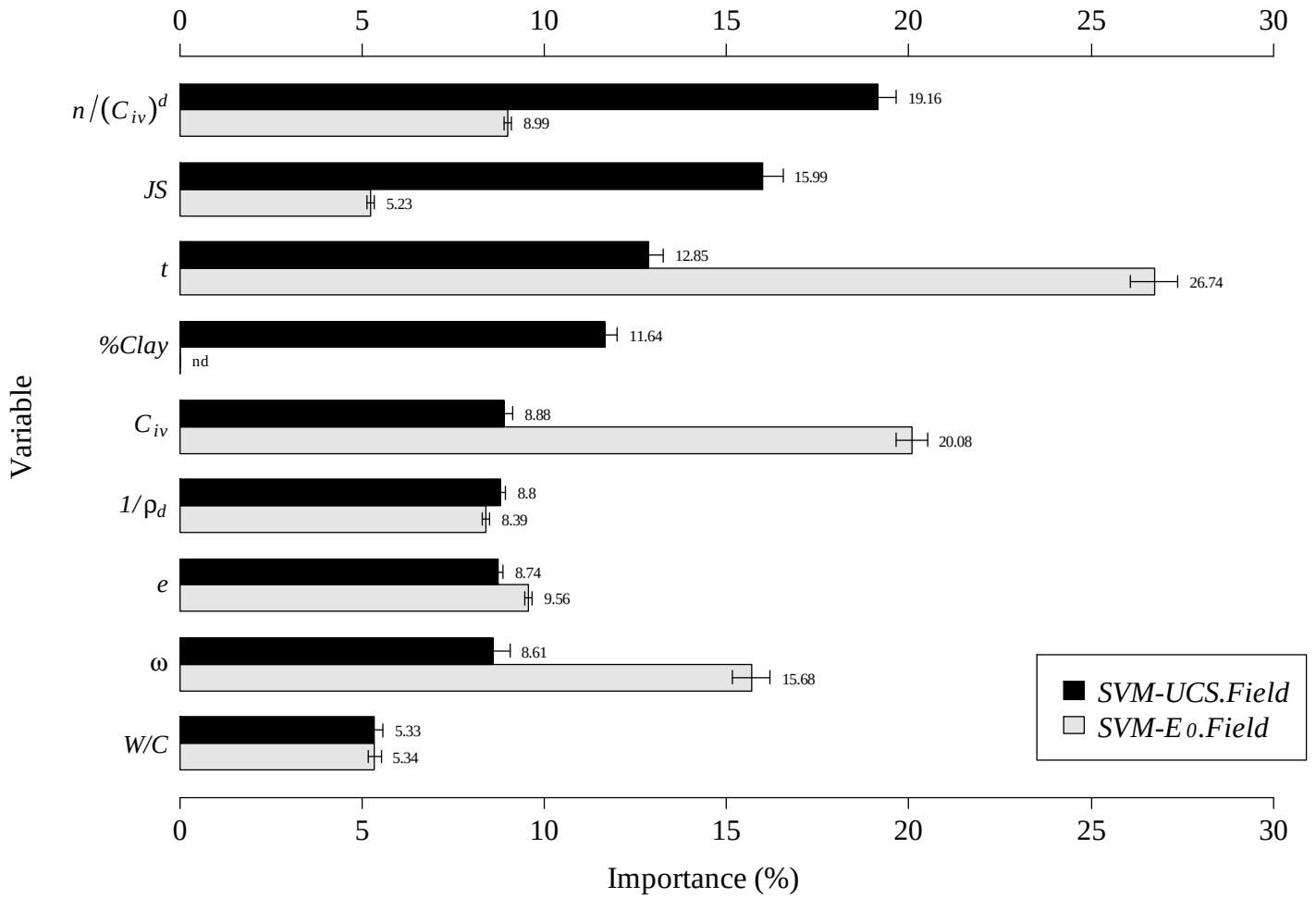


Figure 6.20: Comparison of the relative importance of each variable in UCS and E_0 prediction of *soilcrete* mixtures, according to *SVM-UCS.Field* and *SVM-E_0.Field* models

Table 6.11 summarizes the main statistics of both input and output variables used during this experiment, i.e. to predict E_0 of *soilcrete* mixtures as a function of the respective UCS. For this exercise, we only applied the two DM algorithms that achieved the best global performance throughout this research work, i.e. the algorithms ANN and SVM. Additionally, it was also applied the MR algorithm for a baseline comparison. After training these three algorithms using the database characterized in Table 6.11 and the same hyperparameters and considerations underlined in Section 6.1 (i.e. ANN activation function, model generalization approaches, etc.), the obtained models will be termed as *MR-*E_0UCS*.Field*, *ANN-*E_0UCS*.Field* and *SVM-*E_0UCS*.Field*, respectively. Table 6.12 summarizes the averaged hyperparameters and fitting time values (and respective 95% level confidence intervals according to a t-student distribution) of *MR-*E_0UCS*.Field*, *ANN-*E_0UCS*.Field* and *SVM-*E_0UCS*.Field* models. The predictive capacity of *MR-*E_0UCS*.Field*, *ANN-*E_0UCS*.Field* and *SVM-*E_0UCS*.Field* models is compared in Table 6.13 (mean value and 95% confidence intervals), using MAD, $RMSE$ and R^2 metrics as a performance criteria.

Table 6.11: Summary statistics for both input and output variables of the database used during the experiments performed with the goal to correlate E_0 and UCS of *soilcrete* mixtures

Variable	Minimum	Maximum	Mean	Standard Deviation
JS	1.00	3.00	2.04	0.51
ω	2.50	96.80	36.38	13.34
UCS	0.32	20.27	4.03	3.15
t	9.00	181.00	36.72	35.07
C_{iv}	0.14	0.28	0.21	0.04
$n/(C_{iv})^d$	37.88	78.61	58.00	7.50
E_0	0.06	3.63	0.89	0.68

Table 6.12: Hyperparameters and computation time of *MR-*E_0UCS*.Field*, *ANN-*E_0UCS*.Field* and *SVM-*E_0UCS*.Field* models, used in E_0 prediction of *soilcrete* material

Model	Hyperparameters	time (s)
MR-E_0UCS*.Field*	-	0.85 ± 0.01
ANN-E_0UCS*.Field*	$H = 4 \pm 1$	52.89 ± 0.04
SVM-E_0UCS*.Field*	$\gamma = 0.26 \pm 0.02$, $\epsilon = 0.11 \pm 0.00$	38.61 ± 0.11

Table 6.13: Error metrics of *MR-$E_0UCS.Field$*, *ANN-$E_0UCS.Field$* and *SVM-$E_0UCS.Field$* models, used for E_0 prediction of *soilcrete*, and its comparison with *ANN-E_0.Field* and *SVM-E_0.Field* models (test set values, best values in **bold**)

Model	MAD	RMSE	R^2
MR-$E_0UCS.Field$	0.36 ± 0.00	0.52 ± 0.00	0.41 ± 0.00
ANN-$E_0UCS.Field$	$\mathbf{0.26} \pm 0.00$	$\mathbf{0.43} \pm 0.03$	$\mathbf{0.59} \pm 0.06$
SVM-$E_0UCS.Field$	$\mathbf{0.26} \pm 0.00$	$\mathbf{0.43} \pm 0.00$	$\mathbf{0.59} \pm 0.01$
ANN-E_0.Field	$\mathbf{0.31} \pm 0.01$	0.46 ± 0.01	0.54 ± 0.02
SVM-E_0.Field	0.31 ± 0.00	0.46 ± 0.01	0.53 ± 0.01

Although not very accurate, *ANN-$E_0UCS.Field$* and *SVM-$E_0UCS.Field$* models performs better E_0 prediction when compared to the *ANN-E_0.Field* and *SVM-E_0.Field* models presented and discussed in Section 6.4. Figure 6.21 plots the relationship between E_0 experimental values versus predicted by *ANN-$E_0UCS.Field$* and *SVM-$E_0UCS.Field$* models, corroborating its better accuracy in E_0 prediction of *soilcrete* mixtures when compared to *ANN-E_0.Field* and *SVM-E_0.Field* models (see Figures 6.13 and 6.14 respectively). Indeed, both these new models are able to perform E_0 prediction of *soilcrete* within an absolute deviation lower than 0.5 GPa for 88% of the records, which represent an improvement around 4%. Figure 6.22 compares the performance of *ANN-$E_0UCS.Field$*, *SVM-$E_0UCS.Field$*, *ANN-E_0.Field* and *SVM-E_0.Field* models in E_0 prediction of *soilcrete* mixtures throughout the *REC* curves (Bi and Bennett, 2003). It is shown that *SVM-$E_0UCS.Field$* model is able to predict E_0 of *soilcrete* mixtures more accurately than *ANN-$E_0UCS.Field$*, as well as its superiority when compared with *ANN-E_0.Field* and *SVM-E_0.Field* models.

From these observations can be pointed out that E_0 prediction of *soilcrete* mixtures using UCS as input variable leads to a more reliable results. Hence, it is recommended the use of UCS as an input variable in *soilcrete* stiffness prediction whenever this information is available.

The relative importance of each one of the six input variables, according to *ANN-$E_0UCS.Field$* and *SVM-$E_0UCS.Field$* models, was measured based on a 1-D *SA*. Figure 6.23 illustrates that in both models t and UCS are the two most relevant variables in *soilcrete* stiffness prediction with an total influence around 50%. It is also appealing to observe that the effect of both variables is almost linear, particularly according to ANN-E_0UCS.field model, as depicted in Figure 6.24.

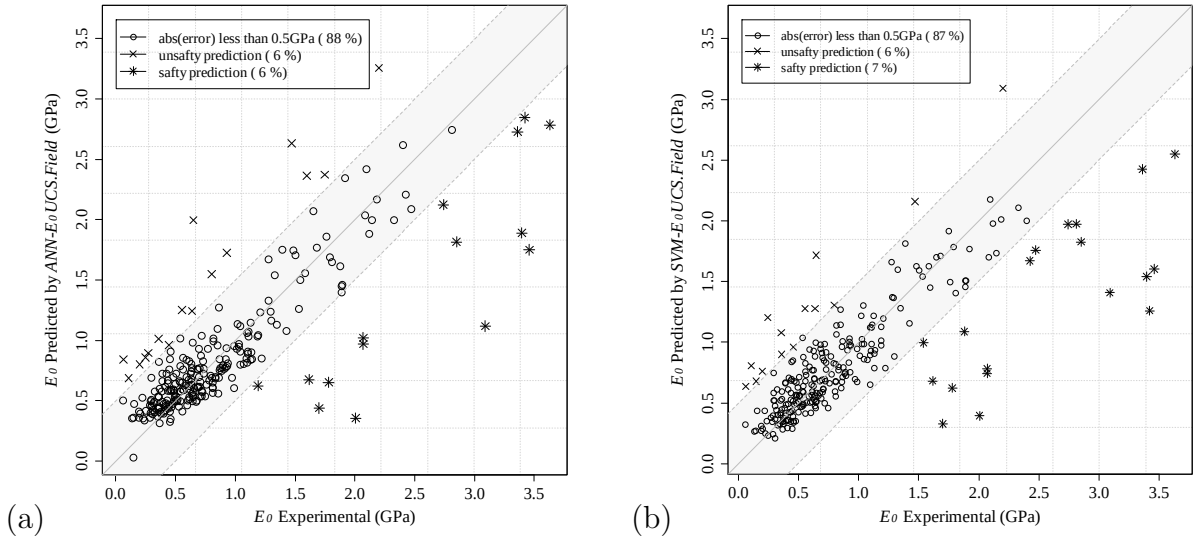


Figure 6.21: Relationship between E_0 experimental versus predicted values by: a) *ANN-*E_0*UCS.Field* model and b) *SVM-*E_0*UCS.Field* model

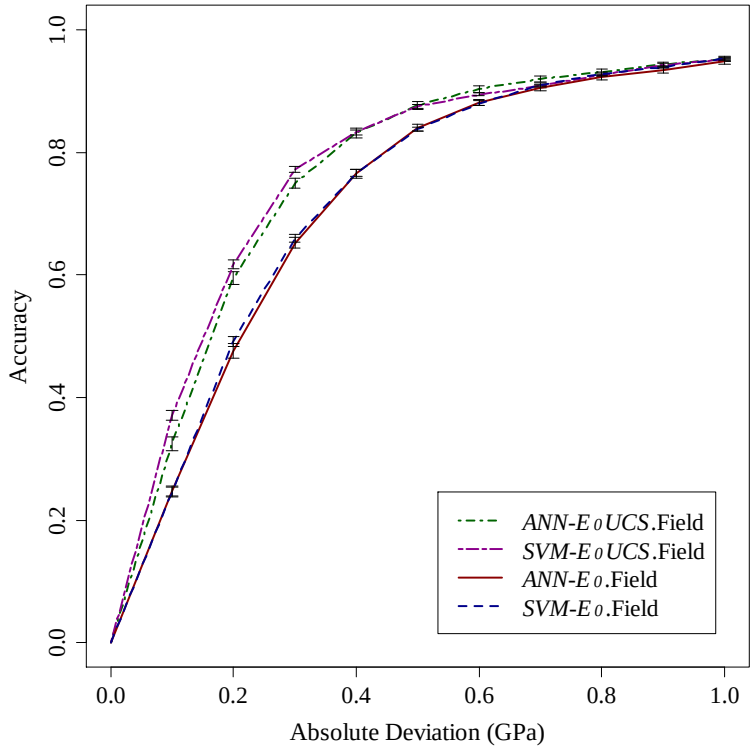


Figure 6.22: *REC* curves of *ANN-*E_0*.Field*, *SVM-*E_0*.Field*, *ANN-*E_0*UCS.Field* and *SVM-*E_0*UCS.Field* models, comparing its performance in E_0 prediction of *soilcrete* mixtures

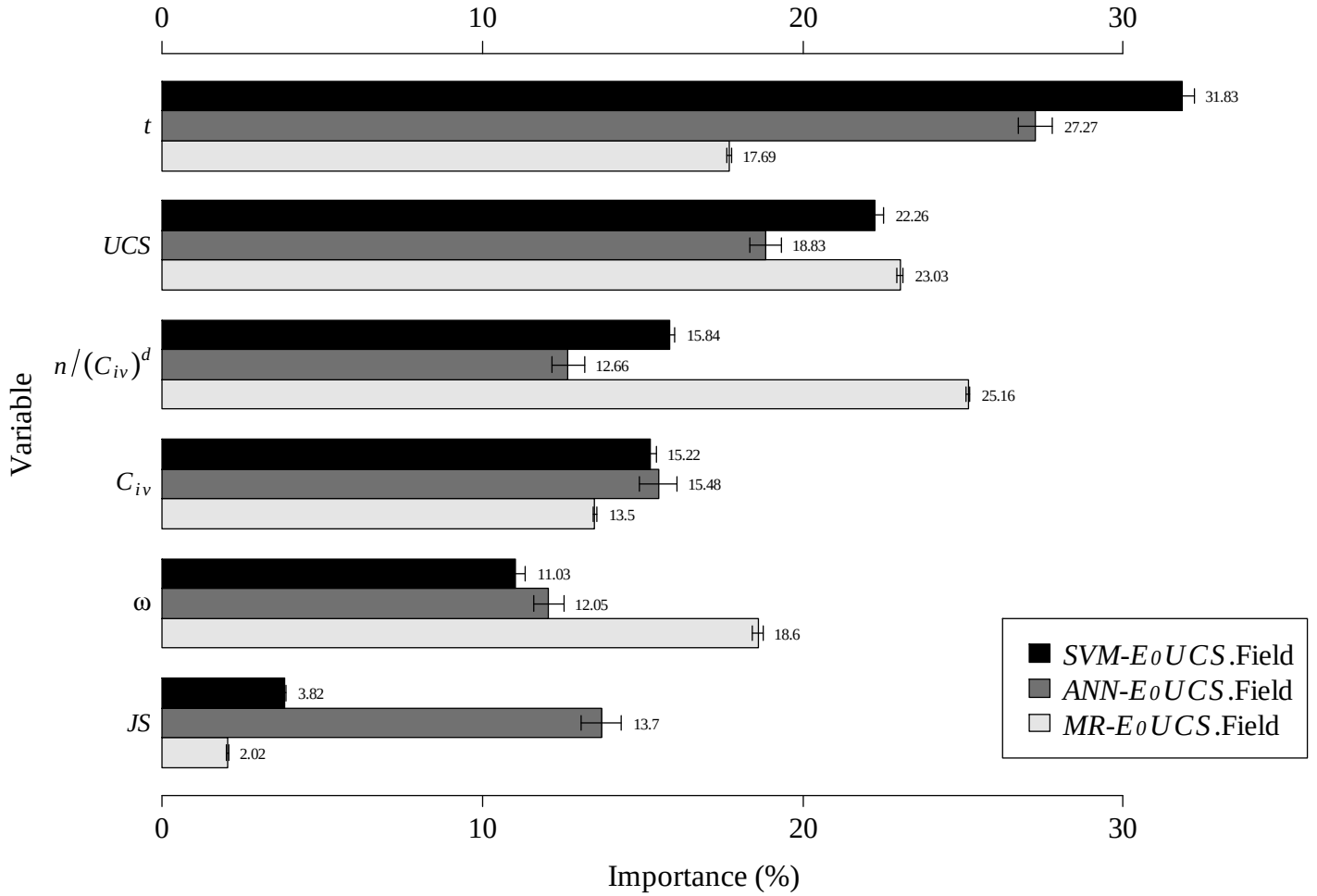


Figure 6.23: Relative importance of each input variable quantified by 1-D *SA*, comparing *MR-E_0UCS.Field*, *ANN-E_0UCS.Field* and *SVM-E_0UCS.Field* models

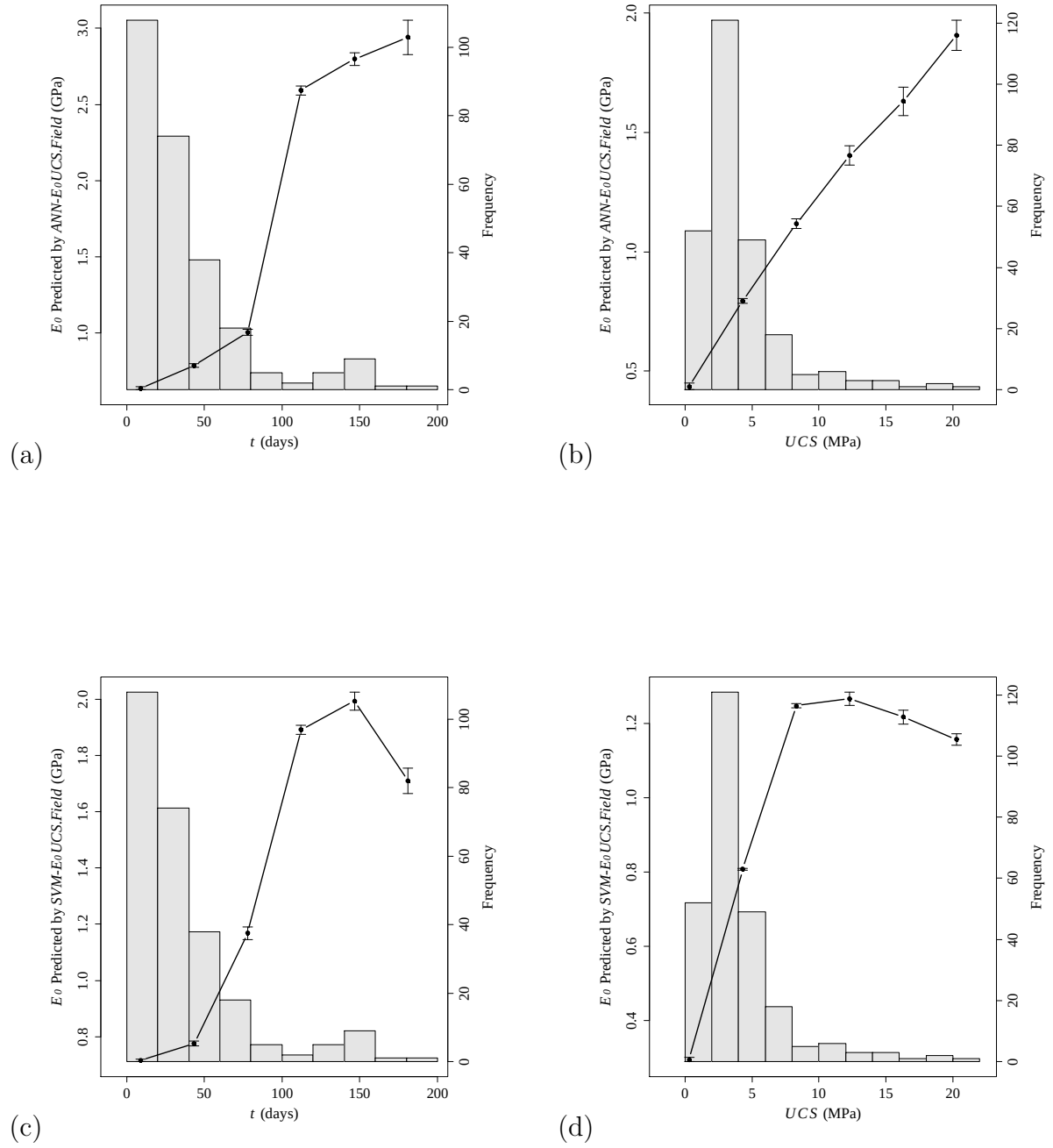


Figure 6.24: *VEC* curves of: a) t according to *ANN-E_0UCS.Field* model; b) UCS according to *ANN-E_0UCS.Field* model; c) t according to *SVM-E_0UCS.Field* model and d) UCS according to *SVM-E_0UCS.Field* model, on *soilcrete* stiffness prediction, quantified by 1-D *SA*

A 2-D *SA* over *ANN-E_0UCS.Field* corroborates the strong influence of t and UCS in *soilcrete* stiffness prediction, as shown in Figure 6.25a that depicts the interaction level between all variables with UCS. The effect of UCS and t interaction is plotted in Figure 6.25b, denoting a uniform influence of both variables in E_0 behaviour of *soilcrete* material.

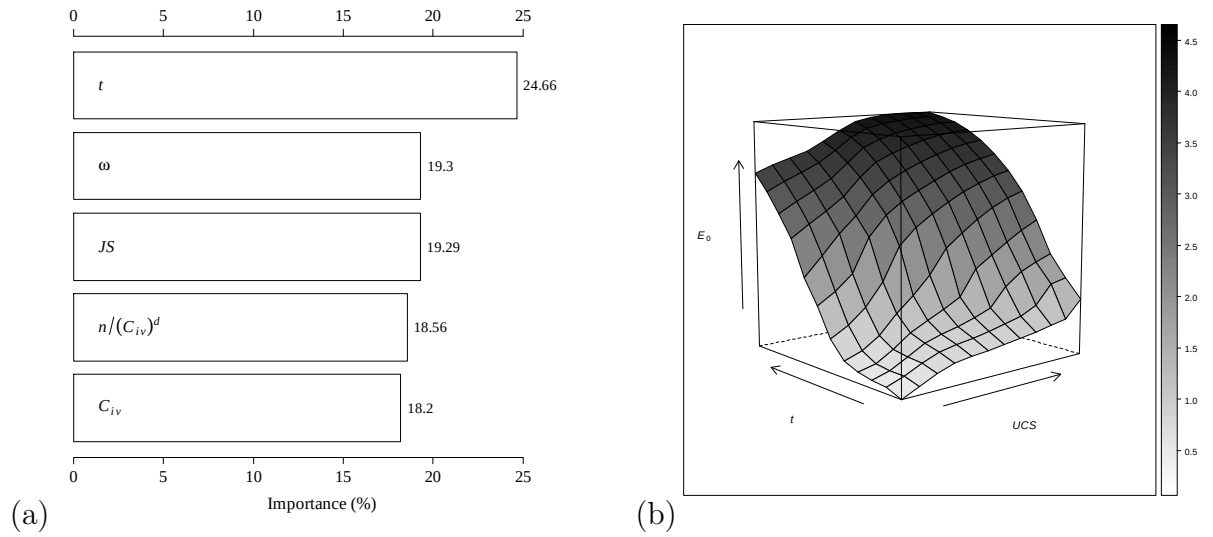


Figure 6.25: 2-D *SA* according to *ANN-E_0UCS.Field* model in E_0 prediction of *soilcrete*: a) interaction level between all variables with UCS and b) *VEC* surface for UCS and t interaction

6.6 Diameter prediction

6.6.1 Model performance

In this section, we present and discuss the proposed models for D prediction, developed through the application *DM* tools. Relating to this task, it should be underlined that the learning process was supported on a database that includes information from test columns, for which the diameter was measured, and project columns, for which the diameter is assumed equal to the test columns, since these columns are built under the same conditions.

Table 6.14 compares the *SVM* models performance of the two *FS* approaches implemented, i.e. forward and backward methods, with the manual selection that took into account the knowledge acquired from literature review and balanced with the information given by *FS* approaches.

Accordingly, Table 6.15 summarizes the main statistics of the database used during the study of *JG* column diameter, i.e. the database that includes just the nine variables assigned in Table 6.14 as MS_{Df1}, which encompasses 632 records (403 from test columns and 229 from project columns).

The average hyperparameters and fitting time values (and respective 95% level confidence intervals according to a t-student distribution) of all *DM* models trained using the set of eight input variables assigned in Table 6.14 as MS_{Df1} are shown in Table 6.16. These models, developed to predict *JG* column diameter will be termed as *MR-D.Field*,

Table 6.14: Comparison of the *SVM* models performance developed using the forward and backward *FS* approaches with the manual selection, aiming to predict D

Var	FFS	BFS	MS_{Df1}
JS	×	✓	✓
FR	✓	✓	✓
WS	✓	✓	✓
Imp_{grout}	×	✓	✓
P_{grout}	×	✓	✓
D_{grout}	✓	✓	✓
$\%Sand$	✓	×	✓
$\%Clay$	×	×	✓
WT	×	✓	×
rmp	✓	✓	×
kg/m^3	✓	✓	×
kg/ml	×	✓	×
W/C	✓	✓	×
ρ_{grout}	×	✓	×
P_{water}	×	✓	×
P_{air}	×	✓	×
D_{water}	×	✓	×
MAD	0.92 ± 0.40	0.97 ± 0.22	0.23 ± 0.22
RMSE	3.69 ± 6.47	4.87 ± 1.78	2.27 ± 3.88
R^2	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

FFS - forward feature selection; BFS - backward feature selection

Table 6.15: Summary statistics for both input and output variables of the database used during the study of D, which contemplates the eight input variables assigned in Table 6.14 as MS_{Df1}

Variable	Minimum	Maximum	Mean	Standard Deviation
JS	1.00	3.00	2.05	0.37
WS	6.00	21.82	10.10	4.22
FR	139.00	577.89	363.30	79.06
D_{grout}	3.80	7.00	4.89	0.84
P_{grout}	140.00	450.00	355.78	85.37
Imp_{grout}	58.06	278.95	213.56	69.12
$\%Sand$	0.01	39.00	22.15	16.79
$\%Clay$	22.50	45.00	32.89	7.28
D	800.00	3008.00	2184.55	432.12

ANN-D.Field and *SVM-D.Field*, and are respectively the result of the training of *MR*, *ANN* and *SVM* algorithms with *JG* column diameter data.

Table 6.16: Hyperparameters and computation time of each *DM* model for *D* prediction

Model	Hyperparameters	time (s)
MR-D.Field	-	1.21 ± 0.02
ANN-D.Field	$H = 8 \pm 1$	115.92 ± 0.82
SVM-D.Field	$\gamma = 1.98 \pm 0.24$, $\epsilon = 2.61\text{E}^{-5} \pm 4.48\text{E}^{-7}$	112.64 ± 0.34

Table 6.17 shows the predictive capacity of all trained models, comparing its performance on *JG* column diameter prediction based on the *MAD*, *RMSE* and R^2 metrics, computed for the test data under a 20-fold cross-validation approach (mean value and 95% confidence intervals). Analysing Table 6.17, it is concluded that *JG* column diameter prediction was correctly learned by both *ANN* and *SVM* algorithms. Indeed, *ANN-D.Field* and *SVM-D.Field* models achieved an $R^2 = 1$ in such task.

Table 6.17: Error metrics of all *DM* models for *D* prediction (test set values, best values in **bold**)

Model	MAD	RMSE	R^2
MR-D.Field	76.97 ± 0.09	125.46 ± 0.14	0.92 ± 0.00
ANN-D.Field	0.83 ± 0.16	2.78 ± 2.58	$\mathbf{1.00} \pm 0.00$
SVM-D.Field	$\mathbf{0.23} \pm 0.22$	$\mathbf{2.27} \pm 3.38$	$\mathbf{1.00} \pm 0.00$

This excellent performance is shown in the Scatterplots shown in Figure 6.26, where the predictions according to *ANN-D.Field* and *SVM-D.Field* models are very close with the experimental ones (diagonal line) for both test and project columns. Figure 6.26a illustrated the difficulty of predicting *JG* column diameter based on linear laws, which is corroborated by the *REC* curves plotted in Figure 6.26d. This figure also illustrates once more the very high accuracy of *ANN-D.Field* and *SVM-D.Field* models in *JG* column diameter prediction. As shown, both *ANN-D.Field* and *SVM-D.Field* models are able to predict almost all records of the database with and absolute deviation lower than 0.5 mm.

6.6.2 Model interpretability

In order to identify what are the most relevant variables in *JG* column diameter prediction, a 1-D *GSA* was performed over *MR-D.Field*, *ANN-D.Field* and *SVM-D.Field* models.

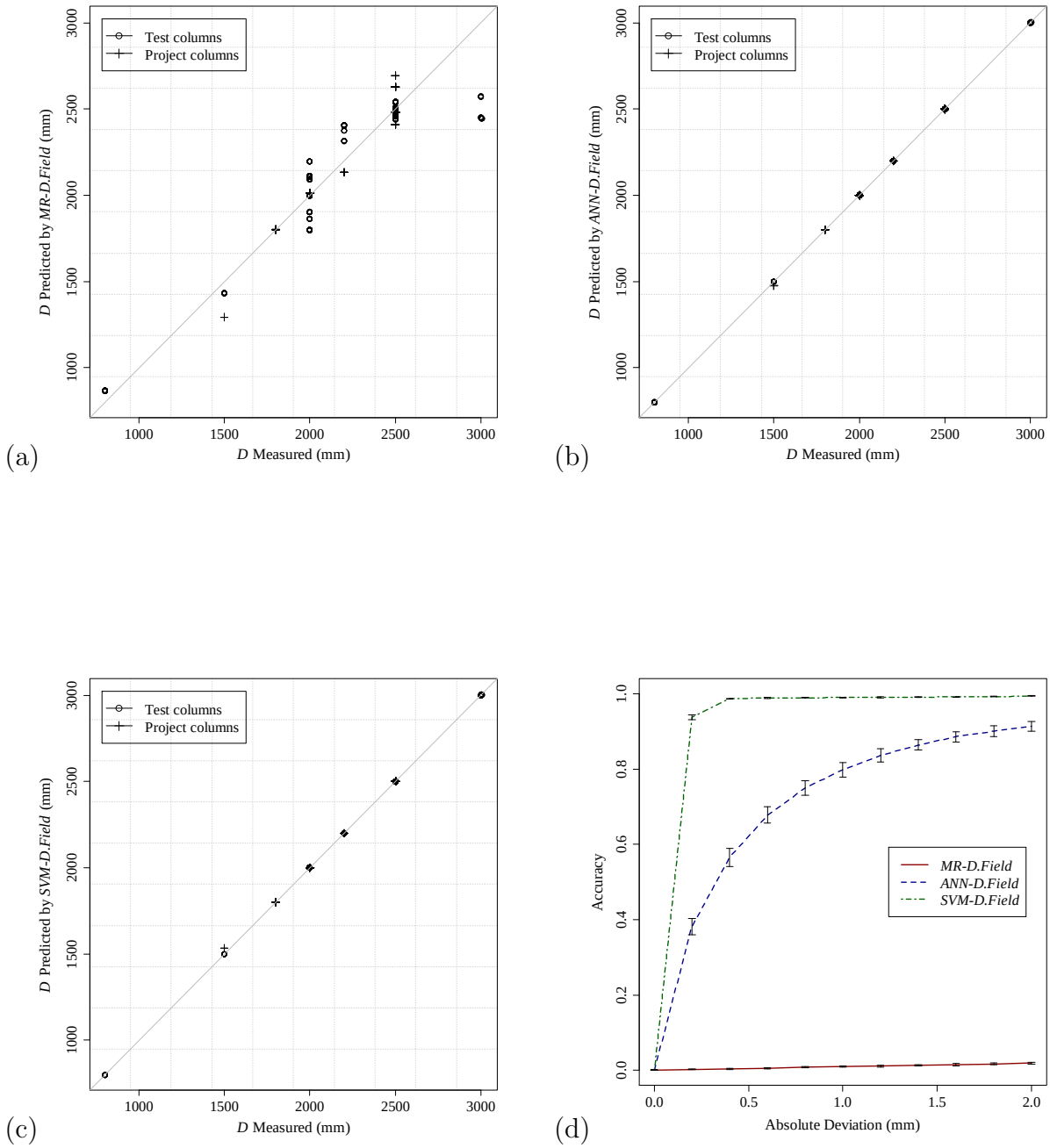


Figure 6.26: Relationship between D measured versus predicted values by: a) *MR-D.Field* model; b) *ANN-D.Field* model and c) *SVM-D.Field* model. In d) it is plotted the *REC* curves of *MR-D.Field*, *ANN-D.Field* and *SVM-D.Field* models, comparing its performance in D prediction

Figure 6.27 compares the relative importance of each variables showing that, although *ANN-D.Field* and *SVM-D.Field* models are both very accurate in *JG* column diameter prediction, they are not guided exactly by the same variables. According to *SVM-D.Field* model, the $\%Sand$, WS, $\%Clay$ and D_{grout} are the for key variables in *JG* column diameter prediction. On the other hand, *ANN-D.Field* model presents an relative importance distribution more uniform where P_{grout} and WS are in the top of the ranking. Looking to the key variables according to each model, we can conclude that *SVM-D.Field* model predicts *JG* column diameter as a function of the soil properties ($\%Sand$ and $\%Clay$ have an total influence around 44%). On the other hand, and according to *ANN-D.Field* model, *JG* column diameter is particularly related with the energy applied during the *JG* soil improvement (soil properties just have an influence around 15%). Together, these two models combine the observations performed by Modoni et al. (2006) on their theoretical approach for *JG* column diameter prediction, i.e. the interaction between the soil and the jet energy on *JG* column diameter development.

Aiming to understand how *ANN-D.Field* and *SVM-D.Field* models learned the effect of the grout jet and soil properties in *JG* column diameter development, a *GSA* was performed over these two models. Accordingly, and based on a 1-D *SA*, the *VEC* curves of P_{grout} and WS were calculated using the *ANN-D.Field* model. Figure 6.28a plots the *VEC* curve P_{grout}, showing that *JG* column diameter decreases when the jet grout pressure increases. This behaviour, apparently not expected, can be explained by the concepts behind the different *JG* systems. Indeed, the grout pressure used in the triple fluid system is normally lower than in single fluid system, as illustrated in Figure 6.28b, because its main function is "just" to mix the fragmented soil with the cement slurry. However, it is known (Essler and Yoshida, 2004) that the diameter of *JG* column built with single fluid system is lower than by triple fluid system, as a result of the highest energy applied in triple system, supplied by the additional water jet involved by pressurized air that cut the soil before apply the grout jet. Therefore, since the effect of the water jet used in double or triple fluid systems is not available to the model, it learned the effect of the jet fluids just using the grout pressure. The effect of the WS and JS in *JG* column diameter prediction is depicted in Figure 6.29, showing, as expected, that the column diameter decreases with the increasing of the WS according to a logarithm law, and increases almost linearly from single to triple fluid system.

The *VEC* curves of $\%Sand$, WS and $\%Clay$ according to *SVM-D.Field* model are plotted in Figure 6.30, with the intention of explaining the effect of soil in *JG* column diameter development. On one hand, it is observed that the *VEC* curve of WS presents the same shape that in *ANN-D.Field* model, i.e. that *JG* column diameter decreases with the increasing of WS.

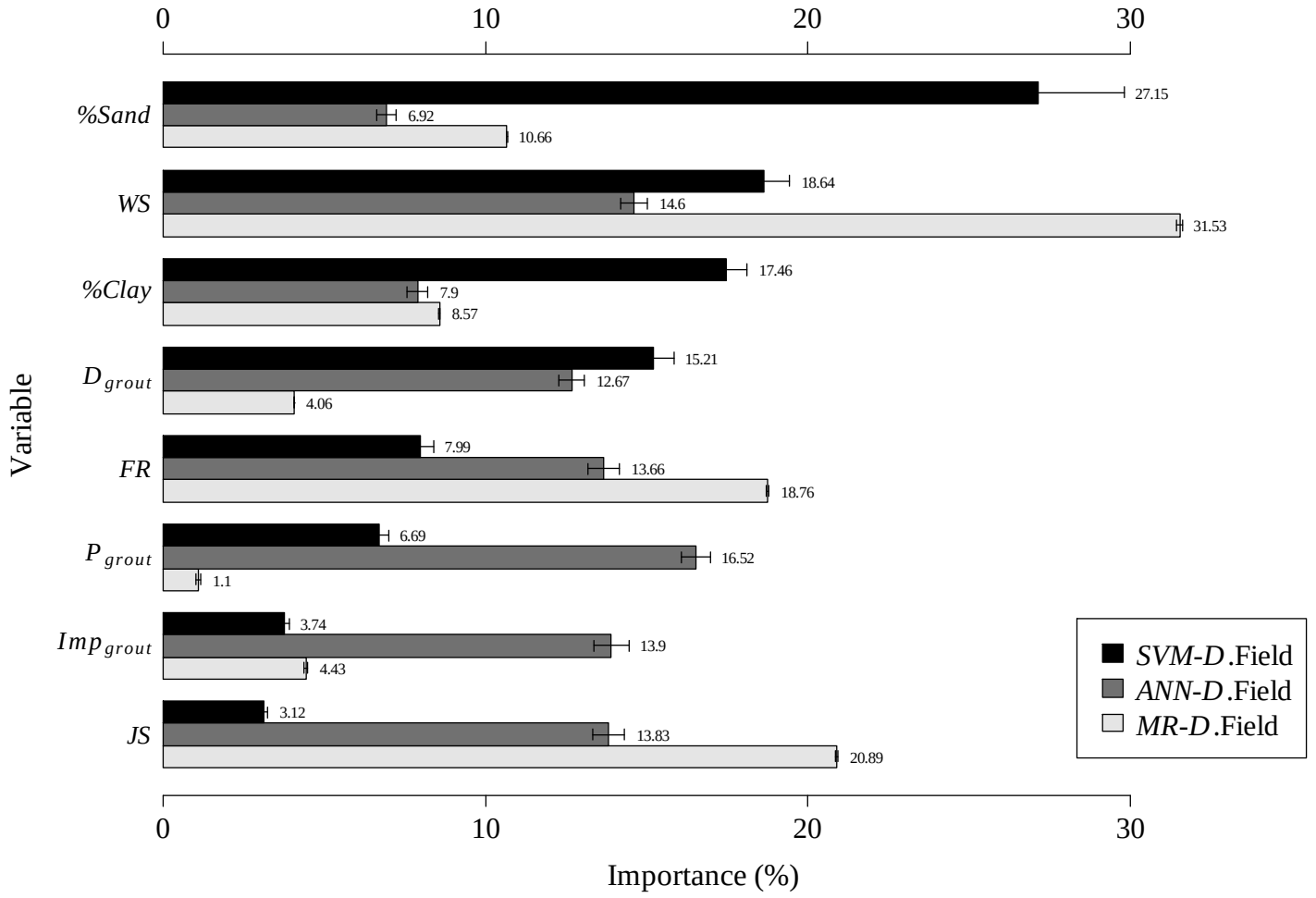


Figure 6.27: Relative importance of each input variable quantified by 1-D *SA*, comparing *MR-D.Field*, *ANN-D.Field* and *SVM-D.Field* models

On the other hand, *VEC* curves of *%Sand* and *%Clay* show that the *JG* columns with the highest diameters are built in sandy soils and that the smallest ones are built in clayed soils. Moreover, comparing these two *VEC* curves, it is pointed out that the decrease of the clay fraction of the soil has a higher impact in the column diameter than the increase of the sand fraction.

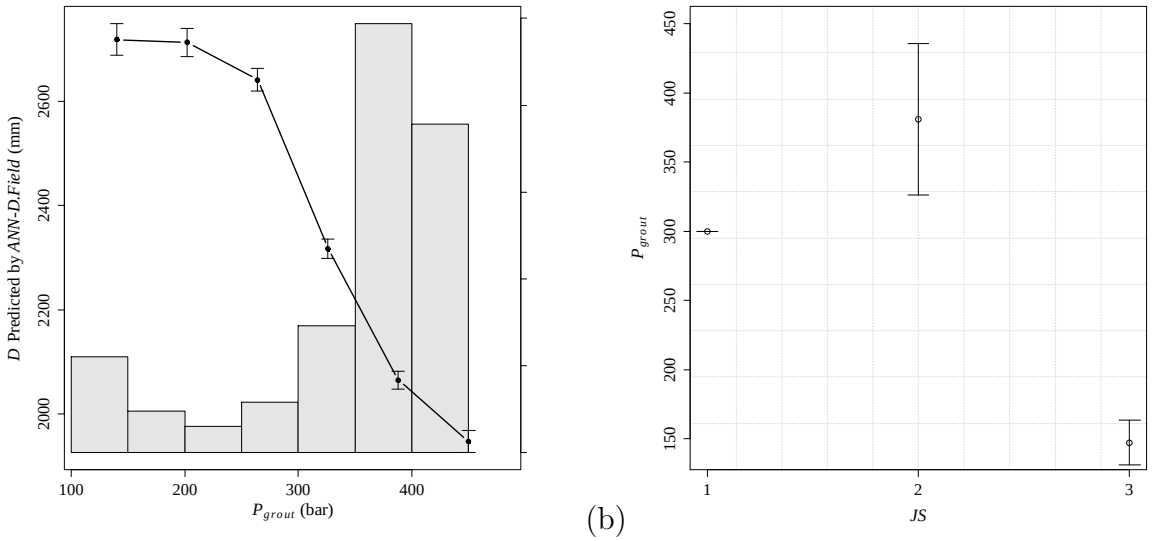


Figure 6.28: 1-D *SA* according to *ANN-D.Field* model: a) vertical averaging of P_{grout} *VEC* curve (points and whiskers) and histogram (in bars) and b) relationship between JS and P_{grout} variables

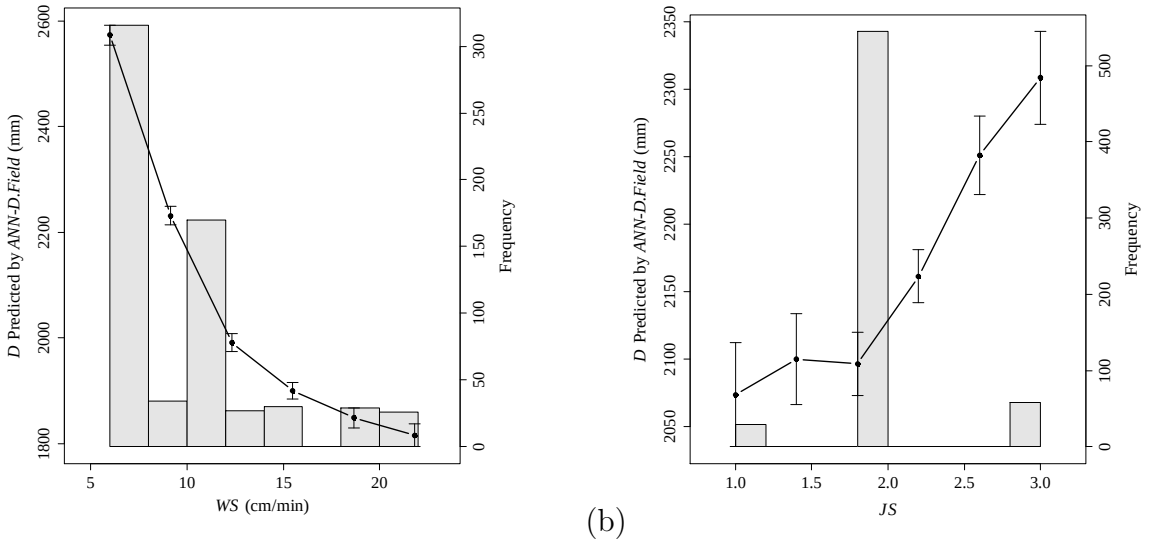


Figure 6.29: Vertical averaging of the *VEC* curves (points and whiskers) and histogram (in bars) according to *ANN-D.Field* model for: a) WS and b) JS variables in D prediction , quantified by 1-D *SA*

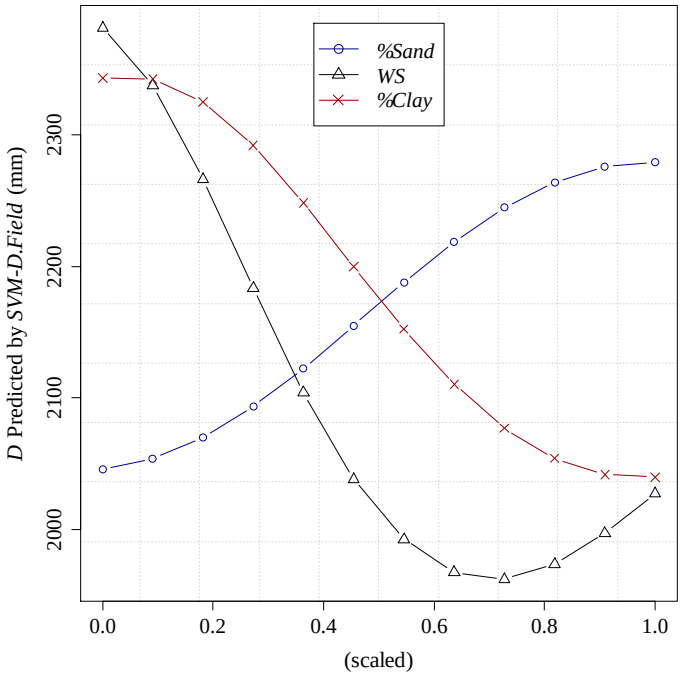


Figure 6.30: *VEC* curves for the three key input variables according to *SVM-D.Field* model in D prediction, quantified by 1-D *SA*

Based on a 2-D sensitivity analysis, it was measured the interaction level between all variables with $\%Clay$ (see Figure 6.31a) and plotted the effect in *JG* column diameter development when $\%clay$ and WS are changed simultaneously. Figure 6.31b shows that the effect of WS is more preponderant in soils with high clay content.

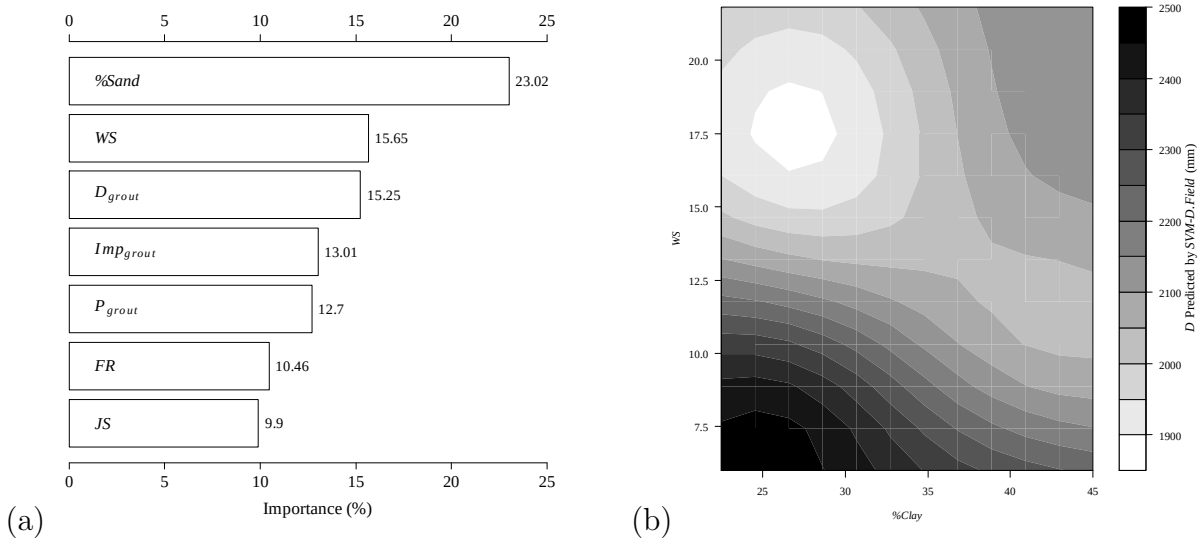


Figure 6.31: 2-D *SA* according to *SVM-D.Field* model in D prediction: a) interaction level between all variables with $\%Cay$ and b) *VEC* contour for $\%Clay$ and WS interaction

6.7 Proposal for jet grouting column diameter design

In the present chapter, some predictive models for *soilcrete* mechanical properties and *JG* column diameter were proposed. In the case of column diameter prediction, one of the most important tasks for *JG* quality control purposes, a high accuracy was achieved. However, due to the high mathematical complexity of the *DM* algorithms applied (e.g. *SVM* algorithm), such models are difficult to understand and implement for practical applications.

For models interpretability, a *GSA* was applied (see Section 6.6.2), where important observations were taken. To facilitate the implementation of the proposed models, namely during the project level, a graphical representation of the proposed model could be very useful. However, due to the high number of variables involved, such representation is complex, being necessary to apply some simplifications to make it possible.

Taking the *SVM-D.Field* model, which achieved a great performance in *JG* column diameter prediction as shown in Section 6.6, Figure 6.32 depicts the relationship between *JG* column diameter built using single fluid system and WS for different combination of the remain input variables, i.e. FR, P_{grout}, Imp_{grout} and D_{grout}, and according to the soil properties. The equivalent representation for double and triple fluid system are plot in Figures 6.33 and 6.34 respectively. For each one of the input variables, particularly WS, FR, D_{grout} and P_{grout}, it was considered the range currently used, as summarized in Table 3.3, but limited to the *SVM-D.Field* model applicability. In these three plots, the doted line represents the relation between WS and *JG* column diameter considering the mean value of each one of the remains input variables and the shaded area represents the envelop of the *JG* column diameter for different combinations of each one of the input variables.

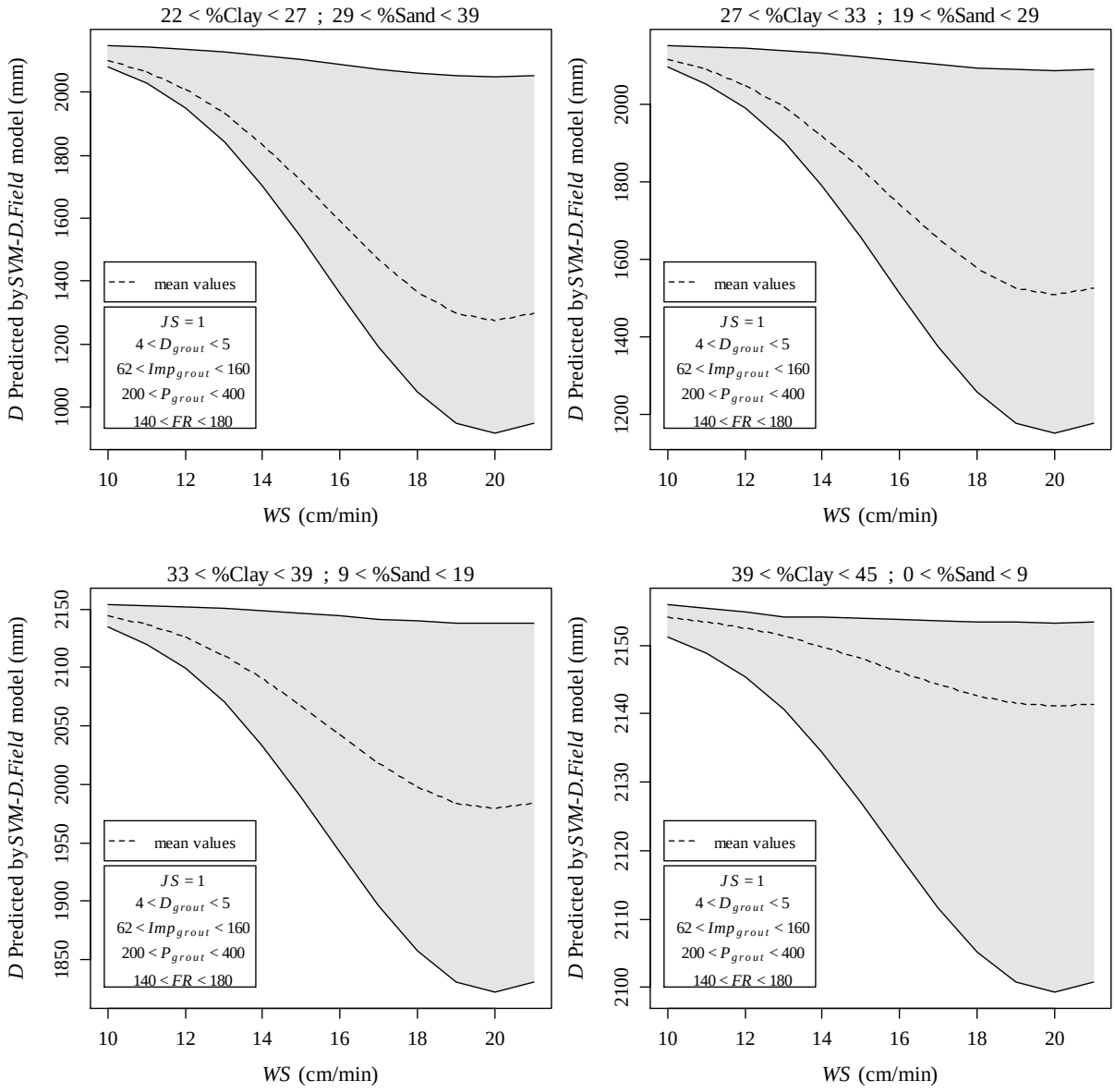


Figure 6.32: Abacus for D design of single fluid system and according to *SVM-D.Field* model

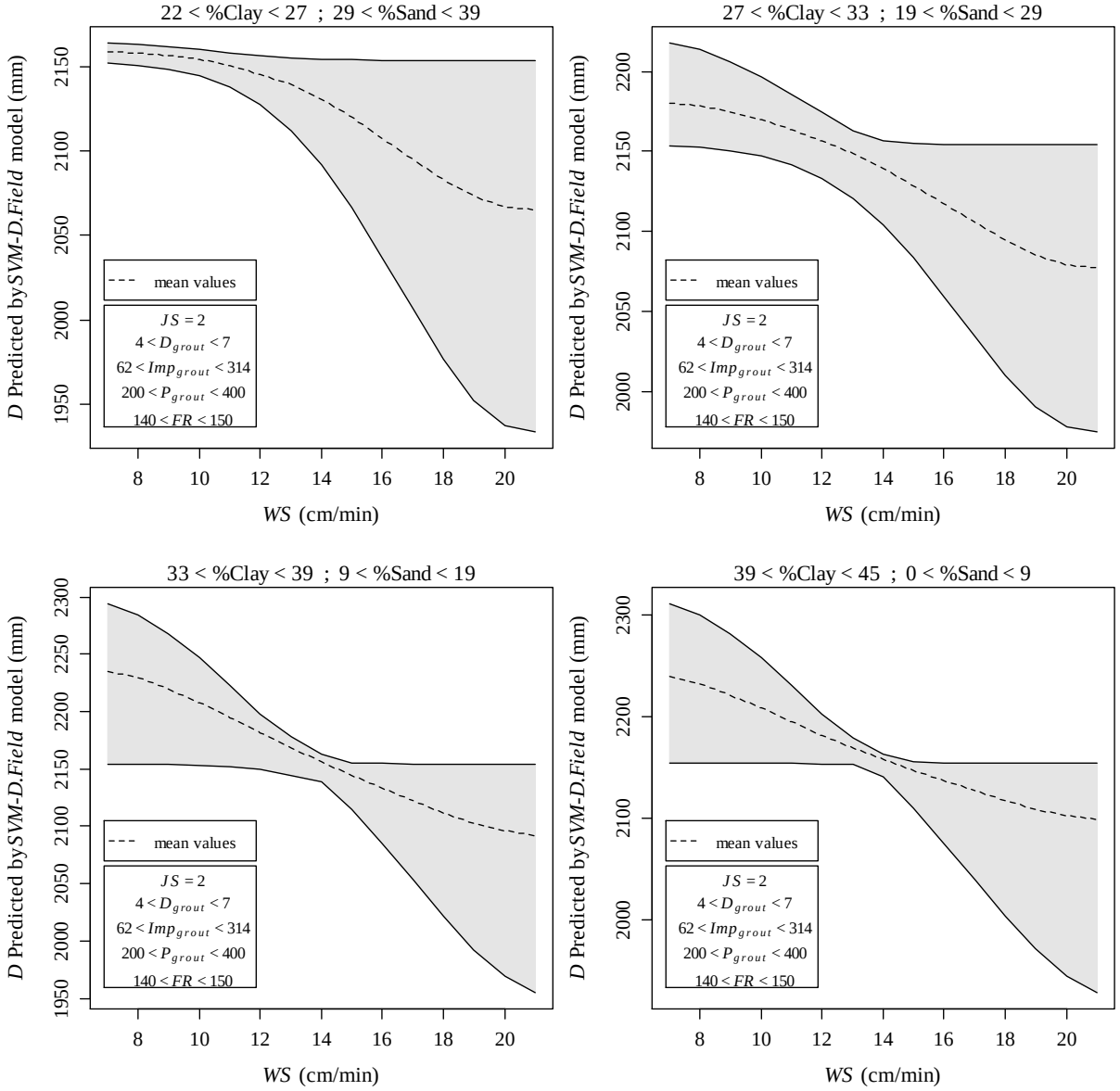


Figure 6.33: Abacus for D design of double fluid system and according to *SVM-D.Field* model

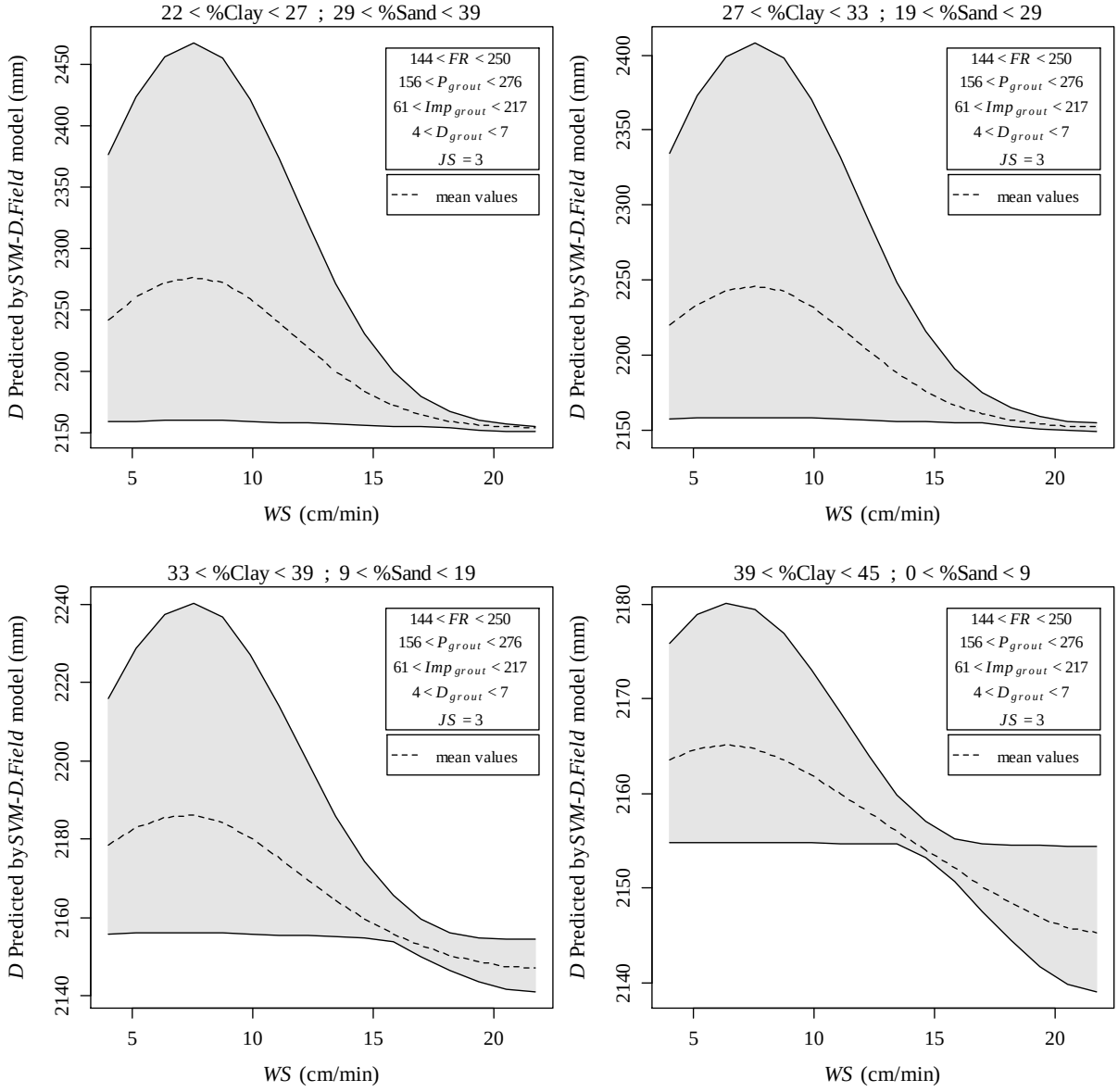


Figure 6.34: Abacus for D design of triple fluid system and according to *SVM-D.Field* model

6.8 Conclusions

Soft soil improvement using *JG* technology is currently applied in many geotechnical works. Quality assessment is usually taken from the *JG* column diameter, particularly of the test columns, and the *soilcrete* mechanical properties (strength and stiffness). Therefore, it is useful to have numerical approaches capable of accurately predicting each one of these elements. However, due to the high number of parameters involved during the soil improvement and the heterogeneity of the soils, such a task is highly complex. As a result, attempts to develop predictive models for *soilcrete* mechanical properties and column diameter of *JG* technology are scarce and have important applicability limitations (summarised in Chapter 3).

In this chapter, some analytical models were proposed for predicting UCS and E_0 of *soilcrete* mixtures and *JG* column diameter through the application of advanced statistical analysis, usually known as *DM* techniques. Although these techniques have good potential for learning complex mappings (as shown in Chapter 5), non-ideal predictive performances were achieved in the experiments conducted for the prediction of the mechanical properties of *soilcrete*. Nevertheless, the proposed models, particularly *ANN* and *SVM*, for UCS and E_0 prediction of *soilcrete* mixtures achieved a performance that can be acceptable for field mixtures study. In particular, it was observed that most of the predictions are above the experimental values, i.e., predictions have a positive safety factor. Moreover, after applying a *GSA* procedure over the most interesting data-driven models, important and useful observations were noted that help one understand *soilcrete* mechanical behaviour. For instance, the exponential effect of t in *soilcrete* strength and stiffness behaviour was observed, and ω and C_{iv} play an important role in E_0 prediction of *soilcrete* mixtures. Moreover, the relation $n/(C_{iv})^d$ and the jet system also exert an important influence on *soilcrete* strength prediction. In addition, it was shown that the behaviour of soilcrete mechanical properties cannot be accurately learned by linear approaches.

Despite its good performance in the *JGLF* mechanical properties study, the *EC2* approach for the prediction of *soilcrete* mechanical properties was fair. Indeed, even considering the material properties at 28 days of curing, the performance was poor, which could be attributed to other variables not contemplated in the model. For example, it was shown through the data-driven models that *soilcrete* porosity is relevant in the study of its mechanical properties.

The preliminary experiments performed to correlate the UCS of laboratory formulations and field mixtures showed that the mean value (by geotechnical work) of UCS of *soilcrete* mixtures at 28 days of curing is approximately 11% higher than the equivalent

laboratory formulation. However, this experiment needs to be further validated when new data are available because it was not satisfied for one of the five geotechnical works considered.

The prediction of *soilcrete* stiffness based on strength values achieved a slightly better performance in comparison to the model where the strength is not considered as an input variable. In this experiment an almost linear relationship between E_0 and UCS for a given sample was observed.

Regarding the predictive models for JG column diameter, an excellent performance was achieved, namely by the SVM and ANN algorithms. Moreover, a GSA was performed over the proposed models, confirming some of the well-known theoretical approaches formulated to describe the development of the JG column diameter, i.e., the effect of the soil resistance and jet action. The importance of the WS and the soil properties, namely its sand and clay content, in the development of the JG column diameter was underscored.

It should also be noted that the models explored in this thesis represent a starting point towards the development of new approaches for more accurate and applicable (in terms of soil types) JG technology design. Moreover, some interesting observations were stressed, contributing a better understanding of the mechanical behaviour of *soilcrete* mixtures as well as the JG column diameter development, which has the potential to improve the technical and economic efficiency of JG technology.

This page was intentionally left blank.

Chapter 7

Main summary

7.1 Synthesis and main conclusions

This thesis studied *Jet Grouting* (*JG*) technology from the point of view of the development of new approaches for its design, i.e., to predict *Uniaxial Compressive Strength* (*UCS*) and stiffness of *Jet Grouting Laboratory Formulations* (*JGLF*) and *soilcrete*[1] mixtures as well as *JG* column diameter. To do so, a literature review was performed to identify the existing approaches for *JG* design and the most relevant variables that can directly or indirectly interfere in the development of strength, stiffness and diameter of *JG* columns during the *JG* process. In addition, research on *Artificial Intelligence* (*AI*)/*Data Mining* (*DM*) tools was performed to develop a background to support the implementation of such tools in the development of the intended *JG* design methodologies.

The main achievements found in each chapter are summarised below.

In Chapter 2, the high learning capabilities and flexibility of *DM* techniques even when addressed to problems of high dimensionality/complexity were highlighted, and the importance of a structured database with sufficient data containing significant attributes for the discovery task was emphasised. Relating to this issue, the noise levels deserve particular attention because it is not current practice in *JG* projects to organise all information related with each *JG* project. Moreover, another important task during a *DM* problem is related to the selection of the best set of input variables, where the implementation of *Feature Selection* (*FS*) algorithms can provide a valuable contribution. Additionally, it was observed that the main drawback related to the application of *DM* techniques to solve complex problems is the model interpretability due to the high mathematical complexity of the algorithms implemented. To overcome this drawback, the application of a novel *Global Sensitivity Analysis* (*GSA*) over the trained models gives important help. With

[1] *Soilcrete* – practical designation for soil-cement mixture resulting from *JG* technology.

this analysis, it was possible to measure the relative importance of each input variable as well as its average effect on the target variable.

In Chapter 3, the importance of *JG* technology as a soft soil improvement method was discussed. This versatile technology can be applied for different purposes, such as groundwater control or support. Despite *JG* being widely used, the actual approaches for *JG* design (mechanical properties and column diameter) are scarce and have important applicability limitations. The main factors for this scenario are the high number of parameters involved and the heterogeneity of the soil. Therefore, demand for advanced tools able to develop design approaches for *JG* design is rising. The answer could lie in the large amount of data related to different *JG* projects that were collected and stored over the last few years. These data, containing information related to the soil properties, *JG* parameters, mechanical properties of the *soilcrete* mixture and *JG* column diameter, can now be analysed by powerful statistical analysis methods usually known as *DM* techniques. These tools are able to analyse complex data and extract useful patterns and trends that can be converted into knowledge/models for implementation in future projects.

A first step toward the development of new and more reliable approaches for *JG* technology design was addressed by *JGLF*. The main achievements of these formulations, normally prepared for large-scale *JG* projects, were based on the databases created in Chapter 4 and were presented and discussed in Chapter 5. The main innovative contributions are as follows:

- *DM* techniques, particularly *Support Vector Machine* (*SVM*) and *Artificial Neural Network* (*ANN*) algorithms, proved to be powerful tools for exploring *JGLF* mechanical properties. Indeed, these tools were able to learn with high accuracy the complex relationships between *JGLF* mechanical properties and their contributing factors. For both UCS and stiffness prediction of *JGLF*, *SVM* achieved a performance higher than 0.93, using R^2 as a performance indicator;

- Based on a *GSA*, it was shown that the relation between the mixture porosity and the volumetric content of cement $(n/(C_{iv})^d)$ is a key variable in both mechanical properties prediction of *JGLF*. Moreover, in the UCS study the age of the mixture (t) and C_{iv} (volumetric content of cement) should also be taken into account. Additionally, it was observed that the soil properties are slightly more relevant in stiffness prediction of *JGLF* than for strength;

- By measuring the average impact of t and C_{iv} in the mechanical properties of *JGLF*, a positive influence following an exponential law, with a concave shape in

the case of t and convex shape for C_{iv}, was observed. On the other hand, the relation $n/(C_{iv})^d$ and the $\%Clay$ (clay content of the soil) have a negative impact in both $JGLF$ mechanical properties development;

- The analytical expressions proposed by Eurocode 2 for predicting the strength and stiffness of concrete can be seen as an interesting alternative for $JGLF$ strength and stiffness predictions. However, because these approaches require information from 28-day laboratory tests, their application is particularly limited to validation purposes;

- An attempt to predict the *elastic Young's modulus* (E_0) based on UCS values was successfully performed through the SVM algorithm, where an almost linear relationship between E_0 and UCS of $JGLF$ was observed. Although there is a practical importance for such an approach (i.e., predict E_0 based on UCS values), this task can be accurately performed by an equivalent model (in terms of performance) using elementary variables as attributes, i.e., without considering UCS.

- The obtained results are a valuable contribution to geotechnical engineers, as the number of $JGLF$ can be reduced. Additionally, a better understanding of the behaviour of JG material based on few variables was achieved. As a result of this knowledge, the quality, speed and cost of JG technology can be improved by efficiently controlling some variables involved in JG technology to achieve the desired result. Furthermore, DM models can be easily updated when new data are available, expanding its applicability in terms of soil types and for a range of JG variables.

Concerning the study addressing JG mixtures collected directly from real JG columns, for which the main achievements were presented and discussed in Chapter 6, the main innovative contributions related with mechanical properties and column diameter prediction are as follows:

- When working with *soilcrete*, DM techniques experienced some difficulties learning the complex relationship between *soilcrete* mechanical properties and their contributing factors. However, particularly for the SVM algorithm, it is still possible to predict UCS and E_0 of *soilcrete* mixtures with considerable accuracy, from 9 to 181 days in advance and for single, double and triple fluid systems;

- Supported by a novel GSA, the development of *soilcrete* mechanical properties followed an exponential law based on the age of the mixture. For UCS prediction the relation $n/(C_{iv})^d$ and jet system (JS) also play an important role, and in stiffness

development a strong influence of the cement and water content of the mixture was observed;

- For a better understanding of the problem at hand, a detailed *Sensitivity Analysis* (*SA*) (e.g., 2-D or higher) was extremely useful. For instance, based on a 2-D *SA* it was shown that t and W/C have a strong interaction in UCS prediction and that the effect of t is more pronounced on *JG* columns built with a single fluid system;

- Although it has shown good performance in *JGLF* mechanical properties study, using *Eurocode 2* (*EC2*) for the prediction of *soilcrete* mechanical properties was fair. Indeed, even when considering the material properties at 28 days of curing, the achieved performance was poor, which could be attributed to other variables not included in the model. For example, it was shown through the data-driven models that the *soilcrete* porosity is relevant in the study of its mechanical properties;

- *Soilcrete* stiffness can be predicted with better accuracy when the UCS of the mixtures is available to use as an input variable in the model. In this circumstance an almost linear relationship between E_0 and UCS is observed;

- The mean value (by geotechnical work) of UCS of *soilcrete* mixtures at 28 days of curing is approximately 11% higher that the equivalent laboratory formulation. This tentative correlation between the UCS in laboratory formulations and *soilcrete* mixtures needs to be further validated when new data are available because this was not satisfied for one of the five geotechnical works considered;

- For *JG* column diameter prediction, two models with high accuracy were developed based on *ANN* and *SVM* algorithms. These models were able to assimilate both the jet action and the soil resistance in the development of *JG* column diameter.

As a final observation, the following conclusion should be stressed:

- *DM* tools were shown to be a powerful instrument for addressing complex geotechnical problems that involve a high number of variables, such as in *JG* technology. Particularly, the *ANN* and *SVM* algorithms were able to learn the complex phenomena involving soil-cement mixtures and are recommended to explore similar problems;

- In addition to the high learning capabilities showed by the applied *DM* tools, it should also be stressed that the proposed models can be further updated when new data are made available, improving its performance and applicability, namely in terms of soil types;

- Another important and useful methodology, representing a complement to *DM* tools, is the application of *GSA* over the trained models. These methodologies can provide a valuable contribution for the models' interpretability, promoting a better understanding of the problem;

- It should be noted that the proposed models are not intended to substitute for the actual approaches, but to complement it. Moreover, independent of the accuracy of the developed models, there will always be an associated error, which needs to be controlled by laboratory and/or field tests;

- Taking into account the achieved results, the proposed models, namely those obtained from the *SVM* algorithm, can be seen as a starting point to describe statistically the actual knowledge related to the behaviour of *JG* mixtures. Moreover, the proposed approaches can be used for either future *JG* project design or quality control purposes. Therefore, it is expected to improve *JG* technical and economic efficiency and to optimise both quality and costs of the soil improvement;

- It should be strongly stressed that all proposed models for strength, stiffness and diameter of *JG* columns, as well as all conclusions, are based on the databases used. This means that, for instance, because all data were collected from just one company, other important variables that were not considered because they are not usually used by the company (e.g., nozzle geometry) may exist. Moreover, the proposed models should only be applied in the same conditions for which they were developed.

7.2 Future Developments

The different models proposed in the present work for mechanical properties prediction of both *JG* laboratory and field mixtures, as well as for *JG* column diameter, gave an important contribution for a better understanding of *JG* technology. However, the applicability of such models in real *JG* project design was not assessed. Therefore, it will be very useful the development of an informatics application supported on the proposed models, allowing its easily implementation and, at the same time assess its practical application and real contribution for *JG* technology.

In the present research, the high learning capabilities of *DM* tools to address *JG* material, particularly to learn the *JG* column diameter, were proved. However, the data used to feed the *DM* algorithms, although consistent and collected from reliable sources, contained some missing data that forced some variables to be omitted as input attributes.

This lack can be seen as one of the causes of the lower performance of the proposed models for *soilcrete* mechanical properties prediction. Therefore, supported by the idea that *DM* tools are able to learn the complex relationships behind *soilcrete* mechanical properties and *JG* column diameter, it is proposed to spend some effort to determine additional variables that improve model performance. Moreover, it will be interesting to have data regarding the Xjet system and to compare their results with the prediction from the other techniques.

According to the literature review, the type of soil is one of the main parameters that influences both *soilcrete* mechanical properties and *JG* column diameter. However, a detailed characterisation of the soil conditions is an expensive task, and, as a result, it is minimised to the vital parameters only. On the other hand, important information related to the soil profile can be taken during the perforation phase in the *JG* technology. Accordingly, the development of an integrated approach able to contemplate the information collected during this *JG* phase and the high learning capabilities of *DM* techniques could represent an important advance for *JG* technology efficiency.

Bibliography

M. Abramento, A. Koshima, and A.C. Zirlis. Fundações: Teoria e trática – reforço do terreno. pages 641–656. São Paulo, 1998.

A. Alonso-Betanzos, E. Castillo, O. Fontenla-Romero, and N. Sánchez-Marono. Shear strength prediction using dimensional analysis and functional networks. In *Proceedings of European Symposium on Artificial Neural Networks (ESANN04)*, pages 251–256. Citeseer, 2004.

ASTM, 1985. *Standard practice for classification of soils for engineering purposes (unified classification system)*. American Society for Testing and Materials, ref. D2487-83, 1985.

M. Azenha, C. Ferreira, J. Silva, A. Gomes Correia, R. Aguilar, and L. Ramos. Continuous stiffness monitoring of cemented sand through resonant frequency. In *2011 GeoHunan International Conference - Emerging Technologies for Material, Design, Rehabilitation and Inspection of Roadway Pavements*, pages 174–183, Hunan, Chine, June 2011. ASCE.

A. Azevedo and M.F. Santos. KDD, SEMMA and CRISP-DM: a parallel overview. In *Proceedings of the IADIS European conf. data mining*, pages 182–185, Amsterdam, Netherlands, July 2008.

J. Bi and K.P. Bennett. Regression error characteristic curves. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 43–50, Washington, DC, USA, August 2003. AAAI Press.

M.L. Brown and J.F. Kros. Data mining and the impact of missing data. *Industrial Management and Data Systems*, 103(8):611–621, 2003. ISSN 0263-5577.

J. Bulkley, S. Gayle, B. Hicks, and R. Stephens. Adding the where to the who. In *Proceedings of the Twenty-Fourth Annual SAS - Users Group International conference*, Miami Beach, Florida, USA, April 1999.

M.F.W. Carletto. *Jet Grouting (Sistema Monofluido): Um Método Teórico Simplificado para a Previsão do Diâmetro das Colunas*. PhD thesis, Escola Politécnica da Universidade de São Paulo, São Paulo, Brasil, Agosto 2009.

J.R. Carreto. Jet grouting. uma técnica em desenvolvimento. In *VII Congresso Nacional de Geotecnia*, pages 1043–1054. VII Congresso Nacional de Geotecnia, 2000.

E. Castillo, A. Cobo, J. Gutirrez, and R. Pruneda. *Functional Networks with Applications: A Neural-based Paradigm*. Springer, 1998. ISBN 079238332X.

E. Castillo, J.M. Gutiérrez, A.S. Hadi, and B. Lacruz. Some applications of functional networks in statistics and engineering. *Technometrics*, 43(1):10–24, 2001. ISSN 0040-1706.

CEB-FIP. *Model Code 1990*. Comité Euro-International du Béton, Vienna, September 1991.

CEN. *Eurocode 2: Design of Concrete Structures - Part 1-1: General Rules and Rules for Buildings*. European Committee for Standardization, Brussels, 2004a.

CEN. *Eurocode 7: Geotechnical design - Part 1: General rules*. European Committee for Standardization, Brussels, 2004b.

P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. *CRISP-DM 1.0: Step-by-step Data Mining Guide*. CRISP-DM Consortium, 2000.

Y. Chen and I.G. Councill. An introduction to support vector machines: A review. *AI Magazine*, 24(2):105–107, 2003. ISSN 0738-4602.

V. Cherkassky and Y. Ma. Practical selection of svm parameters and noise estimation for svm regression. *Neural Networks*, 17(1):113–126, 2004. ISSN 0893-6080.

J. Chou, C. Chiu, M. Farfoura, and I. Al-Taharwa. Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques. *Journal of Computing in Civil Engineering*, 25(3):242–253, 2011.

D. Cook and D.F. Swayne. *Interactive and Dynamic Graphics for Data Analysis: with R and GGobi*. Springer, 2007. ISBN 978-0-387-71761-6. Web site: `http://www.R-project.org`.

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

P. Cortez. Data mining with neural networks and support vector machines using the r/rminer tool. In P. Perner, editor, *Advances in Data Mining: Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining*, pages 572–583, Berlin, Germany, July 2010. LNAI 6171, Springer.

P. Cortez and M. Embrechts. Opening black box data mining models using sensitivity analysis. In *Proceedings of the 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2011)*, pages 341–348, Paris, France, April 2011. IEEE.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4): 547–553, 2009.

P.A.R. Cortez. *Modelos Inspirados na Natureza para a Previsão de Séries Temporais*. PhD thesis, Universidade do Minho, Guimarães, Portugal, 2002.

S. Coulter and CD Martin. Single fluid jet-grout strength and deformation properties. *Tunnelling and Underground Space Technology*, 21(6):690–695, 2006. ISSN 0886-7798.

P. Croce and A. Flora. Jet-grouting effects on pyroclastic soils. *Revista Italiana di Geotecnica*, 2:5–14, 1998.

M. Dash and H. Liu. Feature selection for classification. *Intelligent data analysis*, 1(1-4): 131–156, 1997.

E.A. El-Sebakhy, K.A. Faisal, T. Helmy, F. Azzedin, and A. Al-Suhaim. Evaluation of breast cancer tumor classification with unconstrained functional networks classifier. In *the 4th ACS/IEEE International Conf. on Computer Systems and Applications*, pages 281–287, 2006.

W. Ertel. *Introduction to Artificial Intelligence*. Springer, 2009. ISBN 9780857292988.

Y. Erzin. Artificial neural networks approach for swell pressure versus soil suction behaviour. *Canadian Geotechnical Journal*, 44(10):1215–1223, 2007. ISSN 0008-3674.

R. Essler and H. Yoshida. chapter Jet grouting, pages 160–196. Taylor & Francis, second edition, 2004. ISBN 0203570855.

J. Falcão, A. Pinto, and F. Pinto. Case histories of ground improvement solutions using jet-grouting. Geotechnical news, Tecnasol FGE: Fundações e Geotecnia S.A, 2000.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–54, 1996a.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, 1996b. ISSN 0001-0782.

H. Frohlich and A. Zell. Efficient parameter selection for support vector machines in classification and regression via model-based global optimization. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'05)*, volume 3, pages 1431–1436, Montréal, Québec, Canada, July 2005. IEEE.

F. Gallavresi. Grouting Improvement of Foundation Soils. In *Proceedings of Grouting, Soil Improvement and Geosynthetics*, volume 1, pages 1–38, New York, USA, 1992. ASCE.

GAMS Development Corporation. Welcome to the gams home page, October 2012. URL `http://www.gams.com/default.htm`.

H.N. Gazaway and B.H. Jasperse. Jet grouting in contaminated soils. In *Grouting, Soil Improvement and Geosynthetics (GSP 30)*, pages 206–214, New Orleans, LA, USA, 1992. ASCE.

P. Gazzarrini, M. Kokan, and S. Jungaro. Case history of jet grouting in british columbia. underpinning of SN rail tunnel in north vancouver. Geotechnical news, The Grout Line, 2005.

P. Gazzarrini, M. Kokan, and S. Jungaro. Jet grouting case history - olynpic village-false creek vancouver, bc, canada. In *Proceedings of the 61st Canadian Geotechnical Conference and 9th Joint CGS/IAH-CNC Groundwater Conference*, Edmonton, Canada, September 2008.

S. Gilan, H. Bahrami Jovein, and A.A. Ramezanianpour. Hybrid support vector regression – particle swarm optimization for prediction of compressive strength and rcpt of concretes containing metakaolin. *Construction and Building Materials*, 34:321–329, 2012.

K.G. GmbH. The soilcrete - jet grouting process. Brochure 67-03 E, 2002.

A.T.C. Goh and S.H. Goh. Support vector machines: Their use in geotechnical engineering as illustrated using seismic liquefaction data. *Computers and Geotechnics*, 34(5):410–421, 2007. ISSN 0266-352X.

A. Gomes Correia. Evaluation of mechanical properties of unbound granular materials for pavements and rail tracks. In A. Gomes Correia and Loizos, editors, *Geotechnics in Pavement and Railway Design and Construction: Proceedings of the International Seminar on Geotechnics and Railway Design and Construction*, volume 1, pages 35–60, Athens, Greece, December 2004. MillPress.

A. Gomes Correia, T. Valente, J. Tinoco, J. Falção, J. Barata, D. Cebola, and S. Coelho. Evaluation of mechanical properties of jet grouting columns using different test methods. In M. Hamza et al. (Eds.), editor, *17th International Conference on Soil Mechanics and Geotechnical Engineering (17th ICSMGE)*, pages 2169–2171, Alexandria, Egypt, October 2009. IOS Press.

A. Gomes Correia, P. Cortez, and J. Tinoco. Application of data mining in transportation geotechnics. In *Proceedings of the International Symposium on Advances in Ground Technology and Geo-Information (IS-AGTG)*, pages 25–40, Singapore, December 2011. Research Publishing.

S.R. Gunn. Support vector machines for classification and regression. Technical report, University of Southampton, 1998.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003. ISSN 1532-4435.

L. Hamel. *Knowledge Discovery with Support Sector Machines*. Wiley-Interscience, 2009. ISBN 0470371927.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag New York, second edition, 2009.

S. Horpibulsuk, N. Miura, and TS Nagaraj. Assessment of strength development in cement-admixed high water content clays with abrams' law as a basis. *Geotechnique*, 53(4):439–444, 2003.

C.M. Huang, Y.J. Lee, D.K.J. Lin, and S.Y. Huang. Model selection for support vector machines via uniform design. *Computational Statistics & Data Analysis*, 52(1):335–346, 2007. ISSN 0167-9473.

ICOG, editor. *4th International Conference on Grouting and Deep Mixing*, New Orleans, LA, USA, February 2012. DFI, ASCE.

H. Kanematsu. High pressure jet grouting method. *Civil Construction*, 21(13), 1980.

J.L. Kaushinger, R. Hankour, and E.B. Perry. Methods to estimate composition of jet grout bodies. In *Proceedings of Grouting, Soil Improvement and Geosynthetics*, volume 1, pages 194–204, New York, USA, 1992. ASCE.

S. Kenig, A. Ben-David, M. Omer, and A. Sadeh. Control of properties in injection molding by neural networks. *Engineering Applications of Artificial Intelligence*, 14(6): 819–823, 2001. ISSN 0952-1976.

S. Lai and M. Serra. Concrete strength prediction by means of neural network. *Construction and Building Materials*, 11(2):93–98, 1997. ISSN 0950-0618.

W.K. Langbehn. The jet grouting method: Applications in slope stabilization and landslide repair. Master's thesis, Department of Civil Engineering, University of California, Berkeley, California, USA, 1986.

F.H. Lee, Y. Lee, S.H. Chew, and K.Y. Yong. Strength and modulus of marine clay-cement mixes. *Journal of Geotechnical and Geoenvironmental Engineering*, 131(2): 178–186, 2005.

C. Li, X. Liao, Z. Wu, and J. Yu. Complex functional networks. *Mathematics and computers in simulation*, 57(6):355–365, 2001.

S.H. Liao, P.H. Chu, and P.Y. Hsiao. Data mining techniques and applications. A decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12):11303–11311, 2012.

T. Limprasert. *Analysis and Assessment of Engineering Behavior of Cement Stabilized Clays.* PhD thesis, The Faculty of the Russ College of Engineering and Technology, Ohio University, Athens, Ohio, March 1995.

H. Liu, H. Motoda, R. Setiono, and Z. Zhao. Feature selection: An ever evolving frontier in data mining. In *Proceedings of the Fourth Workshop on Feature Selection in Data Mining*, volume 10, pages 4–13, Hyderabad, India, June 2010.

S. Y. Liu, D. W. Zhang, Z. B. Liu, and Y. F. Deng. Assessment of unconfined compressive strength of cement stabilized marine clay. *Marine Georesources and Geotechnology*, 26 (1):19–35, 2008.

G.A. Lorenzo and D.T. Bergado. Fundamental parameters of cement-admixed clay-new approach. *Journal of Geotechnical and Geoenvironmental Engineering*, 130(10):1042–1050, 2004.

T. Magalhães. Estudo do comportamento mecânico de um solo: Aplicação ao jet grouting. Relatório de projecto individual, Escola de Engenharia, Universidade do Minho, Guimarães, Portugal, 2006.

G. Miki and W. Nakanishi. Technical progress of the jet grouting method and its newest type. In *Proceedings of In situ Soil and Rock Reinformcement International Conference*, pages 195–200, Paris, France, October 1984.

M. Minsky and S. Papert. *Perceptrons*. MIT-Press, Cambridge, 1969.

T. Miranda, A.G. Correia, M. Santos, L.R. Sousa, and P. Cortez. New models for strength and deformability parameter calculation in rock masses using data-mining techniques. *International Journal of Geomechanics*, 11:44–58, 2011.

T.F.S. Miranda. *Geomechanical Parameters Evaluation in Underground Structures. Artificial Intelligence, Bayesian Probabilities and Inverse Methods*. PhD thesis, School of Engineering, University of Minho, Guimarães, Portugal, November 2007.

J.K. Mitchell, T.S. Veng, and C.L. Monismith. Behavior of stabilized soils under repeated loading. Technical report, Department of Civil Engineering, University of California, 1974.

N. Miura, S. Horpibussuk, and T.S. Nagaraj. Engineering behavior of cement stabilized clay at high water content. *Soils Found*, 41(5):33–45, 2001.

G. Modoni, P. Croce, and L. Mongiovi. Theoretical modelling of jet grouting. *Geotechnique*, 56(5):335–347, 2006. ISSN 0016-8505.

M.P. Moseley and K. Kirsch. *Ground Improvement*. Taylor & Francis, second edition, 2004. ISBN 0203570855.

R.A. Muenchen and J.M. Hilbe. *R for Stata users*. Springer Verlag, 2010. ISBN 9781441913173.

T.S. Nagaraj and N. Miura. Induced cementation of soft ground – a parametric assessment. In *Proceedings of the International Symposium on Lowland Technology*, pages 85–97, Saga, Japan, 1996.

T.S. Nagaraj, N. Miura, P.P. Yaligar, and A. Yamadera. Predicting strength development by cement admixture based on water content. In *Grouting and Deep Mixing: 2nd International Conference on Ground Improvement Geosystems (IS Tokyo'96)*, pages 431–436, Tokyo, Japan, 1996.

B.S. Narendra, P.V. Sivapullaiah, S. Suresh, and S.N. Omkar. Prediction of unconfined compressive strength of soft grounds using computational intelligence techniques: A comparative study. *Computers and Geotechnics*, 33(3):196–208, 2006. ISSN 0266-352X.

B. Nikbakhtan and K. Ahangari. Field study of the influence of various jet-grouting parameters on soilcrete unconfined compressive strength and its diameter. *International Journal of Rock Mechanics and Mining Sciences*, 47:685–689, 2010. ISSN 1365-1609.

B. Nikbakhtan and M. Osanloo. Effect of grout pressure and grout flow on soil physical and mechanical properties in jet grouting operations. *International Journal of Rock Mechanics and Mining Sciences*, 46(3):498–505, 2009. ISSN 1365-1609.

B. Nikbakhtan, K. Ahangari, and N. Rahmani. Estimation of jet grouting parameters in shahriar dam, iran. *Mining Science and Technology*, 20(3):472–477, 2010. ISSN 1674-5264.

NOVATECNICA. *Nova Técnica Consolidações e Construções S.A.* São Paulo, 2003. Catálogo Técnico.

A.B. Padura, J.B. Sevilla, J.G. Navarro, E.Y. Bustamante, and E.P. Crego. Study of the soil consolidation using reinforced jet grouting by geophysical and geotechnical techniques. *Construction and Building Materials*, 23(3):1389–1400, 2009. ISSN 0950-0618.

BK Prasad, H. Eskandari, and BV Reddy. Prediction of compressive strength of SCC and HPC with high volume fly ash using ann. *Construction and Building Materials*, 23(1): 117–128, 2009. ISSN 0950-0618.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Autria, 2009. ISBN 3-900051-00-3. Web site: `http://www.r-project.org/`.

E. Rich. *Artificial Intelligence*. Macgraw-hill, 1983. ISBN 9781441913173.

D.E. Rumelhart, G.E. Hinton, and R.J. Williams. *Learning internal representations by error propagation*. MIT-Press, Cambridge, 1986.

S.L. Shen, Y.S. Xu, J. Han, and J.M. Zhang. A ten-year review on the development of soil mixing technologies in china, 2010.

M. Shibazaki and H. Yoshida. Constructing bottom barriers with jet grouting. In *International Containment Techology Conference and Exhibition*, 1997.

M. Shibazaki, H. Yoshida, and Y. Matsumoto. Development of a soil improvement method utilizing cross jet. In *Grouting and Deep Mixing*, pages 707–710. Balkema, 1996.

A.J. Smola. Regression estimation with support vector learning machines. Master's thesis, Technische Universit at Munchen, Munchen, Germany, December 1996.

A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004. ISSN 0960-3174.

M. Swell. Feature Selection. 2007.

T.S. Tan, T.L. Goh, and K.Y. Yong. Properties of singapore marine clays improved by cement mixing. *Geotechnical Testing Journal*, 25(4):422–433, 2002.

M. Terashi and I. Juran. Ground improvement – state of the art. In GeoEng 2000, editor, *An International Conference on Geotechnical and Geological Engineering*, pages 19–24, Melbourne, Australia, November 2000. Technomic Publishing Company.

J. Tinoco, A. Gomes Correia, and P. Cortez. A data mining approach for jet grouting uniaxial compressive strength prediction. In *World Congress on Nature and Biologically Inspired Computing (NaBIC 2009)*, pages 553–558, Coimbatore, India, December 2009. IEEE.

J. Tinoco, A. Gomes Correia, and P. Cortez. Aplicação de técnicas de data mining na previsão do comportamento mecânico de colunas de jet grouting. In A. Gomes Correia et al. (Eds.), editor, *Actas do 12 Congresso Nacional de Geotecnia (12CNG)- Geotecnia e Desenvolvimento Sustentável*, pages 2167–2176, Guimarães, Portugal, Abril 2010a. SPG & UM, SPG & UM.

J. Tinoco, A. Gomes Correia, and P. Cortez. Application of data mining techniques to estimate elastic young modulus over time of jet grouting laboratory formulations. In *1st International Conference on Information Technology in Geo-Engineering (ICITG-Shanghai 2010)*, pages 92–100, Shanghai, Chine, September 2010b. IOS Press.

J. Tinoco, A. Gomes Correia, and P. Cortez. Application of data mining techniques in the estimation of mechanical properties of jet grouting laboratory formulations over time. *Soft Computing in Industrial Applications*, 96/2011:283–292, 2011a. ISSN 1867-5662.

J. Tinoco, A. Gomes Correia, and P. Cortez. Application of data mining techniques in the estimation of the uniaxial compressive strength of jet grouting columns over time. *Construction and Building Materials*, 25(3):1257–1262, March 2011b.

J. Tinoco, A. Gomes Correia, and P. Cortez. A data mining approach for predicting jet grouting geomechanical parameters. In *GeoHunan 2011: Rood Material and New Innovations in Pavement Engineering, Geotechnical Special Publication No. 23 (GSP 23)*, pages 97–104, Hunan, Chine, June 2011c. ASCE.

J. Tinoco, A. Gomes Correia, and P. Cortez. Support vector machines in mechanical properties prediction of jet grouting columns. In Escola de Engenharia, editor, *Atas da Semana da Escola de Engenharia 2011 - Reinventar o Futuro (SEEUM2011)*, pages 1–10, Guimarães, Portugal, October 2011d. UM, Escola de Engenharia•UM. ISBN 978-972-8692-61-2.

J. Tinoco, A. Gomes Correia, and P. Cortez. Uniaxial compressive strength prediction of jet grouting columns using support vector machines. In P. Novais et al., editor, *Proceedings of the European Simulation and Modeling Conference - ESM'2011*, pages 326–330, Guimarães, Portugal, October 2011e. *eti*, EUROSIS.

J. Tinoco, A. Gomes Correia, and P. Cortez. Using data mining techniques to predict deformability properties of jet grouting laboratory formulations over time. *Progress in Artificial Intelligence*, pages 491–505, 2011f.

J. Tinoco, A. Gomes Correia, and P. Cortez. Jet grouting mechanicals properties prediction using data mining techniques. In *Proceedings of the 4th International Conference on Grouting and Deep Mixing*, pages 2082–2091, New Orleans, Louisiana, USA, February 2012a. ASCE.

J. Tinoco, A. Gomes Correia, and P. Cortez. Application of a sensitivity analysis procedure to interpret uniaxial compressive strength prediction of jet grouting laboratory formulations performed by svm model. In Nicolas Denies & Noel Huybrechts, editor, *Proceedings of the International Symposium and Short Courses - IS-GI Brussels 2012*, pages 317–326, Brussels, Belgium, May 2012b.

J. Tinoco, A. Gomes Correia, and P. Cortez. Jet grouting deformability modulus prediction using data mining tools. In Miura et al. (eds), editor, *Proceedings of the 2nd International Conference on Transportation Geotechnics*, pages 168–173, Hokkaido, Japan, September 2012c. Taylor & Francis Group.

J. Tinoco, A. Gomes Correia, and P. Cortez. Previsão do comportamento mecânico de formulações laboratoriais de solo-cimento para colunas de jet grouting com recurso a máquina de vetores de suporte. *Civil Engineering Journal*, 42:43–55, January 2012d.

J. Tinoco, A. Gomes Correia, and P. Cortez. Utilização de máquina de vetores de suporte na previsão do comportamento mecânico de formulações laboratoriais de solo-cimento para colunas de jet grouting. In *Actas do 13 Congresso Nacional de Geotecnia (13CNG) - Pensar e Construir com a Natureza. Uma Visão para a Engenharia*, pages 335–336, Lisboa, Portugal, Abril 2012e. SPG & IST.

W.P. Van Impe, R.D.. Verástegui Flores, P. Mengé, and M. Van den Broeck. Considerations on laboratory test results of cement stabilised sludge. In *Deep Mixing '05: 1st International Conference on Deep Mixing - Best Practice and Recent Advances*, pages 163–168, 2005.

V. Vapnik. *Statistical Learning Theory*, volume 1851. John Wiley & Sons, Inc., 1998. ISBN 0471030031.

V. Vapnik, S.E. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems 9*, volume 9, pages 281–287. Bradford, 1997.

JG Wang, B Oh, SW Lim, and GS Kumar. Studies on soil disturbance caused by grouting in treating marine clay. In Ci-Premier Pte Limited, editor, *2nd Int. Conf. On Ground Improvement Techniques*, pages 521–528, Singapore, October 1998.

JG Wang, B.O.S.W. Lim, and GS Kumar. Effect of different jet grouting installations on neighboring structures. In *Field Measurements in Geomechanics*, pages 511–516. Balkema, 1999.

Z.F. Wang, S.L. Shen, and J. Yang. Estimation of the diameter of jet-grouted column based on turbulent kinematic flow theory. In *4th International Conference on Grouting and Deep Mixing (GROUT 2012)*, New Orleans, Louisiana, USA, February 2012. ASCE.

J.P. Welsh and G.K. Burke. Jet grouting: Uses for soil improvement. *Geotechnical Engineering Congress (GSP N27)*, pages 334–345, 1991.

J.P. Welsh and G.K. Burke. Advances in grouting technology, 1997.

I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, second edition, 2005.

P.P. Xanthakos, L.W. Abramson, and D.A. Bruce. Ground control and improvement. Wiley-Interscience, 1994. ISBN 0471552313.

A. Yamadera, T.S. Nagaraj, and N. Miura. Prediction of strength development in cement stabilized marine clay. In *Short Course on Improvement of Soil Ground, Analysis and Current Research*, pages 141–153, Bangkok, Thailand, 1997.

Y.Q. Zhou, D.X. He, and Z. Nong. Application of functional network to solving classification problems. In *Proceedings of the World Academy of Science, Engineering and Technology*, volume 7, pages 390–393. Citeseer, 2005.

Appendix A

Histograms and main statistics of the numerical variables used in the DM process

A.1 Jet grouting laboratory formulations data

A.1.1 Main statistics and histograms for *UCS* study

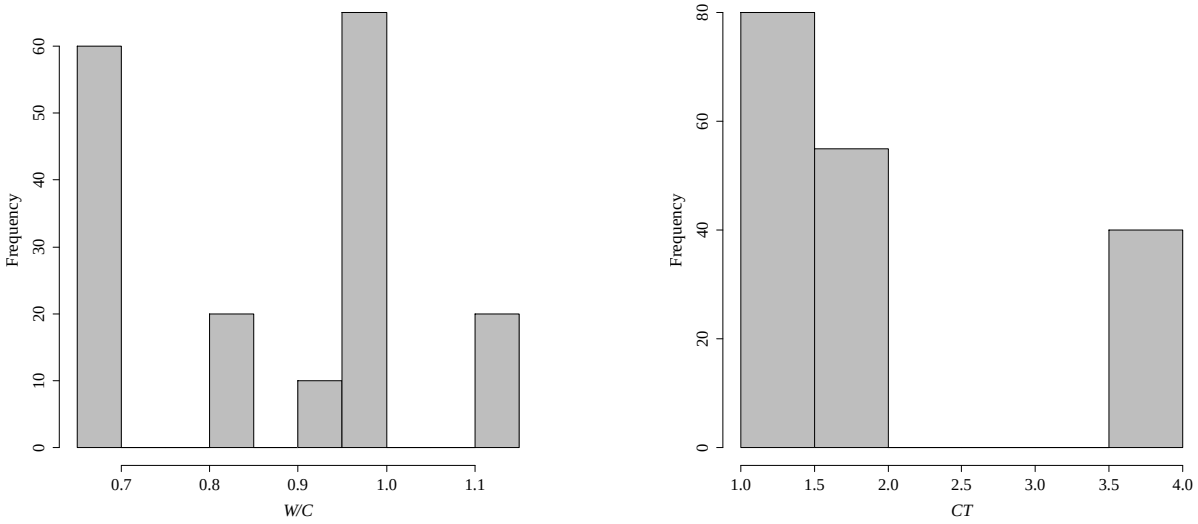
Table A.1: Summary of the input and output variables of database used in *UCS* study of *JGLF*

Variable	Minimum	Maximum	Mean	Standard Deviation
W/C	0.68	1.12	0.88	0.16
CT	1.00	4.00	2	1.17
SCC	32.50	42.50	40.21	4.21
s	0.20	0.25	0.21	0.02
kg/m^3	500.00	1806.00	1010.86	402.81
t (days)	3.00	56.00	21.6	19.24
ρ $(\mathrm{kg} \cdot \mathrm{m}^{-3})$	1484.11	1916.16	1689.77	118.73
ω (%)	28.00	87.00	52.77	16.83
ρ_d $(\mathrm{kg} \cdot \mathrm{m}^{-3})$	807.28	1497.00	1127.44	197.34
$1/\rho_d$ $(\mathrm{m}^3 \cdot \mathrm{kg}^{-1})$	$6.68\mathrm{E}^{-4}$	$1.23\mathrm{E}^{-3}$	$9.16\mathrm{E}^{-4}$	$1.67\mathrm{E}^{-4}$
$\%Soil$	26.02	75.81	52.56	15.21
$\%Cement$	24.19	73.98	47.44	15.21
$\gamma_{s.mixt}$ $(\mathrm{kg} \cdot \mathrm{m}^{-3})$	2758.86	2982.91	2863.49	68.43
e	0.87	2.57	1.63	0.52
n	46.53	71.96	60.50	7.59

Continued on next page

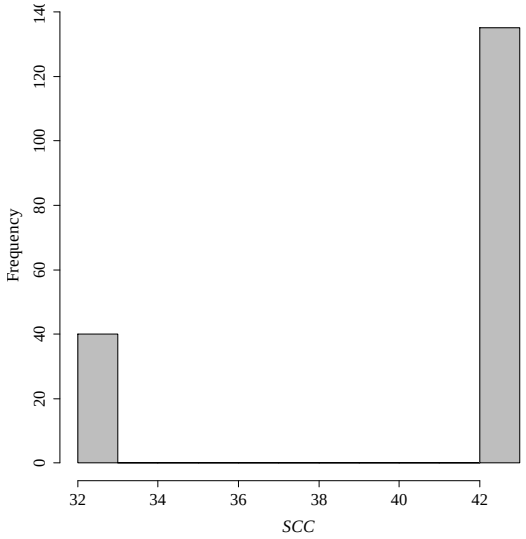
Table A.1 – continued from previous page

Variable	Minimum	Maximum	Mean	Standard Deviation
$1/n$	0.01	0.02	0.02	0.00
ω_{sat} (%)	31.08	88.89	56.62	17.17
S_ω	0.88	0.98	0.93	0.02
C_{iv}	0.21	0.71	0.44	0.15
$n/(C_{iv})^d$	48.83	74.26	62.59	7.26
$\%Sand$	0.00	39.00	13.57	11.54
$\%Silt$	33.00	57.00	50.49	5.49
$\%Clay$	22.50	45.00	35.89	7.74
$\%OM$	0.40	8.30	2.71	1.81
UCS (MPa)	0.76	13.19	5.20	2.73

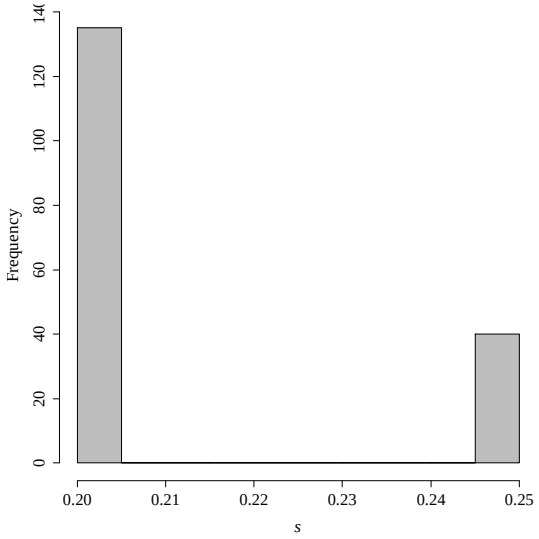


(1) Histogram of W/C (2) Histogram of CT

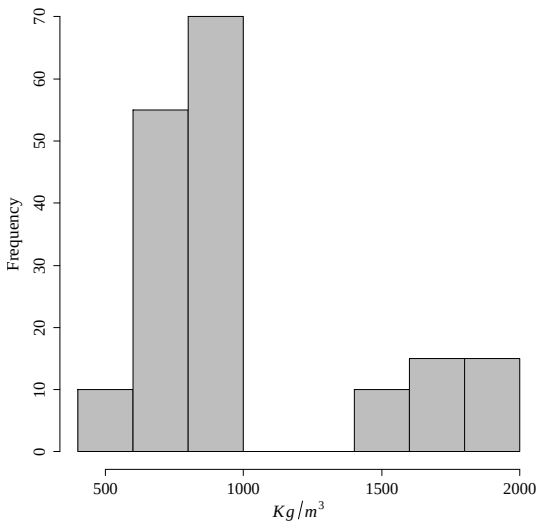
Figure A.1: Histograms of the numeric variables used in UCS study of $JGLF$



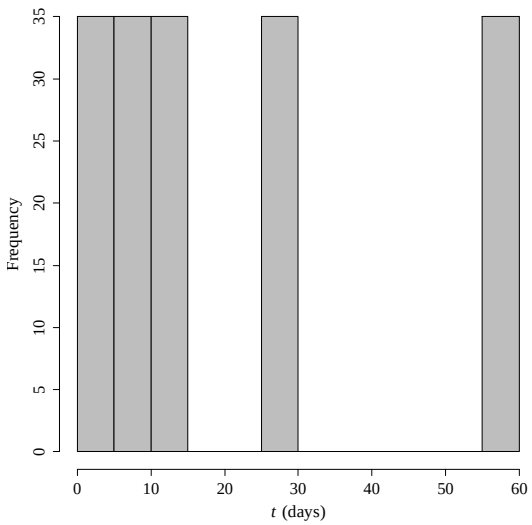
(3) Histogram of SCC



(4) Histogram of s

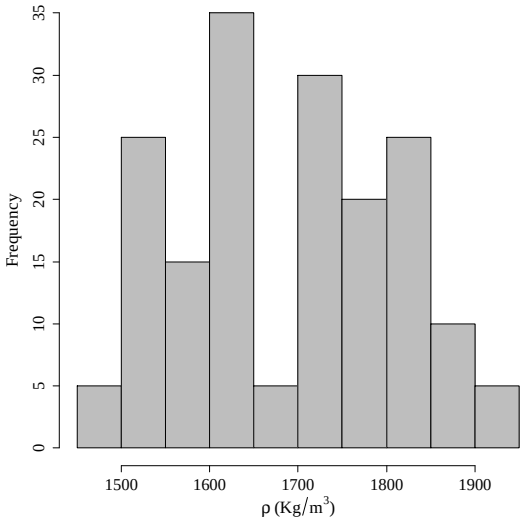


(5) Histogram of kg/m^3

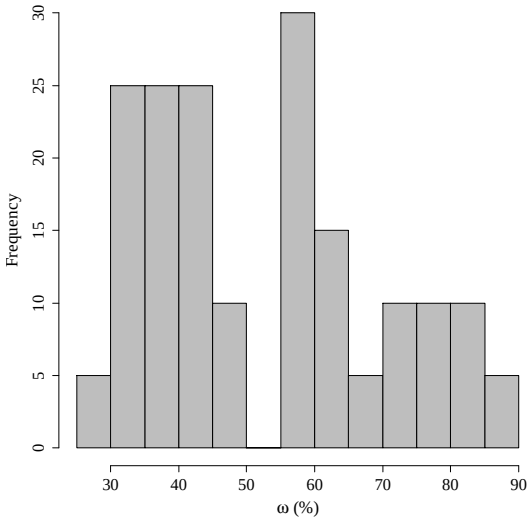


(6) Histogram of t

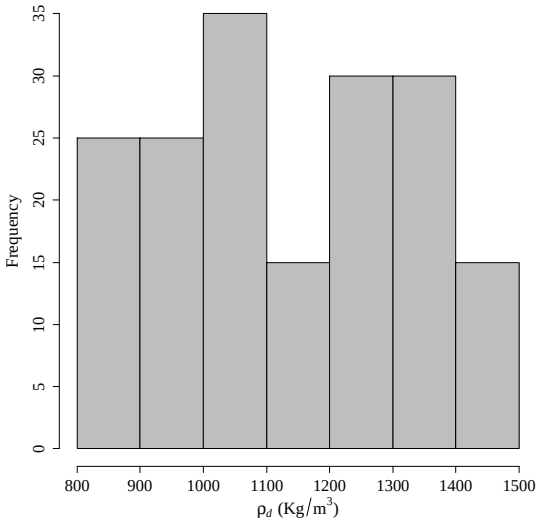
Figure A.1: Histograms of the numeric variables used in UCS study of $JGLF$ (cont'd)



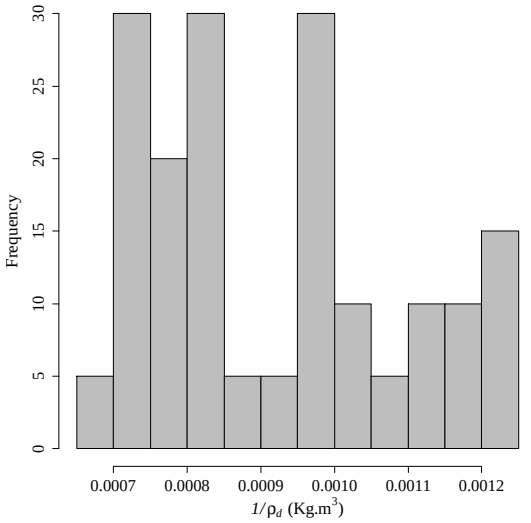
(7) Histogram of ρ



(8) Histogram of ω

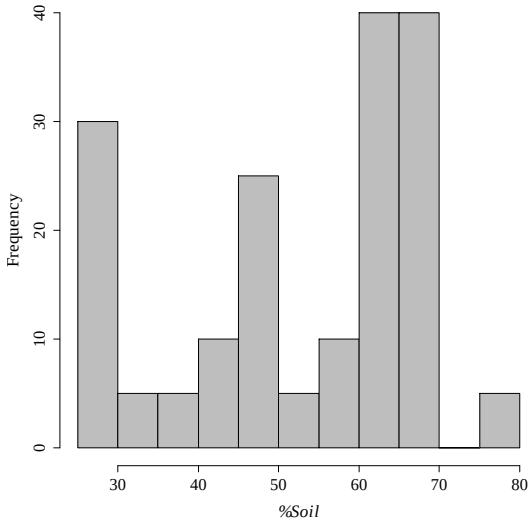


(9) Histogram of ρ_d

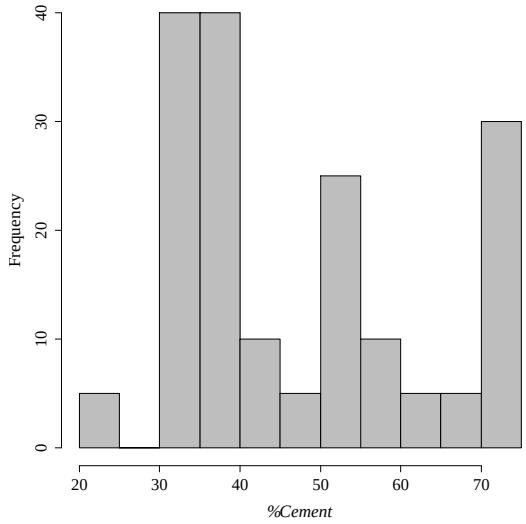


(10) Histogram of $1/\rho_d$

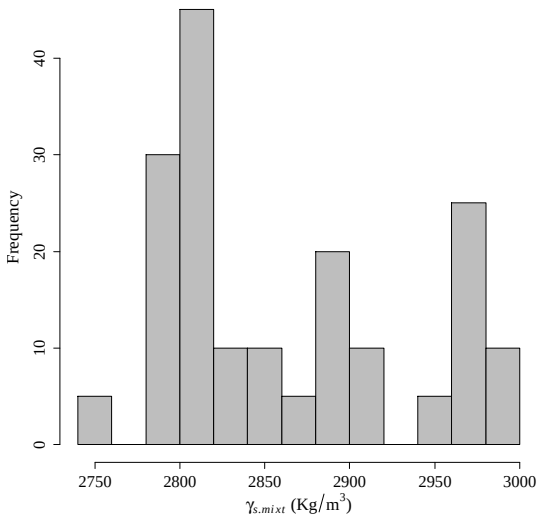
Figure A.1: Histograms of the numeric variables used in UCS study of $JGLF$ (cont'd)



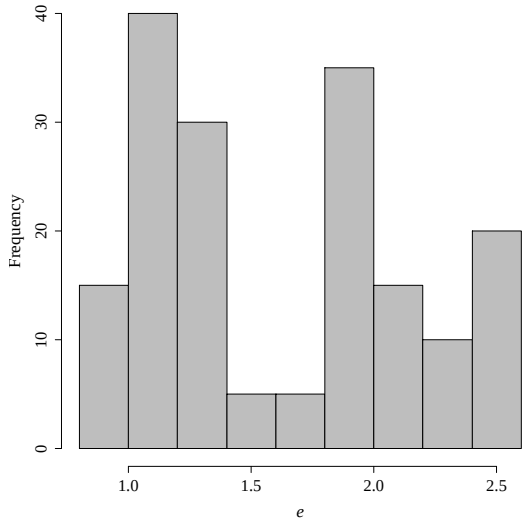
(11) Histogram of $\%Soil$



(12) Histogram of $\%Cement$

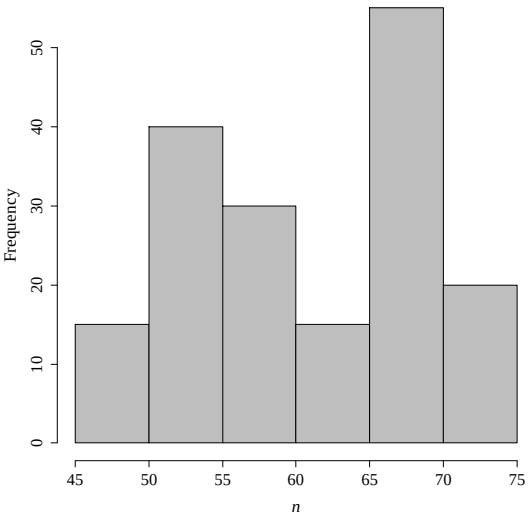


(13) Histogram of $\gamma_{s.mixt}$

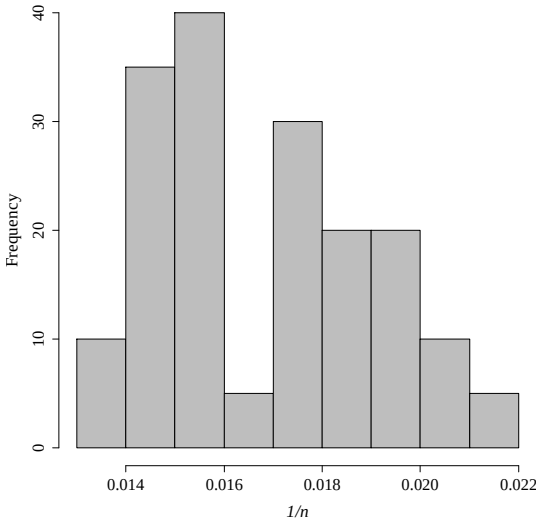


(14) Histogram of e

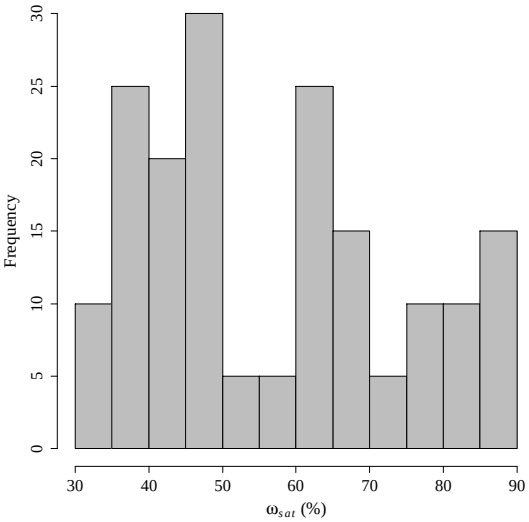
Figure A.1: Histograms of the numeric variables used in UCS study of $JGLF$ (cont'd)



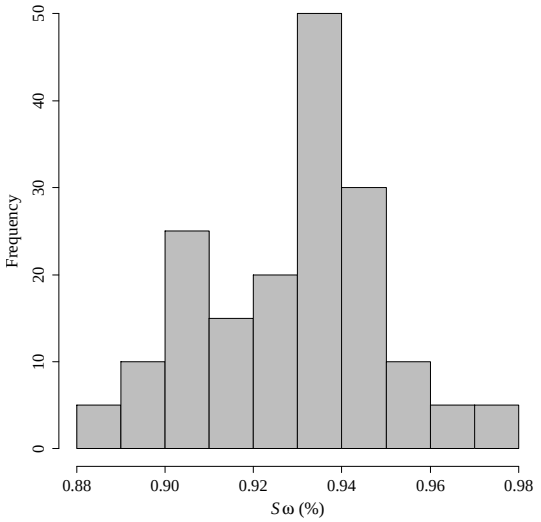
(15) Histogram of n



(16) Histogram of $1/n$

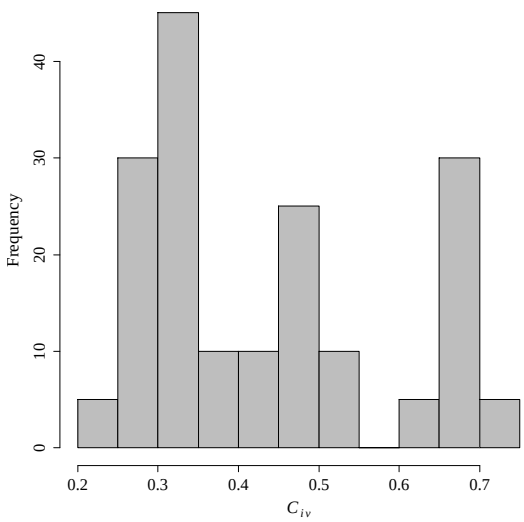


(17) Histogram of ω_{sat}

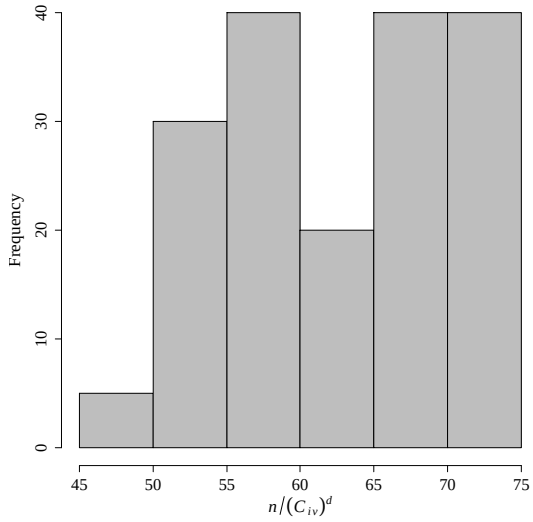


(18) Histogram of S_ω

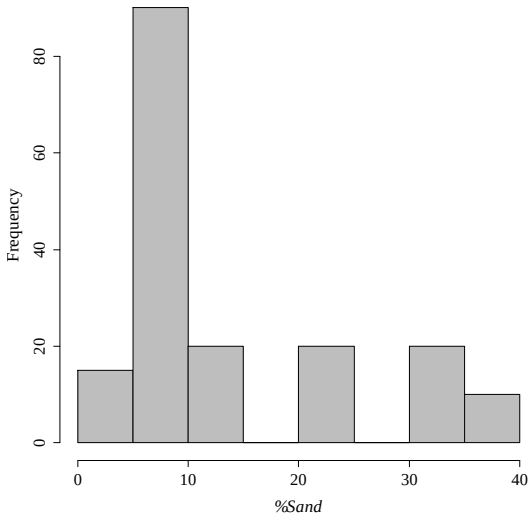
Figure A.1: Histograms of the numeric variables used in UCS study of $JGLF$ (cont'd)



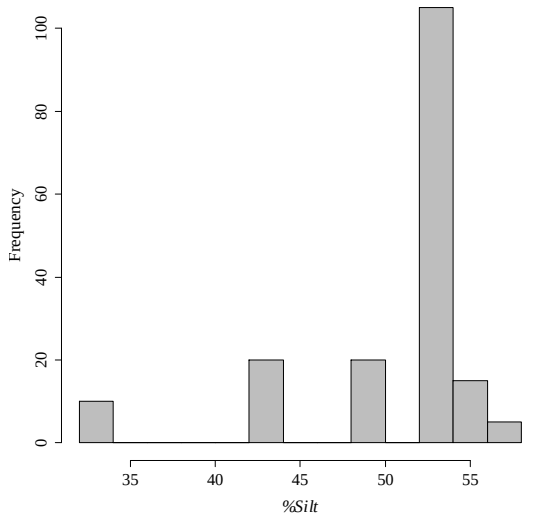
(19) Histogram of C_{iv}



(20) Histogram of $n/(C_{iv})^d$

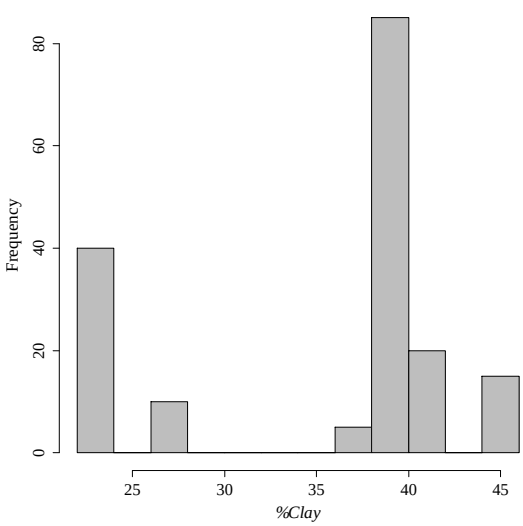


(21) Histogram of $\%Sand$

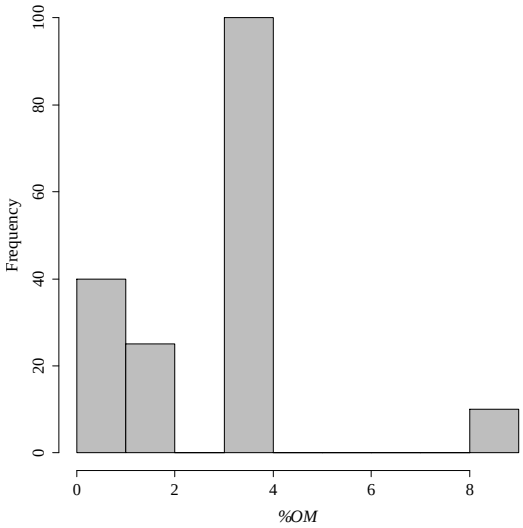


(22) Histogram of $\%Silt$

Figure A.1: Histograms of the numeric variables used in UCS study of $JGLF$ (cont'd)



(23) Histogram of $\%Clay$



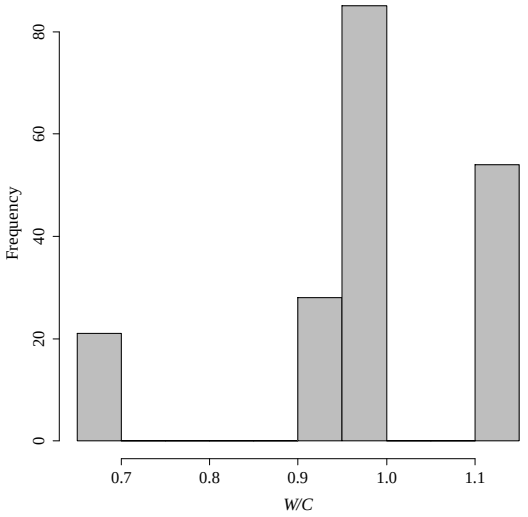
(24) Histogram of $\%OM$

Figure A.1: Histograms of the numeric variables used in UCS study of $JGLF$ (cont'd)

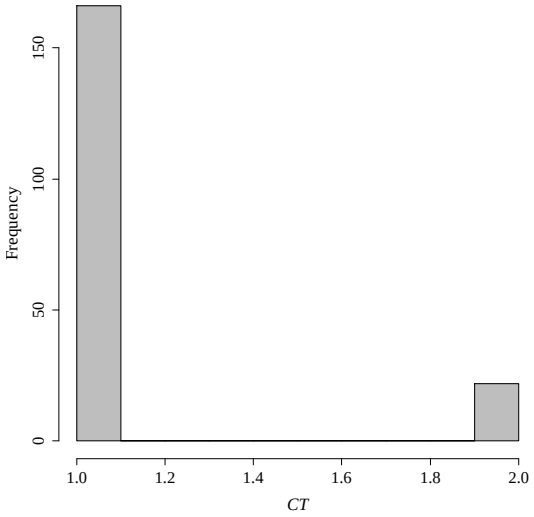
A.1.2 Main statistics and histograms for E_0 study

Table A.2: Summary of the input and output variables of database used in E_0 study of *JGLF*

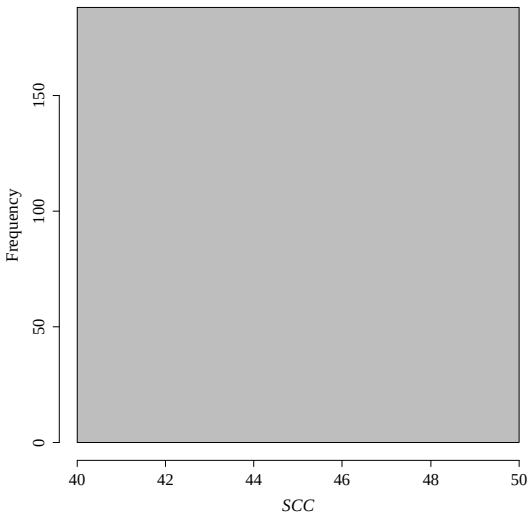
Variable	Minimum	Maximum	Mean	Standard Deviation
W/C	0.69	1.11	0.98	0.12
CT	1.00	2.00	1.12	0.32
SCC	42.50	42.50	42.50	0.00
s	0.20	0.20	0.20	0.00
kg/m^3	500.00	1000.00	790.43	168.00
t (days)	3.00	56.00	20.22	17.07
ρ ($\mathrm{kg \cdot m^{-3}}$)	1478.15	1853.41	1667.10	134.47
ω (%)	29.00	90.00	60.23	19.76
ρ_d ($\mathrm{kg \cdot m^{-3}}$)	822.65	1435.19	1136.03	152.87
$1/\rho_d$ ($\mathrm{m^3 \cdot kg^{-1}}$)	$6.97\mathrm{E}^{-4}$	$1.22\mathrm{E}^{-3}$	$8.97\mathrm{E}^{-4}$	$1.27\mathrm{E}^{-4}$
$\%Soil$	35.14	75.81	54.90	11.48
$\%Cement$	24.19	64.86	45.10	11.48
$\gamma_{s.mixt}$ ($\mathrm{kg \cdot m^{-3}}$)	2758.86	2902.23	2830.60	31.75
e	0.96	2.45	1.54	0.37
n	48.95	71.02	59.83	5.68
$1/n$	0.01	0.02	0.02	0.00
ω_{sat} (%)	34.11	86.33	54.37	12.92
S_ω	0.85	1.49	1.09	0.19
C_{iv}	0.21	0.61	0.41	0.11
$n/(C_{iv})^d$	51.21	73.81	62.03	5.49
$\%Sand$	0.00	39.00	13.44	12.82
$\%Silt$	33.00	57.00	50.57	7.48
$\%Clay$	22.50	45.00	35.85	7.48
$\%OM$	0.40	8.30	3.51	2.28
E_0 (GPa)	0.25	7.89	2.36	1.32



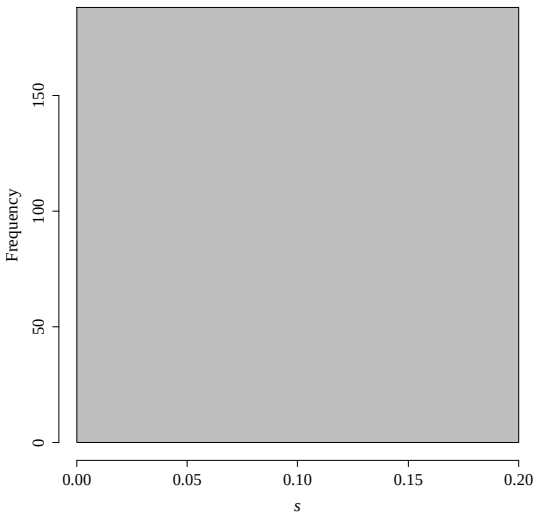
(1) Histogram of W/C



(2) Histogram of CT

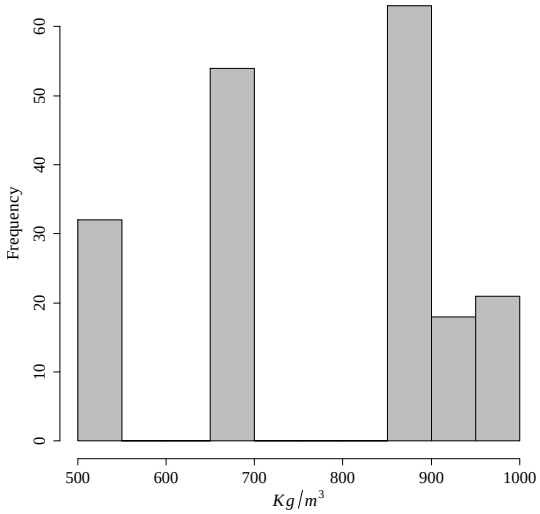


(3) Histogram of SCC

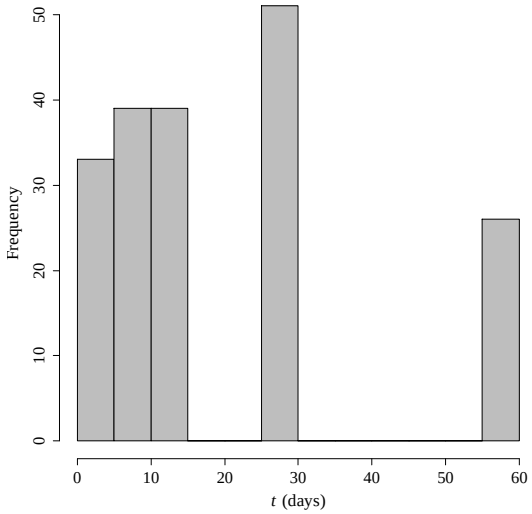


(4) Histogram of s

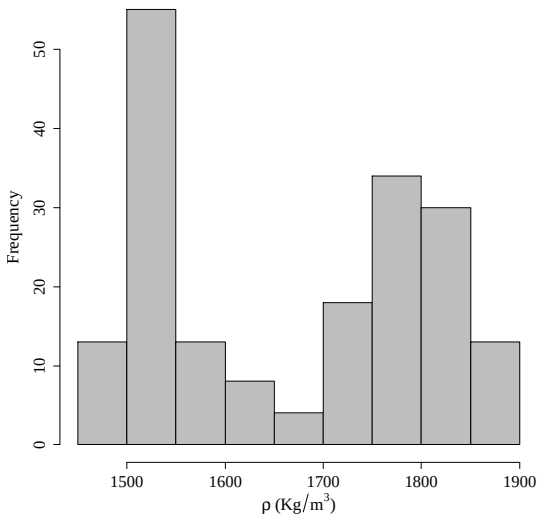
Figure A.2: Histograms of the numeric variables used in E_0 study of $JGLF$



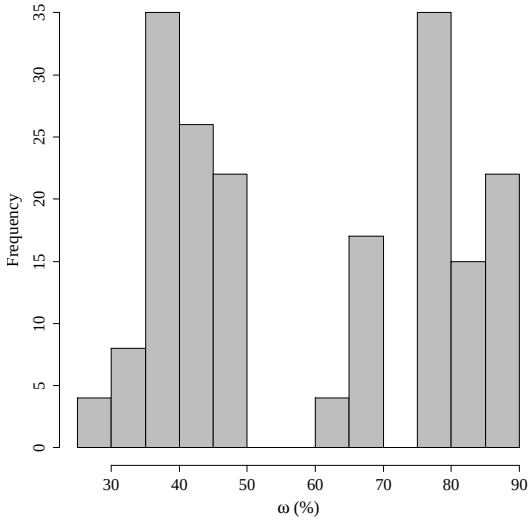
(5) Histogram of kg/m^3



(6) Histogram of t

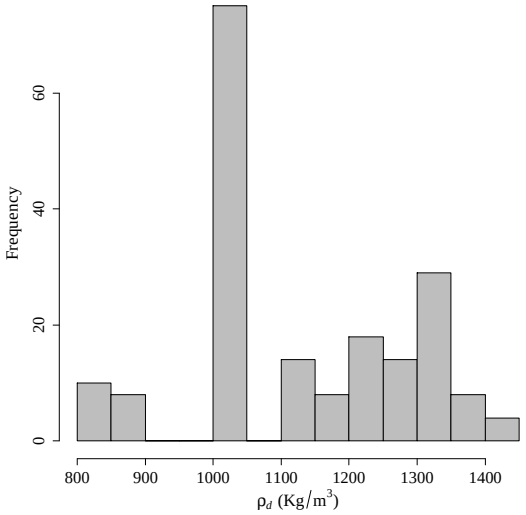


(7) Histogram of ρ

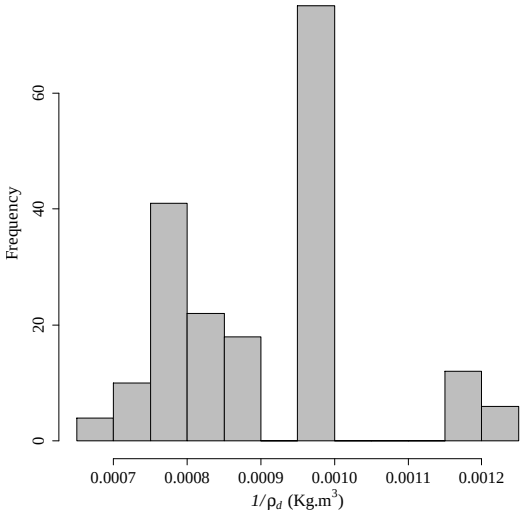


(8) Histogram of ω

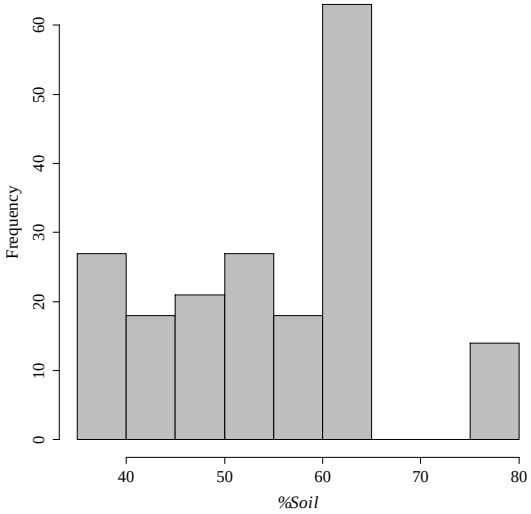
Figure A.2: Histograms of the numeric variables used in E_0 study of *JGLF* (cont'd)



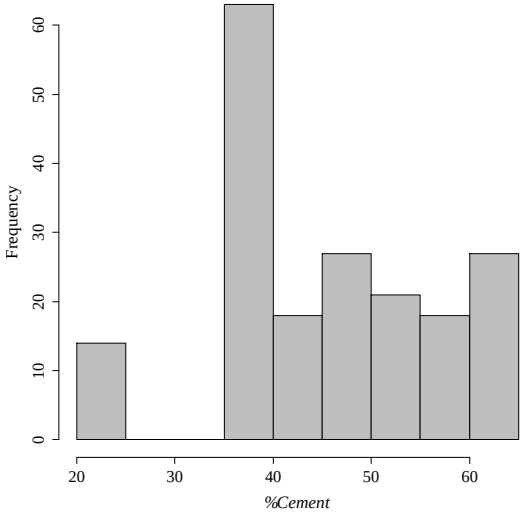
(9) Histogram of ρ_d



(10) Histogram of $1/\rho_d$

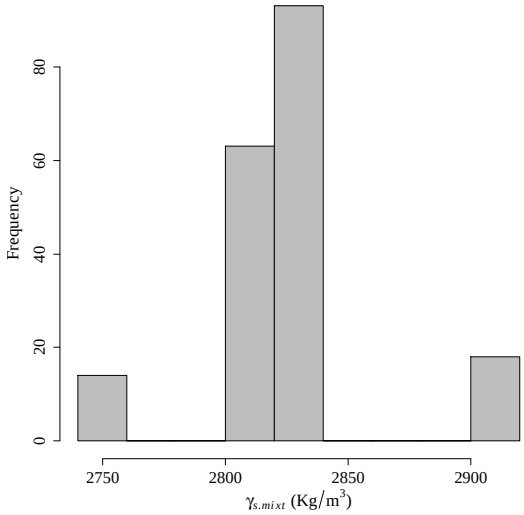


(11) Histogram of $\%Soil$

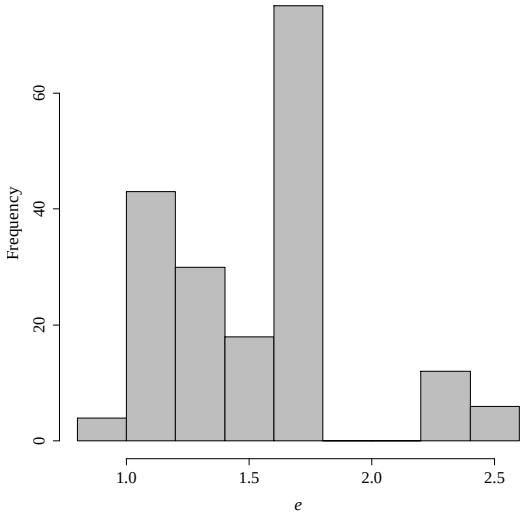


(12) Histogram of $\%Cement$

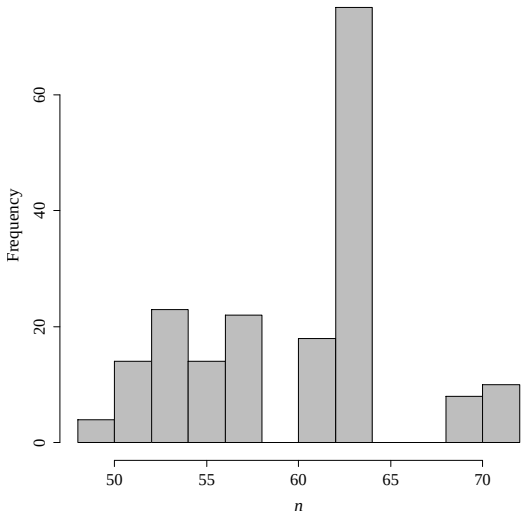
Figure A.2: Histograms of the numeric variables used in E_0 study of $JGLF$ (cont'd)



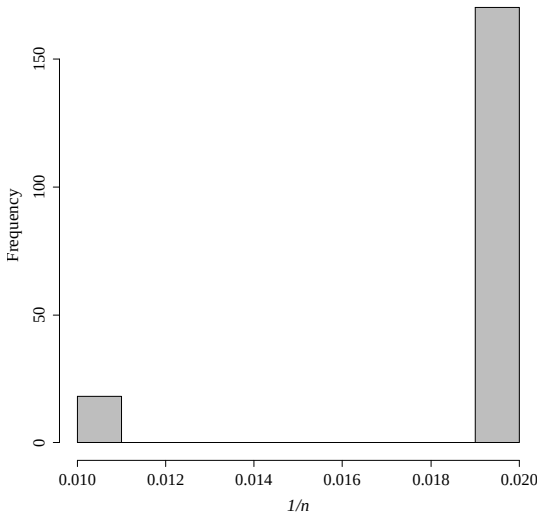
(13) Histogram of $\gamma_{s.mixt}$



(14) Histogram of e

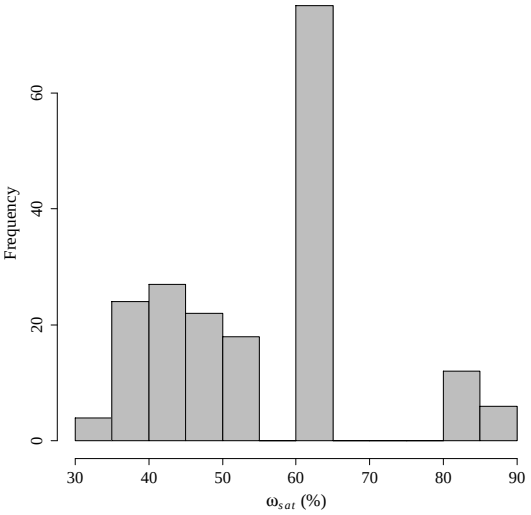


(15) Histogram of n

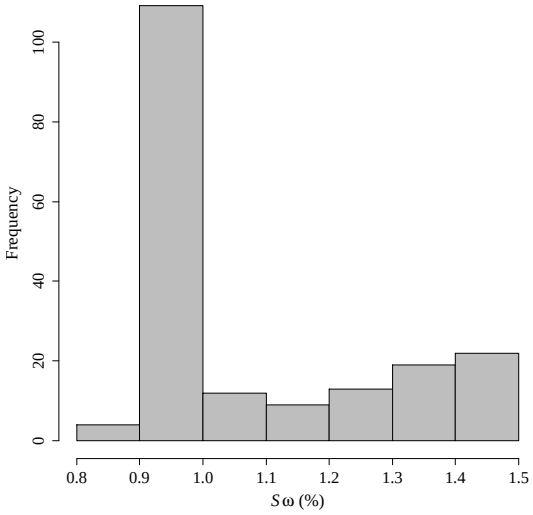


(16) Histogram of $1/n$

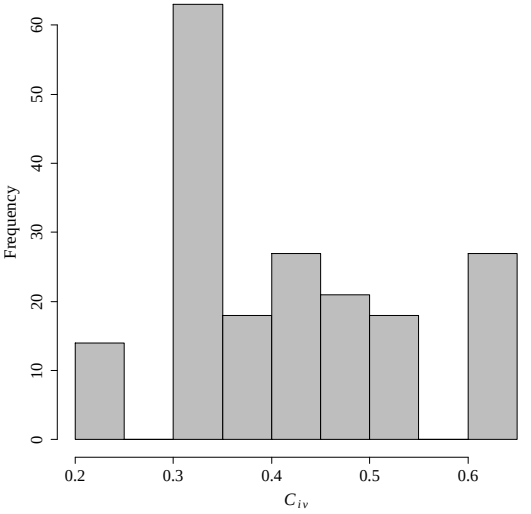
Figure A.2: Histograms of the numeric variables used in E_0 study of *JGLF* (cont'd)



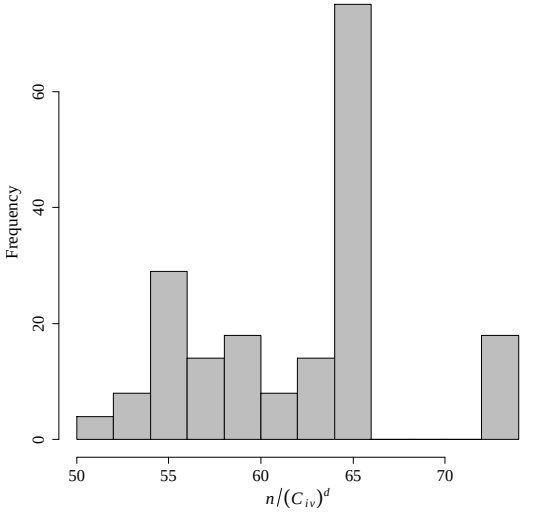
(17) Histogram of ω_{sat}



(18) Histogram of S_ω

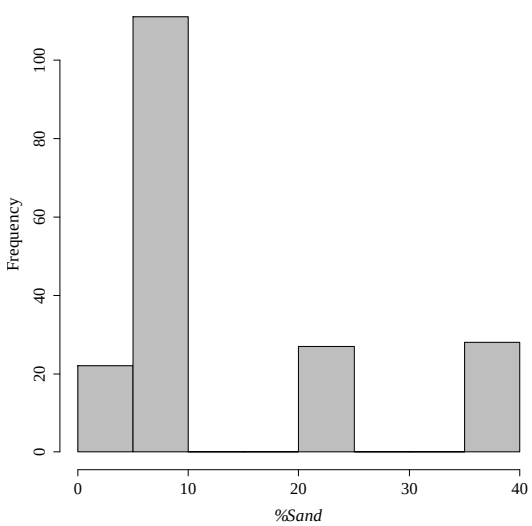


(19) Histogram of C_{iv}

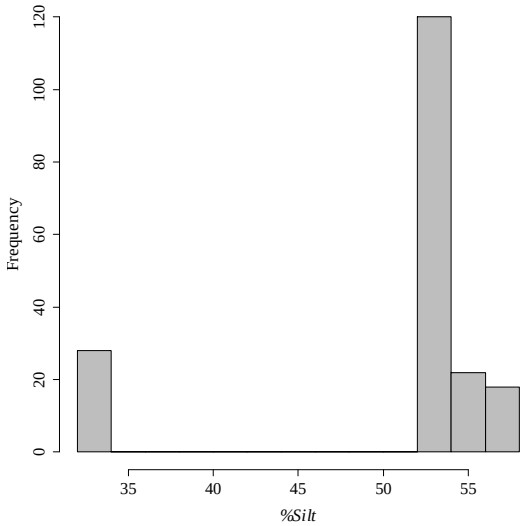


(20) Histogram of $n/(C_{iv})^d$

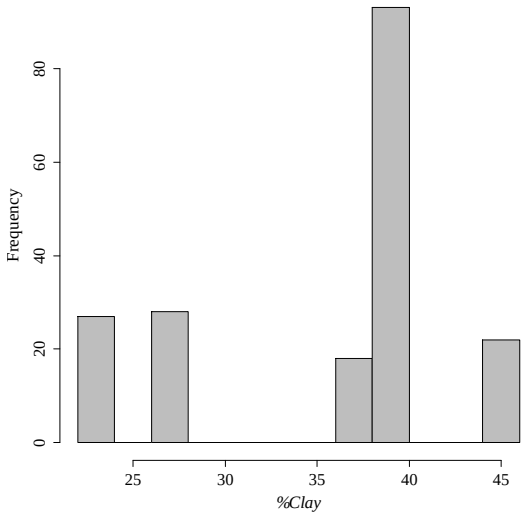
Figure A.2: Histograms of the numeric variables used in E_0 study of *JGLF* (cont'd)



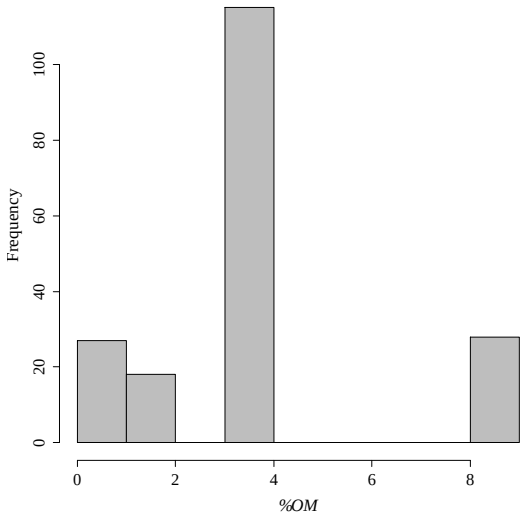
(21) Histogram of $\%Sand$



(22) Histogram of $\%Silt$



(23) Histogram of $\%Clay$



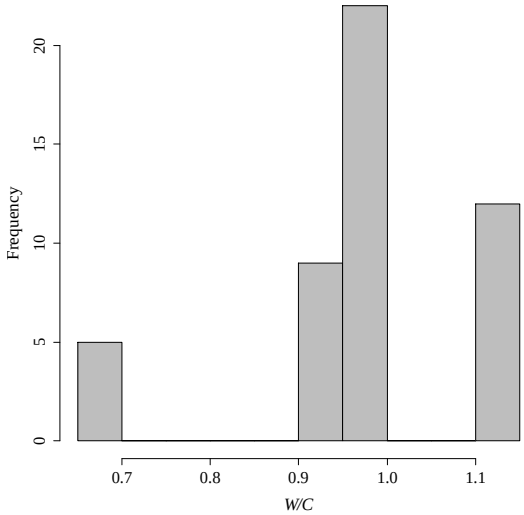
(24) Histogram of $\%OM$

Figure A.2: Histograms of the numeric variables used in E_0 study of *JGLF* (cont'd)

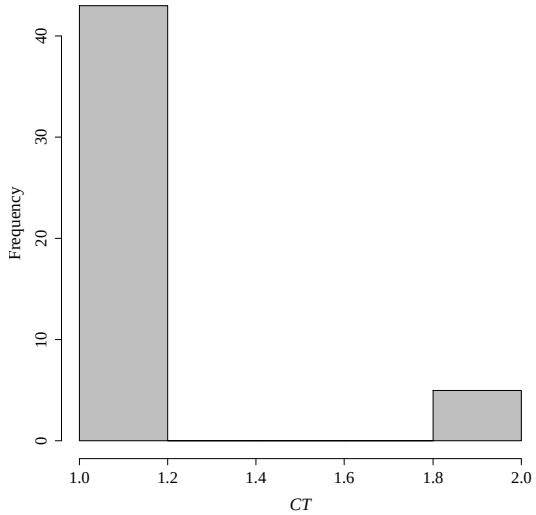
A.1.3 Main statistics and histograms for $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} study

Table A.3: Summary of the input and output variables of database used in $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} study of *JGLF*

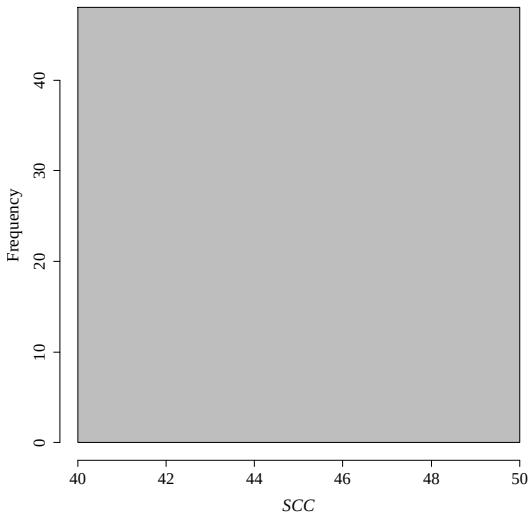
Variable	Minimum	Maximum	Mean	Standard Deviation
W/C	0.69	1.11	0.98	0.12
CT	1.00	2.00	1.10	0.31
SCC	42.50	42.50	42.50	0.00
s	0.20	0.20	0.20	0.00
kg/m^3	500.00	1000.00	783.33	178.15
t (days)	28	84	64.75	19.29
ρ $(\mathrm{kg} \cdot \mathrm{m}^{-3})$	1478.15	1853.41	1674.79	134.64
ω (%)	29.00	90.00	59.06	19.75
ρ_d $(\mathrm{kg} \cdot \mathrm{m}^{-3})$	822.65	1435.19	1139.66	158.19
$1/\rho_d$ $(\mathrm{m}^3 \cdot \mathrm{kg}^{-1})$	$6.97\mathrm{E}^{-4}$	$1.22\mathrm{E}^{-3}$	$8.95\mathrm{E}^{-4}$	$1.34\mathrm{E}^{-4}$
$\%Soil$	35.14	75.81	55.45	11.87
$\%Cement$	24.19	64.86	44.55	11.87
$\gamma_{s.mixt}$ $(\mathrm{kg} \cdot \mathrm{m}^{-3})$	2758.86	2902.23	2830.68	36.37
e	0.96	2.45	1.54	0.39
n	48.95	71.02	59.70	5.90
$1/n$	0.01	0.02	0.02	0.00
ω_{sat} (%)	34.11	86.33	54.20	13.57
S_ω	0.85	1.49	1.08	0.19
C_{iv}	0.21	0.61	0.41	0.12
$n/(C_{iv})^d$	51.21	73.81	61.93	5.7
$\%Sand$	0.00	39.00	14.40	13.67
$\%Silt$	33.00	57.00	49.90	8.32
$\%Clay$	22.50	45.00	35.52	7.40
$\%OM$	0.40	8.30	3.70	2.45
E_{max} (GPa)	1.50	7.00	3.44	1.30
$E_{sec50\%}$ (GPa)	1.50	5.67	3.17	1.11
$E_{tg50\%}$ (GPa)	1.30	4.90	2.76	0.93



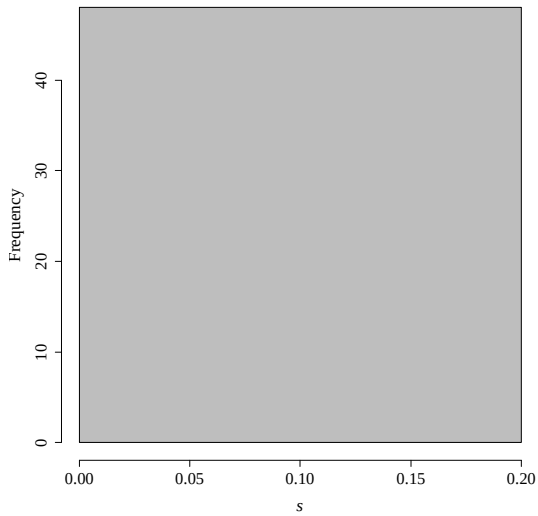
(1) Histogram of W/C



(2) Histogram of CT

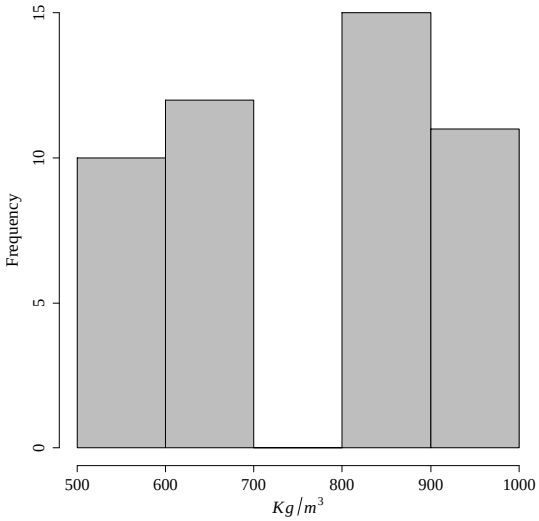


(3) Histogram of SCC

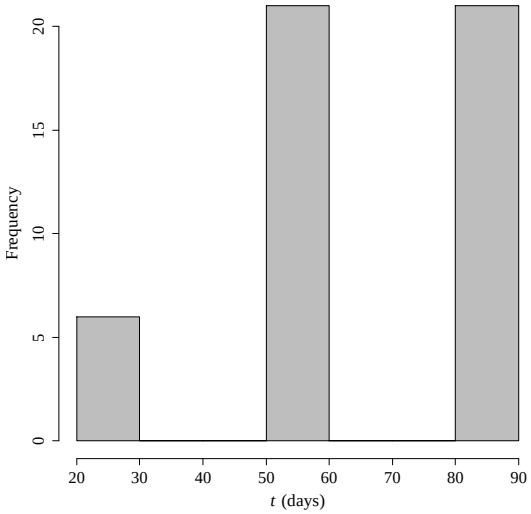


(4) Histogram of s

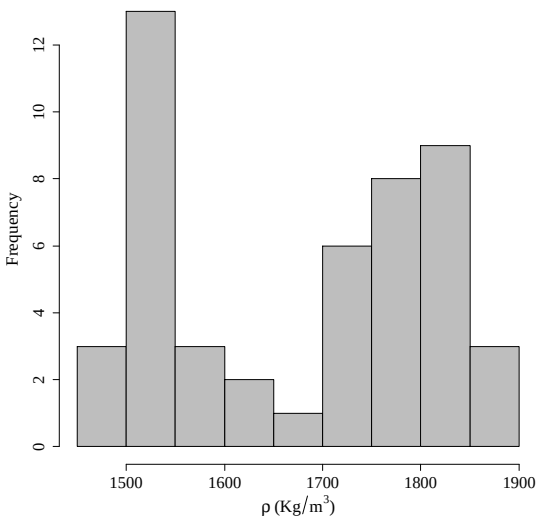
Figure A.3: Histograms of the numeric variables used in $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} studies of *JGLF*



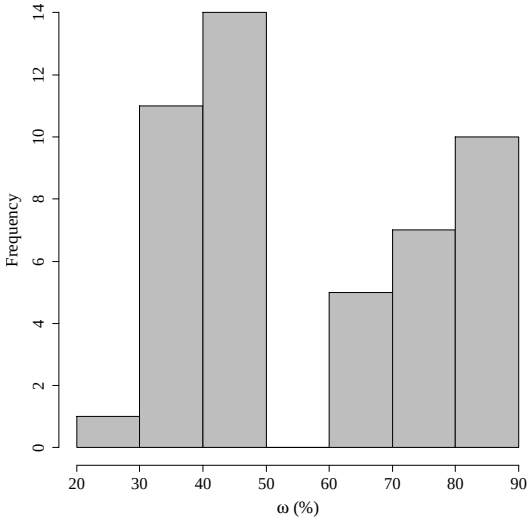
(5) Histogram of kg/m^3



(6) Histogram of t

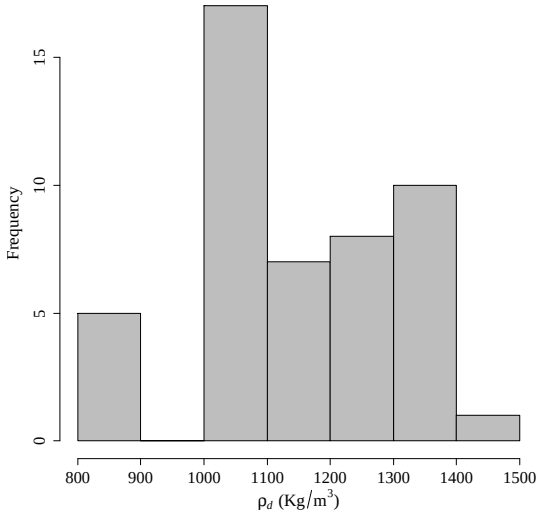


(7) Histogram of ρ

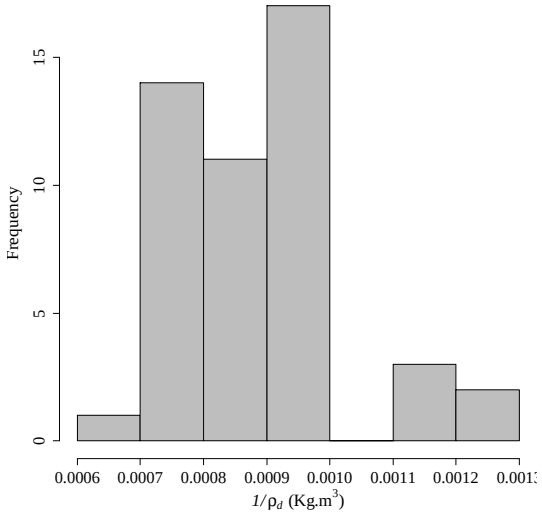


(8) Histogram of ω

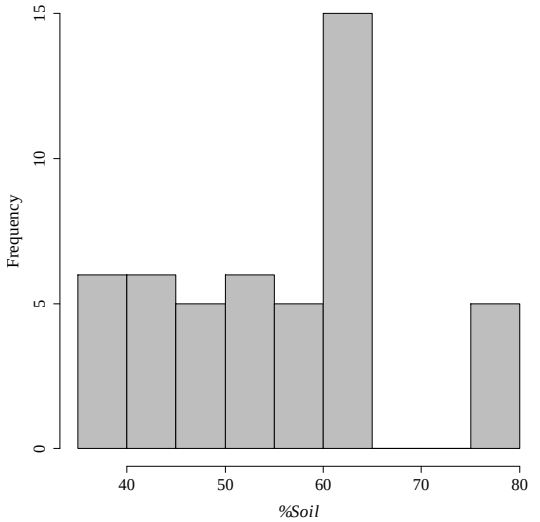
Figure A.3: Histograms of the numeric variables used in $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} studies of *JGLF* (cont'd)



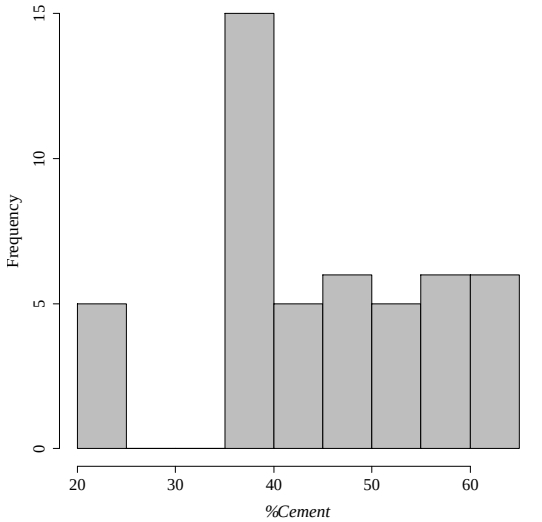
(9) Histogram of ρ_d



(10) Histogram of $1/\rho_d$

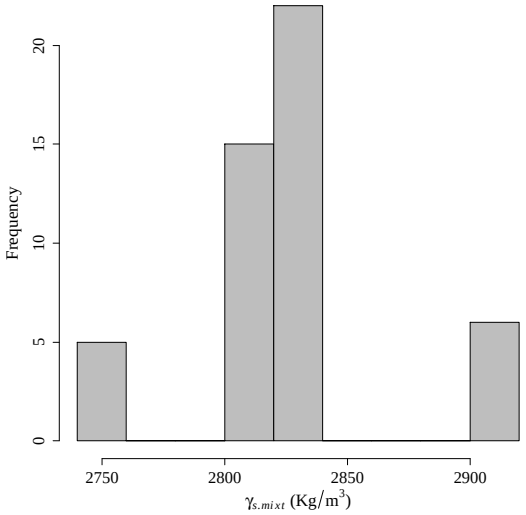


(11) Histogram of $\%Soil$

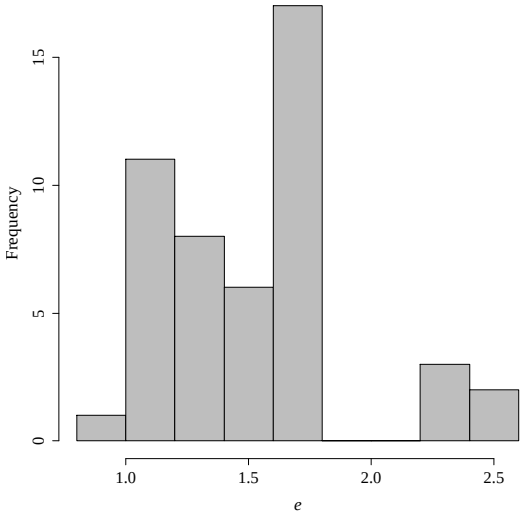


(12) Histogram of $\%Cement$

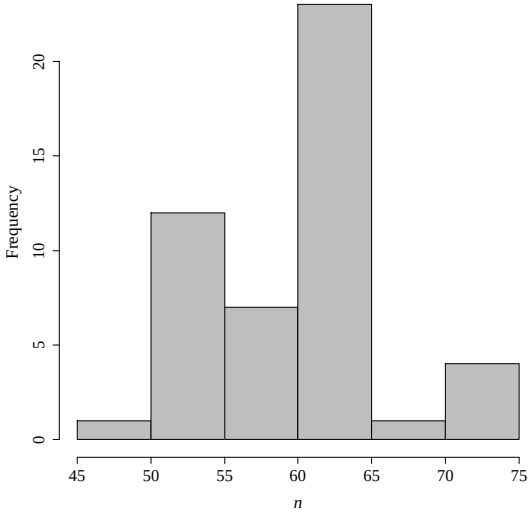
Figure A.3: Histograms of the numeric variables used in $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} studies of *JGLF* (cont'd)



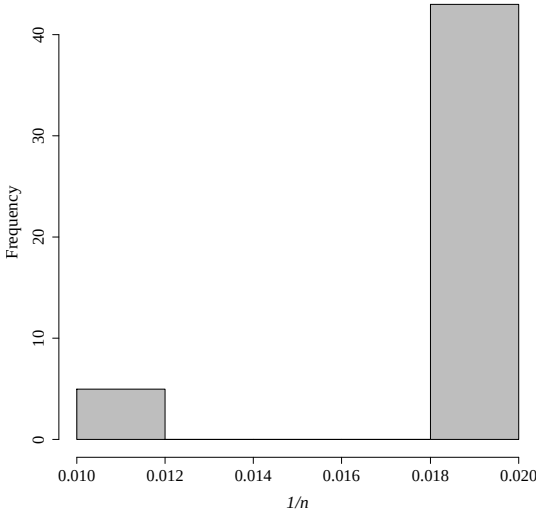
(13) Histogram of $\gamma_{s.mixt}$



(14) Histogram of e

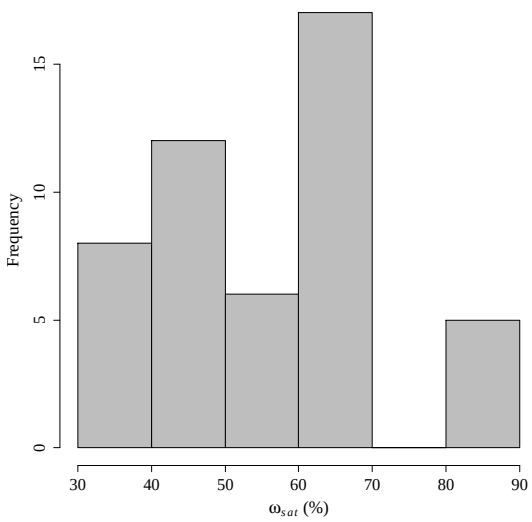


(15) Histogram of n

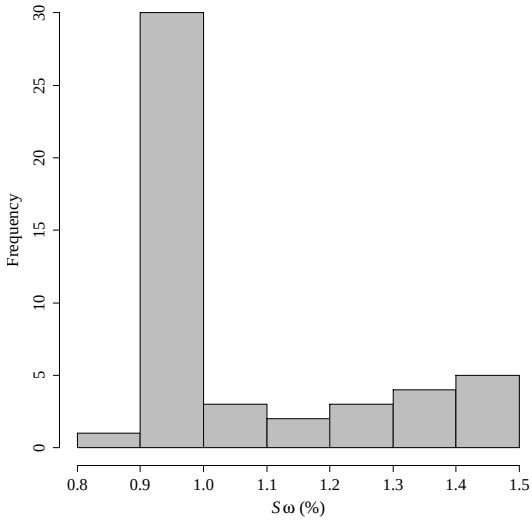


(16) Histogram of $1/n$

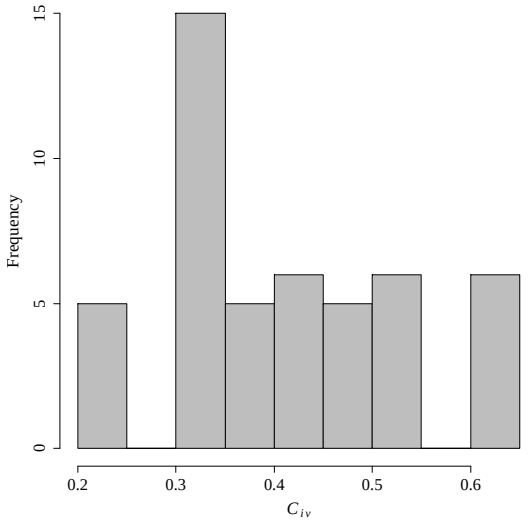
Figure A.3: Histograms of the numeric variables used in $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} studies of *JGLF* (cont'd)



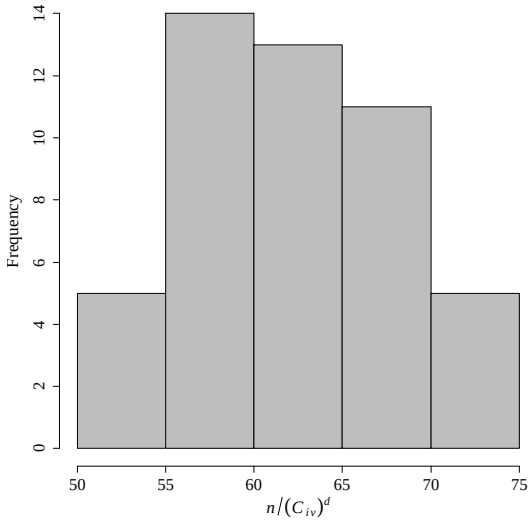
(17) Histogram of ω_{sat}



(18) Histogram of S_ω

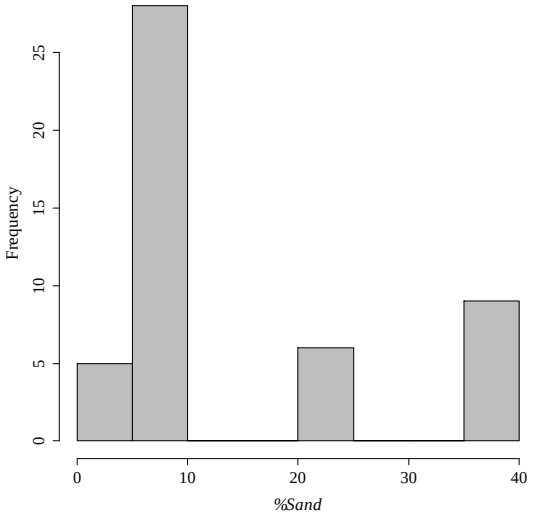


(19) Histogram of C_{iv}

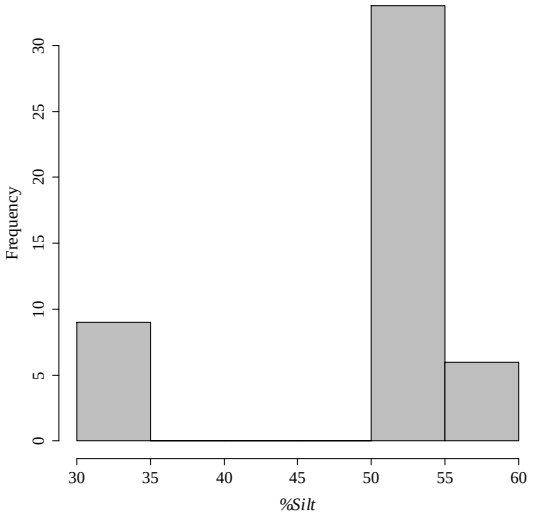


(20) Histogram of $n/(C_{iv})^d$

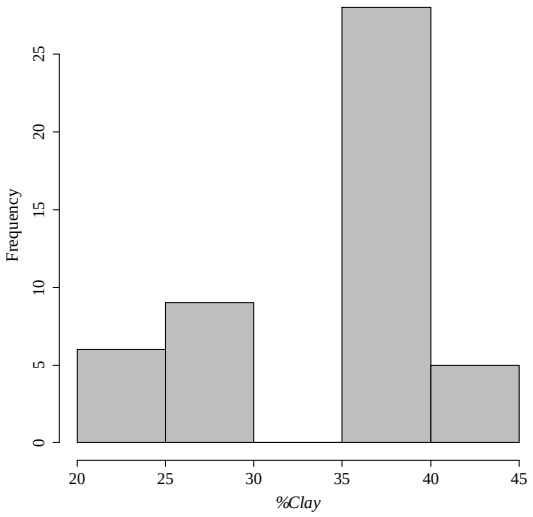
Figure A.3: Histograms of the numeric variables used in $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} studies of *JGLF* (cont'd)



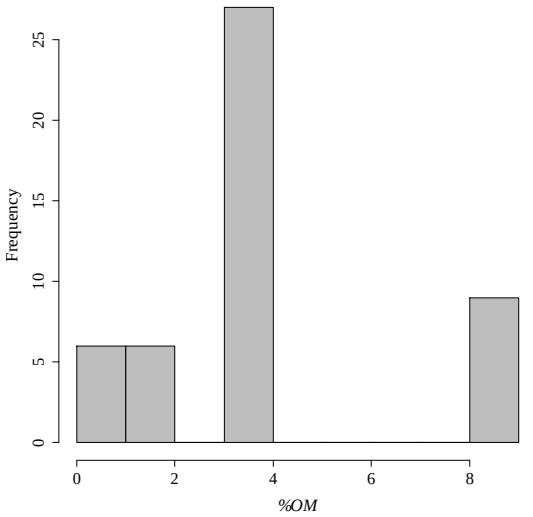
(21) Histogram of $\%Sand$



(22) Histogram of $\%Silt$



(23) Histogram of $\%Clay$



(24) Histogram of $\%OM$

Figure A.3: Histograms of the numeric variables used in $E_{tg50\%}$, $E_{sec50\%}$ and E_{max} studies of *JGLF* (cont'd)

A.2 Jet grouting field samples data

A.2.1 Main statistics and histograms for *UCS* study

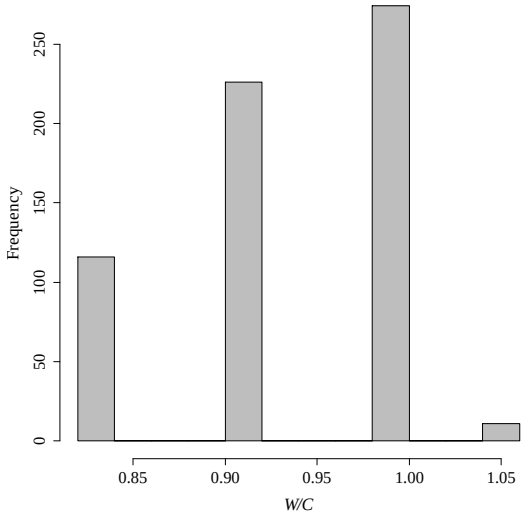
Table A.4: Summary of the input and output variables of database used in UCS study of *soilcrete* mixtures

Variable	Minimum	Maximum	Mean	Standard Deviation
W/C	0.83	1.05	0.94	0.07
CT	1.00	2.00	1.22	0.42
SCC	42.50	42.50	42.50	0.00
s	0.20	0.20	0.20	0.00
kg/m^3	492.00	1194.00	846.08	163.79
kg/ml	600.00	18885	3906.94	3301.32
t (days)	9.00	181.00	47.77	32.22
ρ $(\mathrm{kg}\cdot\mathrm{m}^{-3})$	1000.00	2600.00	1665.42	132.66
ω (%)	2.50	96.80	38.80	12.13
ρ_d $(\mathrm{kg}\cdot\mathrm{m}^{-3})$	693.00	1776.26	1213.17	177.64
$1/\rho_d$ $(\mathrm{m}^3\cdot\mathrm{kg}^{-1})$	$5.63\mathrm{E}^{-4}$	$1.44\mathrm{E}^{-3}$	$8.42\mathrm{E}^{-4}$	$1.22\mathrm{E}^{-4}$
$\%Soil$	72.19	86.30	78.70	3.34
$\%Cement$	13.70	27.81	21.30	3.34
$\gamma_{s.mixt}$ $(\mathrm{kg}\cdot\mathrm{m}^{-3})$	2711.64	2775.13	2745.86	15.01
e	0.56	2.99	1.31	0.34
n	35.91	74.92	55.86	6.43
$1/n$	0.01	0.03	0.02	0.00
ω_{sat} (%)	20.26	108.11	47.78	12.23
S_ω	0.09	2.38	0.81	0.17
C_{iv}	0.18	0.43	0.31	0.06
$n/(C_{iv})^d$	37.88	79.17	59.41	6.88
W_c/C	0.96	2.30	1.53	0.39
s/C	2.60	6.30	3.83	0.90
OM/C	0.02	0.61	0.26	0.17
$OM/C^{W_c/C}$	0.00	0.37	0.18	0.15
ρ_{grout}	1.52	1.59	1.55	0.03
$\%Sand$	0.01	39.00	24.40	16.53

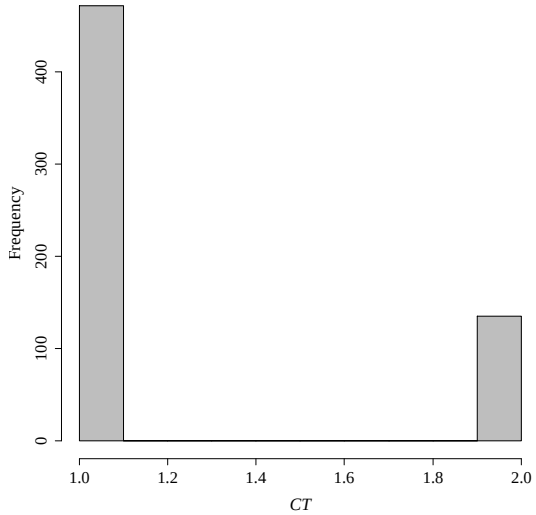
Continued on next page

Table A.4 – continued from previous page

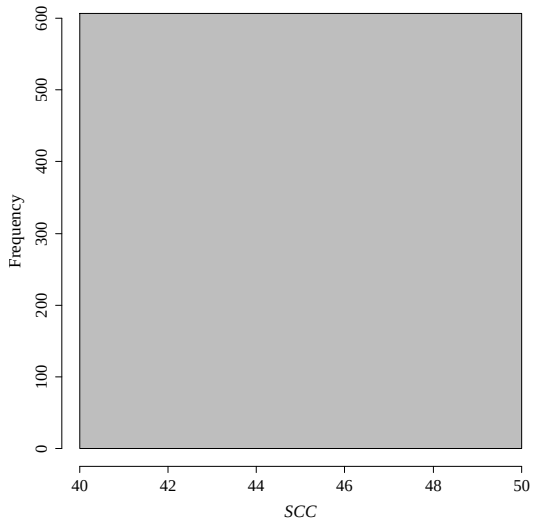
Variable	Minimum	Maximum	Mean	Standard Deviation
$\%Silt$	33.00	57.00	43.23	11.18
$\%Clay$	22.50	45.00	31.84	6.87
$\%OM$	0.40	8.30	5.40	3.22
H (m)	7.00	31.95	22.50	4.97
JS	1.00	3.00	2.05	0.37
WS (cm/min)	6.00	20.87	9.87	3.51
rpm	3.00	10.00	4.83	1.52
WT (s)	11.50	60.00	38.39	12.74
$Step$ (cm)	4.00	6.00	5.66	0.75
FR (l/min)	139.00	577.89	370.86	78.41
D_{grout} (mm)	4.00	7.00	4.84	0.73
N_{Dgrout}	1.00	2.00	1.66	0.47
D_{water} (mm)	0.00	5.00	0.16	0.88
P_{grout} (bar)	140.00	450.00	364.19	82.72
P_{air} (bar)	0.00	10.00	9.08	2.25
P_{water} (bar)	0.00	400.00	36.88	115.82
Imp_{grout} (kg)	58.06	278.95	220.13	66.98
UCS (MPa)	0.32	20.27	3.85	2.61



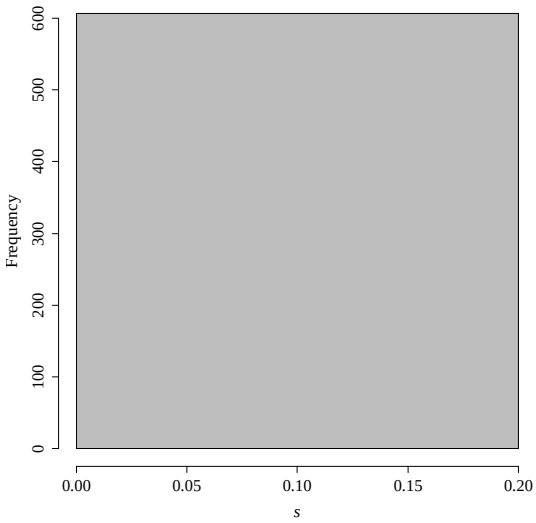
(1) Histogram of W/C



(2) Histogram of CT

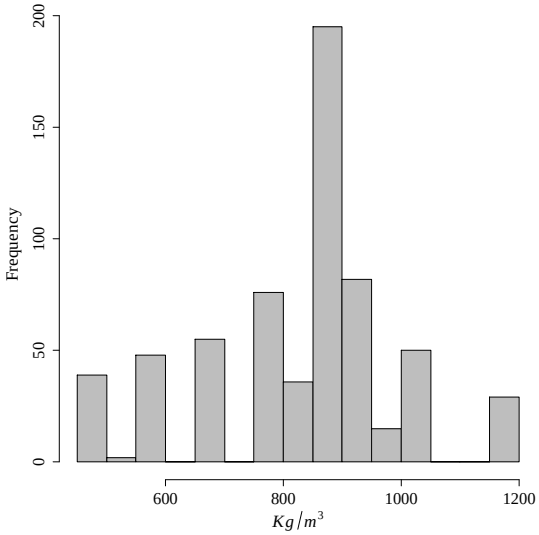


(3) Histogram of SCC

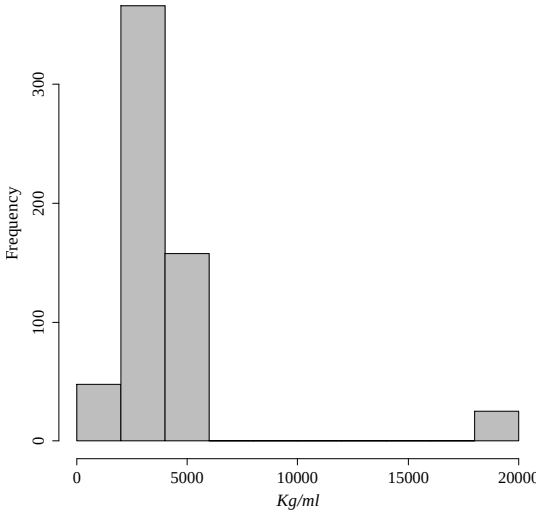


(4) Histogram of s

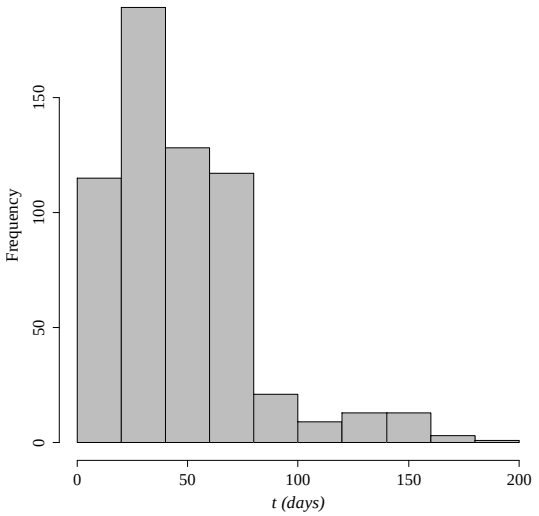
Figure A.4: Histograms of the numeric variables used in UCS study of *soilcrete* mixtures



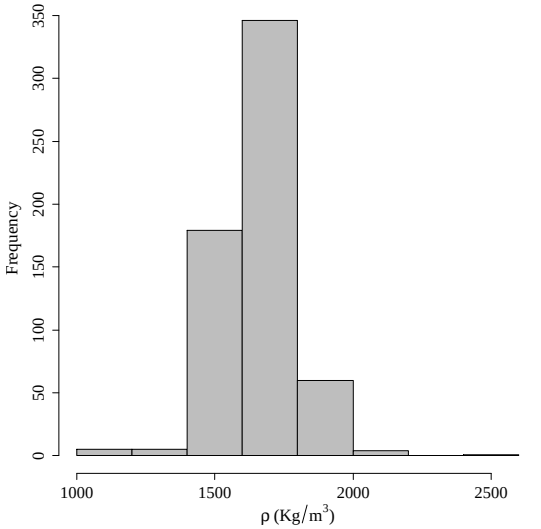
(5) Histogram of kg/m^3



(6) Histogram of kg/ml

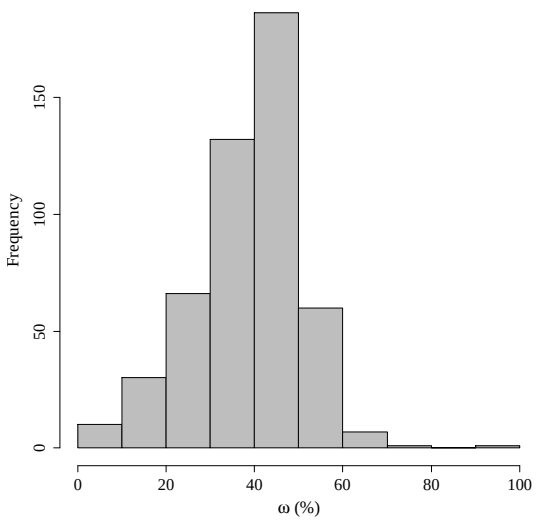


(7) Histogram of t

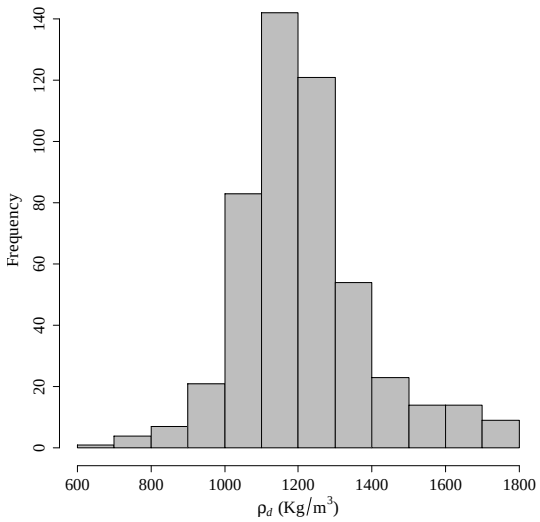


(8) Histogram of ρ

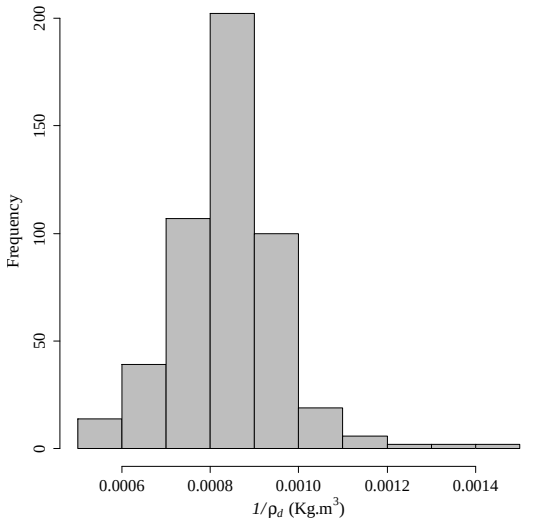
Figure A.4: Histograms of the numeric variables used in UCS study of *soilcrete* mixtures (cont'd)



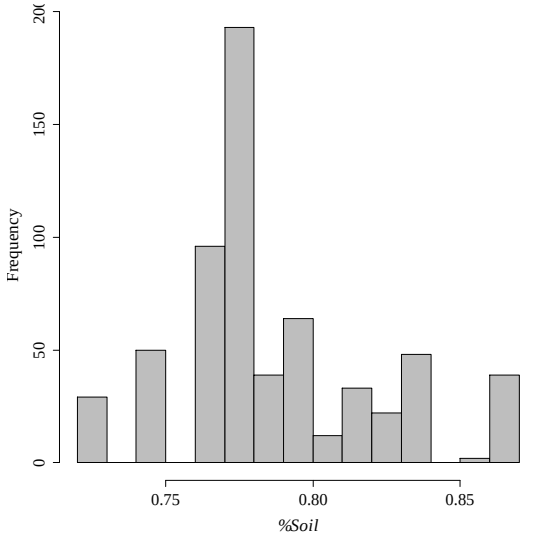
(9) Histogram of ω



(10) Histogram of ρ_d

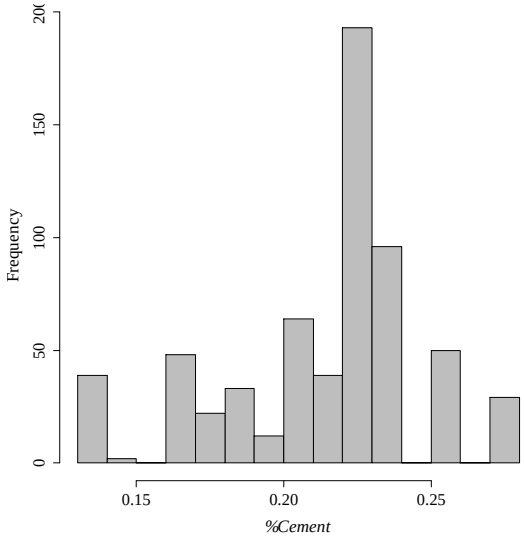


(11) Histogram of $1/\rho_d$

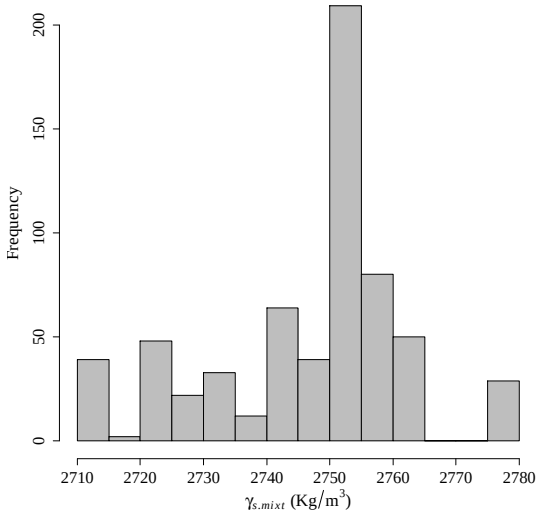


(12) Histogram of $\%Soil$

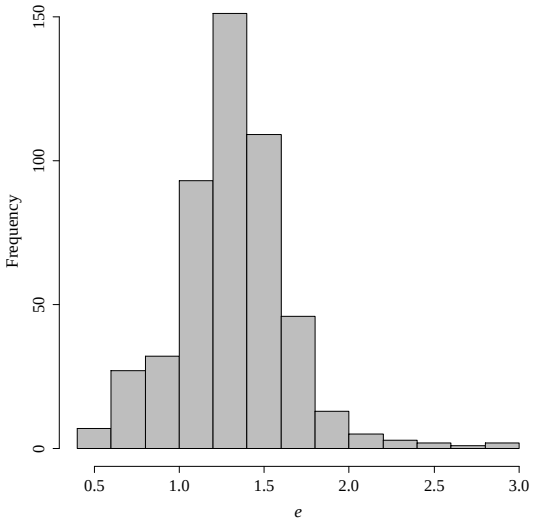
Figure A.4: Histograms of the numeric variables used in UCS study of *soilcrete* mixtures (cont'd)



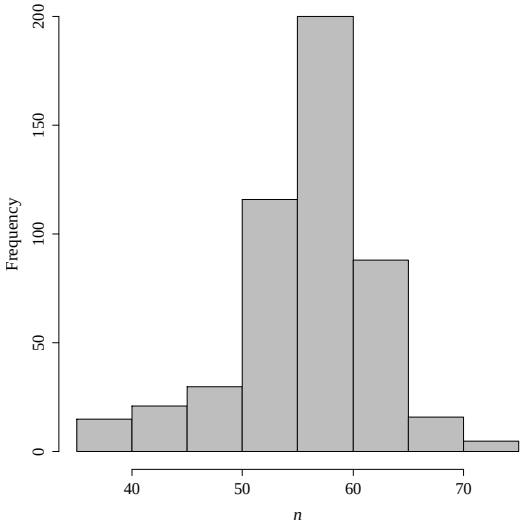
(13) Histogram of $\%Cement$



(14) Histogram of $\gamma_{s.mixt}$

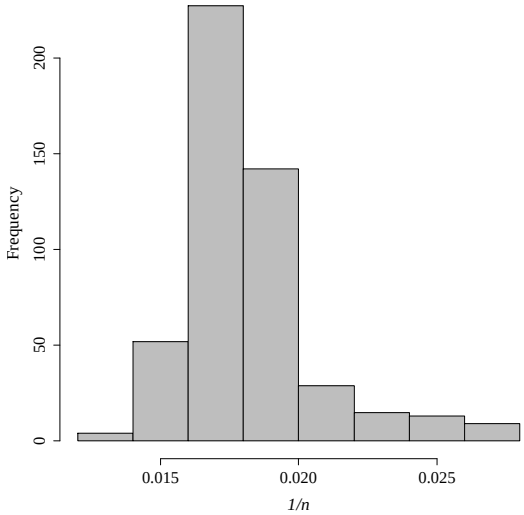


(15) Histogram of e

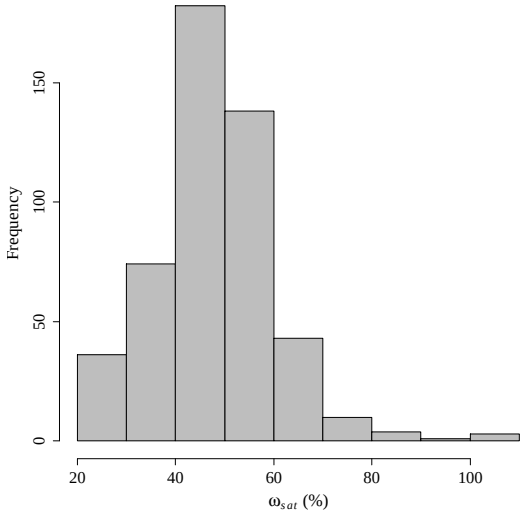


(16) Histogram of n

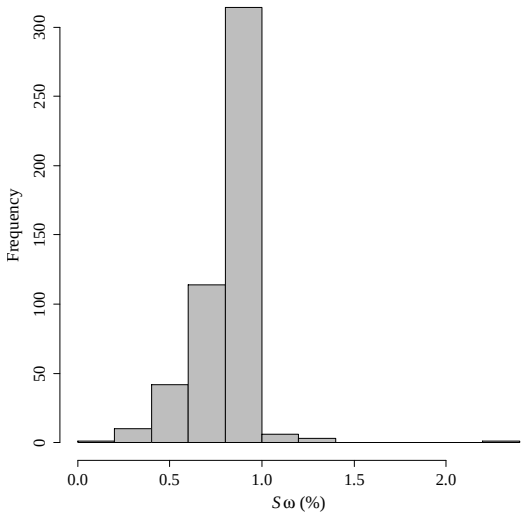
Figure A.4: Histograms of the numeric variables used in UCS study of *soilcrete* mixtures (cont'd)



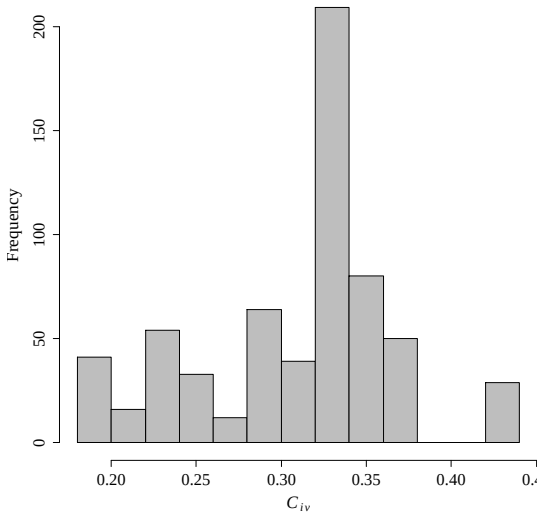
(17) Histogram of $1/n$



(18) Histogram of ω_{sat}

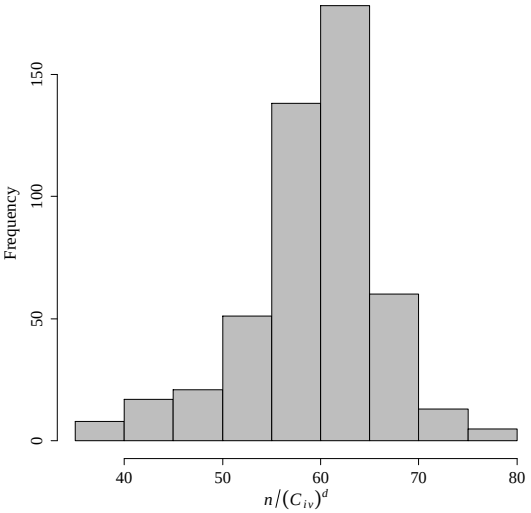


(19) Histogram of S_ω

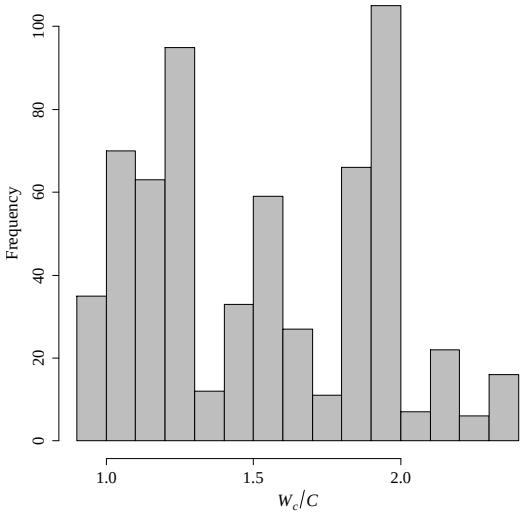


(20) Histogram of C_{iv}

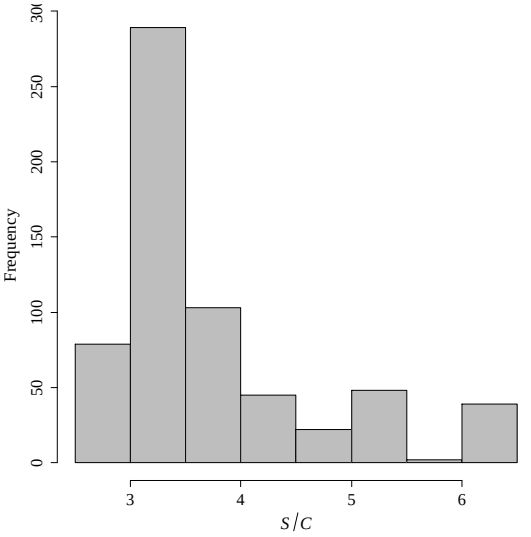
Figure A.4: Histograms of the numeric variables used in UCS study of *soilcrete* mixtures (cont'd)



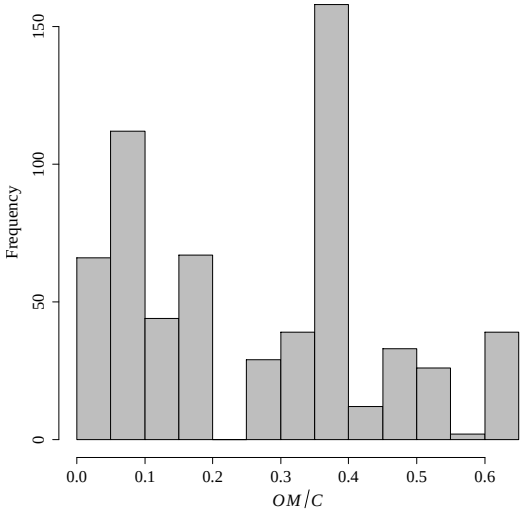
(21) Histogram of $n/(C_{iv})^d$



(22) Histogram of W_c/C

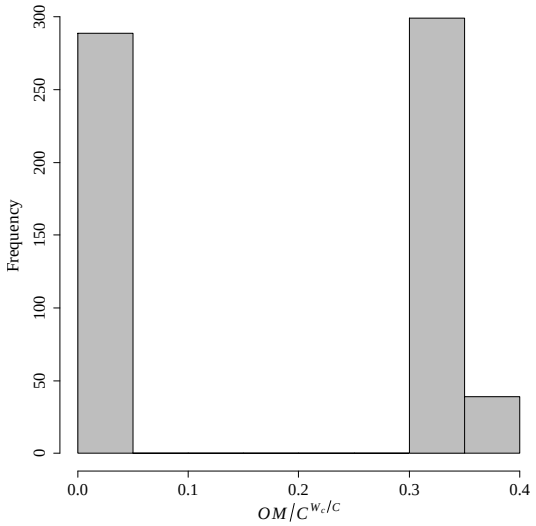


(23) Histogram of S/C

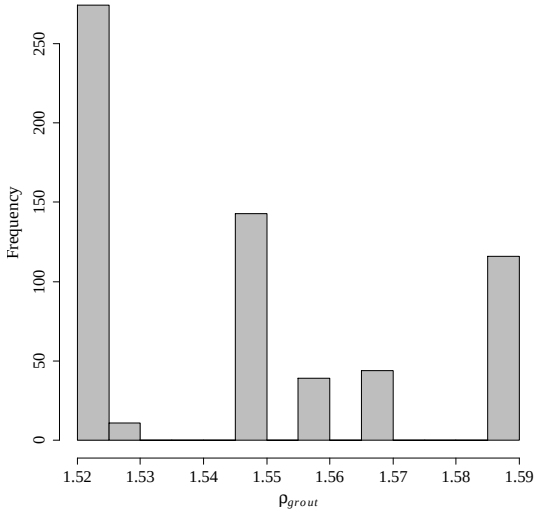


(24) Histogram of OM/C

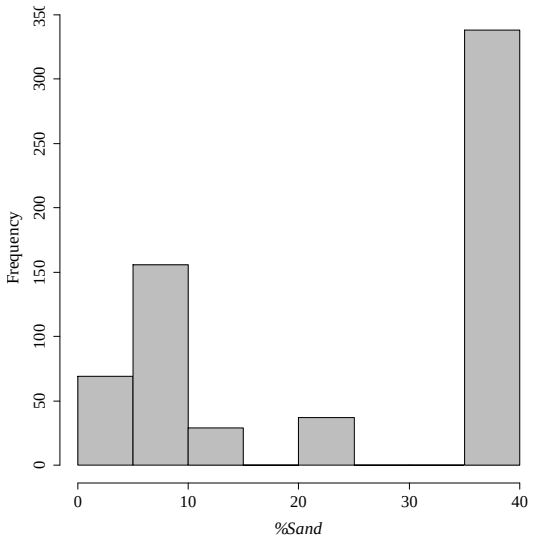
Figure A.4: Histograms of the numeric variables used in UCS study of *soilcrete* mixtures (cont'd)



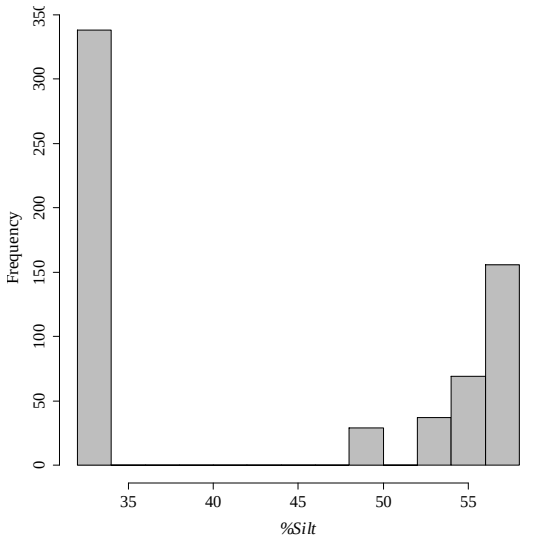
(25) Histogram of $OM/C^{W_c/C}$



(26) Histogram of ρ_{grout}

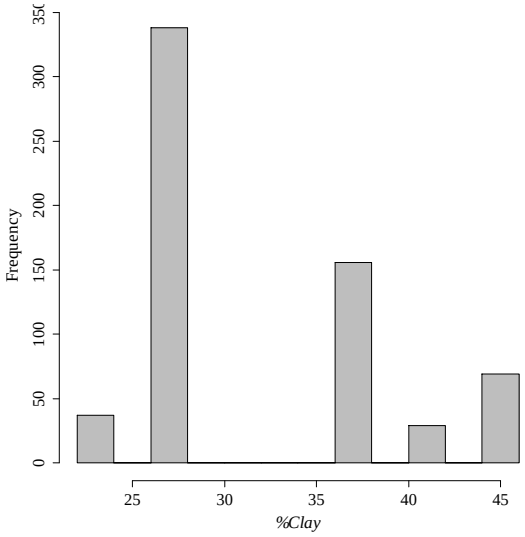


(27) Histogram of $\%Sand$

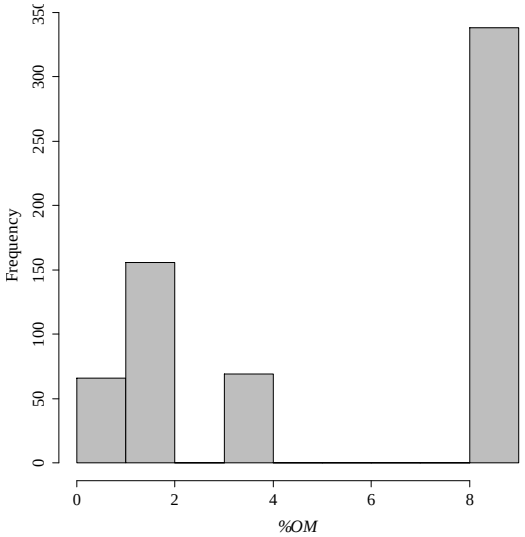


(28) Histogram of $\%Silt$

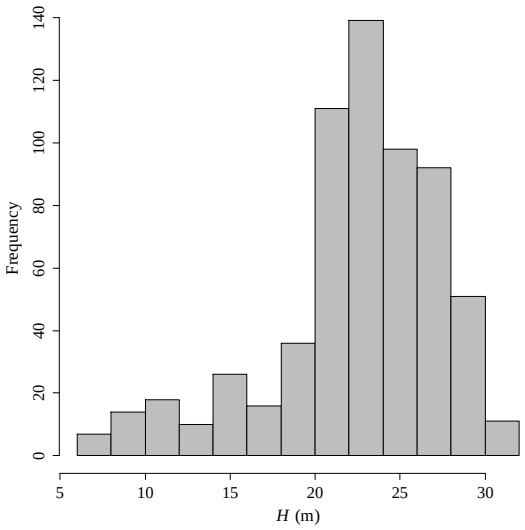
Figure A.4: Histograms of the numeric variables used in UCS study of *soilcrete* mixtures (cont'd)



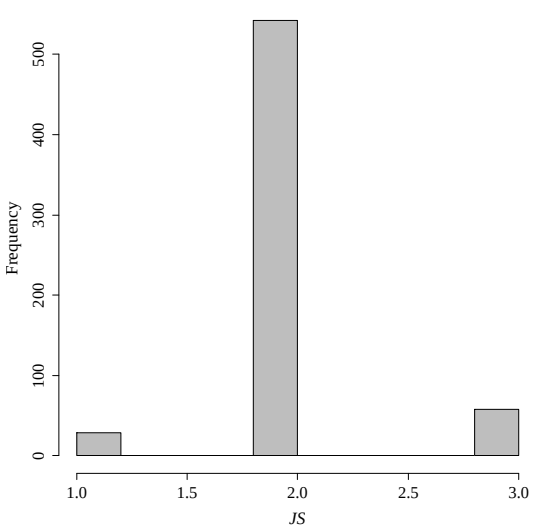
(29) Histogram of $\%Clay$



(30) Histogram of $\%OM$

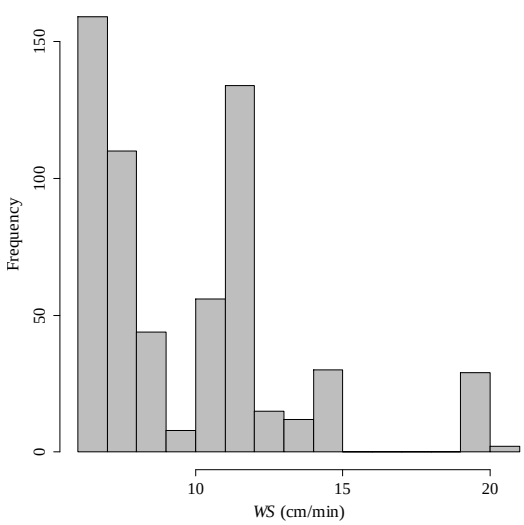


(31) Histogram of H

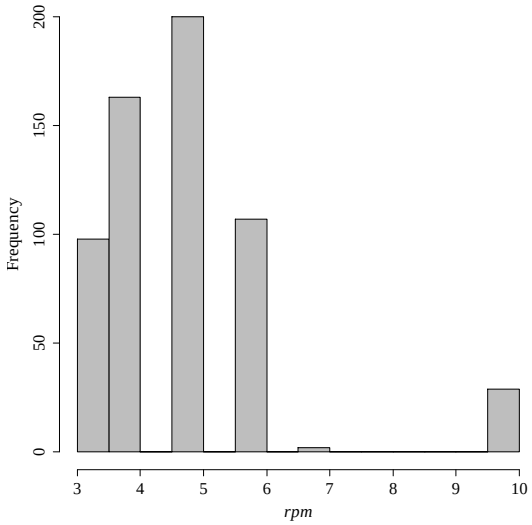


(32) Histogram of JS

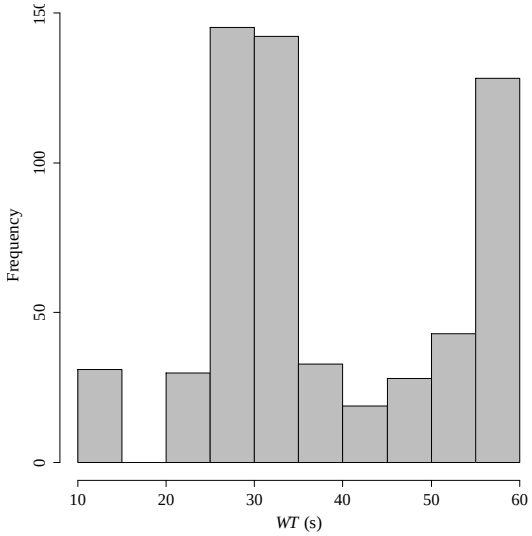
Figure A.4: Histograms of the numeric variables used in UCS study of *soilcrete* mixtures (cont'd)



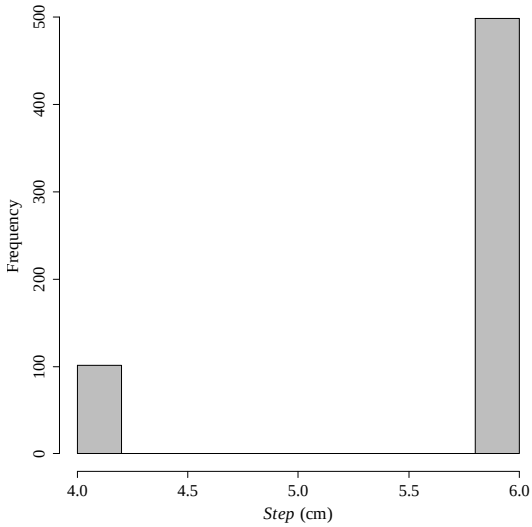
(33) Histogram of WS



(34) Histogram of rpm

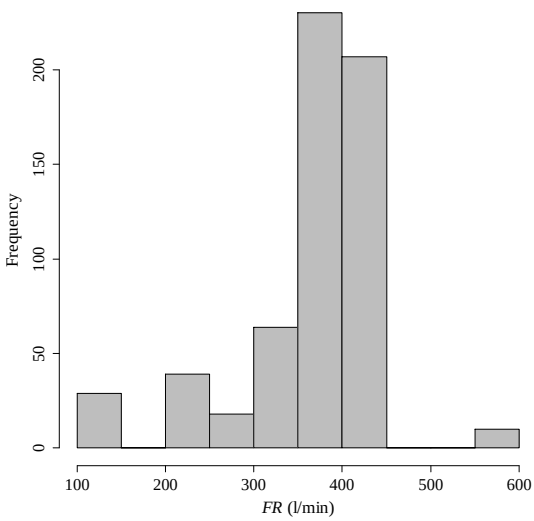


(35) Histogram of WT

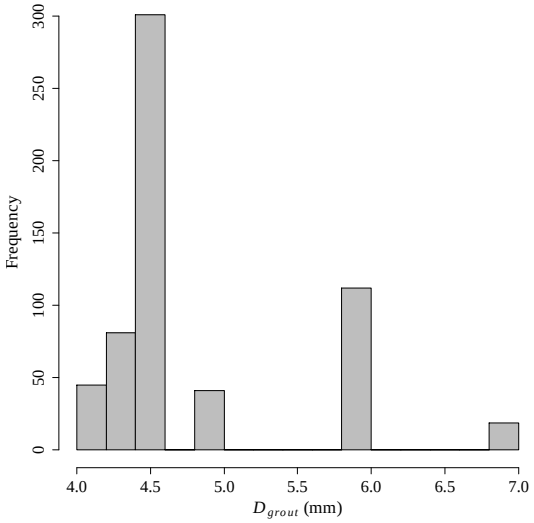


(36) Histogram of $Step$

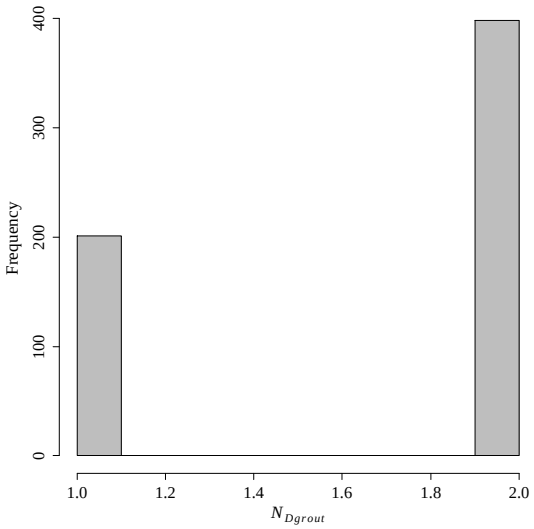
Figure A.4: Histograms of the numeric variables used in UCS study of *soilcrete* mixtures (cont'd)



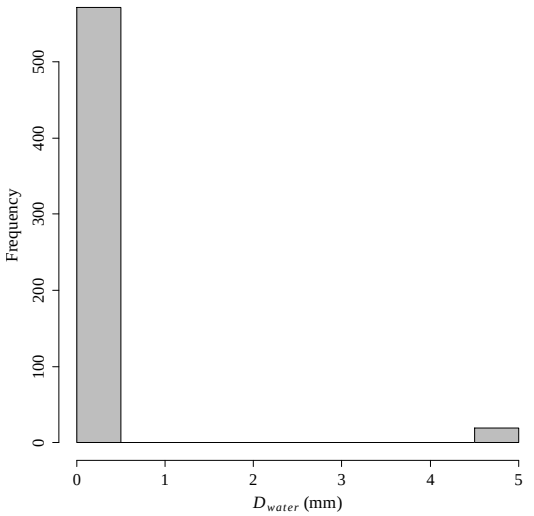
(37) Histogram of FR



(38) Histogram of D_{grout}

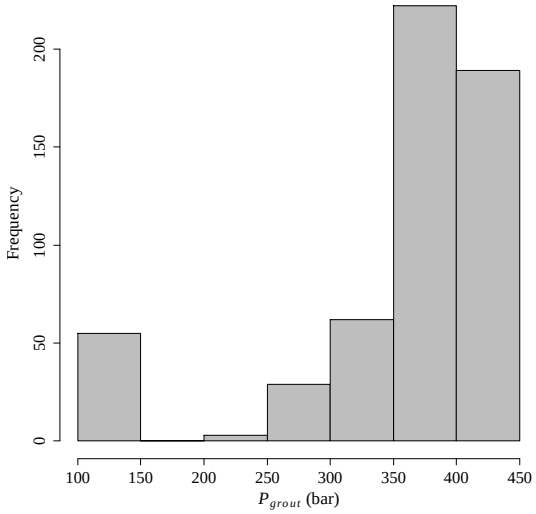


(39) Histogram of N_{Dgrout}

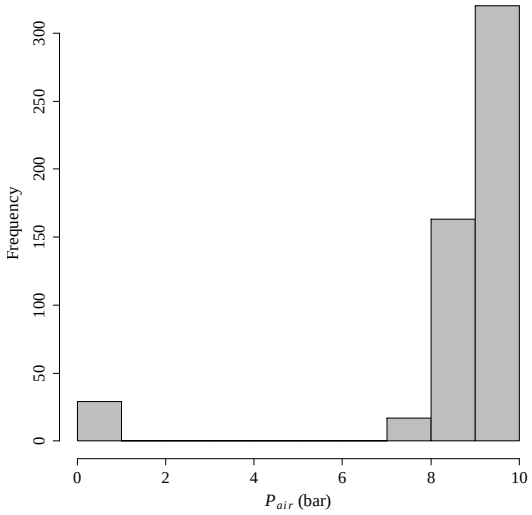


(40) Histogram of D_{water}

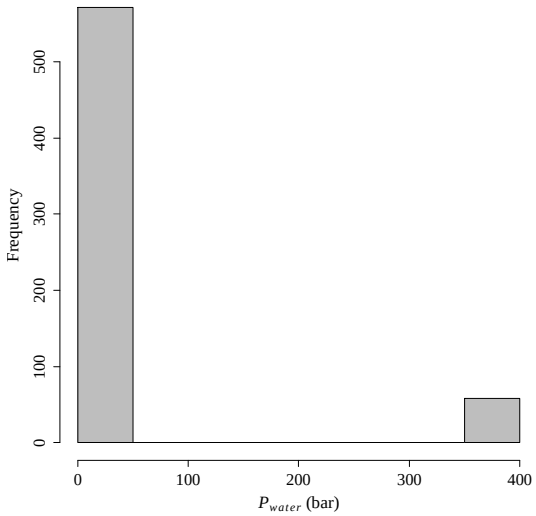
Figure A.4: Histograms of the numeric variables used in UCS study of *soilcrete* mixtures (cont'd)



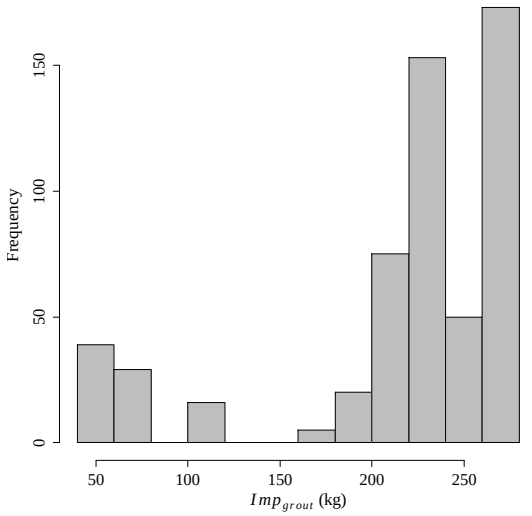
(41) Histogram of P_{grout}



(42) Histogram of P_{air}



(43) Histogram of P_{water}



(44) Histogram of Imp_{grout}

Figure A.4: Histograms of the numeric variables used in UCS study of *soilcrete* mixtures (cont'd)

A.2.2 Main statistics and histograms for E_0 study

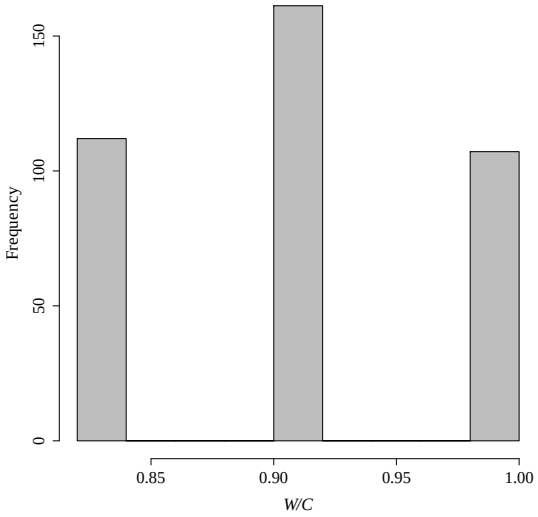
Table A.5: Summary of the input and output variables of database used in E_0 study of *soilcrete* mixtures

Variable	Minimum	Maximum	Mean	Standard Deviation
W/C	0.83	1.00	0.91	0.06
CT	1.00	2.00	1.02	0.15
SCC	42.50	42.50	42.50	0.00
s	0.20	0.20	0.20	0.00
kg/m^3	492.00	1194.00	821.28	186.31
kg/ml	600.00	18885.00	3907.28	4230.68
t (days)	9.00	181.00	41.52	34.10
ρ $(\mathrm{kg \cdot m^{-3}})$	1310.00	2080.00	1677.99	126.34
ω (%)	2.50	96.80	36.41	13.32
ρ_d $(\mathrm{kg \cdot m^{-3}})$	713.92	1776.26	1250.58	193.12
$1/\rho_d$ $(\mathrm{m^3 \cdot kg^{-1}})$	$5.63\mathrm{E}^{-4}$	$1.40\mathrm{E}^{-3}$	$8.18\mathrm{E}^{-4}$	$1.23\mathrm{E}^{-4}$
$\%Soil$	72.19	86.30	79.24	3.79
$\%Cement$	13.70	27.81	20.77	3.79
$\gamma_{s.mixt}$ $(\mathrm{kg \cdot m^{-3}})$	2711.64	2775.13	2743.44	17.04
e	0.56	2.85	1.25	0.33
n	35.91	74.05	54.47	6.94
$1/n$	0.01	0.03	0.02	0.00
ω_{sat} (%)	20.26	103.72	45.38	12.24
S_ω	0.09	1.12	0.79	0.17
C_{iv}	0.18	0.43	0.30	0.07
$n/(C_{iv})^d$	37.88	78.61	58.00	7.49
W_c/C	0.97	2.30	1.48	0.38
S/C	2.60	6.30	4.00	1.02
OM/C	0.08	0.61	0.34	0.16
$OM/C^{W_c/C}$	0.01	0.37	0.24	0.13
ρ_{grout}	1.52	1.59	1.56	0.03
$\%Sand$	0.01	39.00	30.88	14.40
$\%Silt$	33.00	57.00	38.77	10.22
$\%Clay$	27.00	45.00	29.59	4.72
$\%OM$	1.80	8.30	6.77	2.73

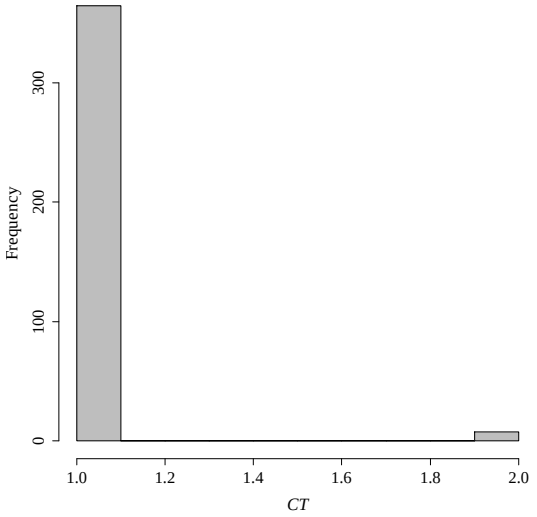
Continued on next page

Table A.5 – continued from previous page

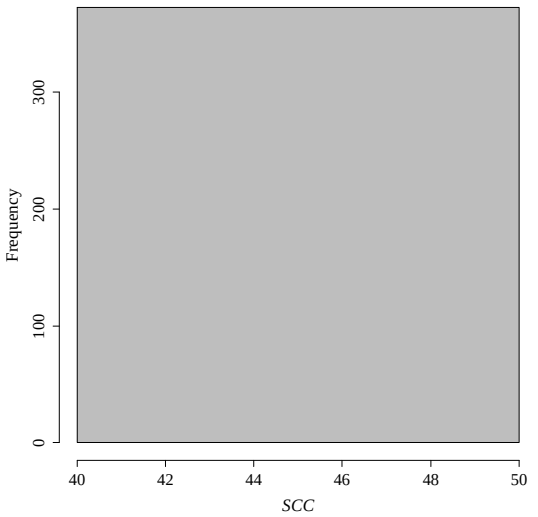
Variable	Minimum	Maximum	Mean	Standard Deviation
H (m)	7.17	31.60	22.81	5.15
JS	1.00	3.00	2.07	0.46
WS (cm/min)	6.00	20.87	11.17	3.72
rpm	3.00	10.00	5.51	1.54
WT (s)	11.50	60.00	32.52	10.78
$Step$ (cm)	4.00	6.00	5.49	0.87
FR (l/min)	139.00	432.00	356.09	89.64
D_{grout} (mm)	4.00	7.00	4.67	0.67
N_{Dgrout}	1.00	2.00	1.72	0.45
D_{water} (mm)	0.00	5.00	0.28	1.14
P_{grout} (bar)	140.00	450.00	359.75	100.97
P_{air} (bar)	0.00	10.00	8.81	2.84
P_{water} (bar)	0.00	400.00	56.84	139.85
Imp_{grout} (kg)	58.06	278.95	215.85	81.79
E (GPa)	0.06	3.88	1.16	0.88



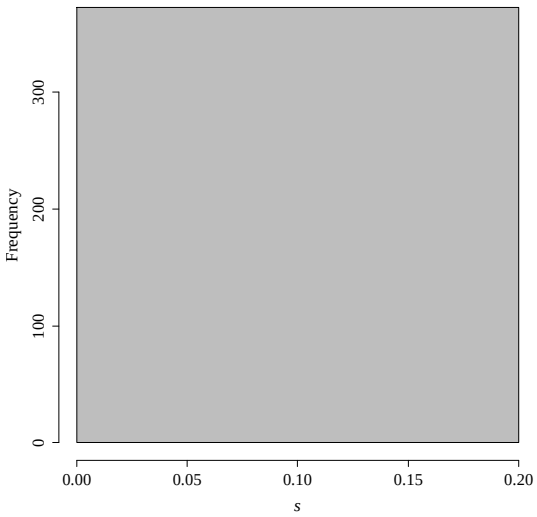
(1) Histogram of W/C



(2) Histogram of CT

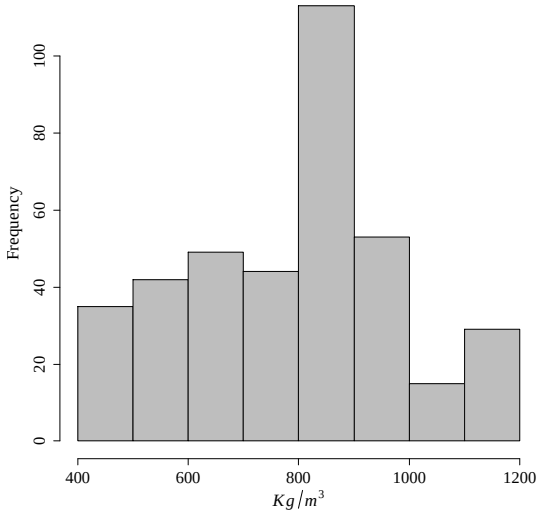


(3) Histogram of SCC

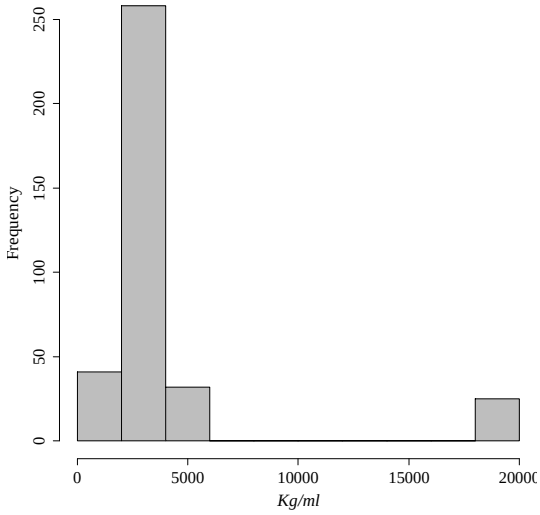


(4) Histogram of s

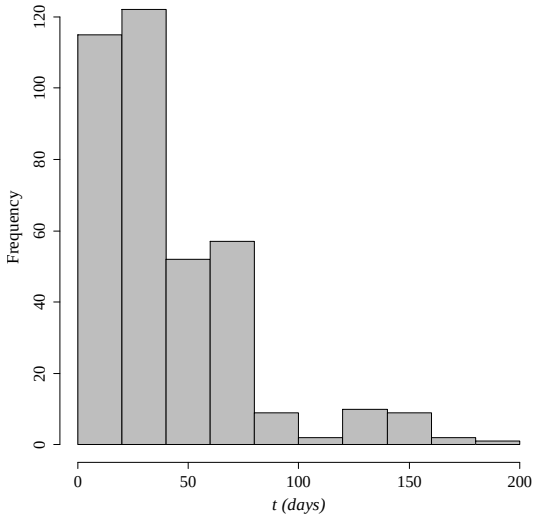
Figure A.5: Histograms of the numeric variables used in E_0 study of *soilcrete* mixtures



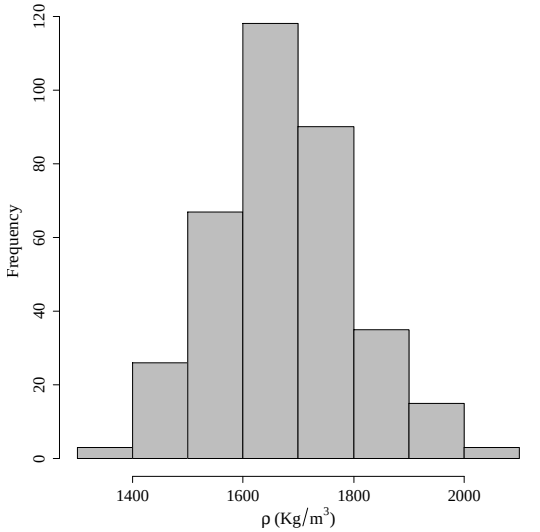
(5) Histogram of kg/m^3



(6) Histogram of kg/ml

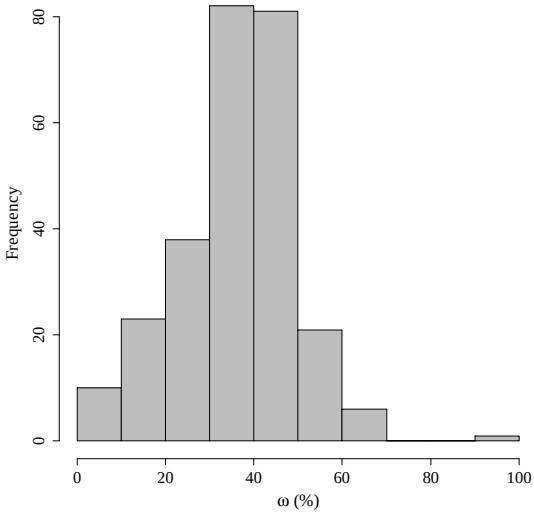


(7) Histogram of t

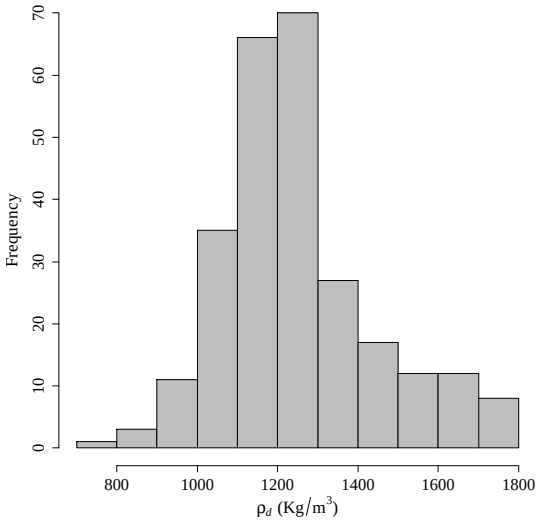


(8) Histogram of ρ

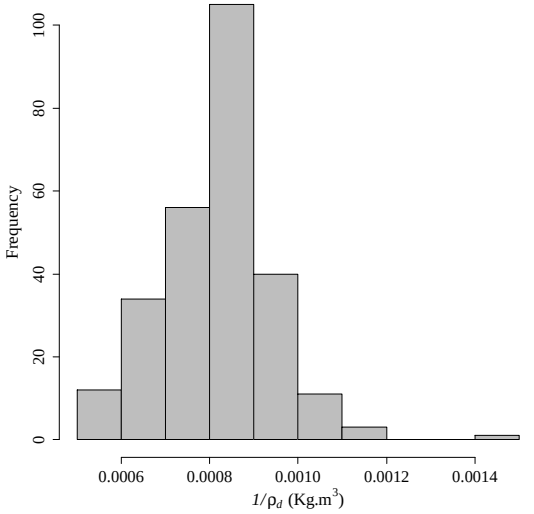
Figure A.5: Histograms of the numeric variables used in E_0 study of *soilcrete* mixtures (cont'd)



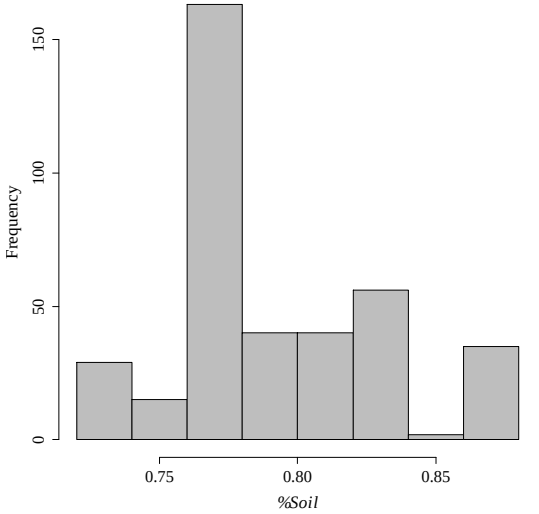
(9) Histogram of ω



(10) Histogram of ρ_d

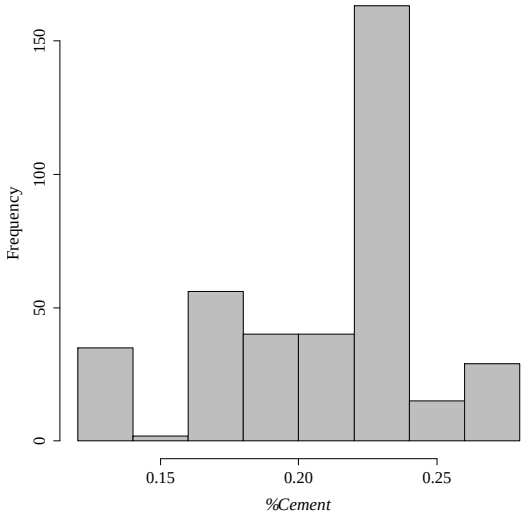


(11) Histogram of $1/\rho_d$

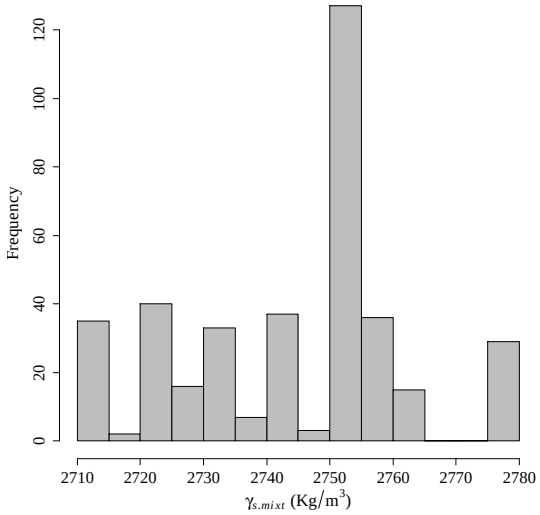


(12) Histogram of $\%Soil$

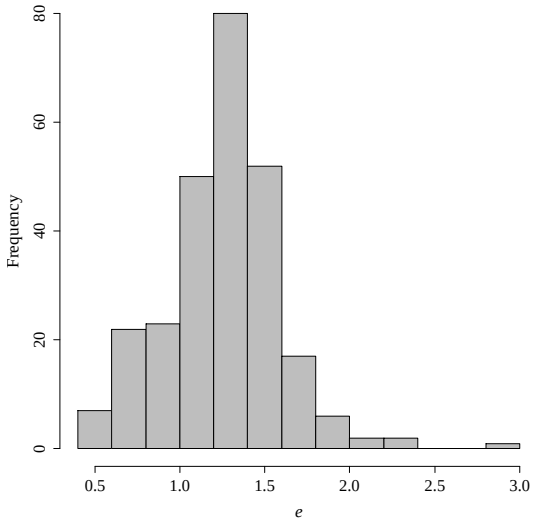
Figure A.5: Histograms of the numeric variables used in E_0 study of *soilcrete* mixtures (cont'd)



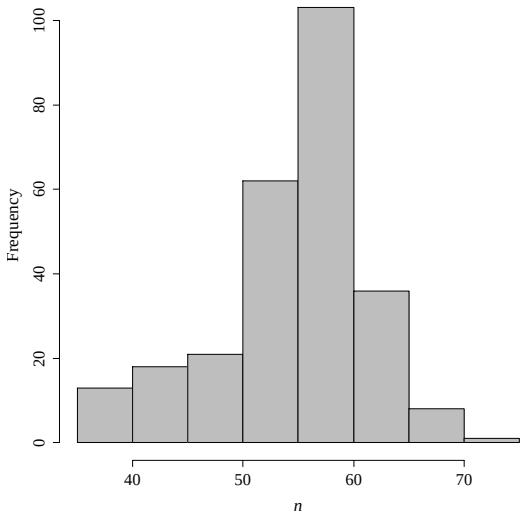
(13) Histogram of $\%Cement$



(14) Histogram of $\gamma_{s.mixt}$

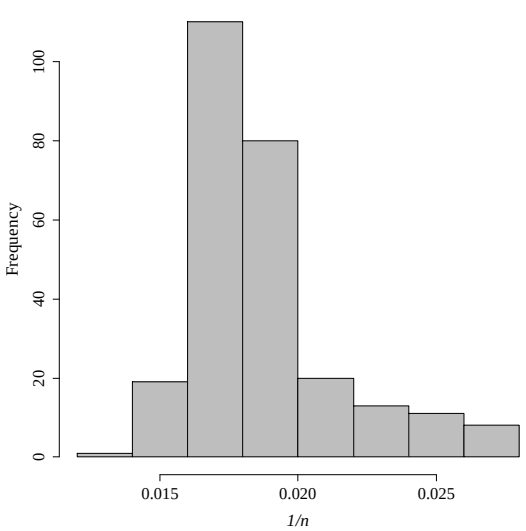


(15) Histogram of e

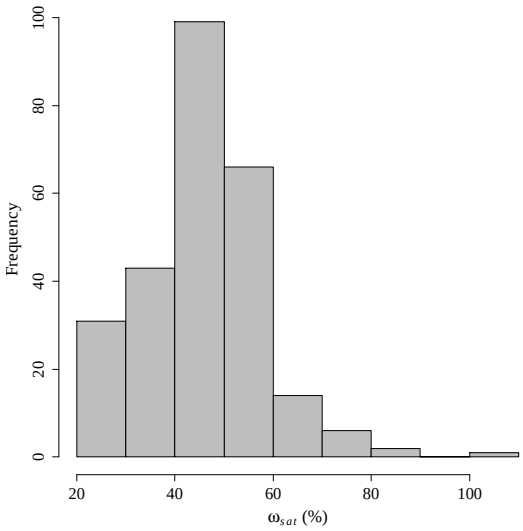


(16) Histogram of n

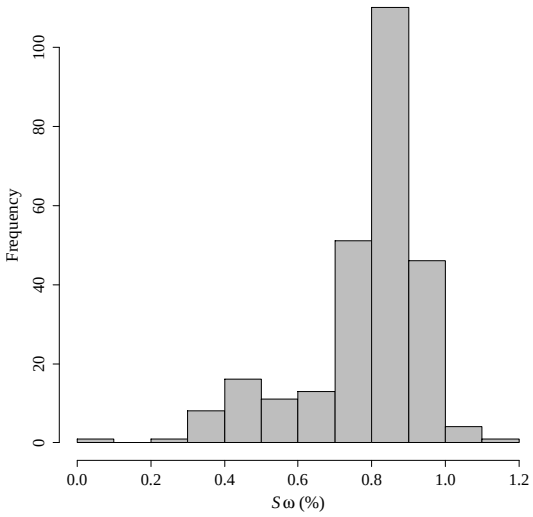
Figure A.5: Histograms of the numeric variables used in E_0 study of *soilcrete* mixtures (cont'd)



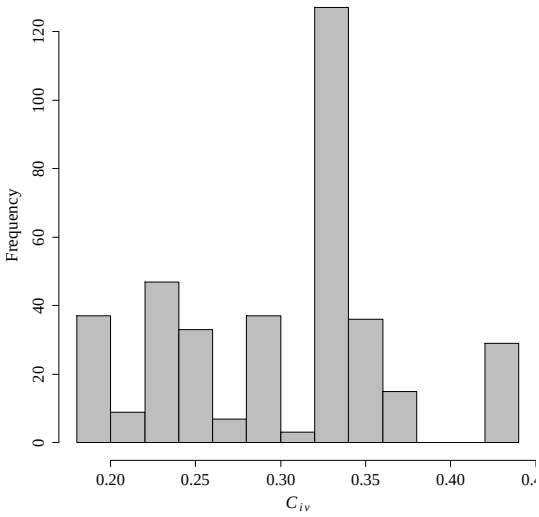
(17) Histogram of $1/n$



(18) Histogram of ω_{sat}

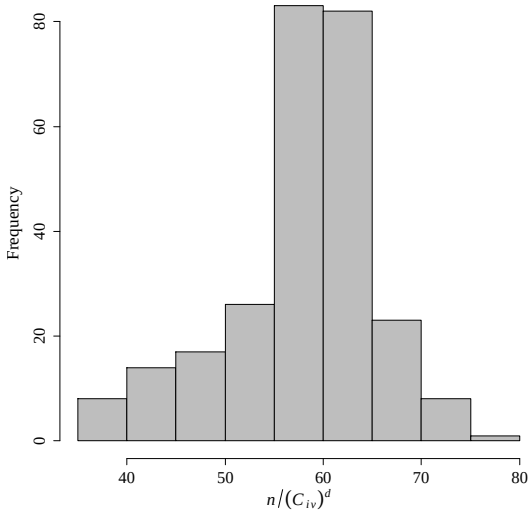


(19) Histogram of S_ω

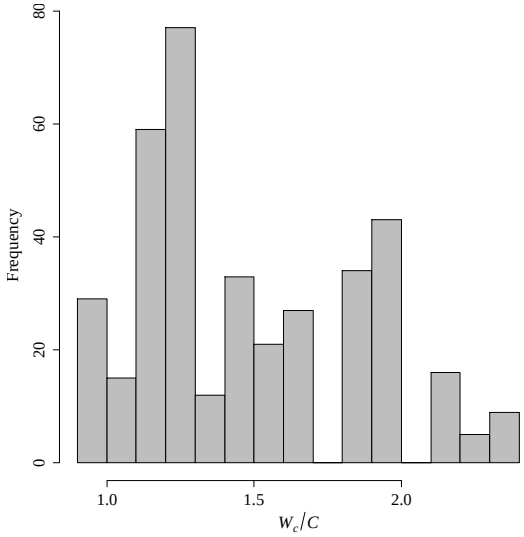


(20) Histogram of C_{iv}

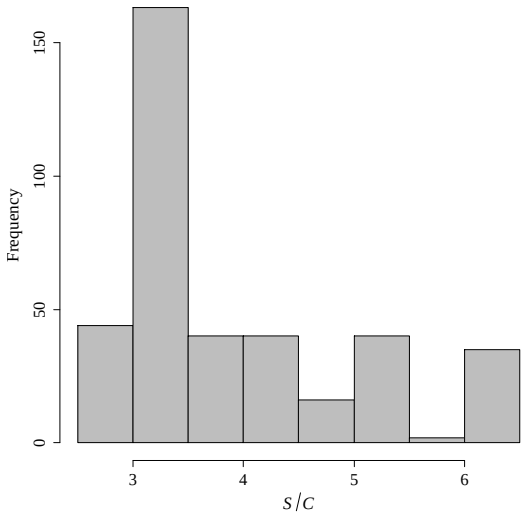
Figure A.5: Histograms of the numeric variables used in E_0 study of *soilcrete* mixtures (cont'd)



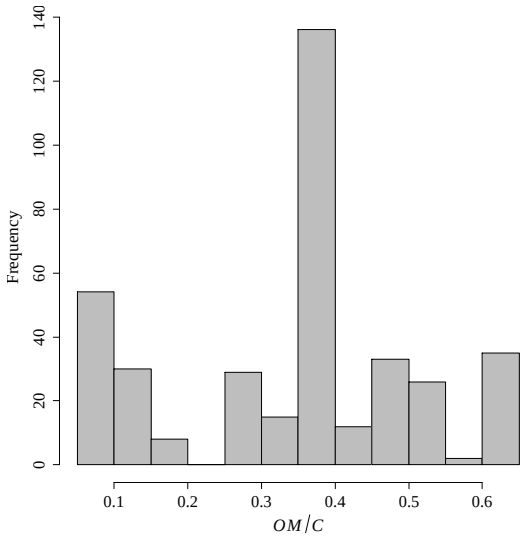
(21) Histogram of $n/(C_{iv})^d$



(22) Histogram of W_c/C

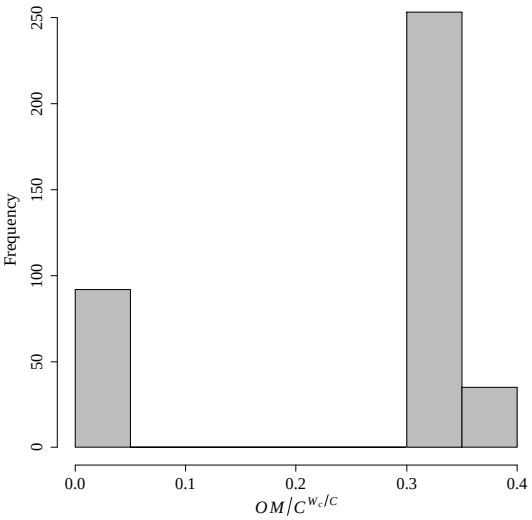


(23) Histogram of S/C

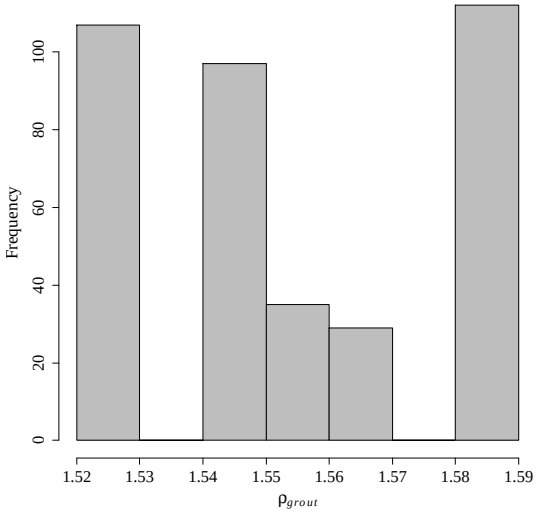


(24) Histogram of OM/C

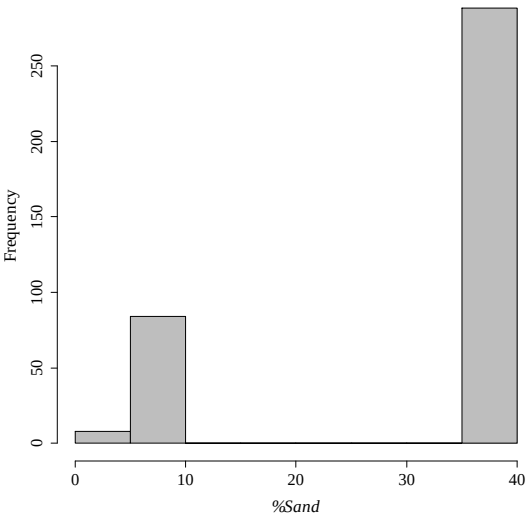
Figure A.5: Histograms of the numeric variables used in E_0 study of *soilcrete* mixtures (cont'd)



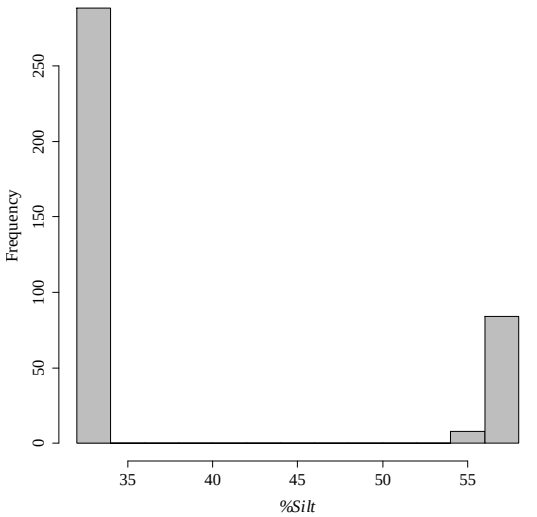
(25) Histogram of $OM/C^{W_c/C}$



(26) Histogram of ρ_{grout}

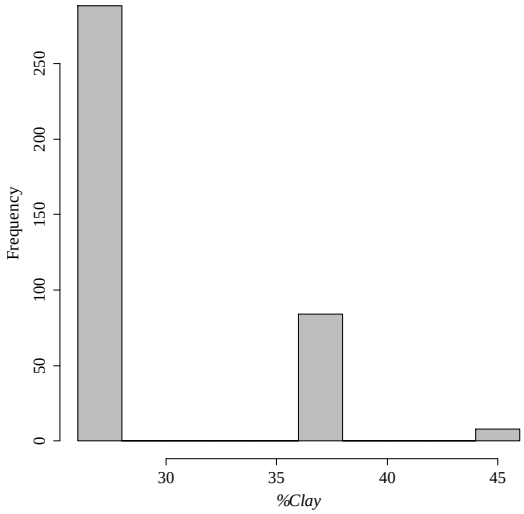


(27) Histogram of $\%Sand$

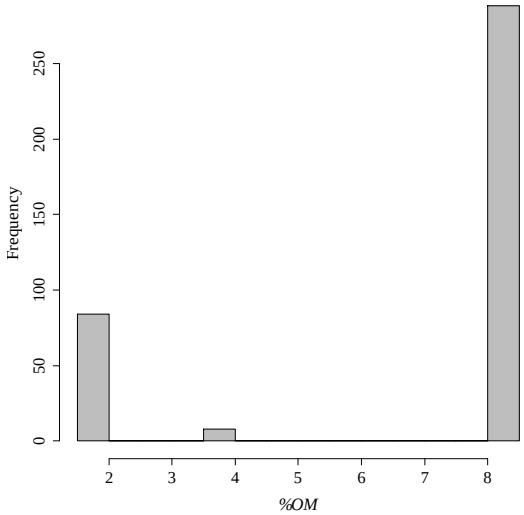


(28) Histogram of $\%Silt$

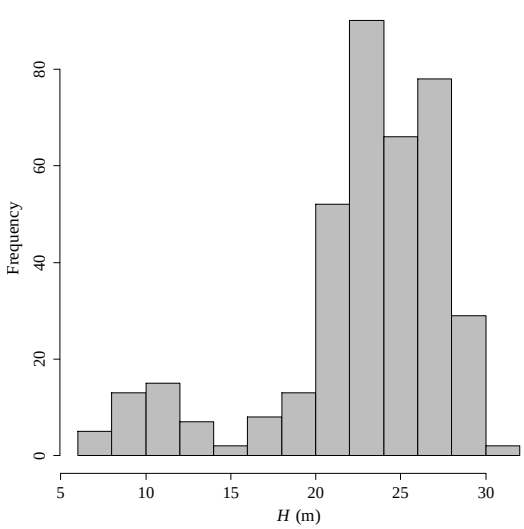
Figure A.5: Histograms of the numeric variables used in E_0 study of *soilcrete* mixtures (cont'd)



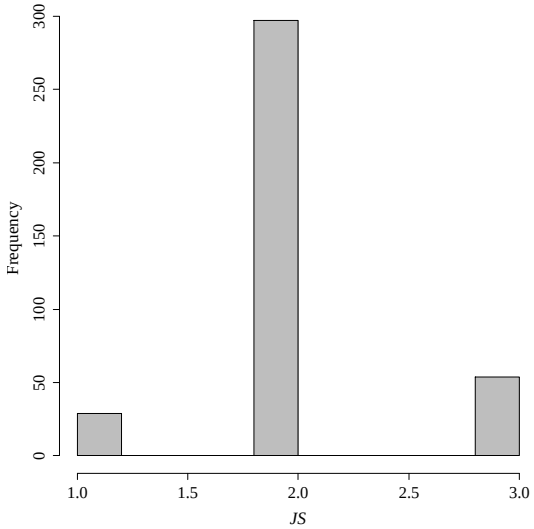
(29) Histogram of $\%Clay$



(30) Histogram of $\%OM$

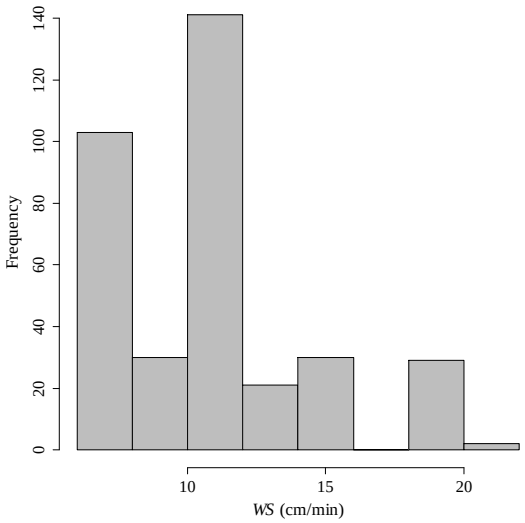


(31) Histogram of H

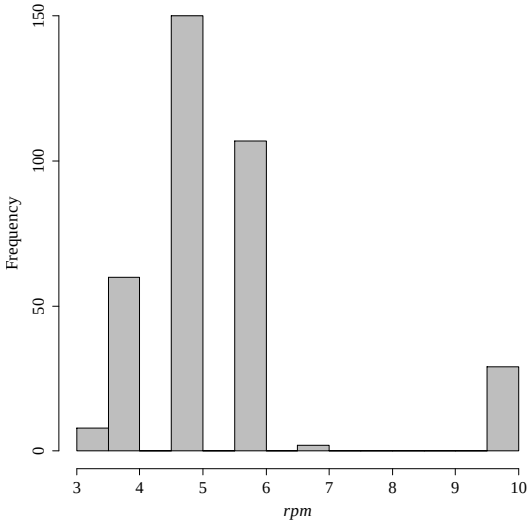


(32) Histogram of JS

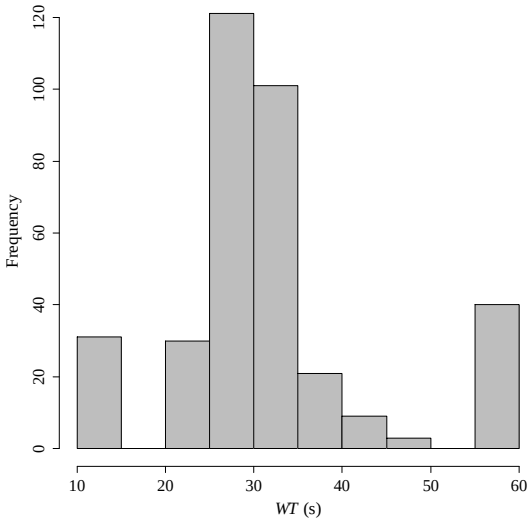
Figure A.5: Histograms of the numeric variables used in E_0 study of *soilcrete* mixtures (cont'd)



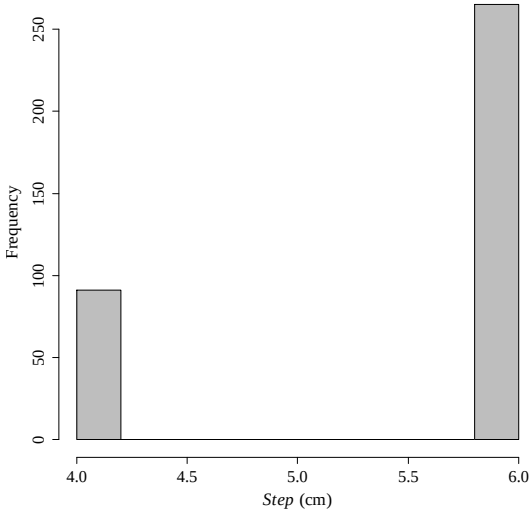
(33) Histogram of WS



(34) Histogram of rpm

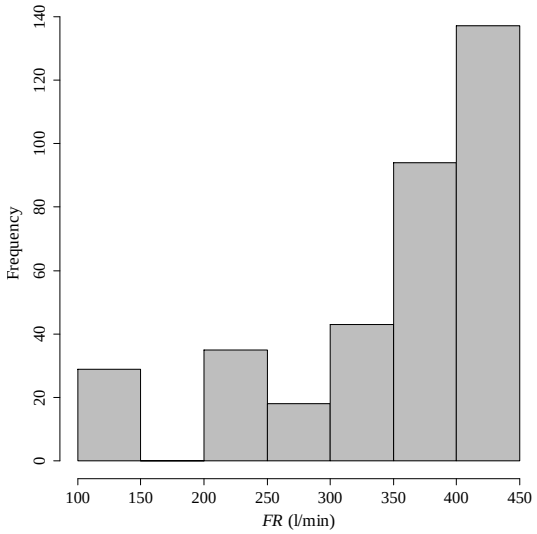


(35) Histogram of WT

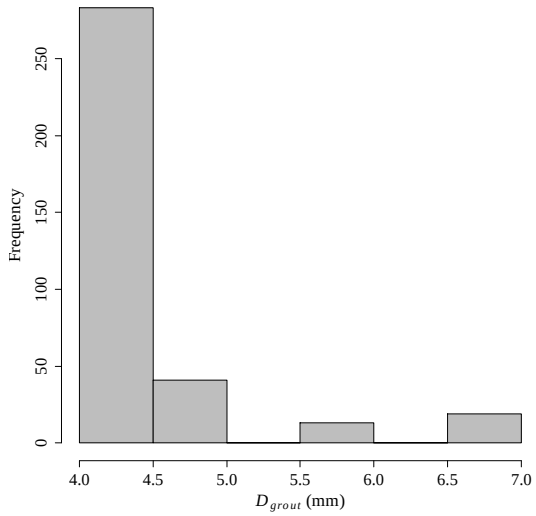


(36) Histogram of $Step$

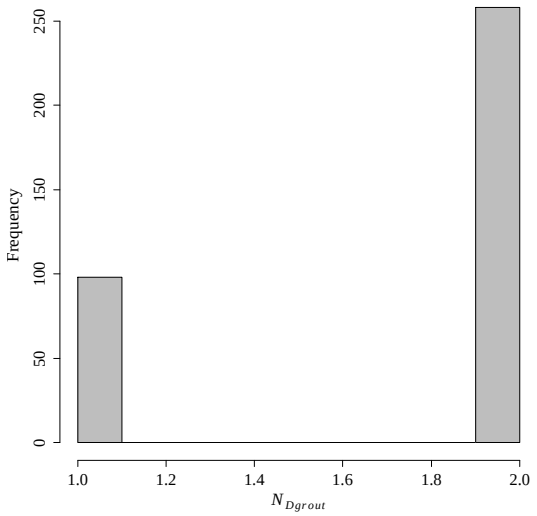
Figure A.5: Histograms of the numeric variables used in E_0 study of *soilcrete* mixtures (cont'd)



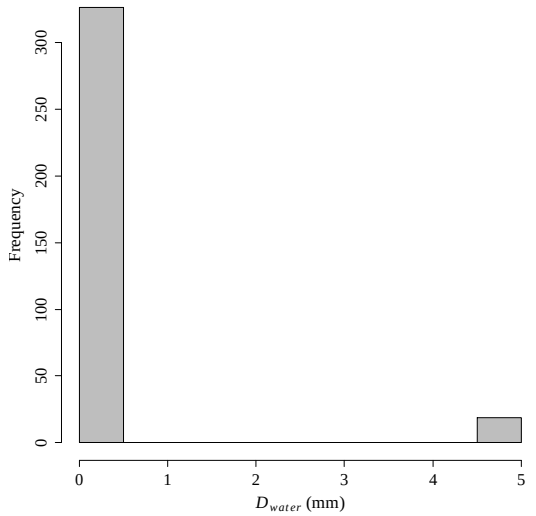
(37) Histogram of FR



(38) Histogram of D_{grout}

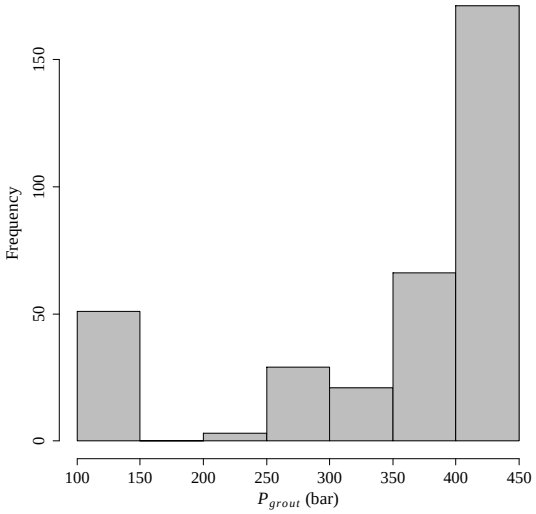


(39) Histogram of N_{Dgrout}

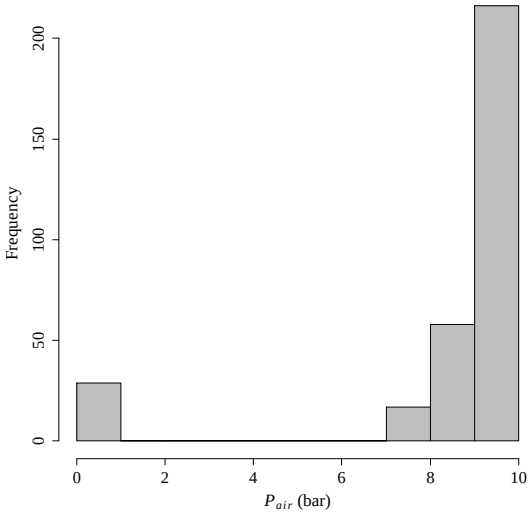


(40) Histogram of D_{water}

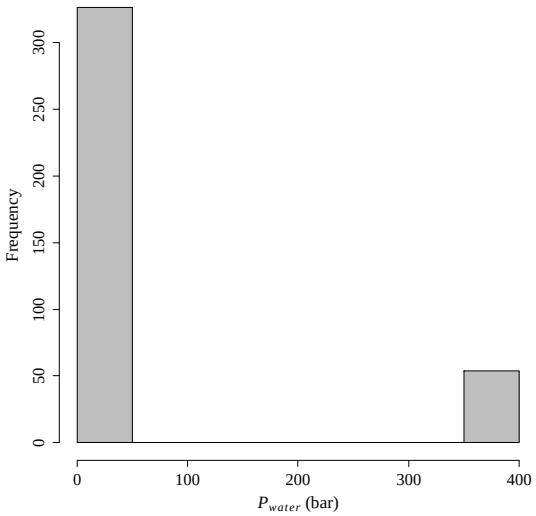
Figure A.5: Histograms of the numeric variables used in E_0 study of *soilcrete* mixtures (cont'd)



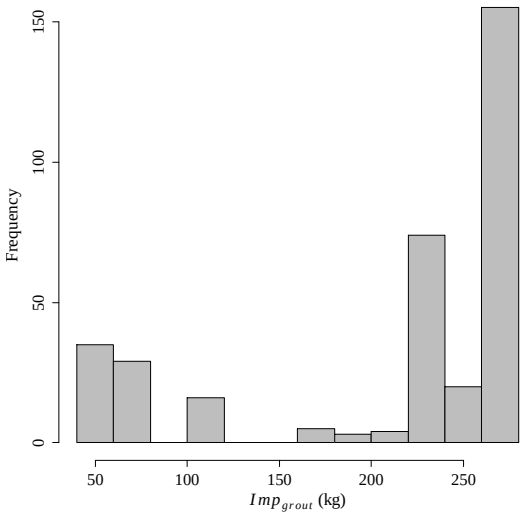
(41) Histogram of P_{grout}



(42) Histogram of P_{air}



(43) Histogram of P_{water}



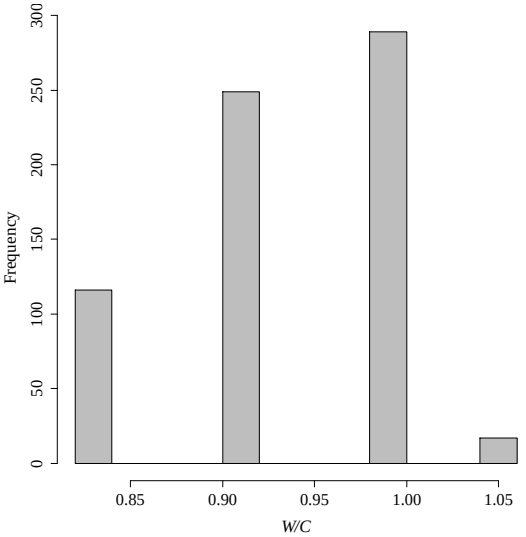
(44) Histogram of Imp_{grout}

Figure A.5: Histograms of the numeric variables used in E_0 study of *soilcrete* mixtures (cont'd)

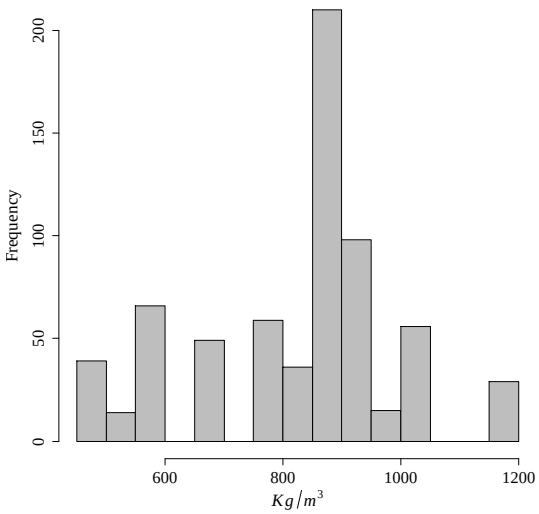
A.2.3 Main statistics and histograms for *D* study

Table A.6: Summary of the input and output variables of database used in D study

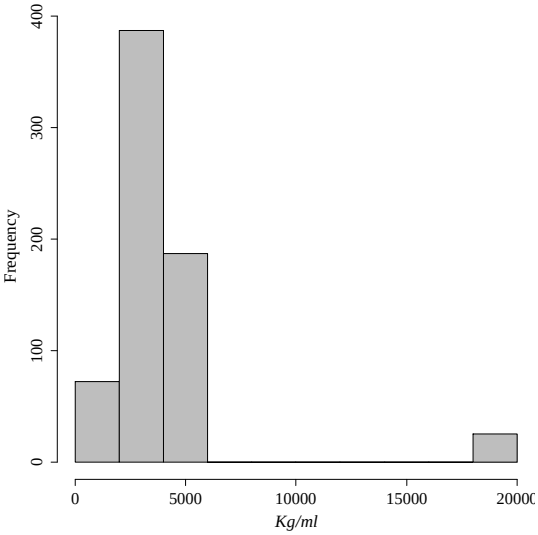
Variable	Minimum	Maximum	Mean	Standard Deviation
W/C	0.83	1.05	0.94	0.07
kg/m^3	492.00	1194.00	840.69	172.06
kg/ml	600.00	18885.00	3826.09	3165.90
ρ_{grout}	1.52	1.59	1.54	0.03
$\%Sand$	0.01	39.00	23.06	16.78
$\%Silt$	33.00	57.00	43.85	11.05
$\%Clay$	22.50	45.00	32.58	7.22
$\%OM$	0.40	8.30	5.31	3.12
JS	1.00	3.00	2.04	0.36
WS (cm/min)	6.00	21.82	10.05	4.10
rpm	3.00	10.00	4.75	1.51
WT (s)	11.00	60.00	38.28	13.63
$Step$ (cm)	4.00	6.00	5.59	0.81
FR (l/min)	139.00	577.89	366.24	77.64
D_{grout} (mm)	4.00	7.00	4.91	0.76
N_{Dgrout}	1.00	2.00	1.63	0.48
D_{water}	0.00	5.00	0.16	0.88
P_{grout} (bar)	140.00	450.00	355.82	85.24
P_{air} (bar)	0.00	10.00	9.16	2.13
P_{water} (bar)	0.00	400.00	36.42	115.16
Imp_{grout} (kg)	58.06	278.95	213.64	69.02
D (mm)	800.00	3008.00	2180.44	420.28



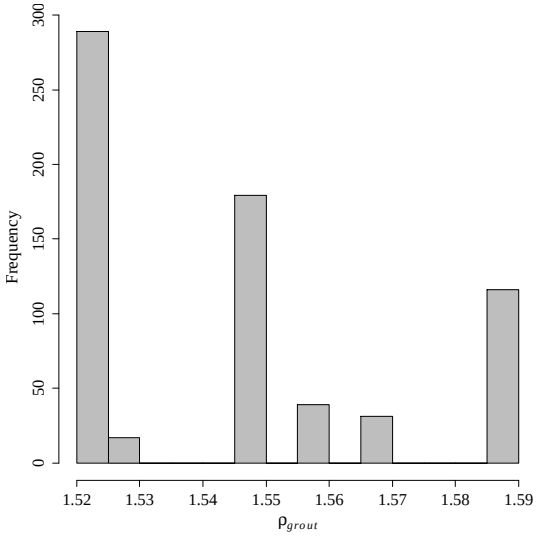
(1) Histogram of W/C



(2) Histogram of kg/m^3

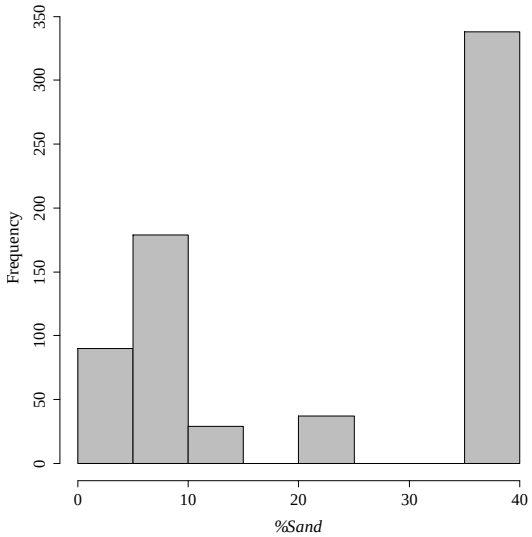


(3) Histogram of kg/ml

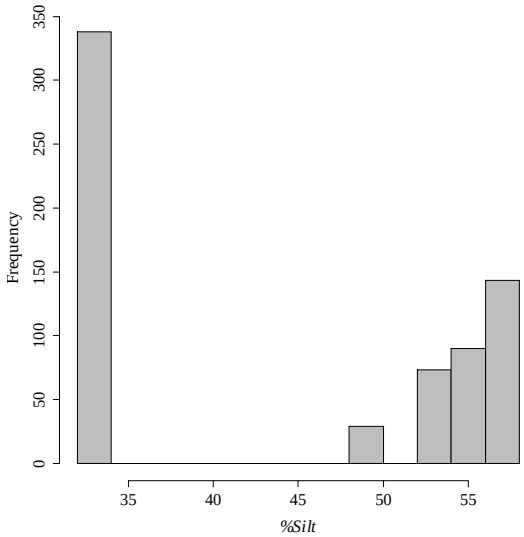


(4) Histogram of ρ_{grout}

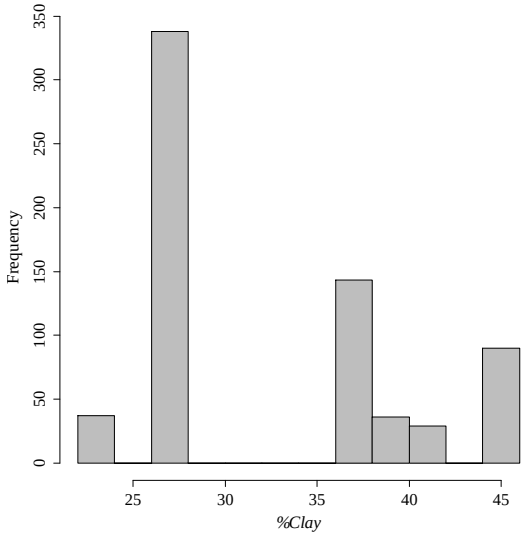
Figure A.6: Histograms of the numeric variables used in D study



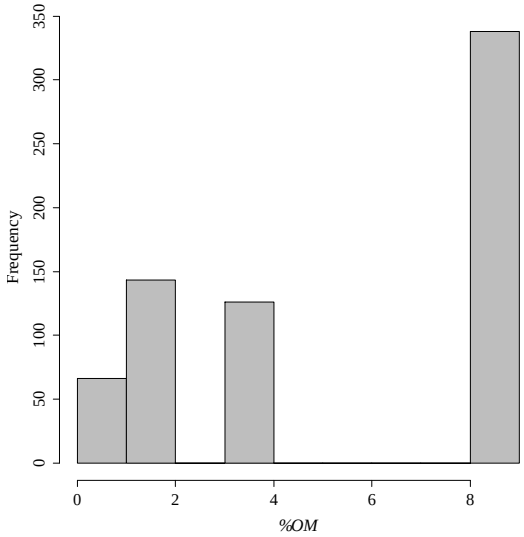
(5) Histogram of $\%Sand$



(6) Histogram of $\%Silt$

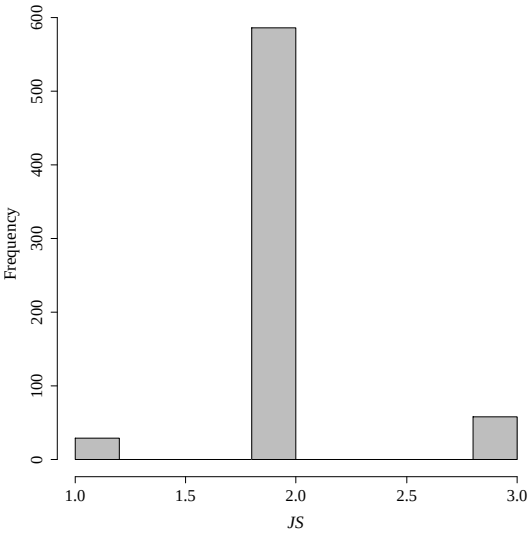


(7) Histogram of $\%Clay$

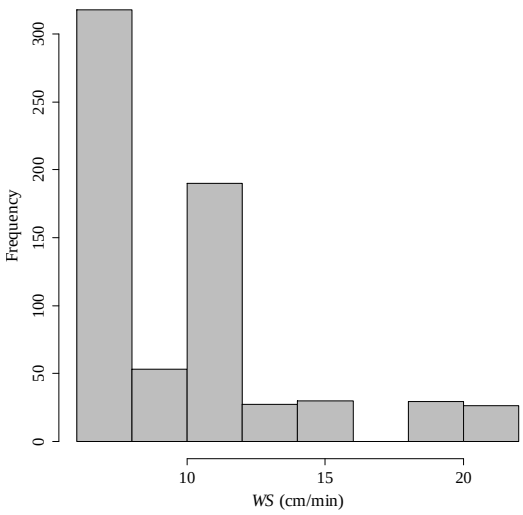


(8) Histogram of $\%OM$

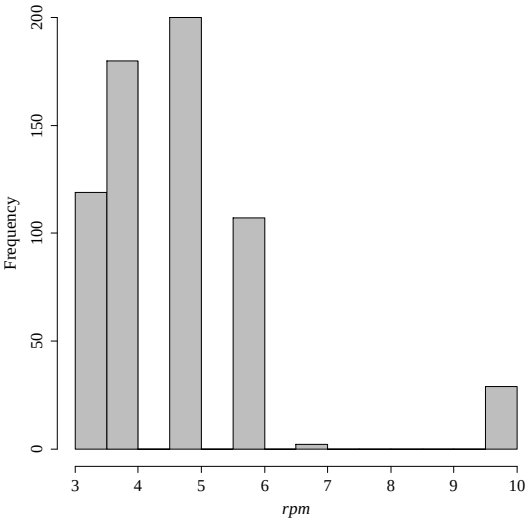
Figure A.6: Histograms of the numeric variables used in D study (cont'd)



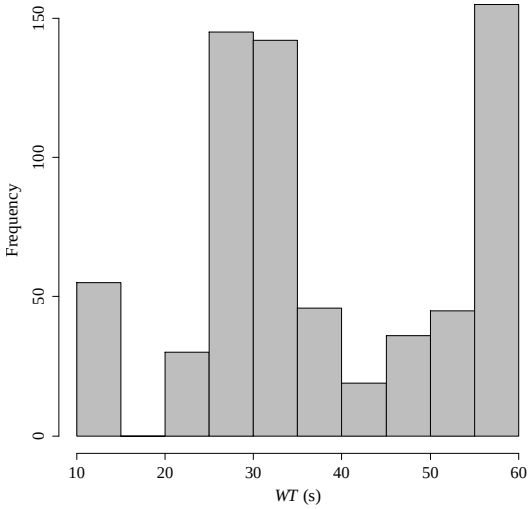
(9) Histogram of JS



(10) Histogram of WS

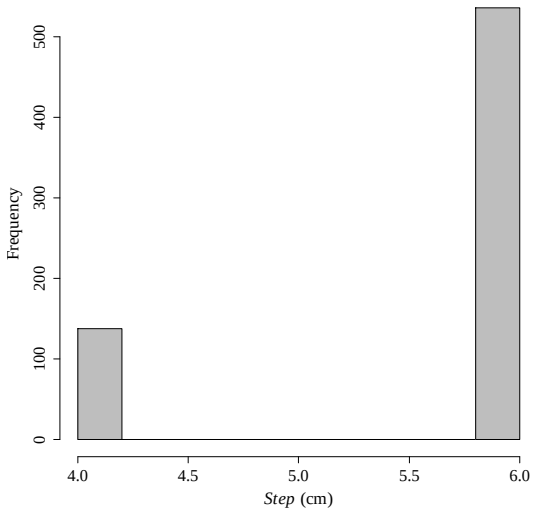


(11) Histogram of rpm

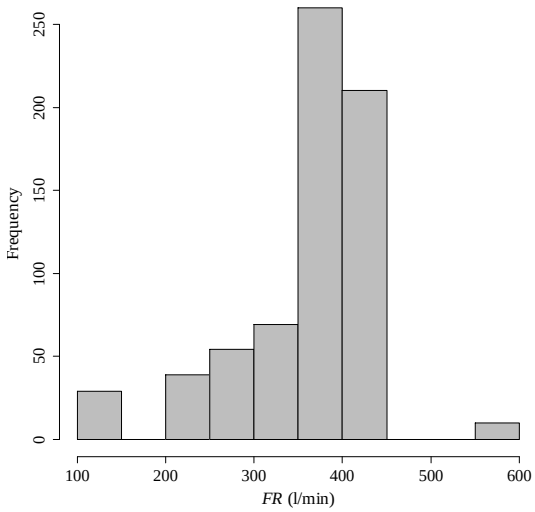


(12) Histogram of WT

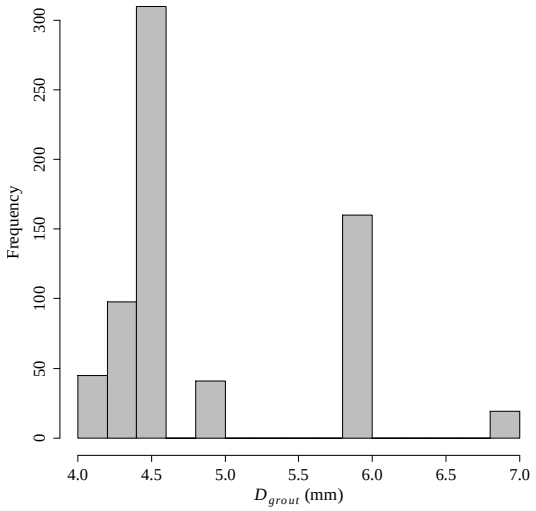
Figure A.6: Histograms of the numeric variables used in D study (cont'd)



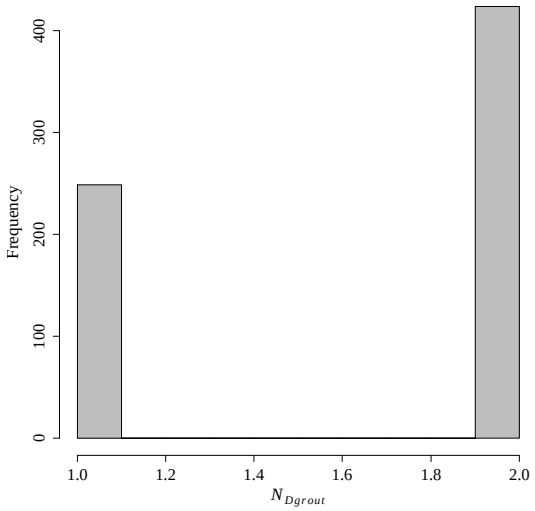
(13) Histogram of $Step$



(14) Histogram of FR

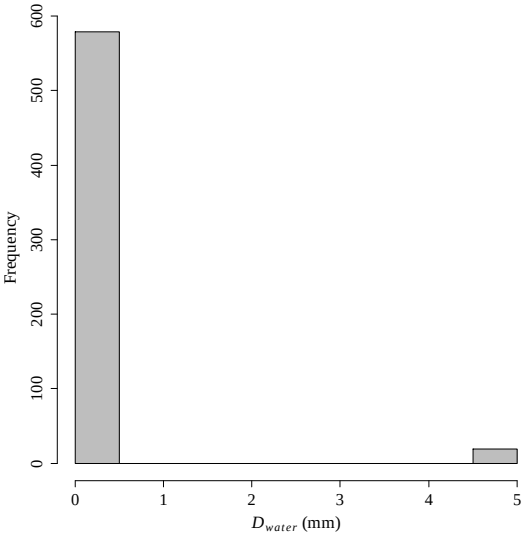


(15) Histogram of D_{grout}

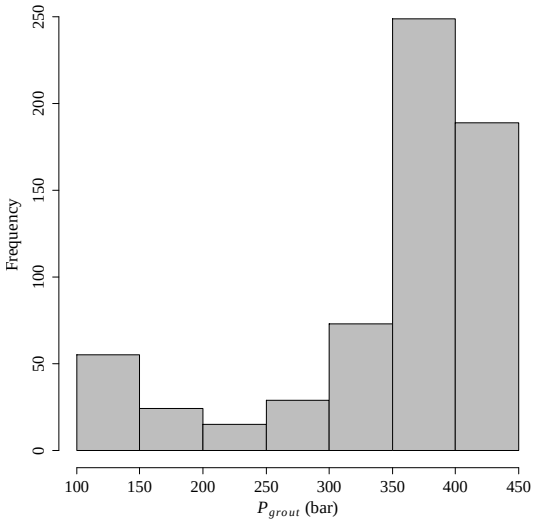


(16) Histogram of N_{Dgrout}

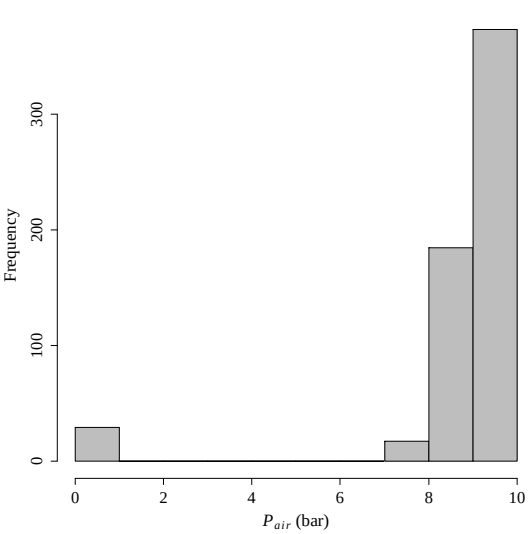
Figure A.6: Histograms of the numeric variables used in D study (cont'd)



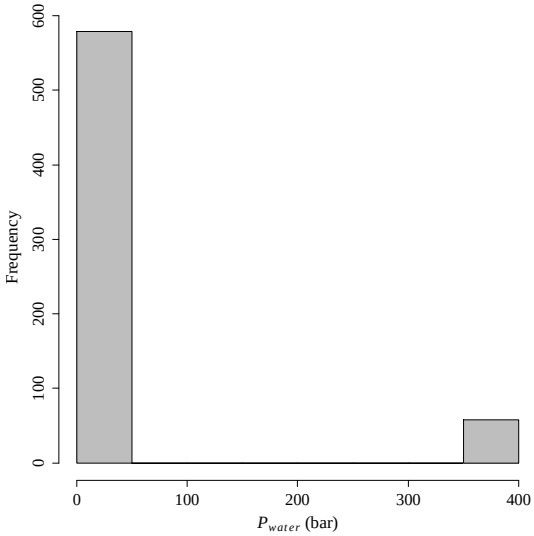
(17) Histogram of D_{water}



(18) Histogram of P_{grout}

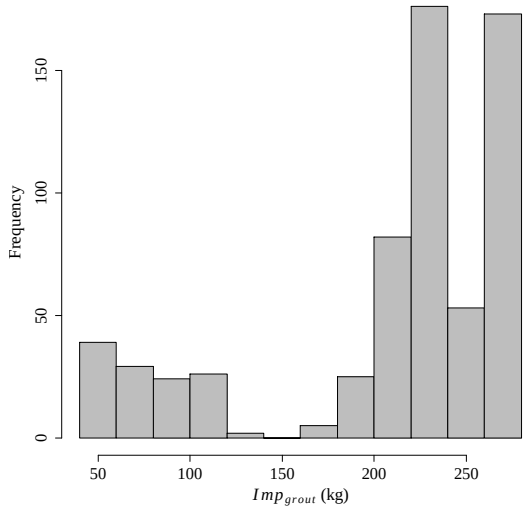


(19) Histogram of P_{air}



(20) Histogram of P_{water}

Figure A.6: Histograms of the numeric variables used in D study (cont'd)



(21) Histogram of Imp_{grout}

Figure A.6: Histograms of the numeric variables used in D study (cont'd)

This page was intentionally left blank.

Appendix B

Mathematical expressions for input variables calculation

Some of the variables used as input in the present research work were not measured or experimentally quantified but calculated according to a given mathematical expression. Following are present are presented the mathematical expressions used for calculate some of the input variables used in this work.

– **Dry density of the mixture** — ρ_d ($\mathrm{kg \cdot m^{-3}}$):

$$\rho_d = \frac{\rho}{1 + \omega/100} \tag{B.1}$$

where ρ is the natural density of the mixture ($\mathrm{kg/m^3}$) and ω is the water content of the mixtures in percentage.

– **Unite weight of the mixture** — $\gamma_{s.mixt}$ ($\mathrm{kg \cdot m^{-3}}$):

$$\gamma_{s.mixt} = G_s^{mixt} \times \gamma_w \qquad \text{where} \qquad G_s^{mixt} = \frac{\%soil}{100} \times G_s + \frac{\%Cement}{100} \times c \tag{B.2}$$

where $\gamma_w = 1000\,\mathrm{kg \cdot m^{-3}}$, $G_s = 2.65$ and $c = 3.1$.

– **Void ratio of the mixture** — e:

$$e = \frac{\gamma_{s.mixt} - \rho_d}{rho_d} \tag{B.3}$$

– **Mixture porosity** — η:

$$\eta = \frac{e}{1+e} \tag{B.4}$$

– **Saturated water content** — ω_{sat} (%):

$$\omega_{sat} = \frac{e}{G_s^{mixt}} \times 100 \qquad \text{where} \qquad G_s^{mixt} = \frac{\%soil}{100} \times G_s + \frac{\%Cement}{100} \times c \tag{B.5}$$

where G_s and c take the values of 2.65 and 3.1 respectively.

– **Degree of saturation** — S_ω:

$$S_\omega = \frac{\omega}{\omega_{sat}} \tag{B.6}$$

– **Volumetric content of cement** — C_{iv}:

$$C_{iv} = \frac{\frac{\frac{\%Cement}{100} \times \rho_d \times \frac{V_{sample}}{1000000}}{3100}}{\left(\frac{\frac{\%Soil}{100} \times \rho_d \times \frac{V_{sample}}{1000000}}{G_s \times \gamma_w}\right) + \left(\frac{\frac{\%Cement}{100} \times \rho_d \times \frac{V_{sample}}{1000000}}{3100}\right)} \tag{B.7}$$

where $G_s = 2.65$, $\gamma_w = 1000\,\mathrm{kg} \cdot \mathrm{m}^{-3}$ and V_{sample} is the volume of the sample in cm^3.

– **Grout impact** — Imp_{grout} (kg):

$$Imp_{grout} = 2 \times \frac{\pi \times D_{grout}^2 \times N_{grout}}{4} \times P_{grout} \tag{B.8}$$

where D_{grout} is mean diameter of grout nozzles in meters and P_{grout} is the grout pressure in MPa.