



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Master of Science in Engineering

**Skyscrapers' Conceptual Cost
Estimation with Non-Local Data
using Artificial Neural Networks
and Bootstrap Aggregating**

by

Wagner Nunes de Andrade Neto

Department of Architecture & Architectural Engineering

The Graduate School

Seoul National University

June 2018

Abstract

Skyscrapers' Conceptual Cost Estimation with Non-Local Data using Artificial Neural Networks and Bootstrap Aggregating

Wagner Nunes de Andrade Neto
Department of Architecture and Architectural Engineering
Graduate School
Seoul National University

In recent years the amount of new developed skyscraper buildings have considerable increased around the world. And due to the high construction costs of those projects, there are a lot of financial risks involved on such developments. On that scenario, accurate early cost estimates are of main importance for the successful completion of those buildings.

Preliminary research has shown that neural networks models can be great tools for accurate conceptual construction cost estimation. However, those techniques are highly dependent on considerable amount of previous similar

data that can be hard to acquire in the case of skyscrapers.

With the intention to try applying the addition of non-local to estimate conceptual costs of skyscrapers, this research studied the use of neural networks in association with bootstrap aggregating (bagging) as a way to deal with the limitations on local data.

Data from 124 projects in 5 different countries was collected and different models were tested to assess the efficiency of using artificial neural networks with added non-local data and bagging. The obtained results showed promising potential for the method as a way to increase conceptual cost estimation accuracy both in countries with already sufficient data, and in countries with limited previous projects data.

Keywords: Skyscraper, Bootstrap Aggregating, Artificial Neural Networks, Construction Cost Estimation

Student Number: 2016-26777

Table of Contents

Chapter 1. Introduction	7
1.1 Research Background	7
1.2 Problem Statement	11
1.3 Research Objectives and Scope	12
1.4 Research Methodology	13
Chapter 2. Preliminary Study	16
2.1 Skyscrapers’ Cost Estimation Overview	16
2.2 Machine Learning Techniques for Construction Cost Estimation	19
2.3 Estimating Skyscrapers’ Costs Using Non-Local Data.....	21
2.4 Artificial Neural Networks and Construction Cost Estimation	24
2.5 Bootstrap Aggregating (Bagging) for Construction Cost Estimation	27
2.6 Summary	29
Chapter 3. Learning from Non-local Data & Variable Selection	30
3.1 Learning from Non-local Data	31
3.2 Relevant Variables & Data Collection	33
3.3 Data Preprocessing.....	36
3.4 Selection of Most Relevant Attributes	39
3.5 Summary	42
Chapter 4. Skyscrapers’ Conceptual Cost Estimation Models Using Non-local Data	43
4.1 Artificial Neural Networks and Bagging Models	43
4.1.1 Modelling Datasets	46
4.1.2 ANN Models	47

4.1.3 ANN and Bootstrap Aggregating Models	49
4.2 Other Techniques Models for Accuracy Comparison	51
4.2.1 Average Cost per Area	51
4.2.2 Linear Regression	52
4.2.3 Case-Based Reasoning	53
4.3 Summary	55
Chapter 5. Models Analysis.....	56
5.1 Models for Comparison	57
5.2 Models with Data from Individual Countries	60
5.3 Simple ANN Model with Data from Combined Countries..	62
5.4 ANN and Bootstrap Aggregating Model with Data from Combined Countries	64
5.5 Use of ANN and Bootstrap Aggregating Model in Countries with Less Data.....	67
5.6 Discussion	69
5.7 Summary	72
Chapter 6. Conclusions.....	73
6.1 Research Summary	73
6.2 Contributions	76
6.3 Limitations and Further Research	77
References	79
Appendix A – Selected attributes for data collection.....	82
Appendix B – Example of cost estimation.....	84

List of Tables

Table 3.1 – Created attribute from previous existing attributes.	38
Table 3.2 - Subset Evaluation results for the individual.....	40
Table 4.1 - Definition of best neural structures	48
Table 4.2 – Summary of developed models.....	50
Table 4.3 - Summary of Linear Equations.....	52
Table 5.1 - Accuracy of average cost/area.	57
Table 5.2 - Accuracy of Linear Regression	58
Table 5.3 - Accuracy of CBR	58
Table 5.4 - Results of individual countries models using ANN	59
Table 5.5 - Results for ANN models for combined and single countries datasets.....	61
Table 5.6 - Results for combined, ANN + bagging and ANN only.	63
Table 5.7 - Results for all countries combined, Australia and Singapore	66

List of Figures

Figure 1.1 - Number of 200-meter-plus buildings completed each year from 1960 to 2017	8
Figure 1.2 - Research Process.....	14
Figure 2.1 - Basic structure of artificial neural networks.	24
Figure 2.2 - Bootstrap aggregating process	26
Figure 3.1 - Strategy behind the use of additional non-local data.	30
Figure 3.2 - Data Preprocessing.	35
Figure 3.3 - Process of selection of final modelling attributes.	38
Figure 4.1 - Framework of the development of ANN and Bagging models	43
Figure 5.1 - Comparison of MAPE results of all studied models.....	65

Chapter 1. Introduction

1.1 Research Background

According to the Council on Tall Buildings and Urban Habitat (CTBUH), skyscrapers are high-rise buildings that are 200 m height or more. Through the years, since the first ever built skyscraper, the number of new constructions of these kind of buildings has varied in cycles of peaks of high and low levels of new developments. However, in recent years, there has been an intense increase in the amount of investment allocated to those kind of projects, as since 2014, every year, the number of new completed skyscrapers reaches a historical high. In 2017, the fourth consecutive record-breaking year, the amount of new completions reached 144 new skyscrapers, and by 2018, it is expected that the number of new skyscraper buildings will reach a new record once again (CTBUH, 2017) . The Figure 1.1 shows the number of new skyscrapers per year. As it can be seen, the tendency is that the amount of skyscrapers projects will keep increasing in the years to come.

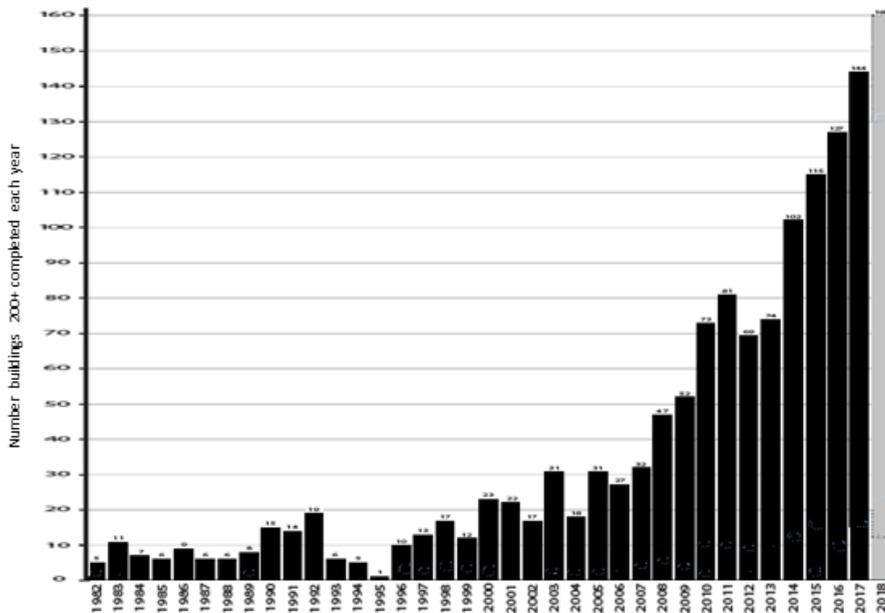


Figure 1.1 – Number of 200-meter-plus buildings completed each year from 1960 to 2017. (adapted from CTBUH, 2017)

The expansion of skyscrapers is not only in matter of number of buildings, but also, according to the CTBUH (2017), the year of 2017 has been the most geographically diverse, with projects been completed across 69 cities in 23 countries, 15 cities and 5 countries more than the previous year of 2016, and with 13 cities completing their first ever skyscraper.

As high rise buildings, these projects are characterized by high investment risks associated to the high costs of construction and complexity of the projects. Most of those buildings reach costs of hundred-millions or billions of dollars, and with the increase in the amount and size of skyscrapers, the risks associated with failure of those projects also increase. Inaccurate initial estimates can

direct the projects to insufficient funds and future failure, for not reaching the low estimated costs, or direct them to not even being made, because of over estimations of needed funds. On that scenario, early cost estimates, or conceptual estimates are of main importance, since it is the base for the decision-making processes of economic feasibility, development strategies and resource commitment (Oberlender and Trost, 2001; Li et al., 2005; Lee et al. 2011).

Conceptual costs estimates can present a challenge for the lack of enough information on the early stages. On that scenario, most estimates require experienced cost estimators responsible for controlling the risks associated with the uncertainties and complexities of construction projects (Ahn, 2016). The high level of uncertainties and complexities in high-rise buildings constructions requires that estimators have specific previous experience with skyscrapers projects. However, even though skyscrapers constructions are more common than ever before, they are still singular projects, and few are the experienced professionals in the world.

Due to the lack of detailed data at early stages and the importance of conceptual cost estimation of construction projects, many approaches have been proposed and studies in the literature that (Li, 1995; Khosrowshahi and Kaka, 1996; Bode, 1998; Perera and Watson, 1998; Skitmore and Ng, 2003; Kim et al, 2004; Wilson, 2005; Lee et al., 2011; Dursun and Stoy, 2016). Among them, machine learning techniques have shown good accuracy results for

construction cost estimation in different kinds of construction projects, while reducing the dependency on the knowledge of a few experts (Kim et al, 2004; An et al., 2007; Arafa and Algedra, 2011; Bayram and Al-Jibouri; Dursun and Stoy, 2016). These approaches show a great potential for use in skyscrapers estimation, and some researchers have already address its use in skyscrapers' construction schedule estimation (Li et al., 2016). However, although powerful, the use of those methodologies is limited by the collection of a considerable amount of similar previous projects data within the studied region what, in most skyscrapers scenarios, can be a great challenge to complete.

1.2 Problem Statement

The amount of new developments of skyscrapers is expected to keep increasing in the future, all around the world. Due to the high construction complexity and the high costs, accurate construction costs estimates at early stages are crucial for the continuous and stable increase and improvement in skyscrapers projects.

Previous research was able to develop accurate machine learning models that relies less on the expertise of experienced professionals for construction cost estimation of different kinds of buildings. However, the amount of research on the use of machine learning techniques on skyscrapers estimations is limited, and there is an intrinsic scarcity of skyscraper similar projects' data on a specific region.

It is important to study the use of machine learning techniques on the conceptual cost estimation of skyscrapers, while developing an accurate model that can adapt to the common scenario of not sufficient similar projects' data in the specific regions.

1.3 Research Objectives and Scope

In order to study the use of machine learning techniques in accurate conceptual cost estimation of skyscrapers, and address the problem of not sufficient local similar data in the skyscrapers scenario, this research proposes the combined use of skyscrapers' project data from different countries in an ensemble model of bootstrap aggregating with artificial neural networks.

The main objective of this research is to develop an accurate model for early cost estimation of skyscrapers' projects that have less dependency on the expertise of few professionals. Also, it investigates the usefulness of variables for conceptual cost estimation of skyscrapers, and the accuracy of different estimating methods.

On the scope of this study, the expression "local data" means data obtained within a same country. Five countries are studied, those being United States of America, China, United Arab Emirates, Singapore and Australia. Although Singapore and Australia are among the countries the ten countries with biggest number of skyscrapers in the world, for the purpose of enough validation data, here they are used as examples of countries with a small amount of skyscrapers' data (eight or less). A small amount of non-skyscrapers high-rise buildings are also used on the modeling in order to increase the amount of available data. And the use of international construction cost index is outside the scope of this work.

1.4 Research Methodology

After this introduction, the research process follows the diagram present in Figure 1.2. It starts by a preliminary study (Chapter 2) that is divided in five main parts. First, a review about skyscrapers, including definition and cost estimation practices and challenges is made. Second, machine learning techniques and their application to construction cost estimations are studied. Then, reviewing skyscrapers and machine learning techniques, it is presented and analyzed the possible challenges on the estimation of construction costs of skyscrapers through machine learning techniques with the addition of data from different countries. Lastly, the next two parts focus on neural networks and bootstrap aggregating methodologies, respectively, as promising solutions for dealing with the use of non-local data.

Following the preliminary study, the idea behind the conceptual cost estimation models studied in this thesis is shown in Chapter 3 with the selection and collection of relevant data. First data is collect from projects from five different countries, the variables selected for collection are chosen according to the preliminary studies and available data in the used sources. After, for optimization, the data is preprocessed and attributes are selected through feature selection techniques in the datasets of the different countries. In Chapter 4, once the data is ready for modelling, different structures of neural networks are tested, until the optimum neural networks structure is found for each of the different

modelling methods. The development of the models is divided about the use of distinct datasets in order to best assess the use of the different techniques. ANN models structures are developed and analyzed in distinct tests sets. In total 9 ANN models are made and compared to each other. All models are developed and tested on WEKA (Waikato Environment for Knowledge Analysis). In addition, three other methodologies are performed to serve as basis for accuracy comparison. Then, initially, non-bagging and bagging models made of individual countries data are tested for comparison with the models made of the combined countries dataset. In sequence, the models utilizing the combined dataset of the three countries with the biggest amount of projects' data are performed and the results compared to the previous models. After, if combined models show improved accuracy over the single country models, then models are developed for the cost estimation of projects in the two countries with less amount of data, in order to assess the adequacy of the methodology in countries than have considerable less that than the previous countries.

After all the models are developed, Chapter 5 analysis the accuracy of the different models in the different countries and datasets, and analysis the overall results of the proposed use of artificial neural networks and bagging method with non-local data.

Then, lastly, Chapter 6 concludes this research with a summary of all this study and its contribution. And it provides insights about the limitation of this research and possible ideas for future research related to this topic.

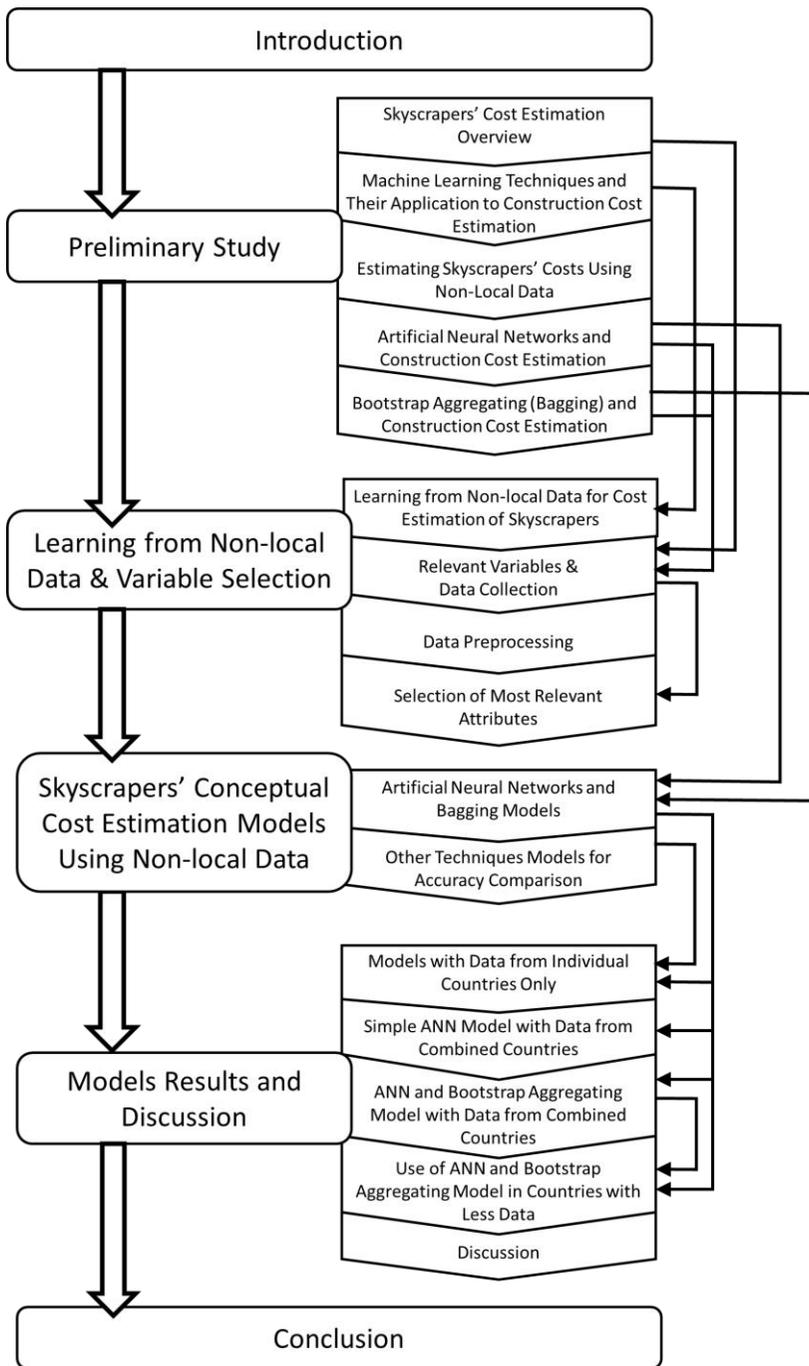


Figure 1.1 – Research Process

Chapter 2. Preliminary Study

This chapter reviews the main concepts necessary for the understanding of the studied models and the challenges associated to it. It starts by reviewing skyscrapers buildings, the definitions, the cost estimation practices and challenges associated to cost estimation of those projects. After, it presents a brief introduction on machine learning techniques and previous research on its use for construction cost estimation. Then it analysis possible issues that the proposed cost estimation might present, and endorses the reasoning behind the chosen proposed methodology. Later, it explains artificial neural networks and bootstrap aggregating, and their uses on estimation research.

2.1 Skyscrapers' Cost Estimation Overview

According to the Council on Tall Buildings and Urban Habitat (CTBUH), skyscrapers are high-rise buildings defined by a height of 200 m or more.

The common practice of cost estimation of those specific types of buildings is based on measuring material quantities from design development drawings, and then applying local unit costs to those quantities, as unit costs of materials, equipment and labor (Lee et al., 2011). In addition, other researchers notice the necessity of complex cost estimation models for accurate cost estimation of high-rise buildings (de Jong and van Oss, 2007; Newton, 2015).

Estimation of skyscrapers' construction costs presents various challenges that arises from the complexities associated with the colossal heights of those projects. De Jong and Wamelink (2008) highlight some extra challenges of skyscrapers that result in higher difficulties of accurate estimations, when compared to lower buildings. Those challenges are as follows:

- More complex plant and distribution systems to compensate for the increased heights;
- Extra structural needs due to the extra wind loading;
- Difficulties in the vertical movement of materials and labor;
- Amplified need of safety measures and liability concerns, increasing the associated risks.

All the above characteristics of skyscrapers projects affect the whole construction process and, consequently, the costs of such buildings. Lee et al. (2011) also notice that the higher the number of floors, the higher is the increase in area unit costs, with floors in greater heights presenting a greater increase in its costs. Furthermore, van Oss (2007) notices how the costs of those high-rise buildings tend to increase by 8% every ten floors, and de Jong (2008) acknowledges that costs models tend to account for those additional costs by applying a "height charge" factor.

Another main issue with estimation of skyscrapers is that cost models are generally based on historical data that can be compared based on specific location, floor size and other similarities in characteristics (Newton, 2015).

However, Newton (2015) highlights that “the problem with high-rise buildings is that there is a limited pool of data to use for an accurate cost comparison” what can jeopardize any attempt to estimate the costs of skyscrapers.

The models studied in this research attempt to account for those challenges in a simplified manner by using machine learning techniques, more specifically, neural networks and bagging. The next sections introduces the concepts involved in such techniques, as well the researched uses in construction cost estimations.

2.2 Machine Learning Techniques for Construction Cost Estimation

Machine learning techniques are techniques based on models that explore algorithms that can learn from data, and make consequent predictions on not yet seen data (Kohavi, 1998). Some of those techniques are Case Based Reasoning (CBR), Support Vector Machines (SVM), Decision Trees, K-Nearest Neighbors, Naïve Bayes, Artificial Neural Networks and some Ensemble methods among many more.

In the field of construction cost estimation, there are numerous studies on the utility of those methods, and on the advantages they bring, especially in increased accuracy of the estimations. An et al. (2007), Cheng et al. (2012), Cheng et al. (2013) used approaches based on Support Vector Machines to estimate construction costs related information, and all obtained promising results. Kim et al. (2004) and Ji et al. (2011) studied the use of Case Based Reasoning (CBR) techniques for construction cost estimation of buildings. Both studies obtained improvement in the accuracy of the estimation when compared to traditional methods, with Kim et al. (2004) obtaining average absolute errors of 4.81% with the CBR method. Other authors researched about the use of neural networks based models, also obtained encouraging high levels of accuracy (Hegazy and Ayed, 1998; Günaydın and Doğan, 2004; Kim et al., 2004; Cheng et al., 2010; Dursun and Stoy, 2016). Kim et al. (2004) achieved

mean absolute error rate of 2.97% of prediction of construction costs estimates, while Cheng et al. obtained 5.9% on conceptual estimates. All these studies show the utility that machine learning techniques can have on the field of construction costs estimation, and indicate the potential for use in the cost estimation of skyscrapers.

The next section introduces the use of Artificial Neural Networks (ANN) on construction estimation, focusing on the previous research on buildings cost estimations.

2.3 Estimating Skyscrapers' Costs Using Non-Local Data

Skyscrapers, as a specific category of high rise buildings, have a limited pool of data (Newton, 2015). These are rare projects, and so, their construction data can be very scarce. With exception of very few countries in the World, it is hard to find nations with more than 50 projects of that kind (CTBUH, 2017). The nature of machine learning requires as much data as is possible to get, while maintaining the quality and ability of that data to represent the reality. In despite of that, it is important to notice that research using machine learning for skyscrapers estimation exists. Li et al. (2008) has successfully performed Case Based Reasoning forecasting of construction schedule of skyscrapers. However, that research was fully based on Chinese skyscrapers, and, since China is the world's country with the most number of skyscraper buildings, it is an exception to the hardships of obtaining enough data from similar projects in the same location.

At this point, it is important to define that in this research "local data" refers to data originated within a same country, while "non-local" data refers to data originated outside from the studied country. This view is used on this research to account for the differences in construction practices, legislations, productivity, materials, architectural preferences and other aspects that can significantly differ among countries.

As stated above, in most places around the world, on a local point of view,

there is a shortage of skyscrapers' data, what would limit the efficient use of machine learning techniques. However, from a global point of view there are enough completed projects that maybe can contain relevant information, even those being from different countries. Thus, using the combined data of projects originated in different countries could provide enough data with some possible relevant information.

However, even if projects in different countries can contain relevant information, it is also important to take into account that possible problems might advent with the use of non-local data. Some of those possible problems are:

- Non linearity between projects in distinct countries;
- Differences in productivity, material costs, legal requirements, construction techniques and materials used;
- And bias towards the country with the biggest amount of data, inputting false information for other countries.

Possible solutions for of those problems are the use of artificial neural networks and the use of bagging methods. First, neural networks are strong at dealing with non-linearity, so they present great potential to solve, not only the natural non-linearity of skyscraper projects, but also the non-linearity between the construction environments of different countries. Second, for dealing with the differences in the construction aspects of the different countries, the addition

of a specific attribute, as the country name, might allow the neural network model to learn and recognize which parameters should be taken into account for each country, according to the information the data of the projects in that specific country contains. And third, bootstrap aggregating can increase the accuracy of neural networks, and can also serve to remove the possible bias towards the projects of a specific country.

Another possible solution would be to use international construction costs indexes to compensate for the difference between countries construction aspects, similarly to what is made for projects within a single country, but different cities. However those indexes can be hard to obtain and to forecast with high precision, and so, they were not taken into account on the scope of this research.

2.4 Artificial Neural Networks and Construction Cost Estimation

“Artificial neural networks are an information processing technology that simulates the human brain and the nervous system” (Boussabaine, 1996). The basic structure of neural networks consists of one input layer, followed by a hidden layer, and lastly, an output layer, see Figure 2.1.

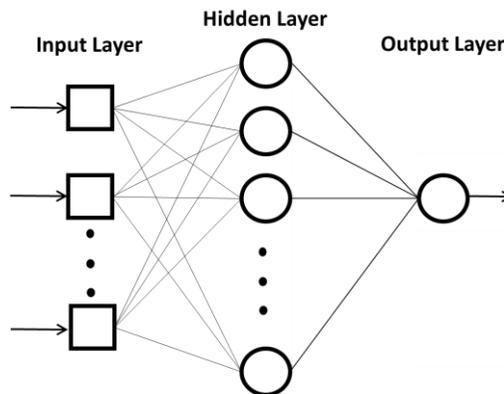


Figure 2.1 – Basic structure of artificial neural networks

Neural networks methods work by learning the relationships between the different parameters of the inputted previous data information (Doğan, 2005). The training of the network is made by establishing weights for the connections between neurons that will result in known output values, based on the input parameters. Those weights, then, retain the information learned through the data that can later be applied on new not yet seen data.

When compared to some other methods, neural networks possess the advantage of producing fairly accurate prediction results, even in situation when proper information is not available or when there are incomplete information or high complex problems (Rafiq et al., 2001). Another advantage is the strong capability of tackling non-linear relationships, as noted by Nerrand et al. (1993) that consider that ANNs can be seen as general non-linear filters.

As introduced in section 2.2, research has been made that testify the applicability of neural networks for construction cost estimation tasks. Specifically on cost estimation of buildings projects. Günaydı and Doğan (2004) used neural networks to perform early costs estimation of structural systems of buildings in Turkey. Despite being limited by the lack of specific design information at the early stages, they were able to obtain high accuracies of 93%. In South Korea, Kim et al. (2004) compared difference cost estimation models for buildings projects, obtaining the highest accuracy level with artificial neural networks modelling techniques. Similarly, Wang and Gibson (2010) also obtained better estimations when using neural networks. Cheng et al. (2010), using data from buildings in Taiwan, developed neural networks models with average error rates of approximately 6%. In Germany, Dursun and Stoy (2016) were also able to achieve acceptable levels of accuracy for conceptual cost estimations of buildings using ANN. Lastly, Sonmez (2011), Bayram and Al-Jibouri (2016) obtained great results when estimating construction costs of building project, in United States and in Turkey, respectively.

All those studies prove the efficacy of using artificial neural networks for forecasting construction cost estimation of buildings. And so, they also indicate the adequacy of using ANN for the estimation tasks of skyscrapers.

Aside from artificial neural networks, another powerful machine learning tool, many times used in association with ANN, is an ensemble method called Bootstrap Aggregating, or Bagging. The next section describes this method and the advantages it can bring.

2.5 Bootstrap Aggregating (Bagging) for Construction Cost Estimation

Bootstrap Aggregating or Bagging predictors is an ensemble method that generates multiple datasets that are used for training multiple models that then are aggregated in a single predictor (Breiman, 1996). This method works by re-sampling with replacement from the original provided dataset. “In sampling with replacement every sample is returned to the data set after sampling. So a particular data point from the observed data set could appear zero times, once, twice, or, more in a given sample” (Sonmez, 2011). The aggregation of the models is made by voting in case of categorical attributes, and by averaging the predictions in case of numerical attributes (Wang et al., 2012). Figure 2.2 demonstrates the overall process of bagging.

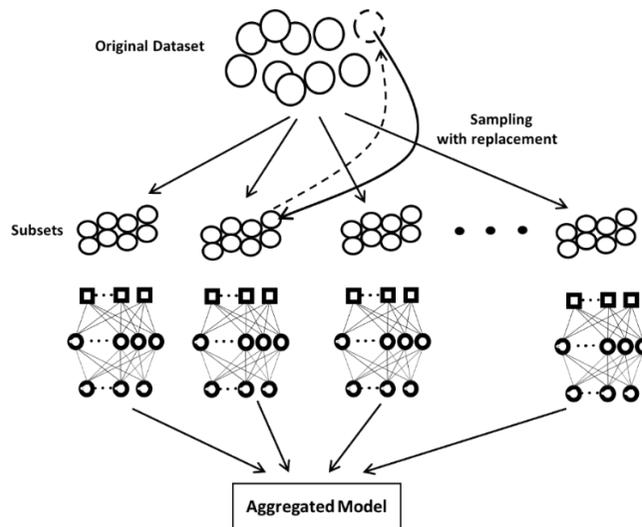


Figure 2.2 – Bootstrap aggregating process

As demonstrated by Breiman (1996) the bootstrap aggregating method is capable of both improving the accuracy of estimates and reducing possible bias on small databases. Bias being a measure of how closely a model can be approximated to a target function, it indicates the error in estimation of a specific model (Lutu, 2010). In addition, Tsai and Li (2008) also notices that this method can be used to improve predictions of neural networks when only sparse data is available. These characteristics make bagging a promising resource for dealing with cost estimation in the conceptual stages, when data tends to be very limited.

On the field of building construction costs estimation, Wang et al. (2012) concluded that neural networks models that are built from bootstrap aggregating have better accuracy and are more robust than models built only of single neural networks. This results shows the possible use of bagging in skyscrapers cost estimation for compensating for the high variation expected from the complexity related to those buildings, especially in the scenario of buildings from different construction backgrounds in distinct countries.

All the so far mentioned research utilizing neural networks or bagging were able to achieve high accuracy using data from projects within a specific country. However, in the specific case of skyscrapers, data within only one country is, most likely, limited, restricting the possible use of such techniques. This research approaches that problem and investigates the use of additional non-local data with the use of artificial neural networks and bootstrap

aggregating as a possible method to compensate for the disparities between projects from different countries and to accurately learn from it.

2.6 Summary

Skyscrapers buildings present considerable challenges for their cost estimations. Among them, it is highlighted the scarcity of similar projects data in the same country. Machine learning methods have been proven in previous research to be an efficient tool for dealing with conceptual cost estimations, nonetheless, the use of these tools are highly dependent on sufficient quality data from previous projects. Considering that fact, the use of extra added data from other countries for cost estimation was presented as a possible approach to this problem. However it was recognized that the use of data from projects in different countries can contribute with non-linearity, noisy data that does not reflect the reality of the projects being estimated, and possible bias towards data dominant countries. Both artificial neural networks (ANN) and bootstrap aggregating (Bagging) were presented as possible solutions for those issues.

In light of what was presented in this chapter, the next chapters introduce the development of a cost estimation model based on the use of combined data from different countries and artificial neural networks and bootstrapping aggregating techniques.

Chapter 3. Learning from Non-local Data & Variable Selection

This Chapter introduces the idea behind the proposed model development for conceptual cost estimation of skyscrapers, based on the use of combined data from different countries and artificial neural networks with bagging. Also, it regards the selection of attributes and collection of relevant data for the modelling steps in Chapter 4.

First it introduces a diagram to explain what is expected from the modelling processes using the addition of non-local data. Then, the data collection stage, showing the selection of relevant attributes, as well as information about the characteristics of the collected data and the sources where they were collected from. Then it details the preprocessing processes used on the data and the selection of the final most attributes used for the modelling stage.

3.1 Learning from Non-local Data

The overall idea behind the proposed use of additional non-local data for the construction cost estimation of skyscrapers is represented in Figure 3.1.

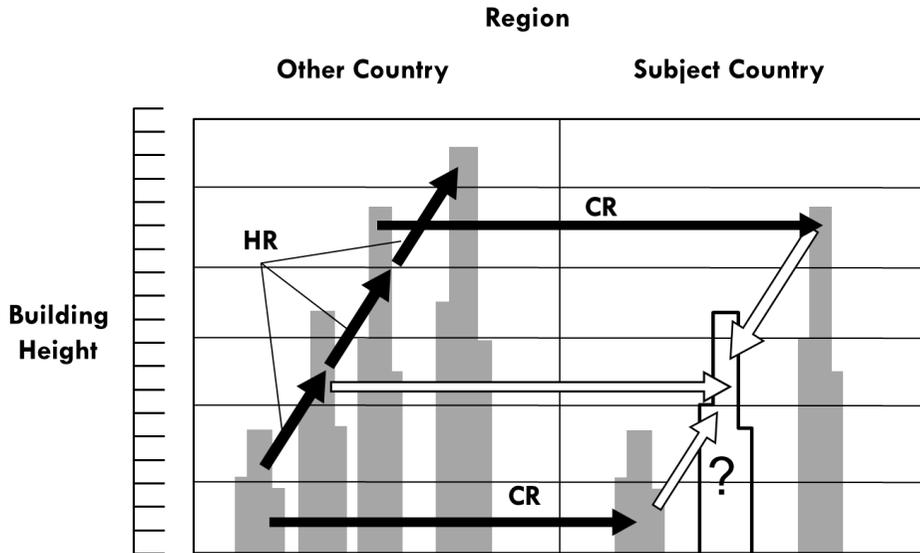


Figure 3.1 – Strategy behind the use of additional non-local data

The diagram in Figure 3.1 represents both the country subject of the estimation, as well as other country from which additional skyscrapers projects data will be acquired from. As shown, the subject country has less completed projects than the other country, and it has a building to be estimated. As presented before in the review of previous research, one of the complexities involved with the estimation of these kind of projects is that different heights present different associated construction costs. However, due to the lack of many completed projects, there is not enough information available within that

country for the algorithm to learn the differences and relations among different heights of buildings, what could bring imprecise cost relations between different heights. On the other hand, the other country, possesses a higher amount of completed buildings, and the information contained in its database should allow the algorithms to learn those relations (represented as the arrows HR). By using a combined dataset from the subject country and the other country, an algorithm could learn the cost relations between similar projects in the two different locations (CR arrows). And by the higher number of projects of different sizes in the other country, it could learn about local costs relations that different heights of buildings have. With those two kinds of relations, it would then be possible to estimate relations in the subject country about different heights, or about cost proportion of similar height buildings in the other country (represented as the white arrows). Ideally, with the addition of other “other” countries, any gap between different sizes of projects could be filled, and more precise relations among costs in different projects could be learnt. Then, through the relations of other countries’ projects with the subject project, a new building cost could be accurately estimated, even with the lack of similar projects on that specific location.

The models on this study were based on the presented above. As to develop those models and assess their accuracy, the next sections in this chapter present the selection, collection and preparation of the data from different countries.

3.2 Relevant Variables & Data Collection

In order to develop the conceptual cost estimation models, first it is necessary to collect relevant data for the task. The primary source for the skyscrapers data is “The Skyscraper Center” a global database of tall buildings ran by the Council on Tall Buildings and Urban Habitats (CTBUH). This database provides information about skyscrapers in many different countries. Possible useful attributes to be collected were selected based on the available variables from that main source (CTBUH), on the challenges associated to cost estimation of skyscrapers, as wind forces, as presented in Chapter 2, and on the previous successful research on conceptual construction costs estimations of buildings and on research on skyscrapers.

From the CTBUH database the attributes selected were: architectural height, height to tip, occupied height, number of floors above ground, number of floors below ground, number of elevators, top elevator speed, tower gross floor area, development gross floor area, number of apartments, number of hotel rooms, number of parking spaces, structural materials used, the building function and completion year.

Some of those attributes are constant among the used attributes for cost estimation models of construction projects, as floor area, number of floors, completion year, number of apartments and building function (Kim et al., 2004; Li et al., 2008; Cheng et al., 2010; Arafa & Alqedra, 2011; Dursun and Stoy,

2016).

From the research about skyscrapers schedule estimation (Li et al., 2008), aside from the previous CTBUH attributes, the attributes green building certification, real GDP of the city, labor productivity and wind speed were selected. And as Kim et al. (2004), this research used the attribute of finishing grade, aside from the previous mentioned ones.

Appendix A shows all the selected attributes and their descriptions.

After defining the important attributes, the data about skyscraper projects was collected from 5 distinct countries, namely, United States of America, China, United Arab Emirates, Australia and Singapore. As the product of the data collection, data of 124 projects was collected. USA was the country with the majority of collected project data with 50 projects in total. China and UAE followed with 28 and 26 projects, respectively. Both Singapore and Australia represent small skyscrapers databases with 9 and 11 projects, respectively.

Data was collected from buildings completed between 2000 and 2017, of three possible uses, residential, office and hotel, and of three possible structural materials, concrete, steel and composite. The data regarding the attributes of the building, as height and structural materials, were collected from the CTBUH database. The remaining attributes were collected from distinct open source databases. The construction costs were collect out of an extensive investigation of specialized media articles, or from direct contact to personnel involved in those projects.

Among the skyscrapers projects a small percentage is made of building between 150-200 m, in order to improve the ability of the model in correlating different levels of buildings' height. The range of the collected buildings height goes from 152 m to 828 m. The Tower gross floor area ranges from 16,763 m² to 459,187 m². Some of the collected attributes as Development gross floor area, or number of parking spots present a considerable amount of missing data, however, those variables were still left in the datasets, since neural networks can still perform with dataset with missing data.

After the collection of the projects' data, the 5 created datasets went through preprocessing for better application of the model.

3.3 Data Preprocessing

Figure 3.2 shows a summary of all the steps of the data preprocessing. Since some attributes of the data were collected from different sources, a considerable portion of the data presented distinction in the unity measures of some numerical attributes, or in the categories of categorical attributes. In order to solve that issue, first all numerical attributes were transformed to a unique unit measurement. More specifically, all wind design speed data was converted to meters per second, and all costs and currency related variables were converted to US\$, according to the average exchange rates at the completion year of the projects.

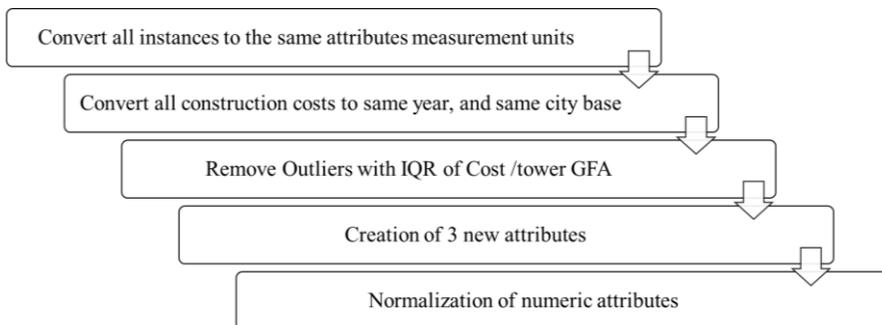


Figure 3.2 – Data Preprocessing

The categorical attribute of Green Building Certification presented a special challenge because of the different certifications applied in different parts of the world. To solve that issues, all different green building certifications were standardized in a 5 categories range from 0 as no certification to 4 as the highest

level of certification, or correspondent to a LEED Platinum. All the other levels of certifications were categorized as a proportion of the 0 to 4 range.

The next step of preprocessing involved bringing all the different costs to a common city and year base, within each country, in order to compensate for cities' construction costs differences and for the time adjustment of those costs. With that purpose, local construction cost indexes were used to convert the construction costs of the buildings projects to the specific base of each individual country. All the projects' costs were converted to costs on the base year of 2010. Each country had a specific base city for which the costs were converted to. New York City for USA, Beijing for China, Dubai for UAE, Sydney for Australia and Singapore as Singapore.

Once all attributes were standardized inside each country, a process of outliers removal was made in each isolated country dataset. The purpose of the outliers removal here is not to remove proper noisy data that can be meaningful on the skyscrapers scenario, but to actually remove possible noisy data due to errors in the input of instances in the database. A new variable of Cost per Tower Gross Floor Area was created for that analysis of outliers. By interquartile method, in total, 8 outliers were removed, 3 in USA, 2 in China, 1 in UAE, 1 in Australia and 1 in Singapore.

After the removal of outliers, 3 new attributes were created from combinations of the collected variables. Those attributes were created to check for possible better fit in explaining the construction costs, than the attributes

they originated from. The table 3.1 shows the 3 new attributes created, namely, real GDP per person, real GDP as base of the biggest GDP inside the country and real GDP per person as base of the biggest GDP per person inside the country.

Table 3.1 – Created attributes from previous existing attributes

New attribute	Description	Creation rule
real GDP per person	Real GDP per person of the city	Real GDP divided by population
real GDP, country's base	Real GDP of the city, compared to the biggest city GDP among the cities in the same country	Real GDP of the city divided by the biggest GDP among the cities in the country
GDP per person, country's base	Real GDP per person of the city, compared to the biggest city GDP among the cities in the same country	Real GDP per person of the city divided by the biggest GDP per person among the cities in the country

In the last stage of preprocessing, in order to perform the artificial neural networks models, all numeric attributes were normalized in a scale of 0 to 1, and then, all nominal and categorical attributes were converted to binary attributes. After the preprocessing, the most relevant attributes were selected.

3.4 Selection of Most Relevant Attributes

The selection of the most relevant attribute was made with the purpose of avoiding overfit. Figure 3.3 shows the process that lead to the selection of the final attributes.

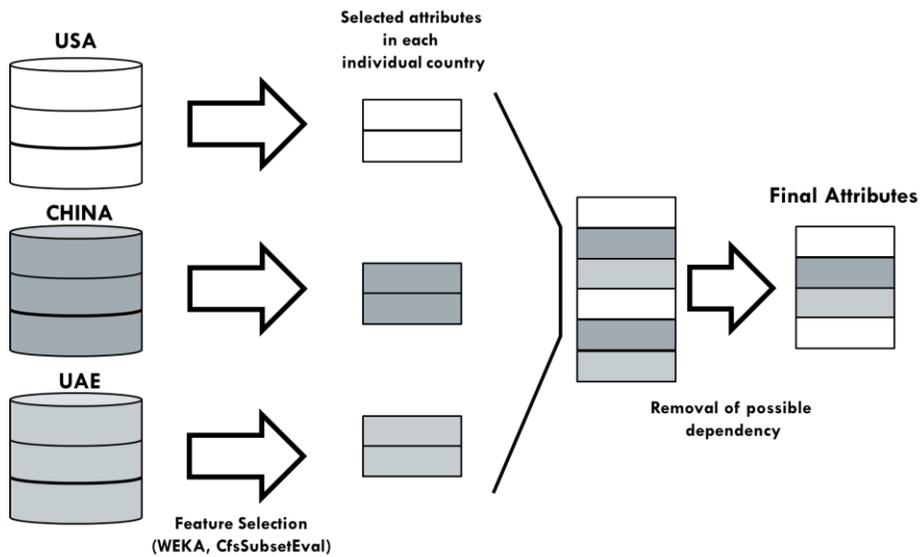


Figure 3.3 – Process of selection of final modelling attributes

The feature selection was made using WEKA's (Waikato Environment for Knowledge Analysis) built-in CfsSubsetEval algorithm by Bestfirst in the countries subsets of USA, China and UAE, since they possess more significant amounts of data. This algorithm evaluates the attributes by checking the individual predictive ability of each individual attribute, while taking into account the redundancies between the different features. In the end, the algorithm chooses features that are highly correlated to the class, but that

present low correlation among themselves. The results obtained for each country are contained in Table 3.2.

Table 3.2 – Subset Evaluation results for the individual countries

Country	USA	CHINA	UAE
Selected Attributes	Real GDP	Real GDP per capita	Development GFA
	Tower GFA	Height: Architectural	Finishing Grade
	Height: Architectural	Structural Material	Height to tip
	Structural Material	Number of parking spaces	Floors Above Ground
		Tower GFA	

In order to be applied in the models with combined data from different countries and to be used for comparison purposes in all the other models with individual countries datasets, the selected attributes for each individual country datasets were combined in a single set of attributes and scanned for possible dependent variables.

Three sets of variables were presenting possible dependency, those being:

- Tower GFA and Development GFA;
- Height:Architectural, Height to tip and Floors Above Ground;
- Real GDP and Real GDP per capita.

The difference between Tower GFA and Development GFA account for only the extra areas of the development that are not direct part of the skyscrapers

towers, and specific Development GFA for each skyscraper can be hard to define in developments composed by two or more skyscrapers. The variable for heights are related to the number of floors above grounds, when considering that there is a minimum height for each floor, so the more floors the increased height. GDP and GDP per capita are directly related, since GDP per capita is calculated as the ratio GDP per population. From each of those groups, only one feature was selected to make part of the final attributes (Tower GFA, Height: Architectural and Real GDP). The final selection of attributes resulted in a dataset structure composed of eight attributes:

- Tower GFA;
- Real GDP;
- Finishing Grade;
- Height: Architectural;
- Number of parking spaces;
- Structural Material;
- Country (added for the application on the combined data only);
- Construction Cost.

3.5 Summary

This Chapter presented the idea behind the models and the whole process of the selection of relevant attributes and data collection. Initial attributes were selected based on availability and previous research on the fields of cost estimation and skyscrapers. The data of 124 projects in five different countries was then, collected from the CTBUH skyscrapers database and other media sources specialized in the construction field. In sequence, the rough database was preprocessed to standardize all the attributes that came from different data sources, and to eliminate outliers, and to prepare the datasets for the modelling processes. With the prepared data, the most relevant attributes were found by feature selection. Chapter 4 introduces the development of the models for analyzing the use of additional projects data from other countries.

Chapter 4. Skyscrapers' Conceptual Cost

Estimation Models Using Non-local Data

After the collection of data from the 5 countries, and the final selection of attributes to be utilized in the modelling processes, the models for analyzing the efficacy of adding data from projects in other countries were developed.

Two main kinds of models were made. First, models based on ANN and Bagging were developed as the main focus of this study. Later, models based in other methods were also performed, in order to provide a base for comparison of accuracy.

4.1 Artificial Neural Networks and Bagging Models

The main point of this research is to analyze the use of an artificial neural network and bootstrap aggregating model to estimate conceptual construction costs, using additional non-local data. The framework of the development of the models using ANN or Bagging is presented in figure 4.1.

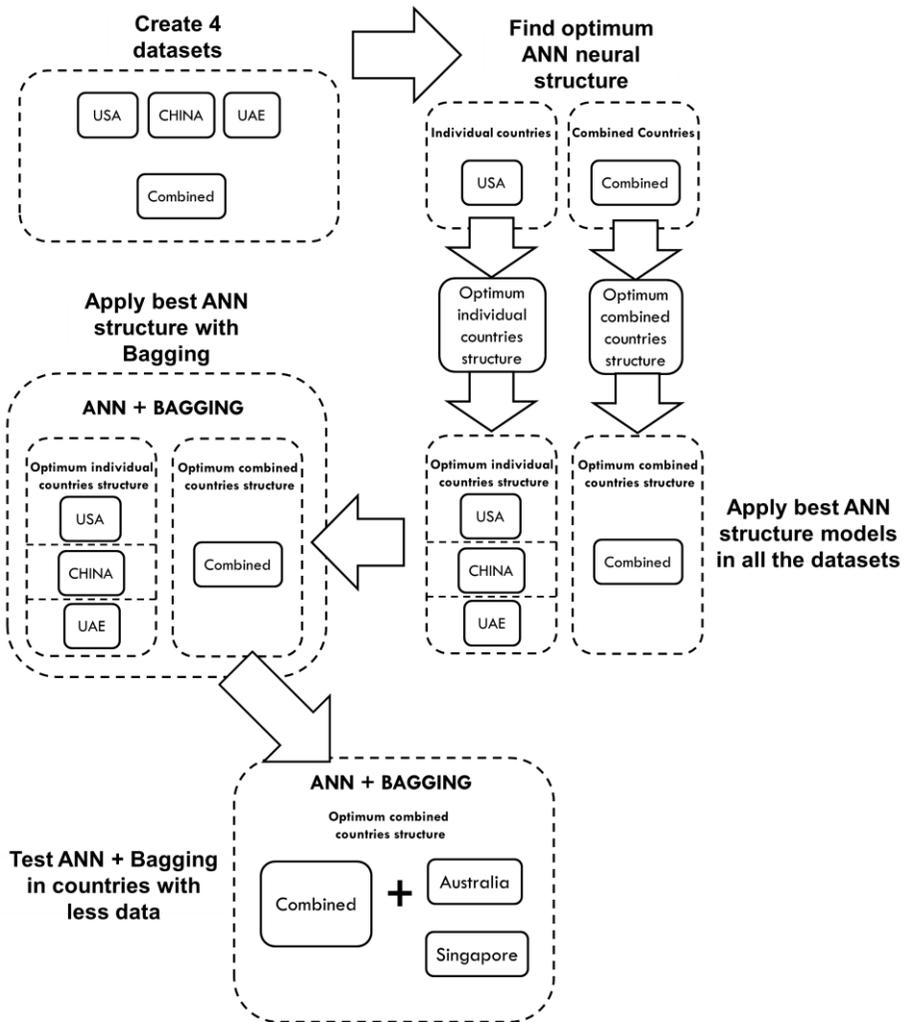


Figure 4.1 – Framework of the development of ANN and Bagging models

Initially, the collected data is arranged in different datasets to be used in the different models to be performed. Then the optimum neural structures for both individual and combined countries datasets were investigated. Once determined the best structures, those were applied in simple ANN models for

all the datasets. After, the same optimum structure was used to develop the ANN and Bagging models for all the datasets, individual and combined countries. Then, with the objective of testing the model in countries with less data, if the ANN and Bagging model for the combined countries dataset showed better results than the use of the individual datasets, the same structure was used with a new dataset of all the countries, including Australia and Singapore.

The six individual countries models were performed to later serve as comparison for the combined dataset models. Those six models were the combination of the three individual country datasets of USA, China and UAE, with two possibilities of machine learning algorithms, simple ANN (without bagging) or ANN plus bagging. The simple artificial neural networks model (without bagging) with the combined data of the three main countries, USA, China and UAE had the purpose to show if the addition of extra non-local data brings an increase in the accuracy of the estimations for the projects in the individual countries. This model also served as a comparison model to assess the effects of bagging when used with the combined data.

The ANN plus bagging model was performed with the combined dataset to assess for increased accuracy compared to the previous models. To assess the extent to which the combination of non-local data, ANN and bagging can bring accurate results for skyscrapers conceptual cost estimation, the model with same parameters, but using a new combined dataset made of the full database, USA, China, UAE, Australia and Singapore was performed. The accuracy of

that model was then checked for Australia and Singapore projects, as representatives of countries with smaller amount of skyscrapers projects.

4.1.1 Modelling Datasets

The collected data was divided in 5 main datasets:

- USA;
- China;
- UAE;
- Combined countries (USA, China and UAE);
- Full dataset (USA, China, UAE, Singapore and Australia).

These datasets were used for the models with ANN and Bagging, and also for the models of other methodologies used for comparison of accuracy.

The USA, China and UAE datasets are made by only the data of projects within the individual country that gives the name to the dataset. While, the combined countries dataset is made of the combination of those countries' projects, and the full dataset is the combined dataset plus the addition of the data of projects from Singapore and Australia.

Once those datasets were prepared, the models were developed according to what is presented in the next sections.

4.1.2 ANN Models

The artificial neural networks models were developed using the MultilayerPerceptron algorithm in WEKA.

In order to analyze the proposed use of non-local data, two basic optimum model structures needed to be found. One structure for estimating models of single isolated countries, and one structure for estimating models of combined countries. For the definition of the single countries model structure, the dataset of USA was used, due to the higher amount of instances. In the case of the combined countries model, the combined dataset composed by all the instances of USA, China and UAE was used.

Both datasets, USA and Combined, were analyzed in different artificial neural networks model configurations. The analysis was made in WEKA, using the MultilayerPerceptron algorithm and cross-validation with 10 folds. Various different neural structures were tested varying the number of neurons in the hidden layer to find the best fit for each of the datasets. A summary of the results is presented in Table 4.1, with the values for the root mean squared error (RMSE), relative absolute error (RAE) and root relative squared error (RRSE).

Table 4.1 – Definition of best neural structures

Neurons in the hidden layer		6	8	9	10	11	12
USA isolated	RMSE	2.95E+08	2.99E+08	3.19E+08	2.51E+08	2.77E+08	3.04E+08
	RAE	71.21%	67.15%	63.77%	59.19%	63.83%	64.97%
	RRSE	61.78%	62.72%	66.80%	52.54%	57.97%	63.81%
Combined countries	RMSE	3.22E+08	3.12E+08	2.78E+08	2.88E+08	2.77E+08	3.58E+08
	RAE	55.29%	53.76%	50.07%	49.27%	49.98%	55.96%
	RRSE	60.94%	58.97%	52.61%	54.42%	52.50%	67.77%

The optimum artificial neural network structures were defined according to the lowest values of RMSE, RAE and RRSE. For single countries model the optimum structure was 7-10-1, and for the combined model 8-11-1.

After the definition of the basic structures, the 4 simple ANN models were developed. Those using the individual USA, China and UAE datasets, and the Combined dataset. All those models were developed and assessed using WEKA and MultilayerPerceptron, with learning rate of 0.3, 500 training times, and cross-validation with 10 folds. The models with datasets of individual countries used a 7-10-1 structure, and models with combined datasets used an 8-11-1 structure. Then, the ANN plus Bootstrap Aggregating models were developed.

4.1.3 ANN and Bootstrap Aggregating Models

The ANN with Bagging models were developed using the same parameters for the simple ANN models presented in the previous section.

Bootstrap Aggregating (Bagging) is an ensemble method that learns from aggregated prediction of multiple generated datasets. Basically, it creates new datasets by sampling with replacement from the provided original dataset. On this case of numerical estimation, bagging aggregates the prediction of the different subsets models by averaging the estimations.

The main parameter for a bagging predictor is the amount and size of subsets. On this study, for the bootstrap aggregating models, 100 subsets of same size were used as the bagging parameters, and the training and testing were also cross-validated with 10 folds. MultilayerPerceptron was used as the ANN base classifier.

The ANN base classifiers kept the same structure of the simple ANN of the previous section. The models with datasets of individual countries used a 7-10-1 structure, and models with combined datasets used an 8-11-1 structure for the ANN, with learning rate of 0.3, 500 training times.

Initially four models were performed with the ANN and Bagging method. 3 models of each individual country (USA, China and UAE), and 1 model for the Combined dataset. Following the results of the combined dataset model, a fifth model was made using the combined data of all the countries including

Australia and Singapore. In total, 9 models of ANN or Bagging were performed. Table 4.2 contains a summary of the models and which dataset and algorithm is used in each.

Table 4.2 – Summary of developed models

Datasets	ANN	ANN + Bagging
USA	Model 1	Model 5
CHINA	Model 2	Model 6
UAE	Model 3	Model 7
COMBINED (USA, CHINA,UAE)	Model 4	Model 8
COMBINED (all countries)		Model 9

Appendix B shows the input variables and their values for five examples of projects that were estimated by the model using non-local data with ANN and bagging. Each example is a project in a different country (USA, China, UAE, Australia and Singapore).

In order to have a base comparison of the performance of the models presented in this section, extra models following different methods were developed. Those models development are presented in the next section.

4.2 Other Techniques Models for Accuracy Comparison

The main objective of this section is to develop models to serve as a basis points for assessment of the accuracy of the ANN and Bagging models. 3 different methodologies are used, those being:

- Average Cost per Area;
- Linear Regression;
- Case-Based Reasoning.

4.2.1 Average Cost per Area

As the simplest base comparison, this study developed estimates using the average cost per area of the buildings in the datasets in each isolated country. The skyscrapers were separated according to its structural material. Then, for each structural material the cost per area was calculated as the total construction cost divided by the tower gross floor area. Each building cost was then estimated according to that average cost.

4.2.2 Linear Regression

Linear regression models were developed using individual countries datasets, as well as the combined dataset. The models were developed in WEKA using the LinearRegression algorithm. The attributes used were the same final selected attributes for all the different datasets. Even in cases when one of the attributes is not significant enough for a specific country, the attribute was still kept for the regression.

The linear equations obtained for USA, China, UAE and Combined datasets are presented in Table 4.3.

Table 4.3 – Summary of Linear Equations

Dataset	Linear Equation
USA	= 0.2162 * Height: Architectural + 0.5508 * Tower GFA - 0.0027 * Real GDP +0.1548 * Class=b + 0.0549 * # parking spaces -0.029 * material=concrete + 0.0433 * material=steel + 0.0044 * material=composite + -0.1306
China	= -0.3208 * Height: Architectural + 0.7056 * Tower GFA - 0.1054 * Real GDP +0.2267 * Class=b -0.1018 * # parking spaces -0.0577 * material=composite + 0.177
UAE	= 0.2881 * Height: Architectural + 0.3366 * Tower GFA - 0.0318 * Real GDP +0.1069 * Class=b + 0.2723 * # parking spaces -0.011 * material=concrete - 0.1142
Combined	= -0.0258 * Country=CHINA -0.04 * Country=UAE + 0.0506 * Country=USA -0.0805 * Height: Architectural +0.6771 * Tower GFA + 0.0678 * Real GDP + 0.1448 * Class=b +0.0686 * # parking spaces -0.0196 * material=concrete -0.1024

4.2.3 Case-Based Reasoning

“Case-based reasoning Case-based reasoning (CBR) is a continuous dynamic learning process that originated from artificial intelligence. It solved new problems by reusing experience of similar previous cases” (Li et al., 2016). It works by calculating the similarity of the new project case to previous completed projects. Then, it ranks the past projects by the similarity to the new one. Then it uses the previous similar projects as recommended solutions for the new problem.

On this research, CBR models were developed for the individual countries datasets. Because of the nature of CBR, using it with the combined dataset would result in only retrieving same country projects, or retrieving projects from other countries, that would results in a mistaken cost solution.

CBR is made by steps. First, it has a retrieval step where it is calculated the weights for the different attributes. On this study, the weights of the different attributes were calculated based on the linear regression analysis of the previous section. For each different dataset, the correspondent linear equation as used as the weight for the projects' variables. Then the similarity between cases was calculated following those weights, and all cases were ranked according to the highest similarity. The percentage similarity for the numerical attributes was calculated using the Euclidean distance as shown in Equation 1.

$$S = \left[1 - \sqrt{\frac{\sum_{i=1}^n W_i^2 \times (N_i - P_i)^2}{\sum_{i=1}^n W_i^2}} \right] \times 100, \quad (\text{Equation 1})$$

Where, S is the percentage similarity, w_i is the weight of attribute i , N_i is the value of the attribute i on the new case and P is the value of attribute i on the previous case.

In the case of nominal attributes, the distance was given as 1 if the attributes were the same, and as zero if the attributes were different.

The most similar cases were reused and proposed as possible solutions. In this research, 4 different amounts of similar cases were tested. For each project, the most 1, 3, 5 and 10 similar cases were retrieved and used to check for the best prediction. In the cases of more than one similar retrieved case, the proposed solution was the average cost of the cases costs. Similar to Kim et al. (2004) that obtained great results for cost estimation with CBR, this research used no case revision, keeping only the direct solution from similar cases.

4.3 Summary

This chapter presented the development of the conceptual cost estimation models. From the attributes selected in Chapter 3, 9 models using ANN or Bagging with individual countries were developed. Those models were developed with the intend to assess the ability of those methodologies to learn from the different countries information and to compared the improvement that the different datasets and techniques (ANN and bagging) can bring to the estimation process. The optimum neural networks structure was defined previously to the development of each of those models, and the models were all teste by cross folding with 10 folds.

Other extra models were develop as basis comparison of the accuracy for the ANN and bagging models. Those models were averaged cost per area, linear regression and case-based reasoning. The next chapter presents the accuracy results of all the models, and discusses the comparison and implications of those results.

Chapter 5. Models Analysis

Following the modeling parameters stipulated in the previous Chapter, nine models ANN or Bagging models were performed, plus models based in different techniques used as base comparison. The accuracy of the results obtained on those models are presented in this chapter and analyzed about the efficiency that adding non-local data and using ANN and bootstrap aggregating can bring to the field of cost estimation of skyscrapers. All models are analyzed for their accuracies measured as the Mean Absolute Percentage Error (MAPE), defined as given in Equation 2.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{Ai - Ei}{Ai} \right|, \quad (\text{Equation 2})$$

Where Ai is the actual construction cost of project i , Ei is the estimated cost of project i , and n is the number of estimated projects.

The analysis is made by first presenting the accuracy results obtained for the other techniques models of section 4.2, used as basis for later comparison. Then, the models with individual countries datasets are analyzed, then the use of combined data with ANN is also investigated. Later, the main model of combined data with ANN and bagging is studied. And lastly, based on the results of the main model, the combined model of all countries, including the two countries with lack of data, is analyzed to look for evidences of possible use of the proposed approach on countries with just a few skyscrapers' data.

5.1 Models for Comparison

As presented before, 3 different methodologies were used as basis of comparison for the estimation of the conceptual construction costs of skyscrapers buildings. The results of each methodology is presented in the order of average cost, linear regression, and then Case-based reasoning.

The error rate obtained by the average cost per area is presented in Table 5.1 below, according to each of the individual countries datasets.

Table 5.1 – Accuracy of average cost/area

Country	Average Cost/Area
	MAPE (%)
USA	56.62
China	62.42
UAE	42.36

The MAPE of all the three countries was high, with the smallest error rate being from UAE of 42.36%. This result shows how the construction of skyscrapers have complexities involved that are not easily taken into account with very simple methods.

Following, the linear regression was used on the individual countries and on the combined dataset. The accuracy results obtained are presented in Table 5.2.

Table 5.2 – Accuracy of Linear Regression

Country	Linear Regression with individual datasets	Linear Regression with combined dataset
	MAPE (%)	MAPE (%)
USA	44.97	36.98
China	32.36	34.90
UAE	64.47	69.63
All Countries Combined	-	44.75

It can be seen that linear regression showed an overall improved accuracy when compared to the average cost/area. However, in the case of UAE, the linear regression represented a decrease in accuracy.

The addition of other countries data, as studied in the combined dataset, resulted in a lower accuracy. As it is, the addition of non-local data seems to have input noise that jeopardized the estimation, both China and UAE had worst accuracy with the combined dataset. The only country that obtained benefit from the addition of non-local data was USA, which is the country with the biggest amount of instances. The bad results of the addition of data from other countries are probably due to non-linear relations between projects in distinct countries, relations that the linear regression is not capable of dealing with.

The CBR was performed only on the individual countries. For each country 4 different amount of similar cases reused were performed to find the one with the best accuracy. The results obtained are shown in Table 5.3.

Table 5.3 – Accuracy of CBR

Country	CBR 1 similar case	CBR 3 similar cases	CBR 5 similar cases	CBR 10 similar cases
	MAPE (%)	MAPE (%)	MAPE (%)	MAPE (%)
USA	27.94	23.14	23.17	28.97
China	47.41	26.27	23.59	26.30
UAE	29.22	29.38	31.55	31.89

The best results were obtained as the average of the costs of 3, 5 and 1 similar cases to USA, China and UAE, respectively. The results obtained are considerable good, with all the countries obtaining error rates of less than 30%. Also, the CBR showed, so far, the best accuracy for all the methods for basis comparison. With the results for comparison, the next sections focus on the artificial neural networks and bootstrap aggregating models.

5.2 Models with Data from Individual Countries

This section presents the models that used datasets of isolated individual countries. Three countries were used for it, USA, China and UAE. Those countries were chosen due to relatively high amount of available data they possess. Two basic models were used. One model of a singular artificial neural networks, and one model of bagging with artificial neural networks. The results from the 10 fold cross-validation are presented in Table 5.4.

Table 5.4 – Results of individual countries models using ANN

Country	ANN		ANN + Bagging	
	MAPE (%)	percentage below 30% error (%)	MAPE (%)	percentage below 30% error (%)
USA	29.63	59.57	26.35	65.95
China	30.56	65.38	20.76	73.07
UAE	37.31	60,00	29.05	64.00

As can be seen, the combination of bagging with the artificial neural networks resulted in higher accuracy for all the countries when compared to the models with only ANN. Not only bagging was able to improve the values of MAPE, but also it increased the amount of projects that were estimated below the acceptable conceptual cost estimation range of 30%, for all the studied countries.

Comparing with the basis models, both ANN and ANN plus bagging showed better results than the linear regression and average cost methods. However, the simple ANN didn't reached the accuracy level of the CBR models.

The bagging and ANN showed similar error rate to the CBR models. But it had slight improvements for China and UAE. It is interesting to notice how the addition of bootstrap aggregating seemed to have influenced more the models of countries with smaller amount of available data. China and UAE had approximately, reduction of 10% and 8% on the error rate with the addition of bagging, while USA, with the largest available dataset, had only a modest improvement of, approximately, 3%. This might happen, because in smaller datasets, there is more potential for a single instance to influence the whole data more strongly than when compared to bigger datasets. And bagging should be able to reduce that bias, and make the models more generalized for the other projects in the dataset. The main purpose of those models in this research is to be used as references for the models presented in the next sections.

5.3 Simple ANN Model with Data from Combined

Countries

The next performed model was the model using the combined dataset of three countries, USA, China and UAE, using only ANN, and no bagging.

The results obtained in that model for the combined dataset and the individual countries are shown in Table 5.5, below. The previous results of ANN model with single countries datasets are also presented to allow for easier comparisons.

Table 5.5 – Results for ANN models for combined and single countries datasets

Country	ANN single countries dataset		ANN combined countries dataset	
	MAPE (%)	percentage below 30% error (%)	MAPE (%)	percentage below 30% error (%)
All countries combined			29.57	58.16
USA	29.63	59.57	29.94	55.31
China	30.56	65.38	27.83	65.38
UAE	37.31	60,00	30.67	56.00

The results obtained with the combined data model indicates the potential of using non-local data in order to improve cost estimation accuracy. The combined MAPE resulted from the combined model (29.57%) is smaller than any of the individual MAPEs of the countries in the isolated datasets.

With exception of United States, that obtained a similar accuracy result, both China and UAE achieved better results with the combined dataset. China

went from a MAPE of 30.56% in the single countries model, to a MAPE of 27.83% in the combined countries model. And UAE had an impressive decrease of almost 7% in the average estimation error, reaching 30.67%.

Comparing with the linear regression using the combined countries dataset, the results obtained from ANN showed how it is possible to obtain useful information from non-local projects. While linear regression was not able to acquire that useful information, ANN showed to be able to adapt for that combined dataset scenario. This fact, reinforces the idea that projects in different countries, probably, present nonlinear relations that the artificial neural networks models are able to learn from.

The results obtained by the simple ANN using combined countries dataset are still not as accurate as the estimations of CBR with the individual countries. Since CBR isolates directly similar projects, this results show that the additional data might be introducing noise that makes the algorithm to have bigger estimation errors.

Also, even though, the MAPE measures presented a tendency to decrease with the addition of other countries data, the percentage of projects below the cut mark of 30% error, decrease in both USA and UAE, and kept constant in China. That shows that the addition of other countries data by itself was not sufficient, at this point, to increase the amount of correctly predicted projects. It probably just improved the accuracy of the estimations of the projects, but not in an enough significant manner.

5.4 ANN and Bootstrap Aggregating Model with Data from Combined Countries

After the previous models results, the main model based on the combined dataset of USA, China and UAE, and using artificial neural networks and bagging was performed.

The results from that model are presented in Table 5.6, in combination with the results obtained for the combined ANN only model.

Table 5.6 - Results for combined, ANN + bagging and ANN only

Country	ANN only combined dataset		ANN +bagging combined dataset	
	MAPE (%)	percentage below 30% error (%)	MAPE (%)	percentage below 30% error (%)
All countries combined	29.57	58.16	20.83	73.47
USA	29.94	55.31	23.50	65.96
China	27.83	65.38	17.35	76.92
UAE	30.67	56.00	19.44	84.00

The proposed model of using data from different countries with ANN and bootstrap aggregating obtained the best results among all the other ANN models. Also, it was the model with the biggest accuracy improvement in comparison to the other models. The overall MAPE considering projects from all the three countries reached the mark of 20.83%, an improvement of more than 3% in accuracy, from the best MAPE value from the ANN + bagging only combined dataset for Chinese projects.

In the individual countries, this method using combined data was able to significantly improve the accuracy of all the three countries, when compared to the same algorithms applied for the single countries datasets. USA had an improvement of almost 3%, China of 3.4% and UAE of impressive 9.61%.

And when compared to the ANN only combined model, not only the addition of bagging increase the accuracy considerably, but also it had a higher increase in accuracy than just the use of bagging had on the single country datasets. This indicates that the combination of bagging and addition of data from different countries possess a great potential for improving the cost estimations.

Another interesting finding is that the model was also able to increase the amount of instances estimated below 30% error, in both countries with smaller amounts of data, China and UAE.

Comparing with the best basis model of CBR, the ANN and bagging method obtained better accuracy for China and UAE projects. For USA both methods, obtained similar results. This finding shows how in the case of availability of enough data, as USA in this study, the ANN and bagging method using addition of non-local data is not necessary beneficial. However as the number of similar projects decrease, there is a great improvement and benefit in using the addition of non-local data using bagging with ANN.

The Figure 5.1 display the MAPE results of the ANN models until this point and the best basis model of CBR. It allows to easily visualize the

improvement provided by the ANN plus bagging model using the addition of non-local data.

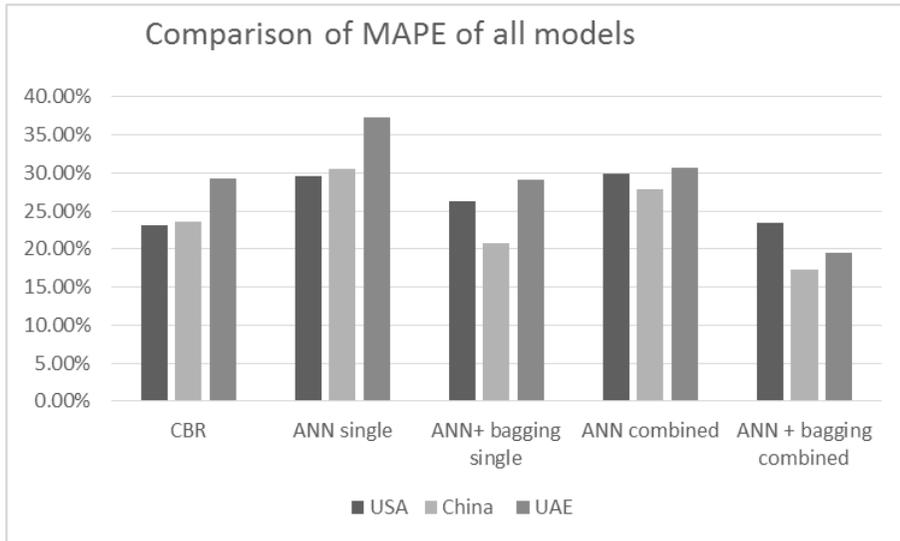


Figure 5.1 – Comparison of MAPE results of all studied models

Based on the successful application of the ANN with bootstrap aggregating method using combined data from different countries. Specially, based on the fact that the improvement in accuracy was more expressive in the countries with less data, a new model was made to check the application of this methodology on countries that present a really reduced amount of data compared to the studied countries.

The next section uses Australian and Singaporean skyscrapers projects data to assess the use of this methodology in a scenario of really reduced data.

5.5 Use of ANN and Bootstrap Aggregating Model in

Countries with Less Data

Based on the promising results obtained in previous section, the data obtained from Singapore and Australia was also added to the combined countries dataset. Then the model was trained and tested against the project cost data from those two countries. CBR models with the individual data from countries of Australia and Singapore were also performed to serve as basis for comparison. The obtained results are shown in Table 5.7.

Table 5.7 – Results for all countries combined, Australia and Singapore

Country	CBR	ANN +bagging combined dataset	
	MAPE (%)	MAPE (%)	percentage below 30% error (%)
Australia	38.48	21.46	80.0
Singapore	65.86	27.10	50.0

These results show good accuracy of the combined dataset modelling, considering the lack of project data from those specific countries. Compared to the CBR, that previously showed good accuracy, for this two countries that possess considerable less data, the error rate increased way over 30%, representing not a reliable estimation method for this scenario. This corroborate the idea that CBR can have issues in dealing with the diversity that the complexity of skyscrapers represent.

In the case of Australia, the accuracy of 21.46% was comparable to the

ones obtained for the countries with more than double the amount of projects' data. And more impressive, the percentage of instances estimated in the 30% or less error range, reached the biggest mark among all the countries, with 80% of the instances being below 30% error.

In the case of Singapore however, even though the MAPE result is within the desired conceptual cost estimation range, it is important to notice that only half of the instances were estimated in that range.

The difference in results between Australia and Singapore might come from the fact that data was collected from different sources, and in doing so, some noise can have been input that caused the difference in behavior among the countries. Having said that, the results achieved are still indicative of the possibility of using this approach even for countries that have only a few examples of completed skyscrapers.

5.6 Discussion

The results of the cost estimation of five specific projects, using the ANN and bagging model with the use of the combined dataset, are shown in Appendix B. One project from each country was selected to demonstrate the input and outputs obtained with the use of the model. The eight attributes selected in section 3.4 are used as input, and then the algorithm returns the output of estimated cost. The output is given as the cost adjusted to the year 2010 and to the base city selected for each country. The error is given as the percentage difference between the actual adjusted cost and the estimated cost. As seen, the error represents the amount of dollars wrongly estimated, and varies from 2 million to 26 million dollars, for the given examples.

Compared to the results of conceptual construction cost estimation for other kinds of buildings projects, in the literature using neural networks (Kim et al., 2004; Wang and Gibson, 2010; Cheng et al., 2010), the results here obtained lack the same high levels of accuracy. However, considering the complexity of skyscraper projects, and the fact of combining different sources for data collection, with possible added noise, the estimations achieved in this research are still satisfactory. And more important they show the potential to increase the quality of construction cost estimations by using the combination of adding non-local data while using bootstrap aggregating.

Adding non-local data showed to be useful in association with neural

networks, despite of the possible problems that it could bring, as discussed before in Chapter 2. This finding indicates that useful information, at least at the conceptual stage, can be contained in data of projects in other countries, with a potential different construction background. Even though this research specifically focused on skyscrapers, it opens space for research about the use of similar strategies in other kinds of construction projects. Projects that are rare, having small number of examples sparsely distributed around the globe, might benefit from such kind of approach.

The results of the first two steps in the modelling process showed that both adding non-local data and using bagging, individually increase the accuracy of the estimations. This finding about bagging confirms the ideas discussed in Chapter 2 of bagging being able to improve accuracy of estimates, especially in neural networks with sparse data (Breiman, 1996; Tsai and Li, 2008). However, compared to the individual improvement that those methods provided, the model at step three, combining both non-local data and bagging, went further, and showed results that indicates that those two solutions, when combined, not only maintain the same level of improvement, but actually show an even more efficient estimation.

Even though, in reality, Singapore and Australia are both among the countries with the biggest numbers of skyscrapers buildings in the world, in this research they were limited to less than ten projects each., considering also some projects below 200 m. The accuracy of the estimations (both MAPEs below

30%), using those two countries limited data as test data, while the reduction of accuracy of CBR, indicates that this approach also has the potential to perform well, not only of improving the data of countries with already some considerable amount of data, but also, to generate adequate estimates for countries with considerable less data points. The meaningfulness of these results lies on the fact that the results in step three are applicable for only eight countries in the world, that have more than 25 completed skyscrapers projects (CTBUH, 2017), while the results in step four doubles the possible spectrum of applicability of the method to a total of sixteen countries around the globe.

Following the promising results of this research, the projects presented in Appendix B can serve as implementation examples for a real scenario. The 8 attributes of a project (Country, Architectural Height, Tower GFA, Real GDP, finishing grade, number of parking spaces and structural material) are input, then the model predicts a cost, using all the available data as training data. The resulting estimated cost is given in dollars, adjusted to the base year of 2010 and the base city for each country. The output cost, can then be transformed to the year of construction and the city of location of the project being estimated, using construction cost indexes, and then converted to the most suitable currency unit.

5.7 Summary

The models development previously described in Chapter 4 were performed and analyzed. The results obtained from the individual countries' models using bagging showed the ability that bagging has to increase the accuracy of the construction cost estimations.

The analysis of the models using the dataset of combined countries demonstrated that data from projects in different countries can present useful levels of information that can be used for improving the accuracy of the estimations.

The model combining the use of additional countries projects' data and bootstrap aggregating obtained significant improved results compared to the other models developed. That indicates the high potential for the application of that method for conceptual cost estimations of skyscrapers.

Analysis with countries with less data was then performed to check if the same methods could be applied with significantly smaller datasets, while maintained a similar level of accuracy. The results although not conclusive still showed the potential for the application of the method.

Chapter 6. Conclusions

6.1 Research Summary

This research focused on cost conceptual cost estimation of Skyscrapers, buildings that reach heights of 200 meters or more. It started by showing the significant increase in the amount of new developments of skyscrapers around the world, in the recent years. Then it showed that because of the high costs related to those constructions, there are a lot of financial risks involved on those projects, and that accurate early cost estimates have a great importance for the successful construction of those developments.

A review of preliminary research was performed and showed that neural networks models can be great tools for accurate conceptual construction cost estimation. But, in view of the high dependence that the accuracy of those techniques have to considerable amount of previous similar data, the application of those models in skyscrapers would be limited due to scarcity of data. However, further review of previous studies showed the potential use of the bootstrap aggregating (bagging) method as strong tool to improve the accuracy of neural networks with scarce data, while also being able to reduce the bias that models made of projects' data from different countries could have towards a data dominant country.

As result of the preliminary findings, with the intention to try applying neural networks to estimate conceptual costs of skyscrapers, this research then

proposed the addition of non-local data, data from projects in other countries, as a way to deal with the limitations on local data.

Skyscrapers' data was then collect from 5 different countries, in a total of 124 projects. Different attributes were chosen to be collect in face of the challenges associated with skyscrapers constructions and conceptual estimation of buildings costs. After extensive analysis and preprocessing, 6 attributes aside from cost and country of construction, were chosen for the development of the models.

In order to study the combined use of adding skyscrapers' data from different countries with artificial neural networks and bagging. Initially, 8 models were made using data from 3 different countries (USA, China and UAE), and used to assess the improvement in accuracy that the proposed model could achieve. Additional methodologies were used as estimations for base comparison. The results corroborated the initial ideas that the ANN with bagging model using a database of combined countries (USA, China and UAE) could significantly improve the conceptual cost estimations of skyscrapers, in all the studied countries.

Based on the satisfactory results, an extra model was performed to assess the use of such methodology in forecasting the costs of skyscrapers in countries with only limited past projects data. A dataset made of 5 countries data (USA, China, UAE, Australia and Singapore) was used, and the adequacy of the model

was tested against the countries with the smallest number of data instances (Australia and Singapore). The results showed potential for the application of the methodology for also countries with a limited pool of data.

Overall the proposed approach of adding non-local data while using artificial neural networks and bootstrap aggregating showed promising results and potential for being applied in conceptual construction cost estimations of skyscrapers buildings.

With the future development and increase in the numbers of skyscrapers around the world, some countries might reach a situation of enough local data, and wouldn't need the use of the methodology studied in this research. However, according to the CTBUH report (CTBUH, 2017), the number of countries building their first skyscraper increases every year, and so it is expected that many countries will also reach the situation of only a few completed projects.

6.2 Contributions

This research is the first to address the use of combined data of projects in different countries to increase the accuracy of artificial neural networks models in the estimation of construction costs of skyscraper buildings. It also demonstrates that bootstrap aggregating can be used in association with the non-local data for even more improved estimations.

It was developed a new accurate model for conceptual cost estimation of skyscraper projects, that method can be used for early estimation of construction costs of complex skyscrapers buildings, without necessarily depending on the expertise of specialists. Even if the accuracy of this model does not reach the accuracy of later stages cost estimating processes, the easy to obtain attributes used in this model allow it to be used by estimating professionals, in the early stages, as an extra tool to support and assess decision making processes of skyscraper projects. This model allows a proper planning and control of overall construction costs, helping to guarantee proper amount of funds before the actual detail design stages of a project development.

The findings on this research also indicate the potential of using the proposed approach of using non-local data in other kinds of construction projects that as skyscrapers projects present lack of similar local data. As example, plant projects that also have limitations of similar projects' data within a same location can benefit from the findings of this research.

6.3 Limitations and Further Research

This research was limited by the possible quality of the data. Since data was collected from difference sources, it is possible that a considerable amount of noise can have been introduced and thus, somehow jeopardized the accuracy of the final results.

Another limitation was that due to difficulties in acquiring cost data of skyscraper projects, Australia and Singapore had to be used as substitutes for countries with not enough examples of skyscrapers.

Further studies should investigate the use of the non-local data with artificial neural networks and bagging method in countries that still haven't built their first skyscraper. Maybe by adding data related to high rise buildings that are not skyscrapers, this approach can potentially still demonstrate good results that would also implicate in a broader use of this methodology, by including countries that are still on the process of building their first skyscrapers.

This research didn't focus on testing specific neural networks methods, but on assessing the capacity of that methodology to learn from the non-local data. It is possible that other algorithms could result in better accuracies. For example, Deep Neural Networks show great potential for the improving even further the estimates obtained in this study. Further study should also focus on improving the neural networks used.

Another possible direction for future research is trying to apply the

approach on plant projects that also have lack of local data, and that could greatly benefit from such approach.

References

- Ahn, J. (2016). *Front-End Cost Estimation by Selective Case-Based Reasoning for Building Construction Projects* (Doctoral dissertation, Seoul National University).
- An, S. H., Park, U. Y., Kang, K. I., Cho, M. Y., & Cho, H. H. (2007). Application of support vector machines in assessing conceptual cost estimates. *Journal of Computing in Civil Engineering*, 21(4), 259-264.
- Arafa, M., & Alqedra, M. (2011). Early stage cost estimation of buildings construction projects using artificial neural networks. *Journal of Artificial Intelligence*, 4(1), 63-75.
- Bayram, S., & Al-Jibouri, S. (2016). Efficacy of estimation methods in forecasting building projects' costs. *Journal of construction engineering and management*, 142(11), 05016012.
- Bayram, S., Ocal, M. E., & Oral, E. L. (2013). Analysis of cost and schedule variances in construction works with artificial intelligence approaches: The case of Turkey. International Student Conference of Civil Engineering.
- Bode, J. (1998). Neural networks for cost estimation. *Cost Engineering*, 40(1), 25.
- Boussabaine, A. H. (1996). The use of artificial neural networks in construction management: a review. *Construction Management & Economics*, 14(5), 427-436.
- Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123-140.
- Cheng, M. Y., Hoang, N. D., & Wu, Y. W. (2013). Hybrid intelligence approach based on LS-SVM and Differential Evolution for construction cost index estimation: A Taiwan case study. *Automation in Construction*, 35, 306-313.
- Cheng, M. Y., Hoang, N. D., Roy, A. F., & Wu, Y. W. (2012). A novel time-dependend evolutionary fuzzy SVM inference model for estimating construction project at completion. *Engineering Applications of Artificial Intelligence*, 25(4), 744-752.
- Cheng, M. Y., Tsai, H. C., & Sudjono, E. (2010). Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry. *Expert Systems with Applications*, 37(6), 4224-4231.
- CTBUH (Council on Tall Buildings and Urban Habitat). (2017). "Year in Review: Tall Trends of 2017". pp. 38-49, Rep.(http://www.skyscrapercenter.com/research/CTBUH_ResearchReport_2017YearInReview.pdf)
- De Jong, P., & van Oss, S. C. F. (2007). High rise costs. In *4TH INTERNATIONAL SCRI SYMPOSIUM. SALFORD: UNIVERSITY OF SALFORD*.
- De Jong, P., & Wamelink, H. (2008, March). Building cost and eco-cost aspects of tall buildings. In *CTBUH 8th World Congress* (pp. 3-5).
- Doğan, S. Z. (2005). Using machine learning techniques for early cost prediction of structural systems of buildings.

- Dursun, O., & Stoy, C. (2016). Conceptual estimation of construction costs using the multistep ahead approach. *Journal of Construction Engineering and Management*, 142(9), 04016038.
- 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.
- Hegazy, T., & Ayed, A. (1998). Neural network model for parametric cost estimation of highway projects. *Journal of Construction Engineering and Management*, 124(3), 210-218.
- Ji, S. H., Park, M., & Lee, H. S. (2011). Cost estimation model for building projects using case-based reasoning. *Canadian Journal of Civil Engineering*, 38(5), 570-581.
- Khosrowshahi, F., & Kaka, A. P. (1996). Estimation of project total cost and duration for housing projects in the UK. *Building and Environment*, 31(4), 375-383.
- 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.
- Kohavi, R. (1998). Glossary of terms. *Machine Learning*, 30, 271-274.
- Lee, J. S., Lee, H. S., & Park, M. S. (2011). Schematic cost estimating model for super tall buildings using a high-rise premium ratio. *Canadian Journal of Civil Engineering*, 38(5), 530-545.
- Li, H. (1995). Neural networks for construction cost estimation. *Building Research and Information*, 23(5), 279-284
- Li, H., Shen, Q. P., & Love, P. E. (2005). Cost modelling of office buildings in Hong Kong: an exploratory study. *Facilities*, 23(9/10), 438-452.
- Li, Y., Lu, K., & Lu, Y. (2016). Project Schedule Forecasting for Skyscrapers. *Journal of Management in Engineering*, 33(3), 05016023
- Lutu, P. E. N. (2010). Dataset selection for aggregate model implementation in predictive data mining (Doctoral dissertation, University of Pretoria).
- Nerrand, O., Roussel-Ragot, P., Personnaz, L., Dreyfus, G., & Marcos, S. (1993). Neural networks and nonlinear adaptive filtering: Unifying concepts and new algorithms. *Neural computation*, 5(2), 165-199.
- Newton, K. O. (2015). *Skyscraper Floor and Cladding Cost Estimator* (Doctoral dissertation, Brigham Young University).
- Oberlender, G. D., & Trost, S. M. (2001). Predicting accuracy of early cost estimates based on estimate quality. *Journal of construction engineering and management*, 127(3), 173-182.
- Perera, S., & Watson, I. (1998). Collaborative case-based estimating and design. *Advances in Engineering Software*, 29(10), 801-808.

- Rafiq, M. Y., Bugmann, G., & Easterbrook, D. J. (2001). Neural network design for engineering applications. *Computers & Structures*, 79(17), 1541-1552.
- Skitmore, R. M., & Ng, S. T. (2003). Forecast models for actual construction time and cost. *Building and environment*, 38(8), 1075-1083.
- Sonmez, R. (2011). Range estimation of construction costs using neural networks with bootstrap prediction intervals. *Expert systems with applications*, 38(8), 9913-9917.
- Tsai, T. I., & Li, D. C. (2008). Utilize bootstrap in small data set learning for pilot run modeling of manufacturing systems. *Expert Systems with Applications*, 35(3), 1293-1300.
- Wang, Y. R., & Gibson Jr, G. E. (2010). A study of preproject planning and project success using ANNs and regression models. *Automation in Construction*, 19(3), 341-346.
- Wilson, A. J. (2005). Experiments in probabilistic cost modelling. *Cost Modelling*, 436.

Appendix A – Selected attributes for data collection

	Attribute	Type of attribute	Description
Project Specific factors	Height: Architectural	Numerical	Height measured from the level of the lowest, significant, open-air, pedestrian entrance to the architectural top of the building, including spires, but not including antennae, signage, flag poles or other functional-technical equipment
	Height: to tip	Numerical	Height is measured from the level of the lowest, significant, open-air, pedestrian entrance to the highest point of the building, irrespective of material or function of the highest element
	Height: Occupied	Numerical	Height is measured from the level of the lowest, significant, open-air, pedestrian entrance to the highest occupied floor within the building.
	Floors Above Ground	Numerical	The number of floors above ground, including the ground floor level, any significant mezzanine floors and major mechanical plant floors
	Floors Below Ground	Numerical	The number of floors below ground should include all major floors located below the ground floor level
	number of elevators	Numerical	The total number of elevator cars (not shafts) contained within a particular building (including public, private and freight elevators).
	top elevator speed	Numerical	The top speed capable of being achieved by an elevator within a particular building, measured in meters per second.
	Tower GFA	Numerical	The total gross floor area within the tower footprint, not including adjoining podiums, connected buildings or other towers within the development.
	Development GFA	Numerical	The total gross floor area within the entire development in which the tower exists, and may therefore be at times the total sum of several towers and related low-rise buildings

	number of apartments	Numerical	Number of apartments contained within the building
	number of hotel rooms	Numerical	Number of hotel rooms contained within the building
	number of parking spaces	Numerical	Number of car parking spaces contained within the building
	Structural Material	Categorical (concrete, steel, composite)	The structure materials used in for the construction of the building (concrete, steel, composite)
	Building Function	Categorical (Residence, office, hotel)	The function for which the building is used for (hotel, residence, office, casino)
	Green Building Certification	Categorical (0, 1, 2, 3, 4)	Green Certifications related to energy efficiency
	Finishing Grade	Categorical (a-high, b-luxury)	The quality of the finishing materials of the building (high or super high)
	Completion Year	Numerical	year of completion of the construction
Local construction environment factors	Real GDP of the city	Numerical	Gross domestic product of the city where the building is constructed, in the year of completion
	Labour productivity of the city	Numerical	The average productivity of the labour in the construction industry of the city, in the year of completion, defined as the ratio between the construction production of the year and the number of employees in the construction industry.
	Wind Speed	Numerical	Design wind speed in the city, at the year of completion, according to local requirements
	Population	Numerical	population of the city in the year of completion
	City	Nominal	City where the building is
	Country	Nominal	Country where the building is
	Construction Cost	Numerical	Final construction cost of the project.

Appendix B – Example of cost estimation

	Country	CHINA	UAE	USA	AUS	SGP
INPUT	Height: Architectural (m)	228	288	244.5	218	242
	Tower GFA (m²)	137,996	107,300	62,263	48,836	79,383
	Real GDP (billions \$)	254	81.6	453	241	264
	Finishing Grade	a	a	b	a	b
	# Parking Spaces	300	725	199	157	180
	Structural Material	composite	concrete	concrete	concrete	composite
OUTPUT	Estimated Cost (million \$)	265	147	366	121	414
	Adjusted Cost (million \$)	263	136	352	112	378
	Error (%)	0.58	8.39	3.92	7.56	9.46