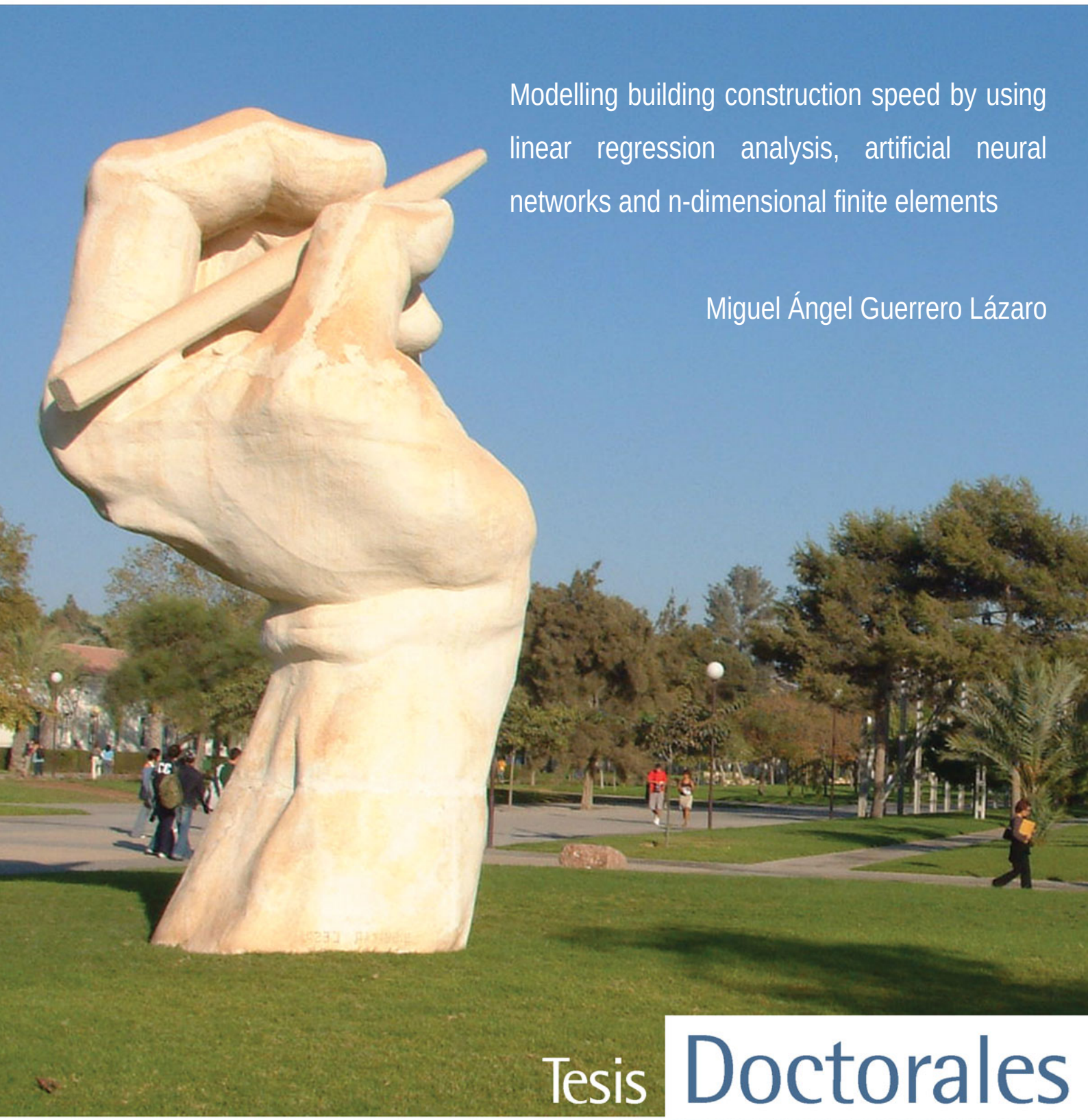
Universitat d'Alacant
Universidad de Alicante



Modelling building construction speed by using linear regression analysis, artificial neural networks and n-dimensional finite elements

Miguel Ángel Guerrero Lázaro

Tesis Doctorales

www.eltallerdigital.com

UNIVERSIDAD de ALICANTE

Universitat d'Alacant
Universidad de Alicante

Departamento de Edificación y Urbanismo
Escuela Politécnica Superior

Modelling Building Construction Speed by Using Linear Regression Analysis, Artificial Neural Networks and *n*-Dimensional Finite Elements

PhD Thesis

Author:
Miguel A. Guerrero Lázaro

Supervisors:
Dr. Andrés Montoyo Guijarro // Dr. Yolanda Villacampa Esteve

January 2016

This research work has been funded by a research grant awarded by the Office of the Vice President for Research, Development and Innovation of the University of Alicante (resolution 22 December 2011).

A mi esposa Loli
y a mi hijo David

"A theory has only the alternative of being right or wrong. A model has a third possibility: it may be right, but irrelevant."

Manfred Eigen (1927-). The Physicist's Conception of Nature, 1973.

Acknowledgements

Aunque la culminación de una tesis es el resultado de un proceso de documentación y experimentación por parte del autor, la misma no sería posible sin la colaboración y ayuda, de una u otra manera, por parte de otras personas. Aunque no resulta nada sencillo expresar de forma apropiada y resumida el sentimiento de gratitud infinita que tengo hacia estas personas, trataré de hacerlo lo mejor posible en las siguientes líneas.

En primer lugar, nunca podré agradecerle lo suficiente a mi director de tesis Andrés Montoyo la oportunidad que me brindó para obtener la beca predoctoral que ha hecho posible el desarrollo de esta tesis. Gracias Andrés por la confianza que depositaste en mí y por todos los buenos consejos que me has dado a lo largo de estos últimos cuatro años para que la investigación que he llevado a cabo llegue a buen puerto.

Si Andrés ha sido un elemento clave para poder culminar esta tesis que decir de Yolanda Villacampa, mi otra directora. No puedo más que expresar mi más profundo agradecimiento y reconocimiento a su inagotable paciencia y total dedicación para resolver mis interminables dudas sobre todo tipo de conceptos matemáticos y estadísticos utilizados durante la investigación. La aportación de su conocimiento como experta en el campo de la modelización matemática ha sido esencial para poder finalizar con éxito la tesis que ahora presento.

Es necesario dar también las gracias a la empresa Soft SA y al Servicio Murciano de Salud (en especial a ti José Antonio) por facilitarme los datos que han hecho posible el desarrollo de esta tesis.

Quiero asimismo expresar mi gratitud a la Universidad de Alicante tanto por la beca predoctoral que me concedió como por la ayuda económica que me proporcionó para poder llevar a cabo una estancia de investigación en la Universidad de Gante (Bélgica).

Por supuesto, no puedo dejar de agradecer a todo el personal que forma parte del Departamento de Lenguajes y Sistemas Informáticos, y en especial a los miembros del Grupo de Investigación en Procesamiento del Lenguaje Natural

y Sistemas de Información, por su total predisposición para ayudarme siempre que lo he necesitado, brindándome su tiempo, sus ideas y su compañía en esos días donde la tesis parecía no tener fin y todo lo veía de color negro. Ha sido un placer y un privilegio poder trabajar con las personas que forman parte este grupo de investigación porque de todas ellas he aprendido algo.

En el capítulo personal, agradezco también a mis padres el esfuerzo que realizaron durante tantos años para que sus hijos disfrutaran de oportunidades que ellos nunca tuvieron, y sin las cuales yo nunca hubiera podido seguir el camino que me ha llevado hasta la culminación de esta tesis. Y tampoco me puedo olvidar de mis hermanos, Eva y Antonio, porque aunque los veo con menos frecuencia de la que desearía, sé que están ahí cuando los necesito.

Por último, no encuentro palabras suficientes para poder expresarle como realmente se merece mi enorme gratitud a Loli, mi mujer, tanto por su apoyo incondicional a mi labor investigadora como por la comprensión y la paciencia que ha tenido conmigo durante los últimos cuatro años, sin las que no hubiera sido capaz de finalizar esta tesis. Ella, sin lugar a dudas, se ha llevado la peor parte de todos los sacrificios que ha supuesto la consecución de este trabajo de investigación, porque convivir con un marido que trata de finalizar una tesis contra el reloj es un "castigo" difícil de soportar. Además, durante el transcurso de la tesis llegó a este mundo nuestro hijo David, que si bien ha supuesto un motivo de inmensa alegría para nuestras vidas, y una constante motivación para tratar de mejorar tanto en lo profesional como en lo personal, también ha requerido de un mayor esfuerzo para poder compaginar su cuidado y educación con nuestra vida laboral, y de nuevo ha sido Loli quien ha asumido con abnegación y sacrificio una mayor carga de responsabilidad en estas tareas. Solo le pido a Dios que me permita en esta vida devolverles con creces toda la atención y el cariño que se merecen y no he podido dedicarles durante el transcurso de esta tesis.

Alicante, enero de 2016

Miguel A. Guerrero

Abstract

The estimation of the time required to construct building projects has been a topic of great interest to many researchers and practitioners. Delays are a common problem in the construction industry and may be motivated by different factors. In this context, prediction of the construction time of building projects at early project phases has been considered a key element for project success. Initially, the construction time of a building is affected by several factors related to project features, although some factors are more crucial than others. Based on these factors, for the purpose of providing proper tools to estimate construction time and minimise the subjectivity in such estimation, to date, most research works have presented parametric models which were built using linear regression analysis (LRA). Nevertheless, there is an increasing trend for using artificial neural networks (ANNs) to develop better predictive models.

In order to produce the best possible predictive models and provide a clearer explanation regarding the relationships that exist between different project scope factors and the construction time of new builds, the research work presented in this thesis used two data sets and three different modelling techniques: LRA, ANNs and a new numerical methodology based on the finite element method (FEM). In particular, this thesis addressed the general assumption that nonlinear modelling techniques are likely to better represent the previously mentioned relationships than LRA. According to available data, predictor variables related to construction costs, gross floor area (GFA), number of floors, and the type of facility were selected to analyse their influence on the duration of the construction process. Additionally, and since that there is no general agreement in the literature regarding which is the most appropriate dependent variable for predicting construction time, both time and speed of construction were analysed to determine which of these offer better predictive models. In this regard, construction speed can be used as a useful and robust benchmark for comparison of contractor performance.

In the case of ANNs, two different types of network architectures were tested: the multilayer perceptron (MLP) and the radial basis function (RFB). The modelling process of MLP networks was divided into five stages: (i) selection of the training methodology, (ii) data division, (iii) design of the initial network structure, (iv) network

optimisation, and (v) validation of the optimised models. MLP networks were used in conjunction with two different training algorithms and five options for calibration data division. In addition, a methodology was defined to obtain optimised MLP networks with an adequate predictive performance. This methodology develops a stepwise trial and error procedure in which a basic MLP network structure**,** with enough consistency**,** is first established and subsequently this initial structure is modified at each step of the proposed optimisation process in order to achieve the best possible network configuration.

This thesis also proposes a framework to evaluate the performance of predictive models which includes five different assessment criteria: (i) verification of compliance with the underlying assumptions regarding the statistical procedure used to obtain the models, (ii) checking the goodness of fit of the models to the data set used for generating them, (iii) validation of models in terms of ability to generalise, (iv) assessment of the balance existing between the ability of a model to generalise and the accuracy obtained with the calibration data, and (v) development of a sensitivity analysis to verify model stability. Finally, a sensitivity analysis was also proposed to evaluate the impact of the construction cost variability, caused by the uncertainty in its estimation, on the performance of predictive models.

The results obtained with this thesis showed that construction speed is a more appropriate dependent variable than construction time to develop predictive models to estimate the construction process duration of building projects, and that such construction speed is affected more by GFA than by construction cost. Furthermore, the FEM-based numerical methodology provided better predictive models than those generated by MLP networks and LRA. In this regard, the findings of this research work support the idea that linear regression models can provide a good starting point from which to search for better predictive models using nonlinear modelling techniques.

The knowledge gained from this thesis will allow for new approaches to be explored in order to better determine the relationships existing between project scope factors and the construction speed of new builds, but always taking into account that the results provided by the models proposed herein are only initial construction speed estimates at early stages of project development, when only basic information is available, and are not intended to replace detailed schedules undertaken by builders.

Resumen

La estimación del tiempo requerido para llevar a cabo un proyecto de edificación ha sido un tema de interés para muchos investigadores y profesionales. Los retrasos son un problema común en la industria de la construcción y pueden estar motivados por causas diversas. En este contexto, la predicción del tiempo de construcción en etapas tempranas del proyecto ha sido considerada un elemento clave para el éxito del mismo. Inicialmente, la duración del proceso de construcción de un proyecto de edificación está afectada por varios factores relacionados con las características del proyecto, aunque algunos factores son más cruciales que otros. Basándose en estos factores, con el propósito de proporcionar herramientas adecuadas para estimar la duración del proceso constructivo y minimizar la subjetivad de tal estimación, la mayoría de los trabajos de investigación desarrollados hasta la fecha han presentado modelos paramétricos que fueron construidos mediante el análisis de regresión lineal (ARL). No obstante, hay una tendencia creciente en los últimos años a utilizar redes neuronales artificiales (RNA) con el propósito de desarrollar modelos con un mayor rendimiento predictivo.

Con el fin de producir mejores modelos predictivos y proporcionar una explicación más clara respecto a las relaciones existentes entre diferentes factores de alcance del proyecto y la duración del proceso constructivo de edificios de nueva planta, el trabajo de investigación presentado en esta tesis ha utilizado dos conjuntos de datos y tres técnicas de modelado diferentes: ARL, RNA y una nueva metodología numérica basada en el método de los elementos finitos (MEF). En particular, esta tesis ha sido desarrollada bajo el supuesto de que técnicas de modelado no lineal podrían representar mejor las mencionadas relaciones que el clásico ARL. De acuerdo con los datos disponibles, como variables predictoras utilizadas para estimar la duración del proceso constructivo se seleccionaron variables relacionadas con los costes de construcción, la superficie construida (SC), el número de plantas y el tipo de instalación. Además, debido a que no existe acuerdo en la literatura relacionada con proyectos de edificación sobre cuál es la variable dependiente más apropiada para predecir la duración del proceso constructivo, con el propósito de determinar que variable proporciona mejores modelos predictivos en esta tesis se utilizaron como variables de respuesta tanto el tiempo como la velocidad de construcción. A este

respecto, la velocidad de construcción puede ser utilizada como un punto de referencia útil y robusto para comparar el rendimiento entre contratistas.

En el caso de las RNA, se probaron dos tipos diferentes de arquitecturas de red: el perceptrón multicapa (PMC) y las RNA de base radial. El proceso de modelado del PMC fue dividido en cinco etapas: (i) selección de la metodología de entrenamiento, (ii) división de los datos, (iii) diseño de la estructura inicial de la red, (iv) optimización de la red y (v) validación de los modelos optimizados. La red PMC fue utilizada junto con dos algoritmos diferentes de entrenamiento y cinco opciones para la división de los datos de calibración. Además, se definió una metodología de optimización con el propósito de obtener redes PMC con un adecuado rendimiento predictivo. La metodología propuesta desarrolla un proceso de ensayo y error por etapas en el que una estructura PMC básica, con suficiente consistencia, se establece primero y después esta estructura inicial es modificada a cada paso del proceso de optimización propuesto para lograr la mejor configuraron posible de la red neuronal.

Esta tesis también propone un marco de evaluación del rendimiento de los modelos predictivos desarrollados que incluye cinco criterios: (i) verificación del cumplimiento de los supuestos subyacentes respecto del procedimiento estadístico utilizado para obtener los modelos, (ii) comprobación de la bondad de ajuste de los modelos a los datos utilizados para su generación, (iii) validación de los modelos en términos de habilidad para generalizar, (iv) evaluación del equilibrio existente entre la habilidad para generalizar de un modelo y la precisión obtenida con los datos de calibración y (v) desarrollo de un análisis de sensibilidad para verificar la estabilidad de los modelos. Finalmente, también se propone un análisis de sensibilidad para evaluar el impacto de la variabilidad de los costes de construcción, causada por la incertidumbre en su estimación, sobre el rendimiento de los modelos predictivos desarrollados.

Los resultados obtenidos con esta tesis muestran que la velocidad de construcción es una variable de respuesta más adecuada que el tiempo de construcción a la hora de desarrollar modelos predictivos para estimar la duración del proceso constructivo de proyectos de edificación, y que dicha velocidad se ve más afectada por la SC que por el coste. Además, la metodología numérica basada en el MEF proporcionó mejores modelos predictivos que los generados mediante redes PMC o utilizando ARL. A este respecto, los hallazgos de esta investigación apoyan la

idea de que modelos desarrollados mediante ARL pueden proporcionar un buen punto de partida desde el que buscar mejores modelos predictivos utilizando técnicas de modelado no lineal.

El conocimiento obtenido con esta tesis permitirá explorar nuevos enfoques con el propósito de determinar mejor las relaciones existentes entre los factores principales de alcance del proyecto y la velocidad de construcción de proyectos de edificación, pero siempre teniendo en cuenta que los resultados proporcionados por los modelos propuestos son solo estimaciones iniciales de la velocidad de construcción en las etapas iniciales de desarrollo del proyecto, cuando solo se dispone de información básica, y tales modelos no pretenden remplazar las programaciones detalladas desarrolladas por el constructor.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
BE	Backward Elimination
BFGS	Broyden-Fletcher-Goldfarb-Shanno
BP	Back-Propagation
BTC	Bromilow's Time-Cost
CBR	Case-Based Reasoning
CG	Conjugate Gradient
CPI	Consumer Price Index
CV	Coefficient of Variation
DFP	Davidson-Fletcher-Powell
FEM	Finite Element Method
FS	Forward Selection
GD	Gradient Descent
GFA	Gross Floor Area
GO	Global Optimisation
HCA	Hierarchical Cluster Analysis
HN	Hidden Nodes
IRF	Index based on Revision Formulas
KW	Kruskal–Wallis
LR	Learning Rate
LRA	Linear Regression Analysis
M	Momentum
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLP	Multi-Layer Perceptron
MLRA	Multiple Linear Regression Analysis
MMRE	Mean Magnitude of Relative Error
MRE	Magnitude of Relative Error

MSE	Mean Squared Error
MW	Mann-Whitney
NRBF	Normalised Radial Basis Function
PI	Performance Index
RBF	Radial Basis Function
RMSE	Root Mean Square Error
RNCC	Retrofit Net Construction Cost
RRQR	Rank Revealing QR
RTO	Regression-Through-the-Origin
SA	Simulated Annealing
SBP	Standard Back-Propagation
SCAWI	Statistically Controlled Activation Weight Initialisation
SCG	Scaled Conjugate Gradient
SSE	Sum of Squared Error
SV	Scaling of Variables
TF	Transfer Functions
UK	United Kingdom
US	United States

Chapter I

Introduction

I. Introduction

I.1 Background

The project success can be defined as "*the degree to which the project goals and expectations are met*" (Sanvido et al., 1992). Although success criteria often change from project to project, and depends on the point of view of each project participant (Sanvido et al., 1992), it is commonly believed that a building project can be seen as successful if the building is completed as scheduled, within the budget and quality standards, and achieves a high level of client satisfaction (Chan, 1998). Consequently, time, cost and quality have been recognised as three of the most important project success criteria. In particular, predicting the construction time of a building project has been of great interest for many researchers and practitioners related to the construction industry (Chan & Kumaraswamy, 1999). One of the main reasons is that the duration of the construction process often serves as a benchmark for assessing the performance of a project and the efficiency of the project organisation (Chan, 1998; Nkado, 1995). In this connection, a company needs to compare its construction time performance regarding the performance of its competitors in order to identify potential negative differences, and, if so, implement strategies to ensure an improvement in its productivity (Walker, 1995).

Nevertheless, delays are a common problem in the construction industry (Doloi et al., 2012; Meng, 2012) and their magnitude varies considerably from one project to another. For example, Bordoli & Baldwin (1998), cited in Chan & Kumaraswamy (2002), reported that there were 50-80% delays on 1627 projects sponsored by the World Bank between 1974 and 1988, together with an average of 23.2 time overrun on the United Kingdom (UK) Government construction projects from 1993 to 1994. Deviations that occur in building projects regarding to their planned duration have become an almost inevitable fact that may be due to many reasons. Moreover, these delays can produce serious economic consequences for the different agents involved in the construction process (Stoy, Dreier, et al., 2007). The owner loses out due to loss of potential income derivatives from the use of the built product and the increase in

overheads caused by an increase in project monitoring time. The contractor also loses due to increased costs in tied-up capital (Al-Khalil & Al-Ghafly, 1999), the specific time-based costs, and the effects of liquidated damages according to their contractual arrangements (Forsythe et al., 2010). Therefore, there are significant economic factors underpinning the importance of compliance with building deadlines. In addition, project duration overestimation may also produce a negative impact on organisation, for example, by depriving it of the opportunity to develop other projects (Chan, 1998).

From a basic preliminary analysis, the delay of a building project may be due to poor performance motivated by different causes such as failure of contractor, changes in project design, or inclement weather (Greenwood & Shaglouf, 1997). But delays can also be caused by initial estimates that are too optimistic about the final duration of works (Ng et al, 2001). Predicting construction time is a very hard task because the completion of the construction process is the result of a combination of many processes and events, planned or unplanned, occurring in a constantly changing environment (Sanvido et al., 1992). Moreover, project managers must cope with increasing challenges in the construction process of building projects as it has become more dynamic and complex (Chan, 1996). Consequently, despite the many advances made in the discipline of project management, in general, over the past decades construction projects have obtained poor performance regarding their planned duration (Chan & Kumaraswamy, 2002; Ng et al., 2001).

In this context, there is no consensus in the literature regarding the identification of factors that affect the construction time of buildings projects (Nkado, 1995). Therefore, there is a need to identify the factors that have more influence on construction time in the building sector and analyse the interrelationships existing between them in order to better understand the complexity of the construction process. In addition, predicting construction time at early project phases has been considered a key element for project success (Dursun & Stoy, 2011b), as better planning at early stages of project development leads to better final project outcomes (Wang et al., 2012). Consequently, there is also a growing need to develop tools that allow project managers to estimate in a reliable manner the duration of the construction process of building projects, even before designs are completed and/or before detailed programmes can be developed (Chan & Chan, 2004).

I.2 Problem Statement and Research Motivation

I.2.1 Construction Process Duration

The construction industry is very diversified, heterogeneous in structure, and complex in product type (Sidwell, 1982). The construction process of a project can usually be divided into three important phases: project conceptualisation, project design and project construction (Chan, 1998). This thesis focuses on the construction phase, understood as the stage where the physical form of a project is created in compliance with design specifications (Chan & Kumaraswamy, 1997) and, in this study, this phase covers the period of time from the beginning of the works up until all planned construction activities are completed. In this regard, it should be noted that throughout this thesis the terms "construction time", "construction duration", “project duration”, and "construction process duration" will be used interchangeably to refer to the duration of the aforementioned construction phase.

Although sometimes planners spend a lot of time calculating construction process duration in a justified manner, there are no fixed rules to do it. Sometimes, deadlines are calculated based on previous experience of planners in similar projects (Chan & Chan, 2004), in which case the estimation of construction time tends to be subjective and highly dependent on the skill and experience of the planner (Chan & Kumaraswamy, 1995). Other times the client imposes the project duration and contractors simply assume that this duration is realistic enough and prepare their offers in line with it (Ng et al., 2001). In the best case scenario, an accurate estimate of the construction time in building projects can be made by developing a detailed schedule. Nonetheless, for this purpose the following, among other things, must be determined: project work breakdown structure, duration of each activity and resource constraints. This information is very difficult to obtain at early stages of project development due to the wide variability existing in construction processes and because, generally, such information is only in the hands of the constructor. The great uncertainty existing in project scheduling arises from the following characteristics (Khodakarami et al., 2007):

- **Uniqueness**. Each project is unique because there is always at least one of the following parameters that changes: targets, resources and/or environment (Vidal & Marle, 2008).

- **Variability**. Trade-off between performance measures like time, cost, and quality.
- **Ambiguity.** Lack of clarity, lack of data, lack of structure, and bias in estimates.

The aforementioned problems, together with both time and labour limitations (Hoffman et al., 2007), often make it impractical to elaborate a detailed schedule at the early stage of project development and, consequently, the estimation of construction time at these stages is more often than not based on individual intuition and experience of the planner (Chan & Kumaraswamy, 1995). However, it is during these stages that decisions made by the project team regarding project duration and costs are of significant importance, since, in the end, these decisions will affect the cost and schedule performance of the project (Wang & Gibson, 2010). Thus, with a view to minimising the subjective effect on the estimation of construction time, at this initial stage of project development when only basic information is available, the agents involved in the project need effective tools that allow them to estimate the duration of the construction process in a simplified but reliable manner (Sousa et al., 2014). In this regard, although it is generally more realistic to calculate the project duration through an activity-based methodology, a good strategy for predicting construction time could be to use in the first place a general forecasting model of the construction process as soon as the project design has been completed and subsequently develop the specific planning details (Chan, 1998).

Although the planning of the construction process differs from project to project, nevertheless it is of a similar nature (Soltani et al., 2015). In this regard, the literature states that, initially, the construction process duration of a building is affected by various factors related to project features, although some factors are more crucial than others. It is generally assumed that planners are able to make reasonable decisions about the influence of these factors on construction time, but the literature and history of building projects suggest otherwise (Boussabaine, 2001a). In the case of building projects, most of the literature has identified the factors included in the category "project scope" as key predictors to estimate construction time (Chan, 1998; Walker, 1995). However, there are contradictory conclusions about what dependent variable and project scope factors are more suitable for modelling construction time. On one hand, although cost and GFA have been the factors most commonly used to define

the project scope, there are no perfect measurement units (Walker, 1994). On the other hand, although Bromilow (1969) and other researchers have also shown that construction time is usually the most appropriate dependent variable to be analysed, some models also used the construction speed as an alternative dependent variable (Love et al, 2005; Stoy, Dreier, et al., 2007; Stoy, Pollalis, et al., 2007). In this regard, the debate revolves around whether it is possible to obtain better forecasting models by using construction time as the dependent variable or if it is more appropriate to consider construction speed (Stoy, Pollalis, et al., 2007).

In summary, for the purpose of developing appropriate forecasting tools there is a need to enhance existing knowledge regarding which project scope factors influence construction time, their level of influence and the type of interrelationship between them. By doing so, it will be possible to generate prediction models that allow project managers to increase the accuracy in initial estimates of construction time and reduce the magnitude of deviations from planned durations.

I.2.2 Modelling Construction Time

Modelling is the process by which we construct a simplified mathematical reality from a more complex physical reality. In order to generate good predictive models, the choice of a suitable modelling technique is of vital importance. Empirical studies have shown that forecasting accuracy changes with the use of different modelling techniques, and this, in turn, increases the problem of selecting the best model (Goh, 1998).

In order to provide proper tools to estimate construction time and minimise the subjectivity of such estimation, in previous research related to building projects most authors have developed parametric models by using linear regression analysis (LRA) and data derived from completed real projects. LRA has usually been considered to be better at making judgements than planners, whereas, planners are much better at selecting information than they are at integrating it (Boussabaine, 2001a). Furthermore, LRA can be used as both an analytical technique and a predictive tool when examining the contribution of potential new variables to the reliability of the estimation, although it is not appropriate to define nonlinear relationships (Kim et al., 2004). In LRA the technique of the least squares is the one most often used to obtain the parameter values involved in regression equations. Due to the linearity of the

relationship between variables, the application of this technique leads to solving a linear system in the parameters. It is worth mentioning that the logarithm form of the dependent and independent variables is the most commonly used transformation which allows the possibility of fitting nonlinear models by using linear regression (Goh, 1998). Bearing in mind this aspect, when indicated in this thesis that a model is linear it means that this model is linear in the parameters, although it may be nonlinear in the variables (e.g., logarithmic transformations).

Although parametric models have shown a good balance between the difficulty of developing estimation models and the forecasting accuracy obtained using such models, they represent a great simplification of the complex relationships that control the duration of the construction process in building projects (Sousa et al., 2014). Moreover, their accuracy and trustworthiness is basically limited by some assumptions inherent to LRA (Jafarzadeh et al., 2014). Nonlinear behaviours are common in complex real systems like the construction phase of a building project, and project scope factors that influence the time variable might not be fully associated in a linear manner. In this regard, many attempts made to date to explain the variance of construction time in a linear manner are questionable (Boussabaine, 2001a). Notwithstanding this, LRA would still give us a good approximation to explain relationships, which are not truly linear, and provide a starting point from which to explore more sophisticated models (Gilchrist, 1984, cited in Chan, 1998).

In contrast to LRA, the nonlinearity of the parameters forces us to use numerical computation techniques in order to find solutions (Bates & Watts, 1988). These techniques are iterative methods that start from an initial value of the parameters and obtain a new value closer to the optimum value at each step. In this regard, the learning process of artificial neural networks (ANNs) is equivalent to minimising a global error function, which is a multivariate function that depends on the synaptic weights available in the network (Møller, 1993). From this perspective, the supervised training of ANNs can be seen as a problem of numerical optimisation (Haykin, 1999). There is a growing tendency in recent years to use ANNs to generate predictive models for building projects (see, e.g., Chen & Huang, 2006; Le-Hoai et al., 2013).

ANNs have been considered as artificial intelligence (AI) techniques of interest to the construction industry because they are likely to improve automation efforts (Moselhi et al., 1991) and provide some advantages over conventional linear

regression methods. These traditional statistical methods are model-driven, which means that before unknown model parameters can be estimated the model structure must first be determined (Maier & Dandy, 2000). On the contrary, ANNs are nonlinear data-driven approaches that can develop nonlinear models without prior knowledge of the relationships existing between input and output variables. Thus, they are a more general and flexible modelling tool for forecasting than the traditional statistical methods (Zhang et al., 1998). The learning ability of ANNs allows for solving complex problems (Günaydın & Doğan, 2004) and modelling linear and nonlinear systems without the need to follow a specific statistical distribution to find the relationships between the variables under consideration (Wang & Gibson, 2010). It is clear that LRA and ANNs have become two competing empirical model-building methods (Smith & Mason, 1997).

I.3 Research Objectives

The main objective of this research was to analyse both the kind and degree of influence of the relationships existing between some of the main project scope factors and the construction time of new builds developed in Spain. By doing so, it will be possible to extend the body of knowledge needed to develop realistic prediction models to be used at early project phases by the main stakeholders involved in the construction process of building projects.

In order to achieve the proposed general objective, the following specific sub-objectives need to be defined:

- Identify the best dependent variable to generate the best predictive models for estimating the duration of the construction process of new builds: construction time vs construction speed.
- Generate models for predicting construction time using LRA and project scope factors.
- Develop a comparative analysis of the influence of GFA and construction costs on the construction time of new builds, as they have been recognised in previous research as the most influential project scope factors.
- Generate models for predicting construction time using ANNs.

- Develop a comparative performance analysis by using different types of ANNs, learning algorithms, and calibration data divisions.
- Generate models for predicting construction time using a novel numerical methodology based on the finite element method (FEM).
- Undertake a comparative evaluation of the best predictive models obtained by using the different modelling techniques proposed in this research work.

I.4 Research Hypotheses

According to the research problem posed in this thesis and the abovementioned objectives, three principal hypotheses were postulated:

(H-01) "*Construction speed is a more appropriate dependent variable than construction time in order to generate predictive models for estimating the duration of the construction process of new builds.*"

(H-02) "*GFA has greater influence on the construction speed of new builds than construction costs.*"

(H-03) "*Considering the same set of project scope factors as predictor variables, nonlinear modelling techniques can generate models to estimate the construction speed of new builds with better predictive performance than that offered by linear regression models.*"

I.5 Outline of Research Methodology

The research conducted in this thesis was based on results obtained by previous research and the identification of gaps in knowledge regarding the construction time of building projects. In order to achieve the objectives established in this thesis, the research made use of a database that contains more than 300 projects with different uses, locations and sizes. It contains a collection of Spanish construction projects classified in different constructive typologies. The research focused only on new builds, and, initially, 168 projects were selected as valid for the study and classified into 7 types of facilities according to their nature and scope. In addition, similarly to Irfan et al. (2011), the generated models were validated using a new sample of 18 health building projects, totally independent of the data set used to develop the predictive models.

By using the aforementioned data sets, three different modelling techniques were tested in the study to generate predictive models and provide a clearer explanation about the influence of some of the main project scope factors on the construction speed of new builds. In doing so, at the same time, the hypotheses proposed in this thesis were also checked. To this end, the research work developed herein can be divided into four main stages:

1. In the first stage, different models presented in the literature for predicting the construction time of building projects, together with the factors used to develop them, were investigated.
2. In the second stage, first of all, considering that there is no general agreement in the literature about which dependent variable is the most appropriate to predict the duration of the construction process, the correlations between two types of response variables (construction time and construction speed) and the predictive variables selected from the data sets were analysed. Then, both log-linear and linear forecast models were generated by using LRA. The project scope factors represented by GFA and construction costs were analysed in a special way.
3. Third, based on the knowledge provided by developed linear regression models, different ANN models were generated to predict the construction speed. In this task, the multilayer perceptron (MLP) network architecture was used in conjunction with two different learning algorithms and five options for data division. During this phase of the research a methodology was defined in order to obtain optimised neural network configurations which produce models with an adequate predictive performance. The proposed methodology develops a stepwise trial and error procedure in which a basic MLP network structure with enough consistency is first set, and subsequently this initial structure is modified at each step of the process in order to achieve the best possible network configuration.
4. Lastly, for the purpose of generating better nonlinear forecasting models a new numerical methodology developed by Navarro-González & Villacampa (2012, 2013) was applied for modelling construction speed. This new numerical method allows for the creation of representation models, from a previously defined relationship, by using FEM. FEM is considered as a numerical approximation method and it has been used to solve problems of both

scientific analysis and engineering. In particular, it is worth noting its application for solving problems of structural mechanics and differential equations which have no exact solution.

I.6 Scope and Benefits of the Study

A lot of research has been developed based on factors that may influence construction time and a wide variety of predictive models generated. Nevertheless, among these studies there are contradictory conclusions. Some studies have compared the performance obtained on projects from different countries (Dursun & Stoy, 2011a; Xiao & Proverbs, 2002) and their results reveal the existence of differences in construction speed according to the physical location where the project is developed. Consequently, the construction time is context-specific (Le-Hoai et al., 2013) and the development of similar predictive models, derived from building projects completed in different countries, has been considered relevant in the literature (Chan, 1998). In this regard, to the best of our knowledge, no research has been conducted in Spain, either to identify factors which influence construction time or to propose models for estimating the duration of building projects at the early stages of project development. At the same time, at each step in the development of the body of knowledge related to the construction time of building projects, this new knowledge must be periodically tested, as circumstances change elements of the theory and causal factors could shift (Walker, 1994). By using the concept of construction speed as the output variable of predictive models, this research also attempts to search for answers to the question why some buildings are constructed faster than others.

Although it is clear that besides the project scope factors used in this research there are other factors that can affect the construction speed of a building project, the use of this type of predictive models to estimate the project duration remains valid. This has been shown in many of the previous research studies that can be found in the literature. The results provided by the proposed models are only initial estimates and are not intended to replace detailed schedules undertaken by builders and owners, where other specific factors, such as the construction methods used in the construction process, are likely to be considered. Notwithstanding the fact that accuracy error may arise from making use of the developed models, their use may be helpful before starting the works once the project scope factors involved in the model

are known in the project. The forecast provided at this time can serve as a control parameter to verify if the building deadlines proposed by any of the agents involved in the construction phase are realistic and, if necessary, to take special measures to modify construction speed. In addition, the determination of the construction time is vital for a proper cash flow planning because, from the contractor's point of view, it facilitates optimal resource allocation, while an enhanced certainty of the project duration also assists the client in contractor selection (Chan & Kumaraswamy, 1995).

It is worth stressing that, as stated previously, the main objective of this thesis is to analyse the relationships existing between some project scope factors and the construction speed of new builds, showing which of the applied modelling methodologies generates models with better predictive performance, rather than develop specific predictive models for their practical application. From this perspective, the year of project execution is no longer decisive, although it is a recognised fact that the coefficients of this type of models should be reviewed periodically to keep them updated according to the evolution of construction techniques used in the building sector (Chan & Kumaraswamy, 1999). In any case, the study results can benefit both the researchers and practitioners of the building sector in obtaining a better understanding of the construction speed of new builds. In the case of practitioners, the early understanding of construction time represents a key element for project success as it can help the stakeholders involved in the construction process to make informed decisions at early stages of project development.

I.7 Thesis Structure

In this chapter, an overview of the proposed research work, including practical and theoretical motivation, objectives, hypotheses, and research methodology, has been presented. In subsequent chapters, the thesis is organised as follows:

- **Chapter II** carries out a review of relevant literature regarding the factors affecting the construction speed of building projects, as well as the statistical models developed in previous research for predicting construction time. A discussion on measures of scope and complexity for building projects is also developed, while the concept of construction speed is analysed in detail, as it underpins this thesis. Basic concepts and applications of ANNs in the construction industry are also dealt with.

- **Chapter III** explains the research methodology adopted in this thesis in order to test the hypotheses proposed in Chapter I and meet the research objectives. Firstly, general concepts about the process to generate mathematical models are discussed. Then, the characteristics of the data sets and variables selected for analysis are presented. Thereafter, the modelling tools selected to develop prediction models are described in detail. The analysed modelling techniques are: LRA, ANNs and a new numerical methodology based on FEM. Lastly, the framework proposed for evaluating the predictive performance of linear and nonlinear models is established.
- **Chapter IV** first examines which dependent variable (construction time or construction speed) is more appropriate to develop linear regression models to estimate the construction process duration of new builds. At the same time, it also analyses which predictor variable has more influence on construction speed and which form of the variables under study is more suitable to obtain valid linear regression models. Subsequently, the best predictive models obtained by using multiple linear regression analysis (MLRA) are presented and evaluated. Finally, a general discussion about the interpretation of the regression coefficients is carried out.
- **Chapter V** presents several nonlinear models generated by using ANNs and the numerical methodology based on FEM. For this purpose, the sets of independent variables that obtained the best predictive performance using the MLRA technique were selected. Specifically, MLP models are generated using the optimisation methodology defined in Chapter III together with two different training algorithms. The influence of different types of calibration data divisions on the performance of MLP models as well as the predictive performance of a different type of ANN architecture are also evaluated. Finally, a comparative analysis of the best predictive models generated using the three methodologies proposed in the thesis is carried out.
- **Chapter VI** draws the general conclusions and the main contributions of this thesis. Moreover, the limitations of this study are outlined and research that could be faced in the future is also addressed.

Chapter II

State of the Art

II. State of the Art

II.1 Introduction

This chapter develops an exhaustive review of the literature related to the estimation of construction time in building projects, for the purpose of providing valuable information about the advances made in the research topic during the last decades and its current state. Firstly, different ways to classify the factors likely to influence the construction time of a building project are analysed. Thereafter, the so-called project scope factors, which are the factors analysed in this thesis, and their relationships with project complexity are discussed in detail. In addition, the concept of construction speed, along with its use as output variable, is also explained, since this concept forms the basis of the research work developed in this thesis. Some of the most important parametric models presented in said literature are also discussed in this chapter. Finally, basic concepts and applications of ANNs in the construction industry, along with a comparison between the modelling methodologies represented by LRA and ANNs, are also introduced.

II.2 Categorisation of Factors Influencing Construction Time

A thorough literature review suggests that the construction time of a project is affected to varying degrees by a great number of factors. However, there is no consensus on this subject. Nkado (1995) indicated that time-influencing factors need to be prioritised, and for that, he first identified from the literature a total of 33 specific factors and classified them into six different categories:

- Client
- Design and specialist consultants
- Contract
- Project
- Site management
- External influences

Subsequently, he selected the ten most important factors using the completed questionnaires received from 29 firms belonging to the National Contractors Group in the UK. These factors were: client's specified sequence of completion, contractor's programming of the construction work, form of construction, client's and designer's priority on construction time, project complexity, project location, buildability of design, availability of construction management team, and completeness and timeliness of project information. According to Nkado (1995), the contractor is able to quickly identify these factors based on project information and, although he does not produce the majority of them, the contractor is in a position to quantify and evaluate their impact on construction time.

Chan & Kumaraswamy (1995) also proposed a range of significant qualitative and quantitative factors influencing construction time which were hierarchically structured. They stated that construction time can be regarded as a function of all these hierarchical factors. The eight factors located at the top of the hierarchical tree were:

- Construction cost / value
- Type of construction
- Location
- Client's and other imperatives / priorities
- Total factor productivity
- Others
- Type of contract
- Post-contractual developments

Chan (1998) classified the factors affecting construction duration into four main categories: project scope, project complexity, project environment, and management attributes. Figure II-1 shows these categories and their main associated factors. The subjective nature of some of these factors, along with the high variability associated with qualitative factors, has made the development of prediction models a challenge (Hoffman et al., 2007). In this regard, the definition of project scope is an essential element in the planning process for achieving an excellent project performance (Wang & Gibson, 2010), since most of the literature has identified the factors included in the category "project scope" as useful and reliable predictors of construction time (Chan, 1998; Walker, 1995). Furthermore, according to Walker (1994), challenges to the project team are derived from scope and complexity.

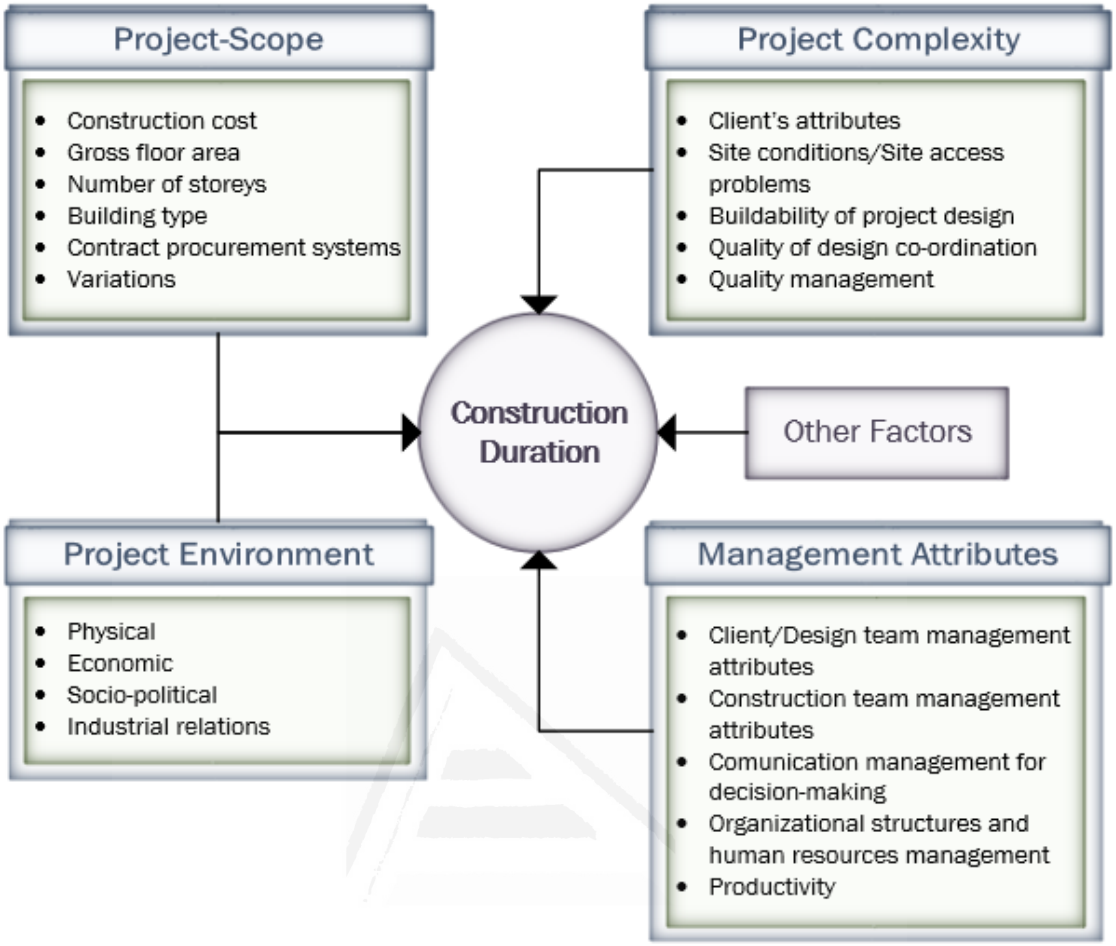


Figure II-1. Categorisation of main factors affecting construction duration (Chan, 1998).

This thesis used the categorisation of factors proposed by Chan (1998) as a basis for analysing both the relationships between factors that influence construction time and the development of predictive models. However, due to the high subjectivity of some of these factors and the unavailability of the data required to study others, the research work focused only on the general study of the group of factors classified as "project scope". The concept of project complexity was also analysed from a general perspective, with special attention to its relationship with project size and the buildability of project design.

In the case of building projects, construction cost, GFA and number of stories were the project scope factors most often used to develop models for estimating construction time (Sousa et al., 2014). However, there are contradictory conclusions in the literature about which of these factors have the greatest influence on construction

time. According to Dursun & Stoy (2012), different predictor variables were employed to describe the construction time in different studies related to building projects because, on one hand, some predictive models were built in different geographical areas and, on the other hand, observational studies are highly influenced by the properties of available data sets, so that different samples could represent distinct populations.

II.3 Project Scope and Project Complexity

II.3.1 Project Scope Factors

In general, in the field of project management the term “project scope” is understood as “*the work that needs to be accomplished to deliver a product, service, or result with the specified features and functions*” (PMI, 2008). Nevertheless, in the case of factors affecting the construction time of building projects, “project scope” has been considered as a measure of project size which transmits a sense of magnitude and can be described in terms of construction cost, area, or volume (Walker, 1994).

Chan (1998) included in the category “project scope” the following factors: construction cost, GFA, number of stories, building type, contract procurement systems, and variations. The features of these factors will be discussed below, but considering the Walker’s definition, construction costs, GFA and number of floors are analysed jointly on the understanding that they are factors that allow for measuring the project size in a real way.

Construction Costs, GFA and Number of Floors

From the time-cost model developed by Bromilow (1969) many models have considered construction cost to be a good measure of project scope. According to Chan & Chan (2004), the construction cost of a building could also suggest a sense of buildability and complexity. Moreover, Bromilow (1969) also indicated that this parameter has the advantage that can reflect the quality and physical size of building projects. After investigating the literature and undertaking statistical analysis of both GFA and cost data, Walker (1994) concluded that construction cost provides the best measure of project scope. He also pointed out that the main advantage of using construction cost as a measure of project scope is that all the elements of a building

can be expressed in terms of cost, but at the same time he also noted that describing the scope of a project without including construction cost can be useful for certain types of projects.

Apart from cost, Chan & Kumaraswamy (1995) also recognised that GFA and number of floors are factors that significantly influence construction time. In the same line of thought, Love et al. (2005) concluded that GFA and number of floors are crucial factors to calculate the construction time of building projects, while cost is a poor predictor to estimate execution time. Similarly, Stoy, Dreier, et al. (2007) indicated that GFA defines the project size in the most appropriate way for building projects. In this regard, GFA and number of storeys meet the aims of being practical to use since this data can be easily obtained from drawings, is less sensitive than cost data and has the potential to be obtained from existing data sources (Forsythe et al., 2010).

Although cost and GFA have been the factors most commonly used to define the project scope, there are no perfect measurement units (Bromilow, 1969; Walker, 1994). On one hand, the problem of construction costs, when used as a predictive variable, is that this data tends to be commercially sensitive (Forsythe et al., 2010) and also that the final cost normally varies from the initial cost estimated before starting the works (Love et al., 2005). This difficulty to set the actual cost of the construction process could impair the ability of a model to predict the final duration when cost is used as predictor variable and restricts the use of construction cost for industry wide benchmarking (Forsythe et al., 2010). Moreover, in some specific cases construction cost need not necessarily affect project duration. For example, let us imagine that there are two building projects with the same design and the same construction processes, and the only difference between them is the quality of materials. In this particular case, the construction cost will be different from each other but the construction time does not have to be different.

On the other hand, the problem of GFA as a unit of measure is that it may also disguise the complexity of certain works, for example in building projects that include a great deal of external works (Walker, 1994), but which could be considered if the construction cost is used as a measure of project scope (Sousa et al., 2014).

Building Type

Regarding the building type, it was related to the concept of project complexity by Sidwell (1982). In this connection, Nahapiet & Nahapiet (1985) showed the difficulty in using construction speed as a measure of construction time performance since different types of buildings show different degrees of complexity in order to carry out the construction process. Every type of building project serves a variety of specific functions, some of which would constitute great difficulty to the contractor, provided that he is not familiar with them or does not have the necessary skills or resources to achieve the required objectives (Xia & Chan, 2012). For example, the construction process of a hospital does not have the same complexity as that of a residential building.

As already stated, the main advantage of using construction costs as a measure of project scope is that all elements of a building can be expressed in a single unit of scope measure, especially in situations where the building project involves different types of uses, for instance hotel and parking (Walker, 1994). In these cases, it might seem more logical to describe the construction time in terms of construction cost rather than separate the building in different parts and use GFA as measure of project scope, particularly when there are shared spaces which are difficult to be defined (Walker, 1994). The same problem can arise in building projects with large external works, where the use of GFA as a measure for project scope might lead to errors for estimating construction time (Walker, 1994).

Bromilow (1969) found that the construction time of buildings of a given value does not depend very strongly on the type of building or its location. Nevertheless, according to the literature, it is expected that different building types would also produce different time-cost relationships (Chan, 1998).

Contract Procurement Systems

There is no consensus in the literature regarding how contracting methods affect the construction time performance. Although some studies indicated that it is possible to shorten the construction time of building projects by using certain types of contracting methods, other researchers concluded that construction speed is not affected by the contractual arrangements (Chan, 1998). The analysis of this factor has been left out of the scope of this research work due both to the wide variety of contract

types used in building projects and to the lack of data available to analyse the potential influence of contracting methods on construction speed.

Variations

Bromilow (1969) and Ireland 1985) also confirmed a link between contract variations and poor construction time performance. Contract variations can affect workflow and change project scope and have been recognised as a common cause of poor construction time performance (Kumaraswamy & Chan, 1995). However, Walker (1994) stated that contract variations can vary in both scope and impact on construction time, and sometimes they may be quickly incorporated into the construction works without causing any effect on project duration through proper management of the construction process. In any case, variations to the contract after the construction process has started make no sense in this research work, as the results provided by the models developed herein are only initial estimates of construction speed at early stages of project development, when only basic information is available.

II.3.2 Project Complexity

Complexity appears to be one of the main reasons of the unpredictability of projects (Vidal & Marle, 2008) and many researchers have recognised the importance of its measurement in construction project analysis (He et al., 2015; Lu et al., 2015). Nevertheless, the term “project complexity” is difficult to define and even more difficult to accurately quantify (Lu et al., 2015). In general, there is no agreement on how to define project complexity (Lu et al., 2015; Vidal & Marle, 2008). Consequently, complexity can be understood in different ways not only in different fields, but also within the same field (Vidal & Marle, 2008). Moreover, the complexity measures would vary in different geographical locations (Xia & Chan, 2012). In this regard, project context is an essential feature of project complexity and, as a result, project complexity should not be analysed without considering the implications of the project context on it (Vidal & Marle, 2008).

Given that it is difficult to quantify precisely project complexity, several studies have focused on identifying and classifying different factors and measures to build frameworks describing project complexity qualitatively (Lu et al., 2015). For instance, Baccarini (1996) classified project complexity into organisational complexity and

technological complexity. Organisational complexity includes the number of hierarchical levels, with vertical and horizontal differentiation. The greater the differentiation, the more complex the organisation (Baccarini, 1996). Technological complexity involves the use of material means, techniques, knowledge and skills (e.g., number of subcontractors or trades involved in a building project).

Complexity challenges related to the client were also investigated by Sidwell (1982) and the findings reported by Chan (1996) indicated that increased complexity would result in decreased client's satisfaction on time and cost, and decreased designer's overall satisfaction. Walker (1994) studied the complexity of building projects as an aggregate measure derived from three qualitative data sources: client and client representative characteristics, project characteristics and environmental characteristics. In particular, project characteristics were defined in terms of: inherent conditions prevailing at the site, buildability of a design solution, quality of design coordination and design detailing, quality management procedures, and access to site and within site.

Vidal & Marle (2008) argued that project complexity can be characterised by using some factors that can be classified into four families: project size, project variety, project interdependence, and project context-dependence. All are necessary, but not sufficient condition for defining project complexity (Vidal & Marle, 2008). Xia & Chan (2012) also identified six crucial complexity measures for building projects, which were ranked in order of importance: (i) building structure and function, (ii) construction method, (iii) the urgency of the project schedule, (iv) project size/scale, (v) geological condition, and (vi) neighboring environment. Generally, these studies built project complexity frameworks from different perspectives, which were based on the assumption that project complexity is linear. However, the behaviour of project complexity is nonlinear (Lu et al., 2015).

According to Williams (2002), overall project complexity can be characterised by two dimensions: structural complexity and uncertainty. In turn, each dimension can be divided into two sub-dimensions (see Figure II-2). On the one hand, the structural complexity is related to the underlying structure of the project. The two sub-dimensions of structural complexity lead to a complex system in which the whole is more than the sum of the parts. In these systems it is very difficult to intuitively infer the behaviour of the system from the behaviour of the sub-elements. On the other hand, uncertainty by

itself might not cause complexity, but when uncertainties arise, either in the goals or the methods, they cause perturbations which produce complex dynamic behaviours in structurally complex systems (Williams, 2002). Thus, in this second dimension the key question is how well defined the project goals are and how well defined are the methods of achieving those goals (Turner & Cochrane, 1993). Uncertainty in the methods to carry out a project will add complexity to the project (Williams, 1999) and different levels of goals will increase the project complexity (Lu et al., 2015)

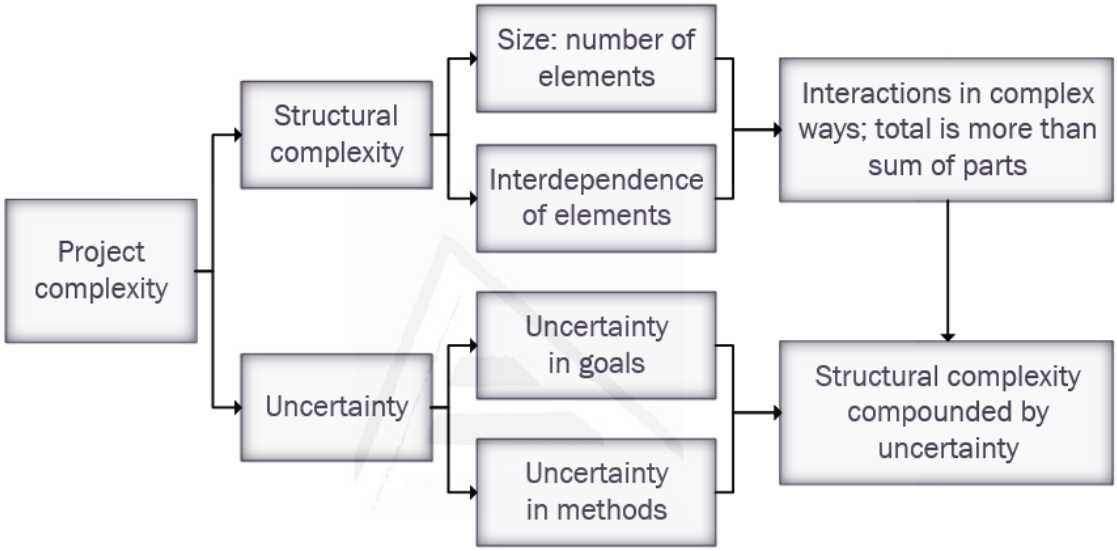


Figure II-2. Dimensions of project complexity (Williams, 2002).

According to the aforementioned studies, there is a general acceptance that project complexity is directly related to the project size. Although a larger project size does not necessarily lead to higher degree of complexity, generally, it is necessary to develop more complex coordination systems. As the size of a project increases, difficulties in coordination among all participants increase, affecting the project complexity in terms of management (Xia & Chan, 2012). Additionally, in the case of building projects, availability of facilities and materials are potential sources of risk associated with large projects (Xia & Chan, 2012).

However, project complexity means something more than a large project (Williams, 2002). Considering only the size of the project is not possible to assess the project complexity, which also depends on the complexity of the relationships between constituent elements and the uncertainties associated with the goals and methods of

the project. In this regard, a project of large scope (measured as GFA or construction cost) may have very little complexity associated with it and represent limited challenges to the project team. On the other hand, a project of smaller scope may represent greater challenges in a very complex environment (Walker, 1994).

In the light of the aforementioned considerations, it is clear that the building construction process has an inherent complexity and uncertainty. However, Walker's (1994) research findings suggest that many factors relating to project and environmental complexity do not affect construction time performance per se. To argue this statement, Walker (1995) explained that very complex projects can be built very quickly and he cited The Empire State Building in New York as an example of building site with severe site access problems that was built in a short time. Only one factor affecting inherent site conditions was found to be significant: water table and geotechnical problems (Walker, 1995). In the same line, Bromilow (1969) indicated that some building projects suffered prolonged delays due to unexpectedly bad ground conditions.

From the perspective of Walker (1994), the project risk to meet construction times is not so much associated with the complexity of project design, but rather with how the builder can overcome different difficulties associated with a complex design in order to construct the building. In this regard, project complexity can have both a negative aspect, in terms of difficulty to be understood and managed, and a positive influence on the project thanks to the emergence of opportunities (Vidal & Marle, 2008).

II.3.3 Buildability of Project Design

The complexity of the form of construction revolves around the concept of "design buildability" and the ease with which builders translate the project design into a constructed form (Walker, 1994). In this regard, "*buildability can be referred to the extent to which a building design facilitates efficient use of construction resources and enhances ease and safety of construction on site whilst the client's requirements are met*" (Lam et al., 2006).

According to Ireland (1985), the inherent complexity of the form of construction arises from special structural features (e.g., the use of long beams or large floor slabs), the use of unusual materials, or the use of complex construction processes,

and it will lead to increased costs and time. The analysis of construction methods as part of the design process, for the purpose of reducing construction time, is usually called “buildability analysis” (Ireland, 1986).

Building construction involves a large number of crews that generally work in a continuing and repeating sequence as they move from one floor to another (Tommelein et al., 1999). Therefore, one factor that influences buildability is the degree to which the design of components, space and access helps the builder to adopt a production-line approach on site (Walker, 1994). This is achieved by increasing the level of repetition, as well as ease of moving work crews from one place to another inside the building when the continuity of some of the works is disrupted. Stretton & Stevens (1990) argued that well planned resource movement and a capacity for uninterrupted work flow are major contributors to achieving acceptable building construction time performance. Disruption to workflow on building sites usually occurs because workers and not products are moved (Walker, 1994). In addition, repetitive activities provide an opportunity to achieve higher productivity because production rates can be improved with additional experience and practice (Arditi et al., 2001).

The fact that construction speed is related to project size, measured as GFA, with greater performance for larger projects (Ireland, 1983; Love et al., 2005; Stoy, Dreier, et al., 2007; Stoy, Pollalis, et al., 2007), might be explained in the light of the builder's capacity to adopt a production line approach and open up multiple workfaces (Walker, 1994). In this connection, Walker (1994) stated that buildability may not be intrinsically a factor, but rather due to construction management planning and control performance which may be enhanced or simplified by a design that favours good workflow. From this perspective, the selection of building system and construction materials for estimating construction speed is less important than favouring continuous workflow and achieving a buildable design (Walker, 1994). Nevertheless, it has been shown that the use of appropriate construction methods and technology, taking into consideration the principles of standardisation and prefabrication, can have a deep impact on the construction time of building projects (Chan & Chan, 2004; Chan & Kumaraswamy, 1995).

According to Ireland (1985), construction process planning during the project design is a managerial action that considers methods and processes of erecting

buildings as a part of the design process, and is an attempt to achieve a design which can be built in the most efficient manner possible. In this regard, planning the construction process during the design stage can be considered as a way of reducing project complexity. Consequently, this action may increase construction speed and reduce project duration. Furthermore, most studies have confirmed the positive relationship between improving the buildability of project design and cost savings, as well as better safety and quality performance (Lam et al., 2006).

II.4 Modelling Construction Time with Parametric Models

II.4.1 Literature Review

The first significant study (and the most cited) related to construction time of building projects was carried out in Australia by Bromilow (1969). By using 329 building projects, this study described mathematically the relationship between the final cost and the actual construction time of a building. Subsequently, the parameters of this time-cost relationship were updated by Bromilow et al. (1980). The resulting model, often called the Bromilow's time-cost (BTC) model (Ng et al., 2001), used construction costs as the main variable to predict project duration, but also considered the type of client (public/private) and year of construction.

The BTC model has been widely recognised as the standard for estimating or benchmarking the duration of construction projects by clients and contractors (Ng et al., 2001). After its development, several studies have tested that the time taken to construct a building project is highly correlated with the construction cost. Ng et al. (2001) conducted a review of the BTC model using a data set with 93 Australian construction projects classified in different types. They found that different parameter estimates were needed for industrial and non-industrial projects, with smaller industrial projects taking less time to complete than the smaller educational and residential projects. Even so, there were no significant differences between the methods of contractor selection or contractual arrangements. Finally, although other model forms were also checked none were found to be superior to the BTC model.

Other studies evaluated the BTC model also in Australia (Ireland, 1983; Walker, 1994), the UK (Kaka & Price, 1991), Hong Kong (Chan, 1999; Chan & Kumaraswamy, 1995) and other countries (Ameyaw et al., 2012; Chan & Kumaraswamy, 1995; Le-

Hoai et al., 2009; Le-Hoai & Lee, 2009). Among these studies there are contradictory conclusions regarding the BTC model performance. Ireland (1985) went on to state that the best predictor of average construction time is construction cost, while Walker (1994) noted that one of the shortcomings of the model is that it does not consider other factors besides cost. In addition, there is no guarantee that the parameters of the BTC model or its form will remain constant over time (Ng et al., 2001). An improvement in productivity will produce an increasing in construction speed and will affect the parameters of the BTC model. Moreover, the BTC model is not expected to be appropriate for all types of building projects and methods of procurement, since different types of buildings might exhibit relationships of the same form, but with quite different constants (Ng et al., 2001).

A lot of research has also been carried out in order to improve the accuracy of the BTC model by using other project scope factors influencing construction time. In addition to construction costs, Chan & Kumaraswamy (1995) also recognised that other specific characteristics of a project, such as GFA and number of floors, are factors that significantly influence project duration. Similarly, Love et al. (2005) concluded that in building construction projects GFA and number of floors are key factors when calculating construction time, while cost is a poor predictor. Some variables derived from project scope factors have also been analysed in the literature. Chan & Chan (2004) developed a model to estimate the duration for a specific type of housing in Hong Kong that included the derived variable represented by the relationship between GFA and number of floors, or what is the same in most cases, the floor area. Some studies have found that the cost/GFA ratio, named as "standard" (construction cost per square meter), is significant to forecast the construction speed of building projects (see, e.g., Stoy, Dreier, et al., 2007; Dursun & Stoy, 2011a). In addition, previous research related to building projects has also detected that construction speed also depends on qualitative variables. In this connection, the type of building project is one of the most used qualitative independent variables (see, e.g., Dursun & Stoy, 2011a; Love et al., 2005).

Forsythe et al. (2010) detected that is unclear in the literature if the variables that have relevance on larger projects, have the same relevance on smaller and differently managed housing projects. He analysed the extent to which GFA and number of levels are important factors in determining the construction time of Australian detached housing projects. Using a dataset of 196 projects the results of this study showed that

GFA and number of floors correlated strongly with estimated construction time, but weakly with actual construction time. Dynamically changing events during the construction process appeared to be the reason for this difference. Further, he stated that these factors affect larger housing projects more significantly than they do smaller projects.

The principal criticism of predictive models based exclusively on project scope factors is that the management of the construction process is too complex to be simplified to the value of a constant conveniently described in a formula, regardless of the accuracy of the statistical methodology used (Walker, 1995). In this regard, besides testing the relationships between construction time and macro-variables such as construction cost, GFA and the number of storeys, Chan & Kumaraswamy (1995) also studied the influence of micro-factors embodying site organisational variables. In particular, they investigated the impact of productivity on construction time by using a case study on a new building construction site. Productivity was defined as a ratio of output to input, i.e. the arithmetical ratio between the amount produced (output) and the amount of any resources used during the process of production (input) (Chan & Kumaraswamy, 1995). The results of this research revealed, as could not be otherwise, that these micro-factors must necessarily affect construction time.

Another attempt to generate more sophisticated models for estimating construction time was made by Nkado (1992). He analysed a sample of 29 commercial, privately funded buildings to create a computerised construction time information system for estimating time planning of buildings at early stages of design. The findings of this research showed that the durations of five primary activity groups (substructure, superstructure, cladding, finishes, and services) and their sequential start-start lag times can be predicted from 12 variables: GFA, area of ground floor, approximate excavated volume, building height, number of storeys, end use, cladding type, presence of atrium, building location, intensity of services and site accessibility. Nevertheless, the programme generated by the system did not include the durations of set-up and external works, which were not considered to be critical in determining the overall construction. Chan & Kumaraswamy (1999) expanded the research developed by Nkado (1992) and also attempted to generate standards for predicting construction time of public housing blocks in Hong Kong by modelling the durations of five primary work packages (piling, pile caps/raft, superstructure, E&M services, and finishes) and

their respective sequential start-start lag times in terms of identified sets of critical factors.

There have also been several research efforts that focused on more qualitative factors, linking non-scope factors to construction time performance. Sidwell (1982) explored the variations in organisation of a building team when developing a building project. For that purpose, a data set of 32 case studies were analysed and the study results suggest that, depending on circumstances, some building teams and project procedures are more likely to lead to the success of the project than others. In particular, the level of managerial control for the project was the most important indicator of success. However, no particular organisational form was found to be better than another per se, rather it is the combined effect of a number of variables that influence the success of a building project. In the same line, Walker (1995) found that construction management team performance plays a vital role in determining construction time performance and also that there is an important relationship between sound client's representative management effectiveness and good construction time performance. Ireland (1985) also used a sample of 25 high-rise office building projects in Australia with the purpose of analysing the effects of managerial actions on the objectives of reducing time, reducing cost and increasing the quality of these projects. He came to the conclusion that increases in variations to the contract, the complexity of the building, the number of storeys, and the extent of industrial disputes strongly increase construction time.

Finally, it must be emphasised that, in total agreement with what was established by Forsythe et al. (2010), most of the qualitative variables raised in the previous discussion may influence construction time, but they may not be well suited for use as predictive variables, due to their inability to be measured in an objective manner. In addition, using too many predictive variables to estimate the construction time of building projects has the risk of creating an overly complex model that may become impractical for its use by project participants and for benchmarking purposes of the construction industry. Therefore, it is necessary to achieve a trade-off between parsimony and prediction accuracy when developing predictive models to estimate construction time at early stages of project development.

II.4.2 Construction Time vs Construction Speed

In order to model the duration of the construction process, time has been the most used dependent variable, but some models have also made use of the construction speed as an alternative dependent variable (Love et al., 2005; Stoy, Dreier, et al., 2007; Stoy, Pollalis, et al., 2007). In this regard, the debate revolves around whether it is possible to obtain better forecasting models using construction time as the dependent variable or if it is more appropriate to consider construction speed (Stoy, Pollalis, et al., 2007).

Different forms of construction speed have been presented in the literature. Chan and Kumaraswamy (1995) pointed out that a unit of built-up area per time may be used as a macro indicator of the construction speed compared to micro indicators such as the productivity standards developed by different crews on a building site. Stoy, Dreier, et al. (2007) defined the construction speed as "*the average progress of construction over the construction duration*", measured as GFA (m^2) per month, while Love et al., (2005) described it as the required time to execute one unit of GFA. Gale and Fellows (1990) adopted a speed indicator of m^2 of built-up area per week (m^2/week) to evaluate the construction speed. Other researchers, like Nahapiet & Nahapiet (1985), used the concept of GFA per week and Ireland (1986) also evaluated the construction speed, but as GFA per 8 h of a working day. As can be observed, construction speed is generally defined as units per time and, consequently, the Love et al. (2005) definition is a non-standard definition. These measures of quantifying construction speed may be called "macro speed indicators" compared to "micro-indicators", such as the productivity, or "meso-indicators", such as average floor cycle times (Chan & Kumaraswamy, 1995).

Construction speed can be used as a useful and robust benchmark for contractor performance comparison purposes, in the same or a different location. Some research works have compared the construction performance obtained with projects from different countries, although international variations in factors such as economic, legal, cultural, technological, managerial and environmental aspects, along with the uniqueness of construction products and the uncertainties in construction process, make international construction comparisons difficult (Xiao & Proverbs, 2002). Gale & Fellows (1990) claimed that the construction speed of the Broadgate project in London (627 m^2/week) was faster by 50% than American building projects. Proverbs et al.

(1998) designed a model building and a structured questionnaire to compare the construction time performance between contractors of France, Germany and the UK. The study results suggest that projected completion times are shortest in France, followed by Germany and the UK. Along the same lines, Xiao & Proverbs (2002) developed a similar study based on a survey of contractors in Japan, the UK, and the United States (US) and found that Japanese contractors achieved shorter construction times and higher levels of time certainty than UK and US contractors. Dursun & Stoy (2011a) conducted an evaluation of the construction speed performance between residential and office projects developed in the UK and Germany. The analysis results indicated that project location causes a significant variation in the mean response when factors related to the type of facility, standard, and height are taken into account.

Another issue to be considered is the possible variation of productivity in the construction industry over time, since an improvement in productivity will produce an increase in construction speed. Consequently, the construction speed of a building will vary depending on the moment when the project is developed. In this regard, Gale & Fellows (1990) mentioned that the average speed increased in the UK from 157 m^2/week to about 169 m^2/week during the 10 years up to 1990. Ng et al. (2001) reported that the productivity of the construction industry in Australia experienced a 9% increase between 1978-1990, while the construction labour productivity grew at an annual rate of 1.9% per year between 1975 and 1990. Chau (1993) developed a similar study in Hong Kong that suggests that the long-term productivity growth was around 2% per year. Ng et al. (2001) also compared the results of their study with previous research to find the extent of changes in construction time performance in Australian construction projects. This comparison showed a great improvement in construction speed over a period of 40 years. In particular, the public sector showed an improvement by up to 132%. One of the principal factors contributing to the increase of construction speed is the extensive standardisation of designs and procedures (Chan & Kumaraswamy, 1995).

Regarding the type of project, Ireland (1983) stated that it can be expected that large buildings will be built faster, per unit area, than small buildings projects and also that buildings of a particular area tend to be more quickly constructed when they have fewer floors. Many construction practitioners have a view that public projects are likely to take longer to be built than similar private projects (Walker, 1994) and some researchers have proved the veracity of this statement (e.g., Chan, 1996; Sidwell,

1982). Kaka & Price (1991) researched this aspect by analysing 801 UK projects and concluded that construction speed in public projects was faster than private ones of similar construction cost. In contrast, the research undertaken by Walker (1995) revealed that government clients had no significantly better or worse construction time performance than private clients.

Based on what has been discussed above, it can be inferred that construction time performance is context-specific (Le-Hoai et al., 2013). Therefore, it is important to understand that the performance of a particular project in terms of construction speed can be assessed only by comparing it with other projects of a similar type, and also that there exist differences of construction speed according to the physical location and the time period in which a project is being carried out.

For contractors the increase in construction speed improves profitability and provides a competitive advantage (Dursun & Stoy, 2011a). At the same time, it plays a vital role in the success of construction projects since durations that are longer or shorter than planned can have a negative impact on project objectives (Stoy, Dreier, et al., 2007), on the understanding that underestimation or overestimation of construction time will lead to misleading information about the expected economic results (Dursun & Stoy, 2011b). The use of indicators to evaluate the construction speed at early project phases could help to increase the probability of project success. However, there is a lack of models supporting this kind of indicators (Stoy, Pollalis, et al., 2007).

On the basis of the foregoing, in this research the construction speed variable was selected as an alternative output variable to generate predictive models. Furthermore, this thesis adopts the definition of construction speed provided by Stoy, Dreier, et al. (2007) and the relationship between construction speed (*CS*) and construction time (*T*) is given by Equation 1.

$$CS(\mathrm{m}^2/\mathrm{month}) = \frac{GFA(\mathrm{m}^2)}{T(\mathrm{months})} \quad (1)$$

II.4.3 Parametric Models

Predicting construction time by using mathematical models and understanding the relationships between construction speed and different project-related factors

represent a problem of constant concern and interest for both researchers and practitioners. Therefore, a large number of predictive models that attempt to estimate the construction time of building projects can be found in the literature, and most of them are based on the use of project scope factors.

Models for predicting construction time have been developed using different approaches. However, linear regression models have been the most used. These models are formulated using statistical analysis of historical data and show the cause-and-effect relationship between the dependent variable (construction time or construction speed) and the independent variables (factors influencing construction time or construction speed). Raftery (1990), cited in Chan (1998), suggested that linear regression models can yield more readily applicable results to traditional construction and be useful in construction time estimations.

In this section of the thesis, some of the most important linear regression models that can be found in the literature are detailed in chronological order, starting from the BTC model developed by Bromilow et al. (1980). For each presented model the predictive variables that make it up and the size of the samples used for its development are analysed, with particular reference to the use or non-use of independent data samples to validate the models. In addition, and for the purpose of achieving greater clarity in the interpretation of the presented predictive models, a common nomenclature has been used on all of them, to the greatest extent possible. For example, the *T* variable is used to designate the construction time in all models, although each model uses a different unit of measure (days, weeks, or months), which is identified individually for each model.

Bromilow et al. (1980)

The model generated by Bromilow et al. (1980) analysed the time-cost data for a total of 419 building projects developed in Australia, of which 290 were government projects and 129 were private ones. However, 13 government and 11 private projects could not be used as they contained serious anomalies. This model describes the average construction time as a function of project cost and is defined by the following equation:

$$T = K \cdot C^B \quad (2)$$

where:

T = duration of the construction period from the date of site possession to practical completion in wording days.

C = final cost of the building project in millions of Australian dollars (A$), adjusted to a price index.

K = a constant indicating the general level of time performance for a $1 million project, when *C* is measured on the June 1979 price basis.

B = a constant describing how the time performance is affected by project size as measured by cost.

The final relationship derived for the different types of building projects is shown in Table II-1.

Type of building	***K***	***B***
Government	358	0.30
Private sector	238	0.33

Table II-1. Statistical data of the Bromilow et al. (1980) model.

The resulting model indicates that the project scope, measured as the total construction cost, partially determines the construction time of a building project. It provided a basis for all project participants to predict a probable duration of the construction process in working days, given the estimated cost of the project. For statistical analysis, the BTC model can be rewritten in the natural logarithmic form as follows:

$$\mathrm{Ln}(T) = \mathrm{Ln}(K) + B\mathrm{Ln}(C) \quad (3)$$

Ireland (1985)

Ireland (1985) used a sample of 25 high-rise office building projects in Australia for the purpose of analysing the effects of managerial actions on the objectives of reducing time, reducing cost and increasing the quality of these projects. The number of floors in the data set ranged from 14 to 50. He came to the conclusion that the best predictor of the average construction time for high-rise commercial buildings is the

construction cost expressed in the form of the BTC model by the following equation (R^2=0.576):

$$T = 219\, C^{0.47} \tag{4}$$

where *T* is the construction time measured in days (excluding days when no work was done, such as Sundays and public holidays, but including days on which strikes were held) and *C* is the construction cost in millions indexed to June 1979.

Besides the BTC model, Ireland (1985) also generated other prediction models. However, due to some missing values in the measurement of some variables, the regression analysis was only developed using subgroups from the data set. By using the largest subgroup (23 cases) three models were built to predict construction time (R^2=0.73), construction time per square metre (R^2=0.61) and building cost per square metre (R^2=0.47). The equations that define these models are:

$$T = 581.8 + 22.6 \cdot COMPINDX - 27.8 \cdot CPDD + 79.0\, GFA \tag{5}$$

$$TPSM = 441.5 - 22.6 \cdot DCOORD2 - 99.4 \cdot GFA + 18.5 \cdot COMPINDX + 46.8\, NS \tag{6}$$

$$CPSM = 3.41 + 0.061 \cdot QUALITY - 0.082 \cdot CPDD + 0.065 \cdot COMPINDX \tag{7}$$

where *TPSM* is days to construct 10,000 m^2, *CPSM* is construction cost in millions per 10,000 m^2, *COMPINDX* is the complexity of the form of construction measured using the constructed score of a questionnaire, *CPDD* is the construction planning during the design phase, *GFA* is measured in m^2, *DCOORD2* is the co-ordination of the design and construction teams at the design-construction interface, *NS* is the number of storeys and *QUALITY* is the architectural quality. According to Ireland (1985), architectural quality is both complex and to some extent subjective and, consequently, it was measured by asking three architectural practitioners to rank the buildings and then adding these rankings to form an interval score.

Ireland (1986)

Ireland (1986) used GFA as a measure of project scope in a research study that compared the construction speed of 15 office and 7 hotel buildings in the US with 10 Australian office and 4 hotel buildings. The study results showed that GFA is a better indicator of project scope than construction cost. The following models were derived from the study:

US offices (n=15; R^2=0.81).

$$CS = 0.042 \cdot GFA^{0.724} \tag{8}$$

US hotels (n=7; R^2=0.93).

$$CS = 24 + 0.0013 \cdot GFA \tag{9}$$

Australian offices and hotels (n=14; R^2=0.92).

$$Log_{10}CS = -5.72956 + 2.96889 \cdot (Log_{10}GFA)^{0.6124} + (2.93390/NS) \tag{10}$$

where *CS* is the construction speed, which is measured as square meters of GFA per the equivalent of a 8-hour working day, and *NS* is the number of levels excluding the roof.

Therefore, these models showed that greater GFA increases construction speed, while greater number of storeys decreases construction speed.

Chan & Kumaraswamy (1995)

Chan & Kumaraswamy (1995) studied 37 government buildings and 36 private building projects developed in Hong Kong. In particular they investigated the relationships between the duration of the construction period and the project scope factors represented by construction cost, GFA and number of storeys. All the projects were used to generate several prediction models.

First of all, the BTC model (Equation 2) was evaluated. Table II-2 shows the results and, according to the coefficient of correlation, public buildings generate a better BTC model than private buildings. The authors of the study justify these results

explaining that there are more controls and standardisation on government projects, which reduce deviations from building project terms. Although the values of *K* and *B* for government and private projects are reasonably comparable with those achieved by Bromilow et al. (1980), it must be noted that when applying the BTC model direct comparison of *K* values is not realistic due to differences in currencies and construction cost levels between different countries (Chan & Kumaraswamy, 1995).

Type of building	***K***	***B***	***R***
Public sector	216.3	0.253	0.79
Private sector	250.9	0.215	0.65

Table II-2. Statistical data of the Chan & Kumaraswamy (1995) model using cost.

Subsequently, Chan & Kumaraswamy (1995) developed a similar model to the BTC model, using GFA instead of construction cost (Equation 11).

$$T = L \cdot GFA^{M} \tag{11}$$

where *T* is the duration of the construction process measured in working days, *GFA* is measured in m^2 and *L* and *M* correspond to the constants *K* and *B* of the BTC model.

The study results (Table II-3) showed that GFA is also a significant project scope factor affecting the construction time of a building and this model obtains similar correlations to that of the BTC model.

Type of building	***L***	***M***	***R***
Public sector	89.8	0.203	0.77
Private sector	153.6	0.150	0.69

Table II-3. Statistical data of the Chan & Kumaraswamy (1995) model using GFA.

The number of storeys was also analysed because of the critical importance of the "floor cycle time" restrictions (Chan & Kumaraswamy, 1995), and a model similar to the BTC model was generated (Equation 12).

$$T = F \cdot NS^{G} \quad (12)$$

where *NS* is the number of storeys and *F* and *G* correspond to the constants *K* and *B* of the BTC model.

The results (Table II-4) showed that the number of storeys has a weaker relationship with the construction time of a building than cost and GFA.

Type of building	***F***	***G***	***R***
Public sector	405.8	0.222	0.60
Private sector	318.5	0.243	0.63

Table II-4. Statistical data of the Chan & Kumaraswamy (1995) model using number of storeys.

Finally, construction cost and GFA were used simultaneously using multiple regression analysis (Equation 13).

$$T = K \cdot C^{B} \cdot GFA^{M} \quad (13)$$

According to Chan & Kumaraswamy (1995), the combined model was confirmed to be significant and they recommended its use when the two factors involved in the model were known. However, it should be noted that the correlation values produced by this model (Table II-5) were similar to those obtained by using each of the two project scope factors in an isolated manner.

Type of building	***K***	***B***	***M***	***R***
Public sector	155.1	0.178	0.068	0.80
Private sector	238.1	0.202	0.012	0.65

Table II-5. Statistical data of the Chan & Kumaraswamy (1995) model using cost and GFA.

Walker (1995)

Walker (1995) generated a model for predicting construction time by using a data set of 33 projects developed in Australia. On one hand, these projects were classified into four different types of construction: new works, refurbishment, mixed new and fit out, and fit out. On the other hand, these also were classified into seven different types of building end-use: office buildings, industrial buildings, education related, hospital, hotels, transport facility and entertainment facilities. The model was defined by the following equation (R^2=0.999):

$$\begin{aligned} T = C^{0.481294} \cdot \exp[&(1.187976 \cdot eot\_act) \\ &- (0.488867 \cdot work\_type) \\ &+ (0.105097 \cdot obj\_qual) \\ &- (0.125269 \cdot cr\_people) \\ &+ (0.079837 \cdot cm\_des\_com) \\ &+ (0.104343 \cdot cm\_IT\_use)] \end{aligned} \quad (14)$$

where *T* is the construction time in workdays (actual days worked), *C* is the construction cost in $000s indexed to January 1990 taken at the mid-point of construction period, *eot_act* is the ratio of extensions of time granted to actual construction period, *work_type* is only applicable if the project is a fit-out, *obj_qual* is the case study's data for the client's representative's objective for high quality of workmanship on a 7 point scale, *cr_people* is the case study's data for the client's representative's people-orientated management style measured on a 1 to 7 point scale, *cm_des_com* is the case study's data for the communications management for decision making between the construction and design team measured on a 1 to 7 point scale, and *cm_IT_use* is the case study's data for the effective use of information technologies by the construction management team measured on a 1 to 7 point scale.

According to Walker (1995), the model indicates that the principal factors affecting construction time are related to management and client, and also that fit-out projects appear to be built faster than non-fit-out projects. In addition, it can also be inferred from the model that an improvement in communications between construction manager and design team adversely affects the construction process duration.

Nevertheless, the author indicated that further analysis needed to be undertaken in order to clarify this unexpected result.

Chan and Chan (2004)

Chan & Chan (2004) analysed a set of project scope factors affecting construction duration based on 56 case studies of a specific type of residential block developed in Hong Kong. In order to estimate the overall construction duration, the following equation was generated without using any new data to validate the model:

$$\log_e T = 3.0264 + (0.1236 \cdot \log_e C\ ) + THS + PF + (1.3E-06 \cdot V) - (0.0003 \cdot \frac{GFA}{NS}) \quad (15)$$

where *T* is the overall construction duration measured in months, *C* is the actual construction cost in HK$M, *THS* is the type of housing scheme (-0.0544 for purchase; 0 for rental), *PF* is the precast facade (0 for with facades; 0.0666 for without facades), *GFA* is measured in m^2, and *V* is the total volume of building in m^3.

The equation gave an R^2 value of 0.777. The interpretation of the model for the authors was that, for the same design and quality required, the larger and more complex the building, the higher will be its construction cost. Regarding the negative coefficient associated with the GFA/number of storeys ratio, it implies that a shorter duration is associated with a larger GFA per floor. The fact that the housing blocks under purchase scheme were built faster was mainly attributed to the need of providing sufficient residential flats to meet a huge demand for home purchase. Another possible reason given by the authors was the use of a strategy proposed by the Hong Kong Housing Authority for shortening construction times, whereby the contractors were responsible for both the design and construction of the building project. Finally, according to Chan & Chan (2004), the use of precast facades in building construction has a high potential to reduce the predicted construction time. This means that the use of appropriate construction methods and technology, taking into consideration the principles of standardisation and prefabrication, can produce a deep impact on the construction time of building projects.

Love et al. (2005)

Love et al. (2005) analysed 161 construction projects that were completed in various Australian States. Four different project types were used in the study: (1) new build, (2) refurbishment/renovation, (3) fit out, and (4) new build/refurbishment. The analysis was performed between project duration (weeks) and the project scope factors represented by the following variables: project type, procurement method, tender type, GFA, and number of stories. All projects were used to generate the models without using any data set to validate them.

Based on the research developed by Ireland (1983), Love et al. (2005) used both GFA and number of storeys in order to predict construction speed measured in terms of time per unit of area. As a result the following model was generated:

$$\mathrm{Log}(T/GFA) = 3.178 - 0.726 \cdot \mathrm{Log}(GFA) + 0.142 \cdot \mathrm{Log}(NS) \quad (16)$$

where *T* is the construction time in weeks.

Based on Equation 16 construction time can be estimated using the following model:

$$\mathrm{Log}(T) = 3.178 + 0.274 \cdot \mathrm{Log}(GFA) + 0.142 \cdot \mathrm{Log}(NS) \quad (17)$$

The interpretation of the model (adjusted R^2=0.96) by the authors was that construction speed tends to decrease when GFA decreases and/or the number of storeys increases. In addition, Love et al. (2005) also argued that cost is a poor predictor of project duration, since it is not possible to know the actual cost of the project before its completion. Therefore, they stated that it is more sensible to predict construction time using GFA and the number of storeys, rather than construction costs.

Stoy, Dreier, et al. (2007)

Stoy, Dreier, et al. (2007) developed a model to forecast the construction speed of building projects by using a data set of 115 German residential buildings corresponding to five different types of use: single-family home, multi-family home, employee apartments, boarding house, and retirement home. All data came from a

single source and 15 projects were randomly selected to validate the model. Finally, they proposed a semi-log regression model (adjusted R^2=0.849) expressing the natural log of construction speed as a function of GFA and the project standard (Equation18).

$$\mathrm{Ln}(CS) = 4.753 + 0.0002 \cdot GFA - 0.001 \cdot Standard \quad (18)$$

where *CS* is the construction speed which represents the GFA(m^2)/construction time(months) ratio, and *Standard* is the Cost(€)/GFA(m^2) ratio.

According to Stoy, Dreier, et al. (2007), the positive correlation between construction speed and GFA means that construction speed increases with the increment of project size, which is defined in an appropriate form by GFA rather than by the construction cost. Their interpretation is that large projects allow for the use of production factors (personnel, machines, and construction processes) in a more efficient way. These production factors can be better optimised due to the repetition of the same construction steps (e.g., in high-rise buildings). In contrast to GFA, the project standard is negatively correlated with construction speed, which means that projects with high standards have a slower construction speed (longer construction duration). Stoy, Dreier, et al. (2007) explained that by creating high quality standards for building projects the construction cost will be considerably increased and works will be of a longer duration, but its impact on construction speed is lower than the one produced by GFA.

Stoy, Pollalis, et al. (2007)

Stoy, Pollalis, et al. (2007) analysed a data set comprising 216 German building projects to model construction speed. These projects were classified into nine types of use: cultural, education, health, industrial and retail, office, residential, school and kindergarten, sport, and other types of use. They created a prediction model using 200 of these projects, while the remaining 16 projects were used for validation purposes. This model incorporated three independent variables represented in the following equation (adjusted R^2=0.913):

$$\mathrm{Ln}(CS) = 6.482 + 0.968 \cdot \ln(GFA_T) - 0.361 \cdot NW - 0.469 \cdot \ln(PD) \quad (19)$$

where *CS* is the construction speed which represents the GFA(m^2)/duration(months) ratio, GFA_T is the project size measured as 1,000 m^2 of GFA, *NW* is the number of winters, and *PD* is the planning duration in months.

According to this model, GFA_T contributes to the largest explanatory content and construction speed decreases with increasing number of winters during the construction process. Regarding the planning duration, Stoy, Pollalis, et al. (2007) explained that this concept "*encompasses the period from the start of determining the fundamentals through the completion of the project by way of documentation*" and its causal relationship with the construction speed variable stems particularly from "soft" factors such as requirements, organisation, etc.

II.5 Modelling Construction Time with ANNs

II.5.1 Artificial Neural Networks

II.5.1.1 Basic Concepts

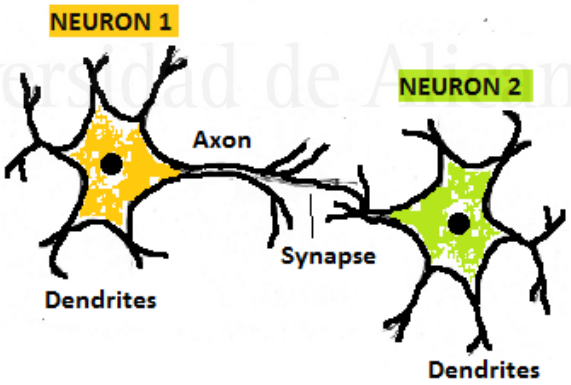


Figure II-3. Operating diagram of biological neurons.

Biological nervous systems process information using a set of units called neurons. These neurons have branches in the form of a tree called dendrites which are responsible for receiving signals from other neurons. The neuron is stimulated through its inputs so that when a certain threshold is reached the neuron is activated

and sends an output signal towards the axon. Synapses are the basic units involved in relationships that arise between neurons. Vesicles are placed at the ends of the synapses containing chemicals called neurotransmitters that help to propagate the signals from one neuron to another. Figure II-3 shows how biological neurons operate.

ANNs are inspired in biological nervous systems and, like them, are composed of a number of interconnected neurons through synaptic weights forming a network (Bhokha & Ogunlana, 1999). They were originally created to solve AI problems, such as speech and hand writing recognition (Fortin et al., 1997). The basic unit of ANNs is the neuron or node. An artificial neuron is an element of calculation interconnected with other neurons and is composed of (see Figure II-4):

- A set of inputs X_j.
- A set of synaptic weights W_{kj}, affecting each neuron input.
- An adder for summing up the inputs, weighted by the respective synapses of the neuron.
- An externally applied bias (b_k), which has the effect of increasing or lowering the net input of the activation function, depending on whether it is positive or negative, respectively (Haykin, 1999).
- An activation function *f* that limits the amplitude of the output of the neuron.
- An output value Y_k.

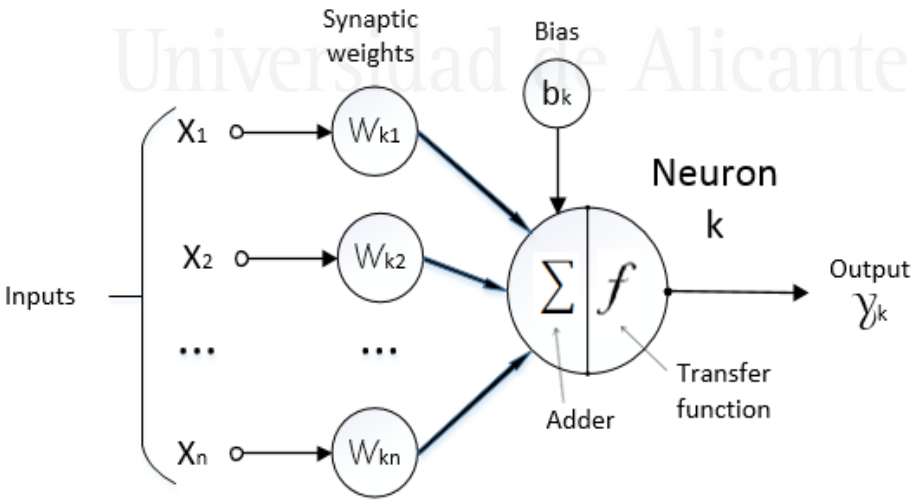


Figure II-4. Operating diagram of an artificial neuron.

According to the scheme presented in Figure II-4, the information provided by inputs is combined taking into account the values of each synaptic weight. This information is modified inside the neurons by a transfer function that produces the node output and which, in turn, can be connected to other neurons in the network. The synaptic weights reflect the strength of the connection between neurons (Castellano, 2009), while the characteristics of the activation functions are important since they define the behaviour of the network (Günaydın & Doğan, 2004). An artificial neuron *k* is mathematically described by Equation 20 (Haykin, 1999).

$$Y_k = f\left(b_k + \sum_{j=1}^{n} W_{kj} X_j\right) \tag{20}$$

where Y_k is the output value, *f* represents the transfer function, b_k is the bias term, *n* is the number of synaptic weights, W_{kj} is the value of the synaptic weight *j* and X_j is the value of the input *j.*

In general, two different classes of network architectures can be identified (Haykin, 1999):

- ***Feedforward neural networks***. This kind of network is acyclic and information propagates forward by the layers of the network through the synaptic weights. There are neither backward connections nor lateral connections. In turn, the feedforward neural networks architectures can be classified into the following two categories: (a) single-layer feedforward networks, which only have an input layer and an output layer, and (b) multilayer feedforward networks that contain and input layer, one or more hidden layers of neurons, and an output layer. If every node in each layer is connected to every other node in the adjacent forward layer the neural network is said to be fully connected. If however some of the synaptic connections are missing from the network, it is said that the network is partially connected.
- ***Recurrent networks.*** A recurrent neural network distinguishes itself from a feedforward neural network in that it has at least one feedback loop. The presence of feedback loops has a deep impact on the learning ability and performance of the network.

Feedforward neural networks are the kind of networks considered in this research work and their nodes are distributed in different layers so that the nodes of a layer are connected only with nodes of the next layer. The input layer is used to introduce input data (independent variables) into the network. Intermediate layers are called hidden layers and the function of the hidden neurons is to extract the useful features from input data (Günaydın & Doğan, 2004). Hidden nodes play a fundamental role in the development of ANN models as they allow neural networks to learn complicated nonlinear relationships between input and output variables (Zhang et al., 1998). The output layer is the last one and it generates the required answer or answers.

Although there are many types of network architectures, within the feedforward neural networks MLP and radial basis function (RBF) networks have been the two architectures most commonly used (Castellano, 2009). Both types of ANNs are considered universal approximators (Cybenko, 1989; Park & Sandberg, 1991). That is, any type of function, which is smooth enough, can be modelled using these architectures.

ANNs need to learn about the problem under study and, to this end, it is necessary to carry out a training process whose objective is to find the set of weight values that will produce the best estimations from the neural network, and the particular procedure used in this process is known as the "learning algorithm" (Haykin, 1999). This procedure is equivalent to the parameter estimation process developed in conventional statistical models (Maier & Dandy, 2000) and, in this sense, the term “training” is equivalent to the term "calibrating” (Fortin et al., 1997). There are several classifications of the learning process but the most common classification is that which makes a distinction between networks with supervised learning and networks with unsupervised learning:

- ***Supervised learning.*** These techniques use training data containing both input data and the expected output values. During the training process, the synaptic weights of neurons are optimised in order to minimise the error between the output provided by the network and the desired output value.
- ***Unsupervised learning.*** These techniques only provide input data, without information on expected outputs. Therefore, the neural network must classify the inputs and outputs based on their similarity to other entries.

Within the supervised learning, the learning based on the error-correction rule is the most common mode of learning. With this mode the adjustment of weights will be carried out depending on the difference between the desired values and those obtained by the network (Castellano, 2009). In this context, convergence is the process of searching a set of weight factors for the ANN model so that prediction errors can be reduced to a minimum (Sun et al., 2003). The theory of nonlinear optimisation is applicable to the training process of feedforward networks (Battiti, 1992).

For the purpose of carrying out the training process of ANNs it is necessary to define an objective function. This objective function, also called the cost function, is minimised during the training process (Maier & Dandy, 2000). Both the sum of squared error (SSE) and the mean squared error (MSE) have been usually used as the objective function, since they are defined in terms of error (Zhang et al., 1998). Nevertheless, MSE is most commonly used (Maier & Dandy, 2000), because it can be easily calculated, penalises large errors, its partial derivative with respect to the synaptic weights can be calculated easily, and it lies close to the heart of the normal distribution (Masters, 1993).

Although the concept of artificial neurons was first introduced by McCulloch & Pitts (1943), it was only after the introduction of the back-propagation (BP) training algorithm by Rumelhart et al. (1986) that ANNs started being widely used in a large number of research fields (Maier & Dandy, 2000).

II.5.1.2 Multilayer Perceptron

The first single-layer perceptron was created by Rosenblatt (1958) and shortly thereafter the delta rule (Widrow & Hoff, 1960) was also developed, which made it possible to find the overall error made during the training process. The combination of several single-layer perceptrons could solve certain nonlinear problems but there was no automatic mechanism to adjust the weights of the hidden layer (Castellano, 2009). The MLP network was presented by Rosenblatt (1962); this is a multilayer feedforward network which can have one or more hidden layers and represents a generalisation of the single-layer perceptron. MLP networks can be fully or locally connected, and although it is common to use the same activation function for each node in the same layer, different activation functions can also be used for each neuron (Castellano, 2009).

Numerous studies about application of ANNs in construction management have showed that MLP are usually the most appropriate type of neural networks for this economic sector (Le-Hoai et al., 2013). This kind of feedforward networks can have one or more hidden layers and its nodes are structured with feedforward connections from the input layer to the output layer (see Figure II-5).

In order to develop appropriate predictive models, determining an optimal MLP network structure is not only one of the most important tasks, but also one of the most difficult (Maier et al., 2010). The configuration of a neural network has a huge impact on its performance (Hegazy et al., 1994) and when designing a MLP network it is necessary to determine the following variables:

- Number of input nodes
- Number of hidden layers
- Number of hidden nodes in each hidden layer
- Number of output nodes
- Activation functions

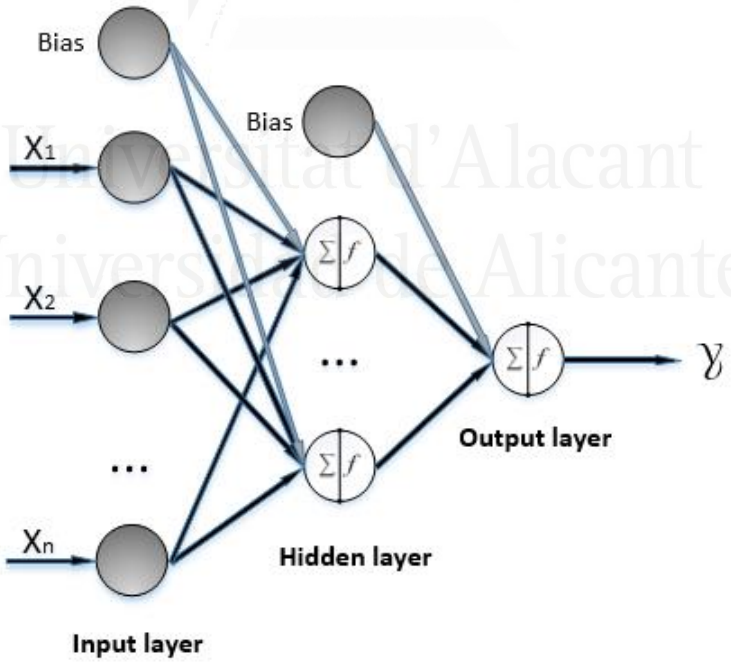


Figure II-5. Architecture of a typical three-layer MLP with one hidden layer.

The MLP network may have more than one node in the output layer, but in the case of estimation problems there is only a single output neuron that produces the

value of the dependent variable. The characteristics of the problem under study determine the number of neurons that will be used in the input layer. In the particular case of prediction models this number corresponds to the number of predictor variables. The critical choice in designing the architecture of a MLP is to select the number of hidden layers and the quantity of hidden neurons in each hidden layer (Maier et al., 2010).

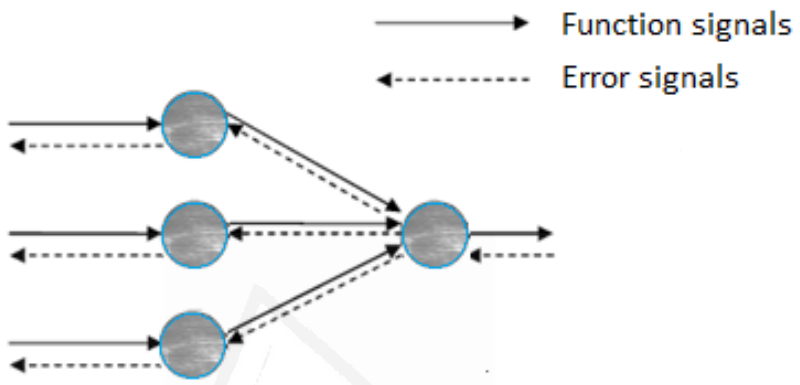


Figure II-6. Diagram of the back-propagation algorithm (Haykin, 1999).

The training process of MLP networks is an unconstrained nonlinear optimisation problem in which the weights are iteratively modified to minimise the error between estimated and actual values. Although the final objective of this process is to find a global solution, currently, there is no algorithm that is able to ensure such a global optimal solution. In the case of MLP networks, the most commonly used training method is the standard back-propagation (SBP) algorithm (Zhang et al., 1998), which is essentially a specific technique for implementing the gradient descent (GD) algorithm (Haykin, 1999). The SBP technique with momentum is adopted by most researchers due to its easy implementation and good ability to generalise (Hegazy et al., 1994). This method, which develops a supervised learning, updates the synaptic weights by performing two paths through the layers of the neural network: (1) forward signal propagation to calculate the error between actual and observed values and (2) backward error propagation to update the synaptic weights (see Figure II-6).

The GD algorithm, also known as steepest descent, is a first-order optimisation algorithm which enables us to find a local minimum of an error function. The weight space and the error function define an error surface. The minimum value of this error

surface is found by moving in the direction of steepest descent at each point. A true GD algorithm will continuously move in the negative direction of the instantaneous gradient. However, in the SBP method the true gradient descent is approximated by taking small steps, which are defined by the learning rate, in the direction of the steepest descent. The operation of the GD algorithm in the SBP method is equivalent to a true GD algorithm as the learning rate approaches zero, in which case the duration of the training process approaches infinity (Wilson & Martinez, 2003).

The correction applied to the synaptic weight connecting a neuron *i* to a neuron *j* is defined by the delta rule through the following equation (Haykin, 1999):

$$\Delta w_{ji} = \eta \delta_j(n) \gamma_i(n) \tag{21}$$

where η is the learning rate, $\Delta w_{ji}(n)$ is the weight correction, $\delta_j(n)$ is the local gradient, and $\gamma_i(n)$ is the input signal of neuron *j*.

The size of the learning rate is critical to the success and efficiency of the learning algorithm (Wilson & Martinez, 2003) and it determines the quantity of weight modification among the neurons during each training iteration. This value ranges between 0 and 1, where a value closer to 1 indicates significant modification, while a value close to 0 indicates little modification (Günaydın & Doğan, 2004). In principle, the use of extreme values for the learning rate should be avoided. The smaller the value of the learning rate, the smaller the changes to the synaptic weights will be. If these changes are too small they will cause a significant decrease in the convergence speed (Haykin, 1999) and the ANN model may be caught on a local error minimum instead of the global minimum (Ghaffari et al., 2006). Conversely, if the value of the learning rate is too high the network may become unstable (Haykin, 1999), which could hamper obtaining the desired convergence due to oscillations of weight changes. A simple way to allow faster learning, but avoiding oscillations, is to modify the delta rule (Equation 21) by including a momentum coefficient (α) to the weight change used in the previous training iteration (Equation 22) (Haykin, 1999).

$$\Delta w_{ji}(n) = \alpha \Delta w_{ji}(n-1) + \eta \delta_j(n) \gamma_i(n) \tag{22}$$

On the other hand, usually, for a given training set two different versions of SBP can be considered:

- ***On-line mode.*** In this mode of learning, the weight updating is performed after use of each training example. The on-line mode is also referred to as sequential or stochastic mode.
- ***Batch mode (off-line).*** In the batch mode, weight updating is performed after the presentation of all the training examples that constitute an epoch.

In general, there is no way of knowing what size of learning rate should be used in both on-line and batch training. At sufficiently small learning rates, on-line and batch training are essentially equivalent (Wilson & Martinez, 2003). The on-line version can safely use a larger learning rate than batch training, which allows it to reach the same level of accuracy in less time (Wilson & Martinez, 2003). Nevertheless, the stochastic nature of this mode makes it hard to establish theoretical conditions for convergence of the algorithm and it is not consistent with the optimisation theory (Møller, 1993). In contrast, by using the batch mode it is possible to obtain an accurate estimate of the gradient vector, so that the convergence to a local minimum can be achieved under simple conditions (Haykin, 1999). In this thesis, when speaking about the SBP algorithm we refer to the batch mode, as it is the only mode which has been used in the development of MLP models.

It is well known that the SBP training methodology's problems are slow convergence, inefficiency, and lack of robustness. Furthermore, it can be very sensitive to the choice of learning parameters like the learning rate and the momentum coefficient (Zhang et al., 1998). In light of the weakness of the SBP algorithm, a number of variations or modifications of it have been proposed in the literature, such as adaptive methods (e.g., Jacobs, 1988), quickprop (Fahlman, 1988), or second-order methods. Second-order methods provide faster convergence, robustness, and good ability to find local minima (Zhang et al., 1998).

II.5.1.3 RBF Networks

Broomhead & Lowe (1988) introduced RBF networks as an alternative approach to MLP networks. The structure of RBF networks is similar to MLP networks with the difference that they are composed of three fixed layers: the input layer, only one hidden layer, and the output layer. The other main differences are the existence of

radial basis functions in the hidden layer neurons and the compulsory use of linear activation functions in the output layer. Each neuron in the hidden layer has a local character determined by the use of radial activation functions with nonlinear basis, which have their own gravitational centres. To perform the local transformation of the input signals, in each hidden neuron *j* a Gaussian basis function $\varphi_j(X)$ is typically used with the following form:

$$\varphi_j(X) = exp\left(-\sum_{i=1}^{P}\frac{1}{2\sigma_{ji}^2}\left(X_i - \mu_{ji}\right)^2\right) \quad (23)$$

where X is the input vector, P is the number of inputs, σ_j is the width of $\varphi_j(X)$ and μ_j is the centre of $\varphi_j(X)$.

The output layer is also fully interconnected with the hidden layer and the output neuron is activated by a continuous linear function. The output neuron (Z_r) results in a linear sum of weights:

$$Z_r = W_{r0} + \sum_{j=1}^{J} W_{rj}\varphi_j(X) \quad (24)$$

where W_{rj} is the weight connecting the output neuron *r* with the hidden neuron *j*, $\varphi_j(X)$ is the *j*th RBF unit and W_{r0} is the threshold of the output unit *r.*

The training of a RBF network is usually a process composed of two stages, where the activation functions are established at the hidden layer during the first stage and the weights connecting the hidden layer to the output layer are directly determined in the second stage (Maier et al., 2010).

II.5.2 ANNs vs LRA

Statistical methods such as LRA have been applied to a wide range of disciplines because of their established methodology, long history of application, availability of software and acceptance among practitioners and academics (Razi & Athappilly, 2005). Linear models hold some advantages because they can be understood in great detail and are easy to implement. However, they may be totally inappropriate if the

underlying function is nonlinear (Zhang et al., 1998). More often than not, the formulation of a nonlinear statistical model to a particular data set is a very difficult task due to the complexity of the real system being modelled. In these cases, a pre-specified nonlinear model may not be general enough to capture all important features of the system and ANNs can be a good alternative approach to identify the real nonlinear function (Zhang et al., 1998). For the purpose of developing ANN models to predict the construction time of building projects, first, planners need to have some knowledge about ANNs and understand the nature of the prediction of construction time at the predesign stage when only little information of the project is available. Then, they must be able to transform such information into the proper forms required by the network (Bhokha & Ogunlana, 1999).

Although ANNs were originated in mathematical neurobiology, several researchers have illustrated their connection to traditional statistical methods and they have been used as a theoretically sound alternative to them (Paliwal & Kumar, 2009). However, a different terminology has usually been used by statistical and ANN modellers, which can be confusing. A useful glossary of commonly used ANN terminologies and their statistical equivalents was given by Maier & Dandy (2000) (Table II-6).

ANN terminology	**Statistical terminology**
Input	Independent variable
Output	Predicted value
Training values	Dependent variables
Errors	Residuals
Training or learning	Estimation
Error function or cost function	Estimation criterion
Pattern or training pairs	Observations
Weights	Parameter estimates
Generalisation	Interpolation and extrapolation

Table II-6. Glossary of ANNs and statistical terminology (Maier & Dandy, 2000).

The functional relationship estimated by ANN models can be written as $y = f(x_1, x_2, \dots, x_p)$ where x_1, x_2,…,x_p are the p independent variables and y is the

dependent variable. In this connection, ANNs are functionally equivalent to a nonlinear regression model (Zhang et al., 1998). Moreover, both parametric and nonparametric regression models can be studied from the perspective of neural networks (Castellano, 2009), so ANNs may be understood as general regression methods (Haykin, 1999). Many classical statistical methods have been rewritten as neural networks. For example, Cheng & Titterington (1994) made a comparative analysis of several ANN models with traditional statistical methods and showed that perceptrons have a strong association with discriminant analysis and regression, while unsupervised networks can be associated with cluster analysis. ANNs are considered as universal functional approximators, because it has been shown that a neural network can approximate any continuous function to any desired accuracy (Cybenko, 1989; Hornik et al., 1989).

ANNs have certain advantages over conventional linear regression methods when generating models to predict construction time of building projects at the early stages of project development. The learning ability of ANNs allows them to solve complex problems (Günaydın & Doğan, 2004) whose analytic or numerical solutions cannot be explicitly represented in mathematical terms (Boussabaine, 2001a; Günaydın & Doğan, 2004) or whose explicit formation causes a loss of sensitivity because of over-simplification (Boussabaine, 2001a). Although any forecasting model assumes that there exists an underlying (known or unknown) relationship between the inputs and the outputs (Zhang et al., 1998), the behaviour of ANNs is established by the data through a learning process, without establishing a rigid model to which data must be adjusted. If a linear relationship between the dependent and independent variables is appropriate, the results of the ANN model will be very similar to those obtained with a linear regression model.

One could argue that linear regression models are a special case of certain types of neural networks. Nevertheless, linear regression is model-driven, which means that the structure of the model has first to be determined before the unknown model parameters can be estimated (Maier & Dandy, 2000). Moreover, both the data used to generate the model and the residuals generated by it must meet several statistical assumptions to be able to validate the generated models. Conversely, ANNs are nonlinear data-driven approaches which can modelling linear and nonlinear systems without having to follow a specific statistical distribution to find out the relationships between the variables under consideration (Wang & Gibson, 2010), i.e., ANNs do not depend on assumptions about functional form, probability distribution, or smoothness

(Smith & Mason, 1997). Thus, they are a more general and flexible modelling tool for forecasting than the traditional statistics methods (Zhang et al., 1998). Although one of the main problems associated to the data-driven modelling approach is that observations are often masked by noise, even when the data to be processed contains errors or is incomplete ANNs can provide meaningful answers (Bhokha & Ogunlana, 1999).

Another benefit is their ability to generalise, being understood as the capacity to produce reasonable outputs for new inputs which have not been used during the learning process (Haykin, 1999). However, since ANNs are data-driven and model-free they may suffer high variance in the estimation, as they may be too dependent on training data (Zhang et al., 1998). The trade-off of any model development is the bias/variance dilemma (Geman et al., 1992). A model that is under-parameterised (or incorrectly parameterised), results in a biased model. A model that is over-parameterised has high variance (fits the training data set well), but has poor ability to generalise (validation data set). This bias/variance trade off becomes particularly obvious when working with small data sets where a smooth form is barely perceptible from the variability of the data. In the case of a simple linear regression model the bias is the assumed linear functional form, while the variance is the determination of the slope and intercept parameters using the training data set. However, for ANN models the bias and variance problem is less clear because neural networks have many more parameters that need to be estimated than traditional statistical models, but are redundancy tolerant (Smith & Mason, 1997).

ANNs do not attempt to replace human experience and judgement. Instead, they introduce the possibility of applying new approaches and methodologies to allow the planner to forecast the construction time in a more productive and efficient manner (Bhokha & Ogunlana, 1999). Although ANNs offer a promising alternative approach to the traditional statistical methods, like the latter they also have weaknesses. It cannot be expected that ANNs will predict everything well for all problems since probably there is not a single best forecasting method that will be equally suitable to all situations (Zhang et al., 1998). On one hand, in the case of static linear processes with little disturbance, ANNs might not be better than linear statistical methods. On the other hand, finding a parsimonious model for a real problem is critical for all statistical methods, but it is particularly important for ANNs as they often suffer overfitting problems, due to the large number of parameters that need to be estimated, and

because parsimonious models also have a more important generalisation ability (Zhang et al., 1998).

Another disadvantage of neural networks is the problem related to the replication of results. Training a neural network is an algorithmic procedure and the results can be replicated as long as one uses identical computer codes, the same initial weights, the same training data, and the same deterministic method of presenting the data during training. However, if one of these parameters is altered, the resulting neural network will almost certainly be different from the original one. This is one of the aspects that converts the development process of ANN models into an "art" (Smith & Mason, 1997).

Although ANNs can model nonlinear relationships between input and output data, the foundations supporting these relationships are unclear. The network parameters (i.e., weights, transfer functions, topology, etc.) cannot explain the modelled relationships by the trained network in a rational way, and this problem reduces confidence in ANN modelling techniques (Boussabaine, 2001a). This is the main reason why the ANN approach has been considered a black-box technique. In this respect, statistical models are clearly superior to ANNs as they allow for proper interpretation of the coefficients of individual variables. Moreover, due to the parametric assumptions of these models, inferences can also be drawn regarding the significance of certain variables (Paliwal & Kumar, 2009). If the objective of modelling is to explain the underlying process that produces the relationships between the dependent and independent variables, it would be better to use a more traditional statistical model. In addition, the knowledge acquisition process of ANNs has also been recognised as a very time-consuming.

According to Paliwal & Kumar (2009), in most of the comparatives studies developed in the literature, ANNs outperformed other methods, or at least worked just as well as them. Nevertheless, these authors stated that most of the papers revised by them seemed not to have used the statistical techniques optimally.

The combination of the features of both ANNs and statistical techniques has also been considered in many studies for the purpose of improving the predictive performance of the models. In this regard, these two techniques can be understood as complementary methods for generating models rather than as competing methods (Paliwal & Kumar, 2009).

II.5.3 Applications of ANNs in the Construction Industry

As has been addressed in previous sections, in order to provide proper tools for estimating construction time and minimise the subjective effects in such estimation, so far most authors have developed parametric models by using LRA and data derived from completed real projects. Although these parametric models have shown a good balance between the difficulty of developing estimation models and the forecasting accuracy obtained with them, it must be pointed out that they represent a great simplification of the complex relationships which control the duration of the construction process in building projects (Sousa et al., 2014). Moreover, their accuracy and trustworthiness is basically limited by some assumptions inherent in the LRA technique (Jafarzadeh et al., 2014). In this regard, it has been proven that problems which involve complex nonlinear relationships can be better solved by ANNs than by conventional methods (Rumelhart et al., 1994). Consequently, there is a growing tendency in recent years to use ANNs to generate better predictive models (e.g., Chen & Huang, 2006; Le-Hoai et al., 2013). They have been considered as an AI research area of interest for the construction industry because they can improve automation efforts (Moselhi et al., 1991).

The use of ANNs in construction management goes back to the early 1980's (ElSawy et al., 2011). Many publications have described the underlying theory to develop ANNs and its applications from different perspectives. A review of the BP algorithm with suggestions on how to develop practical neural network applications was presented by Hegazy et al. (1994). Moselhi et al. (1991) described basic neural networks architectures and discussed their potential applications in construction engineering and management, like estimation, classification, or optimisation tasks. He also highlighted the possibility of integrating ANNs and expert systems. Boussabaine (1996) reviewed ANN techniques and concepts used in construction management that could help researchers to identify opportunities where ANNs are applicable in assisting decision makers. He concluded that, in view of their extensive applications, ANNs would continue making impressive gains in the field of construction management. Paliwal & Kumar (2009) carried out a review of articles that involved a comparative study of feedforward neural networks and statistical techniques used for prediction and classification problems across a variety of applications and areas.

So far, according to Jain & Pathak (2014), ANNs have been used in construction engineering and management for prediction, risk analysis, decision-making, resources optimisation, classification, and selection. Kamarthi et al. (1992) made use of a two-layered BP network to develop a computer system that provided the selection of vertical formwork systems for a given building site. Murtaza & Fisher (1994) presented an approach for decision-making in construction modularisation using an unsupervised neural network with the Kohonen algorithm. Chao & Skibniewski (1994) developed an ANN-based approach for predicting the adoption potential or acceptability of a new construction technology. Hegazy & Moselhi (1994) used ANNs to develop an optimum mark-up estimation model that derived solutions to new bid situations. Elazouni et al. (1997) developed an ANN model using BP to estimate construction resource requirements at the conceptual design stage. Chua et al. (1997) used ANNs to identify the key management factors affecting budget performance in a project. Al-Tabtabai et al. (1997) made use of ANNs to propose a judgment-based forecasting approach to identify schedule variances from a baseline plan for typical construction projects. Adeli & Wu (1998) developed a regularisation neural network to estimate the cost of reinforced concrete pavement.

The great majority of civil engineering applications of ANNs are based on the BP algorithm, mainly because of its simplicity, and prediction has been its most common use in the field of construction (Jain & Pathak, 2014).

II.5.4 Applications of ANNs to Predict Construction Time and Costs

Among the different topics existing in the field of forecasting using ANNs, perhaps the prediction of construction costs has been the subject most addressed by researchers related to the construction industry (see, e.g., Adeli & Wu, 1998; ElSawy et al., 2011; Goh, 1998; Günaydın & Doğan, 2004; Jafarzadeh et al., 2014; Kim et al., 2004; Yeh, 1998). Some of these non-parametric models have pointed out the existing relationships of construction costs with GFA, number of storeys and duration. For example, Günaydın & Doğan (2004) developed a model to predict the cost per square meter of reinforced concrete structural systems in building projects. For that purpose, the authors of the study used a feedforward neural network with only one hidden layer, the SBP algorithm, and cost and design data from thirty projects of 4–8 storey residential buildings in Turkey. 20% of the observations (6 cases) were selected

randomly for model validation, while the rest of the projects were used for training the network. The selected model made use of the following eight design parameters as input variables: total area of the building, the ratio of the typical floor area to the total area of the building, the ratio of ground floor area to the total area of the building, number of floors, the console direction of the building (one or two directions), the foundation system of the building (pier, wall, or slab), the floor type of the building (reinforced concrete or precast concrete), and the location of the core of the building (at the site or middle). Kim et al. (2004) also examined the performance of three modelling techniques for estimating construction costs: MLRA, ANNs, and case-based reasoning (CBR). The actual construction costs of 530 projects of residential buildings were used along with the SBP algorithm to generate predictive models. Eight variables were selected as input data: GFA (m^2), number of storeys, total unit, duration (months), roof types, foundation types, usage of basement, and finishing grades. The data collected was divided randomly into three data subsets: 40 projects for validation purposes, 50 projects for testing and 440 for training the neural network. The best ANN model (with only one hidden layer) gave more accurate estimating results than either the MLRA or the CBR estimating models.

In the same line of thought, Jafarzadeh et al. (2014) developed several ANN models based on significant predictors of the retrofit net construction cost (RNCC). The study used three different feedforward architectures, with one, two and three hidden layers respectively. The SBP algorithm together with data from 158 earthquake-prone schools was used to develop ANN models. 121 schools (75%) were used as the training data set, and the remaining 37 projects (25%) were maintained to constitute the test data set. The logarithmic form of the total area (m^2) and the number of stories were selected as input variables along with five other independent variables. The study results indicated that total area is the key and probably the most important variable contributing to the RNCC prediction.

In contrast to the estimation of construction costs, Bhokha & Ogunlana (1999) identified a notable gap regarding literature on ANNs applied to predict construction time. They applied neural networks for modelling the construction process duration of building projects at the predesign stage. To this end, a data set with 136 building projects was divided into two equal subsets for training and validation purposes. The study made use of the SBP algorithm for training a three-layered ANN (only one hidden layer) and eleven variables were selected as input data: building function,

structural system, functional area, height index, complexity of foundation works, exterior finishing, decorating quality, and site accessibility. Ten of the variables were binary, while only the functional area was a real-value variable. Bhokha & Ogunlana (1999) stated that ANNs show advantages over conventional methods, which use knowledge and experience of experts, in order to significantly improve the construction time estimation.

Le-Hoai et al. (2013) first identified six significant variables influencing construction time performance by using a data set of 70 building projects along with the MLRA technique. These variables were: underground site condition, project management works, estimating works, competency of subcontractor(s), accuracy and completeness of design, and owner's project financing. Subsequently, they compared the prediction accuracy of the models developed by using MLRA with the accuracy of models generated by using MLP networks. The best MLP model used only one hidden layer, the same input variables than the regression model and the training algorithm known as the scaled conjugate gradient (SCG). According to the study results, the authors concluded that although the developed ANN model used a different approach regarding the initial data set, which was divided randomly into two subsets (80% for training and 20% for validation) against the single data set used in generating the regression model, it was able to obtain similar prediction accuracy when compared to the regression model.

Chen and Huang (2006) utilised 132 school reconstruction projects destroyed by an earthquake and demonstrated that the floor area provides a good basis for estimating the cost and duration of school reconstruction projects. In this case, 90% of the collected projects were selected randomly for model development and the remaining 10% was used for validation purposes. The results of this latest study suggested that MLP networks with the SBP algorithm yield better prediction results than regression models.

By using a data set of 75 building projects, Petruseva et al. (2013) compared the BTC model with an ANN model built by using the MLP network. Each modelling methodology made use of a different number of logarithmic input variables. So while the linear regression model took into consideration only one logarithmic dependent variable (real price), the MLP model utilised three logarithmic input variables (real price, contracted time and contracted price). The SCG algorithm was used for training

the MLP models and with both methodologies the initial data set of building projects was divided into two subsets: one for training the model and another to validate it. Nevertheless, authors do not give any indication about the number of projects used in each subset. According to the authors of the study, the comparison results showed that application of MLP models produced significant improvement of the predictive accuracy.

II.6 Conclusion

The comprehensive literature review carried out in this chapter suggests that the construction time of a building project is affected to varying degrees by a great number of factors which need to be prioritised. For this purpose, several categorisations of these factors have been presented in some studies related to building projects. In this regard, most of the literature has identified the factors included in the category "project scope" as key predictors of construction time. Project scope can be understood as a measure of project size. The principal criticism of predictive models based exclusively on project scope factors is that the duration of the construction process is too complex to be estimated using only this type of factors. However, most of the qualitative variables proposed in previous research are hard to incorporate into mathematical models due to the difficulty encountered when attempting to measure them in an objective manner. Consequently, the subjective nature of these qualitative factors has made the development of reliable prediction models a challenge.

In the case of building projects, construction costs and GFA have been the factors most often used to define the project scope and develop models for estimating construction time. Nevertheless, there are no perfect measurement units to define the project scope and there is no consensus in the literature to determine which of these two factors should be used to better estimate the construction time of building projects. On the one hand, construction costs are commercially sensitive and the final cost normally varies from the initial cost estimated before starting the works. On the other hand, the use of GFA may hide the real complexity of the construction works, for example in building projects involving a lot of external works.

It is clear that project size influences the project complexity. However, considering only the size of the project by using GFA or construction costs is not possible to fully assess the complexity of new build projects. A project of a large size might have very

little complexity associated with it. In this sense, the type of building has been related to project complexity and it is expected that different building types will produce different relationships between project scope factors and construction time. Some studies have also found that the cost/GFA ratio, known as “standard”, could be a good indicator of the complexity of building projects. High quality standards will result in lower construction speed. In this respect, also the number of floors has a limited negative impact on construction speed.

Project complexity has also been associated with the form of construction. Nevertheless, the risk of poor construction time performance is not so much associated with the complexity of the form of construction, but rather with how the builder overcomes the difficulties resulting from a complex design. In this connection, it is necessary to take into consideration that an appropriate planning of the resource movement allows for an uninterrupted workflow, which is vital for achieving a good construction time performance in building projects. Therefore, construction planning during the design stage can be considered as a way to reduce project complexity and increase construction speed. In addition, it has been shown that the use of appropriate construction methods and technology, taking into consideration the principles of standardisation and prefabrication, can have an important positive impact on construction speed.

The review of the literature also revealed that there is a debate which revolves around whether it is possible to obtain better forecasting models using construction time as the dependent variable or if it is more appropriate to consider construction speed. In this regard, construction speed can be used as a useful benchmark to compare performance between contractors. However, it is important to understand that there will be differences in construction speed depending on the physical location where a project is built and the moment when such a project is developed. The use of indicators to evaluate the construction speed at early project phases could help to increase the probability of project success, but nevertheless there is a lack of models to support this kind of indicators.

So far, most authors have developed parametric models by using LRA. Although these parametric models have shown good performance, they might represent a great simplification of the complex relationships which control the construction process of building projects. Moreover, their accuracy and reliability is limited by some

assumptions inherent in the LRA technique. In this regard, it has been proven that problems which involve complex nonlinear relationships can be better solved using ANNs than by conventional methods. This has led to an increasing trend in recent years for using ANNs to develop better predictive models, so that LRA and ANNs have become two competing modelling methods. Nevertheless, in contrast to the estimation of construction costs, there is a significant gap regarding the literature related to ANNs applied to predict construction time. Moreover, the few existing studies have used different learning algorithms and different types of data division.

Chapter III

Research Methodology

III. Research Methodology

III.1 Introduction

The purpose of this chapter is to explain and justify the theoretical basis and practical aspects of the research methodology adopted in this thesis in order to achieve the objectives and hypotheses proposed in Chapter I. The development of a research is a process that must follow a logical sequence of steps which form its methodology (Chan, 1998) and during this process observations need to be explained but also explanations must be tested against facts (Walker, 1994). The general research approach adopted in this thesis has been to use a large sample of building projects to develop linear and nonlinear predictive models and use another smaller sample of projects to validate them. At the same time, it was necessary to develop an appropriate statistical analysis of used data sets so that this research work can be replicated in further studies. Furthermore, the research methodology proposed herein was developed from concepts and ideas reported in the literature. In this regard, the works undertaken by Bromilow (1969), Walker (1994), Chan (1998), Love et al. (2005) and Stoy, Dreier, et al. (2007) provided much of the basis for studying the impact of project scope factors on the construction time of building projects.

This chapter is organised as follows: first of all, general concepts about the generation process of mathematical models are discussed in Section III.2 with the purpose of providing a solid basis for selecting the appropriate modelling techniques and research approaches that will be used in this thesis. Secondly, sections III.3 and III.4 argue the hypothesis testing procedures and levels of statistical significance used in this research work. Then, Section III.5 analyses the data sets used for testing the hypothesis established in Chapter I. Subsequently, Section III.6 presents the variables selected for the study, which represent some of the main project scope factors related to building projects. Thereafter, the modelling tools selected to develop predictive models are discussed in Section III.7. And lastly, the framework proposed to evaluate the predictive performance of linear and nonlinear models is explained in detail in Section III.8.

III.2 Generation of Mathematical Models

According to Vidal & Marle (2008), "*a system is an object, which, in a given environment, aims at reaching some objectives by doing an activity while its internal structure evolves through time without losing its own identity*" and, thus, projects can be considered as systems (Vidal & Marle, 2008). When studying a system it is important to know the behaviour of its elements, which also involves studying the relationships between them (Villacampa et al., 2007). Models attempt to represent a system and, in general, a model can be physical or mathematical. Physical models represent physical reality using physical materials, while in some cases mathematical models can reduce a system to an equation which represents to a greater or lesser extent the actual behaviour of the system. According to Williams (2002), a mathematical model "*represents or describes perceptions of a real system, simplified, using a formal, theoretically based language of concepts and their relationships (that enables manipulation of these entities), in order to facilitate management, control, or understanding of that system*". Among other features, a good model should (Williams, 2002):

- Be empirically based, informed by data.
- Be theoretically sound, agreeing with the body of knowledge of management science.
- Be coherent, because the elements of the model should not contradict each other.
- Be simplified, to such an extent that it can be modelled and analysed, although at the same time the model has to be of "requisite variety". The judgement of simplicity versus complexity of the model should be made taking into account the purpose of the model, which also implies the desired degree of replication of actuality.
- Address the complexity of the real system. Otherwise the system will behave in ways not predicted by the model.
- Add value, in order to help us understand the phenomena under study.
- Impact on decisions, implying that a good model will influence management decisions.

Because it is generally very difficult to propose mathematical laws describing the exact operation of a real system, it is useful to develop models from statistical analysis of experimental data related to the system being modelled. Statistics is a simple but an important research tool which uses small samples of a population to make inferences about the statistical characteristics of that population (Leedy & Ormrod, 2010). After inspecting and debugging the databases of building projects and selecting the set of variables to be studied, it is necessary to select the statistical tools that will be applied in order to analyse the relationships among the selected variables and test the hypotheses established in the research. The data analysis technique adopted will depend on the complexity of the research problem (Chan, 1998). There is no point in modelling unless it affects decision-making and the types of decisions that will be affected provide the reasons for building the models and determine how they should be developed (Williams, 2002). In the particular case of this research, the real system under analysis corresponds to the construction process of building projects, while the mathematical models that we intend to develop aim to predict the duration of this process on the basis of existing relationships between the variable representing the construction speed and the main factors of project scope. At the initial stage of project development, when only basic information is available, the proportionate forecast by these models can serve as a control parameter to verify if the construction deadlines proposed by any of the agents involved in the building project are realistic and, if necessary, to take special measures to modify the construction speed.

A mathematical model is an abstract representation of a system and the modeller needs to obtain formally the relationship between a variable *y* and a set of predictor variables $x_1, x_2, \dots, x_n$, which can be expressed from the linear or nonlinear relationship $y = u(x_1, x_2, \dots, x_n)$. The process of modelling highlights gaps in knowledge, and motivates the modeller to try to fill them (Williams, 2002). Generically, one of the basic aspirations being pursued when a mathematical model is built is to obtain a tool to predict different situations of the system being modelled as a function of some particular input parameters (Verdú, 2004). Numerous methodologies have been developed to determine families of linear and nonlinear models. In general, first, all methodologies build mathematical equations that seek to fit to the experimental data as much as possible. Then, the adjustment and the error existing between experimental data and the outputs produced by the model are quantified. According to

Verdú (2004), the process of modelling a system is developed in four phases (Figure III-1):

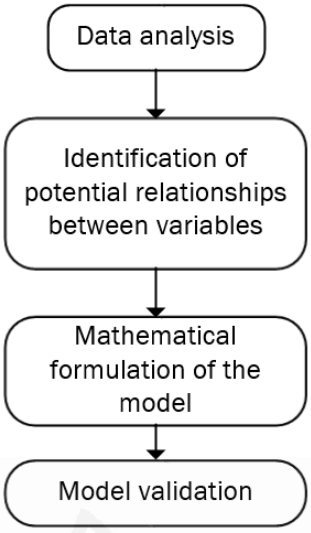


Figure III-1. Outline of the general process to generate predictive mathematical models.

- **Data analysis.** This phase will allow us to identify data which could be ruled out due to erroneous measurements, invalidity of data, etc. In addition, it is also necessary to identify the status and meaning of the data set used to build the models, defining what is being measured, and what it means regarding the real system. Moreover, it is necessary to remember that any data collected is only a sample, and only from a particular moment in time (Williams, 2002).
- **Identification of potential relationships between variables.** This point is left to the skilled in the art to be explained with the model. These experts will validate if the obtained relationships make sense or are valid to build the desired model.
- **Mathematical formulation of the model.** The methodologies used for obtaining mathematical models can be divided into two broad categories: (1) techniques of linear regression analysis and (2) techniques of nonlinear analysis.
- **Model validation.** Once the model has been built, it should first be determined whether such a model is reliable and valid for its application before using it. After the validation process, a finding of invalidity of the model would lead us back to the identification of relationships and their mathematical formulation.

According to Williams (2002), when modelling it is also necessary to take into consideration the issue related to achieving a good trade-off between complexity and simplicity of the model. There are two key balancing principles generally accepted in modelling. The first one is known as “Occam’s razor” and it states that among competing hypotheses that predict equally well, the one with the fewest assumptions should be selected. However, it is necessary to balance this first principle with the principle known as “requisite variety”. In accordance with this second law, in order to avoid overly simplistic models, the predictor variables that make up the model must be able to represent well enough the features of the system that the modeller wants to model and, thus, it must have sufficient variety to replicate the project complexity (Williams, 2002).

III.3 Hypothesis Testing

Hypothesis testing is a fundamental procedure in statistics which allows us to evaluate two mutually exclusive statements (hypotheses) about two or more data sets in order to determine which statement is best supported by the statistical significance. The alternative hypothesis (H_1) is the researcher's hypothesis and states that the difference between groups is greater than would have been expected by chance. Conversely, the null hypothesis (H_0) is the negation of the alternative hypothesis and states that the difference between groups is within normal limits and is not statistically significant. In statistical hypothesis testing, a Type I error (false positive) is the incorrect rejection of a null hypothesis that is true, while a Type II error (false negative) is the failure to reject a null hypothesis that is false.

In the particular case of two samples, the *t*-test has often been used to develop the hypothesis testing and such a parametric test is based on the joint assumption that the populations are normally distributed and have the same variance (Markowski & Markowski, 1990). Many studies has been carried out to evaluate its performance when the assumptions are violated, and the general conclusion of these studies was that for samples of the same size the *t*-test is very robust, although for samples of different size there are situations where it performs poorly (Markowski & Markowski, 1990). Even with conventional significance levels, it has been stated that this test may often make Type II errors if the sample size is small or the assumptions required for the test to be valid are violated (Sprent & Smeeton, 2001). In the case where more

than two groups need to be compared, it is possible to use the well-known analysis of variance (ANOVA) test with which the same assumptions considered in the *t*-test must also be satisfied. Nevertheless, some studies about the robustness of the ANOVA test in the case of violations of normality and variance homogeneity assumptions have also showed that the performance of this parametric test might be compromised in many cases (Lix et al., 1996). Furthermore, the abstract notion of random sampling from an infinite population often works well in practice, but is never completely true (Sprent & Smeeton, 2001) and the parametric assumption of normality is particularly worrisome for small sample sizes with $n<30$ (Chan & Chan, 2004).

During the development of this thesis it has been necessary to perform hypothesis testing on several occasions to compare the differences between two or more groups. To this end, the initial intention was to develop parametric tests. However, in most cases we found the failure of one or more of the assumptions and recommendations cited in the preceding paragraph. That is, we found absence of normality, heterogeneity of variances, samples with large difference in size, and/or samples which are too small, as in the case of comparisons carried out with validation data ($n=18$). According to Lix et al. (1996), two approaches can be considered by researchers to make the comparison of different groups when the assumptions of parametric tests are violated: (a) to apply a transformation to the data or (b) to select an alternative test which is insensitive to the failure to comply with the assumptions. Unfortunately, transformations have a number of limitations, mainly due to issues of interpretation (Lix et al., 1996). With regard to the second approach, if the researcher determines that assumptions of the parametric test are not satisfied, it is possible to use an analogous nonparametric procedure instead. Some of the advantages of nonparametric methods are (Hollander et al., 2014):

- Nonparametric methods require few assumptions about the underlying populations from which the data are obtained. In particular, nonparametric procedures need not satisfy the traditional assumption that the underlying populations are normal.
- Generally, nonparametric procedures are only slightly less efficient than parametric tests when the underlying populations are normal, but they can be much more efficient than parametric tests when the underlying populations are not normal.
- Nonparametric methods are relatively insensitive to outlying observations.

Even if a parametric test does not depend critically on the assumption that samples come from a particular distribution, when in doubt, it may be preferable to develop a nonparametric test which needs weaker assumptions (Sprent & Smeeton, 2001). In the light of the above, in this thesis it was decided to carry out the comparison between groups by developing nonparametric tests. In the case of the *t*-test, the Mann-Whitney (MW) *U* test was conducted instead because it has been used in similar studies related to construction projects (see, e.g., Irfan et al., 2011). The MW *U* statistic is given by the following equations (Hollander et al., 2014):

$$U_1 = n_1 n_2 \frac{n_1(n_1+1)}{2} - R_1 \tag{25}$$

$$U_2 = n_1 n_2 \frac{n_2(n_2+1)}{2} - R_2 \tag{26}$$

where n_1 and n_2 are the sample size for samples 1 and 2 respectively, and R_1 and R_2 are the sum of the ranks from samples 1 and 2 respectively. The smaller value of U_1 and U_2 is the one used when consulting significance tables.

The Kruskal-Wallis (KW) test is the most well-known nonparametric alternative to the ANOVA test (Lix et al., 1996). The KW test extends the MW *U* test when there are more than two groups and it was also used in this thesis.

III.4 Statistical Significance

Statistical significance is related to the need to prove the study hypothesis. Generally, as is the case in this thesis, the research results are based on the analysis of one or several of the infinite samples that can be obtained from the reference population, as these results may differ due to chance. The analysis of the level of significance (*p*) allows us to quantify this variability and know if the research findings can be extrapolated to all cases of the population under study (Rebasa, 2003). The *p*-value is the probability of making a Type I error in a hypothesis testing, which is, in turn, the probability of refusing the true null hypothesis and is inversely related to the probability of making a Type II error (Kim & Ji, 2015). Nevertheless, the acceptance of

a hypothesis does not mean conclusive proof, but rather a provisional approach that can be refuted or verified in subsequent experiments.

Significance levels commonly used in the literature are 1%, 5%, and 10% (Kim & Ji, 2015; Ritter & Muñoz-Carpena, 2013). The 0.05 level is nearly universal, while the 0.01 and 0.10 levels are also widely used (Kim & Ji, 2015). However, there is no scientific basis to select any of these values and no reason to believe that they are optimal (Kim & Ji, 2015). In this regard, the selection of a level of significance should be based on the research context, taking into consideration the risk we want to assume when accepting or rejecting the null hypothesis (Ritter & Muñoz-Carpena, 2013). It is also necessary that the hypothesis has a good theoretical basis, since a hypothesis with a good theoretical basis along with a small *p*-value provides enough arguments to keep to it. Moreover, the rejection of the null hypothesis does not suggest any causality if the study design is not solid. Conversely, a non-significant result does not prove that the null hypothesis is true because it is perfectly possible that the null hypothesis is false when too small a sample is used. In this connection, it should be noted that statistical significance depends on the magnitude of the difference we want to test and the sample size (Rebasa, 2003). Leamer (1978) suggested that the level of significance should be adjusted as a decreasing function of the sample for sensible hypothesis testing.

According to the discussion in the preceding paragraphs, the *p*-value should be selected carefully, always taking into account the context of the research and the characteristics of the sample under study, and not as a magic number. In this thesis, a significance level of 5% was selected in a general manner, since most similar studies related to building projects have also used this significance level (Chan, 1998). Nevertheless, in the case of the statistical significance of independent variables used with LRA the *p*-value was increased to 10% because the influence of the selected independent variables on construction time has been proven in many previous studies.

III.5 Data Sets

The project collection used in this study to generate predictive models belongs to a database containing more than 300 projects with different uses, locations and sizes. It is a collection of construction projects carried out in Spain, which was provided by the Spanish company Soft SA. All projects were developed between the late 1980s and

early 2000s. The research focused on building projects and, in particular, on new builds. Projects relating to industrial buildings and rehabilitation/refurbishment projects were discarded for the analysis. Special sports facilities (swimming pools, soccer fields, etc.) and other singular constructions (bullrings, churches, etc.) were also excluded. By selecting projects that are similar in nature it is possible to reduce the diversity of factors affecting construction time (Sousa et al., 2014). In order to establish a basis for fair comparison, all construction costs were updated to February 2010 by using the date of commencement of work and the coefficients of price revision obtained with the polynomial formulas proposed by the Spanish government for the building sector. Initially, 168 projects were considered as valid for the proposed study.

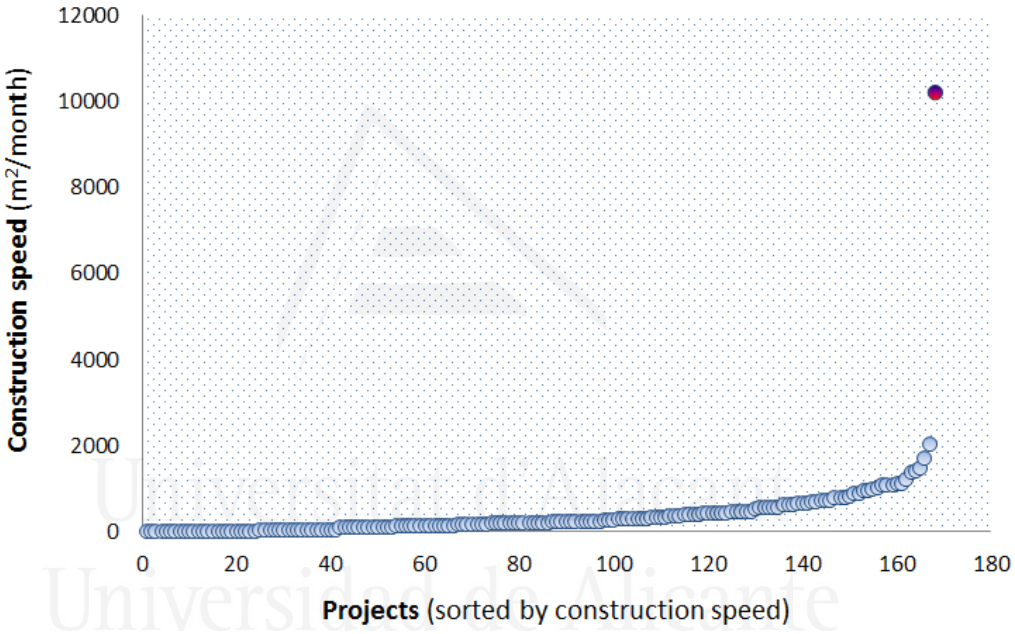


Figure III-2. Graph showing the existence of one outlier (*n*=168).

Notwithstanding the foregoing, the identification of outliers is a necessary process to develop reliable regression models (Chan & Chan, 2004). In the case of linear regression models some cases were detected with statistical values outside the thresholds established in the study and the regression analysis was performed with and without these identified outliers. The removal of these values improved the fit of the models to the projects used for its generation but worsened the accuracy of the models when they were applied to validation projects. In contrast to the linear regression analysis, there is no process formally established to detect outliers in the case of numerical analysis. But after the generation of the first nonlinear models, it

was detected the existence of a project that could be conditioning the predictive ability of the models (see Figure III-2). This view was confirmed by the development of a hierarchical cluster analysis (HCA) with the construction speed variable and the predictor variables representing GFA, number of floors and the cost/GFA ratio. The cluster analysis is a multivariate technique used to classify a set of variables into relatively homogeneous groups so that in each group (cluster) the variables tend to be similar but different from the variables of other groups regarding to some given selection criterion. In particular, HCA is characterised by the development of a hierarchy or tree-shaped structure. The squared Euclidean distance was used as the dissimilarity measure, and the comparison of distances between groups was made by using the method known as the average linkage between groups.

HCA showed the existence of a project that constituted by itself a distinct group with extreme features. The elimination of this outlier enabled us to build nonlinear models with better results of accuracy in both the set of projects used to generate the models and the set of validation projects. By contrast, in the case of linear models with predictor variables in logarithmic form, the removal of this project produced no significant difference of accuracy and adjustment in the validation projects. Consequently, in order to maintain the comparative analysis of methodologies under equal experimental conditions, it was decided to eliminate the project identified as atypical from the research.

Finally, 167 projects were selected to generate predictive models. This group of building projects is referred to hereafter as the calibration data set. These projects were classified into 7 different types of facilities according to their nature and scope. Their features ranged from 1 to 18 total floors, with GFA between 108 and 52,159 m^2, and updated actual costs from € 47,309.08 to € 47,105,476.82 (see Tables III-1 and III-2). Furthermore, similarly to Irfan et al. (2011), the models developed in this thesis were validated using a new sample of 18 health building projects, totally independent from the calibration data set used to generate the models. The set of validation data, hereinafter referred to as validation projects, was provided by the Murcian Health Service, which is the agency responsible for managing the system of public health services in the Autonomous Community of the Region of Murcia (Spain). The features of validation projects ranged from 1 to 3 total floors, with GFA between 500 and 4,302 m^2 and updated actual costs from € 354,405 to € 3,053,197. Their final duration ranged between 8 and 24 months and they were executed between 1992 and 2009.

Type	Description	*n*	Total cost (€)		Total GFA (m^2)		Total number of floors	
			Max.	Min.	Max.	Min.	Max.	Min.
1	Multi-family housing	57	17,980,803.43	456,546.45	41,000.00	633.84	11	2
2	Single family housing	14	674,553.11	82,691.55	383.09	120.00	4	1
3	Offices and commercial	18	47,105,476.82	1,279,724.24	52,259.00	1,216.00	18	1
4	Sports and entertainment	15	3,052,279.92	195,134.05	12,551.92	308.00	5	1
5	Health and wellness	15	5,087,504.25	126,964.35	6,912.00	116.74	6	1
6	Cultural	28	16,856,717.74	313,229.98	23,677.00	374.27	9	1
7	Other uses (not industrial)	20	17,657,541.39	47,309.08	16,451.00	108.00	10	1
	All types	**167**	**47,105,476.82**	**47,309.08**	**52,259.00**	**108.00**	**18**	**1**

Table III-1. Main statistical data of the calibration data set.

Type	Description	*n*	Total cost (€)		Total GFA (m^2)		Total number of floors	
			Median	Std. dev.	Median	Std. dev.	Median	Std. dev.
1	Multi-family housing	57	3,345,112.75	3,220,034.34	8,283.00	7,586.95	7.00	2.02
2	Single family housing	14	203,888.58	160,295.33	231.13	82.93	2.50	0.85
3	Offices and commercial	18	4,583,291.95	12,505,023.55	5,177.50	14,944.13	5.00	5.03
4	Sports and entertainment	15	750,659.68	1,075,098.50	1,439.00	3,020.35	1.00	1.10
5	Health and wellness	15	1,313,781.26	1,497,417.86	1,588.00	2,117.62	2.00	1.51
6	Cultural	28	2,209,193.08	3,821,584.98	3,407.00	5,201.14	3.50	1.57
7	Other uses (not industrial)	20	1,286,737.79	4,884,461.19	1,493.79	4,657.69	4.00	2.81
	All types	**167**	**2,612,707.74**	**5,607,392.29**	**3,561.00**	**8,276.52**	**4.00**	**3.04**

Table III-2. Main statistical data of the calibration data set.

The sample of calibration projects contained only the final construction costs. In this regard, it is necessary to clarify that, apart from using data sets obtained from different databases, the choice of the validation data set was also determined by the fact that the selected validation projects contained planned costs in addition to final construction costs. By using these planned costs, it was possible to carry out a sensitivity analysis to check whether the forecasting accuracy of predictive models is negatively affected when these models use estimated costs in their practical application.

III.6 Tested Variables

Deciding which factors must be considered in the model development is part of the data pre-processing process (Boussabaine, 2001b). The task of variable selection needs to be performed carefully because literature presents general guidelines rather than demonstrating formal procedures. This forces the researcher to use theory, intuition, and common sense when executing this selection process (Dursun & Stoy, 2012). Based on what was mentioned in Chapter II it can be deduced that there is no consensus in the literature on what factors most significantly influence construction time. In addition, some of the variables proposed in previous research cannot be measured objectively and for this reason they could be difficult to introduce in a prediction model (Forsythe et al., 2010). Furthermore, the limitation regarding the type of data available for its study, unfortunately, restricts the possibility of analysing some variables identified in the literature as possible factors influencing construction time.

According to available data, quantitative variables representing the total cost of construction, GFA and the total number of floors (maximum number of floors excluding the roof) were selected for the study. The number of floors above and below ground was also analysed. However, sometimes it may be necessary to use derived variables because some of the significant variables might not be a single entity and they appear as a ratio (Chan & Kumaraswamy, 1999). Similar studies to that proposed have also used these variables (Chan & Chan, 2004; Dursun & Stoy, 2011a; Stoy, Dreier, et al., 2007), and obtained good results by using them. For this reason, two new variables were derived for analysis: (i) total GFA/total number of floors and (ii) total cost of construction/total GFA. All independent variables used in the research are shown in Table III-3.

Variable	Type	Abbreviation	Unit of measure
Total cost of construction	Continuous	*T_Cost*	€
Total gross floor area	Continuous	*T_GFA*	m^2
Total number of floors	Discrete	*T_Floors*	Nominal numbers
Number of floors above ground	Discrete	*A_Floors*	
Number of floors below ground	Discrete	*B_Floors*	
Total gross floor area / total number of floors	Continuous	*T_GFA/T_Floors*	m^2
Total cost of construction / total gross floor area	Continuous	*Standard*	€ / m^2

Table III-3. Predictor variables selected for analysis.

Type	Description	*n*	*Speed* (m^2/month)			
			Max.	Min.	Median	Std. dev.
1	Multi-family housing	57	1,366.67	39.84	435.95	323.39
2	Single family housing	14	26.10	3.24	12.94	6.90
3	Offices and commercial	18	2,009.96	79.10	473.17	630.74
4	Sports and entertainment	15	627.60	30.80	75.27	150.61
5	Health and wellness	15	384.00	6.49	105.87	111.35
6	Cultural	28	926.18	12.73	190.26	197.96
7	Other uses (not industrial)	20	799.53	10.80	72.67	198.10
	All types	**167**	**2,009.96**	**3.24**	**196.78**	**366.76**

Table III-4. Main statistical data of the dependent variable representing construction speed.

Considering that there is no general agreement in the literature about which is the most appropriate dependent variable to predict the construction time, two types of response variables were selected for analysis: construction process duration, referred to as *Time* variable and measured in months, and construction speed, referred to as *Speed* variable and measured in m^2/month. The features of the calibration data for these two dependent variables are included in Tables III-4 and III-5, while Figure III-3 shows the values of construction speed versus the duration of calibration projects.

Type	Description	*n*	*Time* (months)			
			Max.	Min.	Median	Std. dev.
1	Multi-family housing	57	55.00	12.00	20.00	7.99
2	Single family housing	14	37.00	10.00	19.00	7.62
3	Offices and commercial	18	29.00	6.00	16.50	5.63
4	Sports and entertainment	15	53.00	6.00	15.00	12.16
5	Health and wellness	15	49.00	4.00	18.00	11.24
6	Cultural	28	55.00	6.00	15.00	11.54
7	Other uses (not industrial)	20	42.00	10.00	18.00	8.64
	All types	**167**	**55.00**	**4.00**	**18.00**	**9.21**

Table III-5. Main statistical data of the dependent variable representing construction time.

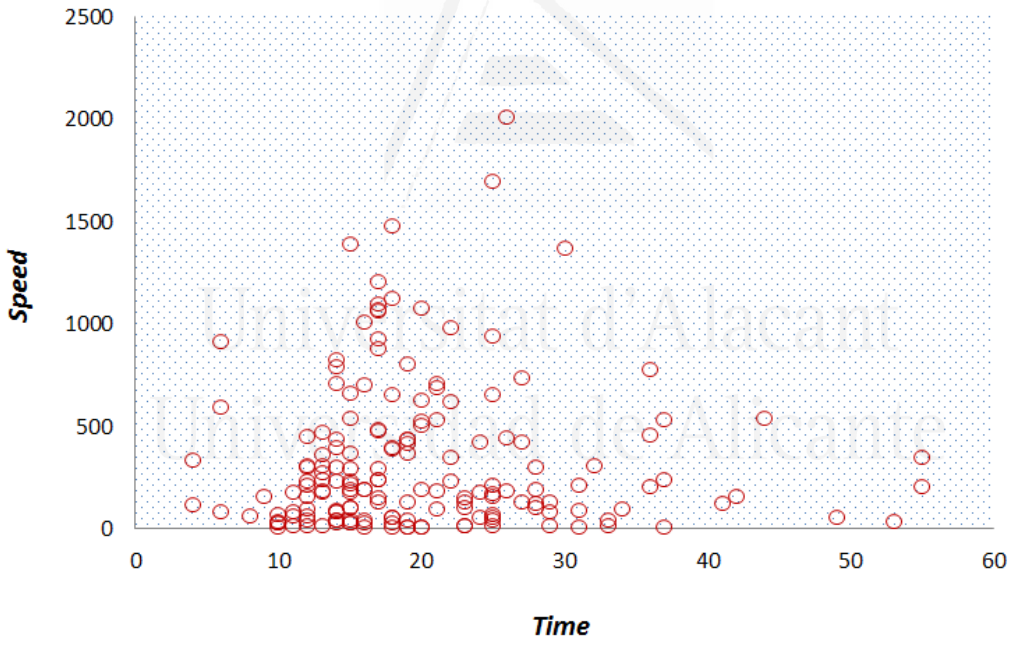


Figure III-3. Construction speed versus construction time of calibration projects (*n*=167).

III.7 Selected Modelling Tools

The construction process of a building is very complex and there are no exact rules to select the factors that are significant enough to explain all the variability of the construction time existing in a sample of building projects. Nevertheless, by using

regression analysis it is possible to identify a small number of variables to explain a significant proportion of this variability (Chan, 1998). Regression analysis has been one of the most widely used statistical tools in many subject areas, because it provides simple methods for investigating functional relationships among different variables (Chatterjee & Hadi, 2006). Such relationships are expressed in the form of an equation connecting the response or dependent variable and one or more predictors or independent variables. Regression techniques work well with both qualitative and quantitative variables (Chan, 1998). However, it is not always easy to derive relationships among the variables under study because, in spite of the name "independent variables", in practice the predictor variables are rarely independent of each other (Chatterjee & Hadi, 2006).

In order to provide proper models to estimate the construction time of building projects, to date, most authors have developed parametric models by using LRA. Although these parametric models have shown a good balance between the difficulty of developing estimation models and the forecasting accuracy obtained using said models, they represent a great simplification of the complex relationships controlling the duration of the construction process in building projects (Sousa et al., 2014). Moreover, their accuracy and trustworthiness is basically limited by some assumptions inherent to LRA (Jafarzadeh et al., 2014). Considering that nonlinear behaviours are common in complex real systems like the construction phase of a building project, it is reasonable to think that the project scope factors that influence the time variable might not be fully associated in a linear manner. In such a case, the nonlinearity of the parameters forces us to use numerical computation techniques in order to find the solution (Bates & Watts, 1988). These techniques are iterative methods that start with an initial value of the parameters and obtain a new value closer to the optimum value at each step. In this connection, the learning process of artificial neural networks (ANNs) is equivalent to minimising a global error function, which is a multivariate function that depends on the synaptic weights in the network (Møller, 1993). With this perspective, the supervised training of ANNs can be seen as a problem of numerical optimisation (Haykin, 1999). In recent years the growing tendency has been to use ANNs to generate predictive models related to building projects (e.g., Chen & Huang, 2006; Le-Hoai et al., 2013).

There are other modelling tools in the literature that allow the generation of nonlinear models from experimental data in order to predict new situations. For

example, by using the MODELHSS methodology (Cortés et al., 2000) it will be possible to obtain nonlinear models expressed in a specific formal language, which are generated from the definition of a set of vocabularies from different orders. The equations obtained using the MODELHSS methodology are characterised as a linear combination of functions defined from the vocabularies, so that their statistical treatment is reduced to linear regression analysis cases since they are linear in the parameters. The use of higher order vocabularies and the generation of families of linear mathematical equations in the parameters, which fit the observed data to a greater or lesser degree, present a problem of spatial and temporal complexity that leads to consider pruning criteria in the generation of equations that have an exponential growth. Another example of methodology to create nonlinear models is known as Polimodel (Verdú & Villacampa, 2008). This methodology allows for the generation of families of nonlinear models in the parameters that can be automatically created. From the point of view of numerical methods, we can cite the methodologies proposed by Pérez-Carrió et al. (2009) and Navarro-González & Villacampa (2012, 2013), which determine bi-dimensional and *n*-dimensional numerical models respectively.

Notwithstanding the above, it has been suggested in the literature that, generally, linear regression models could give us a good approximation to explain relationships that are not truly linear and provide a starting point from which to explore more sophisticated models (Chan, 1998). This assumption has been supported by the results of construction-based studies where ANNs and LRA were compared (see, e.g., Le-Hoai et al., 2013). Taking this into consideration, first of all, different linear regression models were generated in this thesis by using MLRA. Then, the sets of independent variables with the best predictive performance were used to generate nonlinear forecasting models to improve the results obtained by the best linear models. To this end, ANNs and the FEM-based numerical methodology developed by Navarro-González & Villacampa (2012, 2013) were used to create new predictive models.

The rationale for the choice of the three modelling techniques used in this thesis is as follows: firstly, most statistical models presented to date for estimating construction time have been developed by using LRA and, consequently, this is the modelling tool that must work as a benchmark; second, according to the discussion carried out in Chapter II, it is clear that MLRA and ANNs have become two competing empirical

model-building methods and, thus, in order to develop nonlinear models with better predictive performance, it is mandatory the use of ANNs; finally, a novel numerical modelling technique based on FEM has been introduced into the study since it has not been applied previously in order to generate models to estimate the construction speed of building projects. By doing so, we want to test if it is possible to improve the predictive performance which has been obtained to date with the traditional modelling techniques represented by MLRA and ANNs. Each of the three selected modelling techniques is analysed in detail below.

III.7.1 Multiple Linear Regression Analysis

MLRA is a generalisation of simple linear regression and has been the statistical procedure most widely used in the literature to obtain predictive models for estimating the construction time of building projects (see, e.g., Chan & Chan, 2004; Love et al., 2005; Stoy, Dreier, et al., 2007). This study uses MLRA as a research tool to analyse the effects of some project scope factors on construction time and develop forecast models to estimate the most likely time to complete the construction of a building. In MLRA the technique of the least squares is the one most often used to obtain the parameter values involved in regression equations. Due to the linearity of the relationship between variables, the application of this technique leads to solving a linear system in the parameters. A general model of MLRA is given by the following expression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \cdots + \beta_k x_k + \varepsilon \tag{27}$$

where y is the dependent variable, β_i determines the contribution of each independent variable x_i and ε is the random error.

Linear regression estimates the coefficients of the independent variables that best predict the value of the dependent variable. These coefficients are chosen in order that the sum of squared differences between observed and predicted values is minimal so as to minimise the residual variance. The random error can be considered to be positive or negative and, for any adjustment made to the values of independent variables, such an error has a normal probability distribution with mean equal to 0 and variance equal to σ^2. According to Chatterjee & Hadi (2006), two questions need to be answered when formulating a regression model:

- Which variables should be included?
- In what form should they be included?

The procedures available for the selection of the best predictor variables can be classified into two broad categories (Chatterjee & Hadi, 2006): (1) the forward selection (FS) procedure and (2) the backward elimination (BE) technique. There is also a modification of the FS procedure known as the stepwise method. Regarding the second question, data is not always in a form suitable for analysis. Sometimes variables have to be transformed before the regression analysis can be developed in order to meet certain assumptions such as linearity or normality. In parametric tests the distribution function of the response variable plays a vital role (Dursun & Stoy, 2011a). The non-normality of the dependent variable invalidates the standard tests of significance since they are based on the normality assumption (Chatterjee & Hadi, 2006). In this regard, the logarithm form of the dependent and independent variables is the most commonly used transformation which allows the possibility of fitting nonlinear models by using linear regression (Goh, 1998). Nevertheless, as mentioned in Chapter I, a regression model is considered linear when the parameters of the model occur linearly even if the predictor variables happen nonlinearly (Chatterjee & Hadi, 2006).

The computer software SPSS version 20 for Windows was used in this thesis as the statistical tool to carry out the proposed linear regression analysis. From the previously mentioned selection methods the one that offered better results was the BE method. This method starts by including all independent variables in the model and then proceeds to eliminate them one by one according to the established exit criteria. The variable with the smallest partial correlation with the dependent variable is considered first for elimination. If it meets the criteria for removal, then this variable is eliminated. After the first variable is removed, the variable remaining in the equation with the smallest partial correlation is considered next. The process of elimination stops when there are no variables in the model that meet the removal criteria. The BE technique is preferable over the other techniques because it has the advantage of inspecting all variables in the early stages of the model development process (Attalla & Hegazy, 2003). This advantage enables the BE technique to extract more significant variables than other techniques (Lowe et al., 2006).

In order to identify the most appropriate response variable for carrying out MLRA, first of all, this thesis used the Pearson's correlation coefficient to assess the bivariate

correlations between the independent variables under study and the two options of dependent variables (*Time* and *Speed*). The Pearson's correlation coefficient (*r*) measures the linear relationship between two quantitative random variables. Unlike the covariance, this coefficient is independent of the measuring range of the variables. It ranges between the values -1 and 1, where 1 represents a total positive correlation, 0 means no correlation and −1 represents a total negative correlation. Therefore, the magnitude of the relationship is specified by the numerical value of the coefficient, whereas the sign represents the direction of the relationship. In theory, the most effective independent variables to develop a prediction model through LRA are those having a high correlation with the dependent variable.

III.7.2 Artificial Neural Networks

As previously explained in Chapter II, although there are many types of ANN architectures, MLP and RBF networks have been the two architectures most commonly used in the literature. Consequently, this thesis compared the performance of these different types of ANNs when they are used to develop predictive models related to the construction speed of new builds. The methodologies listed below were developed for each of the proposed ANN architectures in order to obtain the best possible prediction models.

III.7.2.1 MLP Networks

III.7.2.1.1 Input Variables

When developing ANN models the task of selecting input variables has a significant impact on model performance (Günaydın & Doğan, 2004; Shahin et al., 2008). This task is especially difficult in the event of a large number of variables if there is no previous knowledge to make a decision about the input variables that must be chosen (Maier & Dandy, 2000). Nowadays, there is no specific systematic way of selecting input variables (Zhang et al., 1998), but several techniques have been developed in the literature to help with the selection process. Zhang et al. (1998) indicated that the selection of input variables should be included in the model construction process (model-based approach). Nevertheless, a potential shortcoming of this approach is that the determination of whether a input variable is significant or not is dependent on the error of a trained model, which is not only a function of the input variables but also of the model structure and the quality of the training process.

Therefore, the network structure should be optimised for each candidate set of input variables (Maier et al., 2010).

In order to overcome the limitations of model-based approaches, a method that is often used is to select the input variables based on a priori knowledge (Maier & Dandy, 2000; Zhang et al., 1998). This kind of approach, known as model-free approach, can use both linear dependence measures, such as correlation, or nonlinear measures of dependence, such as mutual information, to obtain the significant input variables before developing ANN models (Shahin et al., 2008). The Akaike information criterion (AIC) is frequently used for identification of nonlinear models, but nevertheless there is disagreement regarding its use (Zhang et al., 1998).

Taking into consideration that it has been suggested in the literature that linear regression models could provide a good starting point from which to explore more sophisticated models, in this thesis the sets of independent variables which were used to obtain the best linear models were also used for the purpose of generating nonlinear forecasting models to be able to improve the results obtained with linear models. Moreover, in accordance with that stated in this section, the network structure of each candidate set of input variables was optimised independently.

III.7.2.1.2 Hidden Layers, Activation Functions and Data Pre-processing

MLP networks have usually been considered the most appropriate approach for the construction sector (Le-Hoai et al., 2013). Although this kind of feedforward networks can have more than one hidden layer, it has been shown that, generally, one hidden layer is enough to approximate any complex nonlinear function provided that sufficient connection weights are used (Cybenko, 1989). That is one of the main reasons why most authors use only one hidden layer for forecasting purposes (Zhang et al., 1998). Therefore, bearing in mind this consideration, and taking into account that similar studies related to the construction time of building projects have also obtained good results by using a single hidden layer (see, e.g., Bhokha & Ogunlana, 1999; Le-Hoai et al., 2013), this thesis made use of a fixed three-layered MLP structure to create ANN models.

On the other hand, in order to harness the power of neural networks to learn complex or nonlinear relationships among different variables, the use of nonlinear activation functions is recommended at least in the neurons of the hidden layer. These

activation functions introduce a degree of nonlinearity that is very useful for most ANN applications (Zhang et al., 1998). The computation of the local gradient for each neuron of a MLP network requires the knowledge of the derivative of the associated activation function. For this purpose, the transfer function has to be continuous (Haykin, 1999). Examples of continuously differentiate nonlinear transfer functions usually associated with the use of MLPs are the sigmoidal type functions. The two most popular forms of sigmoidal functions are the logistic function, determined by Equation 28, and the hyperbolic tangent function, represented by Equation 29 (Haykin, 1999; Maier & Dandy, 2000).

$$f(x) = [1 + exp(x)]^{-1} \tag{28}$$

$$f(x) = \big(exp(x) - exp(-x)\big)/\big(exp(x) + exp(-x)\big) \tag{29}$$

$$f(x) = x \tag{30}$$

It is important to remember that when using sigmoid transfer functions in the output layer, it is fundamental the scaling of the output values to the limits of such activation functions (between –1 to 1 for the hyperbolic tangent transfer function and 0 to 1 for the sigmoid transfer function) (Shahin et al., 2008). Although there is no consensus in the literature on what activation function should be used in the case of the output layer, it has been suggested that it might be more reasonable to use the linear activation function (Equation 30) when using continuous output variables (Zhang et al., 1998). In such a case, scaling is not obligatory but it is still recommended (Maier & Dandy, 2000). In any case, it is not clear the effect of the use of different activation functions on the performance of ANN prediction models (Zhang et al., 1998). Generally, the same transfer function is used in all layers (Maier & Dandy, 2000; Zhang et al., 1998).

The scaling process of output values is usually independent of the scaling process of input values, but when scaling the output values it is necessary to rescale them in order to interpret the results obtained from the network. Moreover, the model performance should also be evaluated based on the rescaled outputs. However, few studies clearly state whether the performance measures are calculated on the original or transformed scale (Zhang et al., 1998).

Another point that needs to be discussed is the distribution of input data. In most traditional statistical models, the data used to develop prediction models has to be normally distributed so that they can be properly generated. Conversely, it has been suggested in the literature that by using ANNs the probability distribution of input data does not have to be known (Maier et al., 2010). Nevertheless, some authors have indicated that transforming the input data into some known forms (e.g., logarithmic transformation) may be useful to improve the network performance (Shahin et al., 2008).

III.7.2.1.3 Modelling Process of MLP Networks

The configuration of an ANN has a huge impact on its performance, but nevertheless the process of designing an optimal neural network is highly problem-dependent (Hegazy et al., 1994; Moselhi et al., 1991). Moreover, there is no defined methodology to set an optimal configuration and obtain the best generalisation for a given problem in a specific domain. In order to find the best possible solution, the modeller needs to carry out a trial and error process and identify good combinations of network architecture, training algorithm and stopping criteria to avoid overfitting (Piotrowski & Napiorkowski, 2011; Smith & Mason, 1997). The concept of overfitting refers to fitting specific examples too much, which means loosing ability to generalise. On the contrary, overtraining addresses the issue of using too many iterations in the learning process, which leads to overfitting (Amari et al., 1997).

The optimal network structure can be defined as the smallest network that properly captures existing relationships in training data (Maier et al., 2010). However, there is no guide to help us to determine the appropriate configuration of MLP networks and it is not always obvious how to build the network appropriately, since the conditions under which good generalisation can be obtained are not well understood (Hegazy et al., 1994). In fact, the design process of MLPs has been considered more of an art than a science (Hegazy et al., 1994; Xiang et al. , 2005; Zhang et al., 1998).

For the purpose of generating a valid ANN prediction model, first of all, the modeller must decide on the type of network that will be used. The network architecture affects the optimal training and its ability to generalise, and the general criterion in designing neural networks is to start from the simplest structure that can provide the essential consistency and adequacy (Nawari et al., 1999). It should be noted that while the selection of appropriate model geometry is necessary for most

ANN architectures, it is superfluous for others which have a fixed structure (Maier et al., 2010).

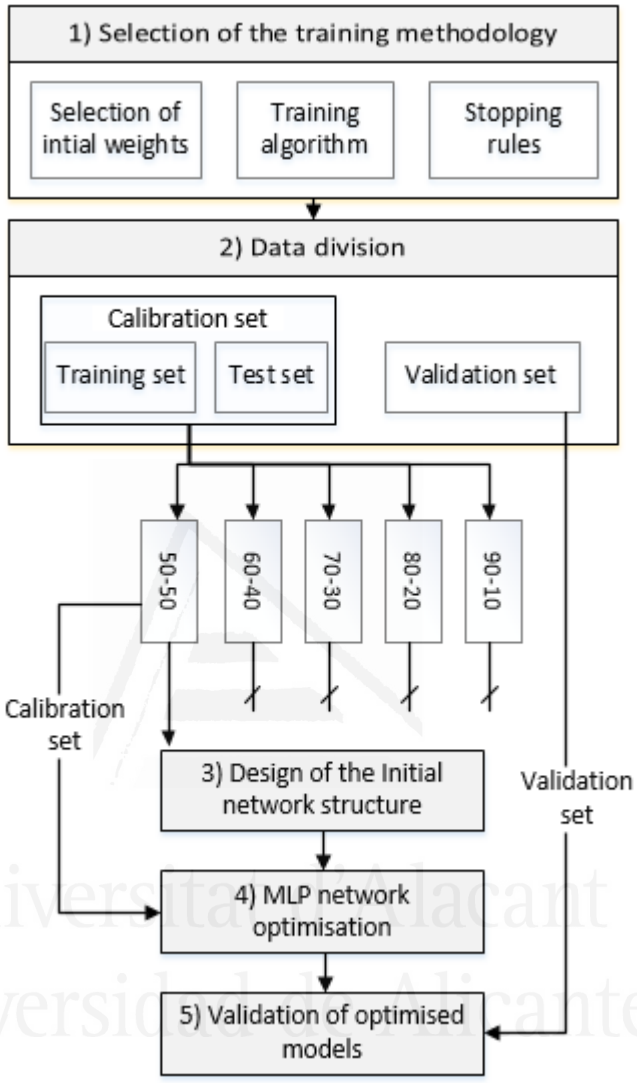


Figure III-4. Modelling process proposed for MLP neural networks.

In order to improve network performance, it is fundamental to use a systematic approach and statistical principles in the development of ANN models (Maier & Dandy, 2000). Systematic approaches, such as pruning and constructive algorithms, have also been generated to obtain automatically optimal network structures. The adaptive method of architecture determination (Ghaboussi & Sidarta, 1998) and Cascade-Correlation (Fahlman & Lebiere, 1990) are a few examples of automatic methods developed to obtain optimal network structures of ANNs. Nevertheless, these methods are usually quite complex and difficult to implement. Furthermore, none of these

methods can guarantee the optimal solution for all real forecasting problems. It has been shown that exhaustive search over the space of network architectures is not possible even for networks of small size and this is the main reason why researchers and practitioners use heuristic strategies that drastically reduce search complexity (Moody, 1994). In this situation, the modeller must try different network configurations to find an optimal setup.

Once it became clear what kind of network architecture and input variables would be used in this thesis, the modelling process of MLP networks was divided in five main steps (Figure III-4): (1) selection of the training methodology, (2) data division, (3) design of the initial network structure, (4) MLP network optimisation, and (5) validation of optimised models. The first four steps are discussed in more detail below, whereas the validation process is encompassed under the overall framework proposed in Section III.8 to evaluate the predictive performance of models.

Selection of the Training Methodology

After the network geometry has been set, it is necessary to train the designed network using a set of examples (training set) representing the problem under study. This training process will enable the network to "learn" what the relationships are between the selected variables.

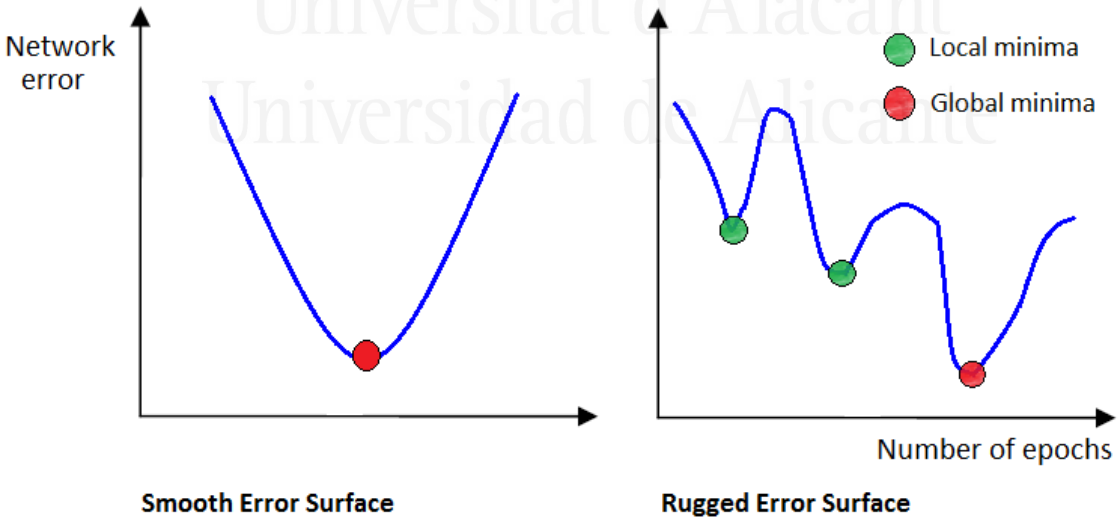


Figure III-5. Error surface with different degrees of ruggedness (Maier et al., 2010).

Determining the combination of weights that minimises the error is a difficult problem, since each combination generally results in a different error surface. The

degree of difficulty in finding the best combination of parameter values that minimises the error is affected by the "ruggedness" of the error surface, on the understanding that "ruggedness" is a measure of the number, spacing and steepness of the craters and valleys in the error surface (Maier et al., 2010). If the error surface is smooth then there are fewer local minima. Conversely, if the error surface is more rugged there are more local minima and the global minima is harder to find (see Figure III-5). The degree of ruggedness of an error surface is usually problem dependent and is affected by the number of model parameters (Maier et al., 2010). Generally, the higher the number of model parameters, the greater the number of local minima. In addition, a larger number of parameters makes it harder to interpret the model and increases the risk of obtaining models with anomalous behaviours. Therefore, according to the above, it is crucial to find the model with the smallest number of parameters that is able to represent with enough accuracy the underlying relationships between the selected variables (Maier et al., 2010).

In the case of MLP networks, among the various training methods available BP is the most widely used due to its easy implementation and good ability to generalise (Hegazy et al., 1994). By using BP we must cope with three essential points required to develop the training process, which are related to (i) the selection of the initial weights, (ii) the training algorithm used to optimise the value of the synaptic weights, and (iii) the stopping rules of the training process. Each point is explained in detail below.

Selection of Initial Weights. As has previously been explained, one of the dangers likely to arise when using methods based on the GD algorithm is that the learning process might converge to a local minimum rather than a global minimum. The existence of different local minima in the error function makes it much harder to train the network and, apart from the learning algorithm, another key question to consider before training MLP networks is where to start the learning process. At the start of the training process it is necessary to assign an initial value to the synaptic weights and one of the main reasons for the long duration of the training process is the lack of proper initialisation methods (Duch et al., 1997). Therefore, it is of great importance to find the right weight initialisation for achieving both an optimum training and very fast learning (convergence speed) (Drago & Ridella, 1992). This has led to analyse the best possible weight initialisation strategies and their effect on the convergence speed by some researchers (Rojas, 1996).

A random set of weights is usually used to initialise the BP algorithm, but since it has been designed to converge on local solutions, when a random starting point is used the possibility of finding the correct combination of weights is based primarily on chance (Duch, 1999). To improve the possibility of finding a global optimum, the training process may be repeated many times, starting from different initial points. Such a multi-start approach is very popular, but sometimes criticised for not being efficient (Piotrowski & Napiorkowski, 2011). A well-known initialisation heuristic for a feedforward network with sigmoidal units is to select its weights with uniform probability from an interval $[-\alpha, \alpha]$. However, very small values of α paralyses the learning process, whereas very large values can lead to saturation of the nodes (Rojas, 1996). Drago & Ridella (1992) used the statistically controlled activation weight initialisation (SCAWI) to find the optimal initial weights and prevent neurons from reaching saturation during an early stage of the training. They determined the maximum magnitude of the weights through statistical analysis. Nguyen & Widrow (1990) speed up the training process by setting the initial weights of the hidden layer so that each hidden node was assigned to approximate a portion of the range of the desired function at the start of training.

On the other hand, by using global optimisation (GO) strategies it is possible to avoid the problem of occasional convergence to local minima in MLP training (see, e.g., Plagianakos et al., 2001; Sexton et al., 1999). However, GO methods also have drawbacks, as they are computationally expensive. In order to overcome these problems, hybrid methods employing both GD and GO algorithms have been developed. The aim of this combination of techniques is to maintain quick convergence using a computationally cheap algorithm, while reducing the likelihood of convergence to poor local minima (Treadgold & Gedeon, 1998). In this line of work, the simulated annealing (SA) appears as a metaheuristic search algorithm for global optimisation problems, in order to find a good approximation to the global optimum of a given function in a large search space. Annealing refers to the process that occurs when physical substances, such as metals, are raised to high energy levels and then gradually cooled until some solid state is reached (Sexton et al., 1999). This notion of slow cooling is implemented in the SA algorithm as a slow decrease in the probability of accepting worse solutions as it explores the solution space. The method was independently described by Kirkpatrick et al. (1983) and by Černý (1985).

The SPSS software used in this research takes a random sample from the training set and applies the alternated simulated annealing and training procedure to derive the set of initial weights. This procedure uses simulated annealing and training alternately up until K_1 times. Simulated annealing is used to break out of the local minimum that training finds by perturbing the local minimum K_2 times. If break out is successful, simulated annealing sets a better initial weight for the next training with the final objective of finding the global minimum by repeating this procedure K_3 times. The procedure is rather expensive for large data sets, so it is only used on a random sample, which means that there will always be a certain proportion of randomness in the development of the network training process.

Training Algorithm. The network training is an unconstrained nonlinear optimisation problem in which the synaptic weights are iteratively modified to minimise the error between estimated and actual values (Zhang et al., 1998). The selection of a particular optimisation algorithm depends on the type of network architecture, available data and the problem being analysed (Boussabaine, 1996). In the case of MLP networks, the GD algorithm with momentum has been the most popularly used training method (Zhang et al., 1998). Nevertheless, the learning process of neural networks is an NP-complete problem (Rojas, 1996) and although the SBP algorithm is very useful for MLP networks, it has three main drawbacks (Castillo et al., 2006):

- **Slow learning speed.** When the error surface is nonlinear, and only the information of the local gradient is available, the learning rate should be kept small enough to ensure a stable convergence, which in turn will increase considerably the duration of the learning process.
- **Convergence to local minima.** It is vital to select the best possible set of initial weights when it comes to developing good prediction models, because when the training process begins the learning algorithm opts for an address and it does not explore other possibilities. If the initial choice is unsuitable, training might finish in a local minimum.
- **Oscillations.** When the minimum of the error function lies in a narrow “valley”, following the gradient direction can lead to wide oscillations of the search process (Rojas, 1996).

In addition, the success of the SBP algorithm depends on learning parameters which have no theoretical basis for selecting them and must be specified by the user (Møller, 1993). In light of these weaknesses, several variations or modifications have been proposed in the literature, such as adaptive methods or second order methods.

In the SBP method, the learning rate, which determines the magnitude of the changes in the synaptic weights for each iteration, is fixed at the beginning of the learning process (Castillo et al., 2006). On the contrary, in adaptive algorithms the size of change is increased whenever the algorithm proceeds down the error function over several iterations and it is decreased when the algorithm jumps over a valley of the error function. Generally, the existing adaptive algorithms differ according to the type of information used to modify the size of change. In this connection, Silva & Almeida (1990) proposed a method which works with different learning rates for each synaptic weight. Jacobs (1988) developed another similar algorithm, named delta-bar-delta, where the acceleration of the learning rates is made with more caution than the deceleration. A variant of these adaptive algorithms, named RPROP, was proposed by Riedmiller & Braun (1993). RPROP performs a local adaptation of the updates of network weights according to the behaviour of the error function. The main difference to other adaptive techniques is that the effect of the adaptation process is not confused by the unpredictable influence of the size of the derivative, but only dependent on the temporal behaviour of its sign.

On the other hand, second order methods have also been proposed in order to increase the convergence speed of MLP networks. Second order means that these methods make use of the second derivatives of the error function, while first-order techniques, like the SBP algorithm, only use the first derivatives (Haykin, 1999). Second-order methods have been considered the most efficient nonlinear methods (Castillo et al., 2006; Zhang et al., 1998) and are among the fastest learning algorithms, although they have a higher computational cost. The most relevant examples of second order methods are the conjugate gradient (CG) algorithms, the quasi-Newton methods and the Levenberg-Marquardt algorithm (Hagan & Menhaj, 1994). Like the SBP, CG algorithms iteratively try to get closer to the minimum, but while SBP always proceeds down the gradient of the error function, a conjugate gradient method will proceed in a direction which is

conjugate to the directions of the previous steps. Consequently, the minimisation performed in one step is not partially undone by the next, as it is the case with SBP. The use of a momentum term to avoid oscillations can be considered as an approximated form of conjugate gradient (Battiti, 1992). Examples of CG algorithms are the Fletcher-Reeves (Fletcher & Reeves, 1964), Polak-Ribiere, Powell-Beale (Powell, 1977), or the SCG algorithm (Møller, 1993). Regarding the quasi-Newton methods, since in most cases it is difficult and expensive to compute the exact Hessian matrix, they use second-order information about the error surface but without requiring knowledge of the Hessian Matrix. Quasi-Newton methods are not as sensitive to accuracy in the line search stage as the CG methods, but CG methods are preferable to quasi-Newton methods in computational terms (Haykin, 1999). The two most common quasi-Newton procedures are the Davidson-Fletcher-Powell (DFP) method and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. Finally, similarly to quasi-Newton methods, the Levenberg-Marquardt algorithm was designed to obtain second-order information about the error surface without having to compute the Hessian matrix. It can be considered a trust-region modification to Gauss-Newton (Hagan & Menhaj, 1994).

In this research, the SCG algorithm (Møller, 1993) has been selected as an alternative training algorithm to develop MLP models. This algorithm is a second-order method that does not contain any user-dependent parameters whose values are crucial for the success of the training process. By using a step size scaling mechanism, SCG avoids a time consuming search per learning iteration, which makes the algorithm faster than other second-order algorithms such as the classic CG algorithm with line search or the BFGS quasi-Newton algorithm (Møller, 1993).

Stopping Rules. BP algorithms cannot be shown to converge, and there are no well-defined criteria for stopping them (Haykin, 1999). In addition, the best trained MLP network is not one whose error is continuously decreasing. One problem of the MLP network is the danger of obtaining a model over-trained, in which case the network memorises the training values. A model that has been over-trained will fit too much the information that has been used in the training process, but will be unable to provide good estimates with new cases which were not used during the learning phase of the network. Therefore, it is necessary to establish certain mechanisms to control the performance offered by the neural network in order to

detect when it is necessary to continue training or when to stop the process. After covering the entire training data set and modifying the synaptic weights, certain stopping rules must be checked. If none of the stopping rules is satisfied, the training process starts again. Several approaches can be used to determine when to stop training:

- After the development of a fixed number of training cycles.
- When the training error reaches a sufficiently small value.
- When no changes in the training process occur for a given number of epochs.
- When the rate of change in the average squared error is small enough.

The effects of these stopping rules are important and, theoretically, the stopping criteria determine when the network has been optimally trained (Maier and Dandy 2000). Nevertheless, the selection of a misguided criterion may lead to stopping the training process prematurely (underfitting, high model bias) or produce overtraining (overfitting, high model variance) (Haykin, 1999; Shahin et al., 2008). This is usually referred to as the bias/variance dilemma (Geman et al., 1992) and the model selection techniques attempt to find the optimal trade-off between bias (undefitting) and variance (overfitting) (Moody, 1994).

The selection of the method that must be used to avoid overfitting will depend on the number of data, number of tests the user is ready to perform and the optimisation algorithm that will be used (Piotrowski & Napiorkowski, 2013). Although there are several methods used to try to prevent overfitting, this research work used the early stopping method, which is the only one that is widely applied in practice (Piotrowski and Napiorkowski, 2013) because it is simple to understand and implement (Prechelt, 1998). It uses a cross-validation approach through a third data set, referred in this study to as test set. The cross-validation method is a popular strategy for algorithm selection and its assessment process. It consists of controlled or uncontrolled data division of a sample, once or several times, where part of data is used for training an algorithm and the remaining data is used for estimating its prediction risk (Stone, 1974). The popularity of cross-validation mostly comes from the universality of the data splitting heuristics (Arlot & Celisse, 2010). A single data division, known as hold-out, produces an estimate of the risk, and averaging over several data divisions produces a cross-validation estimate

(Arlot & Celisse, 2010). The former type of cross-validation is the one used by the early stopping method.

According to the early stopping method, when there is overtraining the prediction error on the test set starts to rise although the error in the training set is still falling. So, the minimum point of error on the test set will be used as the criterion for stopping the training process (Haykin, 1999). Figure III-6 shows the scheme of the early stopping method. The fact that both the training and test errors cannot be simultaneously driven to zero means that the network does not have the capacity to model the function exactly (Haykin, 1999).

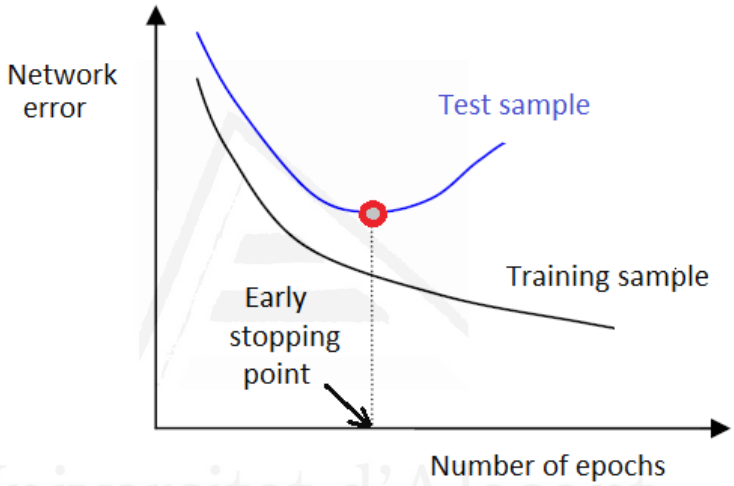


Figure III-6. Early stopping method.

There has been a lot of discussion related to the effectiveness of the use of cross-validation as the stopping criterion (Maier & Dandy, 2000). The problem with the application of this method is that real generalisation curves almost always have more than one local minimum (Prechelt, 1998), so that, even though the error obtained with the test data set might increase at some particular point during the training process, it cannot be ensured that this error will not be lower at a later point if the training process was continued (Maier & Dandy, 2000). Therefore, immediate stopping when the error for test data starts to increase for the first time during optimisation process usually results in underfitting of the neural network (Piotrowski & Napiorkowski, 2013). Consequently, when the early stopping method is used, it is vital to continue training for some time after the error in the test set

first starts to rise (Maier & Dandy, 2000). In this thesis, 100 epochs were selected as maximum number of steps allowed before checking for a decrease in error in the testing sample. If there was no decrease in error after the specified number of steps, then training stopped. In addition, 1,000 cycles of training (epochs) were considered to be enough to reach a good local minimum on the error surface, so that when this maximum number of epochs was exceeded the training stopped.

Data Division

According to what has been set forth in previous sections, for the development of ANN models it is necessary to have, at least, two data sets: a calibration set, which in this thesis is composed of 167 individual cases, and an independent validation set, which in this research work is a fixed group of 18 individual cases. Therefore, this fixed ratio is 90% for calibration data and 10% for validation data. Nevertheless, it is also necessary to incorporate a third data group due both to the use of the early stopping method to avoid the overfitting problem and to the use of the optimisation process proposed in this thesis for MLP networks, which is a process where the network geometry is optimised by trial and error (Maier et al., 2010). Nonetheless, it is common in the literature to use one test set for both validation and testing purposes, particularly with small data sets (Zhang et al., 1998).

Based on the above, the calibration set had to be divided into two new groups of projects: one for learning (training set) and the other one for stopping the training process (test set). With this perspective in mind, it is possible to see the problem of developing neural networks with good ability to generalise from two different views (Haykin, 1999): (1) the size of the training set is fixed and we have to determine the best network structure, and (2) the network structure is fixed, in which case we have to find out what size of training set is the most adequate one.

In most ANN applications, the data sets have been constituted on an arbitrary basis (Shahin et al., 2008) by using a random selection of the data, which introduces the problem of conclusion instability (López-Martín, 2015), i.e., different studies might make different conclusions about the same studied variables (Kocaguneli & Menzies, 2013). Furthermore, it has been suggested that if the model performance does not satisfy the expectations the size of the training set can be modified (Zhang & Fuh, 1998). Most authors make data division based on ratios 90-10, 80-20, etc., and, according to Haykin (1999), the relationship 80-20 between training and test cases

appears to be a sensible choice. In this regard, Shahin et al. (2004) analysed the impact of different ratios of data division on the performance of ANN models related to shallow foundations, and stated that there was no clear relationship between the proportion of data division and the model performance.

Nevertheless, it has also been argued in the literature that the statistical properties of obtained data subsets should be considered as part of any data division process, because when the different subsets represent the same statistical population there is an improvement in model performance (Shahin et al., 2004). On one hand, although the proposed modelling process uses the early stopping method to avoid the overfitting problem, even if one reaches the global minimum, overfitting or underfitting might sometimes go undetected if the test set is not perfectly representative of the problem under study (Prechelt, 1998). In that case, the training process might be stopped at a suboptimal time and, consequently, a suboptimal network geometry or wrong training parameters might be selected (Shahin et al., 2004). Moreover, If the test set is too small, an accurate estimate of the prediction risk cannot be obtained (Moody, 1994). On the other hand, the number of training samples must be sufficiently large to guarantee good generalisation (Boussabaine, 1996), but the use of the cross validation will result in a significant waste of data available for training (Liu et al., 2008). Even in cases where the data set is large, it is necessary to use as much data as possible for training purposes, since the estimation of the error associated with the model variance becomes worse as the training set size is reduced (Günaydın & Doğan, 2004). Therefore, it seems to be that the cross-validation becomes practical only if both data subsets (training and test) are large enough (Moody, 1994).

However, with respect to the sample size, there is no definite rule to know what minimum size is needed for an adequate training process but, in either case, the sample size is constrained by the data availability (Zhang et al., 1998). In this regard, if the samples are limited the number of parameters to be estimated must be minimised (Bhokha & Ogunlana, 1999).

According to Tokar & Johnson (1999), the effect of the size of the training set on the accuracy of ANN models is not as large as the effect of the content of training data. Furthermore, the distribution of training data must be sufficiently dense over the range of expected values to produce an accurate interpolation (Boussabaine, 1996). Thus, apart from the size of samples, when making a data division other factors such

as particular characteristics of the research problem and data type should be taken into consideration (Zhang et al., 1998).

In short, the problem with the use of cross validation and the early stopping method is that both training and test data sets need to be properly divided to assure that they are representative enough of the problem under study and can provide an unbiased estimate of the real generalisation error (Liu et al., 2008).

Data division	Variables	Training data set			Test data set			MW *U* test (*p*-value)
		n	**Mean**	**Standard deviation**	***n***	**Mean**	**Standard deviation**	
50-50	*GFA*	84	6,479.01	8,767.97	83	6,838.22	7,796.90	0.468
	Floors		4.56	3.21		4.84	2.88	0.329
	Standard		728.87	327.70		675.99	295.90	0.325
	Speed		319.11	387.25		342.51	346.74	0.377
60-40	*GFA*	100	7,160.88	9,627.68	67	5,906.29	5,682.83	0.843
	Floors		4.98	3.34		4.28	2.50	0.325
	Standard		735.51	341.35		653.46	258.31	0.137
	Speed		339.57	402.34		317.56	308.56	0.786
70-30	*GFA*	117	6,863.27	8,420.00	50	6,176.14	7,993.10	0.850
	Floors		4.62	2.88		4.88	3.42	0.945
	Standard		699.04	340.44		710.90	237.34	0.272
	Speed		339.76	376.43		309.63	345.82	0.953
80-20	*GFA*	133	6,580.41	8,113.62	34	6,959.28	9,007.66	0.904
	Floors		4.70	3.08		4.71	2.96	0.927
	Standard		699.09	299.24		716.27	364.38	0.902
	Speed		329.57	366.85		335.32	371.88	0.886
90-10	*GFA*	149	6,744.41	8,461.48	18	5,938.45	6,705.52	0.749
	Floors		4.75	3.04		4.28	3.16	0.417
	Standard		707.38	315.08		662.98	295.57	0.478
	Speed		332.88	371.36		313.01	335.32	0.792

Table III-6. Statistical properties of training and test data sets for each proposed data division.

In this thesis, for the purpose of analysing the effect of the calibration data division on the performance of MLP models, five different random divisions were made according to the following ratios between the percentage of training and test data sets: 50-50, 60-40, 70-30, 80-20, and 90-10 (see Figure III-4). The proposed random divisions were made keeping the same percentages of individual cases for each type

of facility (*Type 2*, *Type 3* and *Others*), because it has been recommended to take into account data structure when making the splits (Arlot & Celisse, 2010). In addition, similarly to Shahin et al. (2004), a trial and error process was developed to achieve statistically consistent data subsets. Yet, at the same time, it is also important to pay attention to the domain of applicability of the generated models because usually ANNs cannot be expected to extrapolate successfully beyond the range of values covered by the data set used for training (Hawkins, 2004; Minns & Hall, 1996). Consequently, training data should cover all spectrums of data available (Boussabaine, 1996; Günaydın & Doğan, 2004) and, to this end, the extreme cases (i.e., maximums and minimums) were forced to be included in the training set and not in the test data set.

For each data division, the statistics properties of the training and test data sets are shown in Table III-6. The similarity between the two data sets was statistically tested by using the nonparametric MW *U* test and, as a result, the null hypothesis that these data sets have the same characteristics cannot be rejected for all variables ($p>0.05$) in any of the developed data division. This shows that the training and test data subsets are representative of each other. Furthermore, in the case of the 80-20 data division the value of *p* is particularly high for all variables.

Design of the Initial Network Structure

The importance of choosing a proper network structure depends on the type of architecture (Maier et al., 2010). As has been explained previously in this section, a fixed three-layered MLP structure was selected for modelling the construction speed of new builds (the output node). However, in the case of MLP networks it is also necessary to define the number of neurons in the hidden layer, the activation functions to be used in both hidden and output nodes and the kind of scaling for both input and output data. The methodology proposed in this study for selecting the best possible MLP network structure takes advantage of the knowledge obtained in previous research in order to define proper initial network geometry from which to start the optimisation process.

Regarding the number of hidden neurons, Hegazy et al. (1994) identified three different heuristics used in the literature in order to achieve an optimal number of hidden neurons by using a single hidden layer and the BP algorithm. According to these heuristics, the number of neurons used in the hidden layer should be $0.75m$, m, or $2m+1$, where m is the number of input neurons. These heuristics were later used by

Kim et al. (2004) in the field of construction cost estimation, and the best network performance was obtained by using the latter heuristic rule. Moreover, it has been suggested that it is better to select more rather than fewer hidden neurons (Jafarzadeh et al., 2014). Bearing the above in mind, the proposed optimisation methodology makes use of the equation $2m+1$ to set the number of hidden neurons in the initial network structure.

The set of selected input variables cover different ranges of values, so in order to ensure that all variables receive equal attention during the training process, they should be scaled (Maier & Dandy, 2000). In the field of ANNs the training data has generally been normalised for achieving an effective learning, and this transformation has been recognised as a good way of improving neural network performance (Günaydın & Doğan, 2004). Consequently, initially, the values of all continuous input variables were normalised. On the other hand, according to the good results obtained in a similar study undertaken by Bhokha and Ogunlana (1999), the initial MLP network structure made use of logistic functions in both the hidden layer and the output layer. When using sigmoidal-type functions in the output neurons, such as the logistic functions, the scaling of output data is compulsory because data values must be consistent with the limits of the activation functions utilised in the output layer (Shahin et al., 2008).

MLP Network Optimisation

Designing an optimal MLP network structure is highly problem-dependent (Hegazy et al., 1994; Maier & Dandy, 2000; Zhang et al., 1998) and is not a straight forward process (Bhokha & Ogunlana, 1999), since there are no structured methods to identify what network structure can best approximate the underlying function which defines the relationships between the variables under study. A network is designed for a specific set of inputs as well as outputs, whose number is not limited (Kim et al., 2004). As the design of an optimal network structure is a complex and dynamic process, the time-consuming procedures of trial and error are often used (Zhang et al., 1998). In this sense, according to the data type and the response required by the problem under study, the neural network model is empirically rather than theoretically designed (Günaydın & Doğan, 2004).

The network complexity determines the generalisation properties of the model, since a network which is either too simple or too complex will have poor

generalisation, although complex models are much more prone to overfitting (Piotrowski & Napiorkowski, 2013). The complexity of a neural network model is governed by the number of degrees of freedom, which, in turn, is controlled by the number of adaptive parameters (weights and biases) in the network. In most cases, the use of the minimum structure results in lower computation costs, least requirements on implementation resources, and best generalisation (Xiang et al., 2005).

Once the initial network geometry has been selected, it is necessary to optimise it by using the calibration data set and repeating the training process a certain number of times with different configurations of network structures. The proposed optimisation process involves three stages in the case of the GD algorithm and only two steps for the SCG training algorithm (see Figure III-7). In the case of the GD algorithm it is necessary to develop an extra stage (the second one) due to the need to set the best values of learning rate and momentum coefficient. These parameters determine the speed and stability of the network training respectively (Günaydın & Doğan, 2004). In the proposed optimisation process, initially, a small learning rate of 0.1 was arbitrarily selected along with a momentum coefficient of 0.5.

In each of the process stages it is necessary to assess the performance of every trained configuration and select the most appropriate options. Not everything neural networks learn is useful because real-world data is noisy, distorted, and often incomplete, so that the fact that one configuration ultimately reaches a lower training error than another is not a basis for selecting it (Hammerstrom, 1993). Thus, in the proposed optimisation methodology, the root of the mean square error (RMSE) obtained in the test data set was chosen as the selection criterion. RMSE is the most popular measure of error and has the advantage that large errors receive much greater attention than small errors (Shahin et al., 2008). In addition, this measure is consistent with the error function to be minimised by the BP method during the training process (Jafarzadeh et al., 2014). RMSE is defined by Equation 31, where Av_i is the actual value, Pv_i is the predicted value and n is the number of samples.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(Av_i - Pv_i)^2}{n}} \tag{31}$$

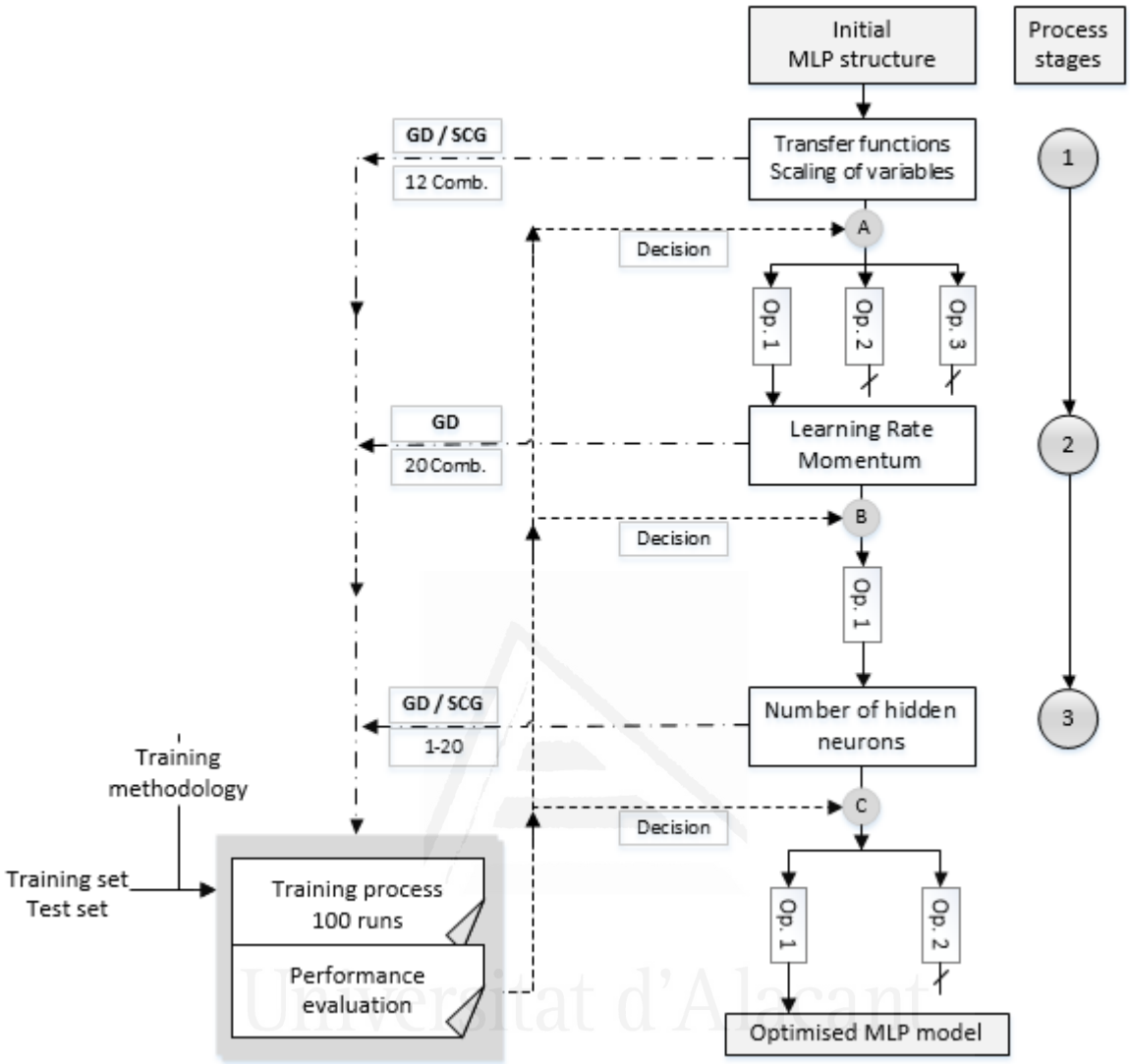


Figure III-7. Optimisation process proposed for MLP networks.

As has already been justified in this section, when using MLP networks with the BP algorithm the starting values chosen for the synaptic weights can be crucial for network performance. Thus, it is sometimes difficult to know if the accuracy results obtained by the trained network are the result of the selected geometry or, conversely, there has been greater influence of the chosen set of initial weights. In this regard, Faraway and Chatfield (1998) suggested that it is advisable to try several different sets of random starting values to see whether consistent results are obtained. Consequently, at each step of the proposed optimisation methodology the training process is repeated one hundred times (i.e., 100 different sets of initial weights) for each tested MLP structure. Moreover, the weights which define the MLP models

depend on the order of examples used in the training process and the initialisation value for the random number generator, because by using the SPSS software the initialisation method of the synaptic weights involves subsampling the training data set. Therefore, in order to carry out a fair comparison, for a given MLP network structure each sequence of runs is executed with the same initialisation value for the random number generator, the same data order, and the same variable order. This way it was possible to obtain the same random sequence of initial synaptic weights for each tested MLP network structure.

Each stage of the optimisation process is described below and as a result of this process, for each data division, six MLP network structures are selected for validation.

***Transfer Functions* (TF) and *Scaling of Variables* (SV).** The first stage in the proposed selection process is to identify the best combination of transfer functions, as it is not clear the effect of different activation functions on the performance of neural networks (Zhang et al., 1998). Consequently, different combinations of transfer functions were tested in this study by using hyperbolic tangent and logistic functions in the hidden layer, and linear, hyperbolic tangent, and logistic functions in the output layer. In addition, due to the strong connection existing between transfer functions and the scaling of data, each possible combination of activation functions was, in turn, tested together with different options for scaling the input variables and the output variable. The SPSS software allows the use of three different types of data scaling: standardised, normalised (values between 0 and 1) and adjusted normalised (values between -1 and 1). All the combinations involve the scaling of variables because when this was not done poor results were obtained.

Furthermore, although the use of the linear transfer function in the output layer provided good RMSE values, it was removed from the experimentation because in the case of projects with low construction speed (*Type 2* projects) some negative values were predicted systematically. Finally, 12 different combinations of transfer functions and scaling of data were used for the purpose of identifying the option that would produce the best MLP network performance (Table III-7). The proposed methodology first identifies the three combinations which offer the best RMSE value in the test data set, and then the network structure for each of the three

selected combinations is optimised independently in the subsequent steps of the optimisation process.

Comb.	Scaling of input variables	Transfer functions		Scaling of the output variable
		Hidden layer	Output layer	
1	Normalised	Tanh	Logistic	Normalised
2	Normalised	Logistic	Logistic	Normalised
3	Normalised	Tanh	Tanh	Adj. normalised
4	Normalised	Logistic	Tanh	Adj. normalised
5	Adj. normalised	Tanh	Logistic	Normalised
6	Adj. normalised	Logistic	Logistic	Normalised
7	Adj. normalised	Tanh	Tanh	Adj. normalised
8	Adj. normalised	Logistic	Tanh	Adj. normalised
9	Standardised	Tanh	Logistic	Normalised
10	Standardised	Logistic	Logistic	Normalised
11	Standardised	Tanh	Tanh	Normalised
12	Standardised	Logistic	Tanh	Normalised

Adj. normalised = Adjusted normalised // Tanh = Hyperbolic tangent

Table III-7. Combinations of transfer functions and scaling of variables (TF-SV).

***Learning Rate* (LR) and *Momentum* (M).** The SCG algorithm does not contain any user-dependent parameters whose values are crucial for the success of the training process (Møller, 1993). On the contrary, in the case of the GD algorithm the training parameters of learning rate and momentum can influence the network performance. These parameters can take on any value between 0 and 1 and it is actually impossible to do an exhaustive search to find the best combination of both parameters. Hence, only some selected combinations of values have been considered by researchers (Zhang et al., 1998). In this line of work, the second stage of the proposed optimisation process uses different combinations of learning rate values {0.01, 0.1, 0.5, 0.7, 0.9} and momentum coefficient {0.3, 0.5, 0.7, 0.9} for each of the three selected network structures. In total, 20 combinations are tested and then the combination with the best performance in the test data set is selected.

***Number of Hidden Nodes* (HN).** The third stage of the optimisation process addresses the issue of determining the optimal number of hidden nodes. There is

no direct and precise way of determining the best number of nodes in each hidden layer. This is why the most common way in determining the number of hidden nodes has been via trial-and-error (Zhang et al., 1998), which is one of the drawbacks of ANNs as it is necessary to spend a lot of time in this task (Kim et al., 2004). For example, Nawari et al. (1999) started with a small number of neurons and then they increased the number until no significant improvement in the neural network performance was achieved.

In general, networks with fewer hidden nodes are preferable as they usually have better generalisation ability and less overfitting problem (Shahin et al., 2008). Moreover, using too many neurons will increase the training time (Nawari et al., 1999). Conversely, networks with too few hidden nodes may not have enough complexity to model the underlying function (Zhang et al., 1998) and often increases the likelihood that the learning algorithm becoming trapped in a local minimum (Shahin et al., 2008). Consequently, it is imperative to use the network with the minimum number of hidden neurons but which achieves a satisfactory performance (Nawari et al., 1999). However, justifying the use of a more complex network structure involves showing that the additional complexity is necessary and that the same model performance cannot be obtained by using a simpler network structure (Hawkins, 2004). Therefore, in this phase the number of neurons is varied from 1 to 20 and, then, the two network structures that obtain the best performance in the test data set are selected. This process is repeated for each of the three networks structures which have previously been chosen.

III.7.2.2 RBF Networks

In this research work, RBF networks were also built in order to explore an alternative approach to the MLP architecture. Unlike the MLP trained with the BP algorithm, the design of RBF networks follows a principled approach (Haykin, 1999). The structure of RBF networks is similar to MLP networks, with the difference that they only have one hidden layer with radial basis functions and use on a mandatory basis linear activation functions in the output layer. In addition, a normalised RBF (NRBF) was also tested in the hidden layer, where the activation of all the RBF hidden neurons (J) is normalised to sum to one. In NRBF networks, the basis function $\varphi_j(X)$ takes the following form:

$$\varphi_j(X) = \frac{exp\left(-\sum_{i=1}^{P} \frac{1}{2\sigma_{ji}^2}\left(X_i - \mu_{ji}\right)^2\right)}{\sum_{j=1}^{J} exp\left(-\sum_{i=1}^{P} \frac{1}{2\sigma_{ji}^2}\left(X_i - \mu_{ji}\right)^2\right)} \quad (32)$$

where X is the input vector, P is the number of inputs, σ_j is the width of $\varphi_j(X)$ and μ_j is the centre of $\varphi_j(X)$.

As with MLP networks, both types of activation functions (RBFs and NBRFs) were tested in the hidden layer together with different options for scaling the input variables and the output variable. Three different types of data scaling were used: standardised, normalised and adjusted normalised. Moreover, in the case of RBF networks, contrary to what happened with MLP networks, not all proposed combinations involved the scaling of variables, as good results were also obtained without scaling them. Finally, 32 different combinations of transfer functions and forms of data were used for the purpose of identifying the option which produces the best RBF network (Table III-8).

Similarly to what was done with MLP networks, for each of the 32 used combinations the number of neurons was varied from 1 to 20. In total, 640 different configurations were tested. The training process of RBF networks is developed by the SPSS software in two different phases:

1. Determination of the basis functions by clustering methods. In this phase the centre and width of each basis function is calculated.
2. Calculation of the synaptic weights of neurons. For the calculated basis functions, the ordinary least-squares regression estimates of the weights are computed. The goal is to minimise the differences between network outputs and the desired outputs. The learning process is guided by minimisation of a computed error function in the network output.

The simplicity of these calculations allows the RBF network to be trained faster than MLP networks. Nevertheless, for the achievement of the same degree of accuracy, the approximation of a nonlinear function by using a RBF network may require a larger number of parameters than using a MLP network (Haykin, 1999).

Comb.	Scaling of input variables	Transfer functions	Scaling of the output variable
		Hidden layer	
1	Standardised	RBF	Standardised
2	Standardised		Normalised
3	Standardised		Adj. normalised
4	Standardised		None
5	Normalised		Standardised
6	Normalised		Normalised
7	Normalised		Adj. normalised
8	Normalised		None
9	Adj. normalised		Standardised
10	Adj. normalised		Normalised
11	Adj. normalised		Adj. normalised
12	Adj. normalised		None
13	None		Standardised
14	None		Normalised
15	None		Adj. normalised
16	None		None
17	Standardised	NRBF	Standardised
18	Standardised		Normalised
19	Standardised		Adj. normalised
20	Standardised		None
21	Normalised		Standardised
22	Normalised		Normalised
23	Normalised		Adj. normalised
24	Normalised		None
25	Adj. normalised		Standardised
26	Adj. normalised		Normalised
27	Adj. normalised		Adj. normalised
28	Adj. normalised		None
29	None		Standardised
30	None		Normalised
31	None		Adj. normalised
32	None		None

Table III-8. Combinations of activation functions and scaling of variables used with RBF networks.

Unlike MLP networks, when using the SPSS software the RBF networks did not suffer from any instability problem related to the initialisation value for the synaptic

weights and, consequently, it was not necessary to repeat the training process several times. Therefore, by keeping the same order for the cases and variables used in the training process, the generated RBF models were equal regardless of the number of different executions carried out.

Finally, it must be observed that the SPSS software allows the use of cross-validation to stop the training process of RBF networks. Hence, following the same criterion established for MLP networks and taking into consideration that the usage of a dynamic early stop to the training process can drastically improve the ability of RBF networks to generalise (Lin et al., 2009), all proposed configurations were tested using the early stopping method. The lowest RMSE value obtained in the test data set was chosen as the criterion for selecting the best model.

III.7.3 FEM-Based Numerical Methodology

In order to overcome the limitations of modelling inherent to the technique of estimation by least squares, and as an alternative procedure for modelling the construction speed to the traditional approaches related to MLRA and ANNs, this research work used a new numerical methodology developed by Navarro-González & Villacampa (2012, 2013) to build representation models based on FEM. The general idea of this method is the division of a continuum in a set of small elements interconnected by a series of points called nodes. FEM can be considered as an approximation numerical method and there is a great deal of publications which develop its application (see, e.g., Lewis & Ward, 1991).

The proposed numerical methodology is an evolution of the methodology developed by Pérez-Carrió et al. (2009), where predictive models can be obtained by using two-dimensional finite elements, and allows us to generate representation models of a function from the interpolation defined in an *n*-dimensional finite element model. The starting point is the existence of a defined relation and the knowledge of its initial conditions. The interpolation of the function involves the application of certain conditions, which in the proposed methodology entails the coincidence of the function in a finite number of points called nodes. These values are determined in ways that minimise an error function previously defined and data dependent. Therefore, it is necessary to solve an optimisation problem that depends on the initial conditions of the problem.

The computational implementation of the proposed methodology enables obtaining families of mathematical models that represent the relationship defined by $z = u(x_1, x_2, \dots\dots, x_n)$ when imposing the initial conditions represented by $\{z_i, x_i^1, x_i^2, \dots\dots, x_i^n\}_{i=1,2,\dots\dots,p}$. This way, it is ensured that the obtained representation models fit the experimental data well enough to not drag systematic errors in the observations and infer new situations reliably, others than those covered by experimental data. Schematically, the generation of a representation model through the proposed numerical methodology requires the development of four steps:

(i) Definition of the geometric model.
(ii) Definition of the model function.
(iii) Definition of the optimisation problem.
(iv) Resolution of the optimisation problem.

This section first presents an explanation on the general functioning of FEM and, then, the four main steps used by the proposed numerical methodology are explained in detail.

III.7.3.1 Finite Element Method

The functioning of certain systems can be analysed by dividing them into different elements and studying their behaviour in isolation. Then, the global system is rebuilt again to study it from the analysis of each of its elements. When it is possible to determine a model using a finite number of elements the problem is called "discreet" whereas if the division is done indefinitely the problem is referred to as "continuous". Various methods have been proposed in the literature to discretise continuous systems. Particularly in the field of engineering, analogies have been made between actual finite discrete elements and finite portions of a continuous domain, which has resulted in the concept of "finite element" (Zienkiewicz & Taylor, 1994).

FEM has been used to solve problems of both scientific analysis and engineering. In particular, it is worth noting its application for solving problems of structural mechanics and differential equations which have no exact solution. Generically, FEM can be considered as a numerical approximation method of continuing problems, so that (Frías, 2004):

- A continuous domain is divided into a finite number of pieces, each of which is referred to as "finite element" and whose behaviour is specified by a finite number of parameters associated with certain characteristic points called nodes. These nodes are points of attachment of each adjacent element.
- The solution of the whole system is determined by assembling of all the elements by following the same rules as the discrete problems.
- The unknowns of the problem are no longer mathematical functions and they become the value of these functions.
- The behaviour within each element is defined by the behaviour of the nodes through appropriate interpolation functions (shape functions).

Therefore, FEM is based on the process of transformation of a continuous domain into an approximate discrete model. This transformation is referred to as the discretisation process. The interpolation of the known values at nodes allows us to know what happens inside the model. Thus, FEM performs an approximation of the values of a function from the knowledge obtained with a finite number of points (Frías, 2004).

A formalisation of the concept of a finite element model is to consider the representation of a function in a region from a finite element model. Generically, it is considered a smooth function $u(x)$ defined on $\overline{\Omega}$, which is the closure of a bounded open $\overline{\Omega} \subset R^n$. Defining a finite element model $\overline{\Omega}$ of $u(x)$ is equivalent to generate an interpolation function.

In general terms, a finite element model that is defined in $\overline{\Omega}$ determines a region $\widetilde{\Omega}$ which is the union of a finite number of bounded subregions $\overline{\Omega}_i$ (finite elements), so that $\widetilde{\Omega}$ matches or is as close as possible to $\overline{\Omega}$. In $\widetilde{\Omega}$ some points called global nodes are identified, while local nodes are identified within each finite element $\overline{\Omega}_i$. Furthermore, in order to join the elements in the model, there must be correspondence between adjacent elements, which results in the fulfilment of certain compatibility conditions.

In the case of the numerical methodology proposed by Navarro-González & Villacampa (2012, 2013), $\widetilde{\Omega}$ is known as the geometric model of finite elements.

III.7.3.2 Definition of the Geometric Model

Initially, it is necessary to define the kind of elements involved in the geometric model and the nodes associated with each element, as well as local and global numbering of both. The elements defined in dimension *n* are *n*-dimensional hypercubes (Figure III-8).

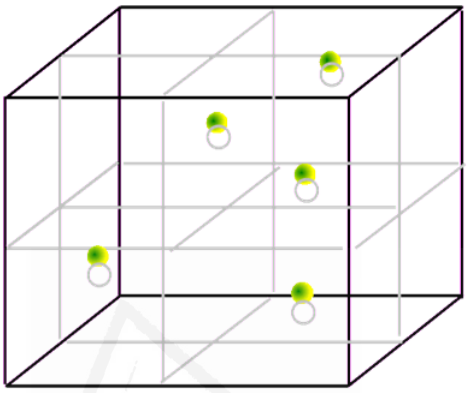


Figure III-8. Geometric model of dimension *n*=3.

Generically, the construction of one of these hypercubes is carried out as the elevation on the additional dimension of a hypercube with dimension *n*-1. As an example, Figure III-9 shows the process to pass from 1 to 2 dimensions and from 2 to 3 dimensions. In this process, the discretisation of the domain representing the set of nodes takes place.

Given the domain $\widetilde{\Omega} = [a_1, b_1] \times ... \times [a_n, b_n]$, the geometric model of this domain is defined in a generic way by the interval [0,1], so that $\widetilde{\Omega} = [0,1] \times ... \times [0,1]$. A finite element model which is defined in $\widetilde{\Omega}$ corresponds with the division of each interval into the same number of subintervals *C*, which defines the complexity of the geometric model. As an example, in the case of two dimensions, the domain $\widetilde{\Omega} = [0,1] \times [0,1]$ could be divided into four elements $\{\overline{\Omega}_1, \overline{\Omega}_2, \overline{\Omega}_3, \overline{\Omega}_4\}$ such that $\widetilde{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2 \cup \overline{\Omega}_3 \cup \overline{\Omega}_4$ and the complexity of this geometric model would be *C*=4. This particular bi-dimensional case has been represented in Figure III-10. Nevertheless, a local coordinate system is always used in order to better manage the calculations made in the elements. So, if we are in two dimensions, when working with any finite element

$\bar{\Omega}_i$ its global coordinates (*x*,*y*) are transformed into local coordinates (*s*,*t*), where *s* and *t* belong to the interval [-1,1].

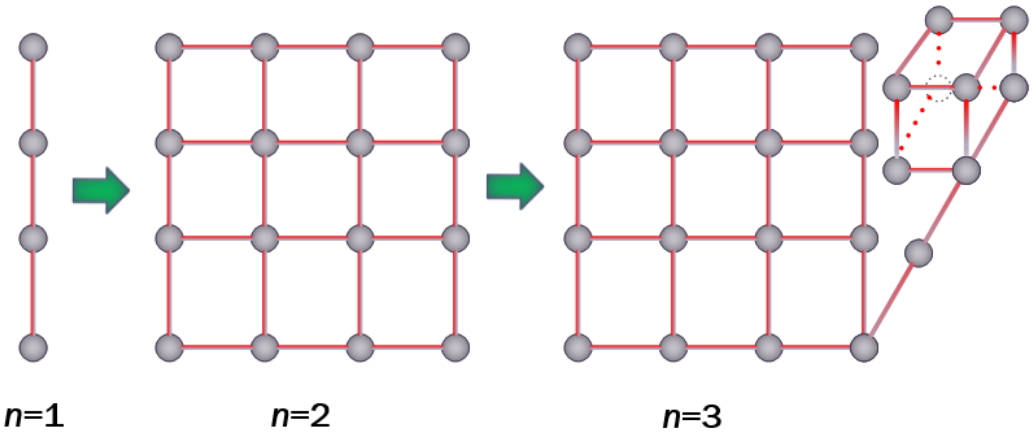


Figure III-9. Domain discretisation representing the set of nodes (Navarro-González & Villacampa, 2012).

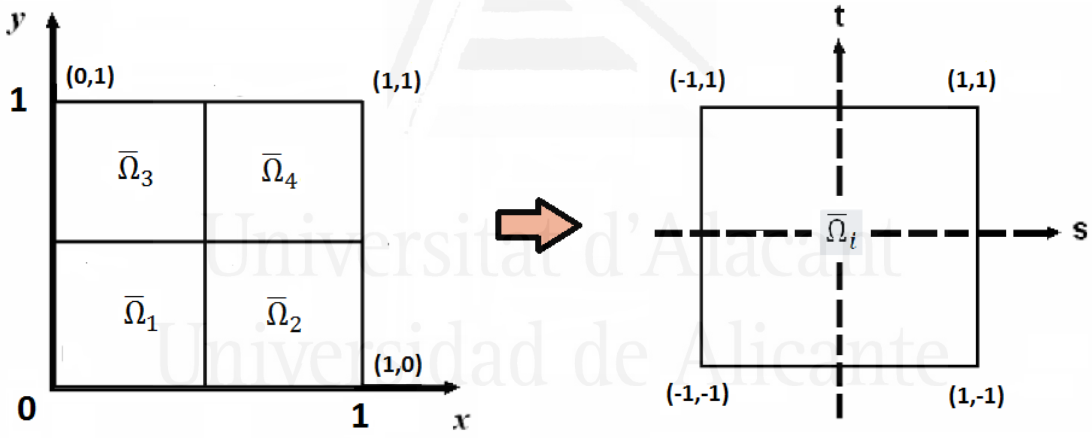


Figure III-10. Transformation of global coordinates into local coordinates.

According to what has been explained in the previous paragraph, each segment of the generic hypercube is locally transformed in the interval [-1,1], so that it can be identified with the domain $[-1,1]^n$. In each one of these elements 2^n nodes are defined. The complexity of the geometric model is the number of intervals in which each segment [0,1] is divided. Therefore, the original domain $[0, 1]^n$ is divided in a set of C^n hypercubics subdomains (Navarro-González & Villacampa, 2013). The elements have a global index between 0 and C^n –1 and, similarly, a global index between 0 and

$(C+1)^n$ -1 is defined for nodes. The process of local numbering of nodes for a generic element can be observed in Figure III-11.

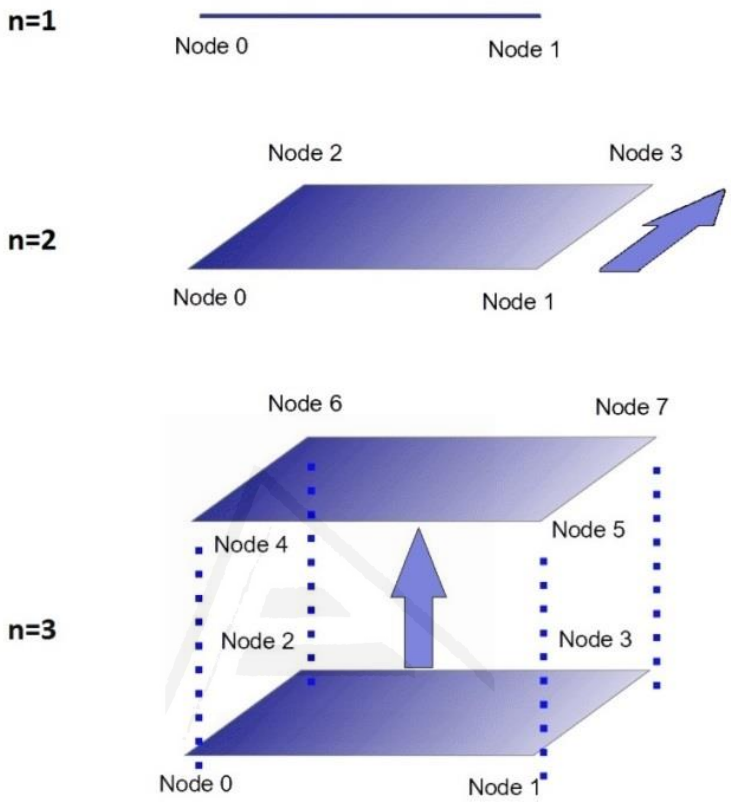


Figure III-11. Process of local numbering of nodes for a generic element.

III.7.3.3 Definition of the Model Function

Once the geometric model has been defined in a hypercube, we proceed to define its function by FEM. The model function, *EF*, is defined for each geometric model considered in the *n*-dimensional hypercube. If in the geometric model a set of nodes $\{n_i\}$ are identified, the function *EF* is defined as an interpolation from the values of these nodes. The interpolation of the function is also defined through a union of local interpolations, as usual in FEM, in which it has been necessary the definition of transformations that improve the efficiency of the interpolation in *n*-dimensional elements.

Given a generic point, $x = (x_1, x_2,\ldots\ldots,x_n)$, and the e_i element that contains it, its local interpolation is the model function in the point, which is defined as:

$$EF(x) = \sum_{n_i=1}^{NN} u_{n_i} N_i(x) \tag{33}$$

where *NN* is the total number of nodes, $N_i(x)$ is the basis function associated to the n_i node, and u_{n_i} is the value of the function in the n_i node.

Particularly, at points $\{P_j\}_{j=1,\dots,NP}$, where the conditions of the problem are defined, and taking into account that the *NP* index is the total number of points, the function is represented by the expression:

$$EF(P) = \sum_{n_i=1}^{NN} u_{n_i} N_i(P) \tag{34}$$

III.7.3.4 Definition of the Optimisation Problem

Considering the points of the initial conditions $\{P_j\}_{j=1,\dots,NP}$, and the values of the observation in these points, $\{Z_j\}_{j=1,\dots,NP}$, the difference between these experimental values and the values obtained from the model function in these points defines an error function as follow:

$$Error = \sum_{j=1}^{NP} \left(\sum_{n_i=1}^{NN} u_{n_i} N_i(P_j) - Z_j \right)^2 \tag{35}$$

The optimisation problem is obtained by imposing the minimisation of this error function regarding the values on u_{n_i}nodes, represented by the following equation:

$$\sum_{n_i=1}^{NN} \sum_{j=1}^{NP} N_s(P_j) N_i(P_j)\, u_{n_i} = \sum_{j=1}^{NP} Z_j N_s(P_j) \tag{36}$$

where the value of *s* ranges from 1 to *NN* forming a system of linear equations.

III.7.3.5 Resolution of the Optimisation Problem

The resolution of the optimisation problem is performed by solving a system of equations. The obtained linear system depends on the distribution of experimental points on the generated geometrical model. When all the existing elements contain an adequate distribution of points, the system is well determined. But the system is usually compatible and undetermined (infinite solutions), and in these cases the following methodology is applied to solve it (Navarro-González & Villacampa, 2012, 2013):

1. Calculating the rank of the matrix based on rank revealing QR (RRQR) factorisation method, with selection of variables.
2. Application of a rigidifying method for determining a solution at the nodes where it is not possible to obtain a general resolution of the system of equations. This method is inspired on the mechanics in a rigid body, where the properties of a function in a node are related to the values of the function in neighbouring points.
3. Solving the resulting system of equations.

III.8 Performance Evaluation of Prediction Models

III.8.1 Accuracy Measures

A measure of accuracy can be defined in terms of the forecasting error, which is the difference between the actual value and the predicted value. Although there are many accuracy measures, each one of them has advantages and drawbacks so that there is no accuracy measure universally accepted by academics and practitioners (Zhang et al., 1998). In this connection, one of the conclusions of the research conducted by Foss et al. (2003) suggests that it is probably futile to search for a universal error measure which can be easily applied to compare a linear regression model with a nonlinear function approximation.

RMSE is one of the most popular measures of error and has the advantage that large errors receive much more attention than small errors (Hecht-Nielsen, 1990). In contrast with RMSE, the mean absolute error (MAE) eliminates the emphasis given to large errors. Both RMSE and MAE are desirable when the evaluated output data are

smooth or continuous (Shahin et al., 2008). According to Hyndman & Koehler (2006), the accuracy measures which are dependent on the scale, such as RMSE, are useful to compare different methods applied to the same dataset but should not be used to compare models coming from datasets with different scales. On the other hand, they also indicate that measures based on percentage errors, such as the mean absolute percentage error (MAPE), have the advantage of being independent of scale and are often used to compare the forecast performance across different data sets. Nevertheless, it has been shown that MAPE is a biased estimator of central tendency of the residuals because it is an asymmetric measure. It has the disadvantage of penalising harder positive errors than negative errors. In other words, MAPE has a limit in the case of underestimation and no limit in the case of overestimation (Miyazaki et al.,1994).

In other fields, such as software engineering, MAPE is known by the name of the mean magnitude of relative error (MMRE) (Jørgensen, 2007), and its use has been severely criticised. The criticism of MAPE was pointed out by Kitchenham et al. (2001) and subsequently by Foss et al. (2003) and Myrtveit et al. (2005). The findings of Foss et al. (2003) suggest that in many cases, when comparing two models, MAPE will select the worst one and it will tend to prefer the model that underestimates, instead of the model that forecast the correct value. In the same line of thought, Shepperd & MacDonell (2012) made an appeal to researchers involved in software engineering prediction systems not to use MAPE as an accuracy indicator because it is unsafe. Moreover, in this type of projects the measure of accuracy named as prediction at level *r* (PRED(*r*)) (Conte et al., 1985) has also been used in a complementary manner. This measure is defined as the ratio between the number of cases with a magnitude of relative error (MRE) lower than the *r* level and the total number of cases. Generally, the criterion PRED(0.25) ≥ 0.75 has been used to accept the validity of a model in the field of software engineering, although some authors decrease this criterion. However, Foss et al. (2003) indicated that measures of accuracy such as PRED(*r*), which is based on relative error, cannot overcome the drawbacks of using MAPE.

In light of the above mentioned deficiencies, although MAPE has been used in the literature to compare the accuracy of different linear regression models related to building projects, we discarded its use in this thesis. Instead, RMSE has been chosen as a measure of accuracy for evaluating the performance and validity of developed

models. RMSE is defined by Equation 31 and it has been widely used in the field of ANNs.

Finally, it should be pointed out that although it might be possible to use multiple accuracy measures, nevertheless, one method judged to be the best along one dimension does not necessarily have to be the best in terms of another dimension (Zhang et al., 1998). Thus, this option could create more confusion when assessing predictive models, rather than add clarity to the research developed in this thesis.

III.8.2 Predictive Performance Evaluation Process

Different models may be suitable to explain the dependent variable under investigation, so that the final step in the development process of forecasting models is to validate the quality of a set of candidate models and select the best one according to one or several criteria (Haykin, 1999). Thus, the evaluation process can be subdivided into two different phases, although they are often developed by the researcher in an almost parallel manner: (i) validation of the models, by checking if they fulfil certain basic assumptions and minimum requirements that make them suitable to be used, and (ii) selection of the most appropriate model based on a set of established parameters that allow the comparison between different models. In the event of a tie, according to the parsimony principle, the simplest model should be selected. The validation of a model is proof or confirmation that such a model is appropriate and effective for the purpose for which it was created (Chan, 1998). Nevertheless, the assessment of prediction methods is complicated by the limitations imposed by sample sizes (Kocaguneli & Menzies, 2013). These restrictions makes both the model selection process and the estimation of prediction risk more difficult, because a limited data set will result in a more severe bias/variance (underfitting/overfitting) trade-off (Moody, 1994).

This section describes the evaluation process proposed to evaluate the performance of the prediction models developed in this thesis. First of all, it is necessary to remark the difficulty of building a joint vision about what parameters and values are required to evaluate the performance of predictive models. The main problem lies in the fact that there is no consensus in the literature regarding the methodology used for validating prediction models. Myrtveit et al. (2005) studied a large number of synthetic data sets, which were generated from distributions found in

a real-world data set, and they found that, as the conditions of their experiments changed, no method was consistently best across every condition. According to Kocaguneli & Menzies (2013), the performance of a modelling methodology depends on: (i) the dataset, (ii) the evaluation method used to assess model accuracy, and (iii) the generation method used to build training and validation data sets.

In the case of linear regression models, similar previous research related to building projects has adopted to date different evaluation procedures and different accuracy measures. Even more important, the sample size and the specific type of collected projects also differ between studies. While some of them have used the same dataset to generate and validate the models, in others the models have been validated through a group of projects not included in the model generation process. For example, Chan & Chan (2004) developed a model to predict the construction time by using a single sample of 56 projects belonging to a very specific type of building. On the contrary, Stoy, Pollalis, et al. (2007) used an initial sample of 216 German building construction projects and 200 projects were used for the generation of a model to predict construction speed, while the remaining 16 projects, drawn randomly from the initial sample, were used to validate the model. Love et al. (2005) also developed prediction models of construction time and construction speed by using a sample of 161 examples with 4 types of building projects (including refurbishment and renovation projects) and they did not use any set of validation projects.

Paliwal & Kumar (2009) developed a review of a large number of articles comparing MLP networks and standard statistical techniques used for prediction and classification problems. They indicated that some of the reviewed studies did not validate the generated prediction models on a new data set that was not used for building the models, so that in these cases it was not possible to obtain an unbiased estimate of the generalisation error. In the case of ANN models, the determining factor when selecting the best model has usually been their ability to generalise rather than the accuracy obtained in the data set used to calibrate it (Jafarzadeh et al., 2014). In this regard, results measured only with training data cannot show the model reliability and ANNs are useful only if they also produce an appropriate performance with data that have not been used for training purposes (Hammerstrom, 1993).

Based what has been stated in the paragraphs above, in this thesis the process used to evaluate the predictive performance was developed by using a general framework incorporating five different assessment criteria:

(i) Verification of **compliance with the underlying assumptions** regarding the statistical procedure used to obtain the models.
(ii) Checking the **goodness of fit** of the models to the data set used for generating them.
(iii) Validation of models in terms of **ability to generalise** by using a new sample of projects, independent of the calibration data set.
(iv) Evaluating **the balance between ability to generalise** (validation set) **and the accuracy obtained with calibration data**.
(v) Developing a sensitivity analysis to verify **model stability**.

Each of the proposed evaluation criteria is discussed in detail below.

III.8.2.1 Fulfilment of Basic Assumptions

Most of the statistical techniques are based on certain basic assumptions whose validity is vital for them. However, it is noteworthy that Paliwal & Kumar (2009) observed that only less than one-third of the papers reviewed by them checked these assumptions. The validity of many of the inferences associated with LRA depends on residuals. Therefore, different tests should be done to check if any of the basic assumptions about the residuals has been violated. In particular, the assumptions regarding the residuals are (Hair et al., 1998):

- They are mutually independent.
- The variance of the residuals is constant.
- They are normally distributed with mean 0.

In addition, it must also be verified that there are no problems of multicollinearity between the explanatory variables, as well as identifying possible outliers. The criteria used in this research to develop the linear regression diagnostics can be found in different sources (see, e.g., Devore, 2003; Hair et al., 1998) and they have been summarised in Table III-9.

Assumption/problem to be tested	Indicator	Accepted values
Multi-collinearity	Tolerance	≥ 0.01
Residual analysis		
Independence	Durbin-Watson	$du < d < (4-du)$
Constant variance	Scatter plot	
Normality	Cumulative probability plot	
	Histogram	
	Kolmogorov-Smirnov test	Sig. ≥ 0.05
Outliers	Standardised residual	≤ 2.0
	Cook`s distance (D)	≤ 1
	Dffits (F)	$\leq 2/\sqrt{(p/n)}$

p = number of variables n = number of projects

Table III-9. Summary of the criteria used to develop regression diagnostics.

Conversely, when we consider nonlinear models the error distribution ceases to be normal and we cannot guarantee it is known, so all the usual techniques used to validate a linear regression model become invalid from a formal point of view (Verdú, 2004). Nevertheless, to the best of our knowledge, there are no studies in the literature about alternative assumptions to be verified when working with nonlinear models.

III.8.2.2 Goodness of Fit

A metric widely used for measuring the adjustment of a model is the coefficient of determination R^2. This coefficient indicates the amount of variation in the dependent variable that is explained by the set of the independent variables which form the developed model. If there is a perfect relationship between the response variable and the independent variables the R^2 value equals 1. Conversely, if there is no relationship between these variables the R^2 value is 0. In addition, in order to take into account the effect of the sample size and the number of predictor variables used in the regression model, it is necessary to calculate the adjusted coefficient of determination (Hair et al., 1998).

Although this coefficient gives us an explanatory value of the model with respect to the dependent variable, it does not provide an interpretation of its predictive ability

(Jafarzadeh et al., 2014). In some cases the idea that conveys the coefficient R^2 can match the meaning transmitted by accuracy measurements, but not others. Hence, in this thesis the adjusted R^2 coefficient is used to measure the goodness of model fit to the calibration data. Nevertheless, in the case of comparison between ANN models the goodness of fit was assessed by using the coefficient of determination without adjustment (see, e.g., Jafarzadeh et al., 2014; Patel & Jha, 2015; Shahin et al., 2005).

It should also be noted that in the particular case of comparison between ANN models which come from the same set of input variables, the similarity between the goodness of fit obtained with the three data sets (training, test, and validation) was also used as another information parameter to select the best models. The use of this parameter is based on results from the study presented by Shahin et al. (2004).

III.8.2.3 Ability to Generalise

Already in the early 30s, Larson (1931), cited in Arlot & Celisse (2010), noted that the formation of an algorithm and the evaluation of its statistical performance on the same data produce an overly optimistic result. During the process of the model generation, the adopted modelling methodology tries to fit the estimates produced by the model as much as possible to the actual observations of the projects used for its generation. But if we take a sample of validation data, independent of the calibration data, it often happens that the model does not fit so well to the new data.

In order to validate a forecast model, the most important performance measure is the prediction accuracy of the model beyond the training data, which is known as the generalisation ability of the model (Zhang et al., 1998). A prediction model generalises well if good results are obtained by applying the model to new data (validation set), not been used during the calibration process of the model (Haykin, 1999; Prechelt, 1998) and is within the limits set by the calibration data (Shahin et al., 2008). Generalisation is a measure of sufficiency of the training data to cover much of the problem's solution space, regardless of the model performance during the calibration process (Hegazy et al., 1994). Moreover, the notion of generalisation ability can also be defined as the prediction risk (Moody, 1994).

In the case of ANN models, if they fits very well to the training set but poorly to new examples the phenomenon is known as overtraining and it means that the network loses its ability to generalise due to memorisation of the training data (Haykin,

1999). Among other factors, the ability of ANNs to generalise is affected by the optimisation algorithm used, the degree with which the training and validation sets represent the population to be modelled, and the stopping criterion used (Maier et al., 2010). According to Maier et al. (2010), what is really important in the case of ANNs is that validation data should not have been used as part of the calibration process.

In light of the above, the approach in this research has been to use a set of independent projects to validate the models generated with the calibration data set. Similarly to Irfan et al. (2011) a new sample of 18 health building projects, totally independent of the group of projects used to generate the prediction models, was used for validation purposes. This set of validation projects allows the control of the initial euphoria of the researcher when a very good fit to the calibration data is found.

III.8.2.4 Analysis of Balance between Ability to Generalise and Accuracy Obtained with Calibration Data

A major problem when it comes to selecting the best prediction model from a set of models arises from the existence of, at least, two different data sets, which have been used to evaluate them. Despite what has been said in the preceding paragraphs, in the related literature certain characteristics have been suggested in order to measure the suitability of the data used for validation. First of all, the validation set should be large enough to be reliable, because otherwise the uncertainty in the prediction risk will inevitably remain after the cross-validation is applied (Hawkins, 2004). If the validation set is too small when compared with the calibration set it could be difficult to obtain an accurate estimate of the prediction risk because the validation set is not perfectly representative of the problem (Prechelt, 1998). Second, poor predictions can be expected if the validation data contain values outside of the range of those used for calibration (Maier & Dandy, 2000; Tokar & Johnson, 1999). Finally, it is also important that the calibration and validation sets are representative of the same population (Maier & Dandy, 2000).

On the other hand, we cannot forget that modelling methodologies depend on data to calibrate the parameters and, consequently, the model performance will strongly depends on its quality (Günaydın & Doğan, 2004). One model could obtain a better accuracy in the validation set than another, but worse accuracy in the calibration data, or vice versa. The problem is that, generally, it is not possible to know what data set

(calibration or validation) has the examples with better quality and in this kind of situation it could be difficult to make a safe decision for choosing the best model.

The independent validation set used in this research is a fixed group of 18 cases, which represents a reduced size of examples utilised for validation, if compared to the size of the data set used to calibrate the models (167 examples). In addition, although the value range of the validation set was within the limits of the value range of the data used to develop the models (interpolation), after conducting the nonparametric MW *U* test it was checked that there was no similarity between the variables of the calibration and validation data sets. Therefore, the accuracy results obtained using the validation set may not be truly representative of the performance of the trained model, even taken into account the fact that the validation set does not contain extreme data points that were not used in the model calibration phase.

According to the above, the comparison between different predictive models was carried out in this study taking into consideration the ability to generalise obtained by the models with the validation data set, but also considering the accuracy results obtained from the calibration data set. For the purpose of evaluating the trade-off between ability to generalise (validation set) and the accuracy obtained with the calibration data, in this research a performance index (PI) (Boussabaine, 2001b) was used (Equation 37).

$$\mathrm{PI} = \frac{\mathrm{RMSE.\,calibration} + \mathrm{RMSE.\,validation}}{2} \tag{37}$$

Although the calibration sample is larger than the validation sample, both data sets are treated equally in Equation 37 because the ability to generalise is more important when evaluating the performance of a model (Boussabaine, 2001b).

In this line of thought, as has been performed in previous research related to the construction sector and the use of ANNs (see, e.g., Jafarzadeh et al., 2014; Shahin et al., 2004), the performance analysis of some predictive models, particularly ANN models, was also evaluated by comparing the percentage reduction of RMSE values obtained with the different data sets under study (test and validation projects when comparing ANN models using the early stopping method).

III.8.2.5 Stability Analysis

An important subject that studies the theory of systems is the sensitivity of the models. Since the seminal study presented by Philips (1959), nonlinear instabilities have worried numerical analysts. The model performance is based on the data set used to build it and the generic study of sensitivity tries to know to what extent the behaviour of the model is altered by changing some data (Villacampa et al., 1998). Based on the definition given by Mesarovic & Takahara (1975), a system is stable if small changes in the initial conditions correspond with small changes in the system behaviour. On the other hand, Shahin et al. (2005) concluded that good performance of ANN models on both calibration and validation data does not guarantee that the developed models will obtain good results in a robust fashion on new similar data sets. For this reason, they proposed a method to test the robustness of the predictive ability of ANN models based on the development of a univariate sensitivity analysis and posterior analysis of how well model predictions are in agreement with the known underlying physical processes of the problem under study.

The effect of the changes undertaken by a sensibility analysis will depend on the type of change made in values, so that different sensitivities can be obtained by using different types of modifications. These different modifications can be obtained by using a statistical design of experiments or by making random samples of possible configurations. Therefore, to develop a sensitivity analysis it is first necessary to define the specific type of disturbances that will be used. In general, there are two kinds of analysis:

- **Univariate analysis.** Modifications are made in the value of one of the dependent variables, while all other variables remain unchanged.
- **Multivariate analysis.** Simultaneous changes are introduced in different variables of the model, allowing highlight the cross effects among them.

In this research work, the development of a stability analysis was considered essential, since there is a significant difference between the number of projects included in the calibration data set and the number of validation projects. As has been shown in the preceding paragraphs there are different ways to study the sensitivity of predictive models in the literature and, in addition, this thesis has studied three different types of modelling techniques. In the case of the models obtained by using LRA, the logarithmic transformation experienced by the variables involves expecting

the smoothness of the function represented by the model. Accordingly, a univariate sensitivity analysis can be used with the models generated by LRA because it allows a clear interpretation of the agreement of the model with the known underlying physical processes of the problem under study. In the case of numerical models and ANNs, the predictor variables of the best selected models have not undergone the logarithmic transformation, that is, they maintain their natural form. The nonlinear functions which represent these models cannot be assumed to have smoothness (differentiability), because, even if there is continuity in some points of such functions they do not have to be smooth. Thus, in the case of numerical models based on FEM and ANNs it is necessary to use another methodology to analyse the sensitivity of the models.

In accordance with the above, in the validation process of predictive models two types of stability analysis are proposed in this thesis depending on the modelling technique to be used:

- For **LRA models** we developed a univariate sensitivity analysis. By visual inspection of graphs, it is possible to examine how well model predictions are in concordance with the known underlying physical processes of the problem under study. That is, the proposed stability analysis aims to verify if the estimations provided by the model are consistent with these underlying processes. This analysis is carried out within the value range of the variables that is valid for applying the predictive model. For that purpose, we first calculate the mean values for independent variables using the sample of calibration projects. These mean values define an “average building” around which revolves the proposed stability analysis. Then, the value of each predictor variable is sequentially modified within a range defined by the limit features of the projects used to develop the model, while the value of the other independent variables remains fixed according to the calculated mean values. The model stability is verified through visual inspection of the graphs representing forecast values versus the variation of values in the independent variables. Anomalous fluctuations in any of these graphs may indicate unstable forecast models within the range of values set for using the model.

- In the case of **FEM-based numerical models and ANNs**, a multivariate analysis is carried out to test the model sensitivity to small variations introduced in the experimental data. This stability analysis intends to determine if small changes in the value of the initial conditions (original values of the variables that have served as the basis for building the model) cause large changes in values predicted by the developed model. Since the initial values of the state variables are always affected by uncertainty, a model that is very sensitive to small changes in the values will present a seemingly chaotic behaviour, reaching very different states when initial experimental data is modified slightly. In a model whose value is defined by the function $y = (v_1, v_2, \dots\dots, v_n)$ and represented by the expression *F(V)*, a V^{ε} modification performed in the data will generate a value in the model $y' = F(V')$. The variation produced in the model is obtained by calculating the distance between *F(V)* and *F(V')*, which we denote by $\|F(V'^{\varepsilon}) - F(V)\|$. The estimation of this error is calculated by averaging the results obtained when modifying the observed values a certain number of times. Thus, for a perturbation V^{ε} we say that a model is stable if and only if $\|F(V'^{\varepsilon}) - F(V)\| \leq \delta$, where δ is a value set by the modeller and expresses the acceptable range of variation.

  As it has been mentioned above, the effect of perturbations depends on the type of change introduced in the values, so different sensitivities can be obtained with different configurations of modifications. In this research work, the stability verification of nonlinear models was carried out on the basis of a multivariate perturbation of the initial data for up to 20%. The Monte Carlo simulation method is used for generating 80 random modifications of the observed values, varying with the same percentage rate of all experimental data through the expression $[x' = x \cdot (1 + z \cdot \varepsilon)]$, where *x* is the observed value, *x'* is the perturbed value, ε is the value of the relative perturbation, which in this research varies from 0 to 0.2, and *z* is a value that follows a normal distribution N(0,1). The stability of the models is verified by calculating the mean value of the existing distance between *F(V)* and *F(V')* through Equation 38, where ε is the value set for the perturbation and *k* is the number of values in the sample for each variable $v_1, v_2, \dots\dots, v_n$.

$$\|F(V'^{\varepsilon}) - F(V)\| = \sum_{i=1}^{k} \frac{\left|F\left(v'_{1,i}, v'_{2,i}, \ldots\ldots, v'_{n,i}\right) - F(v_{1,i}, v_{2,i}, \ldots\ldots, v_{n,i})\right|}{k} \quad (38)$$

III.8.3 Model Sensitivity to Cost Variability

This thesis also analysed the sensitivity of the selected prediction models versus the variability of construction cost caused by the uncertainty in its estimation. The proposed analysis was carried out from two different approaches, which jointly interpreted allow us to appreciate how the aforementioned variability can affect the predictive accuracy of models. The scheme of the developed sensitivity analysis is shown in Figure III-12.

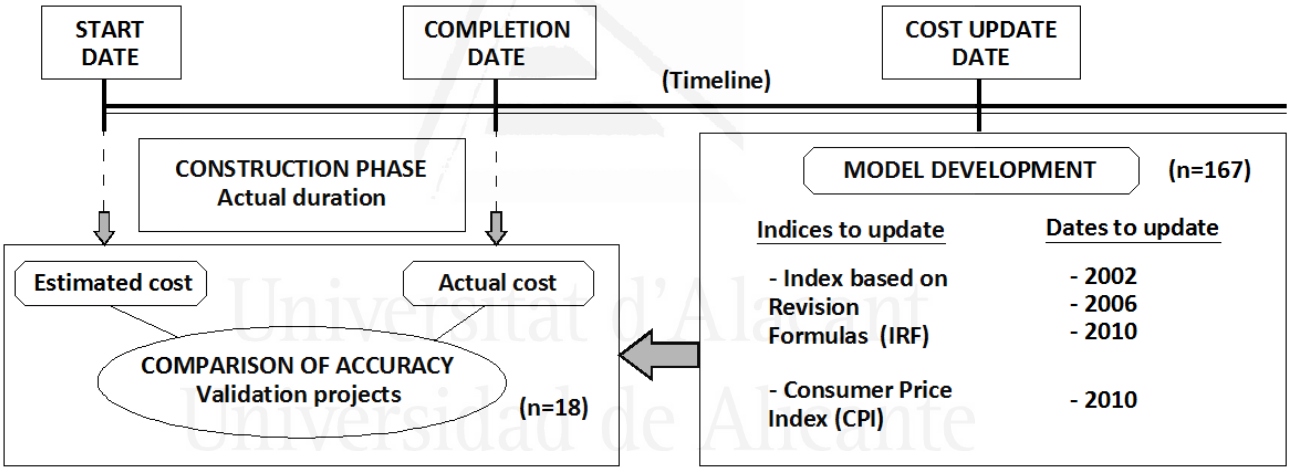


Figure III-12. Scheme to analyse model sensitivity to cost variability.

On one hand, generally, projects undertaken at different time points are used to develop prediction models. To take into account the effects of inflation (Irfan et al., 2011), and establishing a common basis that allows a balanced comparison, construction costs must be adjusted to a certain date over time by using an index to update prices. To the best of our knowledge, in the literature there is no methodology defined for this process and different dates and indices can be used to update prices. With this perspective in mind, three different dates to update construction costs were selected to develop the best linear regression model: February 2010, February 2006

and February 2002. These dates to update prices were used along with coefficients of price revision based on polynomial formulas used by the Spanish public administration in construction projects. In this thesis, these coefficients were referred to as IRF (Index based on Revision Formulas). Additionally, a general index, popularly known as CPI (Consumer Price Index), was also selected to update prices. The CPI was used only with the update date of February 2010.

On the other hand, in the planning phase of works it is not possible to know what will be the actual cost to complete the construction stage, but nevertheless it is in this phase when the practical application of a forecast model makes sense. Although the sample of projects used in the development of the models contained only the final costs of construction, the sample of validation projects had both the actual cost and the costs estimated before starting the works. This is because the two samples are derived from totally different databases. By using these cost data, it was possible to develop an accuracy comparative analysis of the models, based on the type of cost used with them.

The aim of the proposed sensitivity analysis is twofold: (1) to evaluate the statement made by Love et al. (2005), since they considered that construction cost is an inadequate variable to predict construction time because its final value varies regarding the initial estimate, and (2), taking into account that predictive models have been developed using actual construction costs, to check whether their forecasting accuracy is negatively affected when these models are tested using the estimated costs of validation projects.

The first approach of the proposed sensitivity analysis was only developed with the best linear regression model, while the second approach, apart from the best linear regression model, was also applied to both the best ANN model and the best FEM-based numerical model.

III.9 Conclusion

In this chapter, theoretical basis and practical aspects of the research methodology adopted in this thesis have been explained and justified in order to achieve the objectives and hypotheses proposed in Chapter I. The developed research work focuses on new builds, so that projects relating to industrial buildings,

rehabilitation and singular constructions were excluded for the study. Finally, 167 projects were considered valid for generating predictive models and they are referred in this thesis as the calibration dataset. Another sample of 18 projects totally independent of the calibration data was selected to validate the predictive ability of the generated models. According to available data, quantitative variables related to construction costs, GFA and number of floors were selected for research. In addition, considering that there is no general agreement in the literature about which is the most appropriate dependent variable to predict the duration of the construction process, two types of response variables were proposed for analysis: construction time and construction speed.

Three modelling methodologies were selected to generate predictive models: MLRA, ANNs, and a FEM-based numerical methodology. The rationale for the choice of these three modelling techniques was the following: first, most statistical models presented so far for estimating construction time have been developed by using LRA and, consequently, this is the modelling tool that must work as a benchmark; second, it is clear that MLRA and ANNs have become two competing empirical model-building methods and, thus, in order to develop nonlinear models with better predictive performance it is mandatory the use of ANNs; finally, a novel numerical modelling technique based on FEM was introduced into the study since it has not been applied previously in order to generate models to estimate the construction speed of building projects. The three modelling methodologies were described in detail in this chapter.

MLP and RBF neural networks have been the two types of ANN architectures most commonly used in the literature and their predictive performance was compared in this thesis. In the case of MLP networks, the modelling process was divided in five main steps: (1) selection of the training methodology, (2) data division, (3) design of the initial network structure, (4) MLP network optimisation, and (5) validation of optimised models. An optimisation process was elaborated to develop MLP networks, which involves three stages in the case of the GD algorithm and only two steps for the SCG training algorithm. Moreover, with the purpose of analysing the effect of the calibration data division on the performance of MLP models, five different types of data division were analysed.

Although MAPE has been used in previous research to compare the accuracy of different linear regression models related to building projects, in light of some

deficiencies detected in the literature, its use was discarded in this thesis. Instead, RMSE was selected as the measure of accuracy for evaluating the performance and validity of the developed models. The process proposed to evaluate the predictive performance of the forecasting models includes five different assessment criteria: (i) verification of compliance with the underlying assumptions regarding the statistical procedure used to obtain the models, (ii) checking the goodness of fit of the models to the data set used for generating them, (iii) validation of models in terms of ability to generalise, (iv) assessment of the balance existing between ability to generalise and the accuracy obtained with the calibration data, and (v) development of a sensitivity analysis to verify the model stability.

Finally, a sensitivity analysis was also proposed to analyse the impact of the variability of construction costs on the predictive performance of the generated models.

Chapter IV

Linear Regression Models

IV. Linear Regression Models

IV.1 Bivariate Correlations

In order to identify the most appropriate response variable for carrying out LRA, first the Pearson's correlation coefficient was used to assess the bivariate correlations between the variables under consideration. Table IV-1 shows the correlations found between the response variables raised in the research and the independent variables selected for analysis. When construction time is used as a response variable, the analysis results showed weak correlation values with the independent variables. On the contrary, using the construction speed as a dependent variable we obtained very high correlation values (all significant at the 0.01 level), especially with the variable related to GFA and, to a lesser degree, with the variable representing the construction cost.

		T_Cost	***T_GFA***	***T_Floors***	***A_Floors***	***B_Floors***	***T_GFA/ T_Floors***	***Standard***
Time	*r*	0.274**	0.275**	0.225**	0.176*	0.257**	0.195*	0.055
	Sig.	0.000	0.000	0.003	0.023	0.001	0.011	0.477
Speed	*r*	0.794**	0.916**	0.703**	0.688**	0.519**	0.726**	-0.363**
	Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000

*. The correlation is significant at the 0.05 level. - **. The correlation is significant at the 0.01 level.

Table IV-1. Summary of correlations obtained between independent and response variables.

IV.2 Linear Regression Models

The results observed with the bivariate correlations were first endorsed by developing simple linear regression models with the two predictor variables that obtained higher correlation values: *T_GFA* and *T_Cost.* As can be seen in Table IV-2, the models which use the time variable yield very low values of the coefficient of

determination, whereas using the construction speed variable such values are much higher.

Calibration					
Dependent variable	**Independent variable**	**Adjusted R^2**	***F*-Value**	**Sig.**	**Durbin-Watson statistic**
Time	*T_Cost*	0.069	13.351	0.000	2.029
	T_GFA	0.070	13.459	0.000	2.051
Speed	*T_Cost*	0.629	282.199	0.000	1.547
	T_GFA	0.839	865.816	0.000	1.878

Table IV-2. Predictive performance of models developed by using simple LRA.

These results provide strong evidence to support the first hypothesis (H-01) proposed at the beginning of this thesis, that is, "*construction speed is a more appropriate dependent variable than construction time in order to generate predictive models for estimating the duration of the construction process of new builds*". Furthermore, the *T_GFA* variable produced the highest value of the coefficient of determination.

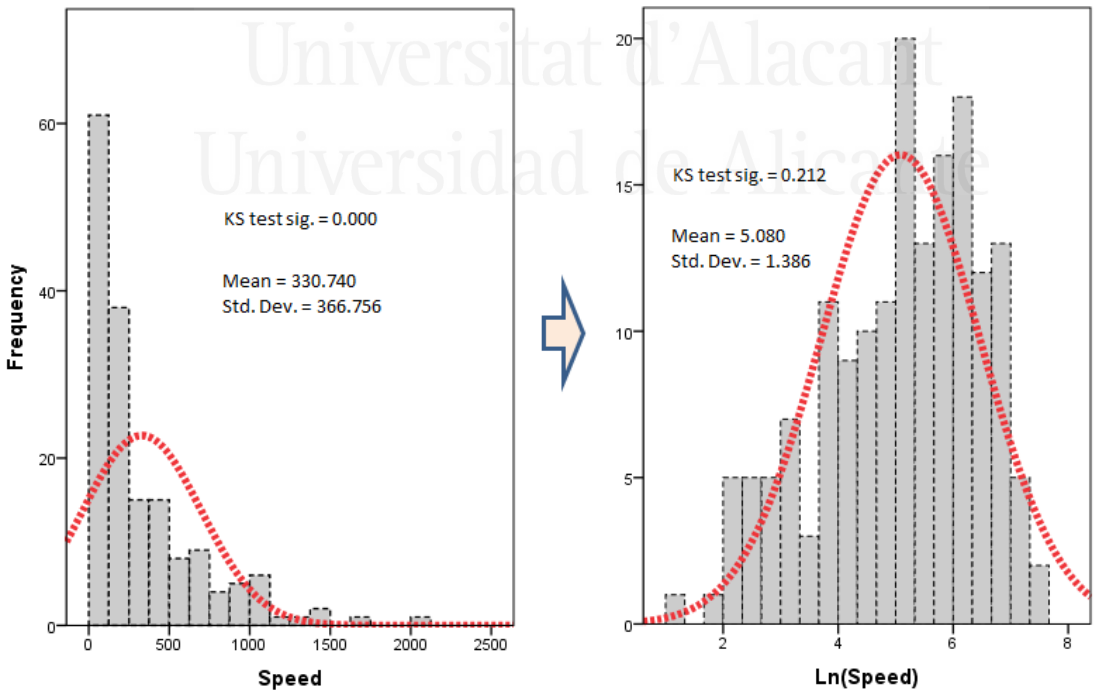


Figure IV-1. Frequency distribution diagrams of construction speed.

In parametric modelling techniques, the distribution function of the response variable plays a vital role (Dursun & Stoy, 2011a). Consequently, considering that the analysis of correlations and the first lineal models identified construction speed as the most appropriate dependent variable to develop the multiple regression analysis, we proceeded to test the hypothesis of normality of it. By using frequency distribution plots and the Kolmogorov–Smirnov test, it was found that the values of construction speed did not conform to the normal distribution. In contrast, the desired fit was obtained by using the logarithmic form of construction speed (see Figure IV-1). These results agree with those presented by Trietsch & Baker (2012) and Trietsch et al. (2012), who showed that time is log-normally distributed when expressed as a multiple of estimated time.

When we tried to use the response variable in its natural form, the regression analysis offered models with high R^2 values, but they did not meet with basic assumptions of LRA. In contrast, by using the logarithmic form of construction speed we obtained models that were consistent with the basic assumptions of the developed regression analysis. In particular, Table IV-3 shows the performance data of four simple LRA models, with and without logarithmic transformation of variables.

Model	**Calibration**						**Validation**	**PI**
	Adj. R^2	***F*-Value**	**Sig.**	**Durbin-Watson statistic**	**Adj. R^2 (after tr.)**	**RMSE**	**RMSE**	
Only *T_Cost* without tr.	0.629	282.199	0.000	1.547	0.629	222.108	95.492	158.800
Only *T_Cost* after log tr.	0.808	701.261	0.000	1.718	0.533	249.187	43.238	146.213
Only *T_GFA* without tr.	0.839	865.816	0.000	1.878	0.839	146.293	32.544	89.419
Only *T_GFA* after log tr.	0.907	1628.287	0.000	1.973	0.845	143.389	25.936	84.663

Table IV-3. Predictive performance obtained with *T_GFA* and *T_Cost*, without transformation and after logarithmic transformation.

The first two models were built using only construction cost as predictive variable of construction speed, while the other two models were generated using only GFA.

The logarithmic form of the model with construction cost as predictor variable corresponds to the BTC model. According to the presented results, after the measures of accuracy and goodness of fit were transformed back to its natural scale, it is evident that GFA clearly outperforms the predictive accuracy provided by construction costs. Moreover, in the case of GFA, although the values of the coefficient of determination were similar, it can be said that the logarithmic transformation was useful as the prediction error was reduced in both the data set used to calibrate the model and the data set used for validation purposes. But what is even more important is that the logarithmic model meets the basic assumptions of LRA, while the same does not occur with the model that uses the variables without any transformation.

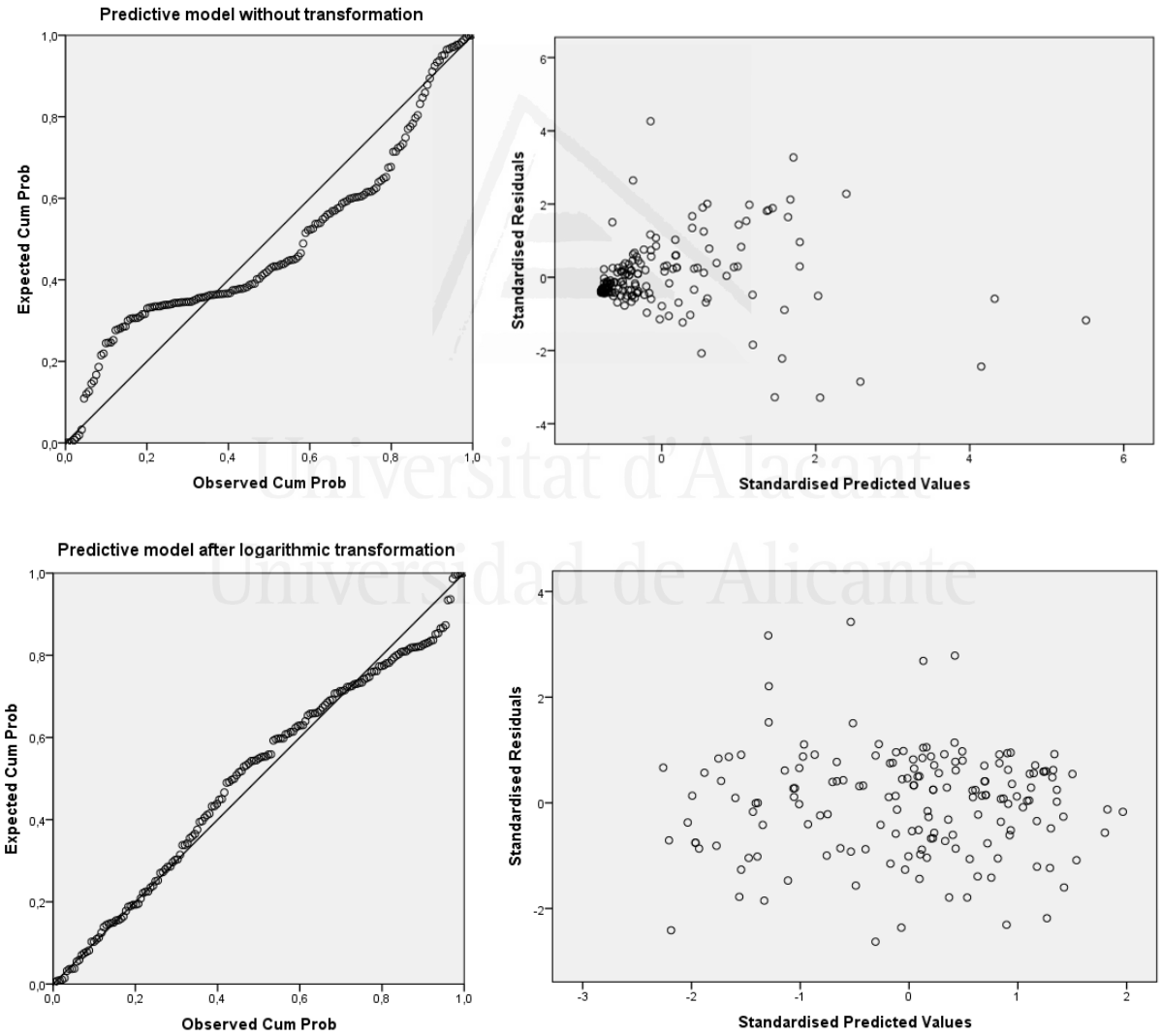


Figure IV-2. Normal Q-Q charts and scatter plots of standardised residuals for the model which uses only the *T_GFA* predictor variable, before and after logarithmic transformation.

Figure IV-2 shows Q–Q charts and scatter plots for the residuals of the two models developed with GFA as predictor variable. As we can see in the charts, the residuals do not follow a normal distribution before transformation, but they are apparently normal after transformation. In the same way, the scatter plots show that the residual variance is not constant before transformation, but it is constant after transformation.

In order to test the second hypothesis (H-02) proposed in this thesis, the predictive accuracy of the two logarithmic models was compared using the MW *U* test. The result of this test showed that there was statistically significant difference in the predictive accuracy of these two models (p<0.05). Therefore, it seems reasonable to assert that "*GFA has greater influence on the construction speed of new builds than construction costs*".

Once both the most significant predictor variable (*T_GFA*) and the most appropriate response variable (*Speed*) were identified, the next step in the research was to try to incorporate new independent variables into the initial simple regression model. According to the preliminary results obtained using MLRA, a logarithmic model, which contains the main project scope factors used in the literature, was selected as a basis for the study. The selected model, named as *Base* model, besides obtaining a high value of goodness of fit (see Table IV-4) and better predictive accuracy with the validation data than the model which uses only the *T_GFA* variable, resulting in a reduction of the RMSE value of 9.61%, provides a clear and coherent understanding of relationships between the variables that define it. The *Base* model is represented by Equation 39.

$$ln(Speed) = -0.609 + 0.928\ln(T\_GFA) - 0.167\ln(T\_Floors) - 0.232\ln(Standard) \quad (39)$$

where *Speed* is the construction speed, *T_GFA* is the total gross floor area, *T_Floors* is the total number of floors and *Standard* is the total cost of construction (*T_Cost*) divided by *T_GFA*.

Table IV-4 shows a value of 0.838 (after transformation of errors) for the adjusted R^2 coefficient obtained with the *Base* model, which represents a good fit of the model to the projects used in its generation. This value means that about 84% of variations in the analysed projects are explained by the independent variables that are part of the

model. Additionally, by replacing some of the predictor variables in the *Base* model, we obtained two equivalent models. These models do not provide greater values of goodness of fit and accuracy, but, as will be seen later, they help us to interpret more clearly the meaning and significance of the project scope factors involved in the *Base* model. The difference between these derived models (see Table IV-5) lies in: substitution of the *T_GFA* variable by the *T_Cost* variable (*Derived_1*) and substitution of the *T_Floors* variable by the *T_GFA/T_Floors* variable (*Derived_2*).

Model	Calibration						Validation	PI
	Adj. R^2	*F*-Value	Sig.	Durbin-Watson statistic	Adj. R^2 (after tr.)	RMSE (after tr.)	RMSE	
Base	0.914	588.505	0.000	1.933	0.838	145.982	23.443	84.713
Derived_1	0.914	588.504	0.000	1.933	0.838	145.982	23.443	84.713
Derived_2	0.914	588.505	0.000	1.933	0.838	145.982	23.443	84.713
Improved_1	0.916	453.980	0.000	1.959	0.858	135.944	20.884	78.414
Improved_2	0.921	386.456	0.000	2.055	0.884	122.508	19.250	70.879
Improved_3	0.920	475.934	0.000	2.036	0.882	124.206	19.543	71.875

Table IV-4. Summary of the most important statistical data of MLRA models.

Considering the standardised coefficients of these three equivalent models (see Table IV-5), it is possible to deduce that GFA and construction cost are the variables with the greatest influence on construction speed. Delving deeper into the issue, the substitution of the *T_GFA* variable by the *T_Cost* variable (*Derived_1* model) produces no change of fit and accuracy regarding the *Base* model. However, the development of a simple LRA, in an individual way with each variable, clearly showed a higher influence of GFA on construction speed (see Table IV-3).

The possible existence of specification errors in the *Base* model, motivated by omission of important variables, was also analysed. For that purpose, firstly, new independent variables were added to the *Base* model, chosen from among the proposals in Table III-2, and then new analyses of linear regression were carried out.

Model	Independent variables	Substituted or added variables		Regression coefficients		Standardised coefficients	*t*	Sig.
				B	**Std. error**	**Beta**		
Base	(Constant)			-0.609	0.707		-0.861	0.391
	Ln(*T_GFA*)			0.928	0.034	0.980	27.231	0.000
	Ln(*T_Floors*)			-0.167	0.065	-0.085	-2.566	0.011
	Ln(*Standard*)			-0.232	0.089	-0.069	-2.616	0.010
Derived_1	(Constant)	*T_GFA* is substituted by *T_Cost*		-0.609	0.707		-0.861	0.391
	Ln(*T_Cost*)			0.928	0.034	0.878	27.231	0.000
	Ln(*T_Floors*)			-0.167	0.065	-0.085	-2.566	0.011
	Ln(*Standard*)			-1.161	0.081	-0.343	14.287	0.000
Derived_2	(Constant)	*T_Floors* is substituted by *T_GFA /T_Floors*		-0.609	0.707		-0.861	0.391
	Ln(*T_GFA*)			0.761	0.048	0.804	15.703	0.000
	Ln(*T_GFA/ T_Floors*)			0.167	0.065	0.129	2.566	0.011
	Ln(*Standard*)			-0.232	0.089	-0.069	-2.616	0.010
Improved_1	(Constant)	*A_Floors* is added in its natural form		-0.844	0.706		-1.195	0.234
	Ln(*T_GFA*)			0.927	0.034	0.978	27.516	0.000
	Ln(*T_Floors*)			-0.375	0.112	-0.190	-3.356	0.001
	Ln(*Standard*)			-0.195	0.089	-0.057	-2.180	0.031
	A_Floors			0.074	0.033	0.122	2.275	0.024
Improved_2	(Constant)	*Type* is added		0.754	0.760		0.992	0.323
	Ln(*T_GFA*)			0.850	0.041	0.897	20.976	0.000
	Ln(*T_Floors*)			-0.118	0.066	-0.060	-1.778	0.077
	Ln(*Standard*)			-0.359	0.092	-0.106	-3.885	0.000
	Type	*Type 2*	yes	-0.276	0.136	-0.055	-2.028	0.044
		Type 3	yes	0.373	0.108	0.084	3.462	0.001
		Others	yes	0.000				
Improved_3	(Constant)	*Type* is added and *T_Floors* is removed		1.093	0.741		1.476	0.142
	Ln(*T_GFA*)			0.799	0.029	0.844	27.647	0.000
	Ln(*Standard*)			-0.372	0.093	-0.110	-4.009	0.000
	Type	*Type 2*	yes	-0.356	0.129	-0.071	-2.752	0.007
		Type 3	yes	0.378	0.108	0.085	3.486	0.001
		Others	yes	0.000				

Table IV-5. Statistical data of the regression coefficients of MLRA models.

By using the number of floors above ground (*A_Floors*) in its natural form, a new model (*Improved_1*) was developed with better fit and greater accuracy than the *Base* model. The *Improved_1* model is a semi-logarithmic model, unlike the other models. Secondly, the possibility that certain types of projects with extreme average speeds (low or high) could be conditioning the accuracy of the *Base* model was also studied. Figure IV-3 shows the average construction speed for each type of facility and, compared to the rest of the buildings, *Type 2* projects (single family housing) have low construction speeds, while the Type *3* projects (offices and commercial) show high construction speeds. With this in mind, the set of calibration projects was classified into three different types of facilities, where the classification of both *Type 2* and *Type* 3 projects was kept under the same conditions and the rest of building projects was regrouped in a single group named as *Others* (see Figure IV-4). In addition, the construction speed of each group was compared with each other by using the KW test. It was necessary to develop this nonparametric test as the assumption of equality of variances is not met and there is a large difference among the size of the three groups. The result of this test showed that the difference among the construction speed of the three groups is statistically significant (p=0.000).

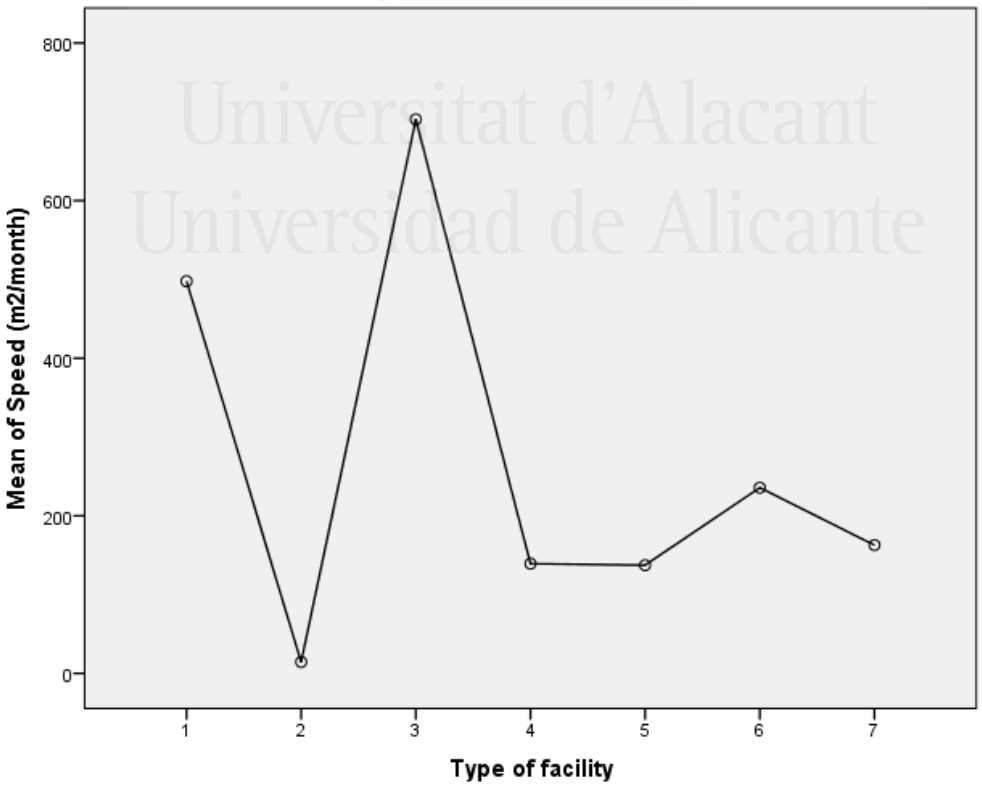


Figure IV-3. Graph representing the mean value of construction speed for each type of facility.

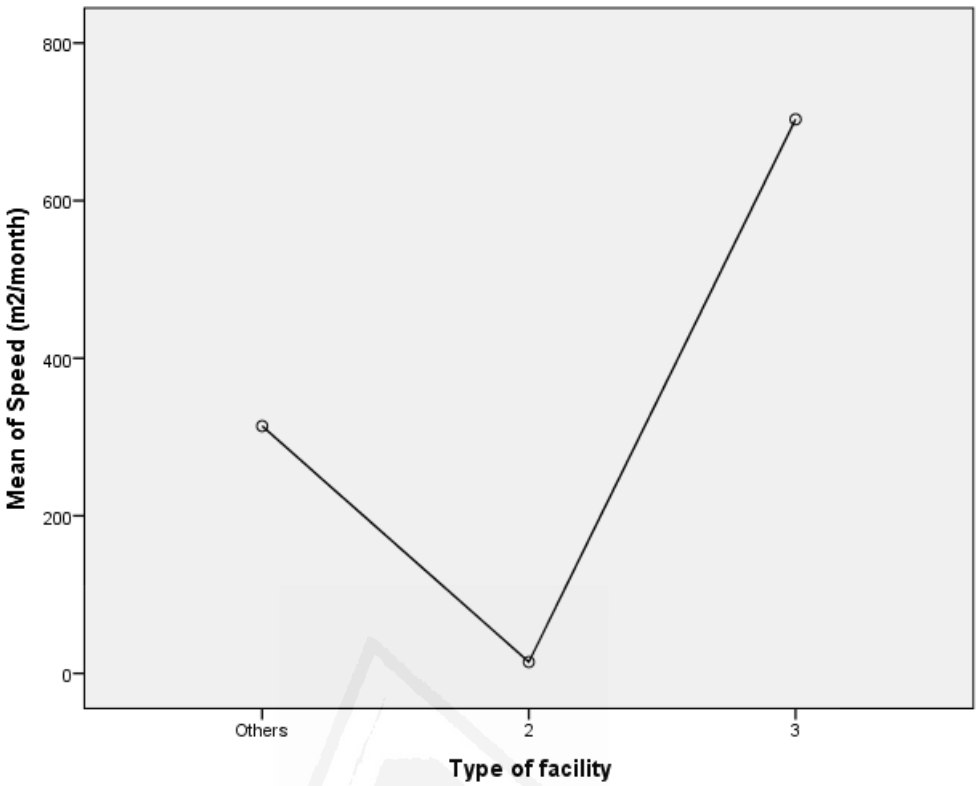


Figure IV-4. Regrouping of calibration projects based on their construction speed.

Taking into consideration the above, two dummy variables were added into the *Base* model to consider the possible effect of projects with extreme-speed characteristics on the estimation of construction speed. In MLRA, this kind of variables takes the value of one, if the project under consideration belongs to the group determined by the variable, and zero otherwise (see, e.g., Love et al., 2005). The development of a new regression analysis with these dummy variables resulted in a new model (*Improved_2*), also with better outcomes of fit and accuracy than those obtained with the *Base* model. It should be noted that the incorporation into the *Base* model of a variable representing the type of facility (*Type*) weakens the statistical significance of the *T_Floors* variable, with the *p*-value slightly exceeding 0.05, but below the limit value of 0.10 established in Chapter III for the development of LRA models. In this regard, it is important to note that the variable representing the number of floors of a building has been used in many parametric models presented in previous research, pointing in all cases the same effect on the construction speed of building projects (see, e.g., Chan & Chan, 2004; Chan & Kumaraswamy, 1995; Ireland, 1986; Love et al., 2005). Nonetheless, we proceeded to build another log-linear model

(*Improved_3*) using the *Type* variable but eliminating the *T_Floors* variable. This new model produced slightly worse accuracy results than the *Improved_2* model in both calibration and validation data, but all predictor variables in this model have a significance level (*p*-value) of less than 1%.

The most significant data about the previously described models are shown in Tables IV-4 and IV-5. It should also be stressed that the constant of the developed models has a *p*-value greater than 0.10. In this regard, it has been established in the literature that models with no intercept should be used provided that they are consistent with the subject matter theory (Chatterjee & Hadi, 2006). Nevertheless, the use of a regression-through-the-origin (RTO) model is controversial, because even when theory bans the use of a constant, it is necessary to consider the observed range of data (Eisenhauer, 2003). Furthermore, Kutner et al. (2005) suggested that using a RTO model is not a safe practice and advised using an intercept model instead. They argue that if the regression line goes through the origin, unless the sample size is small, there will be no detrimental effects in using an intercept model. Conversely, if the regression line does not go through the origin, the use of an intercept regression model will avoid potential difficulties arising from forcing the regression line through the origin when it is not appropriate. In addition, the elimination of the constant in the best two models (*Improved_2* and *Improved_3*) produced a reduction of the accuracy values in both calibration and validation data. Therefore, bearing all of these points in mind, it was decided to keep the constant inside all models.

IV.3 Stability Analysis

The stability analysis proposed in Chapter III was carried out with the *Base* model and the three improved models (*Improved_1*, *Improved_2*, and *Improved_3*). For that purpose, first the mean values for the independent variables of the selected models were computed by using the sample of calibration projects. Figure IV-5 shows the mean values calculated for the independent variables. These mean values define an “average building” around which revolves the stability analysis. Table IV-6 shows the range of allowed values for each independent variable, while the value of the remaining independent variables stays fixed using the calculated mean values. The limit values of each interval were defined by the extreme features of the calibration

projects. The model stability is verified through visual inspection of the graphs that depict forecast values versus the variation of values in the independent variables. Anomalous fluctuations in any of these graphs may indicate unstable forecast models within the range of values set for using the model.

Independent variable	Maximum	Mean	Minimum
T_GFA	52,259.00	6,657.54	108.00
T_Floors	18	5	1
Standard	2,355.18	702.59	229.77

Table IV-6. Range of values used in the stability analysis developed with linear regression models.

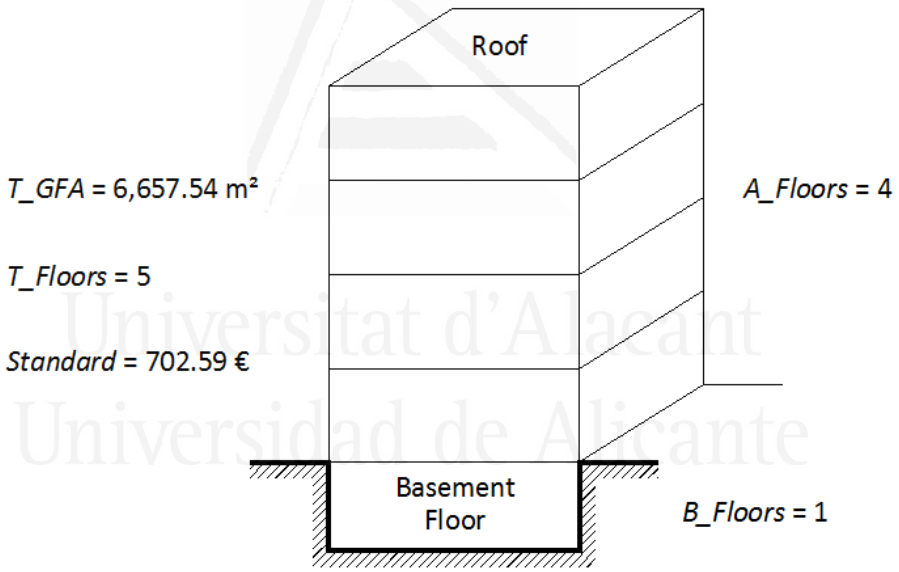


Figure IV-5. Features of the "average building" used as a basis for stability analysis.

In the case of the *Improved_1* model, the number of floors below ground (*B_Floors*) remained constant in the analysis using its mean value (*B_Floors*=1, or what is the same, one basement floor). In this way, the variable representing the number of floors above ground (*A_Floors*) varied in parallel with the variable which represents the total number of floors (*T_Floors*), and its value was always one unit less (*A_Floors* = *T_Floors* − 1). Figures IV-6, IV-7 and IV-8 show graphs depicting the forecasted values for each of the four selected models against the scheduled variation

of values in the independent variables. Figure IV-9 represents how the forecasted construction speed varies in the *Improved_2* model versus the *T_Floors* variable, and according to the used type of facility (*Type 2*, *Type 3*, or *Others*).

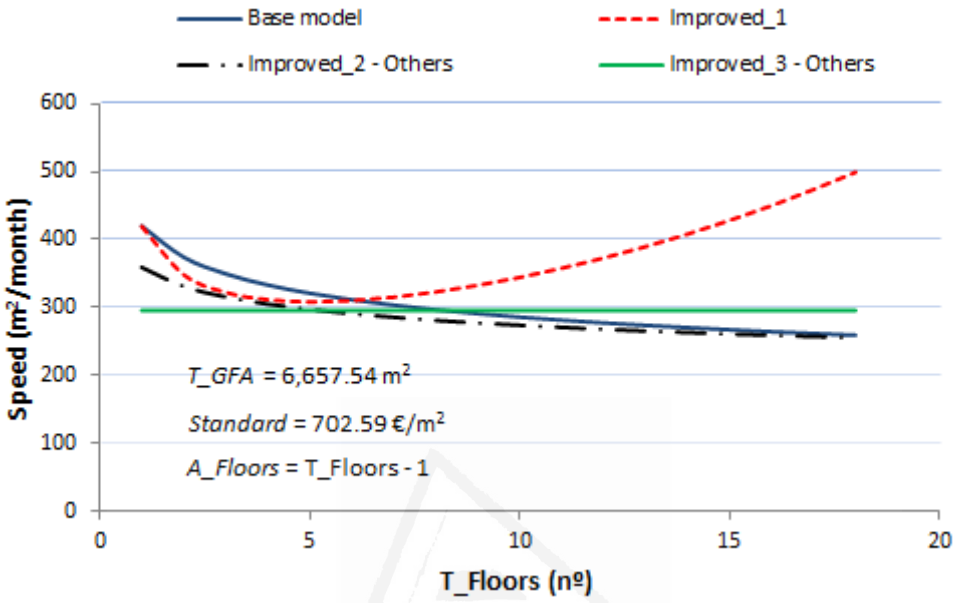


Figure IV-6. Graph representing the predicted values of construction speed versus the variation of number of floors.

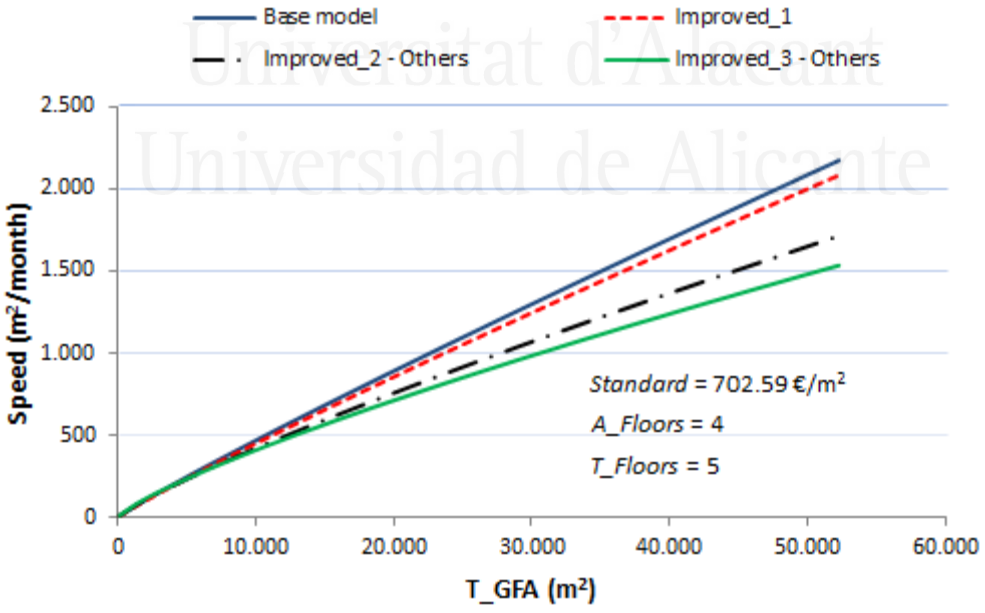


Figure IV-7. Graph representing the predicted values of construction speed versus the variation of GFA.

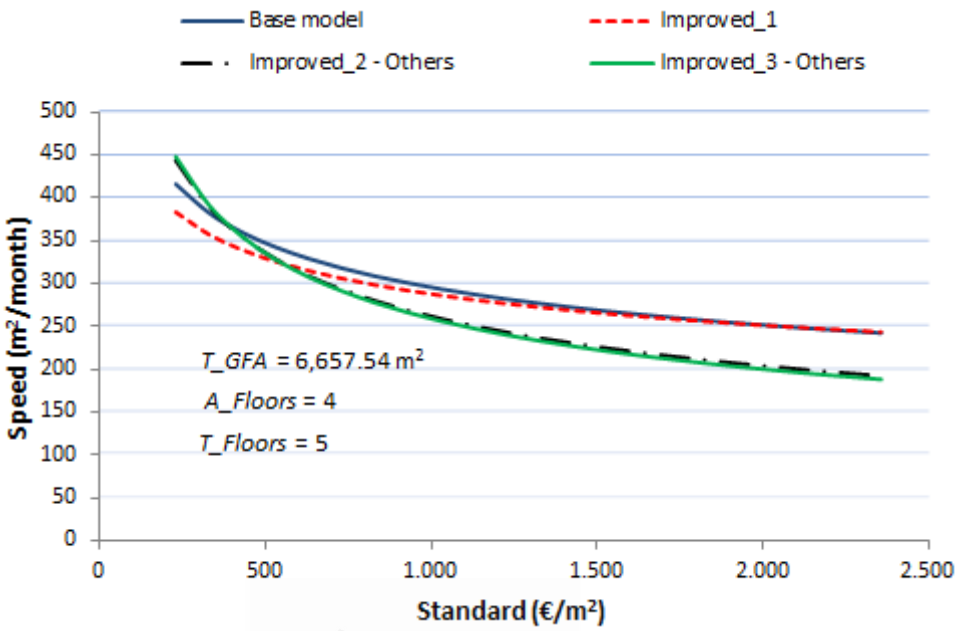


Figure IV-8. Graph representing the predicted values of construction speed versus the variation of the *Standard* variable.

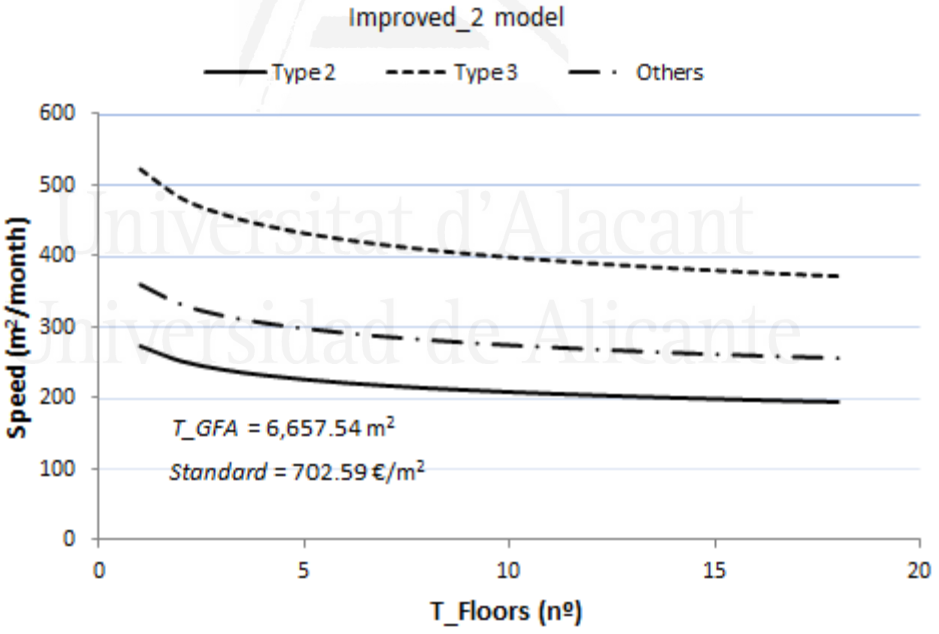


Figure IV-9. Graph representing the predicted values of construction speed versus the variation of number of floors by using the *Improved_2* model.

Considering the displayed graphs, the *Base* model along with the *Improved_2* and the *Improved_3* models provide a stable behaviour when the values of the independent variables were modified within the range defined for the stability test. On

the contrary, the *Improved_1* model shows an anomalous behaviour when the *T_Floors* variable is modified. As we can see in Figure IV-6, initially, in this model the construction speed decreases proportionally with an increase in *T_Floors*, but from the fifth floor onwards this relationship is reversed and it ultimately leads to the absurd result that, remaining the same conditions in the rest of project parameters, a building with 18 floors is built in less time than a one-story building. Although the *Improved_1* model is semi-logarithmic, unlike the rest of the presented models, it is necessary to emphasise that other similar research has presented semi-logarithmic models as valid models (see, e.g., Chan & Chan, 2004; Stoy, Dreier, et al., 2007; Stoy, Pollalis, et al., 2007). However, no stability test was developed to show the validity of such semi-logarithmic models.

Taking into consideration the results of goodness of fit, ability to generalise, and stability, the *Improved_2* model was selected as the most appropriate log-linear model to estimate the construction speed of new builds. With this model the construction speed is represented in logarithmic form by Equation 40 and in its natural form using Equation 41. Given that a dummy variable assumes arbitrary discrete values to represent each class, in statistical terms the *Type* variable is simply a constant differentiating the mean construction speed of each type of facility (Sousa et al., 2014).

$$\ln(Speed) = 0.754 + 0.850\ \ln(T\_GFA) - 0.118\ \ln(T\_Floors) - 0.359\ \ln(Standard) + Type \quad (40)$$

$$Speed = e^{0.754} \cdot e^{0.850\ \ln(T\_GFA)} \cdot e^{-0.118\ \ln(T\_Floors)} \cdot e^{-0.359\ \ln(Standard)} \cdot e^{Type} \quad (41)$$

where *Type* represents the type of facility and is equal to -0.276 for *Type 2* projects, 0.373 for *Type 3* projects and 0 for *Others*.

The practical application of the proposed forecast model would require calculating the construction time (*Time*) and, to that end, it is necessary to carry out a transformation of Equation 41. Using the properties of logarithms, the proposed transformation is expressed by Equation 42.

$$Time = 0.470 * (T\_GFA)^{0.150} * (T\_Floors)^{0.118} * (Standard)^{0.359} * Type \quad (42)$$

where *Type* is equal to 1.318 for *Type 2*, 0.689 for *Type 3* and 1 for *Others*.

IV.4 Regression Diagnostics

The development of a reliable regression model requires the verification of compliance with certain underlying assumptions, as well as identifying possible influential cases. These assumptions along with the criteria used to check them were explained in Chapter III.

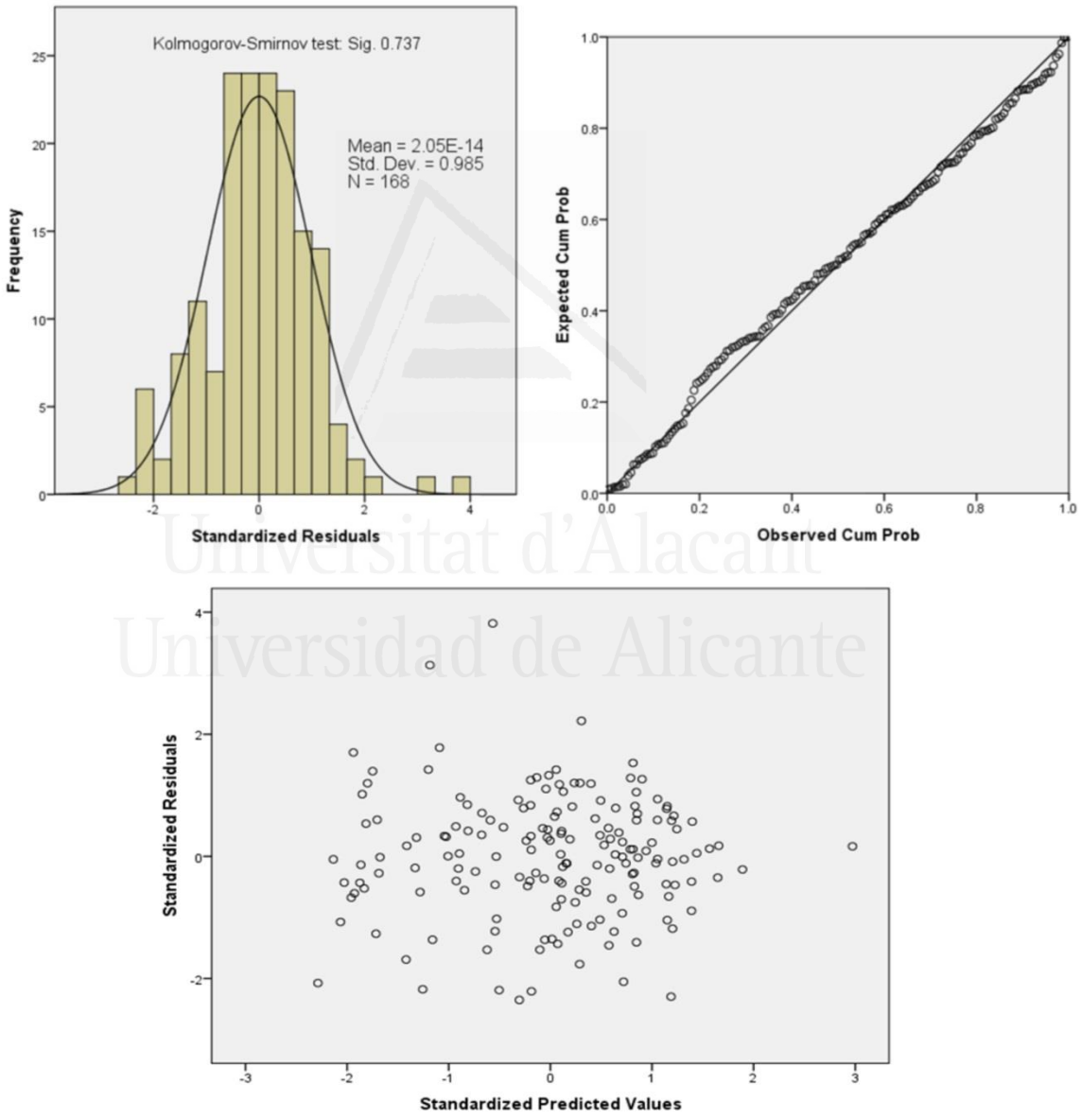


Figure IV-10. Residual analysis of the *Improved_2* model.

Figure IV-10 shows the graphs and diagrams corresponding to the residual analysis conducted with the selected forecasting model. Both the Durbin–Watson statistic (see Table IV-4) and the residual scatter plot show a lack of correlation between the residuals. Regarding the assumption of equal variances, it can be deduced from the scatter plot that the residual variation shows neither trends nor pronounced increases or decreases. Furthermore, almost all values are within a range of two standard deviations.

The joint analysis of the histogram, the cumulative probability plot and the Kolmogorov–Smirnov normality test allows us to state that the normality hypothesis of the residuals has not been violated. Multicollinearity tests undertaken with the SPSS program did not find problems with the independent variables and the tolerance values were higher than 0.01.

Ten projects were also detected with statistical values outside the thresholds set in the study and the regression analysis was developed with and without the identified outliers. The removal of these values did not improve the accuracy of the selected model (see Table IV-7), so that the outliers remained in the calibration data.

Model	Calibration						Validation	PI
	Adjusted R^2	F-Value	Sig.	Durbin-Watson statistic	Adjusted R^2 (after tr.)	RMSE (after tr.)	RMSE	
Improved_2 without outliers	0.921	386.456	0.000	2.055	0.881	124.031	20.654	72.343

Table IV-7. Statistical data obtained with the *Improved_2* model when potential outliers are removed.

IV.5 Model Sensitivity to Cost Variability

According to the proposed research scheme (Figure III-12), a sensitivity analysis was conducted to assess the accuracy of the selected LRA model depending on: (a) the cost update process used in its development and (b) the kind of costs used in its application. All validation projects, except one that fulfilled the estimated cost, ended

with cost overruns. The maximum deviation of the actual cost with respect to the estimated cost was 23.66%, while the mean deviation was 12.34%.

Price Index	Update date	Calibration		Validation		MW *U* test (*p*-value)
		Adjusted R^2 (after tr.)	RMSE (after tr.)	RMSE Estimated cost	RMSE Actual cost	
CPI	2010	0.888	120.555	23.091	19.689	0.628
IRF	2010	0.884	122.508	21.539	19.250	0.501
	2006	0.884	122.700	21.516	19.327	0.481
	2002	0.884	122.626	21.531	19.247	0.501
		KW test (*p-value*)		0.989	0.964	

Table IV-8. Results of the sensitivity analysis carried out for the *Improved_2* model.

Table IV-8 summarises the results of the proposed analysis, including the *p*-value of the nonparametric tests of significance that were conducted to measure differences of predictive accuracy in terms of absolute error. Using the KW test for the results obtained with the three dates to update and the two revision indices, *p*-values were found to be much higher than 0.05. In addition, the MW *U* test was used to analyse the accuracy differences obtained with each model according to the cost type used with the validation projects. The results of this last test indicated that the accuracy differences were also not statistically significant at the 5% level ($p \geq 0.05$).

Based on the results of the developed sensitivity analysis, it can be stated that: (a) there are no statistically significant differences in the accuracy of the selected MLRA model, regardless of the index and the date used to update the construction costs, although the model accuracy worsens if we use a generic price revision index rather than a specific one of the construction sector; and (b) although, on average, the forecasting accuracy is worse when using the costs estimated at the beginning of the project instead of the actual costs observed at the end of the works, the differences in accuracy obtained do not become statistically significant. Therefore, these results show also that, although the proposed model has been obtained using the actual costs of a set of projects, the model works well using the estimated costs, which are the only ones that make sense for predicting the duration of a new project.

IV.6 Interpretation of Regression Coefficients

Table IV-5 shows the regression coefficients obtained with the various linear models discussed throughout this chapter. The standardised regression coefficients (beta coefficients) are directly comparable and a variable has more importance in the regression equation the higher the absolute value of its beta coefficient. Nevertheless, in a linear regression model the effect of the independent variables on the response variable changes significantly according to the shapes adopted by these variables. Considering the logarithmic form of Equation 38, the regression coefficients of the independent variables indicate the percentage increase in the construction speed when such variables increase by 1%.

The beta coefficients obtained with the proposed forecast model indicate that GFA is the most influential factor on the construction speed of new builds and, consequently, in their construction time. According to Stoy, Dreier, et al. (2007), the positive sign of the *T_GFA* variable could indicate that large projects enable more efficient use of production factors and in them the construction speed is higher than in smaller projects. The fact that GFA is the most determinant variable in the model does not invalidate the fact that the *T_Cost* variable also influences the duration of a building project in an important way, although to a lesser degree. As shown in Tables IV-4 and IV-5, the substitution in the *Base* model of the *T_GFA* variable by the *T_Cost* variable determined an equivalent model (*Derived_1*) with the same fit and accuracy as the *Base* model.

The incorporation into the model of a variable (*Standard*) derived from both GFA and cost allows us to interpret that, as a result of a higher construction quality defined by this parameter, the construction speed is negatively affected. That is, the higher the quality level of a project, the greater its complexity and, therefore, more difficulty to develop the building process (Chan & Chan, 2004; Stoy, Dreier, et al., 2007).

In the case of the *T_Floors* variable, a negative regression coefficient in the model indicates that the construction speed of a building will be lower as the number of floors increases. This claim is consistent with the results of the study conducted by Love et al. (2005), but if we compare this variable with GFA, the magnitude of its standardised regression coefficient indicates a significantly weaker relationship with construction speed.

The effect on construction speed that produces the increase in the number of floors could also be explained by using the derived variable *T_GFA*/*T_Floors*. By inspecting the data presented in Tables IV-4 and IV-5, we can see that when replacing the *T_Floors* variable by the *T_GFA*/*T_Floors* variable in the *Base* model we obtained a forecast model (*Derived_2*) with equivalent fit and accuracy. The positive coefficient of this variable in the *Derived_2* model is consistent with the results of the study developed by Chan & Chan (2004), who presented a model where a larger value of the ratio GFA/number of floors involved a shorter duration of the works. The authors argued that a greater GFA per floor means a higher learning effect of the workers, accelerating the construction speed per floor. In other words, an increase of this ratio produces at the same time an increase of the construction speed and a reduction of the final duration. However, other interpretations are also possible, such as when floor areas are large, in which case multiple crews or larger teams can work on each floor at the same time.

The variable that represents the type of facility (*Type*) produces the following two effects on the construction speed provided by the model: (i) if the project belongs to *Type 3* (offices and commercial) there is an increase of 45.2% in the construction speed and (ii) if the project belongs to *Type 2* (single family housing) the *Type* variable produces a reduction of 24.1% in the construction speed. It should be noted that the addition to the *Base* model of a variable that represents the project type weakens the statistical significance of the regression coefficient for the *T_Floors* variable. This is because in the sample of analysed projects the two kinds of buildings represented by the *Type* variable, besides corresponding to projects that have extreme construction speeds (low with *Type 2* and high with *Type 3*), represent buildings also with extreme values in the number of floors (low buildings with *Type 2* and high buildings with *Type 3*).

IV.7 Conclusion

In order to identify the most appropriate response variable for carrying out MLRA and develop appropriate estimation models, first, in this chapter the Pearson's correlation coefficient was used to assess the bivariate correlations existing between the independent variables under study and the two selected dependent variables (*Time* and *Speed*). When construction time is used as the response variable the

analysis results showed weak correlation values with the independent variables under investigation. On the contrary, by using construction speed as dependent variable high correlation values were produced, especially with the *T_GFA* variable. In fact, this factor produced the simple linear regression model with the highest predictive performance. Therefore, in this chapter the first two hypotheses proposed at the beginning of the thesis were tested positively.

It was also found that the values of construction speed did not conform to the normal distribution, which is necessary in order to develop an appropriate LRA. In contrast, the desired fit was obtained by using the logarithmic form of construction speed. Consequently, a logarithmic model was selected as a basis to study the linear relationships between the variables under consideration. This model, which is referred to as *Base* model in the study, contains some of the main project scope factors used in the literature: GFA, total number of floors, and the GFA/cost ratio, which is referred to as *Standard* in this thesis.

From the *Base* model two new equivalent models (*Derived_1* and *Derived_2*) and three improved predictive models (*Improved_1*, *Improved_2*, and *Improved_3*) were built. The *Improved_1* model incorporated a new predictor variable in its natural form representing the number of floors above ground (*A_Floors*), while the *Improved_2* and *Improved_3* models took into consideration certain types of facilities with extreme average construction speeds. In particular, *Type 2* projects (single family housing) had low construction speeds, while *Type 3* projects (offices and commercial) showed high construction speeds. Consequently, two dummy variables were added to the development of MLRA. In theory, the inclusion of dummy variables into prediction models is an excellent option, even if the samples for some types of projects are small (Sousa et al., 2014). Calibration projects were grouped into three different categories taking into consideration their average construction speed and each category represents one or several specific types of facilities that, in theory, have a similar project complexity.

The development of MLRA revealed that both GFA and construction costs are project scope factors of great importance to predict the construction speed of new builds. Although individually GFA has a greater influence on construction speed than costs, by combining both factors in a derived variable (*Standard*) it is possible to achieve predictive models with higher performance. This result suggests the idea that

both variables define different relationships with the construction speed of a building but complementary.

The stability of the *Base* model and the three improved models was verified through visual inspection of the graphs representing forecast values versus the variation of values in the independent variables. Only the *Improved_1* model, which is a semi-log model unlike the rest of the selected models, showed an anomalous behaviour when modifying the *T_Floors* variable, while the rest of selected models offered a stable behaviour. Therefore, special attention should be given to these models in future research.

Taking into consideration the predictive performance obtained with all linear regression models developed in this chapter, the *Improved_2* model was selected as the most appropriate linear model to estimate the construction speed of new builds. Regression diagnostics were carried out with the selected model and they confirmed its reliability. Furthermore, based on the results provided by the sensitivity analysis developed in this study, it can be stated that although the proposed model has been obtained using the actual costs of a set of calibration projects, the model works well using the estimated costs, which are the only ones that make sense for predicting the duration of a new project.

The *Improved_2* model allows us to interpret that in the case of new builds, GFA is the factor that best defines the project scope, understood as a measure of project size. In this regard, the study results seem to confirm that large projects enable more efficient use of production factors, so these projects are constructed with higher construction speed than smaller projects. This might be due to the fact that in large projects the constructor has the opportunity to use multiple crews working at the same time in different zones of the building site. However, the construction speed of building projects is negatively affected when the cost/GFA ratio increases, indicating that a higher construction quality (*Standard*) will result in greater project complexity and lower construction speed. In the case of the number of floors, the *Improved_2* model also suggest that, within the limit values established by the calibration projects, the building works are developed horizontally at a higher speed than vertically. That is, the construction speed of a building will be lower as the number of floors increases, although the magnitude of the standardised regression coefficient indicates a weak relationship of the number of floors with construction speed. This weak relationship

may be due to the fact that a greater GFA per floor might mean a higher learning effect of the workers, accelerating the construction speed per floor, or that multiple crews or larger teams can work on each floor at the same time. Lastly, the *Improved_2* model also reveals that different types of facilities represent different project complexity and construction speed.

Chapter V

Nonlinear Models

V. Nonlinear Models

V.1 Introduction

As it has already been justified in Chapter III of this thesis, linear regression models can provide a starting point from which to explore more sophisticated models. Therefore, the sets of independent variables that obtained the best predictive performance using the MLRA technique were selected to generate nonlinear forecasting models with better accuracy. In Chapter IV, the BE method identified four variables that made a statistically significant contribution to the prediction of construction speed: GFA, number of floors, *Standard* and type of facility. In particular, three sets of variables formed from these four predictor variables were chosen to develop nonlinear models. They represent the linear models named in Chapter IV as *Base*, *Improved_2* and *Improved_3*. That is, these sets of variables incorporate the same predictive variables than those used by the aforementioned linear models. Furthermore, in order to check the effect of transformations on the input variables, and given that the models developed by using ANNs and the FEM-based numerical methodology do not require compliance with the assumptions related to MLRA, each set of variables was used with both logarithmic and natural forms of the variables.

In this chapter of the thesis, first, several models generated by using ANNs are presented. After that, the FEM-based numerical methodology was also applied to develop new nonlinear models. Finally, for the purpose of verifying the third hypothesis proposed in Chapter I, the predictive performance of the best linear regression model was compared to the performance of the best nonlinear models.

V.2 ANN Models

MLP models were first created by using the selected sets of variables along with the 80-20 calibration data division and the GD learning algorithm. These models were developed using the optimisation methodology defined in Chapter III. This methodology follows a stepwise trial and error procedure in which from an initial

network structure with enough consistency six optimised models are generated. By doing so, it was possible to identify the set of variables that produced the prediction models with better predictive performance. After that, the influence of different types of calibration data divisions was also analysed. Lastly, in order to produce better forecasting models, a different training algorithm (SCG) and a different type of network architecture (RBF) were also tested by using the best set of variables and the best calibration data division.

V.2.1 MLP Models Using the GD Training Algorithm

Tables V-1, V-2 and V-3 show the performance data of the six MLP models generated with the proposed optimisation methodology and the three sets of variables under study. In addition, the variables were introduced into the neural network using two different forms: natural and logarithmic.

Data div.	Form	Model	TF SV	LR-M	HN	Training		Test		Validation	
						R^2	RMSE	R^2	RMSE	R^2	RMSE
80-20	Log	*GD-1*	6	0.1/0.5	6	0.833	149.358	0.850	142.115	0.826	23.519
		GD-2			7	0.831	150.024	0.857	138.584	0.820	23.968
		GD-3	5	0.1/0.5	6	0.846	143.499	0.862	136.147	0.846	22.149
		GD-4			7	0.843	144.853	0.857	138.445	0.867	20.557
		GD-5	9	0.1/0.5	4	0.836	147.874	0.850	141.855	0.841	22.508
		GD-6			5	0.848	142.418	0.847	143.377	0.863	20.860
80-20	Natural	*GD-1*	6	0.1/0.5	2	0.816	156.953	0.885	124.420	0.656	33.087
		GD-2			5	0.836	148.066	0.884	125.022	0.662	32.808
		GD-3	5	0.1/0.5	2	0.835	148.486	0.893	119.789	0.609	35.292
		GD-4			3	0.816	156.735	0.886	123.854	0.492	40.218
		GD-5	11	0.1/0.5	2	0.835	148.480	0.887	123.230	0.836	22.815
		GD-6			17	0.845	143.668	0.882	125.752	0.780	26.455

Table V-1. Performance and characteristics of MLP models obtained with the set of variables named as *Base* and the GD training algorithm.

Regarding the models developed with the *Base* set (Table V-1), we can see that, in the case of using the variables in its natural form, it was possible to obtain a better

predictive performance with the test data set than using the variables with logarithmic transformation. However, these results are not confirmed with the validation data, where by using the logarithmic form of the variables it is possible to generate models with greater predictive accuracy than using variables in its natural form, whose performance is quite poor with the exception of one of the models. The models built with the set of variables referred to as *Improved_3* (Table V-2) have a similar behaviour than those developed using the *Base* set, with the only difference that, in general, the accuracy obtained with both calibration and validation data is better.

Data div.	Form	Model	TF SV	LR-M	HN	Training		Test		Validation	
						R^2	RMSE	R^2	RMSE	R^2	RMSE
80-20	Log	*GD-1*	5	0.1/0.5	4	0.886	123.243	0.876	129.246	0.894	18.353
		GD-2			5	0.885	124.032	0.876	129.042	0.899	17.915
		GD-3	11	0.5/0.3	4	0.874	129.910	0.876	129.107	0.868	20.535
		GD-4			17	0.867	133.271	0.878	128.026	0.862	20.940
		GD-5	2	0.5/0.5	3	0.872	130.847	0.862	136.098	0.872	20.225
		GD-6			11	0.873	130.196	0.865	134.557	0.852	21.716
80-20	Natural	*GD-1*	2	0.1/0.5	5	0.883	125.180	0.908	111.030	0.838	22.696
		GD-2			6	0.889	121.767	0.906	112.099	0.849	21.929
		GD-3	6	0.1/0.5	13	0.861	136.234	0.914	107.181	0.556	37.584
		GD-4			9	0.885	123.895	0.913	107.876	0.780	26.489
		GD-5	8	0.1/0.5	3	0.881	126.109	0.914	107.189	0.690	31.405
		GD-6			15	0.891	120.800	0.907	111.449	0.801	25.143

Table V-2. Performance and characteristics of MLP models obtained with the set of variables named as *Improved_3* and the GD training algorithm.

In the case of the *Improved_2* set (Table V-3), the models developed with variables in its natural form produce better predictive performance with both training and test data than the models generated with logarithmic variables. Nonetheless, in this case the results are also confirmed with the validation data, where two models clearly outperform the predictive accuracy offered by the rest of the models. These models were named as *GD-1* and *GD-4* and are highlighted in Table V-3.

Data div.	Form	Model	TF SV	LR-M	HN	Training		Test		Validation	
						R^2	RMSE	R^2	RMSE	R^2	RMSE
80-20	Log	*GD-1*	11	0.9/0.5	17	0.874	129.522	0.881	126.287	0.868	20.519
		GD-2			20	0.884	124.379	0.884	124.599	0.896	18.186
		GD-3	5	0.7/0.7	16	0.879	127.164	0.877	128.286	0.885	19.132
		GD-4			18	0.892	119.952	0.878	127.899	0.900	17.801
		GD-5	7	0.01/0.7	4	0.889	121.738	0.875	129.543	0.902	17.640
		GD-6			6	0.888	122.072	0.876	129.001	0.905	17.431
80-20	Natural	*GD-1*	9	0.1/0.5	13	0.923	101.236	0.913	108.137	0.914	16.551
		GD-2			11	0.921	102.684	0.912	108.780	0.895	18.273
		GD-3	2	0.5/0.3	9	0.898	116.520	0.906	112.151	0.844	22.278
		GD-4			6	0.895	118.459	0.905	113.154	0.918	16.162
		GD-5	11	0.1/0.5	5	0.918	104.596	0.919	104.230	0.826	23.518
		GD-6			11	0.916	105.937	0.913	108.353	0.846	22.124

Table V-3. Performance and characteristics of MLP models obtained with the set of variables named as *Improved_2* and the GD training algorithm.

V.2.2 Influence of Data Division

After identifying the set of variables named as *Improved_2* as the set that produces the best predictive performance with the 80-20 data division, the next step in the research consisted of testing whether by using other types of data division with the calibration projects it could be possible to generate models with better predictive performance. With this aim in mind, as explained in Chapter III, five different random divisions were made according to the following ratios between the percentage of training and test data sets: 50-50, 60-40, 70-30, 80-20, and 90-10 (see Figure III-4).

For each data division the performance of six MLP models developed with the GD training algorithm and the proposed optimisation process is presented in Table V-4. As can be seen, even when the statistical properties of the data subsets are taken into consideration, the results indicate that there are significant differences depending on which proportion of data division is used for training and testing.

Data div.	Form	Model	TF SV	LR-M	HN	Training		Test		Validation	
						R^2	RMSE	R^2	RMSE	R^2	RMSE
50-50	Natural	*GD-1*	9	0.1/0.5	7	0.910	115.193	0.888	115.227	0.880	19.562
		GD-2			4	0.911	114.998	0.884	117.249	0.887	18.937
		GD-3	11	0.1/0.5	12	0.905	118.472	0.886	116.282	0.811	24.500
		GD-4			14	0.912	114.326	0.885	116.784	0.801	25.180
		GD-5	1	0.1/0.5	6	0.896	124.030	0.880	119.394	0.862	20.996
		GD-6			10	0.897	123.627	0.880	119.528	0.819	24.001
60-40	Natural	*GD-1*	11	0.1/0.5	1	0.830	165.076	0.855	116.803	0.817	24.106
		GD-2			12	0.906	123.007	0.852	117.852	0.808	24.702
		GD-3	10	0.1/0.5	1	0.830	165.253	0.848	119.284	0.824	23.671
		GD-4			15	0.900	126.724	0.846	120.173	0.815	24.238
		GD-5	9	0.1/0.5	1	0.830	165.019	0.850	118.675	0.760	27.644
		GD-6			12	0.910	119.962	0.844	121.035	0.746	28.462
70-30	Natural	*GD-1*	12	0.1/0.3	13	0.832	153.658	0.909	103.463	0.200	50.469
		GD-2			17	0.812	162.571	0.907	104.348	0.684	31.737
		GD-3	7	0.1/0.7	11	0.778	176.571	0.908	103.589	0.356	45.263
		GD-4			15	0.810	163.201	0.907	104.329	0.134	52.510
		GD-5	11	0.01/0.3	20	0.889	124.715	0.921	96.077	0.790	25.853
		GD-6			18	0.904	115.918	0.918	98.279	0.786	26.085
80-20	Natural	*GD-1*	9	0.1/0.5	13	0.923	101.236	0.913	108.137	0.914	16.551
		GD-2			11	0.921	102.684	0.912	108.780	0.895	18.273
		GD-3	2	0.5/0.3	9	0.898	116.520	0.906	112.151	0.844	22.278
		GD-4			6	0.895	118.459	0.905	113.154	0.918	16.162
		GD-5	11	0.1/0.5	5	0.918	104.596	0.919	104.230	0.826	23.518
		GD-6			11	0.916	105.937	0.913	108.353	0.846	22.124
90-10	Natural	*GD-1*	6	0.7/0.9	3	0.868	134.636	0.834	156.132	0.482º	40.624
		GD-2			8	0.848	144.373	0.827	159.554	-0.001	56.443
		GD-3	9	0.1/0.5	2	0.887	124.678	0.832	157.182	0.669	32.460
		GD-4			5	0.900	116.878	0.822	162.001	0.622	34.681
		GD-5	11	0.1/0.5	5	0.877	129.885	0.823	161.412	0.750	28.238
		GD-6			17	0.908	112.198	0.819	163.379	0.882	19.361

Table V-4. Performance and characteristics of MLP models obtained with the set of variables named as *Improved_2* and the GD training algorithm, using 5 different types of data division.

The best agreement between the performance obtained with the training and test data sets was achieved by the MLP models which were developed using the 80-20

and the 50-50 data divisions. This consistency between the performances obtained with the calibration data subsets was, in turn, reflected in obtaining better accuracy results in the validation data set.

By focusing our attention on the accuracy results obtained from the validation data set, which is equal for all developed MLP models, it is possible to state that the best ability to generalise was obtained by using the 80-20 data division.

V.2.3 MLP Models Using the SCG Training Algorithm

Once both the best set of predictive variables and the best data division were identified, the developed optimisation process was used again with the SCG training algorithm, which was proposed in this study as an alternative to the standard GD algorithm. Table V-5 shows the performance data of the six MLP models produced by the SCG algorithm and the optimisation process.

Data division	Form	Model	TF SV	HN	Training		Test		Validation	
					R^2	RMSE	R^2	RMSE	R^2	RMSE
80-20	Natural	*SCG-1*	5	7	0.919	104.200	0.924	100.710	0.885	19.163
		SCG-2		16	0.922	102.381	0.923	101.686	0.847	22.041
		SCG-3	6	14	0.913	107.583	0.923	101.669	0.799	25.291
		SCG-4		5	0.915	106.391	0.921	102.979	0.879	19.659
		SCG-5	9	8	0.952	79.765	0.932	95.268	0.515	39.297
		SCG-6		3	0.917	105.244	0.928	98.617	0.780	26.459

Table V-5. Performance and characteristics of MLP models obtained with the set of variables named as *Improved_2* and the SCG training algorithm.

According to the selection criteria established in this research, the *SCG-1* network was chosen as the best prediction model generated by the SCG training algorithm. When comparing this model with the best models obtained with the GD algorithm, it is possible to see that the *SCG-1* model produced greater accuracy in the test data set (RMSE=100.710), resulting in a reduction of the RMSE value of 6.87% and 11.00% regarding the *GD-1* and *GD-4* models respectively. Nevertheless, this model

developed a worse generalisation capability using the validation data set (RMSE=19.163), which means an increase of 15.78% and 18.57% in the RMSE value regarding the *GD-1* and *GD-4* models respectively. In addition, the two models developed by using the GD training algorithm achieved better coherence among the coefficients of determination in the three data subsets. Therefore, the study results seem to show that the traditional GD training algorithm is able to develop models for predicting construction speed with better predictive performance than those generated by using the SCG algorithm.

V.2.4 RBF Models

RBF networks were also generated in this thesis to explore an alternative approach to the MLP architecture and see whether it is possible to improve the predictive ability of the set of variables named as *Improved_2*. Thirty two different combinations of activation functions and scaling of variables were used to develop predictive models, and for each of these combinations the number of neurons was varied from 1 to 20. The 80-20 data division was incorporated to the calibration process, due to the use of the early stopping method.

Table V-6 shows the performance data of the best RBF model. It can clearly be seen that this RBF network offered worse predictive performance than that obtained with the best MLP network and the GD training algorithm.

Data division	Form	TF SV	HN	Training		Test		Validation	
				R^2	RMSE	R^2	RMSE	R^2	RMSE
80-20	Natural	32	8	0.804	161.761	0.890	121.700	0.645	33.624

Table V-6. Performance and characteristics of the best model developed using RBF networks.

V.2.5 Stability Analysis and Selection of the Best ANN model

A large number of ANN models have been presented throughout this section. These models were developed by using two types of architectures, different types of data divisions and several training algorithms. According to the results presented in this section, two MLP models clearly outperformed the predictive accuracy offered by the rest of the developed ANN models. These models, named as *GD-1* and *GD-4*, were built with the MLP architecture, the 80-20 division for the calibration data, the set of variables referred to as *Improved_2*, and the GD training algorithm. Their predictive performance data can be seen in Table V-7, where the results of the calibration group involve both training and test data sets, and in Table V-3.

Set of input variables	Form	Model	Calibration		Validation	PI	Stability (% var.)
			Adjusted R^2	RMSE	RMSE		
Improved_2	Natural	*GD-1*	0.921	102.678	16.551	59.615	3.66%
		GD-4	0.897	117.398	16.162	66.780	3.00%

Table V-7. Predictive performance of the best MLP models.

The *GD-4* model obtained a better RMSE value in the validation set (RMSE=16.162) than the *GD-1* model (RMSE=16.551) resulting in a slight reduction of the RMSE value of 2.35%. However, the *GD-1* model achieved a better RMSE value in the test set (RMSE=108.137) than the *GD-4* model (RMSE=113.154), obtaining in this way a greater reduction of the RMSE value (4.43%). In addition, the *GD-1* model presented a slightly better agreement among the coefficients of determination obtained with the three data sets and the lowest performance index (PI=59.615), with an overall reduction in the RMSE value of the calibration data of 12.54%.

Both models were subjected to the stability analysis presented in the assessment methodology proposed in Chapter III. As we can see in Table V-7, random perturbations of up to 20% in the training data set generated a variation in the models that did not exceed 3.66% (*GD-1*) and 3.00% (*GD-4*), which means that the selected

models are little sensitive to variations introduced in the initial observations and, thus, they can be considered stable (see Figure V-1).

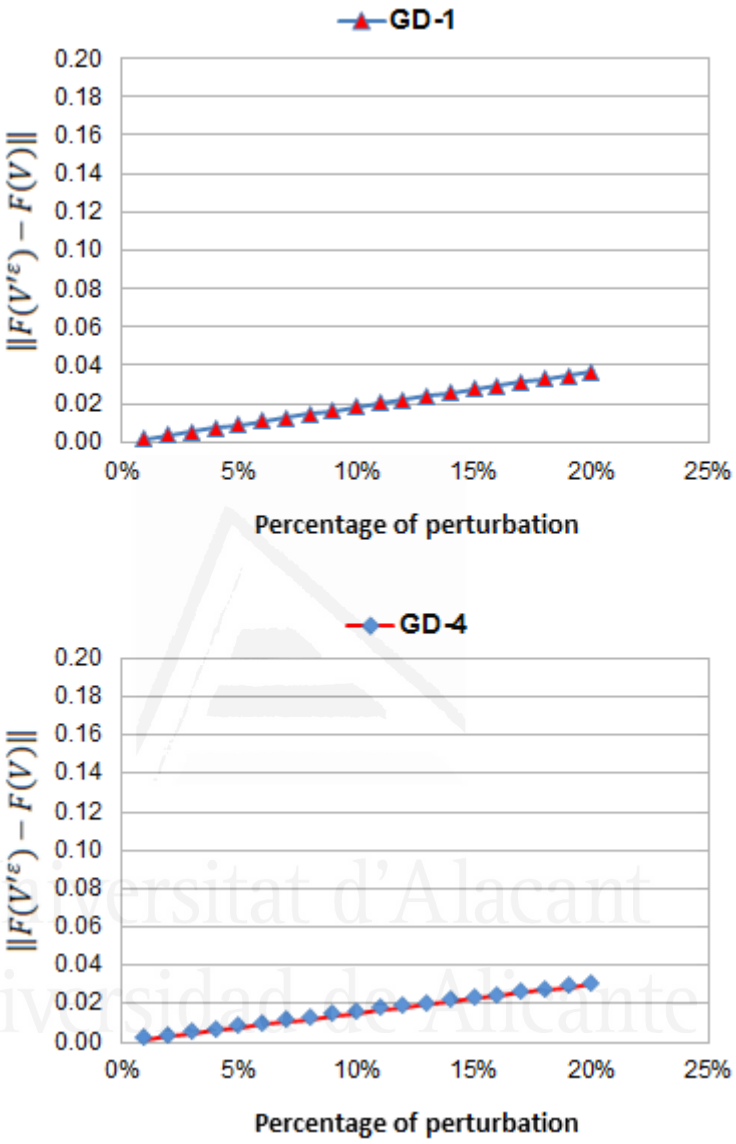


Figure V-1. Graphs showing the stability analysis applied to the best MLP models.

Finally, taking into account all the performance measures presented in this section, the *GD-1* model, which was built with the MLP architecture and the 80-20 calibration data division, was chosen as the best ANN model. Figure V-2 shows the network topology [6-13-1] of the chosen MLP model, with 6 input nodes (including one neuron for each category of the qualitative variable representing the type of facility), 13 hidden nodes and 1 output node (construction speed).

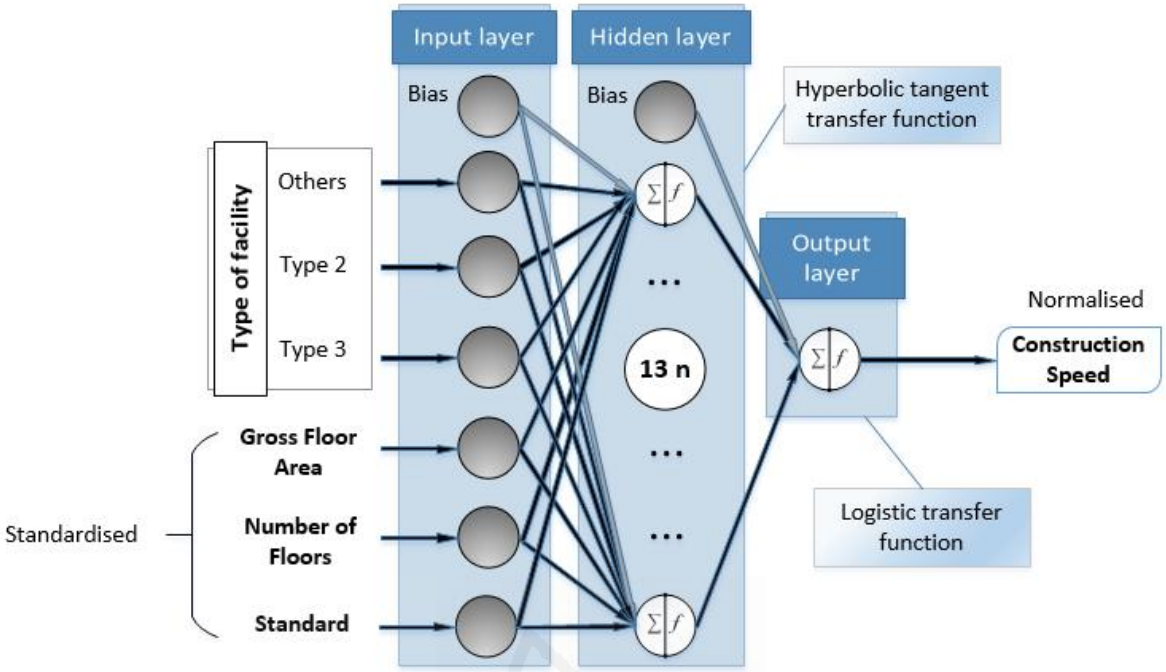


Figure V-2. Topology of the best MLP network.

V.2.6 Model Sensitivity to Cost Variability

Similar to what was done with the best linear regression model, a nonparametric test was also conducted to study how the variability of the *T_Cost* variable affects the predictive performance of the best ANN model. In the case of nonlinear models, only the accuracy differences obtained according to the cost type used with the validation projects were analysed. The results of the MW *U* test indicated that the accuracy differences are not statistically significant at the 5% level (see Table V-8). However, it must be emphasised that the percentage increase in the RMSE value is notably higher in the best ANN model (21.87%) than in the best MLRA model (11.89%).

Price Index	Update date	Calibration		Validation		MW *U* test (*p*-value)
		Adjusted R^2	RMSE	RMSE Estimated cost	RMSE Actual cost	
IRF	2010	0.921	102.678	20.171	16.551	0.864

Table V-8. Results of the sensitivity analysis conducted with the best ANN model.

V.2.7 Analysis of the MLP Network Optimisation Process

V.2.7.1 Influence of the Network Structure and Training Parameters

This research made use of so-called coefficient of variation (CV) in order to evaluate the influence of the network structure and learning parameters on the accuracy obtained with the test data set during the optimisation process. The CV value is defined as the ratio between the standard deviation (σ) and the mean of the population (μ), and is usually expressed as a percentage by using the Equation 43.

$$\mathrm{CV} = \frac{\sigma}{\mu} \tag{43}$$

The CV shows a better interpretation of the extent of variability of a variable than the standard deviation. The higher the CV value, the greater the heterogeneity of the variable. In the case of the RMSE value this means a larger effect of the parameter under study on the prediction error.

Stage	Decision Parameter	TF SV	LR-M	HN	Test data set (80-20)				
					RMSE				
					Min.	Max.	μ	σ	CV
1	TF-SV (12 Comb.)	-	0.1/0.5	13	108.137	123.940	117.141	4.183	3.57%
2	LR-M (20 Comb.)	9	-	13	108.137	108.137	108.137	0.000	0.00%
		2	-	13	113.704	113.713	113.713	0.002	0.00%
		11	-	13	113.719	113.719	113.719	0.000	0.00%
3	HN (1 to 20)	9	0.1/0.5	-	108.137	140.859	112.044	7.046	6.29%
		2	0.5/0.3	-	112.151	140.331	116.764	5.809	4.97%
		11	0.1/0.5	-	104.230	139.318	112.275	6.827	6.08%

Table V-9. Statistical data related to the lowest RMSE value obtained in the test data set (80-20) by using different configurations at each stage of the optimisation process.

The selected 80-20 data division and the GD training algorithm were used to develop the proposed analysis of influence. Table V-9 shows the CV values calculated for the lowest RMSE value obtained in the test data set with different configurations of

MLP networks. The CV value was calculated at each stage of the proposed optimisation process. According to the obtained results, the number of hidden neurons (CV between 4.97% and 6.29%) was the parameter that most affected the accuracy of MLP networks during the optimisation process, while selecting the combination of transfer functions and scaling of variables affected the accuracy of the developed networks to a lesser extent (CV=3.57%). Moreover, it is noteworthy that the values of learning rate and momentum coefficient used to train the neural networks had no influence in the accuracy obtained in the test data set (CV=0.00%).

V.2.7.2 Effect of the Set of Initial Weights

As already noted above, by using the BP algorithm the goodness of the training process is sensitive to the selection of a particular set of initial weights. In order to assess the importance of this choice, the accuracy achieved in the test data set (80-20 data division) with the GD algorithm was evaluated after 100 different executions by using the initial network structure and the best optimised network structure. Due to the existence of randomness in the selection process of initial weights, these executions are equivalent to the selection of different sets of initial weights. Table V-10 shows the statistical data related to the RMSE values obtained in the test data set.

MLP structure	**TF SV**	**LR-M**	**HN**	**Test data set (80-20)**				
				RMSE				
				Min.	**Max.**	***μ***	***σ***	**CV**
Initial network structure	2	0.1/0.5	13	113.704	142.257	127.570	7.206	5.65%
GD-1	9	0.1/0.5	13	108.137	126.464	119.195	3.123	2.62%

Table V-10. Statistical data related to the RMSE values obtained in the test data set (80-20) after 100 executions by using the initial network structure and the best optimised network structure.

According to the presented results, in the case of the initial network structure the variability of the RMSE values (CV=5.65%) was similar to the variability produced by the selection of the number of hidden neurons. Nevertheless, this variability was reduced in the case of the best optimised network structure (CV=2.62%). Thus, even when the used software applies the alternated simulated annealing method in an

attempt to select the best set of initial weights, it can be inferred that such a selection continues to have an important role in obtaining appropriate MLP models to predict the construction speed. Furthermore, it appears that the degree of impact of the initial weights on the training process of MLP networks may be reduced by using an optimised network structure.

As can be seen in Table V-10, another point of interest is the fact that the number of neurons selected for the initial network structure, before starting the optimisation process, is the same as that used by the best MLP model. This finding suggests that the use of the equation $2m+1$ to set the number of hidden neurons in a MLP structure can provide good predictive models.

V.2.8 Importance Analysis of the Input Variables

The study of the effect of each input variable on ANN models is one of the most critical aspects of their application due to the fact that the parameters obtained by the network do not have a practical interpretation, unlike traditional linear regression models. As a result, ANNs are regarded as black-box techniques. However, several methods have been proposed since the late 1980s in order to understand the effect produced by the network's input variables on the output variable. These methods can be divided into two types of methodologies: (i) analysis based on the magnitude of synaptic weights and (ii) sensitivity analysis (Montaño, 2002).

In the case of the SPSS software, it is possible to perform a sensitivity analysis for the purpose of calculating the importance of each predictor variable in determining the construction speed. This analysis is based on the combined use of the training and testing samples. The importance of an input variable is the measure of how much change there is in the value predicted by the ANN model when the value of this input variable is modified. In addition, the normalised importance of a variable is the result of dividing the importance of such a variable by the maximum value of importance of all input variables, and it is expressed as a percentage value. Table V-11 and Figure V-3 show the results obtained with the best MLP model when performing the proposed sensitivity analysis.

Input variable	Importance	Normalised Importance	Ranking
T_GFA	0.585	100.00%	1
T_Floors	0.094	16.04%	4
Standard	0.131	22.45%	3
Type	0.190	32.48%	2

Table V-11. Importance of each of the input variables used with the best MLP model.

According to the obtained results, it is clear that the best MLP model is dominated by GFA, followed by the type of facility, which has acquired greater importance in the MLP model than the *Standard* variable. However, by using this sensitivity analysis is not possible to know the direction of the relationships between the variables under study.

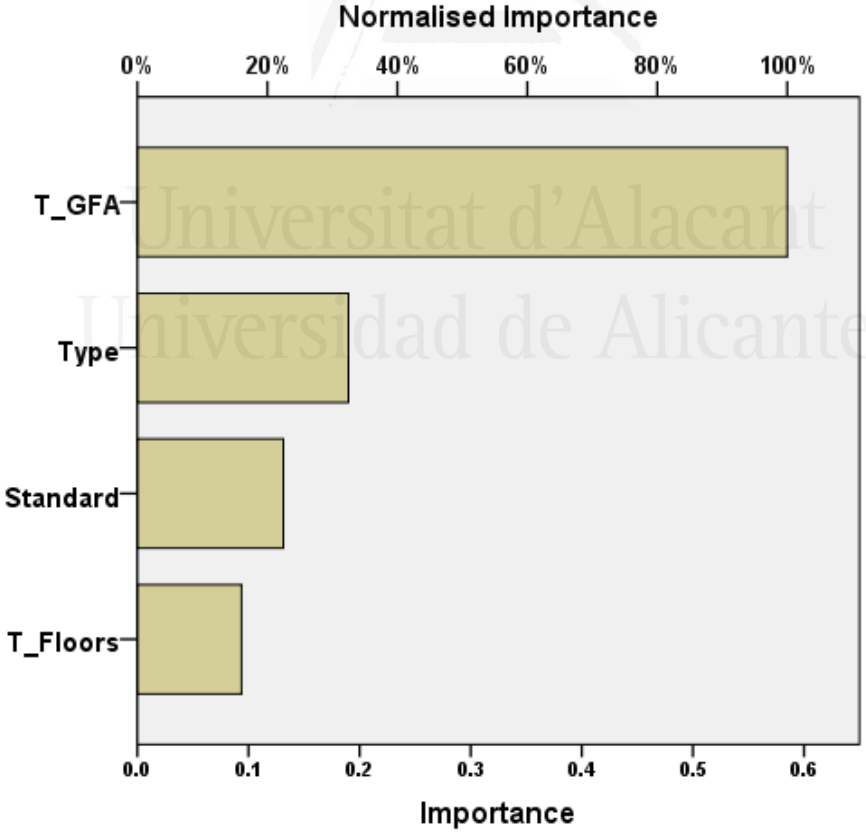


Figure V-3. Importance of each of the input variables used in the best MLP model.

V.3 FEM-Based Numerical Models

V.3.1 Selection of the Best Set of Input Variables

The first step in search of models with better goodness of fit and predictive ability was to generate families of numerical models by using four different complexities (*C*): 10, 20, 50 and 100. Furthermore, each family was generated using both the natural and the logarithmic form of the variables. As a result, eight models were built for each of the three sets of variables selected after the development of linear regression analysis. The main data of predictive performance obtained by the three families of numerical models are listed in Table V-12. In the case of the models with transformed variables, the performance data have been calculated after the estimates were transformed back to its natural scale. It should be noted that the numerical methodology does not generate a direct mathematical equation but instead creates reference values from which you can build the function value at any point.

In order to interpret more clearly the performance data obtained with each of the families of developed numerical models, the accuracy values (RMSE) have been represented in two graphs incorporated in Figures V-4 and V-6, in the case of the calibration projects, and Figures V-5 and V-7, in the case of the validation data. If we observe the RMSE values obtained with the calibration data set, in general, the best accuracy results in both cases (transformed and untransformed variables) are achieved by using higher complexity. That is, the greater the complexity, better values of goodness of fit and accuracy are obtained by the models generated, although the increase in accuracy is much higher in the case of using untransformed variables.

However, if we analyse the results in the case of the validation projects the behaviour of the models is different depending on whether transformation of variables is carried out or not. On the one hand, the accuracy in the families of logarithmic models (Figure V-5) follows a reverse trend when using the validation projects, since using a higher complexity we obtain lower accuracy. On the other hand, the accuracy in the families of models with untransformed variables (Figure V-7) reaches an inflection point where the trend is reversed. From a certain complexity an increase of it does not translate into an improvement of the ability to generalise when we apply the models to the validation projects.

Set of input variables	Form	*C*	Calibration		Validation
			Adjusted R^2	RMSE	RMSE
Base	Log	10	0.865	195.362	17.934
		20	0.940	139.482	26.665
		50	0.988	103.504	66.660
		100	0.996	93.898	86.296
	Natural	10	0.760	177.572	52.667
		20	0.861	135.222	27.216
		50	0.970	63.120	13.948
		100	0.995	25.931	27.981
Improved_2	Log	10	0.815	155.562	19.058
		20	0.744	182.781	26.072
		50	0.780	169.375	67.749
		100	0.787	166.582	86.326
	Natural	10	0.838	145.579	60.623
		20	0.910	108.454	32.854
		50	0.972	60.542	16.020
		100	0.993	30.127	45.616
Improved_3	Log	10	0.813	156.773	18.150
		20	0.852	139.236	21.284
		50	0.869	131.384	28.387
		100	0.891	119.575	38.371
	Natural	10	0.808	158.645	91.167
		20	0.897	116.302	53.550
		50	0.944	85.954	24.793
		100	0.976	56.595	31.573

Table V-12. Main performance data obtained with families of numerical models.

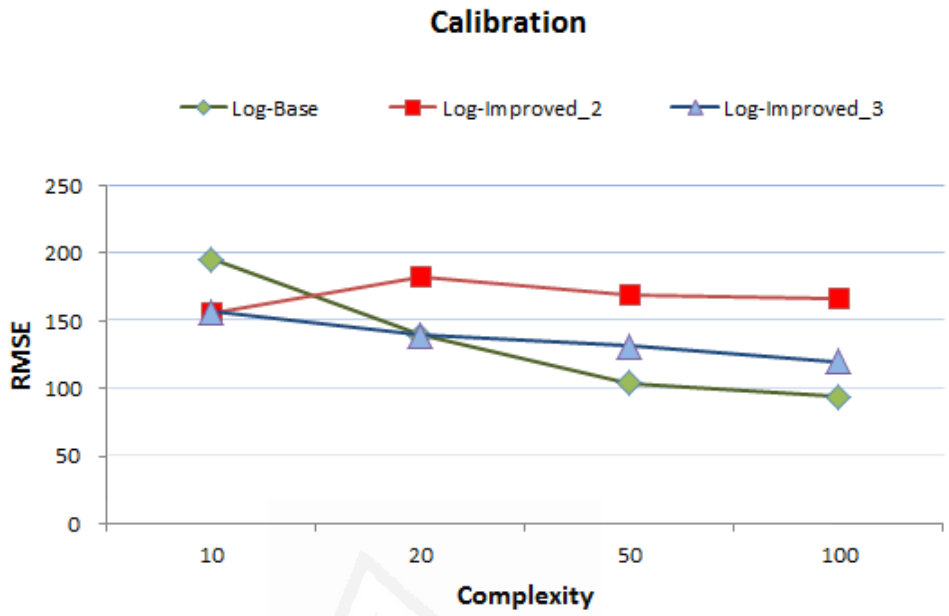


Figure V-4. RMSE values obtained with the set of calibration projects by developing numerical models with logarithmic transformation of variables.

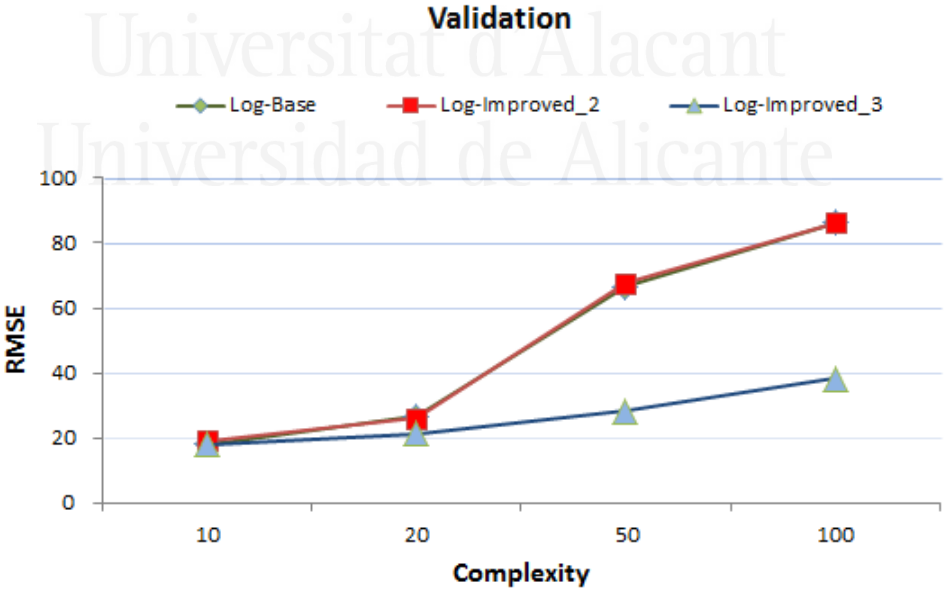


Figure V-5. RMSE values obtained with the set of validation projects by developing numerical models with logarithmic transformation of variables.

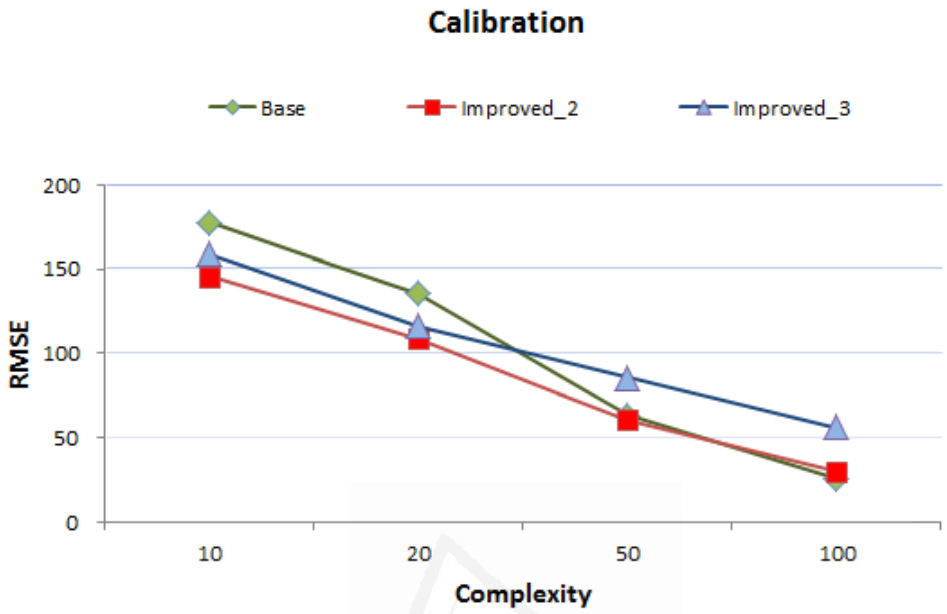


Figure V-6. RMSE values obtained with the set of calibration projects by developing numerical models without logarithmic transformation of variables.

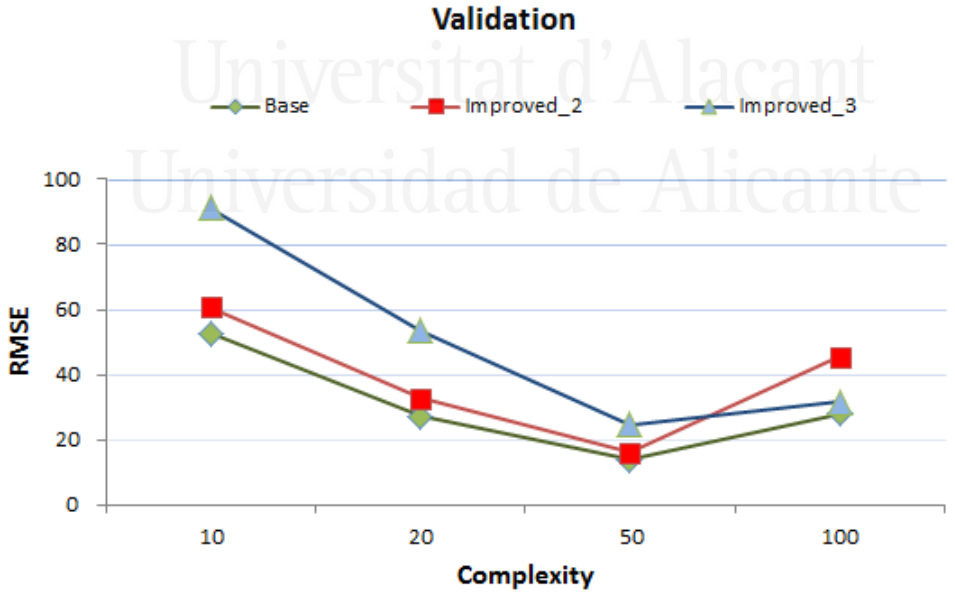


Figure V-7. RMSE values obtained with the set of validation projects by developing numerical models without logarithmic transformation of variables.

According to the criteria set in this thesis, in order to evaluate the performance of the generated models, the process to select the best numerical model requires finding the best possible trade-off between the results obtained with the calibration data set and the results achieved by applying the models on the validation projects. As can be inferred from the graphs presented above, initially, the best equilibrium with the best ability to generalise is achieved by using the family of numerical models which uses the input variables in their natural form and a complexity of 50. Furthermore, in this particular point of complexity, the sets of variables named as *Base* and *Imprioved_2* outperform the accuracy offered by the set of variables referred to as *Improved_3*.

Set of input variables	**Form**	***C***	**Calibration**		**Validation**
			Adjusted R^2	**RMSE**	**RMSE**
Base	Natural	30	0.913	106.975	16.567
		35	0.933	93.751	13.692
		40	0.949	81.433	12.410
		45	0.961	71.807	12.745
		50	0.970	63.120	13.948
		55	0.976	56.015	15.529
		60	0.981	49.744	17.517
		65	0.985	44.893	19.870
		70	0.988	40.372	21.289
Improved_2	Natural	30	0.937	90.308	21.107
		35	0.949	81.669	17.475
		40	0.958	74.371	14.821
		45	0.966	66.648	14.544
		50	0.972	60.542	16.020
		55	0.977	55.040	18.165
		60	0.981	49.931	21.479
		65	0.984	45.775	25.375
		70	0.987	41.891	28.615

Table V-13. Main performance data obtained with families of numerical models.

Consequently, for the purpose of increasing the predictive performance of the numerical models, new families were generated by employing smaller intervals of complexity around 50 with the sets of variables named as *Base* and *Improved_2* (without log transformation). Table V-13 and Figure V-8 show the predictive performance obtained with each of the new families of models generated with complexities between 30 and 70.

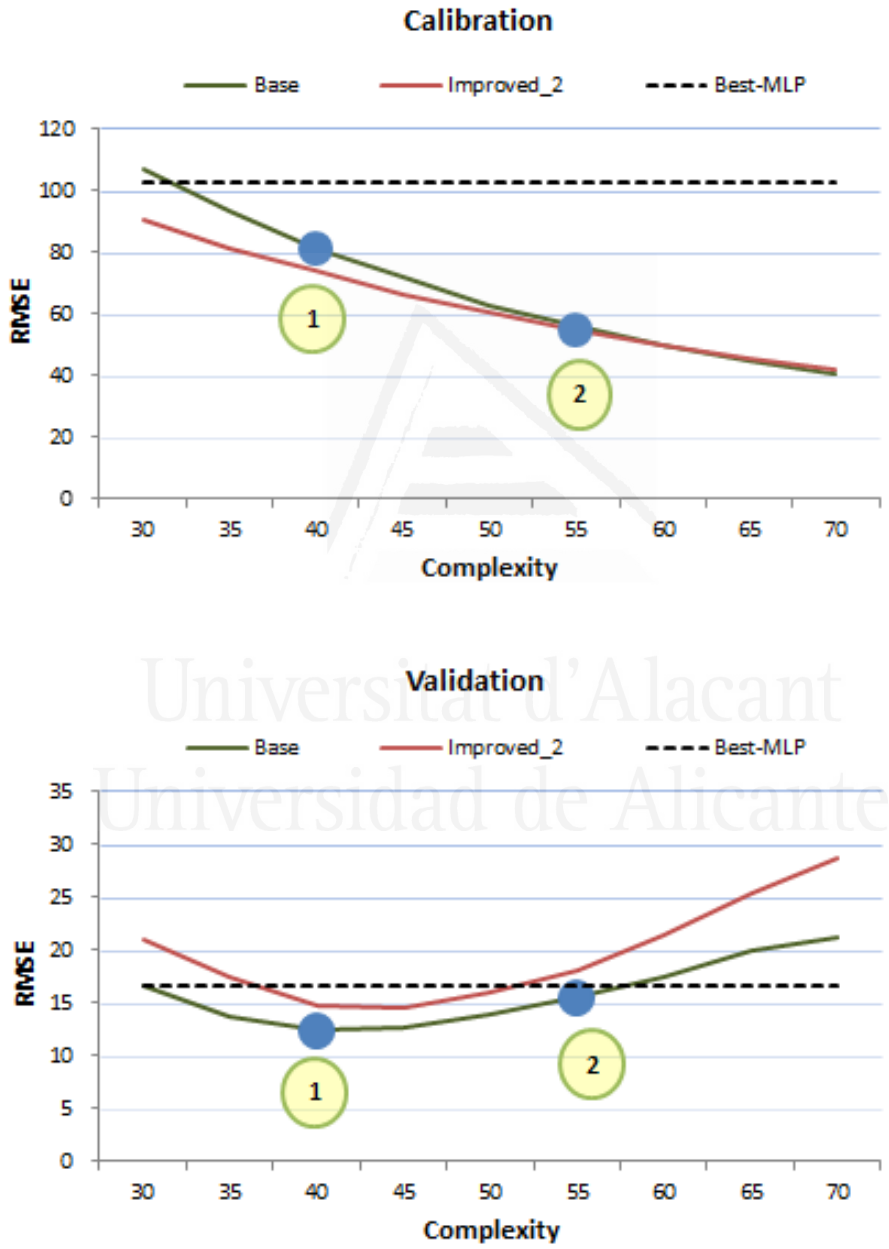


Figure V-8. Best points of balance within the family of selected numerical models.

Although the *Improved_2* set achieves greater accuracy with the calibration data when it comes to complexities below 60, nevertheless, the *Base* set demonstrates a greater ability to generalise when it is applied to the group of validation projects, regardless of the selected complexity. Therefore, this set of variables was chosen to generate the best possible numerical model.

V.3.2 Selection of the Best FEM-Based Numerical Model

Once the set of input variables (*Base*) that produces the best predictive performance has been identified, the next step is to select the best model generated from the use of different complexities. For this purpose, considering the performance obtained with the best ANN model as a benchmark, two criteria were established for selecting the best numerical models:

1. Numerical model that obtains the minimum accuracy error in the validation projects, provided that this model improves the accuracy obtained with the best MLP model in the calibration data set.
2. Numerical model that obtains the minimum accuracy error in the calibration data set, provided that the model improves the ability to generalise obtained with the best MLP model in the validation data set.

Having reached this point, it is necessary to remember that with the implementation of the proposed numerical methodology we obtain families of numerical models using different complexities in the geometric model. The maximum complexity that can be used depends on the number of input variables and the size of the calibration sample used. Unlike ANNs, the goodness of the generated predictive model does not depend on the randomness of certain parameters, such as the selection of the set of initial weights in the case of MLP networks. Therefore, the criterion of minimum error in the validation set cannot be understood as a stop criterion in the calibration process, as occurs with MLP networks when the early stopping method is used. Thus, it is no necessary to divide the calibration data set when using the FEM-based numerical methodology.

As we can see in Figure V-8, the fulfilment of the set criteria led to the selection of two models with complexities 40 (*FEM-1*) and 55 (*FEM-2*), whose predictive performance values are shown in Table V-14. Although the *FEM-1 model* obtains the best ability to generalise (RMSE=12.410), nevertheless the *FEM-2* model got the

highest value of goodness of fit (R^2=0.976), the lowest accuracy with the calibration data (RMSE=56.015), with greater percentage reduction when comparing with the best MLP model (45.45%), and the lowest performance index (PI=35.772),

Set of input variables	Form	Model	*C*	Calibration		Validation	PI	Stability (% var.)
				Adj. R^2	RMSE	RMSE		
Base	Natural	*FEM-1*	40	0.949	81.433	12.410	46.922	1.84
		FEM-2	55	0.976	56.015	15.529	35.772	2.28

Table V-14. Predictive performance of the best FEM-based numerical models.

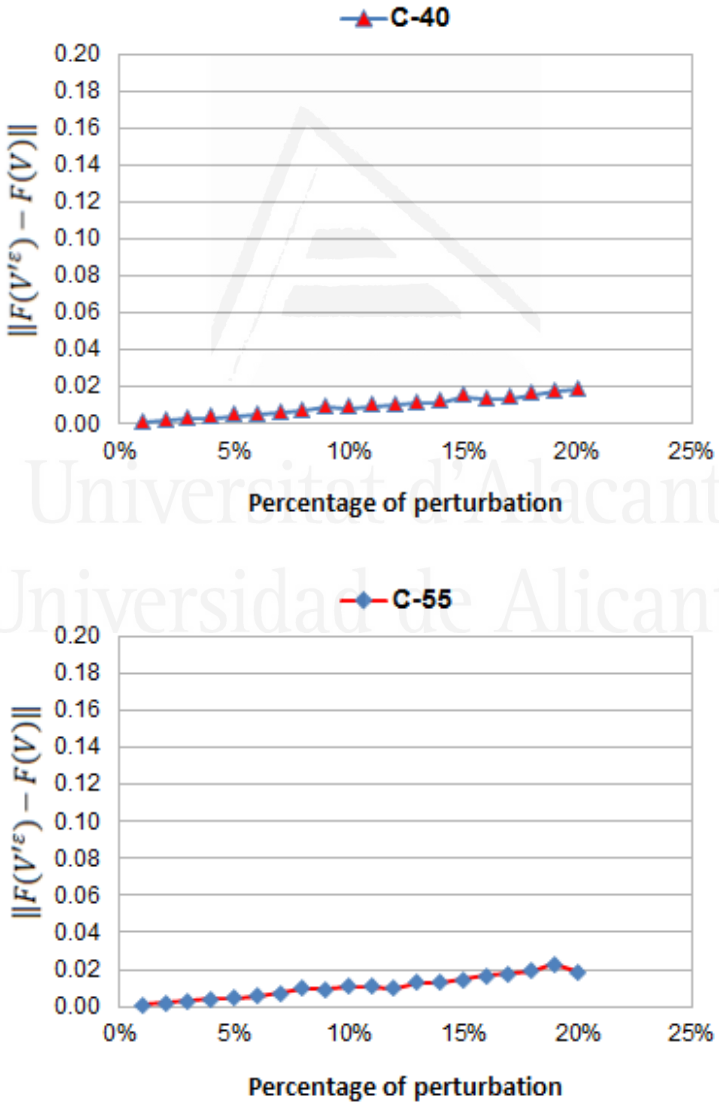


Figure V-9. Graphs showing the stability analysis applied to the best FEM-based numerical models.

Regarding the proposed stability analysis, Figure V-9 shows the obtained results by graphs representing the variations in the estimated values versus the used perturbation percentage. Random perturbations of up to 20% in the original data generated a variation in the models that did not exceed 1.84% (*FEM-1*) and 2.28% (*FEM-2*), which means the selected models are little sensitive to variations introduced in the initial observations and, similarly to the selected MLP network, it can be said that both models are stable.

Lastly, taking into consideration the overall predictive performance, the *FEM-2* model was chosen as the best FEM-based numerical model.

V.3.3 Model Sensitivity to Cost Variability

Similar to what was carried out with the best ANN model, a nonparametric test was also conducted to study how the variability of the *T_Cost* variable affects the predictive performance of the best FEM-based numerical model. The results of the MW *U* test showed that the accuracy differences were not statistically significant at the 5% level (see Table V-15). However, it should be noted that the percentage increase in the RMSE value is notably higher in the best FEM-based numerical model (33.23%) than in both the best MLRA model (11.89%) and the best ANN model (21.87%).

Price Index	Update date	Calibration		Validation		MW *U* test (*p*-value)
		Adjusted R^2	RMSE	RMSE Estimated cost	RMSE Actual cost	
IRF	2010	0.976	102.678	20.689	15.529	0.097

Table V-15. Results of the sensitivity analysis conducted with the best FEM-based numerical model.

V.4 Comparative Analysis of the Best Predictive Models

The main results of predictive performance for the three selected models are provided in Table V-16. It is clear that the best FEM-based numerical model (*Best_FEM*) achieves better results of goodness of fit (R^2=0.976), performance index (PI=35.772) and, to a lesser extent, ability to generalise (RMSE=15.529) than both the best linear model (*Best_MLRA*) and the best ANN model (*Best_ANN*). What is more,

the selected numerical model uses less number of predictor variables. On the contrary, in the case of linear regression and MLP networks both techniques used the same number of predictive variables, although the best MLP model obtains higher values of goodness of fit and ability to generalise.

Model	Set of input variables	Form	Calibration		Validation	PI
			Adj. R^2	RMSE	RMSE	
Best_MLRA	*Improved_2*	Log	0.884	122.508	19.250	70.879
Best_ANN	*Improved_2*	Natural	0.921	102.678	16.551	59.615
Best_FEM	*Base*	Natural	0.976	56.015	15.529	35.772

Table V-16. Overall performance of the best selected models.

The results of predictive accuracy (calibration + validation), measured in terms of absolute error, were also compared with each other using the KW test. It was necessary to develop this nonparametric test as the assumptions of normality and equality of variances are not met. The result of this test showed that the difference among the predictive accuracy of the three selected models was statistically significant (p=0.000). In addition, pairwise comparisons were carried out in order to know which pairs of models differ significantly. In post hoc analysis, one wishes to control the experiment-wise Type 1 error, i.e., the probability of rejecting at least one pair hypothesis given all pairwise hypotheses are true. For this purpose, the SPSS software calculates adjusted p-values through the expression $p_{adj} = pK(K-1)/2$, where K represents the number of groups being compared.

Sample 1	Sample 2	Test statistic	Std. error	Std. test statistic	Sig.	Adj. Sig.
Best_FEM	*Best-MLRA*	53.805	16.673	3.227	0.001	0.004
Best_FEM	*Best-MLP*	83.951	16.673	5.035	0.000	0.000
Best_MLRA	*Best-MLP*	-30.146	16.673	-1.808	0.071	0.212

Null hypothesis = Sample 1 and Sample 2 are the same

Table V-17. Results of pairwise comparisons.

The results of the pairwise comparisons are collected in Table V-17. The evidence indicates that the average prediction error of the best FEM-based numerical model is significantly different from the average prediction error offered by both the best MLRA model and the best MLP network. Consequently, from the analysis of these results it can be concluded that the selected FEM-based numerical model performed more effectively than the other two models in estimating the construction speed of new builds. On the contrary, the difference between the predictive accuracy of the best linear regression model and the best MLP network was not statistically significant, and it is not possible to assert that the best ANN model is superior to the best linear regression model.

According to the above results, there is reasonable evidence to support the third hypothesis (H-03) proposed in this thesis. That is, "*considering the same set of project scope factors as predictor variables, nonlinear modelling techniques can generate models to estimate the construction speed of new builds with better predictive performance than that offered by linear regression models*".

V.5 Conclusion

This thesis addressed the general assumption that nonlinear modelling techniques are likely to better represent the relationships existing between important project scope factors and the construction speed of new builds. For testing this hypothesis, on the basis of the best predictive models developed using the MLRA modelling technique, in this chapter new predictive models were developed through the use of three sets of input variables (*Base*, *Improved_2* and *Improved_3*) and two nonlinear modelling techniques, ANNs and a new numerical methodology based on FEM.

The development of the MLP models presented in this chapter pointed out significant differences on their performance, depending on which proportion of data division was used for optimising the network structures. The 80-20 and 50-50 data divisions achieved the best performance concordance between training and test data sets. This agreement between the results obtained with the calibration data sets resulted, in turn, in an improvement in the ability to generalise (validation data) of the MLP models. The best predictive models were obtained by using the set of variables named as *Improved_2* in their natural form and the 80-20 data division. Despite the problems identified in the literature regarding the effectiveness of the traditional GD

algorithm and its dependency on training parameters, the experimentation carried out here appear to suggest that this algorithm is capable of building better predictive models than the SCG algorithm. In the same line, the predictive performance offered by RBF networks was worse than that achieved with the best MLP network and the GD training algorithm.

After the analysis of the optimisation process of MLP networks, it was found that the number of hidden neurons and the set of selected initial synaptic weights were the parameters that most influenced the accuracy of MLP networks, while the combination of transfer functions and scaling of variables had a lesser impact on the performance of neural networks. However, it seems that the impact of the initial weights on the model performance might be reduced by using an optimised network structure. Moreover, contrary to what has been reported in many other studies, in this research work the training parameters of the GD algorithm had virtually no impact on the optimisation process of the network structures.

A sensitivity analysis was also performed for the purpose of calculating the importance of each predictor variable in determining the construction speed by using the best MLP model. The results of this analysis showed that GFA has the greatest influence on construction speed, supporting the second hypothesis proposed in this thesis.

Families of FEM-based numerical models were also built using four different complexities, and each family was generated using both the natural and the logarithmic form of the predictor variables. The best predictive performance was achieved by using the input variables in their natural form. In particular, contrary to what happened with ANN models, the families of models generated with the set of input variables named as *Base* outperformed the predictive performance provided by the families of models which were generated making use of the sets of input variables referred to as *Improved_2* and *Improved_3*. That is, the FEM-based numerical methodology does not make use of the variable representing the type of facility to achieve the best predictive performance, which could be interpreted as a clear proof that the numerical methodology based on FEM works more effectively than MLRA and ANNs. The stability of the best numerical models was also verified positively.

Also noteworthy is the fact that the selected nonlinear models are more sensitive than linear regression models to the variability of construction costs. This could be due

to the fact that the standard variable has more influence on construction speed when the relationships between both variables are modelled nonlinearly. However, this variability must be taken into account in the practical application of this type of predictive models.

Finally, a comparative study of the predictive performance of the three best models developed with each of the modelling methodologies used in this thesis was made. The results of pairwise comparisons showed that the average prediction error of the best FEM-based numerical model was significantly different from the average prediction error offered by both the best MLRA model and the best MLP network. According to these results, there is reasonable evidence to support the third hypothesis proposed in this thesis. On the contrary, the accuracy differences between the best linear regression model and the best MLP network were not statistically significant. However, it must be pointed out that due to the large number of different configurations that can be used to develop MLP networks, and the lack of fixed rules to determine the best network configuration, there can be no assurance that the selected MLP network was the best possible. In this connection, the main modelling advantage of the FEM-based numerical methodology with respect to MLP networks is that, with the exception of choosing the complexity of the geometric model, it does not depends on any user parameter for generating good predictive models, as it happens when using MLP networks.

Chapter VI

General Conclusions, Study Limitations and Future Work

VI. General Conclusions, Study Limitations and Future Work

VI.1 General Conclusions

VI.1.1 Verification of Research Hypotheses

The prediction of the construction time at early stages of the development of building projects, when only little information is available for project managers, has been considered a key element for project success. Most of the literature has identified the project scope factors as key predictors of construction time, but there are contradictory conclusions about what factors have the greatest influence. Although construction costs and GFA are the most commonly factors used to define the project scope, there are no perfect measurement units. On the other hand, previous research has shown that the concept of construction speed is a useful benchmark to compare the contractor performance and can be used as an alternative output variable to estimate construction time. In this connection, there is a debate which revolves around whether it is possible to obtain better forecasting models using the construction time as the dependent variable or if it is more appropriate to consider the construction speed.

In order to provide proper tools for estimating construction time and minimise the subjectivity in such estimation, to date, most research works have presented parametric models which were built using LRA. However, although these parametric models have shown a good balance between the difficulty of developing estimation models and the predictive accuracy obtained with them, they might represent a great simplification of the complex relationships which control the development of building projects. Thus, the project scope factors that influence on construction speed might not be fully associated in a linear manner and nonlinear modelling techniques are likely to better represent the relationships existing between these factors and construction speed.

Bearing in mind the above, three principal hypotheses were postulated in this thesis:

(H-01) "C*onstruction speed is a more appropriate dependent variable than construction time in order to generate predictive models for estimating the duration of the construction process of new builds.*"

(H-02) "*GFA has greater influence on the construction speed of new builds than construction costs.*"

(H-03) "*Considering the same set of project scope factors as predictor variables, nonlinear modelling techniques can generate models to estimate the construction speed of new builds with better predictive performance than that offered by linear regression models.*"

In order to test these hypotheses, linear and nonlinear predictive models were developed in this thesis by using several project scope factors as predictor variables to estimate the construction speed of new builds developed in Spain. The sets of variables that achieved the best predictive performance using MLRA were also used with ANNs and a novel FEM-based numerical methodology in order to develop nonlinear models with better predictive performance. It is worth stressing that the models generated in this thesis have not been presented to be used in a practical way, but rather to analyse the relationships between the variables under study, as well as to identify which modelling methodology yields better prediction models.

The first two hypotheses were tested positively in Chapter IV. Thus, it can be concluded that, on the one hand, construction speed is the most appropriate dependent variable to develop predictive models for estimating the duration of the construction process of new builds, and, on the other hand, this construction speed is affected more by GFA than by construction costs. Therefore, GFA is the best indicator of project scope and the best predictor of the average construction speed of new builds.

Chapter V provides clear evidence to support the third hypothesis, because using the numerical methodology based on FEM it is possible to develop nonlinear models with better predictive performance than that offered by linear regression models.

VI.1.2 Main Contributions

According to the research methodology and results presented in preceding chapters, the main contributions of this thesis are:

- ***Development of a linear regression model which provides a clear understanding of the relationships between some project scope factors and the time required to complete the construction process of new builds developed in Spain***. In line with the factors most often used in the literature, the best linear regression model developed in this thesis was a logarithmic model which made use of a categorical variable representing the type of facility along with three quantitative variables corresponding to the following project scope factors: GFA, the cost/GFA ratio (*Standard*) and total number of floors.

  According to the different linear models presented herein, it can be inferred that both GFA and construction costs are project scope factors of great importance to predict the construction time of building projects. However, although individually GFA has a greater influence on construction speed than construction costs by combining both factors in the *Standard* variable it is possible to develop models with better predictive performance. This finding suggests the idea that both variables define different relationships with construction speed but complementary. Projects with a large GFA enable more efficient use of production factors and, therefore, greater construction speed. However, the construction speed of building projects is negatively affected when the cost/GFA ratio increases, indicating that a higher construction quality (*Standard*) will result in greater project complexity and lower construction speed. The study results also seem to indicate that, within the limit values established by the features of the data used in the study, the building works are developed horizontally at a higher speed than vertically. The incorporation into the model of a variable representing the type of facility revealed that the construction speed follows a different pattern in some specific kinds of building projects.
- ***Development of an optimisation methodology which allows obtaining configurations of MLP networks with an appropriate predictive performance***. The proposed optimisation methodology follows a stepwise trial and error procedure in which from an initial network structure with enough consistency six optimised MLP models are generated. The set of variables that generated the best

linear regression model was also able to develop the best MLP model. However, the latter model used the variables in its natural form, while the linear regression model had to develop a logarithmic transformation of the variables to fulfil the underlying assumptions of LRA.

In spite of the problems identified in the literature regarding the effectiveness of the traditional GD algorithm and its dependency on training parameters, the results of this thesis seem to suggest that this algorithm is able to generate better predictive models than the SCG training algorithm and RBF networks. Nevertheless, it is necessary to point out that due to the fact that the SCG algorithm does not contain any user-dependent crucial parameter for development of the training process, by making use of this algorithm the time required to apply the optimisation process proposed for MLP networks is reduced considerably.

- ***Evaluation of the influence of calibration data division on the predictive performance of MLP networks.*** The experiments carried out with MLP networks showed that the type of data division used for calibrating this kind of networks has a direct impact on their predictive accuracy. In particular, the best predictive models were obtained by using the 80-20 data division for calibration data and this result is consistent with the suggestions made by other authors (see, e.g., Haykin, 1999).
- ***Evaluation of the influence of network structure, training parameters and the set of initial weights on the predictive performance of MLP networks.*** This research work found that the number of hidden neurons and the set of selected initial weights were the parameters that most influenced the predictive accuracy of MLP networks. Nevertheless, it seems that the impact of the selected initial weights on the model predictive performance might be reduced by using an optimised network structure. In addition, contrary to what has been reported in many other studies, in this research work the training parameters of the GD algorithm had virtually no impact on the accuracy of the generated MLP models.
- ***Application of the FEM-based numerical methodology to develop predictive models related to the construction time of new builds.*** This numerical methodology is a novel modelling technique which has not been applied previously in the literature for the purpose of estimating the construction speed. As with the best ANN model, the best numerical models were obtained using the variables in its natural form. However, in this case, the set of variables with the best predictive

performance did not contain the categorical variable representing the type of facility, which could be interpreted as clear proof that the numerical methodology based on FEM works more effectively than MLRA and ANNs.

- ***Development of a stability analysis to validate predictive models.*** Traditionally, goodness-of-fit and accuracy values have been used in previous research to validate both linear and nonlinear forecasting models. However, this thesis also proposed the development of a stability analysis as an additional element of model validation.

  For LRA models, a univariate sensitivity analysis was developed by examining how well model predictions are in agreement with the known underlying physical processes of the problem under study. The results of this analysis showed as a linear model with a high degree of goodness of fit and accuracy might have an anomalous behaviour, which makes it unusable for its application in actual practice.

  In the case of FEM-based numerical models and ANNs, a multivariate analysis was carried out to test the model sensitivity to small variations introduced in the experimental data. The results of this analysis showed that the best nonlinear models were stable against random perturbations conducted on calibration data.

- ***Development of a sensitivity analysis to evaluate how construction cost variability can affect the predictive performance of models***. Construction costs tend to be commercially sensitive and the final cost normally varies from the initial cost estimated before starting the works. Although in this thesis the use of estimated costs versus the use of actual costs within the *Standard* variable, on average, resulted in worse accuracy values, the variation in accuracy was not statistically significant. Furthermore, attempts to disregard the cost as a predictive variable provided us with models with worse fit and accuracy. It is also necessary to point out that the selected nonlinear models were more sensitive to the cost variability than the best linear regression model, so that such variability must be taken into account in the practical application of this type of predictive models.

- **Development of a comparative analysis between three different modelling techniques used to estimate the construction speed of new builds**. This thesis proved that by using the FEM-based numerical methodology is possible to obtain families of nonlinear models that allow us to predict the construction speed of building projects with greater accuracy than models created with MLRA and

ANNs. Moreover, the results of the statistical analysis carried out in Chapter V showed that the average prediction error of the best FEM-based numerical model was significantly different from the average prediction error offered by both the best MLRA model and the best MLP network. On the contrary, the accuracy differences between the best linear regression model and the best MLP model were not statistically significant.

However, it must be pointed out that due to the large number of different configurations that may be used to develop MLP networks by using the SBP training algorithm, and the lack of fixed rules to determine the best network configuration, there can be no assurance that the selected MLP network was the best possible. In this regard, the main modelling advantage of the FEM-based numerical methodology with respect to ANNs is that, with the exception of choosing the complexity of the geometric model, it does not depends on any user parameter for generating good predictive models, as it happens when using MLP networks.

It is also necessary to remark that linear regression models allow a clear interpretation of how the predictor variables work in the models. In contrast, both ANNs and the FEM-based numerical methodology are black-box techniques and it is not possible to explain in a clear way the influence of the input variables on the output variable. Therefore, if the goal of modelling is to explain the underlying processes that control the relationships between the variables under study, it would be better to use LRA.

Lastly, it is important to note that the findings of this thesis support the idea that linear regression models can provide a good starting point from which to search for better nonlinear predictive models. In this regard, the three modelling techniques analysed in this research work can be understood as complementary methods for generating the best possible predictive model rather than as competing modelling techniques. The research methodology developed herein provides a new framework to generate and evaluate linear and nonlinear predictive models, broadening the tools available to produce useful forecasting models for practitioners and researchers of the building sector. However, although the knowledge gained from this research will allow for new approaches to be explored in order to better determine the relationships between project scope factors and construction speed, it is necessary to understand that the results provided by the type of models analysed in this thesis are only initial

estimates at early stages of project development and are not intended to replace detailed schedules undertaken by builders and owners.

VI.2 Study Limitations

First of all, the research work presented in this thesis examined only some factors that may influence the construction time of new builds. Previous studies have demonstrated that there are other qualitative factors affecting construction time, such as the management effectiveness or labour productivity, which could not be analysed in this study due to the lack of available data.

Second, it should be noted that the projects used in this research have been executed in a wide range of time and represent different types of facilities. It is therefore logic to think that different conclusions could be obtained if new models were generated with projects developed in a shorter time interval or belonging to a very specific type of facility. Moreover, the study made use of a limited number of projects and it is always possible that there was incorrect information in the samples of projects under study, in which case they may have caused distortions in the coefficients and parameters of the models and in its forecasting ability. Therefore, the models presented in this study should be analysed with caution, as more research is needed to evaluate its validity on new data sets and verify that the results remain stable.

On the other hand, it is also worth mentioning that uncertainty is a feature inherent to the construction industry. Thus, the actual duration of the construction phase may be influenced by multiple risks and uncertainties arising from various sources related to the features of the environment in which a building project is developed, and they cannot be collected within deterministic models of prediction such as those proposed in this research. It is also necessary to note that before starting the works, and during its execution, a contractor can choose different types of worker teams, number of shifts, type of machinery, and construction methods to complete project activities. These decisions ultimately affect construction time, regardless of the type of facility, GFA, construction costs, or number of floors of the building under study.

Another issue to consider is the possible variation of productivity in the building sector over time. The construction time of a building also varies depending on the moment when the project is developed. Consequently, the coefficients and parameters

of predictive models must be reviewed periodically to keep them updated according to the evolution of construction techniques used in the building sector. But even after being aware of the accuracy errors which are likely to arise when using the type of models which have been developed in this thesis, its use may be helpful before starting the works once the project scope factors involved in the model are known. The proportionate forecast at this stage can serve as a control parameter to verify if construction deadlines proposed by any of the agents involved in the construction phase are realistic and, if necessary, take special measures to increase the construction speed.

VI.3 Future Work

The author of this thesis has the intention of conducting further research related to the estimation of construction speed using new data sets and also by applying other modelling methodologies such as “super vector machine” or “deep learning”. On the other hand, as it was identified in the state of the art of this thesis, the literature established that the complexity of the building design is also related to the construction time. Moreover, construction speed has also been related to the way in which the builder plans the movement of resources on the building site and the selection of appropriate construction methods, which, in turn, are affected by the complexity of the building design. In this regard, there should be more specific factors related to the complexity of the form of buildings and the way in which the builder organises the construction process to overcome this complexity that could be incorporated into predictive models for increasing their accuracy. The study of such factors and their relationships with the construction speed of new builds will also be the focus of our future research work.

However, the unavailability of data from completed real projects limits the possibility of improving the type of predictive models developed in this thesis. Moreover, often, when data concerning the characteristics of buildings and their construction process is provided, it is incomplete or contains inaccurate information. In the absence of real data containing adequate information on a large number of variables that can influence the construction speed, a logic step to improve the reliability of predictive models based on mathematical equations might be the development of these models from the information generated by simulating the

construction process. Therefore, another area of future research could be to develop predictive models based on data generated by simulating the construction process, which could be compared to the results obtained with predictive models developed by using actual data.

VI.4 Relevant Publications

- Guerrero, M. A., Villacampa, Y., & Montoyo, A. (2014). Modeling construction time in Spanish building projects*. International Journal of Project Management, 32*(5), 861–873.
- Guerrero, M. A., Villacampa, Y., Navarro-González, F. J., & Montoyo, A. (2015). Modelling building construction speed by using n-dimensional finite elements, neural networks and linear regression analysis. *Automation in Construction*. It has been determined to have a potential for possible publication. Under 2nd review*.*
- Guerrero, M. A., Villacampa, Y., & Montoyo, A. (2015). Performance comparison of artificial neural networks in modeling the construction speed of building projects. *Journal of Construction Engineering and Management.* Submitted.

References

Adeli, H., & Wu, M. (1998). Regularization neural network for construction cost estimation. *Journal of Construction Engineering and Management*, *124*(1), 18–24.

Al-Khalil, M. I., & Al-Ghafly, M. A. (1999). Delay in public utility projects in Saudi Arabia. *International Journal of Project Management*, *17*(2), 101–106.

Al-Tabtabai, H., Kartam, N., Flood, I., & Alex, A. P. (1997). Expert judgment in forecasting construction project completion. *Engineering, Construction and Architectural Management*, *4*(4), 271–293.

Amari, S-I., Murata, N., Muller, K-R., Finke, M., & Yang, H. H. (1997). Asymptotic statistical theory of overtraining and cross-validation. *Neural Networks, IEEE Transactions on*, *8*(5), 985–996.

Ameyaw, C., Mensah, S., & Arthur, Y. D. (2012). Applicability of Bromilow's time–cost model on building projects in Ghana. In *4th West Africa Built Environment Research Conference*, 24-26 July 2012, Abuja, Nigeria.

Arditi, D., Tokdemir, O. B., & Suh, K. (2001). Effect of learning on line-of-balance scheduling. *International Journal of Project Management*, *19*(5), 265–277.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Attalla, M., & Hegazy, T. (2003). Predicting cost deviation in reconstruction projects: artificial neural networks versus regression. *Journal of Construction Engineering and Management*, *129*(4), 405–411.

Baccarini, D. (1996). The concept of project complexity - A review. *International Journal of Project Management*, *14*(4), 201–204.

Bates, D., & Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications* (2nd ed.). New York, USA: John Wiley & Sons.

Battiti, R. (1992). First and second-order methods for learning: between steepest descent and Newton's method. *Neural Computation*, *4*, 141–166.

Bhokha, S., & Ogunlana, S. O. (1999). Application of artificial neural network to forecast construction duration of buildings at the predesign stage. *Engineering, Construction and Architectural Management*, *6*(2), 133–144.

Bordoli, D. W., & Baldwin, A. N. (1998). A methodology for assessing construction project delays. *Construction Management and Economics*, *16*(3), 327–337.

Boussabaine, A. H. (1996). The use of artificial neural networks in construction management: a review. *Construction Management and Economics*, *14*(5), 427–436.

Boussabaine, A. H. (2001a). Neurofuzzy modelling of construction projects' duration I: principles. *Engineering, Construction and Architectural Management*, *8*(2), 104–113.

Boussabaine, A. H. (2001b). Neurofuzzy modelling of construction projects' duration II: application. *Engineering, Construction and Architectural Management*, *8*(2), 114–129.

Bromilow, F. J. (1969). Contract time performance expectations and the reality. *Building Forum*, *1*(3), 70–80.

Bromilow, F. J., Hinds, M. F., & Moody, N. F. (1980). AIQS survey of building contract time performance. *Building Economist*, *19*(2), 79–82.

Broomhead, D. S., & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems*, *2*(3), 321–355.

Castellano, M. (2009). *Modelización estadística con redes neuronales. Aplicaciones a la hidrología, aerobiología y modelización de procesos*. (Phd thesis) Universidad da Coruña, Spain.

Castillo, E., Guijarro-Berdiñas, B., Fontenla-Romero, O., & Alonso-Betanzos, A. (2006). A very fast learning method for neural networks based on sensitivity analysis. *Journal of Machine Learning Research*, *7*, 1159–1182.

Černý, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, *45*(1), 41–51.

Chan, A. P. C. (1996). *Determinants of project success in the construction industry of Hong Kong*. (PhD thesis) University of South Australia, Australia.

Chan, A. P. C. (1999). Modelling building durations in Hong Kong. *Construction Management and Economics*, *17*(2), 189–196.

Chan, A. P. C., & Chan, D. W. M. (2004). Developing a benchmark model for project construction time performance in Hong Kong. *Building and Environment*, *39*(3), 339–349.

Chan, D. W. M. (1998). *Modelling construction durations for public housing projects in Hong Kong*. (PhD thesis) The University of Hong Kong, Hong Kong.

Chan, D. W. M., & Kumaraswamy, M. M. (1995). A study of the factors affecting construction durations in Hong Kong. *Construction Management and Economics*, *13*(4), 319–333.

Chan, D. W. M., & Kumaraswamy, M. M. (1997). A comparative study of causes of time overruns in Hong Kong construction projects. *International Journal of Project Management*, *15*(1), 55–63.

Chan, D. W. M., & Kumaraswamy, M. M. (1999). Forecasting construction durations for public housing projects: a Hong Kong perspective. *Building and Environment*, *34*(5), 633–646.

Chan, D. W. M., & Kumaraswamy, M. M. (2002). Compressing construction durations: lessons learned from Hong Kong building projects. *International Journal of Project Management*, *20*(1), 23–35.

Chao, L., & Skibniewski, M. (1994). Estimating construction productivity: neural network-based approach. *Journal of Computing in Civil Engineering*, *8*(2), 234–251.

Chatterjee, S., & Hadi, A. S. (2006). *Regression Analysis by Example* (4th ed.). Hoboken, New Jersey, USA: John Wiley & Sons.

Chau, K. W. (1993). Estimating industry-level productivity trends in the building industry from building cost and price data. *Construction Management and Economics*, *11*(5), 370–383.

Chen, W. T., & Huang, Y-H. (2006). Approximately predicting the cost and duration of school reconstruction projects in Taiwan. *Construction Management and Economics*, *24*(12), 1231–1239.

Cheng, B., & Titterington, D. M. (1994). Neural networks: a review from a statistical perspective. *Statistical Science*, *9*(1), 2–30.

Chua, D., Kog, Y., Loh, P., & Jaselskis, E. (1997). Model for construction budget performance—neural network approach. *Journal of Construction Engineering and Management*, *123*(3), 214–222.

Conte, S. D., Dunsmore, H. E., & Shen, V. Y. (1985). Software effort estimation and productivity. *Advances in Computers*, *24*, 1–60.

Cortés, M., Villacampa, Y., Mateu, J., & Usó, J. L. (2000). A new methodology for modelling highly structured systems. *Environmental Modelling & Software*, *15*(5), 461–470.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314.

Devore, J. L. (2003). *Probability and Statistics for Engineering and the Sciences* (6th ed.). Duxbury, Massachusetts, USA: Duxbury Press.

Doloi, H., Sawhney, A., Iyer, K. C., & Rentala, S. (2012). Analysing factors affecting delays in Indian construction projects. *International Journal of Project Management*, *30*(4), 479–489.

Drago, G. P., & Ridella, S. (1992). Statistically controlled activation weight initialization (SCAWI). *Neural Networks, IEEE Transactions on*, *3*(4), 627–631.

Duch, W. (1999). Alternatives to gradient-based neural training. In *Fourth Conference on Neural Networks and Their Applications*, May 1999, Zakopane, Poland.

Duch, W., Adamczak, R., & Jankowski, N. (1997). Initialization and optimization of multilayered perceptrons. In *Third Conference on Neural Networks and Their Applications* (pp. 99–104), October 1997, Kule, Poland.

Dursun, O., & Stoy, C. (2011a). An evaluation of construction speed performance for buillding construction projects in UK and Germany. In *Management and Innovation for a Sustainable Built Environment*, 20 – 23 June 2011, Amsterdam, The Netherlands.

Dursun, O., & Stoy, C. (2011b). Time–cost relationship of building projects: statistical adequacy of categorization with respect to project location. *Construction Management and Economics*, *29*(1), 97–106.

Dursun, O., & Stoy, C. (2012). Determinants of construction duration for building projects in Germany. *Engineering, Construction and Architectural Management*, *19*(4), 444–468.

Eisenhauer, J. G. (2003). Regression through the origin. *Teaching Statistics*, *25*(3), 76–80.

Elazouni, A., Nosair, I., Mohieldin, Y., & Mohamed, A. (1997). Estimating resource requirements at conceptual design stage using neural networks. *Journal of Computing in Civil Engineering*, *11*(4), 217–223.

ElSawy, I., Hosny, H., & Abdel Razek, M. (2011). A neural network model for construction projects site overhead cost estimating in Egypt. *International Journal of Computer Science Issues, 8*(3), 273–283.

Fahlman, S. E. (1988). Faster-learning variations of back-propagation: an empirical study. In *Connectionist Models Summer School* (pp. 38–51), 17-26 June 1988, San Mateo, CA, USA: Morgan Kaufmann.

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In *Advances in Neural Information Processing Systems 2* (pp. 524–532), San Francisco, CA, USA: Morgan Kaufmann.

Faraway, J., & Chatfield, C. (1998). Time series forecasting with neural networks: a comparative study using the air line data. *Applied Statistics*, *47*(2), 231–250.

Fletcher, R., & Reeves, C. M. (1964). Function minimization by conjugate gradients. *Computer Journal*, *7*, 149–154.

Forsythe, P., Davidson, W., & Phua, F. (2010). Predictors of construction time in detached housing projects. In *35th Annual Meeting of the AUBEA*, 14-16 July 2010*,* Melbourne, Australia.

Fortin, V., Ouarda, T. B. M. J., & Bobée, B. (1997). Comment on "The use of artificial neural networks for the prediction of water quality parameters" by H.R. Maier and G.C. Dandy. *Water Resources Research*, *33*(10), 2423–2424.

Foss, T., Stensrud, E., Kitchenham, B., & Myrtveit, I. (2003). A simulation study of the model evaluation criterion MMRE. *Software Engineering, IEEE Transactions on*, *29*(11), 985–995.

Frías, E. (2004). *Aportaciones al estudio de las máquinas eléctricas de flujo axial mediante la aplicación del método de los elementos finitos*. (PhD thesis) Universidad Politécnica de Cataluña, Spain.

Gale, A. W., & Fellows, R. F. (1990). Challenge and innovation: The challenge to the construction industry. *Construction Management and Economics*, *8*(4), 431–436.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

Ghaboussi, J., & Sidarta, D. E. (1998). New nested adaptive neural networks (NANN) for constitutive modeling. *Computers and Geotechnics*, *22*(1), 29–52.

Ghaffari, A., Abdollahi, H., Khoshayand, M. R., Bozchalooi, I. S., Dadgar, A., & Rafiee-Tehrani, M. (2006). Performance comparison of neural network training algorithms in modeling of bimodal drug delivery. *International Journal of Pharmaceutics*, *327*(1-2), 126–38.

Gilchrist, W. (1984). *Statistical Modelling*. Chichester, Sussex, UK: John Wiley & Sons.

Goh, B-H. (1998). Forecasting residential construction demand in Singapore: a comparative study of the accuracy of time series, regression and artificial neural network techniques. *Engineering, Construction and Architectural Management*, *5*(3), 261–275.

Greenwood, D. J., & Shaglouf, A. A. (1997). Comparison between planned and actual durations in medium sized building projects. In *13th Annual ARCOM Conference* (pp. 233–241), 15-17 September 1997, Cambridge, UK.

Günaydın, H. M., & Doğan, S. Z. (2004). A neural network approach for early cost estimation of structural systems of buildings. *International Journal of Project Management*, *22*(7), 595–602.

Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. *Neural Networks, IEEE Transactions on*, *5*(6), 989–993.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate Data Analysis* (5th ed.). Upper Saddle River, New Jersey, USA: Prentice Hall.

Hammerstrom, D. (1993). Working with neural networks. *IEEE Spectrum*, *30*(7), 46–53.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *22*(1), 1–12.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation* (2nd ed.). Upper Saddle River, New Jersey: Prentice Hall.

He, Q., Luo, L., Hu, Y., & Chan, A. P. C. (2015). Measuring the complexity of mega construction projects in China—A fuzzy analytic network process analysis. *International Journal of Project Management*, *33*(3), 549–563.

Hecht-Nielsen, R. (1990). *Neurocomputing*. Boston, MA, USA: Addison-Wesley.

Hegazy, T., Fazio, P., & Moselhi, O. (1994). Developing practical neural network applications using back-propagation. *Computer-Aided Civil and Infrastructure Engineering*, *9*(2), 145–159.

Hegazy, T., & Moselhi, O. (1994). Analogy-based solution to markup estimation problem. *Journal of Computing in Civil Engineering*, *8*(1), 72–87.

Hoffman, G., Thal, A., Webb, T., & Weir, J. (2007). Estimating performance time for construction projects. *Journal of Management in Engineering*, *23*(4), 193–199.

Hollander, M., Wolfe, D. A., & Chicken, E. (2014). *Nonparametric Statistical Methods* (3rd ed.). Hoboken, New Jersey, USA: John Wiley & Sons.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Ireland, V. (1983). *The role of managerial actions in the cost, time and quality performance of high-rise commercial building projects*. (PhD thesis) University of Sydney, Australia.

Ireland, V. (1985). The role of managerial actions in the cost, time and quality performance of high-rise commercial building projects. *Construction Management and Economics*, *3*(1), 59–87.

Ireland, V. (1986). A comparison of Australian and US building performance for high-rise buildings. *School of Building Studies, University of Technology Sydney, Sydney, Australia*.

Irfan, M., Khurshid, M. B., Anastasopoulos, P., Labi, S., & Moavenzadeh, F. (2011). Planning-stage estimation of highway project duration on the basis of anticipated project cost, project type, and contract type. *International Journal of Project Management*, *29*(1), 78–92.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks*, *1*(4), 295–307.

Jafarzadeh, R., Ingham, J., Wilkinson, S., González, V., & Aghakouchak, A. (2014). Application of artificial neural network methodology for predicting seismic retrofit construction costs. *Journal of Construction Engineering and Management*, *140*(2), 4013044.

Jain, M., & Pathak, K. (2014). Applications of artificial neural network in construction engineering and management - A review. *International Journal of Engineering Technology, Management and Applied Sciences*, *2*(3).

Jørgensen, M. (2007). A critique of how we measure and interpret the accuracy of software development effort estimation. In *1st International Workshop on Software Productivity Analysis and Cost Estimation* (pp. 15–22), 4 December 2007, Nagoya, Japan.

Kaka, A. P., & Price, D. F. (1991). Relationship between value and duration of construction projects. *Construction Management and Economics*, *9*(4), 383–400.

Kamarthi, S., Sanvido, V., & Kumara, S. (1992). Neuroform—Neural network system for vertical formwork selection. *Journal of Computing in Civil Engineering*, *6*(2), 178–199.

Khodakarami, V., Fenton, N., & Neil, M. (2007). Project scheduling: improved approach to incorporate uncertainty using Bayesian networks. *Project Management Journal*, *38*(2), 39–49.

Kim, G-H., An, S-H., & Kang, K-I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment*, *39*(10), 1235–1242.

Kim, J. H., & Ji, P. I. (2015). Significance testing in empirical finance: a critical review and assessment. *Journal of Empirical Finance*, *34*, 1–14.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680.

Kitchenham, B., Pickard, L. M., MacDonell, S. G., & Shepperd, M. (2001). What accuracy statistics really measure [software estimation]. *Software, IEE Proceedings*, *148*(3), 81–85.

Kocaguneli, E., & Menzies, T. (2013). Software effort models should be assessed via leave-one-out validation. *Journal of Systems and Software*, *86*(7), 1879–1890.

Kumaraswamy, M. M., & Chan, D. W. M. (1995). Determinants of construction duration. *Construction Management and Economics*, *13*(3), 209–217.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). New York, USA: McGraw-Hill/Irwin.

Lam, P. T. I., Wong, F. W. H., & Chan, A. P. C. (2006). Contributions of designers to improving buildability and constructability. *Design Studies*, *27*(4), 457–479.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, *22*(1), 45–55.

Leamer, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York, USA: John Wiley & Sons.

Leedy, P. D., & Ormrod, J. E. (2010). *Practical Research: Planning and Design* (9th ed.). Upper Saddle River, New Jersey, USA: Prentice Hall.

Le-Hoai, L., & Lee, Y. D. (2009). Time-cost relationships of building construction project in Korea. *Facilities*, *27*(13/14), 549–559.

Le-Hoai, L., Lee, Y. D., & Cho, J. W. (2009). Construction of time-cost model for building projects in Vietnam. *Korean Journal of Construction Engineering and Management*, *10*(3), 130–138.

Le-Hoai, L., Lee, Y. D., & Nguyen, A. T. (2013). Estimating time performance for building construction projects in Vietnam. *KSCE Journal of Civil Engineering*, *17*(1), 1–8.

Lewis, P. E., & Ward, J. P. (1991). *The Finite Element Method: Principles and Applications*. Boston, CA, USA: Addison-Wesley.

Lin, C. L., Wang, J. F., Chen, C. Y., Chen, C. W., & Yen, C. W. (2009). Improving the generalization performance of RBF neural networks using a linear regression technique. *Expert Systems with Applications*, *36*(10), 12049–12053.

Liu, Y., Starzyk, J. A., & Zhu, Z. (2008). Optimized approximation algorithm in neural networks without overfitting. *Neural Networks, IEEE Transactions on*, *19*(6), 983–995.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance “F” test. *Review of Educational Research*, *66*(4), 579–619.

López-Martín, C. (2015). Predictive accuracy comparison between neural networks and statistical regression for development effort of software projects. *Applied Soft Computing*, *27*, 434–449.

Love, P., Tse, R., & Edwards, D. (2005). Time–cost relationships in Australian building construction projects. *Journal of Construction Engineering and Management*, *131*(2), 187–194.

Lowe, D., Emsley, M., & Harding, A. (2006). Predicting construction cost using multiple regression techniques. *Journal of Construction Engineering and Management*, *132*(7), 750–758.

Lu, Y., Luo, L., Wang, H., Le, Y., & Shi, Q. (2015). Measurement model of project complexity for large-scale projects from task and organization perspective. *International Journal of Project Management*, *33*(3), 610–622.

Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software*, *15*(1), 101–124.

Maier, H. R., Jain, A., Dandy, G. C., & Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. *Environmental Modelling & Software*, *25*(8), 891–909.

Markowski, C. A., & Markowski, E. P. (1990). Conditions for the effectiveness of a preliminary test of variance. *The American Statistician*, *44*(4), 322–326.

Masters, T. (1993). *Practical Neural Network Recipies in C++*. San Diego, CA, USA: Academic Press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

Meng, X. (2012). The effect of relationship management on project performance in construction. *International Journal of Project Management*, *30*(2), 188–198.

Mesarovic, M. D., & Takahara, Y. (1975). *General Systems Theory: Mathematical Foundations*. New York, USA: Academic Press.

Minns, A. W., & Hall, M. J. (1996). Artificial neural networks as rainfall-runoff models. *Hydrological Sciences Journal*, *41*(3), 399–417.

Miyazaki, Y., Terakado, M., Ozaki, K., & Nozaki, H. (1994). Robust regression for developing software estimation models. *Journal of Systems and Software*, *27*(1), 3–16.

Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, *6*(4), 525–533.

Montaño, J. J. (2002). *Redes neuronales artificiales aplicadas al análisis de datos*. (Phd thesis) Universitat de les Illes Balears, Spain.

Moody, J. (1994). Prediction risk and architecture selection for neural networks (pp. 147–165). Springer-Verlag.

Moselhi, O., Hegazy, T., & Fazio, P. (1991). Neural networks as tools in construction. *Journal of Construction Engineering and Management*, *117*(4), 606–625.

Murtaza, M., & Fisher, D. (1994). Neuromodex—Neural network system for modular construction decision making. *Journal of Computing in Civil Engineering*, *8*(2), 221–233.

Myrtveit, I., Stensrud, E., & Shepperd, M. (2005). Reliability and validity in comparative studies of software prediction models. *Software Engineering, IEEE Transactions on*, *31*(5), 380–391.

Nahapiet, H., & Nahapiet, J. (1985). A comparison of contractual arrangements for building projects. *Construction Management and Economics*, *3*(3), 217–231.

Nahapiet, J., & Nahapiet, H. (1985). The management of construction projects - Case studies from the USA and UK. *Chartered Institute of Building, UK*.

Navarro-González, F. J., & Villacampa, Y. (2012). A new methodology for complex systems using n-dimensional finite elements. *Advances in Engineering Software*, *48*, 52–57.

Navarro-González, F. J., & Villacampa, Y. (2013). Generation of representation models for complex systems using Lagrangian functions. *Advances in Engineering Software*, *64*, 33–37.

Nawari, N. O., Liang, R., & Nusairat, J. (1999). Artificial intelligence techniques for the design and analysis of deep foundations. *Electronic Journal of Geotechnical Engineering*, *4*.

Ng, S. T., Mak, M. M. Y., Skitmore, R. M., Lam, K. C., & Varnam, M. (2001). The predictive ability of Bromilow's time-cost model. *Construction Management and Economics*, *19*(2), 165–173.

Nguyen, D., & Widrow, B. (1990). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In *International Joint Conference on Neural Networks* (pp. 3:21–26), 17-21 June 1990, San Diego, CA, USA.

Nkado, R. N. (1992). Construction time information system for the building industry. *Construction Management and Economics*, *10*(6), 489–509.

Nkado, R. N. (1995). Construction time-influencing factors: the contractor's perspective. *Construction Management and Economics*, *13*(1), 81–89.

Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: a review of applications. *Expert Systems with Applications*, *36*(1), 2–17.

Park, J., & Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, *3*, 246–257.

Patel, D., & Jha, K. (2015). Neural network model for the prediction of safe work behavior in construction projects. *Journal of Construction Engineering and Management*, *141*(1), 4014066.

Pérez-Carrió, A., Villacampa, Y., Llorens, J., & García-Alonso, F. (2009). A computational algorithm for system modelling based on bi-dimensional finite element techniques. *Advances in Engineering Software*, *40*(1), 30–40.

Petruseva, S., Zujo, V., & Zileska-Pancovska, V. (2013). Neural network prediction model for construction project duration. *International Journal of Engineering Research & Technology*, *2*(11).

Philips, N. A. (1959). An example of nonlinear computational instability. *The Atmosphere and the Sea in Motion, Ed. by B. Bolin, Rockefeller Institute, New York*, 501–504.

Piotrowski, A. P., & Napiorkowski, J. J. (2011). Optimizing neural networks for river flow forecasting – Evolutionary Computation methods versus the Levenberg–Marquardt approach. *Journal of Hydrology*, *407*(1-4), 12–27.

Piotrowski, A. P., & Napiorkowski, J. J. (2013). A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling. *Journal of Hydrology*, *476*, 97–111.

Plagianakos, V. P., Magoulas, G. D., & Vrahatis, M. N. (2001). Learning in multilayer perceptrons using global optimization strategies. *Nonlinear Analysis: Theory, Methods & Applications*, *47*(5), 3431–3436.

PMI (2008). *A Guide to the Project Management Body of Knowledge: PMBOK Guide* (4th ed.). Project Management Institute, Inc.

Powell, M. J. D. (1977). Restart procedures for the conjugate gradient method. *Mathematical Programming*, *12*(1), 241–254.

Prechelt, L. (1998). Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, *11*(4), 761–767.

Proverbs, D., Holt, G. D., & Olomolaiye, P. O. (1998). Factors impacting construction project duration: a comparison between France, Germany and the U.K. *Building and Environment*, *34*(2), 197–204.

Raftery, J. (1990). Markets in theory and practice: the determination of price for construction projects. In *Symposium on BUILDINGS: Aspects of the Development Prices* (pp. 18–21). Tsinghua University, Beijing, China.

Razi, M., & Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications*, *29*(1), 65–74.

Rebasa, P. (2003). Entendiendo la “p < 0,001.” *Cirugía Española*, *73*(6), 361–365.

Riedmiller, M., & Braun, H. (1993). A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In *IEEE International Conference on Neural Networks* (pp. 586–591), 28 March 1993 - 01 April 1993, San Francisco, CA, USA.

Ritter, A., & Muñoz-Carpena, R. (2013). Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology*, *480*, 33–45.

Rojas, R. (1996). *Neural Networks: A Systematic Introduction*. Berlin, Germany: Springer-Verlag.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408.

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, USA: Spartan Books.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Parallel distributed processing: explorations in the microstructure of cognition, Vol. 1. In D. E. Rumelhart, J. L. McClelland, & C. PDP Research Group (Eds.), (pp. 318–362). Cambridge, MA, USA: MIT Press.

Rumelhart, D. E., Widrow, B., & Lehr, M. A. (1994). The basic ideas in neural networks. *Communications of the ACM*, *37*(3), 87–92.

Sanvido, V., Grobler, F., Parfitt, K., Guvenis, M., & Coyle, M. (1992). Critical success factors for construction projects. *Journal of Construction Engineering and Management*, *118*(1), 94–111.

Sexton, R. S., Dorsey, R. E., & Johnson, J. D. (1999). Beyond back propagation: using simulated annealing for training neural networks. *Journal of Organizational and End User Computing*, *11*(3), 3–10.

Shahin, M. A., Jaksa, M. B., & Maier, H. R. (2008). State of the art of artificial neural networks in geotechnical engineering. *Electronic Journal of Geotechnical Engineering*, *8*, 1–26.

Shahin, M. A., Maier, H. R., & Jaksa, M. (2004). Data division for developing neural networks applied to geotechnical engineering. *Journal of Computing in Civil Engineering*, *18*(2), 105–114.

Shahin, M. A., Maier, H. R., & Jaksa, M. B. (2005). Investigation into the robustness of artificial neural networks for a case study in civil engineering. In *16th International Congress on Modeling and Simulation, MODSIM 2005* (pp. 79–83), December 2005, Melbourne, Australia.

Shepperd, M., & MacDonell, S. (2012). Evaluating prediction systems in software project estimation. *Information and Software Technology*, *54*(8), 820–827.

Sidwell, A. C. (1982). *A critical study of project team organisational forms within the building process*. (PhD thesis) Aston University, Birmingham, UK.

Silva, F. M., & Almeida, L. B. (1990). Speeding-up backpropagation. In *R. Eckmiller (Ed.), Advanced Neural Computers* (pp. 151–158). North-Holland, Amsterdam, Netherlands.

Smith, A. E., & Mason, A. K. (1997). Cost estimation predictive modeling: regression versus neural network. *The Engineering Economist*, *42*(2), 137–161.

Soltani, M. M., Motamedi, A., & Hammad, A. (2015). Enhancing Cluster-based RFID Tag Localization using artificial neural networks and virtual reference tags. *Automation in Construction*, *54*, 93–105.

Sousa, V., Almeida, N., & Dias, L. (2014). Role of statistics and engineering judgment in developing optimized time-cost relationship models. *Journal of Construction Engineering and Management*, *140*(8), 4014034.

Sprent, P., & Smeeton, N. C. (2001). *Applied Nonparametric Statistical Methods* (3rd ed.). London, UK: Chapman & Hall/CRC.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(2), 111–147.

Stoy, C., Dreier, F., & Schalcher, H-R. (2007). Construction duration of residential building projects in Germany. *Engineering, Construction and Architectural Management*, *14*(1), 52–64.

Stoy, C., Pollalis, S., & Schalcher, H-R. (2007). Early estimation of building construction speed in Germany. *International Journal of Project Management*, *25*(3), 283–289.

Stretton, A., & Stevens, G. (1990). PREDICTE: an expert system for estimating indicative construction times for multi-storey buildings at concept stages. In *CIB 90 International Symposium on Building Economics and Construction Management* (pp. 2: 590–601), 14-21 March 1990, Sydney, Australia.

Sun, Y., Peng, Y., Chen, Y., & Shukla, A. J. (2003). Application of artificial neural networks in the design of controlled release drug delivery systems. *Advanced Drug Delivery Reviews*, *55*(9), 1201–1215.

Tokar, A., & Johnson, P. (1999). Rainfall-runoff modeling using artificial neural networks. *Journal of Hydrologic Engineering*, *4*(3), 232–239.

Tommelein, I., Riley, D., & Howell, G. (1999). Parade game: Impact of work flow variability on trade performance. *Journal of Construction Engineering and Management*, *125*(5), 304–310.

Treadgold, N. K., & Gedeon, T. D. (1998). Simulated annealing and weight decay in adaptive learning: the SARPROP algorithm. *Neural Networks, IEEE Transactions on*, *9*(4), 662–668.

Trietsch, D., & Baker, K. R. (2012). PERT 21: Fitting PERT/CPM for use in the 21st century. *International Journal of Project Management*, *30*(4), 490–502.

Trietsch, D., Mazmanyan, L., Gevorgyan, L., & Baker, K. R. (2012). Modeling activity times by the Parkinson distribution with a lognormal core: Theory and validation. *European Journal of Operational Research*, *216*(2), 386–396.

Turner, J. R., & Cochrane, R. A. (1993). Goals-and-methods matrix: coping with projects with ill defined goals and/or methods of achieving them. *International Journal of Project Management*, *11*(2), 93–102.

Verdú, F. (2004). *Un algoritmo para la construcción múltiple de modelos matemáticos no lineales y el estudio de su estabilidad*. (PhD thesis) University of Alicante, Spain.

Verdú, F., & Villacampa, Y. (2008). A computational algorithm for the multiple generation of nonlinear mathematical models and stability study. *Advances in Engineering Software*, *39*(5), 430–437.

Vidal, L-A., & Marle, F. (2008). Understanding project complexity: implications on project management. *Kybernetes*, *37*(8), 1094–1110.

Villacampa, Y., Uso, J. L., Vives, F., & Cortés, M. (1998). *An introduction to the sensitivity and stability in risk models*. (W. P. M. Publications, Ed.). Southampton, UK.

Villacampa, Y., Verdú, F., & Pérez, A. (2007). A stability theory for model system. *Kybernetes*, *36*(5/6), 683–696.

Walker, D. H. T. (1994). *An investigation into factors that determine building construction time performance*. (PhD thesis) Royal Melbourne Institute of Technology, Australia.

Walker, D. H. T. (1995). An investigation into construction time performance. *Construction Management and Economics*, *13*(3), 263–274.

Wang, Y-R., & Gibson, G. E. (2010). A study of preproject planning and project success using ANNs and regression models. *Automation in Construction*, *19*(3), 341–346.

Wang, Y-R., Yu, C-Y., & Chan, H-H. (2012). Predicting construction cost and schedule success using artificial neural networks ensemble and support vector machines classification models. *International Journal of Project Management*, *30*(4), 470–478.

Widrow, B., & Hoff, M. (1960). Adaptive switching circuits. In *Western Electronic Show and Convention* (pp. 4: 96–104). Institute of Radio Engineers.

Williams, T. M. (1999). The need for new paradigms for complex projects. *International Journal of Project Management*, *17*(5), 269–273.

Williams, T. M. (2002). *Modelling Complex Projects*. Chichester, Sussex, UK: John Wiley & Sons.

Wilson, D. R., & Martinez, T. R. (2003). The general inefficiency of batch training for gradient descent learning. *Neural Networks : The Official Journal of the International Neural Network Society*, *16*(10), 1429–51.

Xia, B., & Chan, A. P. C. (2012). Measuring complexity for building projects: a Delphi study. *Engineering, Construction and Architectural Management*, *19*(1), 7–24.

Xiang, C., Ding, S. Q., & Lee, T. H. (2005). Geometrical interpretation and architecture selection of MLP. *Neural Networks, IEEE Transactions on*, *16*(1), 84–96.

Xiao, H., & Proverbs, D. (2002). Construction time performance: an evaluation of contractors from Japan, the UK and the US. *Engineering, Construction and Architectural Management*, *9*(2), 81–89.

Yeh, I. (1998). Quantity estimating of building with logarithm-neuron networks. *Journal of Construction Engineering and Management*, *124*(5), 374–380.

Zhang, G., Eddy Patuwo, B., & Y. Hu, M. (1998). Forecasting with artificial neural networks: the state of art. *International Journal of Forecasting*, *14*(1), 35–62.

Zhang, Y. F., & Fuh, J. Y. H. (1998). A neural network approach for early cost estimation of packaging products. *Computers & Industrial Engineering*, *34*(2), 433–450.

Zienkiewicz, O. C., & Taylor, R. L. (1994). *El Metodo de los Elementos Finitos: Formulación Básica y Problemas Lineales* (4th. ed.). Barcelona, Spain: McGraw-Hill/CIMNE.

ANNEX A. Conclusiones en castellano

A.1 Verificación de las hipótesis de investigación

La estimación del tiempo de construcción en la fase inicial del desarrollo de proyectos de edificación, cuando sus responsables disponen todavía de poca información, ha sido considerada un elemento clave para finalizar con éxito dichos proyectos. La mayor parte de la literatura ha identificado a los denominados factores de alcance del proyecto como predictores clave de la duración de la fase de construcción, sin embargo existen conclusiones contradictorias sobre qué factores ejercen mayor influencia. Aunque los costes de construcción y la superficie construida (SC) son los factores que se han utilizado más comúnmente para definir el alcance del proyecto, no existen unidades de medición perfectas. Por otro lado, investigaciones anteriores han mostrado que el concepto que representa la velocidad de construcción es un punto de referencia útil y fiable para comparar el rendimiento productivo entre contratistas y además puede ser utilizado como una variable de respuesta alternativa para estimar la duración del proceso constructivo. En este sentido, existe un debate en la literatura sobre si es posible obtener mejores modelos predictivos utilizando el tiempo de construcción como variable dependiente o si es más apropiado considerar la velocidad de construcción para este propósito.

Con objeto de proporcionar herramientas adecuadas para estimar el tiempo de construcción y minimizar la subjetividad en tal estimación, hasta la fecha, la mayoría de los trabajos de investigación presentados en la literatura han desarrollado modelos paramétricos utilizando técnicas basadas en el análisis de regresión lineal (ARL). Sin embargo, aunque este tipo de modelos ha mostrado un buen equilibrio entre la dificultad para desarrollarlos y la precisión predictiva obtenida con ellos, representa una gran simplificación de las complejas relaciones que controlan la ejecución de proyectos de edificación. Por tanto, los factores de alcance del proyecto que influyen en la velocidad de construcción podrían no estar completamente asociados de una forma lineal, en cuyo caso mediante el desarrollo de modelos no lineales sería posible

representar mejor las relaciones existentes entre dichos factores y la duración del proceso constructivo.

De acuerdo con lo establecido anteriormente, en esta tesis se postularon tres hipótesis principales:

(H-01) "*La velocidad de construcción es una variable dependiente más apropiada que el tiempo de construcción para generar modelos predictivos que permitan estimar la duración del proceso constructivo de edificios de nueva planta*."

(H-02) "*La SC tiene una mayor influencia sobre la velocidad de construcción de edificios de nueva planta que los costes de construcción*."

(H-03) "*Considerando el mismo conjunto de factores de alcance del proyecto como variables predictoras, las técnicas de modelado no lineal pueden generar modelos para estimar la velocidad de construcción de edificios de nueva planta con mayor rendimiento predictivo que el ofrecido por los modelos de regresión lineal*."

Con el propósito de verificar estas hipótesis, se desarrollaron modelos lineales y no lineales utilizando varios factores de alcance del proyecto para estimar la velocidad de construcción de proyectos de edificación de nueva planta desarrollados en España. A fin de desarrollar modelos no lineales con mejor rendimiento predictivo, el conjunto de variables que obtuvo los mejores resultados con el ARL fue también utilizado con redes neuronales artificiales (RNA) y una novedosa metodología numérica basada en el método de los elementos finitos (MEF). Es necesario resaltar que los modelos generados en esta tesis no han sido presentados para ser utilizados de manera práctica, sino más bien para analizar el tipo de relaciones existentes entre las variables evaluadas, así como también para identificar que metodología de modelado produce mejores modelos predictivos.

Las dos primeras hipótesis fueran verificadas positivamente en el capítulo IV. Por consiguiente, puede concluirse que, por un lado, la velocidad de construcción es la variable de respuesta más apropiada para desarrollar modelos predictivos que permitan estimar la duración del proceso de construcción de edificios de nueva planta, y, por otro lado, esta velocidad de construcción se ve afectada más por la SC que por el coste de construcción. En consecuencia, la SC es el mejor indicador del alcance del proyecto y el mejor predictor de la velocidad media de construcción.

El capítulo V proporciona una evidencia clara que soporta la tercera hipótesis planteada en la investigación, ya que utilizando la metodología numérica basada en el MEF es posible desarrollar modelos no lineales con mejor rendimiento predictivo que el ofrecido por los modelos de regresión lineal.

A.2 Principales aportaciones

De acuerdo con la metodología de investigación y los resultados presentados en los capítulos precedentes, las principales aportaciones de esta tesis han sido:

- ***Desarrollo de un modelo mediante análisis de regresión lineal múltiple (ARLM) que proporciona una clara comprensión de las relaciones existentes entre algunos de los principales factores de alcance del proyecto y el tiempo requerido para completar la fase de construcción en edificios de nueva planta desarrollados en España***. En concordancia con los factores más utilizados en la literatura, el mejor modelo de regresión lineal se corresponde con un modelo logarítmico que utiliza una variable categórica que representa el tipo de instalación, junto con tres variables cuantitativas correspondientes a los siguientes factores de alcance del proyecto: SC, relación coste/SC (*Estándar*) y número total de plantas.

  De acuerdo con los distintos modelos lineales presentados, se puede inferir que tanto la SC como los costes de construcción son factores de alcance del proyecto de gran importancia para estimar la velocidad de construcción en edificios de nueva planta. Sin embargo, aunque individualmente la SC tiene una mayor influencia sobre la velocidad de construcción que los costes, mediante la combinación de ambos factores con la variable *Estándar* es posible desarrollar modelos con un mayor rendimiento predictivo. Este hallazgo sugiere la idea de que ambas variables definen diferentes relaciones con la velocidad de construcción pero complementarias. Proyectos con una gran SC permiten un uso más eficiente de los factores de producción y, por tanto, una mayor velocidad de construcción. Sin embargo, esta velocidad se ve negativamente afectada cuando se incrementa el valor de la relación entre el coste y la SC, indicando en este caso que una mayor calidad de la construcción se traducirá en una mayor complejidad y una menor velocidad de construcción. Los resultados del estudio parecen también indicar que, dentro de los valores límite establecidos por las características de los

datos utilizados en la investigación, las obras de construcción se desarrollan horizontalmente a una velocidad más alta que verticalmente. No obstante, la incorporación en el modelo de regresión lineal de una variable que representa el tipo de instalación reveló que la velocidad de construcción sigue un patrón diferente en algunos tipos específicos de proyectos de edificación.

- ***Desarrollo de una metodología de optimización que permite la obtención de configuraciones de RNA perceptrón multicapa (PMC) con un adecuado rendimiento predictivo***. La metodología de optimización propuesta desarrolla un proceso de ensayo y error por etapas en el que desde una estructura de red inicial con suficiente consistencia se generan seis modelos PMC optimizados. El conjunto de variables utilizado para generar el mejor modelo PMC fue el mismo utilizado para generar el mejor modelo de regresión lineal. No obstante, el modelo PMC utilizó las variables en su forma natural mientras que el modelo de regresión lineal tuvo que desarrollar una transformación logarítmica de las variables para poder cumplir con los supuestos subyacentes del ARLM.

  A pesar de los problemas identificados en la literatura respecto a la efectividad del algoritmo de entrenamiento tradicional de gradiente descendente (GD) y su dependencia de los parámetros utilizados durante el proceso de aprendizaje, los resultados de esta tesis parecen sugerir que utilizando este algoritmo con las redes PMC es posible generar mejores modelos predictivos que aplicando el algoritmo del gradiente conjugado escalado (GCE) o las RNA de base radial. Sin embargo, es necesario señalar que debido a que el algoritmo del GCE no contiene parámetros importantes dependientes del usuario para el desarrollo adecuado del proceso de entrenamiento, utilizando este algoritmo se reduce considerablemente el tiempo requerido para aplicar el proceso de optimización propuesto para las redes PMC.
- ***Evaluación de la influencia de la división de los datos de calibración sobre el rendimiento predictivo de las redes PMC.*** La experimentación desarrollada con redes PMC mostró que el tipo de división de datos utilizado para calibrar este tipo de redes tiene un impacto directo sobre su precisión predictiva. En particular, se obtuvieron los mejores modelos de predicción utilizando una división 80-20 para los datos de calibración y este resultado es consistente con la sugerencias realizadas por otros autores (ver, por ejemplo, Haykin, 1999).

- ***Evaluación de la influencia de la estructura de red, los parámetros de entrenamiento y el conjunto de pesos iniciales sobre el rendimiento predictivo de las redes PMC.*** Este trabajo de investigación halló que el número de neuronas ocultas y el conjunto de pesos seleccionados inicialmente fueron los parámetros que más influyeron en la precisión predictiva de las redes PMC. No obstante, parece que el impacto de la selección de los pesos iniciales sobre el rendimiento predictivo de un modelo PMC podría reducirse utilizando una estructura de red optimizada. Además, de manera contraria a lo que ha sido establecido en muchos otros estudios, en esta investigación los parámetros de entrenamiento del algoritmo de GD no tuvieron ningún impacto sobre la precisión de los modelos PMC generados.
- ***Aplicación de una metodología numérica basada en el MEF para desarrollar modelos predictivos relacionados con la estimación del tiempo de construcción en edificaciones de obra nueva.*** Esta metodología numérica es una técnica de modelado novedosa que no ha sido aplicada previamente en la literatura con el objeto de estimar la velocidad de construcción. Al igual que el mejor modelo obtenido utilizando RNA, los mejores modelos numéricos fueron obtenidos utilizando las variables en su forma natural. Sin embargo, en este caso el conjunto de variables que ofreció el mejor rendimiento predictivo no contenía la variable categórica que representa el tipo de edificación, lo que podría significar una prueba de que la metodología numérica basada en el MEF funciona de manera más efectiva que el ARLM y las RNA.
- ***Desarrollo de un análisis de estabilidad para validar los modelos predictivos.*** Los valores de bondad de ajuste y precisión han sido tradicionalmente utilizados para validar tanto modelos lineales como modelos no lineales. Sin embargo, esta tesis también propone el desarrollo de un análisis de estabilidad como elemento adicional de validación de los modelos.

  Para modelos desarrollados con ARL se llevó a cabo un análisis univariante de sensibilidad examinando la bondad del acuerdo entre los valores estimados por los modelos y los procesos físicos subyacentes conocidos del problema estudiado. Los resultados de este análisis mostraron como un modelo lineal con un alto grado de ajuste y precisión podría tener un comportamiento anómalo, que lo hace inutilizable para su aplicación práctica.

En el caso de los modelos numéricos basados en el MEF y las RNA se desarrolló un análisis multivariante para probar su sensibilidad a pequeñas variaciones introducidas en los datos experimentales. Los resultados de este análisis demostraron que los mejores modelos no lineales fueron estables frente a perturbaciones aleatorias desarrolladas en los datos de calibración.

- ***Desarrollo de un análisis de sensibilidad para evaluar como la variabilidad de los costes de construcción puede afectar el rendimiento predictivo de los modelos.*** Los costes de construcción son sensibles a efectos comerciales y el coste final normalmente varía respecto al coste estimado inicialmente antes de comenzar las obras. Aunque en esta tesis el uso del coste estimado en comparación con el uso del coste final dentro de la variable *Estándar*, en promedio, produjo peores valores de precisión, la variación de precisión no fue estadísticamente significativa. Además, los intentos realizados para no tener en cuenta el costo como una variable predictora produjeron modelos con peor ajuste y precisión. También es necesario señalar que los modelos no lineales seleccionados fueron más sensibles a la variabilidad del coste que el mejor modelo de regresión lineal, de modo que tal variabilidad debe tenerse en cuenta en la aplicación práctica de este tipo de modelos predictivos.
- ***Desarrollo de un análisis comparativo entre tres técnicas de modelado diferentes utilizadas para estimar la velocidad de construcción de edificios de nueva planta.*** Esta tesis demostró que mediante la utilización de la metodología numérica basada en el MEF es posible obtener familias de modelos no lineales que nos permiten predecir la velocidad de construcción con mayor precisión que modelos creados mediante el ARLM o las RNA. Además, los resultados del análisis estadístico desarrollado en el Capítulo V mostraron que el error medio de predicción del mejor modelo numérico fue significativamente diferente del error medio de predicción ofrecido tanto por el mejor modelo desarrollado mediante el ARLM como por la mejor red PMC. Por el contrario, las diferencias de precisión entre el mejor modelo de regresión lineal y la mejor red PMC no fueron estadísticamente significativas.

  Sin embargo, debe señalarse que, debido al gran número de posibles configuraciones diferentes que pueden utilizarse para desarrollar redes PMC con el algoritmo de entrenamiento de GD, y la falta de reglas fijas para determinar la mejor configuración de la red neuronal, no existe ninguna garantía de que la red

PMC seleccionada fuese la mejor posible. En este sentido, la principal ventaja de modelado que ofrece la metodología numérica basada en el MEF con respecto a las RNA es que, con excepción de la elección de la complejidad del modelo geométrico, dicha metodología no depende de ningún parámetro a definir por el usuario para generar buenos modelos, como si sucede al utilizar las redes PMC.

También es necesario destacar que los modelos de regresión lineal permiten una interpretación clara de cómo funcionan las relaciones entre las variables que forman parte del modelo. Por el contrario, tanto las RNA como la metodología numérica basada en el MEF son técnicas de modelado denominadas "cajas negras" y con ellas no es posible explicar de una manera clara la influencia de las variables de entrada sobre la variable de salida de los modelos desarrollados. Por lo tanto, si el objetivo de la modelización es explicar los procesos subyacentes que controlan las relaciones entre las variables en estudio, es mejor utilizar el ARL como técnica de modelado.

Por último, es importante señalar que las conclusiones de esta tesis apoyan la idea de que los modelos de regresión lineal pueden proporcionar un buen punto de partida para la búsqueda de modelos no lineales con mayor rendimiento predictivo. En este sentido, las tres técnicas de modelado analizadas en este trabajo pueden entenderse como métodos complementarios para la generación del mejor modelo de predicción posible en lugar de diferentes técnicas de modelado que compiten entre sí. La metodología de investigación desarrollada en este documento proporciona un nuevo marco para generar y evaluar modelos predictivos lineales y no lineales, ampliando las herramientas disponibles para producir modelos de predicción útiles para los profesionales e investigadores del sector de la edificación. Sin embargo, a pesar de que los conocimientos adquiridos en esta investigación permitirán explorar nuevos enfoques para determinar mejor las relaciones existentes entre los factores de alcance del proyecto y la velocidad de construcción, es necesario entender que los resultados proporcionados por el tipo de modelo analizado en esta tesis son sólo estimaciones iniciales en las primeras etapas de desarrollo del proyecto y no pretenden sustituir a las programaciones detalladas realizadas por el constructor.

A.3 Limitaciones del estudio

En primer lugar, el trabajo de investigación presentado en esta tesis ha examinado sólo algunos de los factores que pueden influir en el tiempo de construcción de edificios de nueva planta. Estudios anteriores han demostrado que hay otros factores cualitativos que también afectan a la duración del proceso de construcción, tales como la eficacia en la gestión o la productividad del trabajo, que no han podido ser analizados en esta tesis por la falta de disponibilidad de datos.

En segundo lugar, hay que señalar que los proyectos utilizados en esta investigación han sido ejecutados durante un amplio rango de tiempo y representan diferentes tipos de instalaciones. Por tanto, es lógico pensar que podrían obtenerse conclusiones diferentes si se generaran nuevos modelos con proyectos desarrollados en un intervalo de tiempo más corto o pertenecientes a un tipo muy específico de instalación. Además, esta investigación hizo uso de un número limitado de proyectos y siempre es posible que hubiera información incorrecta en las muestras de los proyectos estudiados, en cuyo caso podrían haber causado distorsiones en los coeficientes y parámetros de los modelos y en su capacidad de predicción. En consecuencia, los modelos presentados en este estudio deben ser analizados con cautela, ya que es necesario llevar a cabo más investigaciones para evaluar su validez sobre nuevos conjuntos de datos y verificar que los resultados obtenidos se mantienen estables.

Por otro lado, cabe señalar que la incertidumbre es una característica inherente a la industria de la construcción. Por tanto, la duración real de la fase de construcción puede estar influenciada por múltiples riesgos e incertidumbres provenientes de diversas fuentes relacionadas con las características del entorno en el que se desarrolla un proyecto de edificación, y estos factores no pueden introducirse en modelos de predicción deterministas como los que se proponen en esta investigación. También es necesario tener en cuenta que antes de iniciar las obras, y durante su ejecución, un contratista puede seleccionar diferentes tipos de cuadrillas de trabajadores, turnos de trabajo, maquinaria y métodos de construcción con el propósito de completar las actividades del proyecto. Estas decisiones, en última instancia, afectan a la duración del proceso constructivo, independientemente del tipo de edificación, SC, costes de construcción o número de plantas.

Otra cuestión a considerar es la posible variación de la productividad en el sector de la edificación con el paso del tiempo. La duración del proceso de construcción de un edificio también varía en función del momento en el que se desarrolla el proyecto. En consecuencia, los coeficientes y parámetros de los modelos deben ser revisados periódicamente para mantenerlos actualizados de acuerdo con la evolución de las técnicas de construcción utilizadas en el sector de la edificación. Pero aun siendo conscientes de los errores de precisión que pueden derivarse de la utilización del tipo de modelos desarrollados en esta tesis, su uso puede ser útil antes de comenzar las obras, una vez que se conozcan los factores de alcance del proyecto involucrados en el modelo. Las estimaciones proporcionadas en este momento pueden servir como un parámetro de control para verificar si los plazos de construcción propuestos por cualquiera de los agentes implicados en la fase de construcción son realistas y, si es necesario, tomar medidas especiales para aumentar la velocidad de construcción del proyecto.

A.4 Trabajos futuros

El autor de esta tesis tiene la intención de llevar a cabo nuevas investigaciones en relación con la estimación del tiempo de construcción en proyectos de edificación, utilizando para ello nuevos conjuntos de datos y aplicando otras metodologías de modelado tales como “super vector machine” o “deep learning”. Por otro lado, tal y como se identificó en el estado del arte de esta tesis, la literatura indica que la complejidad del diseño de un edificio también está relacionada con la velocidad de construcción. Además, el tiempo de construcción también se ha relacionado con la manera en que el constructor tiene previsto llevar a cabo el movimiento de los recursos en la obra y la selección de métodos de construcción adecuados al tipo de proyecto, los cuales, a su vez, se ven afectados por la complejidad del diseño del edificio. En este sentido, deben existir factores más específicos relacionados con la complejidad de la forma de los edificios y la manera en que el constructor organiza el proceso de construcción para superar dicha complejidad, que puedan incorporarse en modelos predictivos con la finalidad de aumentar su precisión. El estudio de estos factores y sus relaciones con la velocidad de construcción de edificaciones de nueva planta será también el foco de nuestros futuros trabajos de investigación.

Sin embargo, la falta de datos provenientes de proyectos reales finalizados limita la posibilidad de mejorar los modelos predictivos presentados. Además, con frecuencia, cuando se proporcionan datos relacionados con las características de los edificios y su proceso de construcción, dichos datos están incompletos o contienen información incorrecta. En ausencia de datos reales que contengan información adecuada sobre un gran número de variables que puedan influir en la velocidad de construcción, un paso lógico para mejorar la fiabilidad de modelos predictivos basados en ecuaciones matemáticas podría ser el desarrollo de dichos modelos a partir de la información generada mediante la simulación del proceso de construcción. Por tanto, otra área para desarrollar futuras investigaciones podría ser el desarrollo de modelos predictivos basados en datos generados mediante la simulación del proceso de construcción y su comparación con los resultados obtenidos con modelos predictivos desarrollados utilizando datos correspondientes a proyectos reales finalizados.