Application of Data Mining and Big Data Analytics in the Construction Industry

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of the Ohio State University

By

Behzad Abounia Omran, M.S.

Graduate Program in Food, Agricultural and Biological Engineering

The Ohio State University

2016

Dissertation Committee:

Dr. Qian Chen, Advisor

Dr. Ann D. Christy

Dr. Cathy Xonghui Xia

Dr. Jia Liu

Abstract

In recent years, the digital world has experienced an explosion in the magnitude of data being captured and recorded in various industry fields. Accordingly, big data management has emerged to analyze and extract value out of the collected data. The traditional construction industry is also experiencing an increase in data generation and storage. However, its potential and ability for adopting big data techniques have not been adequately studied. This research investigates the trends of utilizing big data techniques in the construction research community, which eventually will impact construction practice.

For this purpose, the application of 26 popular big data analysis techniques in six different construction research areas (represented by 30 prestigious construction journals) was reviewed. Trends, applications, and their associations in each of the six research areas were analyzed. Then, a more in-depth analysis was performed for two of the research areas including construction project management and computation and analytics in construction to map the associations and trends between different construction research subjects and selected analytical techniques.

In the next step, the results from trend and subject analysis were used to identify a promising technique, Artificial Neural Network (ANN), for studying two construction-related subjects, including prediction of concrete properties and prediction of soil erosion quantity in highway slopes. This research also compared the performance and applicability of ANN against eight predictive modeling techniques commonly used by other industries in predicting the compressive strength of environmentally friendly concrete.

The results of this research provide a comprehensive analysis of the current status of applying big data analytics techniques in construction research, including trends, frequencies, and usage distribution in six different construction-related research areas, and demonstrate the applicability and performance level of selected data analytics techniques with an emphasis on ANN in construction-related studies. The main purpose of this dissertation was to help practitioners and researchers identify a suitable and applicable data analytics technique for their specific construction/research issue(s) or to provide insights into potential research directions.

This dissertation is dedicated to my parents

for their endless love, support, and encouragement

Acknowledgments

With these few words, I am expressing my sincere gratitude to everyone who, in one way or another, supported me in this accomplishment. I want to thank my advisor Dr. Chen for her advice, guidance, and above all patience throughout this process. I also want to thank my committee members, Dr. Ann Christy, Dr. Cathy Xonghui Xia, and Dr. Jia Liu, for their guidance and technical support throughout this process.

Vita

2001..B.S. Civil Engineering, University of Mazandaran

2005... M.S. Construction Engineering and Management, University of Tehran (UT)

2009..M.E. Construction Engineering and Management, Texas A&M University

2011 to present ....Graduate Research Associate, Construction System Management, Department of Food, Agricultural, and Biological Engineering, The Ohio State University

Publications

Omran, B. A., Chen, Q., and Jin, R. (2016). "Comparison of Data Mining Techniques for Predicting Compressive Strength of Environmentally Friendly Concrete." *J. Comput. Civ. Eng.*, 10.1061/(ASCE)CP.1943-5487.0000596, 04016029.

Omran, B. and Chen, Q. (2016). "Trend on the Implementation of Analytical Techniques for Big Data in Construction Research (2000–2014)." *Construction Research Congress 2016*: pp. 990-999. doi: 10.1061/9780784479827.100.

Omran, B. A., Chen, Q., and Jin, R. (2014). "Prediction of Compressive Strength of 'Green' Concrete Using Artificial Neural Networks." *Proc., 50th ASC Ann. Int. Conf.*, Associated Schools of Construction (ASC), Windsor, CO.

Cao, W., Omran, B.A., Lv, R., He, H., Yanga, X., Leia, L., Chen, Q., Tian, G. (2016). "Studying slope protection effects of vegetation communities for Xinnan highway in china." Submitted to *Ecological Engineering.*

Field of Study

Major Field: Food, Agricultural, and Biological Engineering Specializing in Construction

Systems Management

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Introduction

In recent years, the digital world has experienced an explosion in the magnitude of data and information being captured and recorded in different fields. Technology is growing fast and businesses and industries are now more information-intensive than before. Companies are dealing with a huge amount of information regarding their customers, suppliers, and operations. Also, millions of networked sensors embedded in devices (e.g., mobile phones, automobiles, etc.) are being used to sense and collect data for further use (McKinsey Global Institute [MGI] 2011).

In addition, some existing and emerging concepts like Internet of Things (IOT), smart grids, remote sensing and automation will continue to drive and even accelerate the growth of data. Storing, managing, and analyzing such a huge amount of data cannot be simply done by using traditional databases and techniques. Instead, it requires a new class of advanced technologies. Under such a circumstance, big data management has recently emerged to address this deficiency.

The construction industry, to some degree, is experiencing the same trend as other industries regarding its data generation, storage, and management. Recent technological advances (e.g., networked sensors and cameras) allow construction managers to closely monitor and control project activities and processes, 24 hours a day and 7 days a week, if necessary. This means that nowadays more data is created and available for projects compared to the past. However, there is no clear indication that the construction industry has or is acquiring the ability to automatically extract and analyze these huge amounts of data, so appropriate feedback can be provided for current and/or future projects. Big data analytics, which has been commonly used in many other industries (healthcare, retail, etc.), could be a promising solution to the above-mentioned problem and assist construction managers in performing a more comprehensive and holistic data exploration in their projects with reasonable cost and time.

1.2 Problem Statement, Rationale, and Significance

The construction industry has long been suffering from the poor performance of labor and resource utilization (see Figure 1.1). The uniqueness and complexity of construction projects are two of the reasons that prevent managers from effectively using previous projects data to improve the productivity of their new projects. The failure and success of a construction project depend on series of events that are correlated with each other through many interactions and interconnections. These correlations are not quite clear and hard to recognize without considering and analyzing a large number of elements (e.g., weather conditions, construction methods, etc.).

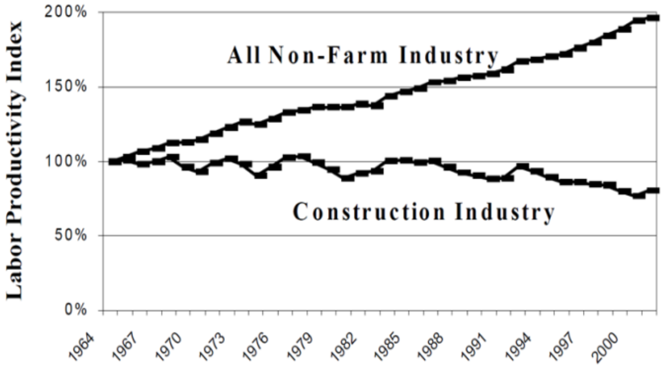


Figure 1.1. Construction industry performance (with permission from Teicholz 2004)

Indeed, with today's technology advances and the increasing amounts of data and information collected and stored each day, it is now possible for construction managers to perform a more comprehensive analysis of their projects to explore the real deficiencies that caused the current productivity issue. For example, big data management techniques such as data mining have been used to explore the reasons behind the low productivity (Kim 2008; Wipro 2013), although the potential for the construction industry to gain values from these techniques was deemed to be lower than many other industries.

According to MGI 2011 (Figure 1.2), the construction industry had the lowest productivity growth of any industry within 2000 to 2008 period. They also claimed that this industry has the lowest potential to explore big data value. This could be a result of the limited amounts of data that were collected and stored in the construction industry in comparison to the other industries, data not digitalized in most cases, or lack of ability for processing and analyzing this data to generate valuable results.

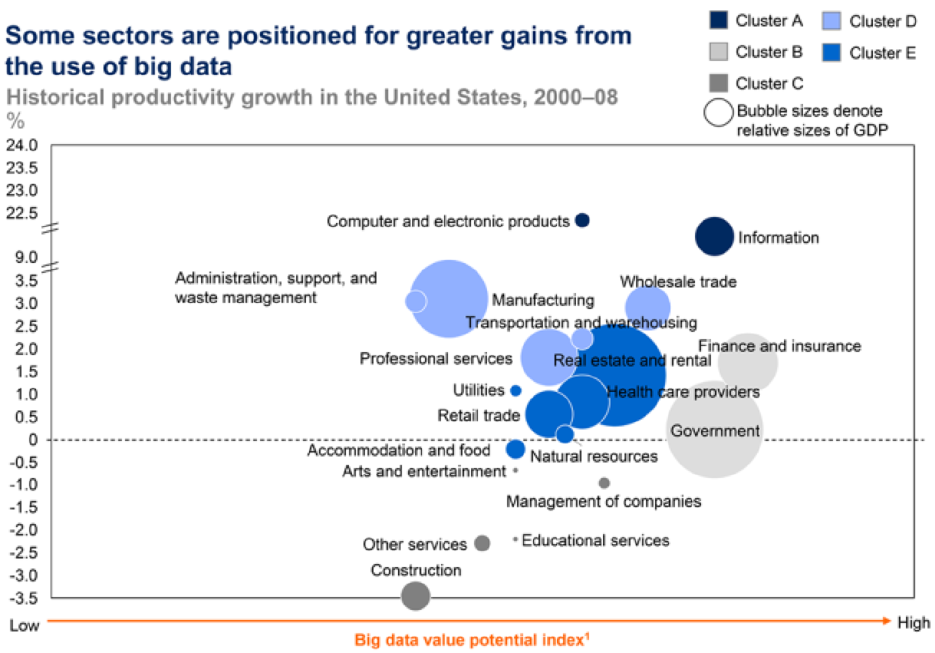


Figure 1.2. Productivity growth vs. big data gains potential (with permission from McKinsey Global Institute)

As discussed before, in recent years construction experiences the same trend as other industries in term of the growth of data. Companies start to digitalize project documents and events. Appendix D shows a list of data and documents commonly record in the construction industry. The industry has seen a rapid increase in the application of Building Information Modeling (BIM) and use of mobile devices and sensors on job sites. However, there has been very little research on how the companies in the construction industry can process, analyze, and apply this huge amount of data in a timely fashion to generate value, attain competitive advantages, or evaluate and improve their productivity and performance, which motivated this dissertation research.

1.3 Scope and Limitation of the Research

This research was set to investigate the data analytics techniques that can be used for big data analysis. Consequently, other data analytics techniques as well as big data technologies are out of the scope of this research. Furthermore, the investigation was limited to the publications within the 30 selected, prestigious journals in the construction-related area.

The scope of this dissertation confined with the intersection of management science, computer science, and construction research as shown in Figure 1.3.

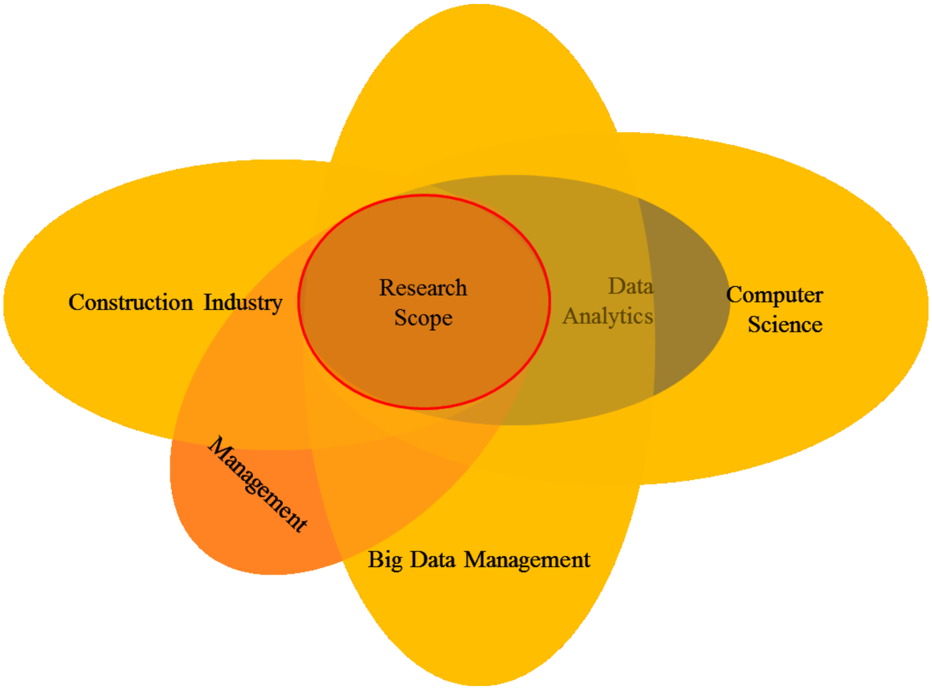


Figure 1. 3. Dissertation Scope

1.4 Research Goal and Objectives

This research aims to explore the applicability and evaluate the potential of using big data analytics techniques in the construction industry. The specific objectives are described below:

- Perform a comprehensive analysis of current status and trends of big data analytics in the construction research community to support the utilization of data mining and other big data analytics techniques in construction-related areas.
- Develop a data-driven subject-oriented application map for the adoption of big data analytics and data mining based on the dataset generated in the previous step.
- Demonstrate the use of the application map by providing two examples of applying data analytics to construction related research subjects
- Evaluate the performance of ANN against eight other predictive modeling techniques in the construction-related research area

1.5 Research Methodology

For the purpose of this research, first a literature review of the big data management and data mining was performed to identify the existing analytical techniques and models in various research communities. As a result, a list of 26 big data analytics techniques was identified and targeted for this research.

Then, a comprehensive search over 30 prestigious journals in 6 different construction-related areas, including 1) *concrete and construction materials,* 2) *building energy and performance,* 3) architectural *research,* 4) *infrastructure research*, 5) *construction project management,* and 6) *computation and analytics in construction,* was performed to generate a dataset containing the metadata information related to the existing publications that have applied at least one of the selected big data analytics techniques in addition to the titles and abstracts for those publications.

The dataset was cleaned, sorted and filtered to eliminate the incomplete and incorrect data, and then used to analyze and evaluate the current status and trends of application of big data analytics techniques in the construction research community.

A manual investigation of the main topics and subjects of publications related to the areas of *construction project management* and *computation and analytics in construction* was performed, which provides the categorization of common subject topics studied in those publications. Then, common patterns and relationships between these research subjects and the adopted big data analytics techniques were identified and mapped in a tabular format.

Using the generated application map, Artificial Neural Network (ANN) was selected and applied as one promising big data analytics technique for two specific construction research subjects including the prediction of concrete properties and the prediction of soil erosion in highway slopes. Different numbers and settings of input variables, parameter tuning, and cross-validation were carefully examined to improve the prediction performance of tested ANN models.

The prediction performance of nine different predictive modeling techniques including ANN was examined and analyzed using the dataset of environmentally friendly concrete. Extensive hand-tuning and 10-fold cross validation was also used to improve the prediction performance. The results were compared to that coming from previous related research.

1.6 Research Significance and Contributions

The construction industry is one of the biggest industries in many countries. According to the United States Census Bureau (2013), the annual value the U.S. construction industry put in place in 2006 was equal to 1.167 trillion dollars. Considering the magnitude of money and resources that are involved in this industry, any minor improvement in its productivity can lead to a huge impact on the U.S. economy. Despite this fact, the construction industry, as reported by many researchers, has a descending trend in productivity within the last few decades.

Big data analytics and data mining as discussed in this dissertation could help the construction industry detect the hidden patterns, risks, and improvement opportunities in its project delivery process, which could increase the efficiency and productivity of this industry. The value of the data became common knowledge throughout the construction industry. Many of the construction companies began to collect, record, and document available data regarding their projects and operations. The ascending trend in the application of Building Information Modeling platforms enabled these companies to document every interaction among the stockholders of the projects, and companies are expanding their information technology department to grasp every bit of competitive advantage they can achieve by utilizing this data. Despite all the expenditures and efforts spent for acquiring valuable data, taking advantage of this data and extracting valuable information from it are still limited to the level of knowledge that managers and decision-makers have regarding what they can do with this asset.

The knowledge gap between the industry and the research community in construction is wider than that of many other industries. In other words, there is lack of close interactions between academia and industry, so practitioners have less awareness of existing data analytics techniques that can help them further explore the value of their collected data to advance field operations. In addition, many researchers working in the construction area still limit themselves with conventional data analysis techniques such as basic statistical analytics in performing their studies.

The most significant contribution of this research was to address these gaps by performing a comprehensive analysis of existing application of big data analytics techniques for construction related subjects. This has not been performed by the construction research community. This research makes several contributions to the existing body of knowledge through:

- Providing a statistical and numerical analysis of the application of 26 big data analytics techniques in the publications from 30 prestigious journals in six different areas of construction research.
- Performing a trend analysis for the application of both the overall big data analytics techniques and each of the 26 individual data analytics techniques over the past 15 years using trend lines, treemaps, and box-and-whisker plots.
- Establishing an inclusive list of categories and subcategories of construction related research subjects that applied at least one of the 15 selected big data analytics

techniques in two selected construction research areas as *construction project management* and *computation and analytics in construction* as well as performing a comparative study of similarities and differences between these two areas.

- Generating treemaps of the research subjects identified in the previous step vs. the 15 selected big data analytics techniques and analyzing the frequency of application of these techniques in each of the individual construction research subjects.

The other major contributions were made through investigating applications of selected data analytics techniques to bridge the gap between the industry and the research community. These contributions include:

- Investigating the application of ANN to predict the compressive strength of concrete made with alternative materials such as fly ash, Haydite lightweight aggregate, and Portland limestone cement. The research tested different settings and numbers of input variables based on different datasets. The results will help researchers and practitioners in the concrete industry predict the strength of this type of concrete with higher accuracy.
- Developing and validating a data-driven predictive ANN model for predicting soil erosion in highway slopes based on the type of vegetation communities, rainfall events, and soil characteristics. The results of the research confirmed that ANN has an acceptable accuracy for soil erosion prediction. In addition, the extensive hand-tuning performed in this study demonstrates the possibility of further improving the prediction performance of ANN models.

- Performing an in-depth analysis and comparative study of the prediction performance of ANN and eight selected data mining techniques in studying environmentally friendly concrete. The unique set of seven data mining models was selected for exploring the prediction performance of four regression tree models (M5P, REPTree, M5-Rules, and decision stump) against other three more advanced models (multilayer perceptron, support vector machines, and Gaussian processes regression). This seemed to be the first time that Gaussian processes regression was examined for predicting concrete strength. In addition, two commonly used ensemble methods (additive regression and bagging) were tested by adopting each of the seven individual models as the base classifier to explore the possibility of improving prediction accuracy. The ultimate goal was to promote the use of data mining techniques for determining the compressive strength or other properties of new types of concrete while reducing the need for extensive experiments. This shift will not only save time and money for the industry but also facilitate the use of new materials.

1.7 Dissertation Organization

The remainder of this dissertation was organized in the following ways. Chapter 2 provides a literature review of the main concepts and elements of data mining and big data analytics. Chapter 3 presents a comprehensive investigation of the application of 26 big data analytics techniques in six construction-related research areas. Chapter 4 demonstrates the applicability and performance of ANN as a predictive modeling technique for two specific construction-related research subjects. Chapter 5 presents a comparative study of nine commonly used data mining techniques in predicting compressive strength of environmentally friendly concrete. Chapter 6 summarizes this research and makes recommendations for future research in the emerging area of big data analytics.

Chapter 2. Literature Review

This chapter first some of the main concepts of data mining such as definition and process as well as it's applications in the construction industry were introduced. Next, the big data management, its definition, opportunities, challenges and its main elements including process, framework, techniques and technologies were discussed and the previous applications of big data management in the construction industry were reviewed. Lastly, the Artificial Neural Network (ANN), its different types and parameters have been briefly discussed.

2.1 Data Mining

According to Hand et al. (2001), data mining is defined as the science of extracting valuable information from large datasets. He further defined data mining as a new area which connects many of the other disciplines, including statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, and others. It is usually seen as the main part of knowledge discovery in databases (KDD) process.

2.1.1 Knowledge Discovery in Databases

Knowledge discovery refers to a wide range of process with the purpose of finding useful information and patterns in data and includes processes such as data selection, preprocessing, transformation, data mining, evaluation, and possibly interpretation of the extracted patterns and information. Cross-Industry Standard Process for Data Mining (CRISP-DM) also provides its version of KDD phases as:

1) **Business Understanding:** Understanding the project objectives from a business perspective and preliminary planning of data mining process to achieve these objectives.
2) **Data Understanding:** It includes exploration and description of collected data and understating of its quality and origin.
3) **Data Preparation:** The process of transforming the data from its initial raw data form to final dataset that is ready to be used in modeling tools and software and could be consist of tasks such as data selection, data cleaning, and data integration.
4) **Modeling:** the primary objective of data mining is to fit the data to a model in order to explore information and patterns that may not be apparent when looking at raw data. This phase includes selecting appropriate modeling techniques, generating appropriate model by running the modeling tools, and assessment of the accuracy of created model.
5) **Evaluation:** In this phase, the generated model will be evaluated against business objectives to identify applicability and usefulness of the model and identify the missed

objectives. Usefulness, return on investment (ROI), accuracy, space, and time are some of the main data mining metrics that can be used for this purpose.

6) **Deployment:** This phase includes implementation of data mining process, and generated model for its intended purpose such as prediction and classification.

Consequently, data mining can be seen as an application of algorithms to extract the information and patterns derived by the KDD process. Depending on the objectives of the analysis, data mining could be categorized into different groups of function as shown in table 2.1.

Table 2. 1. Data mining general objectives and tasks

	Objectives	**Functions**
Data Mining	Predictive	Classification Regression Time Series Analysis Prediction
	Descriptive	Clustering Summarization Association Rules Sequence Discovery

Summarization concerns with providing a simpler and more compact representation of the dataset that can be used in visualization and report generation. Sequence discovery deals with finding the statistical patterns for the sequential data. The other 6 functions will be discussed in Section 2.4.1 along with big data analytics techniques. As mentioned in the previous sections, many of the big data techniques were employed by researchers as data mining techniques to analyze smaller datasets and extract more specific goals. In fact, data mining covers many of the models and techniques used in big data analysis, although some fundamental differences exist between these two.

Big data comes from somewhat new types of data sources that have previously not been analyzed for insight (Lumpkin 2013*).* Carter (2011), in a white paper developed by IDC, discusses that in Big Data Management, content or even consumption of data is not a big concern. Instead, analysis of the data and how that needs to be done are most important. The difference between traditional data and big data was probably best explained in a white paper developed by Howard (2012). According to this paper, the data that represented within big data contain raw data while the data that have traditionally been used in various applications are transactional data which have been processed and stored for a specific purpose. Next, we will discuss some of data mining applications in the previous research related to the construction industry.

2.1.2 Data Mining in Construction

Data extraction and retrieval is one of the data mining areas that have been studied by many researchers in the construction industry (Lin et al. 2006; Lee et al. 2009; Elghamrawy et al. 2010). Lin et al. (2006) proposed a knowledge assisted approach to improve the AEC (architecture, engineering, and construction) product information transactions based on information exploration throughout the web. They employed three different models, including a statistical approach, a semantic approach, and a statistical/semantic hybrid approach, for this purpose. Elghamrawy et al. (2010) performed a four-step methodology to manage construction documents, including a template document metadata, a semantic ontology, a RFID framework, and an ontological indexing mechanism. They used RFID-based semantic contexts for retrieving the documents related to specific concepts on the job site. Lee et al. (2009) introduced a system to capture ideas from past value engineering processes, so the ideas could be applied to solve issues which may occur in current projects. The proposed system consists of both a matching algorithm and a ranking algorithm, which together would be able to find and recall the most related approaches regarding a recorded issue in previous value engineering documents.

Data mining techniques can also be applied to search and identify an intended sketch or image (Brilakis et al. 2005; Yu et al. 2013). A Content-Based Search Engine has been developed by Brilakis et al. (2005) to manage construction image database and retrieve useful images. An approach similar to clustering, called Blind Relevance Feedback, was applied to identify the intended images. Most recently, Yu et al. (2013) implemented a content-based text mining technique to generate a model for computer-aided design (CAD) document retrieval. They used a Vector Space Model (VSM) to match the similarities between CAD documents. In addition, a scope narrow down search was employed to rank available documents and identify the intended one.

Exploring and retrieving the tacit knowledge of experts was another subject that has been studied by researchers (Woo et al. 2004; Lin 2008; Tserng et al. 2008). Specifically, Woo et al. (2004) discussed tacit knowledge in the AEC industry and proposed a dynamic knowledge map, which can assist in the application of expert's tacit knowledge. Lin (2008) used a people-based map to provide an assistant experience management system for construction projects. This system helps share expert's tacit knowledge and experience during the construction phase. In another attempt, a project knowledge management framework for tunnel construction was suggested (Tserng et al. 2008). In their work, a methodology was developed to collect and record the knowledge generated in a specific project and to provide a directional experience anthology framework, which could fill the gap between existing knowledge organization systems and practical requirements.

2.2 Big Data Definition

According to the Research Trends Special Issue on Big Data (2012), the term “big data”, coined by Roger Magoulas in 2005, first appeared in a 1970 article on atmospheric and oceanic soundings to refer to large amounts of data generated by a particular project. However, it was not until 2005 that “big data” was formally considered a research and science topic. There are multiple definitions of big data exist in the literature (Banerjee 2012). For instance:

- According to O’Reilly (Dumbill 2012), “Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures.” They suggest that some alternative methods should be used to extract value from these data.
- MGI defines “big data” as a dataset whose “size is beyond the ability of typical database software tools to capture, store, manage, and analyze.” In its 2011 report, the organization claimed that the definition of “big”, in this term, can

vary over time or by sector, depending on available software tools and the sizes of datasets which are common in a specific industry. As a result, big data could "range from a few dozen terabytes to multiple petabytes" (MGI 2011).

- IDC, in its 2012-2015 forecasts, describes Big Data technologies as a "new generation of technologies and architectures designed to extract value, economically, from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis" (IDC 2012).

In general, big data refers to data that surpasses the processing capacity of conventional data management systems and software tools to capture, store, manage, and analyze (MGI 2011, Dumbill 2012). This could be due to the size of the data, its incompatibility with the structures of existing databases and tools, etc. (Dumbill 2012).

IDC's definition of big data is consistent with the common Three V's framework that has emerged to define big data. Many researchers used three V's or, in some occasions, four V's to capture the characteristics of big data. Ylijoki and Porras (2016) claimed that the original, so-called 3V definition of big data as volume, velocity, and variety was first presented by Laney (2001) and became the fundamental dimensions of big data. They also provided a frequency analysis of most common characteristics used in the definitions of big data in the 62 papers selected in their study. Based on their research, volume, variety, velocity, value and veracity are the most frequent features that have been used to define the big data. Although they identified 17 different definitions of big data and claimed that most of these definitions intend to expand the standard 3V's definition to capture the technical and business aspects of big data. For instance, Gartner (2012) defines big data as high-volume, high-velocity, and high-variety information assets but also adds the need for cost-effective and innovative forms of information processing to its definition. Below is a brief description of the 3 essential V's, including volume, variety, and velocity:

- **Volume** refers to the size of data. The data is now more than text data as we can find data in the format of video, music, sensors data, large images and many others. It is very common to have Terabytes and Petabytes of the storage system for enterprises.
- **Velocity** refers to the rate at which data is generated. Stored, analyzed and acted upon.
- **Variety** refers to various types of data that could be structured, unstructured, or semi-structured and needs different approaches to managing.

Other V's that were suggested by researchers include, but are not limited to, **value**, **veracity**, **visualization** or **visibility**, and **variability** (Ylijoki and Porras 2016). As a result, big data does not have exact size, range or definition and it varies over time or by sector depending on the available technologies/software tools or the sizes of common datasets in individual industries (MGI 2011).

2.3 Opportunities

Many studies have predicted the huge impact of big data on various areas in the near future (MGI 2011, Economist Intelligence Unit [EIU] 2011, IDC 2012). For instance, MGI (2011) emphasizes the important role of big data in economic and states that big data can be beneficial for both private commerce and national economies. EIU (2011), in a sponsored research, conducted a global survey of 586 senior executives. The results show a strong link between effective data management strategies and financial performance of their organizations. As concluded in the study, big data is changing the way how companies (regardless of their sizes and industry sectors) operate and compete.

Big data technologies are already being used in many different industries and sectors. MGI (2012) provides a summary of applications and potential for big data in five industries, namely health care in the United States, public sector administration in the European Union, retail industry in the United States, global manufacturing, global personal location data. They stated five general ways that big data can be used to create value for an organization. These five ways include:

1) Creating transparency
2) Enabling experimentation to discover needs, expose variability, and improve performance
3) Segmenting populations to customize actions
4) Replacing/supporting human decision making with automated algorithms
5) Innovating new business models, products, and services

They also claimed that effective use of big data can significantly enhance the public sector's productivity.

In March 2012 president Obama administration announced the "Big Data Research and Development Initiative." The initiative is supposed to help advance the fields of science and engineering by enhancing the ability of extracting knowledge from large and complex collections of data. (The White House 2012). Many US government agencies acknowledge the opportunities that appear as a result of applying the big data science and start ongoing programs to utilize this advancement to help achieve their agencies' missions (Fact Sheet 2012).

2.4 Challenges

The fast advancement and growth in existing data not only bring many opportunities but also generate many challenges that need to be addressed. Many researchers discussed these challenges and the actions that need to be taken to address them. These challenges are associated with nature of big data and the aforementioned V's (Laney, 2001), in addition to the different phases of big data life cycle. For instance, Chen et al. (2014) list the challenges of big data in six areas of data capture, storage, Data transmission, data curation, data analysis and, data visualization. Lack of training for data specialists and managers is another challenge that was stated by the researchers (Manyika et al., 2011; Hilbert, 2016). According to Manyika et al. (2011), by 2018, the United States will need 160,000 more professionals with deep analytical skills. Infrastructure access, cost and expenditures, and security are other challenges that have been discussed in the previous research (Chen et al., 2014; Hilbert, 2016).

2.4.1 Big Data Techniques vs. Technologies

To better understand the scope of this research, it is important to know the definition of big data techniques and big data technologies as well as their differences. Wactlar (2012) specified three main areas of big data research: 1) the collection, storage, and management of big data, 2) data analytics, and 3) data sharing and collaboration. MGI (2011), on the other hand, centered its big data discussions on two specific terms as “techniques” and “technologies.” According to MGI (2011), in the data management field techniques are generally defined as types of analysis used to process and extract information from raw data while technologies are mostly referring to methods, software, and platforms used for strategic data management.

Many of existing big data techniques originated from the fields of mathematics, statistics, computer science, and other specialty areas. Despite their capability to handle big data, some of them (e.g., A/B testing and regression) can work effectively for smaller datasets. Previous research has attempted to categorize existing big data techniques. For example, Gandomi and Haider (2014) classified these techniques into 5 categories, namely

- **Text analytics:** Also refer to text data mining including the techniques that extract information from textual data such as social network, emails, corporate documents and news through the application of natural language processing (NLP), sentiment analysis, information extraction (IE), text summarization and other analytical methods.
- **Audio analytics:** As it is clear from the name audio analysis or speech analytics concern with extracting information from unstructured audio data. According to Gandomi and Haider (2014), the transcript-based approach and phonetic-based approach are the two common technologies that are being used in speech analytics.
- **Video analytics:** Also known as video content analytics (VCA), it aims to analyze video streams to detect, monitor or extract certain information such as temporal and spatial events, object detection, and motion detection. It is part of computer vision

and widely used in many industries such as health-care, retail, and automotive. The two common approaches for video content analysis are server-based architecture and edge-based architecture.

- **Social media analytics:** According to Gandomi and Haider (2014), social analytics is mostly applied to marketing and has two main categories of content-based analytics (which focuses on extracting insight from data posted by users on social media) and structure-based analytics (which uses network and graph theories to detect the relationships between the users).
- **Predictive analytics:** Using a variety of analytical techniques such as predictive modeling, machine learning, and data mining to predict an outcome in the future based on the historical and current data available data.

Wang (2015) organized recent techniques of big data into resampling-based methods, divide and conquer methods, and sequential updating methods. There are few other categorizations for big data analytics techniques but none is more comprehensive than the list provided by MGI (2011). MGI in its 2011 report listed 26 common data analytics techniques and claimed that although this is not a complete list all the techniques presented are applicable to big data analysis. These big data techniques and a brief description of them (according to MGI 2011) are provided below.

A/B testing: Known as split testing or bucket testing, it compares a control set with many other test sets to define what strategy will enhance a certain objective variable.

Association rule learning: A set of techniques to identify significant relationships and rules between variables in a dataset with a large sample size.

Classification: Techniques used to categorize new data points and assign them to proper groups by using previously defined categories and training sets associated with them.

Cluster analysis: Techniques to identify similar data points in a diverse dataset and group them together in a way that similar objects with the same previously unknown characteristics would be placed in the same group.

Crowdsourcing: Obtaining data, ideas or other needed services by seeking help from a large group of people (especially online community).

Data fusion and data integration: The process of integrating and analyzing data from multiple sources to obtain knowledge for an object in more efficient ways than just using a single source of data.

Data mining: Techniques used to discover hidden patterns from a dataset. It is an interdisciplinary area of science, which combines different statistics and machine learning techniques to analyze and extract the patterns.

Ensemble learning: Using multiple learning algorithms to obtain better predictive performance for a specific problem. This is a type of supervised learning.

Genetic algorithm: A revolutionary algorithm that is inspired by the process of natural selection in a way that solution can merge and mutate. It is a heuristic search that regularly produces useful results to optimize the solution.

Machine learning: Considered a subfield of computer science, machine learning concerns designing algorithms to help automatically learn and identify various patterns based on historical/empirical data.

Natural language processing: Methods that combine computer and linguistics science to analyze human natural language.

Neural networks: A computational system inspired by biological neural networks, which consists of simple, highly interconnected processing elements (nodes or neurons) that work together to solve specific problems, especially for finding nonlinear patterns.

Network analysis: Techniques used to distinguish the associations among discrete nodes of a graph or a network.

Optimization: Consisting of various numerical techniques used to select best element or parameters for a system or process to improve its performance based on a number of objectives.

Pattern recognition: Methods used to identify patterns and regularities in the data. Pattern recognition, machine learning, data mining and some other terms have overlaps with each other. Classification techniques are examples of pattern recognition.

Predictive modeling: Techniques that are used to predict an outcome or its probability. Regression is an example of predictive modeling techniques.

Regression: Statistical methods that focus on estimating the association between dependent and independent variables and can be used for different forecasting or prediction purposes.

Sentiment analysis: Techniques that use natural language processing or other computational linguistics to extract subjective information from a set of texts.

Signal processing: Techniques that analyze or transfer signals and help extract information or distinguish and eliminate the noises. They could be used in modeling for time series analysis or applying data fusion.

Spatial analysis: Using analytical techniques to analyze the topological, geometric, or geographic properties of a data set.

Statistics: The science of the collection, organization, and interpretation of data, which also consists of designing surveys and experiments. As a science, statistics overlaps with many other methods and techniques discussed in this study.

Supervised learning: Machine learning techniques that applied labeled data to generate a function or identify a relationship. Classification is an example for supervised learning.

Simulation: Techniques that imitate or model a real life process or system behaviors, especially over time. They can be used for forecasting and scenario planning.

Time series analysis: Set of methods that analyze sequential data sets or data over a continuous/successive time interval to extract meaningful patterns and information.

Unsupervised learning: Machine learning techniques that apply unlabeled data to find a relationship or hidden structure in a data set. Cluster analysis is an example of unsupervised learning.

Visualization: Methods used for creating images, diagrams, or animations to improve the understanding and communication of big data analyses.

In addition to suitable big data techniques, appropriate computing technologies and platforms are also needed to support the process of collecting and extracting value from big data. While a single computer is sufficient for small-scale data analysis, data mining or data sharing in a medium- or large-scale database is likely beyond its capacity and usually relies on parallel computing or collective mining (Menandas and Joshi 2014). As a result, big data processing frameworks usually run on cluster computers with a high-performance computing platform (Wu et al. 2014). There are several technologies and platforms for the application of big data. Below are examples of the common technologies and platforms used to collect, manage, and evaluate big data and support big data techniques.

Hadoop: Hadoop is an open-source, Java-based, highly scalable software framework for distributed computing. It is an Apache project inspired by Google's MapReduce and

Google File System and one of the most well-established software platforms that can be used for processing huge datasets across clusters of computers.

MapReduce: A software framework introduced by Google and developed by yahoo for processing huge datasets through breaking down complex problems into many sub-problems.

Cassandra: Initiated by Facebook, Cassandra is another open source distributed and structured storage system for managing large scale data without compromising performance.

R: It is a free open-source programming language and environment for statistical computing and graphics that allows the user to distribute, change, and improve the software. It is a free version of S programming language (Originally developed by Bell Labs) under General Public License (GNU) and one of the most widely used platforms for developing statistical software.

Bigtable: It is a distributed storage system based on Google File System for storing large-scale structured data on a cluster of community hardware.

HBase: An open source, non-relational, distributed database, HBase is the Hadoop database consist of a very large table (Same as Google's Bigtable) that can apply HDFS (Hadoop Distributed File System).

Apache Mahout: It is another Apache project and consists of a library of scalable machine-learning and data mining algorithms mostly focused on collaborative filtering, clustering, and classification.

Hadoop Distributed File System (HDFS): As defined by Apache, it is a highly fault-tolerant distributed file system designed to reliably store, stream and run on a cluster of low-cost hardware.

Cloud Computing: In general, cloud computing means using the internet and a network of remote servers to store, manage and process data instead of personal computer.

Mashup: Often used on web-based content, is integration and application of data from different sources to create new services that were not the original reason for producing those raw source data.

Metadata: It is the data that provides basic information such as time, date, and the person responsible for creation or modification of other data. It is commonly used for information stored in data warehouses.

Despite the huge number of existing big data techniques and technologies, researchers are still joining the efforts to develop new or improve existing methods and technologies due to the rapidly evolving nature of this emerging field, the variety of needs/applications, and the huge impact big data is expected to make. The focus of this research is on big data techniques and their applications in construction-related research and not the technologies and platforms.

2.5 Big Data in Construction

In addition to what was mentioned above, considering the focal research areas in the automation of construction processes, safety monitoring/control, resource management, etc. (e.g., Zhai 2009, Zheng 2012), the construction industry should expect a significant increase in the amount of data that will be generated and recorded in the near future. However, as pointed out by Bloom et al. (2012), despite that a large amount of information has been generated since the emergence of building automation, limited efforts have been made to process these data. Fortunately, due to the increasing popularity of big data, the attempts to apply big data management in the construction industry have been increasing in recent years.

For example, Alderon Iron Ore Corp. adopted big data technologies to manage more than 150 different types of project documents (e.g., drawings, models, contracts, and procurement materials) to provide a real-time document access for geographically spread project members (Constructech 2012). The U.S. Veterans Administration used SPARC520, a system that is able to handle big data challenge from the cloud, to identify effective actions for saving energy in its 11-story Washington, D.C. headquarters building (John 2012). According to the organization's Chief Executive Officer, Eric Bowman, the system consists of 107 clamp devices to meter electricity, water, and natural gas consumption. Since each clamp device is able to collect 27 different types of data every 10 seconds, this would generate about 2 billion data points per year (John 2012).

These are a few examples that show the construction industry has started to recognize and employ the power of data analytics and data-driven decision-making. AEC Big Data

(2013) ranked big data the number one among four enabling technologies that are causing organizations including AEC companies to transform their strategies and operations. A survey of 838 construction professionals by Sage (2014) revealed that although 75% of the survey participants were not familiar with the term “big data,” approximately the same percentage of them believed that management of big data should become one of the most important tasks their IT sections/departments offer.

The results of this preliminary research show that many of the big data techniques mentioned by MGI (e.g., Classification, Statistics) commonly have been used by researchers in the construction industry. Some other techniques such as pattern recognition or cluster analysis were less frequently used and a few techniques, e.g., ensemble learning or Crowdsourcing, were rarely applied or had no existing application in the construction related research area. One of the objectives of this research is to analyze these techniques regarding their applications in the construction industry and organize them into a framework that can be beneficial for future research in this area.

2.6 Artificial Neural Network

ANN is a computational system consisting of simple, highly interconnected processing elements (neurons) that work together to solve specific problems (Caudill 1987). It is an algorithm inspired by research in biological nervous systems to generate a simplified model of how the brain works (Rumelhart et al. 1994). The first neural network was proposed by two physiologists, Warren McCulloch and Walter Pitts (1943), and since then many other models have been introduced by other researchers. Self-organizing Mapping (SOM), Radial

Basis Function (RBF), Multilayer Perceptron (MLP) and Neuro-Fuzzy are among the most commonly used ANN models and have been recently studied by researchers in various fields with high accuracy (Bishop, 2006). Some common learning algorithms employed within these ANN models included backpropagation, reinforcement learning, lazy learning, etc. In recent years, ANN has been extensively used for several purposes, including, but not limited to, estimation, pattern recognition, classification, function approximation, and forecasting.

Figure 2.1 shows the basic structure of an artificial neural network. Each neuron will receive one or more inputs. The inputs will be multiplied by their weight, and summed together and with the bias (threshold). The weighting and bias values will be initially chosen as random numbers and then adjusted according to the results of the training process (Atici 2011). The output of each neuron will be generated based on the significance of the summation value and by the means of a predefined specific activation function. Early ANNs generated simple binary outputs in this way, but later it was found that continuous output functions are more flexible. The most common activation functions used by researchers include, but are not limited to, Unipolar Sigmoid Function, Bipolar Sigmoid Function, Hyperbolic Tangent Function, Radial Basis Function, and Conic Section Function (Bishop, 2006).

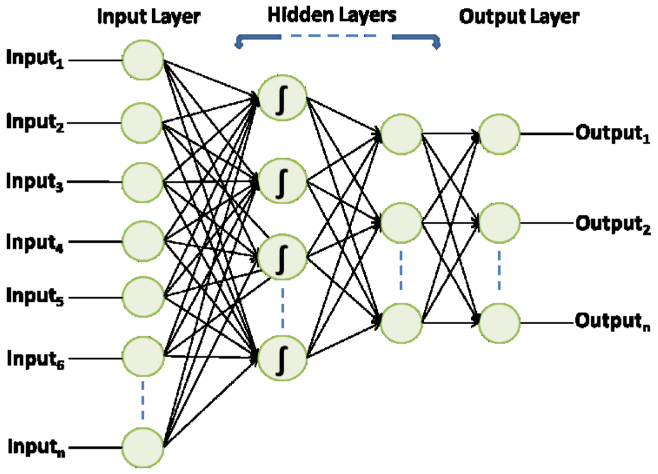


Figure 2. 1. Structure of ANN models

This study applied an MLP model for both research subjects. MLP (Multilayer Perceptron) is a feed-forward neural network developed by Rosenblatt (1958). It is one of the first and most frequently used models in machine learning.

MLP can have one or more hidden layers, depending on the type and complexity of the problem to be solved, although a single hidden layer with a sufficient number of neurons is usually good enough to model many problems. In multi-hidden layer cases, the output from each hidden layer is treated as an input for the next hidden layer. There is no general rule for choosing the number of neurons in the hidden layer. However, it should be large enough to correctly model the problem of interest, but be kept sufficiently low to ensure generalization of the network and to avoid the over-fitting problem (Alshihri 2009). Some studies related the number of hidden layer neurons to the number of variables in the input and output layers or defined an upper bound for it. However, these rules cannot guarantee the generalizability of the networks (Alshihri 2009; Atici 2011). A common way to select the appropriate number of neurons in each hidden layer is to perform a parametric analysis of the network and check the accuracy of the results. The use of a validation set can also help improve the generalization and avoid the over-fitting problem.

Chapter 3. Implementation of Big Data Analytical Techniques in Construction Related Research (2000-2015)

3.1 Introduction

Nowadays, businesses and industries are far more information-intensive than ever before due to the fast development of information technology (IT), which enables them to collect, process, and store data more easily. Particularly, companies are dealing with a huge amount of information regarding their customers, suppliers, and operations on a daily basis. Millions of networked sensors embedded in various devices (mobile phones, automobiles, etc.) are being used to continuously sense and collect data for further use (McKinsey Global Institute [MGI] 2011). In addition, some emerging concepts such as Internet of Things (IoT) and smart cities are expected to produce tremendous amounts of data in the near future (Violino 2013).

According to a report by the International Data Corporation (IDC 2012), the amount of information created and replicated in 2011 passed 1.8 zettabytes (ZB) (one ZB equals to1.6 trillion gigabytes). This was more than 10 times increase in five years compared to 1.61 ZB in 2006. IDC also predicted that this number would reach 40 ZB by 2020. Despite the huge amounts of data available, only approximately 0.5% of these data is being analyzed to extract any form of useful information (Gantz and Reinsel 2012).

As pointed out by IDC (2012), storing, managing, and analyzing large amounts of data cannot be simply done by using traditional data management and analysis technologies and techniques. Instead, it requires a new class of advanced technologies and methods. To address such a critical need, the field of big data management has emerged. During the period of 2006-2011, academic attention toward big data management was increasing significantly. An analysis by Research Trends (2012) showed that the number of papers related to big data had been increased approximately 11 times over the same period.

The construction industry, although at a slower pace, is experiencing a similar trend and expansion with the quantity and quality of the recorded data. Part of this growth comes from digitalized project documents and databases, increased use of building information modeling (BIM), and widely deployed mobile devices on the job sites, which continuously capture the information related to real-time project activities and progresses (Venkatraman and Yoong 2009, Davies and Harty 2013). However, so far there is no clear indication that the construction industry has or is acquiring the ability to process, analyze, and apply these data in a timely fashion to improve its project performance and efficiency.

Considering the potential knowledge/technology transfer from academia to industry, this study aims to investigate the current status of applying big data analysis techniques in construction research, including trends, frequencies, and usage distribution in six different construction-related research areas, followed by a more detailed analysis of the application of selected data analysis techniques in two specific research areas including construction project management and computation and analytics in construction. This study first provides a brief review of big data management and its application in the construction industry. Then, literature on the application of the 26 selected big data analysis techniques in construction research were analyzed, and the related research findings and discussion are presented. The results of this research can not only provide a better understanding of the application of these techniques in existing construction-related studies but also help practitioners and researchers to identify suitable analytic techniques for their specific research topics/problems.

3.2 Research Methodology

As mentioned, the main objective of this research is to investigate the trend of applying big data analytics in the construction research community. The purpose is to learn how familiar the research community is with big data techniques. So a perception can be formed in terms of how well the researchers in this field have prepared themselves to address this coming challenge.

This research adopted the list of 26 data analysis/big data techniques from MGI (2011) for its investigation. It surveyed 30 prestigious journals from six different subareas of construction related research, including 1) *concrete and construction materials,* 2) *building energy and performance,* 3) architectural *research,* 4) *infrastructure research*, 5) *construction project management,* and 6) *computation and analytics in construction*. For each of these areas, 4-6 journals were selected based on their reputation, impact factor, and relevance to the specific research field. Each selected journal was investigated separately and in joint with other journals regarding the application of big data techniques in their papers. The Engineering Village database was used as the main source to retrieve the related publications in these 30 journals. The time frame for this search was between January 1, 2000, and September 30, 2015, a period of nearly 16 years.

In order to identify papers that had applied at least one of the 26 data analysis techniques, the searching process was performed using each technique's name and some other combinations and derivations of words that could help distinguish these techniques inside the articles.

The limitations of this research lie in the search process. The authors aware that some of the big data analytics techniques listed by MGI (2011) have overlap with each other that could affect the accuracy of the research to some degree. In addition, due to the generality of some of these technique's name, and the potential of using other alternative names (that have not identified), it is possible to miss some of the papers related to specific techniques or misclassified others. To improve the accuracy, the searching process was limited to the topics, abstracts, and keywords of the papers and a final manual inspection were performed on the results.

3.3 Results and Discussion

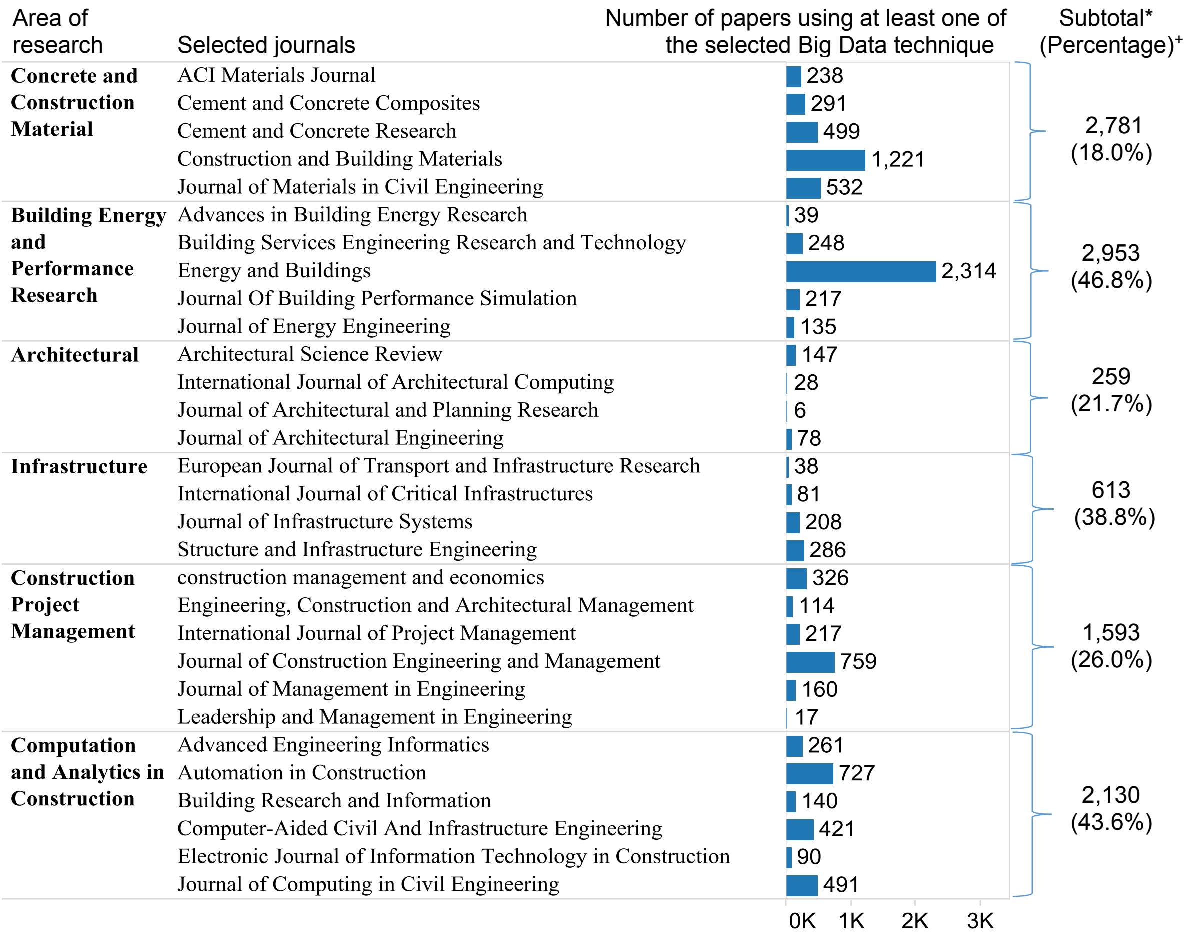
3.3.1 Publications Related to Big Data Techniques

The literature survey began by searching key terms (including "big data" and "construction industry/buildings") and related CAL classification codes in the subject/title/abstract of all the publications included in the database. This led to 48 publications. Based on the goal of this research, manual verification was performed to remove papers not coming from the construction research community (i.e., only counting papers published in construction-related journals or at least one of their co-authors was affiliated with this field). As a result, 10 articles were identified and the topics they study are listed in Table 3.1. The fact that all of these articles were published in the recent three years suggests that big data is still very novel in the construction industry, but researchers started to recognize its importance and value by putting in research efforts. The other 38 publications were authored by researchers in other fields (e.g., computing and information sciences) and published in non-construction related fields.

Table 3. 1. Summary of big data papers authored by construction researchers

Publications	**Main topics**
Sanyal and New 2013	Computer simulation in tuning building energy models
Akhavian and Behzadan 2015	Construction equipment activity recognition
Du et al. 2014	Benchmarking BIM performance
Mahadevan et al. 2014	Structural health monitoring (for concrete structures)
Whyte et al. 2015	Information management (e.g., change management)
Zhu and Ge 2014	Big data for green building
D'Oca and Hong 2015	Occupancy schedules learning using data mining
Elhaddad and Alemdar 2015	Efficient management of big datasets
Tomé et al. 2015	Space-use analysis for buildings
Wong and Zhou 2015	Enhancing environmental sustainability through green BIM

The results of this research on the application of each of the 26 big data techniques show that, in total, these techniques appeared 14,849 times within the 30 selected journals. It is important to notice that many of the publications applied more than one of these techniques. After filtering the identical papers, 10,329 unique papers were identified, each of which at least adopted one of these 26 data analytics techniques. Figure 3.1 presents the selected research areas, associated journals, and the results of the database search.



Area of research	Selected journals	Number of papers using at least one of the selected Big Data technique	Subtotal* (Percentage)+
Concrete and Construction Material	ACI Materials Journal	238	2,781 (18.0%)
	Cement and Concrete Composites	291	
	Cement and Concrete Research	499	
	Construction and Building Materials	1,221	
	Journal of Materials in Civil Engineering	532	
Building Energy and Performance Research	Advances in Building Energy Research	39	2,953 (46.8%)
	Building Services Engineering Research and Technology	248	
	Energy and Buildings	2,314	
	Journal Of Building Performance Simulation	217	
	Journal of Energy Engineering	135	
Architectural	Architectural Science Review	147	259 (21.7%)
	International Journal of Architectural Computing	28	
	Journal of Architectural and Planning Research	6	
	Journal of Architectural Engineering	78	
Infrastructure	European Journal of Transport and Infrastructure Research	38	613 (38.8%)
	International Journal of Critical Infrastructures	81	
	Journal of Infrastructure Systems	208	
	Structure and Infrastructure Engineering	286	
Construction Project Management	construction management and economics	326	1,593 (26.0%)
	Engineering, Construction and Architectural Management	114	
	International Journal of Project Management	217	
	Journal of Construction Engineering and Management	759	
	Journal of Management in Engineering	160	
	Leadership and Management in Engineering	17	
Computation and Analytics in Construction	Advanced Engineering Informatics	261	2,130 (43.6%)
	Automation in Construction	727	
	Building Research and Information	140	
	Computer-Aided Civil And Infrastructure Engineering	421	
	Electronic Journal of Information Technology in Construction	90	
	Journal of Computing in Civil Engineering	491	

* Total number of papers using at least one selected big data technique in each research area
+ Percentage of these papers over the total number of papers in the selected journals for each research area

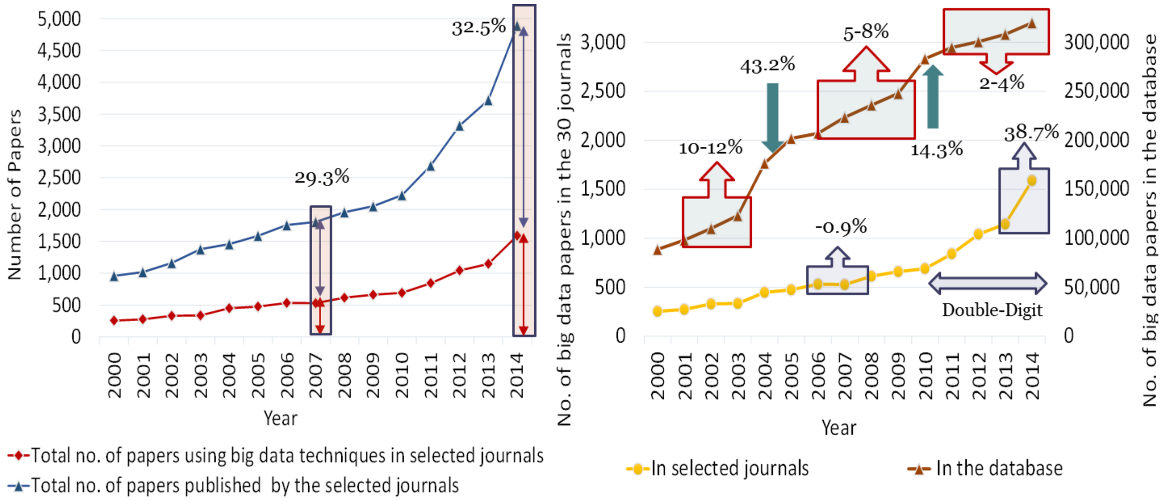
Figure 3. 1. Statistics on papers using big data analytical techniques in selected journals

The information summarized in Figure 3.1 shows that the construction research community has already used some of the selected big data techniques in previous studies to process and analyze available datasets, even though big data is still an emerging concept. It also shows that the frequency of using these techniques in each individual research area varied to some degree. While there is an apparent reason to see more papers (2130, 43.6%) from the field of computation and analytics in construction, it is interesting to learn that 46.8% and 38.8% of the papers in the areas of building energy and performance and infrastructure, respectively, have employed at least one selected analytic technique in their studies. Comparatively, the other three areas had lower adoption rates for big data techniques. It should be noted that the application of data analysis techniques in the field of concrete and construction materials had the lowest ratio (only 18.0%), which indicates a need for improvement.

Figure 3.2a illustrates the overall trends for the application of the 26 big data techniques in the 30 selected journals within the past 15 years compared to the total number of published papers in these journals during the same period. It appears that the number of papers that have applied these techniques has been increasing steadily from 2000 to 2014, but during the same time the total number of papers published in these journals has also increased. By calculating the percentages of papers using big data techniques in the total number of papers published by these journals in each year, this research found that the percentages range from 24.4% to 32.5%. Since 2004, the annual rates slightly fluctuated around 30% with the lowest as 29.3% in 2007 and the highest as 32.5% in 2014. Basically, the advancement in applying big data techniques to construction research has been slow.

Figure 3.2b displays the overall trends for the application of big data techniques in selected journals versus their application in all the engineering research fields included in the database. While both trend lines are going up, the increasing patterns for each trend line are different. For big data usage in construction research, the increase rates fluctuated largely over the years from the lowest of -0.9% in 2007 to the highest of 38.7% in 2014. However, the past four years all had double-digit increases. For the entire engineering research fields, the increase rates were decreased periodically from around 10-12% during 2001-2003, 5-8% during 2007-2009, to only 2-4% during 2011-2014, except for the two unusual, big jumps happening in 2004 (with an increase rate of 43.2%) and 2010 (with an increase rate of 14.3%).

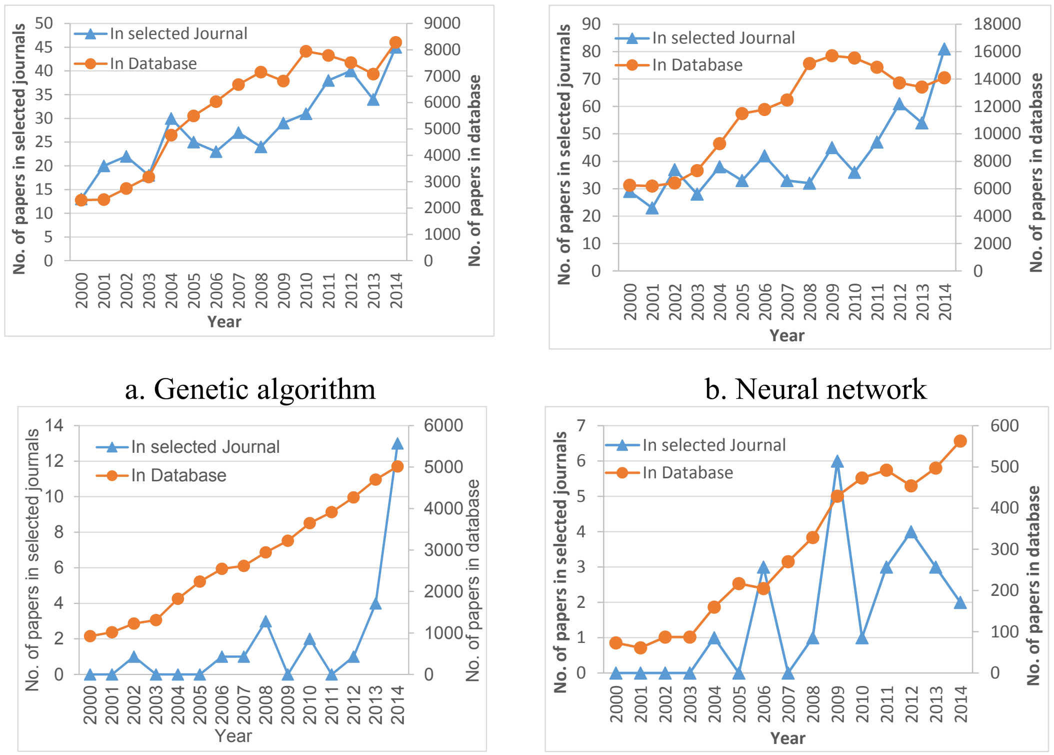
To provide a deeper understanding, trend analysis was performed for each of these 26 techniques (Appendix E). Some examples of this analysis are displayed in Figure 3.3. Note that the displayed trend lines cannot be directly compared due to the different scales used for their associated axes. However, the overall growing trends and fluctuations between years can be easily distinguished. When the number of papers related to a technique was low (e.g., Figures 3.3c and 3.3d), the results had to be carefully interpreted or no trend could be generated due to the limited data.



a. In the selected journals

b. The selected journals vs. the database

Figure 3. 2. Frequency and overall trend for papers applying big data techniques



a. Genetic algorithm

b. Neural network

c. Ensemble learning

d. Spatial analysis

Figure 3. 3. Frequency and trend for the application of big data techniques in selected journals vs. all the journals in the database

3.3.2 Popularity of Big Data Techniques in Each Studied Research Area

Table 3.2 ranks the popularity of the 26 analytical techniques in each area of research according to the number of papers that were applied these techniques in that area. It appears that simulation was the most popular technique in four of the areas including computation and analytics in construction, building energy and performance, architecture, and infrastructure while in the fields of construction project management, and concrete and construction material, statistics and predictive modeling were the most frequently used techniques, respectively. It is also noticeable that simulation, predictive modeling, optimization, statistics, and regression are the top five highest ranked techniques in all of the six areas except for the computation and analytics in construction, which has neural networks and genetic algorithm ranked as fourth and fifth, and the architectural research area, which has visualization ranked the fifth.

The results of research (Appendix B) also suggest that the area of computation and analytics in construction has the highest share for 17 techniques, such as machine learning and pattern recognition, while the other highest usages are associated with the areas of building energy and performance (three techniques: time series analysis, simulation and optimization), concrete and construction materials (two techniques: statistics and predictive modeling), and construction project management (two techniques: regression and network analysis).

Table 3. 2. Popularity ranking of big data techniques in each of the six selected areas

Data Analysis Technique	Computation and Analytics in Construction	Construction Project Management	Concrete and Construction Material	Building Energy and Performance	Architectural	Infrastructure
Simulation	1	2	2	1	1	1
Predictive modeling	3	5	1	2	2	2
Optimization	2	4	5	3	4	3
Statistics	7	1	3	4	3	4
Regression	9	3	4	5	6	5
Neural networks	4	8	6	6	8	6
Classification	8	6	7	8	7	8
Genetic algorithm	5	7	10	7	9	9
Visualization	6	11	8	10	5	12
Time series analysis	13	10	13	9	11	10
Data mining	10	13	17	11	15	15
Network analysis	19	9	11	13	16	7
Pattern recognition	11	14	12	16	12	13
Signal processing	15	16	9	18	17	11
Cluster analysis	16	12	14	12	18	16
Data fusion or data integration	12	15	15	17	19	21
Machine learning	14	21	18	14	10	18
Ensemble learning	18	19	16	15	13	17
Spatial analysis	17	17	19	21	14	14
Natural language processing	20	18	20	20	20	20
Unsupervised learning	22	20	21	22	21	19
Crowd sourcing	23	22	22	19	22	22
Supervised learning	21	23	23	23	23	23
A/B testing	24	24	24	24	24	24
Association rule learning	25	25	25	25	25	25
Sentiment analysis	26	26	26	26	26	26

Figure 3.4 visualizes the usage percentage of each data analysis technique among all the identified papers in each area that used at least one of these 26 techniques. It can be seen that the five aforementioned techniques account for approximately 62% (for the area of computation and analytics in construction) to 80% (for the area of concrete and construction material) of the total papers identified in each research area.

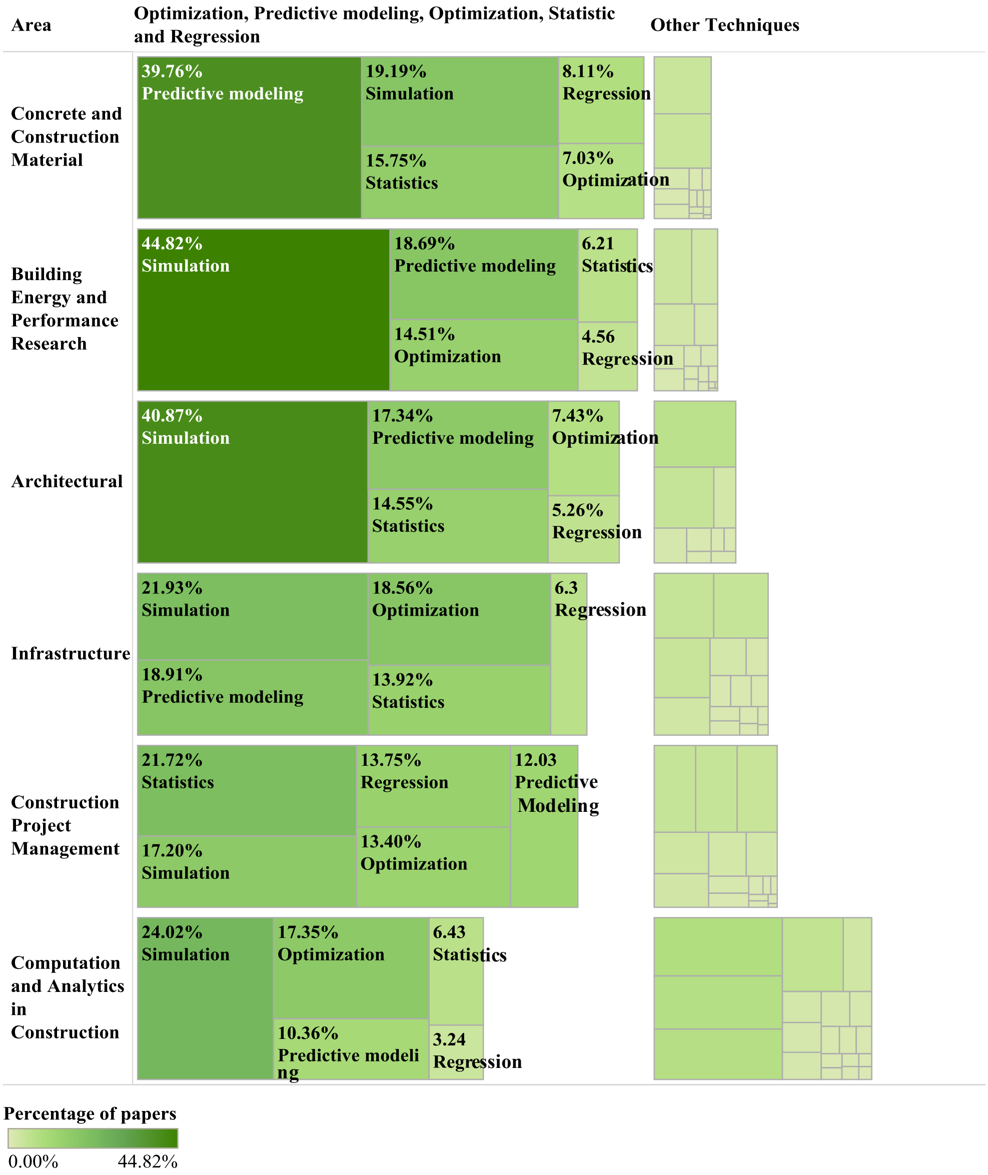


Figure 3. 4. Percentage of papers using the top five most frequently applied techniques compare to the usage of other techniques in each research area

The bar chart in Figure 3.5 presents the percentage of papers that used each of 26 techniques over the total number of papers in 30 journals. According to this chart, simulation, predictive modeling, optimization, statistics, and regression were applied in 11.85, 8.76, 5.57, 4.85, and 2.78 percent of all the papers published by these 30 journals, respectively. In fact, the authors believe that if it was not because of the search methodology limitation, these percentages could have been larger than what is presented here. These five techniques are not only the most popular ones among these 26 techniques; their scopes may also overlap with some other analytic techniques. For example, regression is a subset of the statistical method and also included in the predictive modeling.

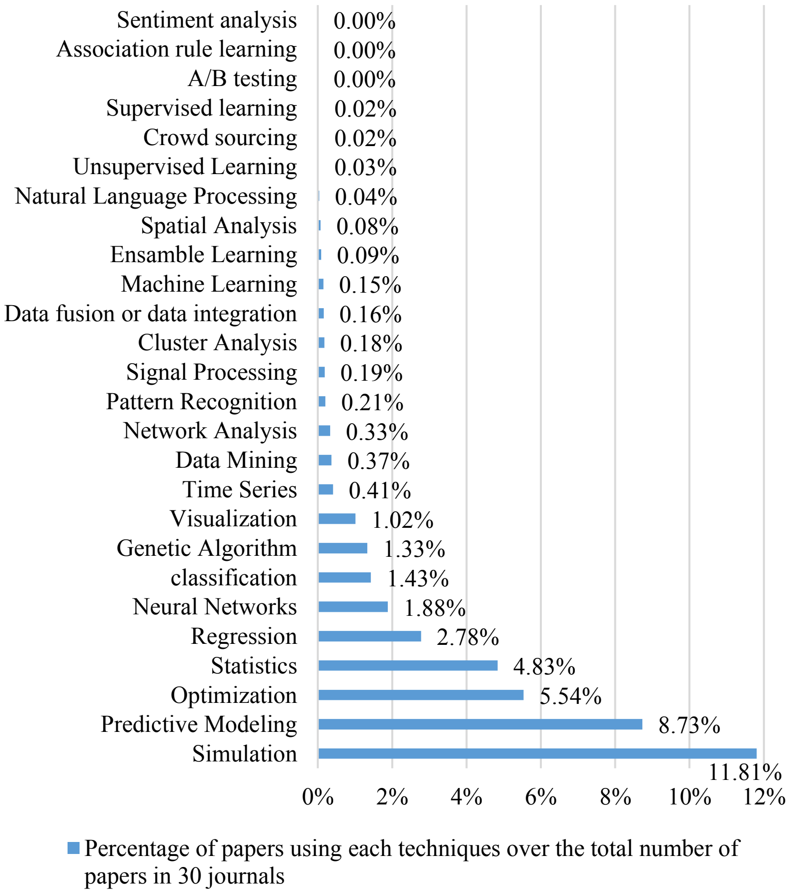


Figure 3. 5. Percentage of papers using each technique in 30 journals

Figure 3.5 also shows that six techniques, including supervised learning, unsupervised learning, association rule learning, crowdsourcing, A/B testing, and sentiment analysis, had a very small sample size of papers, i.e., less than 10 papers, or in the cases of A/B testing and sentiment analysis, even zero in the entire 30 journals. Using scientific judgment, one can claim that analyzing these two groups of analytical techniques (the top five and the lowest six methods) would less likely yield very meaningful usage patterns regarding their applications. As a result, the remaining part of this study focused on the middle 15 analytic techniques.

As shown in Figure 3.6, of these 15 techniques, the neural networks technique had the highest applications in four research areas as computation and analytics in construction, building energy and performance, concrete and construction material, and infrastructure. Visualization in architectural research and classification in construction project management were the techniques with the highest share of papers. Overall, neural networks, classification, genetic algorithm, visualization, time series, data mining, and network analysis were the seven most frequently used data analysis techniques with more than 100 examples of applications (ranging from 117 to 668 papers) in the surveyed journals.

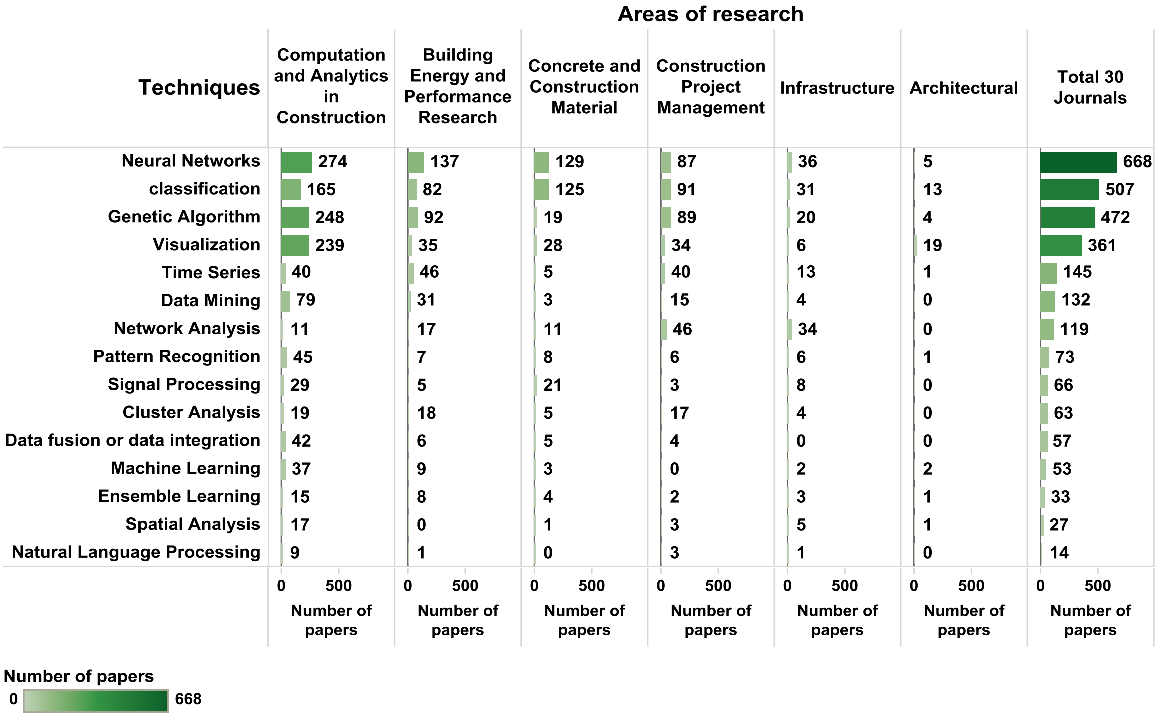


Figure 3. 6. Analytical techniques vs. research area for the selected 15 analytical techniques

Performing in-depth analysis of these 15 techniques in all of the six research areas will require huge amounts of time and effort. This research focused its analysis on two areas, namely computation and analytics in construction and construction project management, which are closely related to the authors' research areas and expertise. In this way more meaningful analysis could be performed and better interpretation of research results could be achieved. Further investigation on other research subareas will be carried out in future research.

3.4 Time-Trend Analysis

Time-trend analysis was performed on the publications within the areas of computation and analytics in construction and construction project management for the 15 selected techniques. Figure 3.7 uses a color table to illustrate the distribution of papers related to each of the techniques and their publication years. A darker color denotes a higher number of records (Frequency of application of top five most used methods in different periods of time can be find in Appendix C).

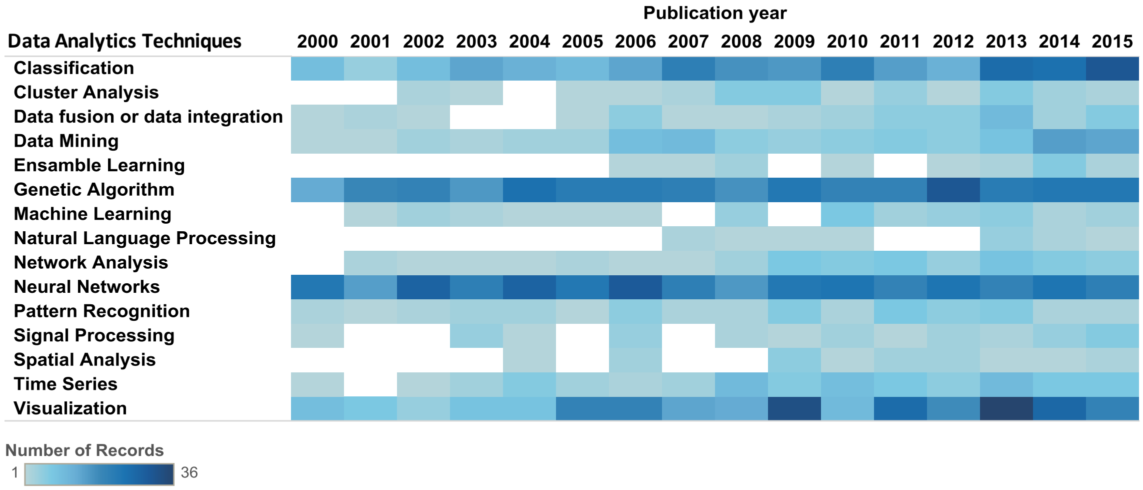


Figure 3. 7. Data analysis techniques vs. publication years

The consistent presence of dark color cells for genetic algorithm and neural networks implies that these two techniques had strong applications throughout the time span covered by this research (from 2000 to September 2015), although it seems that the application of neural networks had already reached its peak 10 years ago. For techniques such as classification and visualization, a gradual increase in their applications can be seen over the years. In addition, while some techniques such as pattern recognition and data mining had been applied for a long time, the implementation of some other techniques (including natural language processing, ensemble learning, and spatial analysis) in these two areas of research started more recently.

Figure 3.8 is a box-and-whisker plot that shows statistics (e.g., range, median, and maximum) about the number of papers that applied each of the 15 data analytical techniques in individual years. For the classification and visualization techniques, their number of publications in each year fluctuated significantly throughout the past 15+ years. For other three techniques, i.e., data mining, genetic algorithm, and neural networks, their numbers fluctuated considerably, but not as significantly as the first two. For the remaining 10 techniques, their annual publication numbers ranged from 0 to10 papers. It is also evident that nine out of these 10 techniques had at least one year in which they did not appear in any publication within these 30 journals.

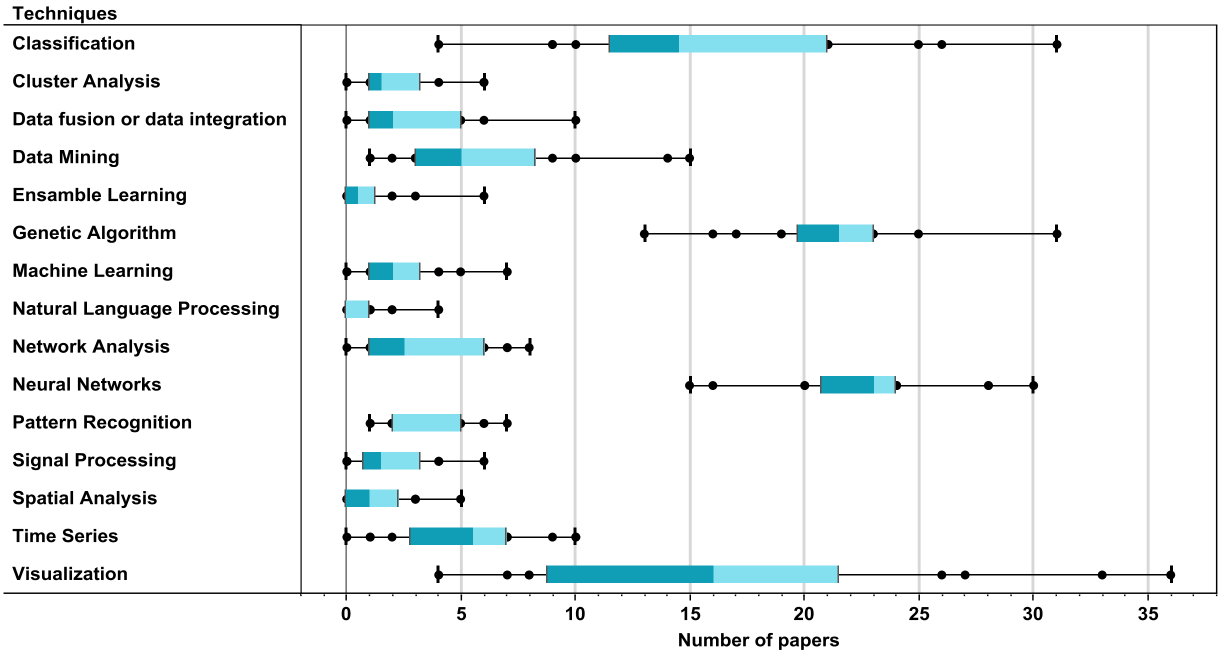


Figure 3. 8. The box-and-whisker plot for publication frequency related to each data analytical technique over the years

Tables 3.3 and 3.4 present the equations that represent the linear trend lines generated for individual analytical techniques based on the number of papers that applied these techniques in each year (Table 3.3), and the ratio of these papers to the total number of papers published by these journals in each year (Table 3.4). The techniques are sorted in descending order based on the slope of their trend lines. That means, for the techniques at the top of the list, e.g., visualization and classification in Table 3, the increase in the number of papers was faster than the ones listed at the bottom. In fact, neural networks located at the very bottom of the list had an overall negative trend during the studied period although the slope was very gentle, implying a slightly decreasing popularity.

Table 3. 3. Trend line equations based on number of papers using individual techniques

Technique*	Equation
Visualization	y = 1.4426x - 2879.1
Classification	y = 1.25x - 2493.4
Data mining	y = 0.6912x - 1381.7
Time series	y = 0.5206x - 1040.1
Network analysis	y = 0.4603x - 920.6
Genetic algorithm	y = 0.4603x - 902.98
Data fusion or data integration	y = 0.3941x - 788.32
Cluster analysis	y = 0.2353x - 470.1
Signal processing	y = 0.2176x - 434.93
Machine learning	y = 0.2103x - 419.85
Ensemble learning	y = 0.2103x - 421.1
Pattern recognition	y = 0.1632x - 324.51
Spatial analysis	y = 0.1559x - 311.68
Natural language processing	y = 0.1353x - 270.85
Neural networks	y = -0.075x + 173.13

* Techniques are sorted based on their equations' slope in descending order

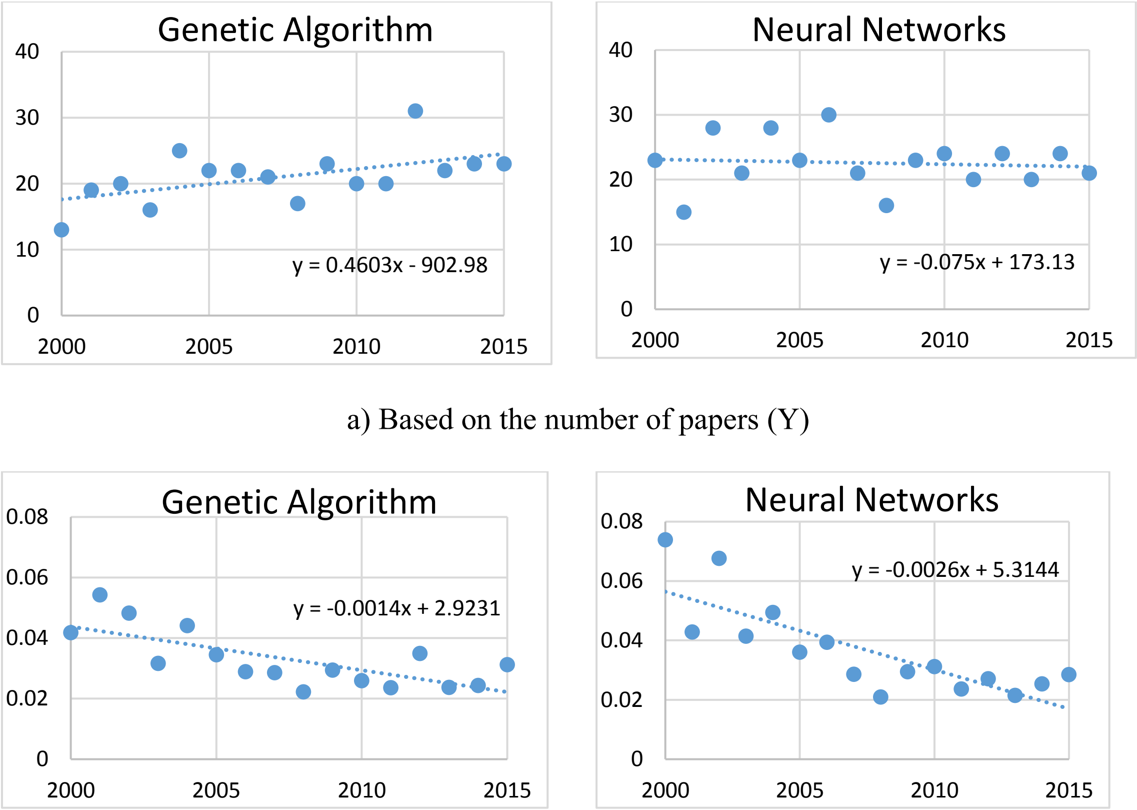
Table 3. 4. Trend line equations based on the ratio of papers using the techniques to total number of papers published in 30 selected journals each year

Technique*	Equation
Visualization	Percentage = 0.0008x - 1.4987
Data mining	Percentage = 0.0006x - 1.2637
Classification	Percentage = 0.0005x - 1.0634
Time series	Percentage = 0.0005x - 0.9465
Network analysis	Percentage = 0.0004x - 0.8679
Data fusion or data integration	Percentage = 0.0003x - 0.6589
Signal processing	Percentage = 0.0002x - 0.3458
Spatial analysis	Percentage = 0.0002x - 0.3621
Cluster analysis	Percentage = 0.0002x - 0.4349
Ensemble learning	Percentage = 0.0002x - 0.4778
Natural language processing	Percentage = 0.0002x - 0.3067
Machine learning	Percentage = 0.0001x - 0.2548
Pattern recognition	Percentage = -0.00003x + 0.0658
Genetic algorithm	Percentage = -0.0014x + 2.9231
Neural networks	Percentage = -0.0026x + 5.3144

* Techniques are sorted based on their equations' slope in descending order.

Table 3.4 provides additional information regarding the trend in the annual application rate of each technique when considering the change in the overall number of papers published by the studied journals each year. The difference between these two tables shows that the trend in annual application rate was different than the trend in the annual number of publications for each of these techniques. For instance, even though classification experienced the second highest increase in the number of papers it was data mining that had the second highest application growth rate in the studied research areas during the past 15+ years.

It appears that neural networks, genetic algorithm, and pattern recognition are losing their share of papers in these research areas even though two of the techniques (Genetic algorithm and pattern recognition) have seen an increasing trend in their number of papers. Figure 3.9 is an example that can help to better understand this subject.



a) Based on the number of papers (Y)

b) Based on application ratio in investigated journals (Y)

Figure 3. 9. Trend lines for neural networks and genetic algorithm

3.5 Analysis of the Research Subjects Studied by Each of the Analytical Techniques

To provide a more in-depth analysis for the two selected areas, the papers associated with these two areas were inspected in more detail and for each paper, a suitable research subject was assigned according to its research purpose. As a result of this process, 178 subcategories of subjects were identified. These subcategories were reviewed to further group them to a more manageable number of categories (Appendix A). The final list of research subjects consists of 84 categories that are presented in Figure 3.10.

Out of these 84 research subjects, three of them including general research related to the construction industry, project management, and lean construction were strictly associated with the area of construction project management. Forty-five of the research subjects are solely observed in the journals related to computation and analytics in construction and the remaining 36 categories were common subjects in both areas of research. Table 3.5 presents the 45 research subjects related to the area of computation and analytics in addition to the number of papers in each subject.

Table 3. 5. Research subjects solely related to the area of computation and analytics

Research Subjects	No. Of Papers	Research Subjects	No. Of Papers	Research Subjects	No. Of Papers
Traffic and Transportation Management	73	Construction Simulation	13	Spatial Analysis	6
Pavement Evaluation	29	Document Identification and Management	12	Environmental Issue (Air and Water Pollution)	6
Structural analysis	24	Process Management	12	Fault detection	5
Predicting Concrete Properties	24	Water distribution systems	11	Robotics	5
System Identification and Analysis	22	Non-Construction Related Paper	11	Construction litigation	4
Damage detection	20	Positioning System	10	Slope stability	4
Accident Prevention and Road Safety	20	Tower Crane Operation	9	Space Use Analysis	4
Building energy Performance	20	Excavation	8	Aggregate classification	3
Text processing	20	Soil Typological Classification	8	Deterioration Detection	2
Action/Object Recognition and Image Processing	20	Conflict Management	7	Rework Analysis	2
Seismology	19	Thermal Analysis and HVAC Analysis	7	Tunneling	2
Material Analysis	18	Lifecycle Analysis	6	Delay Analysis	2
Other Modeling Research	17	Corrosion Detection	6	Geotechnical engineering	2
Crack Detection	14	Innovation Assessment	6	Problem Solving	2
Building information modeling BIM	14	Disasters Management and Emergency Response	6	Transaction Deletion	2

Research Subjects	Cluster Analysis	Classification	Data Fusion or Data Integration	Data Mining	Ensemble Learning	Genetic Algorithm	Machine Learning	Natural Language Processing	Neural Networks	Network Analysis	Pattern Recognition	Signal Processing	Spatial Analysis	Time Series	Visualization	Research Subjects	Cluster Analysis	Classification	Data Fusion or Data Integration	Data Mining	Ensemble Learning	Genetic Algorithm	Machine Learning	Natural Language Processing	Neural Networks	Network Analysis	Pattern Recognition	Signal Processing	Spatial Analysis	Time Series	Visualization
Model and Algorithm Improvement	0	2	0	5	0	17	4	0	21	0	3	0	0	4	3	Project Collaboration and Communication	0	3	0	0	0	0	0	0	1	0	0	0	0	0	8
Design	1	5	1	3	2	35	3	1	8	1	3	1	1	3	33	Soil Typological Classification	0	3	0	0	0	0	0	0	4	0	1	0	0	0	0
Traffic and Transportation Management	4	7	3	3	1	21	1	0	18	1	1	3	0	7	3	Conflict Management	0	2	1	0	0	1	0	0	0	0	1	0	0	0	2
Infrastructure Management	2	1	2	4	1	17	1	0	11	0	1	2	2	1	2	Positioning System	0	1	4	0	0	2	1	0	0	0	1	0	0	0	1
Pavement Evaluation	1	4	0	2	1	2	1	0	11	0	1	3	0	1	2	Production Management	0	2	0	1	0	7	0	1	3	0	0	0	0	0	0
Defect Detection	1	9	1	3	0	1	1	0	6	0	4	4	1	0	1	Project Planning	0	0	0	0	0	1	0	0	3	1	0	0	0	1	2
Knowledge Discovery and Information managemen	0	24	8	18	1	1	3	0	7	0	3	1	1	2	19	Resource Management	0	0	0	0	0	6	0	0	0	1	0	0	0	0	1
Project Scheduling	0	2	0	0	0	18	0	0	3	0	0	0	0	0	6	Corrosion Detection	0	3	0	0	0	0	0	0	2	0	1	0	0	0	0
Structural analysis	0	1	0	0	0	7	0	0	13	0	1	0	0	1	1	Innovation Assessment	1	2	0	1	0	0	0	0	2	0	0	0	0	0	0
Economical Analysis	0	6	2	6	1	16	0	0	12	1	0	0	0	0	2	Miscellaneous Topics	0	1	0	1	0	1	0	0	2	0	0	0	0	0	3
Monitoring and Controlling Projects	0	3	3	3	0	1	3	0	7	0	5	2	1	6	19	Fault detection	0	1	0	0	0	0	1	0	2	0	0	0	0	0	1
Seismology	0	0	0	0	1	1	0	0	8	1	0	2	0	2	4	Robotics	0	1	0	0	0	0	0	0	1	0	0	0	0	0	3
Damage detection	1	4	0	0	0	3	1	0	4	0	2	3	0	2	0	Bidding Strategy	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0
System Identification and Analysis	1	4	1	2	2	6	1	0	4	0	0	0	0	0	1	Disasters Management and Emergency Response	0	0	1	0	0	1	0	0	1	0	0	0	0	0	3
Decision Making	0	1	0	1	0	4	0	0	5	0	1	0	1	0	4	Human Resource Management	0	1	0	0	0	4	0	0	0	0	0	0	0	0	0
Material Analysis	0	7	0	2	1	2	0	0	3	0	1	1	0	0	1	Other Modeling Research	0	3	1	1	1	0	0	0	4	0	0	0	0	2	5
Safety Management and Occupational health	0	3	2	0	0	0	0	0	0	0	0	1	2	0	12	Procurement Management	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
Accident Prevention and Road Safety	1	4	0	0	0	3	0	0	8	0	2	1	0	1	0	Review	0	1	0	0	0	2	0	0	2	0	0	0	0	0	1
Building energy Performance	1	3	2	5	1	2	2	0	1	0	0	1	0	0	2	Tower Crane Operation	0	0	0	0	0	1	0	0	1	0	0	1	0	0	6
Crack Detection	0	2	0	0	0	3	0	0	5	0	2	1	0	0	1	Construction litigation	0	1	0	1	0	0	0	0	2	0	0	0	0	0	0
Layout Planning	1	0	0	0	0	13	1	0	1	0	0	0	2	0	5	Dispute Management	0	1	0	2	1	0	1	1	3	0	0	0	0	0	0
Productivity Analysis	0	3	5	0	0	0	1	0	6	0	1	0	0	0	0	Market/Customer Analysis	2	0	0	1	0	0	0	0	2	0	0	0	0	0	0
Artificial Reality	0	2	0	0	0	0	1	0	2	0	1	0	0	0	23	Slope stability	0	0	0	0	0	2	0	0	1	0	0	0	1	0	0
Water distribution systems	0	0	0	0	0	6	0	0	4	1	0	0	0	0	0	Space Use Analysis	0	1	1	0	0	1	0	0	0	0	0	0	1	0	0
Risk Analysis and Management	0	1	0	0	0	1	0	0	4	1	0	0	0	0	2	Spatial Analysis	0	1	0	0	0	0	0	0	1	0	0	0	4	0	0
Text processing	0	8	0	3	0	0	2	4	2	0	1	0	0	0	0	Thermal Analysis and HVAC Analysis	1	0	1	0	0	1	0	0	3	0	0	0	0	1	0
Predicting Concrete Properties	0	1	0	1	0	5	1	0	13	0	1	0	0	1	1	Deterioration Detection	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Project Performance Analysis	0	0	0	0	0	4	0	0	5	0	0	0	0	0	4	Rework Analysis	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0
Building information modeling BIM	0	1	0	1	0	1	1	0	0	0	0	0	0	0	10	Tunneling	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
Network Modeling and Optimization	0	0	0	0	0	8	0	0	2	1	0	0	0	0	0	Aggregate classification	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0
Sustainable Development and Waste Management	0	2	0	0	1	3	0	0	2	0	0	0	0	1	5	Delay Analysis	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
Construction Simulation	1	0	0	0	0	0	0	0	0	0	0	0	0	0	12	Environmental Issue (Air and Water Pollution)	0	1	0	0	0	1	0	0	3	0	0	0	0	1	0
Process Management	0	1	0	2	0	1	1	0	2	1	0	0	0	1	3	Geotechnical engineering	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Project Execution and Operation	0	0	0	0	0	3	0	1	2	0	0	1	0	0	5	Organizational Issue	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Education	0	1	0	0	0	0	0	0	0	0	0	1	0	0	8	Problem Solving	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
Action/Object Recognition and Image Processing	0	7	0	1	0	1	3	0	2	0	3	0	0	0	3	Transaction deletion	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Document Identification and Management	0	7	0	1	0	0	1	0	1	0	1	0	0	0	1	Construction Equipment/Machinery Managemer	0	1	0	1	0	1	0	0	1	0	0	0	0	0	0
Excavation	0	1	0	0	0	0	0	0	4	0	1	0	0	0	2	Contract Manangement	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Life cycle Analysis	0	0	0	0	0	1	0	0	2	0	0	0	0	1	2	Construction Industry	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Maintenance and Facilities Management	0	0	3	0	0	1	0	0	3	0	0	0	0	1	1	Lean Construction	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Contractor qualification	0	2	0	0	0	0	0	0	5	0	0	0	0	0	0	Project Management	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Non Construction Related Paper	0	1	0	2	0	2	0	1	4	0	1	0	0	0	0	Stakeholder Management	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

Figure 3. 10. Research subject vs. analytical techniques

Figure 3.11 present the research subjects found in both areas in addition to the portion of the papers related to each of the areas. It can be seen that for the 17 of the subjects, the area of CPM has the highest share of papers, while for the other 19 subjects, the computation and analytics area has the majority of the share.

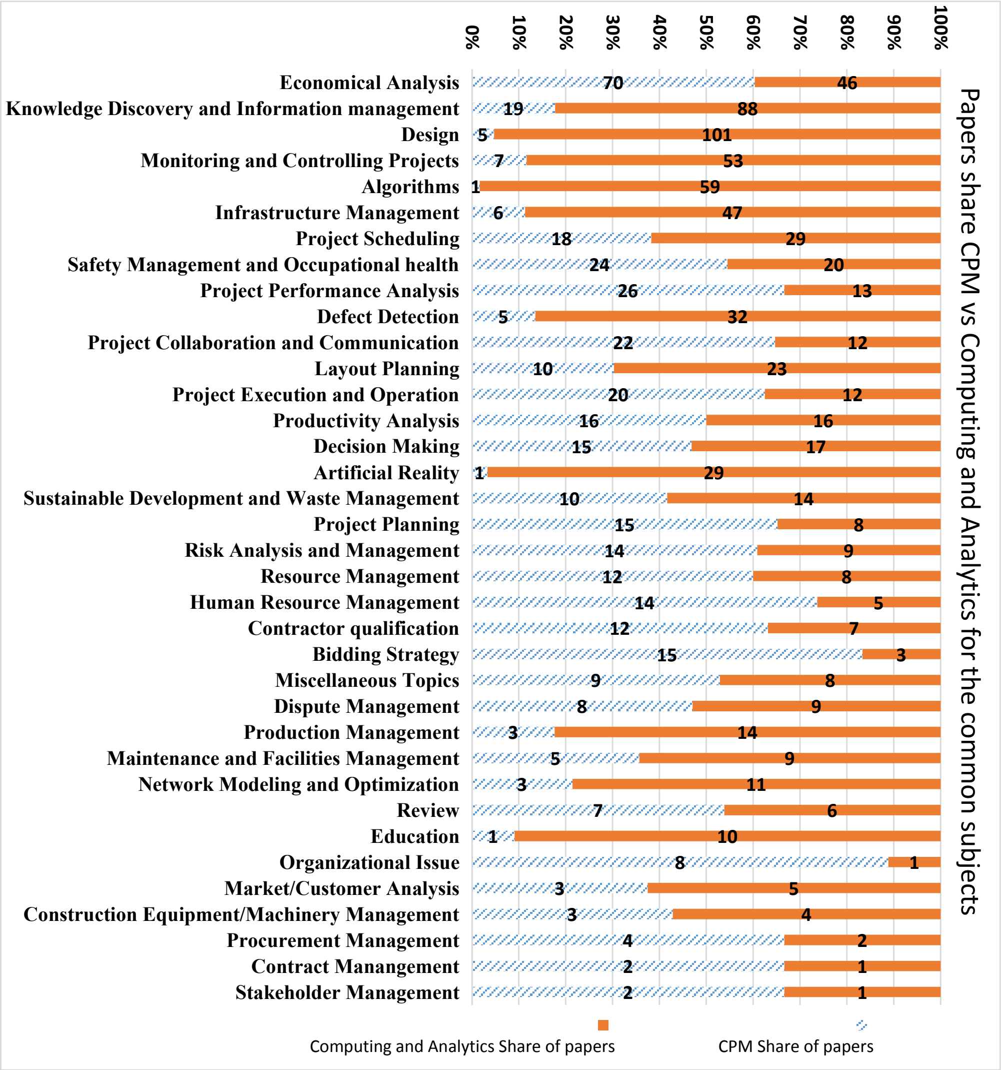


Figure 3. 11. Common subjects found in both construction project management and computation and analytics areas

3.6 Subjects vs. Area

Figure 3.12 presents the top 50 percent of the topics for the two research areas and for the combined dataset including both. There are three subjects with more than one hundred applications of these 15 techniques.

The first subject is the economic analysis, which includes subcategories like project cost analysis, cost-benefit analysis, investment analysis, time-cost analysis, and asset management. This research subject had the highest application of these 15 techniques both in the construction project management and the entire dataset. Economic analysis was the subject of research for almost 16 percent of the research in the construction project management area and 6.8 percent of the papers in the combined dataset.

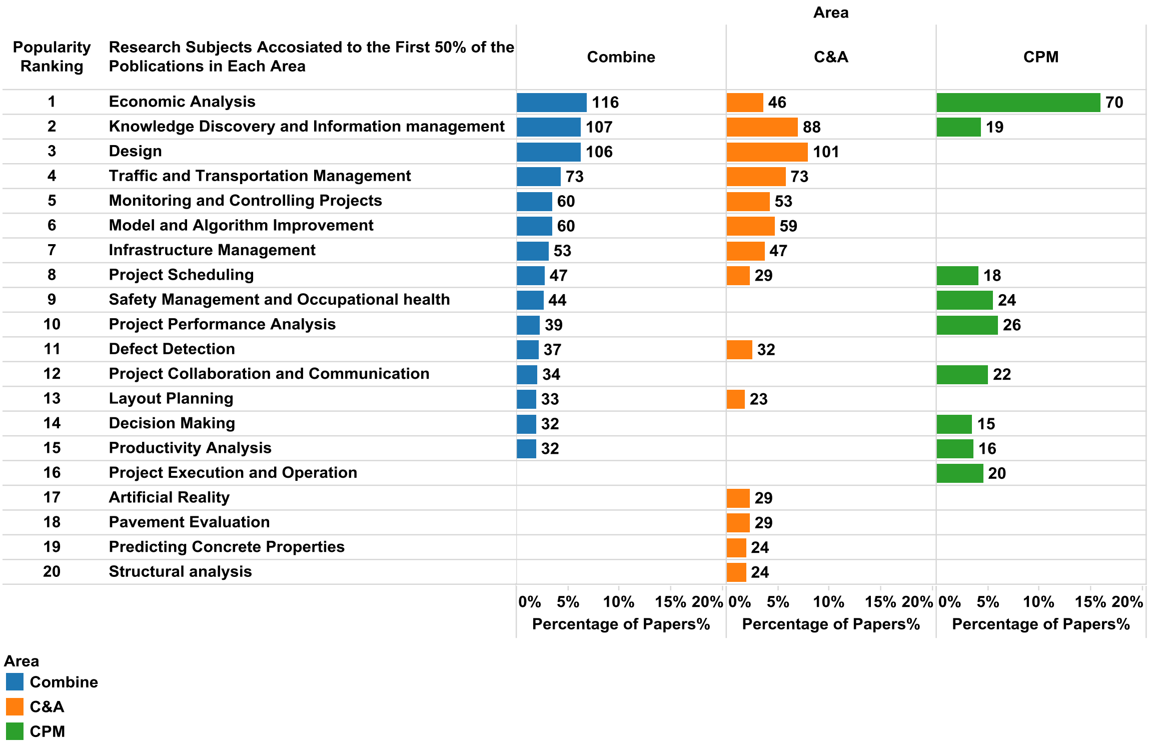


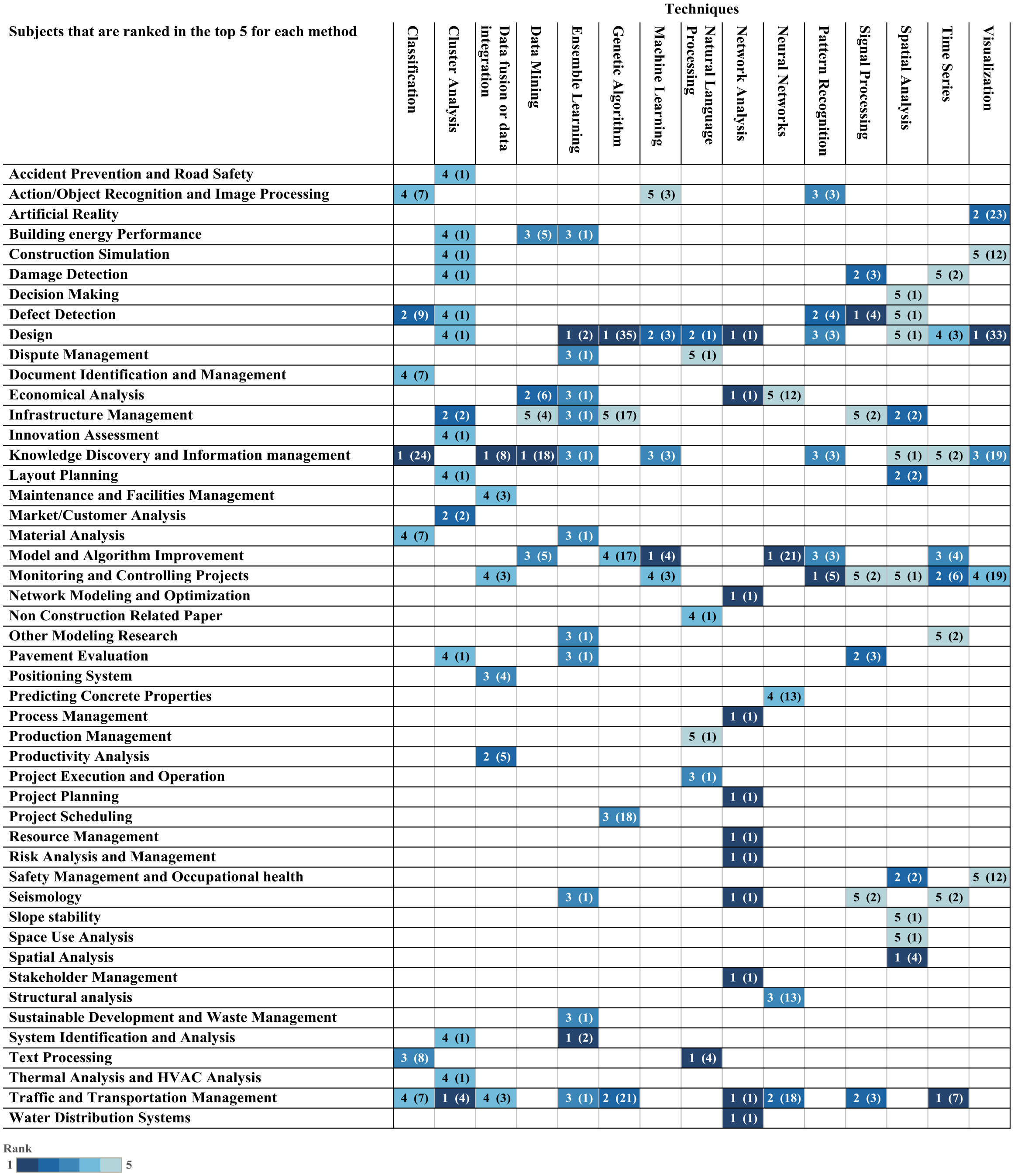
Figure 3. 12. Top 50 percent of the subjects for each research area

The second highest number of papers is related to the subject of knowledge discovery and information management, which covers subcategories such as data collection, data visualization, knowledge discovery, knowledge management, data integration, and information technology. The third subject with more than one hundred examples of applications is the design that includes design analysis, architectural design, collaborative design, structural design, highway design and green building design. The design was the most popular subject of research (using the selected 15 data analytical techniques) in the area of computing and analytics in construction with 101 related papers while there were only 5 papers in the construction project management area with this subject.

It is notable that only three of the 15 highest popular subjects including economic analysis, knowledge discovery and Information management, and project scheduling are listed in top 50 percent of the subjects in both areas of research. It also appears that number four most popular subject, which is traffic and transportation management, is not part of the construction project management scope of research since there was not any paper related to this subject in the construction project management area. As mentioned before, the area of computation and analytics had the broader range of research topics consists of 81 different subjects while the area of construction project management had 39.

3.6.1 Top Five Subjects for Each Technique

Figure 3.13 presents the top five research subjects that applied each of the 15 selected analytical techniques. Notice that the number of subjects is more than five for some of the techniques. This is because some of the subjects have the same number of papers for a specific technique and consequently have the same rank. This is particularly more visible for the network analysis since there were 11 subjects that applied this technique only once. As a result, all the 11 subjects were ranked first and presented in this table.



Subjects that are ranked in the top 5 for each method	Techniques														
	Classification	Cluster Analysis	Data fusion or data integration	Data Mining	Ensemble Learning	Genetic Algorithm	Machine Learning	Natural Language Processing	Network Analysis	Neural Networks	Pattern Recognition	Signal Processing	Spatial Analysis	Time Series	Visualization
Accident Prevention and Road Safety		4 (1)													
Action/Object Recognition and Image Processing	4 (7)						5 (3)				3 (3)				
Artificial Reality															2 (23)
Building energy Performance		4 (1)		3 (5)	3 (1)										
Construction Simulation		4 (1)													5 (12)
Damage Detection		4 (1)										2 (3)		5 (2)	
Decision Making													5 (1)		
Defect Detection	2 (9)	4 (1)									2 (4)	1 (4)	5 (1)		
Design		4 (1)			1 (2)	1 (35)	2 (3)	2 (1)	1 (1)		3 (3)		5 (1)	4 (3)	1 (33)
Dispute Management					3 (1)			5 (1)							
Document Identification and Management	4 (7)														
Economical Analysis				2 (6)	3 (1)				1 (1)	5 (12)					
Infrastructure Management		2 (2)		5 (4)	3 (1)	5 (17)						5 (2)	2 (2)		
Innovation Assessment		4 (1)													
Knowledge Discovery and Information management	1 (24)		1 (8)	1 (18)	3 (1)		3 (3)				3 (3)		5 (1)	5 (2)	3 (19)
Layout Planning		4 (1)											2 (2)		
Maintenance and Facilities Management			4 (3)												
Market/Customer Analysis		2 (2)													
Material Analysis	4 (7)				3 (1)										
Model and Algorithm Improvement				3 (5)		4 (17)	1 (4)			1 (21)	3 (3)			3 (4)	
Monitoring and Controlling Projects			4 (3)				4 (3)				1 (5)	5 (2)	5 (1)	2 (6)	4 (19)
Network Modeling and Optimization									1 (1)						
Non Construction Related Paper								4 (1)							
Other Modeling Research					3 (1)									5 (2)	
Pavement Evaluation		4 (1)			3 (1)							2 (3)			
Positioning System			3 (4)												
Predicting Concrete Properties										4 (13)					
Process Management									1 (1)						
Production Management								5 (1)							
Productivity Analysis			2 (5)												
Project Execution and Operation								3 (1)							
Project Planning									1 (1)						
Project Scheduling						3 (18)									
Resource Management									1 (1)						
Risk Analysis and Management									1 (1)						
Safety Management and Occupational health													2 (2)		5 (12)
Seismology					3 (1)				1 (1)			5 (2)		5 (2)	
Slope stability													5 (1)		
Space Use Analysis													5 (1)		
Spatial Analysis													1 (4)		
Stakeholder Management									1 (1)						
Structural analysis										3 (13)					
Sustainable Development and Waste Management					3 (1)										
System Identification and Analysis		4 (1)			1 (2)										
Text Processing	3 (8)							1 (4)							
Thermal Analysis and HVAC Analysis		4 (1)													
Traffic and Transportation Management	4 (7)	1 (4)	4 (3)		3 (1)	2 (21)			1 (1)	2 (18)		2 (3)		1 (7)	
Water Distribution Systems									1 (1)						

Figure 3. 13. Top 5 research subjects that were investigated by each technique

Note: The numbers presented in figure 10 are the rank (First number) and the number of papers (Second number inside parenthesis) identified for each subject in regard to each specific technique.

The subject of design had the highest application for three of these analytical techniques including genetic algorithm, visualization, and ensemble learning. Although ensemble learning also had the same number of applications in the field of system identification and analysis, knowledge discovery and information management had the highest usage of classification, data fusion and integration, and data mining techniques. Traffic and transportation management was the most frequent subject for cluster analysis and time series analysis. Model and algorithm improvement was the most popular objective for machine learning and artificial neural networks. Pattern recognition mostly used in the research related to monitoring and controlling the projects. Signal processing had the highest application in defect detection. Spatial analysis was mostly used for spatially related investigations and finally, natural language processing was the technique most frequently used for the text processing.

Table 3.6 provides information regarding the number of subjects that used each of these 15 techniques. It appears that artificial neural network and classification had the broadest range of applications and were involved with 67 and 66 different research subjects, respectively. On the other hand, natural language processing had the most limited range, with only 8 research subjects.

Table 3.7 is a crosstab that tabulates the frequency of the subjects that applied each certain number of analytical techniques. For instance, it can be seen that there were three subjects that only applied one of these analytical techniques. These subjects are deterioration detection, geotechnical engineering, and transaction deletion and the techniques they used are the artificial neural network for the first two and classification for the latest.

Table 3. 6. Number of the subjects that applied each of the 15 analytical techniques

Techniques	Number of Research Subjects
Artificial Neural Networks	67
Classification	66
Genetic Algorithm	57
Visualization	50
Data Mining	36
Time Series	32
Pattern Recognition	31
Cluster Analysis	24
Machine Learning	24
Network Analysis	21
Data Fusion or Data Integration	20
Signal Processing	18
Ensemble Learning	14
Spatial Analysis	12
Natural Language Processing	8

Table 3. 7. Number of the subjects that used each certain number of analytical techniques

Number of Techniques	Number of Research Subjects
1	3
2	8
3	10
4	14
5	9
6	12
7	8
8	8
9	2
10	3
11	2
12	1
13	2
14	0
15	2

The design and knowledge discovery and information management were the two research subjects that had at least one example of using each of these 15 analytical techniques. It is interesting to know that the economic analysis, which had the highest number of papers, only applied nine of these analytical techniques.

3.7 Analysis of Papers with More than One Analytical Techniques

There are many papers that applied more than one analytical technique. A comprehensive analysis was performed to specify these papers, identify the analytical techniques, and find any probable relationship or connection between them. Table 3.8 shows the frequency of the number of techniques used in the papers in each area of research. There was only one paper that applied eight of the investigated analytical techniques, two papers with seven techniques and nine that applied six of the techniques.

Table 3. 8. Frequency of the paper vs the number of techniques applied

Number of Techniques Used in One Paper	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Computation and Analytics in Construction	1289	580	187	49	17	6	1	1
Concrete and Construction Material	2190	472	102	12	3	1	1	0
Construction Project Management	1137	339	89	22	5	1	0	0
Building Energy and Performance Research	2010	729	170	33	10	1	0	0
Infrastructure	415	158	33	7	0	0	0	0
Architectural	200	54	5	0	0	0	0	0

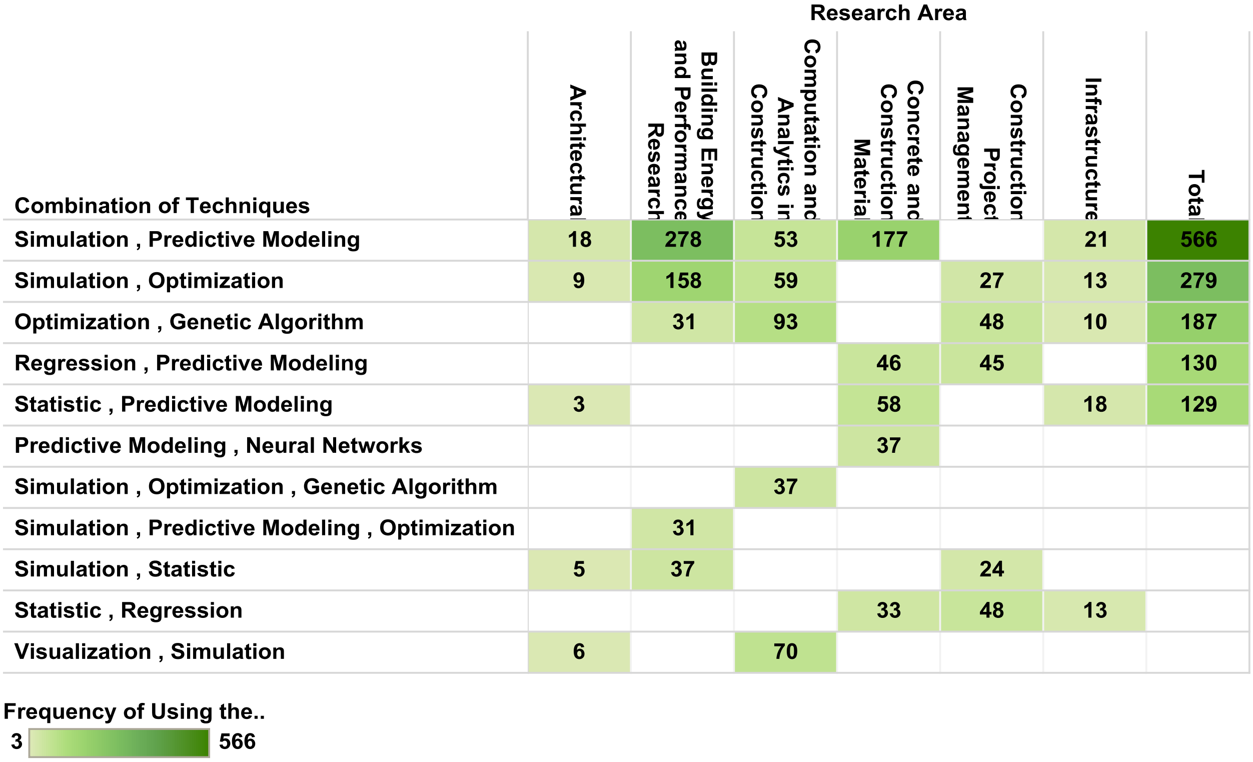
In addition to the 26 individual analytical techniques, 361 different combinations of applying multiple techniques were identified in the investigated journals. As it is presented in Table 3.9, the area of computation and analytics in construction had the highest combination of techniques with 237 different combinations while the architecture research area had the fewest one with only 19. Furthermore, the areas of building energy and performance by 134, construction project management by 105, concrete and construction material by 79, and infrastructure by 61 different combinations each had their share of applying multiple techniques in their publications.

Table 3. 9. Number of unique combinations of techniques in each research area

Research Area	Number of Unique Combinations of Techniques	Number of combinations over total number of papers in each area	Maximum Number of Techniques that Applied in a paper in each area
Computation and Analytics in Construction	237	11.1%	8
Concrete and Construction Material	79	2.8%	7
Construction Project Management	105	6.6%	6
Building Energy and Performance Research	134	4.5%	6
Infrastructure	61	10.0%	4
Architectural	19	7.3%	3

Despite using the fewest number of combinations, the two areas of architecture, and infrastructure had the second and third highest rate for the number of combinations over the total number of papers. Area of concrete and construction material had the lowest rate in this regard. Table 3.9 also presents the maximum number of techniques that were used in a paper in each research area. Area of computation and analytics in construction experienced the highest number of techniques applied in one paper (8 techniques) while the maximum number of papers used in the area of architecture was only 3 techniques.

Figure 3.14 presents the top five most popular combinations of the techniques in each of the six research areas in addition to the total database. Simulation and predictive modeling were the two techniques that have been used together more than any other combination of techniques. Their combination was the most popular one in four of the research areas including architecture, building energy and performance, concrete and construction material, and infrastructure. Optimization and genetic algorithm were the most popular combination of techniques in the area of computation and analytics in construction while the same combination as well as the combination of statistics and regression had the highest presence in the area of construction project management.



	Research Area						
Combination of Techniques	Architectural	Building Energy and Performance Research	Computation and Analytics in Construction	Concrete and Construction Material	Construction Project Management	Infrastructure	Total
Simulation , Predictive Modeling	18	278	53	177		21	566
Simulation , Optimization	9	158	59		27	13	279
Optimization , Genetic Algorithm		31	93		48	10	187
Regression , Predictive Modeling				46	45		130
Statistic , Predictive Modeling	3			58		18	129
Predictive Modeling , Neural Networks				37			
Simulation , Optimization , Genetic Algorithm			37				
Simulation , Predictive Modeling , Optimization		31					
Simulation , Statistic	5	37			24		
Statistic , Regression				33	48	13	
Visualization , Simulation	6		70				

Frequency of Using the..
3 566

Figure 3. 14. Top five most popular combinations of the techniques in each research area

A time-trend analysis has been performed for the 84 research subjects. It was hard to distinguish any clear trend due to the limit number of papers per each subject in different years. Despite this fact, there were a few observations that could be useful for specific purposes. For instance, while there were a few subjects with steady number of papers during the research period (e.g. Design, or model and algorithm improvement, etc.), there were many others such as building information modeling (BIM) and building energy performance that had a more recent uplift in the application of these techniques which could be an indication of future research opportunity in these areas.

3.8 Conclusion

This research studied the application of 26 popular big data analytical techniques in the construction research community within the past 15 years. Six different construction research areas including computation and analytics in construction, construction project management, concrete and construction materials, building energy and performance, architectural research and, infrastructure research were selected and 4 to 6 related journals have been assigned to each of these areas. The papers were identified using the keyword search on the topics, abstracts, and keywords of the papers in the selected journals and a final manual check performed on the results.

This research identified 10329 different papers in these six research areas. Trends, directions, and status of application of these analytical techniques in different areas have been analyzed and the results were presented and described using various figures and tables. In addition, a more in-depth analysis was performed for two of these selected areas including construction project management and computation and analytics in construction. As a result of this study, 178 subcategories and 84 categories of subjects were identified and were assigned to the research papers in these areas and finally the relationships and trends between research subjects and analytical techniques have been discussed and illustrates in different figures and tables.

In order to make this research feasible, the authors focused their study of research subjects on only two of the research areas. Future research could be performed for the other areas of construction research. In addition, creating a framework for the automation of the entire process could turn this research to a valuable lesson learn extraction tool and would be a rewarding topic for future research.

Chapter 4. Application of Artificial Neural Network (ANN) in Construction Research – Case Studies

4.1 Introduction

This chapter aims to provide examples (i.e., case studies) of using the data mining techniques, in particular, artificial neural network (ANN), in construction research area. Two different research subjects including prediction of concrete properties and soil erosion in highway slopes were selected for this purpose. According to the analysis results presented in Chapter 3, ANN was ranked first among the 15 data analytical techniques. Also, it is one of the most popular techniques for predicting concrete properties (more than 54% of the papers, i.e., 13 out of 24 papers we reviewed) and one of the three data analysis techniques (the other two techniques were the genetic algorithm and spatial analysis) applied for studying slope stability. Neural network also had the highest application between the 15 selected data analytic techniques analyzed in entire 30 journals and had the highest share of papers in four of the six selected research areas. It is popular and has high familiarity in the research community.

The advantages of using ANN models lie in the fact that it is a self-adaptive model, which does not require any known equations, and can capture both linear and non-linear functional relationships between the input and output variables (Zhang 1998; Omran et. al. 2016). It does not need much formal statistical training to create and use. It is also a well-known and well-established technique that can consider all possible interactions even among the input variables (Tu 1996). As a result, ANN was selected to be further studied in this dissertation.

The purpose of this chapter was to examine the artificial neural network's ability and accuracy in predictions for the aforementioned two research subjects. Two different datasets were applied in this research. The first dataset includes the information and test results for 144 samples of concrete with different settings or testing ages (Jin 2013). The second one is a dataset containing test results for 442 settings of highway soil erosion with different test sections, rainfall events, and vegetation communities (Cao et al. 2016). A brief description for ANN models and their structure is provided in the case study. For information regarding the background and fundamental and structural details related to ANN, please refer to Chapter 2 Section 2.6. Detail related to each case study including data properties, literature review, research methodology, model development, results, and discussions are presented in their related sections.

4.2 Case Study One: Prediction of Compressive Strength of Environmentally Friendly Concrete Using Artificial Neural Networks

With its growing emphasis on sustainability, the construction industry is more interested in applying environmentally friendly concrete, also known as “green” concrete, in its construction projects. Among other benefits, concrete made with alternative or recycled waste material can reduce pollution and energy use, as well as lower the cost of concrete production. However, the impacts of these alternative materials on concrete properties have not been fully understood, which limits the wide applications of environmentally friendly concrete in practice. This study investigates the application of ANN to predict the compressive strength of concrete made with alternative materials such as fly ash, Haydite lightweight aggregate, and Portland limestone cement. A feed-forward Multilayer Perceptron (MLP) model was applied for this purpose. To determine the accuracy and flexibility of this approach, two different input methods (relative and numerical) were tested on the generated ANN models. The results showed that concrete made of Portland limestone cement had slightly better CS than concrete made of Portland cement. Generally, both input methods provided adequate accuracy to predict CS. It was also observed that a proper MLP model with one hidden layer and sufficient neurons (depending on the input variables and type of cement) could effectively predict the CS of environmentally friendly concrete.

4.2.1 Introduction

The construction industry has observed an increasing shift toward sustainability in recent years. Many companies are proactively using or are required by their clients to use more environmentally friendly building materials and/or processes to reduce the negative environmental impact from construction activities. Environmentally friendly concrete, in this study, is defined as concrete produced using alternative and/or recycled waste materials. This type of concrete is increasingly becoming a common element that can be used to help the construction industry achieve long-term sustainability, although the impact of these alternative or recycled waste materials on various concrete properties has not been fully understood.

The compressive strength of concrete is one of the most important properties in concrete design. Many experiments have been undertaken to study the CS of environmentally friendly concrete that is made of alternative and/or recycled waste materials (Yang et al. 2005; Etxeberria et al. 2007; Kevern et al. 2011). Despite some progress, the available data for such concrete is far from adequate due to the emergence of various alternatives or recycled waste materials and the complexity of concrete mixture design. Not only is more research needed to advance the understanding of the properties of environmentally friendly concrete, but practical tools for designing such concrete are necessary for it to be widely implemented.

Differing from the traditional experimental approach, some researchers have proposed mathematical or statistical models to predict the CS of concrete given its mixture or based on the fresh concrete properties (Atici 2011). The statistical modeling approach is limited in that the underlying relationships between selected variables have to be known for the researchers to build an acceptable model. In contrast, ANN is a self-adaptive method that can learn and capture the linear or non-linear functional relationships among the variables even when such relationships are hard to identify (Zhang 1998). Due to this advantage, some studies have employed ANN to predict the CS of concrete (Topçu and Saridemir 2007; Saridemir et al. 2009; Atici 2011) and the results of these studies have generally confirmed ANN to be a powerful method for this application.

This study aims to investigate the application and performance of ANN as a tool to provide a more accurate estimation for the CS of environmentally friendly concrete. The ultimate goal, if not totally eliminating the need for the experimental determination of the CS or other concrete properties in the future, is to significantly reduce such a need, which will save time and money for the industry. This is extremely helpful for implementing new materials since extensive experimental data may not be available for them. This chapter first introduces a unique composition of "green" concrete, based on which the structure of the generated ANN models (e.g., type of activation function, the number of hidden layers and nodes, etc.) is optimized. Then, it compares the prediction accuracy of these models based on two different input methods (i.e., relative and numerical), which has not been attempted in existing studies.

4.2.2 Environmentally Friendly Concrete

Traditionally, the four main ingredients used to make concrete are water, cement, fine aggregate (sand) and coarse aggregate, although this can be changed depending on the specific properties (e.g., higher compressive or tensile strength, more durability, or lighter volumetric mass density) that are needed for concrete. In these cases, some alternative materials will be added or used to replace certain amounts of the traditional ingredients. For "green" concrete, the commonly used alternative materials are those that contain recycled contents, reduce greenhouse gasses in their production, reserve natural resources, and are locally available to decrease transportation costs or improve material performance during their life cycles. For this research, Portland limestone cement (PLC), Haydite lightweight aggregate (LWA), and fly ash (FA) Class F were selected as environmentally friendly alternatives to the traditional ingredients. These three alternative materials are chosen based on the literature review and the results of a survey that was performed by the research team to identify industry interests in using environmentally friendly concrete. The results of the survey were presented in a different paper and are not included within the scope of this study.

4.2.3 Properties of Alternative Concrete Materials

PLC is an eco-friendly alternative to Portland cement (PC). It is produced by blending PC with limestone, or inter-grinding PC clinker, limestone, and calcium sulfate (Thomas et al., 2010). PLC can significantly reduce CO_2 emissions during cement manufacturing by reducing the clinker content in PC (Kenai et al., 2004). According to *Concrete Monthly* (2004), incorporating 2.5% limestone in the PC can lead to an annual reduction of 11.8 trillion BTUs in energy use, 2.5 million tons reduction in CO_2 emissions, and 190,000 tons reduction in cement kiln dust in the U.S. The PLC Type GUL used in this research was acquired from the Lafarge cement plant located in Ontario, Canada.

FA is a byproduct from coal-fired power plants. It is the commonly used mineral admixture for general purpose concrete. As the most commonly used Supplementary Cementitious Material (SCM) in the concrete industry, FA Class F was adopted in this study. The chemical and physical analyses were provided by the local supplier and met requirements specified by ASTM C 618 and AASHTO M 295.

Haydite LWA is produced by expanding shale in a rotary kiln, at temperatures over 1000°C. It was originally developed in 1908 and patented in 1918, and since then has been used in many different applications such as concrete masonry, high-rise buildings, and precast and pre-stressed concrete elements. According to the Expanded Shale, Clay and Slate Institute (ESCSI, 2007) some of the advantages of using Haydite LWA include: higher strength and durability of the concrete products, aesthetic value, more feasible design, and improvement in thermal performance. In this study, Haydite size B with a maximum size of 3/8 inch (which was comparable to pea gravel) was acquired from a local hydraulic press brick company.

The conventional concrete materials used as control group were: PC type (I/II), with a 28-day CS at 5.54 ksi; brown sand as fine aggregate (fineness modulus at 2.48); and pea gravel with maximum size at 3/8 inch. Micro Air was used as the air-entraining agent (AEA) to increase the air content in the concrete batches.

4.2.4 Application of Artificial Neural Networks (ANN) in Previous Research

Applying ANN in solving construction-related problems has become a significant area of research in recent years. For concrete-related research, ANN has been used to predict fresh and hardened properties of concrete products (Alshihri 2009; Saridemir et al. 2009; Abdeen 2010; Atici 2011; Khan 2012). Specifically, Saridemir et al. (2009) employed ANN and fuzzy logic to predict the effect that using ground granulated blast furnace slag would have on the CS of concrete. A comparison between multivariable regression analysis and ANN approaches provided by Atici (2011) identified the effectiveness of these methods for predicting the strength of mineral admixture concrete. Khan (2012) also developed an ANN model for predicting several properties of high-performance concrete, including CS, tensile strength, gas permeability, and chlorination penetration values. ANN has proven to be effective on the prediction of properties of locally produced LWA concrete (Abdeen 2010) and structural lightweight concrete (Alshihri 2009).

4.2.5 Experimental Design and Data Collection

In this study, 36 different batches of concrete were mixed. Each batch contained different substitution rates (SRs) of FA (0%, 20%, 30% or 40% by weight) and Haydite LWA (0%, 33%, 67% or 100% by volume) in addition to the use of different Types (PC or PLC) and quantities of cementitious materials. In this way, the effect of the alternative materials on the CS of "green" concrete can be examined more accurately. Besides these three variables, the actual water-cement (W/C) ratio, sand-cement (S/C) ratio, sand-coarse aggregate (S/CA) ratio, the amount of AEA (ml per Kg of cement) and the concrete curing age were selected as influential variables for the ANN models to be generated. This study also attempted to use the numerical method for input variables. Table 4.1 shows the range, mean, and standard deviation of the quantity of each raw ingredient used in the experiment.

Table 4. 1. Concrete mixture data set (for one cubic yard or 0.7645 cubic meters of concrete)

Parameter	Min	Max	Mean	StdDev
Age (day)	3.00	90.00	32.00	35.05
Water (Kg)	161.03	161.03	161.03	0.00
PC or PLC (Kg)	173.27	403.70	264.67	78.32
FA (Kg)	0.00	161.48	61.01	55.52
Sand (Kg)	566.99	689.46	587.40	45.96
Pea gravel (Kg)	0.00	573.79	366.86	177.23
Haydite (Kg)	0.00	281.68	101.60	86.97
Micro air (ml)	85.76	103.51	94.64	8.93

All concrete mixed in the experiment was assumed to be air-entrained (intended to be used outdoors in cold climates) by adding AEA into the mixtures and with pea gravel or an LWA of a similar size. The intended slump was 5-6 inches and the air content was 6-7%. Given this information and the selected guideline ACI 211.2 (2004), the amount of water required for each cubic meter of the mixture was calculated to be 210.6 Kg. Concrete was mixed in a laboratory mixer and the whole process of making, pouring and curing concrete was performed based on the ASTM C 31/C 31M – 06 guideline. Three 4-inch-by-8-inch cylinders from each batch of the concrete mixture were tested for CS in each of four curing ages of 3, 7, 28 and 90 days. The same was performed for tensile strength.

4.2.6 Modeling Methodology and Setting

Figure 4.1 shows the basic structure of the ANN models. They consist of an input layer, one or more hidden layers, and an output layer. The symbol “7-3-1” represents 7, 3, and 1 neuron(s) in the input, hidden, and output layers, respectively.

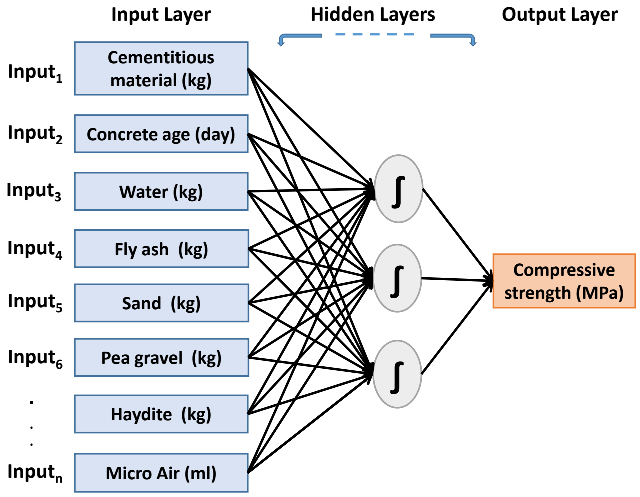


Figure 4. 1. The basic structure of the created ANN models (7-3-1).

In this study, the Weka GUI-based workbench toolbox was used to generate the required MLP model. A feed-forward back propagation learning algorithm was selected for the optimization of the networks. A unipolar sigmoid function was selected as the activation function. It is a non-linear logistic function, which gives the network flexibility in modeling more complicated relationships. The learning rate of 0.3 and the momentum value of 0.2 were selected for the purpose of this study. The training process was set for 500 epochs and the validation threshold was defined as 20 times.

In the literature, researchers either selected the relative or numerical method to input the variables for their ANN models in studying the CS of concrete (Saridemir et al. 2009; Alshihri 2009; Abdeen 2010; Atici 2011; Khan 2012). This study examined both methods on the generated ANN models to assess which form of inputs would lead to better results. Specifically, the relative method used W/C ratios, SRs of FA Class F, SRs of Haydite LWA, S/C ratios, S/CA ratios, amount of AEA (ml per Kg of cement), and the curing age of concrete as inputs. The numerical method used the curing age in days; weight (Kg) of water, PC or PLC, FA, sand, pea gravel, and Haydite LWA; and the volume of micro air (ml). This MLP has the CS of concrete (MPa) as the only output.

4.2.7 Performance Measures

The models were trained with different parameters and/or variables. Their prediction accuracy was evaluated and compared based on four frequently used performance measurements in previous studies: R, R^2, RMSE, and MAE. R, RMSE, and MAE are formulated as:

$$\mathrm{R} = \frac{\sum_{i=1}^{n}(P_i - \mu_P)(A_i - \mu_A)}{\sqrt{\sum_{i=1}^{n}(P_i - \mu_P)^2 \sum_{i=1}^{n}(A_i - \mu_A)^2}} \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - A_i)^2}{n}} \tag{2}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|P_i - A_i|}{n} \tag{3}$$

Where A_i and P_i represent the actual and predicted compressive strength of concrete samples related to data point *i*, respectively, *n* is the total number of data points in the validation set(s), μ_A is the mean value of observations, and μ_P is the mean value of predictions.

A 10-fold cross-validation was used in this study to minimize the bias associated with the random sampling of the training and holdout data samples in regular validation methods. The cross-validation is a technique that evaluates the expected accuracy and validity of a predictive model by dividing a dataset into different subsets and evaluating the accuracy of the model for each of those subsets. In general, a k-fold cross-validation includes the following steps:

1) Splitting the dataset into K subsets of equal size (K folds)
2) In each run, training the model on all the subsets except one
3) Evaluating the prediction accuracy by using the left out subset to test the trained model
4) Repeating steps 2 and 3 for K times and each time leaving a different fold for testing
5) Calculating the final performance measurements by averaging the performance measurements from each of the K runs.

This would improve the generalization and reliability of the performance measurements obtained for models under testing.

4.2.8 Analytical Results and Discussion

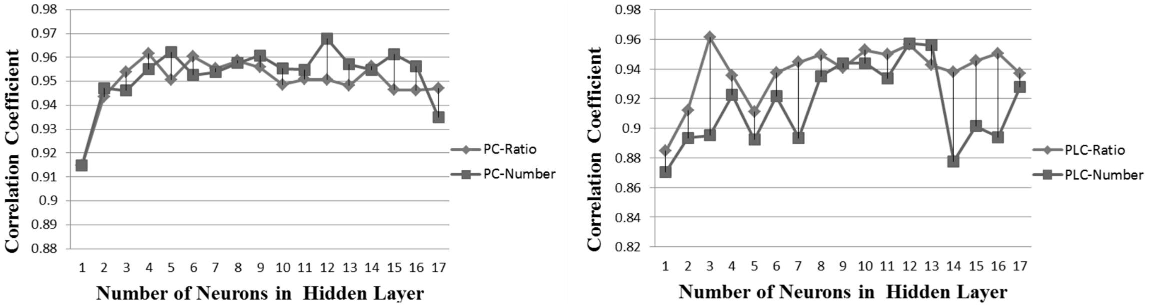
Before starting to model the problem, a simple statistical paired T-Test was performed on the available datasets to determine whether the use of PLC instead of PC has any impact on the CS of concrete. Table 4.2 shows the result of this paired T-Test, which suggests a significant difference between the average CS of PC and PLC concrete. It shows that with

95% confidence, the average CS of concrete samples made with PLC is 2.76 to 4.36 MPa higher than the average CS of concrete samples made with PC. Because of this difference, the ANN model was trained separately for the dataset of concrete made with PC and PLC.

Table 4. 2. T-Test: Paired two sample for means

Statistical item	CS (MPa) for PLC concrete	CS (MPa) for PC concrete
Mean	37.10912557	33.54779098
Variance	227.6586244	195.3950672
Observations	72	72
Hypothesized mean difference	0	
t Stat	8.857249121	
P(T<=t) one-tail	2.16E-13	
t Critical one-tail	1.666599658	
P(T<=t) two-tail	4.31E-13	
t Critical two-tail	1.993943368	

The network was trained several times with different numbers of hidden layers and different numbers of neurons in the hidden layers. The experimental output was compared with the predicted results, by the means of the three previously mentioned performance measurements. Figure 4.2 illustrates the correlation coefficients of the different ANN models trained for PC or PLC concrete datasets. In Figures 4.2a and 4.2b, results associated with the PC-ratio and PLC-ratio were generated based on the relative input method, while PC-Number and PLC-Number represent the results related to the numerical input method. It can be observed in Figure 4.2a that for PC concrete the numerical input method performs slightly better than the relative method. A maximum R was achieved based on a network with 12 neurons at the hidden layer. On the other hand, Figure 4.2b shows that for PLC concrete the relative input method gives better correlation and the optimum number of neurons in the hidden layer was 3.

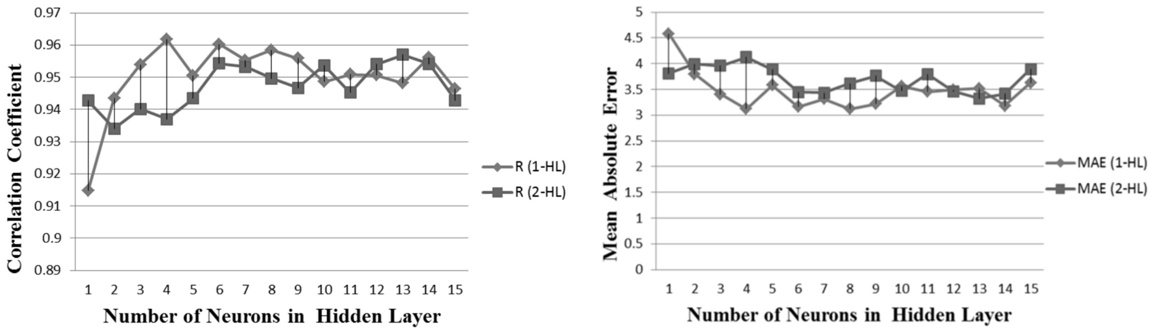


(a) PC-ratio vs. PC-number

(b) PLC-ratio vs. PLC-number

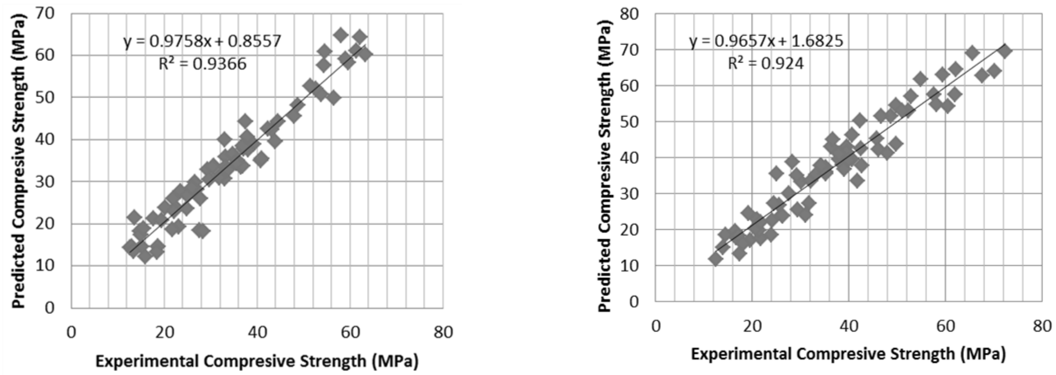
Figure 4. 2. The correlation coefficient (R) trends for the numerical and relative input methods.

Figure 4.3 shows how R and MAE change between ANN models with one hidden layer (1-HL) and 2 hidden layers (2-HL). The results are for PC concrete and related to the relative input method. Better predictions (less error and higher correlation between the predicted and actual results) were found for the 1-HL MLP models. It is also observed that the optimum result was reached by using 4 neurons at the hidden layer. Figure 4.4 shows the correlation between the experimental and predicted results for the numerical and relative input methods. The acceptable R-squared of the trend line indicates that the predicted and actual data are fairly close to each other.



(a) Correlation coefficient (R) (b) Mean Absolute Error (MAE)

Figure 4. 3. Prediction performance vs. number of neurons in 1- and 2-HL ANN



(a) For PC-number (b) For PLC-ratio

Figure 4. 4. Correlation between predicted and actual results for the two input methods.

Table 4.3 below presents a summary of the performance measurements achieved by the suggested ANN models for PC and PLC concrete. Inputs and structures for the networks are also presented in this table. All results are based on cross-validation analysis. The validation set was also used in line with the training and test sets to decrease the probability of the over-fitting problem.

Table 4. 3. Performance results for the generated ANN models

Database	Input variable	No. of optimum neurons in the HL	R	RMSE	MAE
PC-Ratio	Age - FA - HLWA - S/C - S/CA - AEA - W/C	(7-4-1)	0.9617	3.8167	3.1194
PC-Number	Age - PC - FA - HLWA - PG - S - AEA - W	(8-12-1)	0.9678	3.5418	2.7553
PLC-Ratio	Age - FA - HLWA - S/C - S/CA - AEA - W/C	(7-3-1)	0.9612	4.202	3.4335
PLC-Number	Age - PC - FA - HLWA - PG - S - AEA - W	(8-12-1)	0.9573	3.6474	4.5515

To avoid the multicollinearity problem, the seven input attributes were analyzed. The Weka Attribute Selector result suggested that for the relative input method, a subset of concrete curing age, FA, LWA, S/C and S/CA can give the best merit for this problem. The correlation analysis between the seven input variables revealed that, except for concrete curing age, FA, and LWA, other attributes of the model have some level of correlation with each other. An F-Test also confirmed this correlation. On the other hand, a Stepwise Analysis suggested that eliminating correlated attributes would not affect the performance of the models, especially since a 10-fold cross-validation test had been used to estimate the networks' performance.

In response to these correlations, and the need to eliminate the multicollinearity problem, six new different network structures were defined in this study. Table 4.4 shows the structure and the results of performance analysis for each of the six ANN models. The results suggested that there is not a significant performance loss due to the elimination of the input attributes. It is worth noting that these analyses were only based on the results of the experiments performed in this study to determine the effects of the selected alternative materials on the CS of “green” concrete. As a result, eliminating any of the other variables could reduce the generalizability of the model and is, therefore, not recommended.

Table 4. 4. Performance results for the six new neural networks

Database	Input Variable	No. of optimum neurons in the HL	R	RMSE	MAE
PC-Ratio	Age - FA - HLWA - S/C	(4-4-1)	0.9614	3.8721	3.1169
PC-Number	Age - PC - FA - PG	(4-9-1)	0.9641	3.767	2.9479
PC-Number	Age - PC - FA - HLWA	(4-9-1)	0.9646	3.7397	2.9841
PLC-Ratio	Age - FA - HLWA - S/C	(4-8-1)	0.9554	4.5052	3.5321
PLC-Number	Age - PLC - FA - PG	(4-8-1)	0.9542	4.6015	3.7758
PLC-Number	Age - PLC - FA - HLWA	(4-9-1)	0.9514	4.7021	3.7406

4.2.9 Conclusions

This chapter evaluated the application of ANN to predict the CS of “green” concrete made with FA, PLC, and Haydite LWA. The generated MLP models were tested separately for PC and PLC concrete to improve their accuracy. Moreover, the different input methods (numerical and relative) were investigated for the created ANN models. The results showed that MLP is a useful tool for predicting the CS of the studied types of concrete. It is efficient enough for both input methods although the numerical method has a small advantage for PC concrete and the relative method is slightly better for PLC concrete. It was also concluded that 1-HL MLP models provide better accuracy for the prediction of the CS compared to the 2-HL MLP models, although the number of optimum neurons in the hidden layer could vary depending on the type and number of the inputs. This chapter also analyzed the significance of the input variables and the correlation between them. Results showed that an MLP with four independent input variables and the proper number of neurons in the hidden layer could eliminate the multicollinearity problem between variables and still be accurate enough to predict the CS of concrete, even though it is not recommended

The scope of this research was limited to predicting the CS of “green” concrete studied, but it could be expanded to the other properties of “green” concrete such as tensile strength, durability or concrete slump. Moreover, there are several other ANN models that can be evaluated, which could also be a topic for future research.

4.3 Case Study Two: Predicting Soil Erosion in Xinnan Highway Slopes Using ANN

4.3.1 Introduction

This case study aimed to develop a data-driven predictive model for prediction of soil erosion in highway slopes based on the type of vegetation communities, rainfall events, and soil characteristics. Artificial Neural Network (ANN), which is one of the most popular predictive models, was applied for this prediction. The dataset used in this research was retrieved from Cao et al., 2016. ANN model was examined first with 13 input variables and then with a reduced number of variables to eliminate the effect of multicollinearity. Extensive hand tuning was performed to improve the prediction models. The result of this research confirmed that artificial neural network has an acceptable accuracy for prediction of rainfall soil erosion in highway slopes.

4.3.2 Background

Cao et al. (2016) investigated the effect of various vegetation communities on the protection of highway slopes against rainfall soil erosion. According to Cao et al. (2016), four sections of Xinnan highway at Henan province in China were selected for their research. Section A and D were fill slopes while section B and C were cut slopes. The amounts of runoff and soil erosion in these sections for various rainfall events were recorded. Each section had eight different vegetation communities consist of various combinations of eight different plants. The plants were all native and identified according to a survey to be more suitable for the highway slopes. These plants are including; 1) *Amorpha fruticosa*, 2) *Puracantha fortuneana*, 3) *Festuca elata*, 4) *Medicago sativa L.*, 5) *Vitex negundo* var.

heterophyllla, 6) *Euonymus fortunei*, 7) *Trifoliumrepens Linn*, and 8) *Cynodon dactylon* × *Cynodon transvadlensis*. Table 4.5 shows the eight vegetation communities investigated in Cao et al. (2016), including one without any plantation coverage.

Table 4. 5. Vegetation communities

Vegetation community no.	**Plants**	**Vegetation community no.**	**Plants**
1	*C. dactylon ×C. transvadlensis +F.elata+ A.fruticosa*	5	*C. dactylon ×C. transvadlensis+ F.elata+ E. fortune*
2	*C. dactylon ×C. transvadlensis+ F.elata+ M. sativa L.*	6	*C. dactylon ×C. transvadlensis+ F.elata+ T. repens*
3	*C. dactylon ×C. transvadlensis+ F.elata+ V.negundo*	7	*C. dactylon ×C. transvadlensis+ F.elata*
4	*C. dactylon ×C. transvadlensis+ F.elata+ P. fortuneana*	8	*No plantation coverage*

The study was performed within a period from May 15th to Sep 30th of 2005. Over this period 15 different rainfall events were recorded in section A and 16 different rainfall events were recorded in other three sections.

4.3.3 Introduction to the Dataset

The data set applied for the modeling process contains 442 samples based on different test sections, rainfall events, and vegetation communities. The data that were collected includes sections, plant community, coverage, rainfall duration (h), rainfall amount (mm), rainfall intensity (mm/h), growth rate of herbs in each month (cm/d), growth rate of bushes in each month (cm/d), slope gradient, above ground biomass (kg/m^2), Hydraulic conductivity

(K) (mm/min) for three different soil depth (including 0-10 cm, 10-20 cm, and 20-30 cm), and average soil erosion of each rainfall.

To preserve the generality of the developed model for future use, experimental sections (e.g. Sections A, B, C, and D) were not considered a variable in the modeling, even though the analysis showed the improvement of prediction accuracy when test section was added as an additional variable. In this study, 10-fold cross-validation was used to evaluate the performance of the models, which will also improve the validity and generalization of the model.

4.3.4 Previous Research

Prediction of soil erosion is not a new topic. There are several types of erosion models created and used by researchers for this purpose. The most common models are empirical and semi-empirical models such as Universal Soil Loss Equation (USLE) and its revisions developed by the U.S. Department of Agriculture (Wischmeier and Smith 1978; USDA 2016). The Water Erosion Prediction Project (WEPP) is a physically based erosion simulation model developed to replace USLE (Flanagan 2007). Some researchers used the Kinematic-Wave Modeling to model the watershed erosion and sediment yield (Lopes 1987; Woolhiseret al. 1990).

ANN has also been previously used by researchers in predicting soil erosion. Harris and Boardman (1998) applied expert systems and neural networks to generate an alternative soil erosion model. Licznar and Nearing (2003) compared the prediction results between an ANN model and a WEPP model and found that the ANN model performed generally better.

They used an ANN with 10 input variables including intensity and duration of precipitation, canopy cover, interrill cover, effective hydraulic conductivity, adjusted interrill soil erodibility K_i, adjusted baseline rill erodibility, the number of days since last disturbance, slope steepness, and slope length. They found that the achieved correlation coefficients of predictions ranged from 0.7 to 0.9, and the type of transfer function and the number of neurons did not affect the accuracy of the results. Xinyu et al. (2015) offered a fuzzy neural network model for the prediction of soil erosion with input variables as precipitation, runoff depth and flood rate of a small watershed, and output variable as sediment effluent. However, their samples were limited to 9 measurements. A few other researchers either used ANN to develop more accurate prediction models or evaluate ANN's accuracy in predicting sediment (Abdollahzadeh 2011; Mount and Abrahart 2011). Although previous research suggested that ANN might be a viable technique for predicting soil erosion, more research is needed to investigate the reliability of such prediction when dealing with different data sets and input variables.

Based on the literature review, the set of input variables used in this research for the prediction of soil erosion is unique and has not been applied in any previous research. This research also provides a better understanding of the model tuning process and its value on the accuracy of predictions.

4.3.5 Modeling Methodology and Setting

Correlation analysis was performed in SPSS to reveal the relationship between soil erosion and influencing factors considered in this research. Multilayer perceptron, a feedforward Artificial Neural Network (ANN) developed by Rosenblatt (1958), was selected to model the complex relationships between soil erosion and variables considered in this research. This study used the WEKA workbench toolbox (Univ. of Waikato 2015) to generate multilayer perceptron models. Figure 4.5 presents the basic structure of ANN models created for this research. The term (13-7-2-1) shows that this model has 13 input variables, 7 and 2 nodes in its first and second hidden layers, respectively, and one output variable. The input variables selected for the modeling include two nominal variables (i.e., plant community and month) and 11 numeric variables, namely coverage (%), rainfall duration (h), amount (mm) and intensity (mm/h), growth rates of herbs and bushes in each month (cm/d), slope gradient, AGB (kg/m^2), K (0-10cm) (cm/min), K (10-20cm), and K (20-30cm).

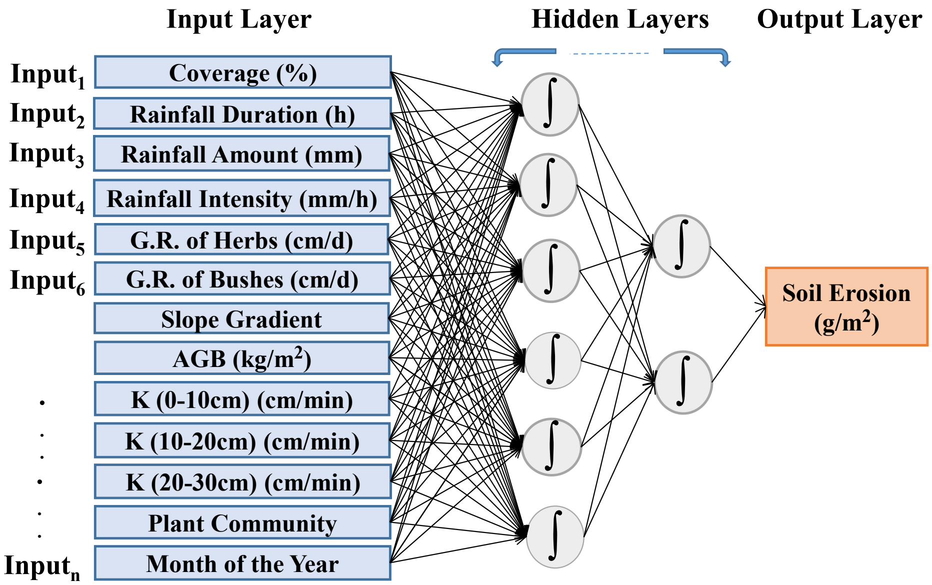


Figure 4. 5. The basic structure of the created ANN models (13-7-2-1)

The parameter setting of an ANN model could affect its prediction accuracy. Extensive hand tuning was performed in this study to identify the parameter settings that led to the highest prediction accuracy based on each of the three performance measurements used while avoiding overfitting issues. The parameters examined in the tuning process include the number of hidden layers, the number of nodes in each hidden layer, learning rate, and momentum. The prediction accuracy was assessed based on the following three performance measurements: correlation coefficient (R), root mean squared error (RMSE) and mean absolute error (MAE).

To avoid the multicollinearity problem, two different factor selection methods (stepwise regression in JMP and Weka attribute selection) were applied to explore the possibility of reducing the number of input variables while preserving the accuracy of the model. Accordingly, the ANN models with the reduced number of input variables were generated and tested for comparison.

4.3.6 Results and Discussion

4.3.6.1 Correlation Analysis

The results of correlation analysis are displayed in Table 4. In terms of independent variables, for rainfall events recorded in this study, both rainfall duration and intensity had significant positive relationships with rainfall amount. A significant negative relationship existed between rainfall duration and intensity. For the vegetation community, significant positive relationships existed among coverage, AGB, and growth rates of herbs and bushes.

Table 4. 6. Correlation among soil erosion, rainfall events and characters of vegetation community

	1	2	3	4	5	6	7	8	9	10	11	12
1. Rainfall duration	1											
2. Rainfall amount	.21**	1										
3. Rainfall intensity	-.43**	.68**	1									
4. Coverage	-.11*	0.36	-0.17	1								
5. growth rate of herbs	-0.07	-.10*	-0.02	.48**	1							
6. growth rate of bushes	-0.05	-.09*	-0.04	.35**	.25**	1						
7. AGB	-.13**	-.33**	-.14**	.79**	.47**	.49**	1					
8. Slope gradient	-0.08	-.29**	-.27**	0.04	0.00	0.05	0.09	1				
9. K (0-10 cm)	0.04	0.01	0.00	-.15**	-0.09	0.00	-.11*	-.18**	1			
10. K (10-20 cm)	-0.09	-.30**	-.27**	0.04	-0.01	0.05	0.09	1.00**	-.22**	1		
11. K (20-30 cm)	-.10*	-.31**	-.27**	0.06	-0.01	0.04	.10*	.98**	-.29**	1.00**	1	
12. Soil erosion of each rainfall	-0.02	.69**	.60**	-.49**	-.29**	-.13**	-.37**	-0.07	.25**	-0.09	-.11*	1

**. Correlation is significant at the 0.01 level.

*. Correlation is significant at the 0.05 level.

The soil erosion of each rainfall was found to have significant positive relationships with rainfall amount and intensity. The result is similar to the findings in previous research, e.g., Zhou et al. (2016). Significant negative relationships existed between soil erosion and the four characteristics of vegetation community. The negative correlation between soil erosion/sediment/loss rate and coverage was found in Bochet and García-Fayos (2004), Zhou et al. (2006); Martínez-Zavala et al. (2008), and Liu et al. (2015), but not in Zhou et al. (2016). Pimentel and Krummel (1987) also revealed a similar negative correlation between soil erosion and AGB. This research confirms that K at both 0-10 cm and 20-30 cm soil layers played an important role in controlling soil erosion (significant at 0.01 and 0.05 level, respectively).

4.3.6.2 Prediction Results

Table 5 presents the structures and prediction performance of ANN models (with 13 input variables) identified in this study to have achieved the highest prediction accuracy in the tuning process (referred to as the best models). It appears that the structure of the best model associated with each individual performance measurement varied. In general, these best ANN models all had good prediction performance.

Table 4. 7. Structure and prediction performance of the best model associated with each individual performance measurement (13 input variables)

The best model for each measurement	Model structure	Learning rate/momentum	Correlation coefficient	Mean absolute error	Root mean squared error
Model with the highest R	(13-8-4-1)	0.2/0.1	0.9776*	2.8567	4.5606
Model with the lowest MAE	(13-7-2-1)	0.2/0.1	0.9716	2.7676*	4.8317
Model with the lowest RMSE	(13-10-4-1)	0.2/0.1	0.9713	2.9035	4.5549*

The highest prediction accuracy achieved for each performance measurement

Figure 4.6 illustrates the residuals of the predictions for the best model according to R, against actual values of soil erosion. It can be seen that when the actual soil erosion amounts are small the residuals are also small.

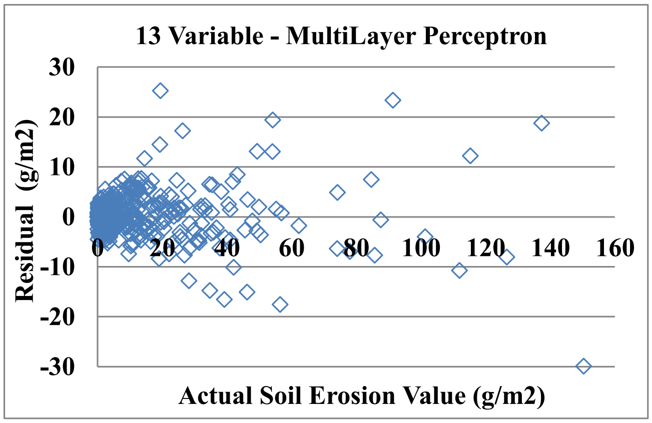


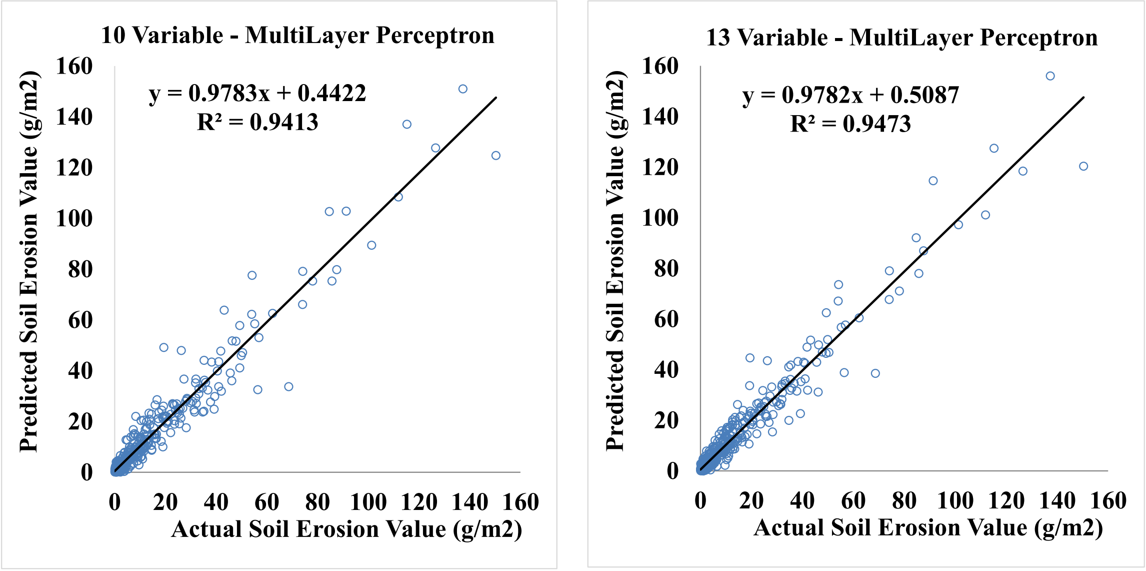
Figure 4. 6. Residuals versus actual soil erosion value

This study found that three of the 13 input variables, including rainfall duration, slope gradient, and K (10-20 cm) had insignificant effects on the accuracy of the model and could be removed. The attribute selector in Weka also suggested K (20-30 cm) as another variable that could be eliminated. However, to avoid losing any valuable information, this research included this variable in the adjusted model while eliminating the other three. The same hand-tuning process was performed for these adjusted models. The structure and prediction accuracy for the best-adjusted model are displayed in Table 6, which shows that the same model achieved the highest accuracy based on all three criteria.

Table 4. 8. Structure and prediction accuracy of the best- adjusted models based on each performance criteria (10 input variables)

Best model for each criterion	Model structure	Learning rate/momentum	R	MAE	RMAE
Model with the highest Prediction accuracy based on all 3 criteria	10-17-1-1	0.2/0.1	0.9727	2.8198	4.8198

Figure 4.7 displays the predicted versus actual values of soil erosion associated with different rainfall events based on the best models (according to R) identified for 13 (Figure 4.7b) and 10 input variables (Figure 4.7a). Both models achieved acceptable R^2 values although the model with 13 input variables has a slightly higher R^2 (0.9473).



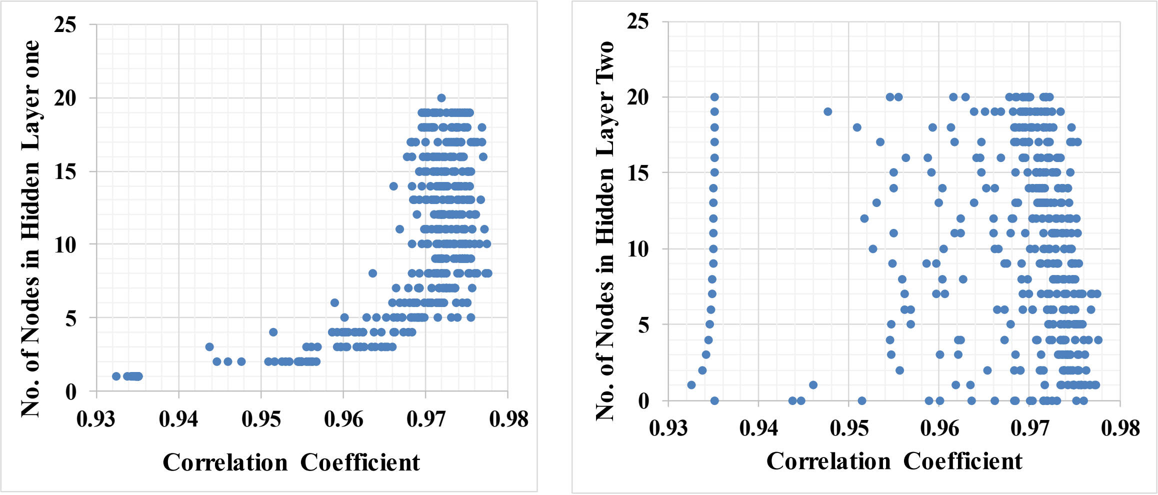
(a) (b)

Figure 4. 7. Prediction vs actual soil erosion values for the best models with a) 10 input variables and b) 13 input variables.

4.3.6.3 Observations on the Hand-Tuning Process

ANN is a computational system consisting of simple, highly interconnected processing elements (nodes or neurons) that work together to solve specific problems (Caudill 1987). ANN models usually consist of an input layer, one or more hidden layers (dependent on the type and complexity of the problem to be solved), and an output layer. The layers can have different numbers of nodes and there is no general rule for choosing the number of nodes in the hidden layer. Some studies attempted to relate the number of hidden layer neurons to the number of variables in the input and output layers or to define an upper bound for it. However, these rules cannot guarantee the generalizability of the networks (Omran et al. 2016). As a result, the structure and parameters (e.g., number of nodes, momentum, learning rate, etc.) of an ANN model need to be fine-tuned to achieve better prediction accuracy.

Certain relationships were observed between the number of nodes in the hidden layers and the prediction accuracy. Figures 4.8 show R values achieved by 400 different model settings with varying number of nodes in either the first or the second hidden layer based on ANN models with 13 input variables. For example, in Figure 4.8a, the data points aligned with each other horizontally denote the model settings with the same number of nodes in the first hidden layer ranging from 1-20 while varying the number of nodes in the second hidden layer from 0-20.



(a)

(b)

Figure 4. 8. Correlation coefficient vs. number of nodes in a) first and b) second hidden layers

Figure 4.8a shows that correlation coefficient of the predictions generally improves as the number of nodes in the first hidden layer increases (up to 8 nodes). Beyond 8 nodes, correlation coefficient remains in the similar range. It is notable in Figure 4.8b that the correlation coefficient values on the same row are more spread out compared with Figure 4.8a, which means the prediction accuracy of these models is affected more by the number of nodes in the first hidden layer than that in the second hidden layer. It seems that the accuracy of prediction declines slightly with the increase in the number of nodes in the second hidden layer.

Most of the ANN models with higher prediction accuracy found in the tuning process were models with 2 hidden layers despite the number of input variables used. In addition, for models (based on 10 input variables) with the lowest RMSE values, the top 5 of them all have only one node in their second hidden layer. The majority of models with top performance (e.g., 18 out of the top 20 in R) have a lower number of nodes (no more than 5) in the second hidden layer.

4.3.7 Conclusion for Case Study Two

The highest correlation coefficient achieved in this case study (i.e., using ANN models for soil erosion prediction) is 0.9776, resulting from the model with thirteen input variables, 10 nodes in the first hidden layer and two in the second hidden layer. The highest correlation coefficient observed for models with 10 input variables is 0.9727, representing a very small performance loss. With 13 input variables, the best models identified using different performance criteria are different in the model parameter setting. With 10 input variables, the same model was identified as the best model according to all the three performance criteria.

4.4 Conclusion and Summary

This chapter presents and evaluates the performance of ANN as a predictive model to be applied in two different research areas related to the construction industry, including the properties of environmentally friendly concrete and vegetation community and soil erosion for highway slopes. The findings show that the basic ANN models can be successfully used for the prediction in both subject areas with acceptable accuracy.

Chapter 5. Comparison of Data Mining Techniques for Predicting Compressive Strength of Environmentally Friendly Concrete

This chapter aims to provide a more in-depth analysis for prediction performance of the artificial neural network compared to some of the other common data mining techniques. To do so, the author investigated and compared the performance of nine data mining models in predicting the compressive strength of the previously introduced environmentally friendly concrete containing three alternative materials as fly ash, Haydite® lightweight aggregate, and Portland limestone cement. These models include three advanced predictive models (multilayer perceptron, support vector machines, and Gaussian processes regression), four regression tree models (M5P, REPTree, M5-Rules, and decision stump), and two ensemble methods (additive regression and bagging) with each of the seven individual models used as the base classifier. The results of this chapter offer valuable insights on improving the use of these models for property prediction of concrete.

5.1 Introduction

Using alternative materials in concrete may positively or negatively impact its properties (Khalaf and Devenny 2004; Yang et al. 2005; Berry et al. 2011). Research is thus needed to thoroughly understand the potential influence of these materials. Since the compressive strength is one of the most important concrete properties, many experiments have been conducted to study the compressive strength of environmentally friendly concrete (Yang et al. 2005; Etxeberria et al. 2007; Kevern et al. 2011). Since statistical modeling has its limitations in estimating the underlying relationships between the inputs and outputs of forecasting models in more complicated cases (Zhang 1998), recent studies have shown an increasing trend toward the application of machine learning techniques in predicting concrete compressive strength (Topçu and Saridemir 2007; Saridemir et al. 2009; Atici 2011; Aiyer et al. 2014; Akande et al. 2014; Omran et al. 2014). The results from these studies demonstrate a great potential of this approach, which warrants further investigation.

The research presented in this chapter compared the use of seven individual machine learning models, including M5Prime (M5P), REPTree, M5-Rules, decision stump, multilayer perceptron, SMO regression (SMOreg), and Gaussian processes regression, in predicting the compressive strength of environmentally friendly concrete. It also tested two commonly used ensemble methods (additive regression and bagging) by adopting each of the seven individual models as the base classifier to explore the possibility of improving prediction accuracy. The ultimate goal was to promote the use of data mining techniques for determining the compressive strength or other properties of new types of concrete while reducing the need for extensive experiments. This shift will not only save time and money for the industry but also facilitate the use of new materials. The unique set of seven data mining models was selected for exploring the prediction performance of four regression tree models against other three more advanced models. This also seemed to be the first time that Gaussian processes regression was examined for predicting concrete strength. This research used four performance measures, namely correlation coefficient (R), the coefficient of determination (R^2), root means squared error (RMSE), and mean absolute error (MAE), to assess prediction accuracy of generated models. R^2 was used to compare models examined in this research and previous studies.

This research first introduces the unique type of environmentally friendly concrete studied in this research and then reviews previous research efforts in modeling and predicting compressive strength of concrete. A brief description of all the data mining models examined in this research is presented. After describing the research methodology and experimental settings, this chapter presents the results and analysis as well as the findings of this research.

5.2 Related Work in Modeling and Predicting Concrete Properties

The experimental determination of the compressive strength of concrete, especially for concrete containing alternative materials, is known to be time-consuming and costly. On the other hand, using simple linear regression models for prediction has limited accuracy and flexibility (Yeh 1998; Deepa et al. 2010). As a result, recent years have seen an increasing interest in using more advanced data mining techniques for predicting concrete properties.

Artificial neural network (ANN) has been used to predict fresh and hardened properties of high-performance concrete (Khan et al. 2013) and LWA concrete (Alshihri et al. 2009;

Abdeen and Hodhod 2010). The results of these studies have generally confirmed ANN to be a powerful method for such applications. Another widely used data mining method, Support Vector Machines (SVM), has also been used to predict properties of hardened concrete, such as compressive strength, tensile strength, and elastic modulus (Gupta 2007; Yan et al. 2013; Yazdi et al. 2013; Aiyer et al. 2014; Akande et al. 2014). In other attempts, both ANN and SVM were applied in conjunction with fuzzy logic to improve the accuracy and reliability of prediction (Nataraja et al. 2006; Saridemir et al. 2009; Cheng et al. 2012). Some other predictive models, e.g., ensembles of decision trees in Erdal et al. (2013), were also examined for predicting the compressive strength of different types of concrete. While these studies have led to more accurate predictions compared to traditional regression techniques, more reliable, applicable, and practical models are yet to be discovered (Chou et al. 2011).

A comparison between multivariable regression analysis and ANN made by Atici (2011) identified the effectiveness of these methods for predicting the strength of mineral admixture concrete. With the increasing use of advanced data mining techniques in concrete property prediction, a few other comparative studies were conducted to evaluate the performance of multiple data mining models, mostly focused on the compressive strength prediction of high-performance concrete. For example, Deepa et al. (2010) examined ANN, linear regression, and M5P tree model for their accuracy and time performance. Similarly, Chou et al. (2011) evaluated ANN, SVM, multiple regression, multiple additive regression trees, and bagging regression trees. So far, very few studies have compared multiple data mining methods in predicting the compressive strength of environmentally friendly concrete. This study aims to fill the aforementioned gap and provide a more accurate and reliable tool to predict the compressive strength of a unique type of environmentally friendly concrete made with PLC, Haydite LWA, and FA.

5.3 Predictive Data Mining Techniques Examined in This Research

The research was performed in two steps: 1) Examining the prediction accuracy of seven individual data mining models, including the four common regression tree models (M5P, REPTree, M5-Rules, and decision stump) and three more advanced predictive models (multilayer perceptron, SMOreg, and Gaussian processes regression), and 2) Examining the prediction accuracy of two commonly used ensemble methods (additive regression and bagging), in which each of the aforementioned models was used as the base classifier to evaluate the effects of boosting. Kotsiantis et al. (2006) defined three mechanisms for the ensemble of regression models: 1) *using a single machine learning model with different subsets of training data*, 2) *using a single learning method with different training parameters*, and 3) *using different machine learning methods*. The second step of this research adopted the first two mechanisms by using a single machine learning model as base classifier for the ensemble models. Studying multiple classifiers for the ensemble models can be a subject for future research. A brief review of these data mining models and selected parameters is presented below.

5.3.1 Regression Tree Models

Regression tree models have long been used in data mining as a supervised learning technique, and have been widely applied to numeric prediction. Compared to some of the state-of-the-art models, regression tree models may have lower prediction accuracy, but usually perform faster and are easier to interpret. This research examined four commonly used regression tree models as described below.

5.3.1.1 M5P

M5P is a reconstruction of the M5 algorithm introduced by Quinlan (1992) for generating a tree of regression models from empirical data (Wang and Witten 1997). In an M5P model, at each branch, the tree stores a linear regression model that predicts the class values of the portion of the dataset that reaches the leaf. The dataset splits into different portions according to certain attributes of the data. Standard deviation (SD) is usually used as a criterion that determines which attribute is the best for splitting the dataset at each node. The attribute to be chosen is the one that has the maximum expectation to reduce error. The process stops when a very small change happens in class values or only a few instances remain. The tree will then be pruned back and a smoothing process will be performed in the end to compensate sharp discontinuities between adjacent linear models (Quinlan 1992).

5.3.1.2 REPTree (Reduced Error Pruning Tree)

REPTree (Reduced Error Pruning Tree) is a fast decision tree learner that builds a decision/regression tree by using information gain or variance as decision features for splitting the data at the nodes. Then the generated regression tree is pruned back using the reduced-error with back over-fitting technique (Witten and Frank 2005). In the context of decision trees, the term “information gain” is usually equivalent to the expectation value of the Kullback–Leibler divergence of a conditional probability distribution (Garcia et al. 2002). For numeric attributes, REPTree sorts the values once at the start of the run, and then uses the sorted list to calculate the right splits in each tree node.

5.3.1.3 M5-Rules

M5-Rules is an algorithm that uses divide-and-conquer to generate decision lists (ordered sets of the if-then rule) for regression problems. Holmes et al. (1999) used decision lists to make a more compact and understandable model tree compared to previous models. Decision lists can work with both continuous and nominal variables. M5-Rules uses the M5 algorithm to build a model tree, makes a rule from the best leaf, and then works on other instances that are left in the dataset according to the generated rule.

5.3.1.4 Decision Stump

Decision Stump is a machine learning model that only consists of a one-level decision tree. It has one internal node (called root node), which is immediately connected to nodes in branches (referred to as terminal nodes). In a decision stump, a prediction is made according to the value of a single input attribute. Regression is performed based on the mean squared error where each root node represents an attribute in an instance to be evaluated, and each branch represents a value that the node can take (Iba and Langley 1992). Decision stump is usually used as a component of a boosting algorithm to improve prediction accuracy.

5.3.2 Multilayer Perceptron (ANN)

ANN is a computational system consisting of simple, highly interconnected processing elements (nodes or neurons) that work together to solve specific problems (Caudill 1987). It is an algorithm inspired by research in biological nervous systems to generate a simplified model of how the brain works (Rumelhart et al. 1994). ANN models usually consist of an input layer, one or more hidden layers, and an output layer, each of which can have a different number of nodes. Each node under the hidden layer(s) will receive one or more inputs. The inputs will be multiplied by their weights and summed together and with the bias (threshold). The weighting and bias values will be initially chosen as random numbers and then be adjusted according to the results of the training process (Atici 2011). The output of each node will be generated based on the significance of the summation value and by the means of a predefined specific activation function (Bishop 2006).

5.3.3 SMOreg-based SVM

SVM is a supervised learning model developed by Cortes and Vapnik (1995). It has been intensively used in many data mining problems for both classification and regression purposes. In an SVM algorithm, the training set is first mapped to an n-dimensional feature space by using a kernel mapping procedure. Then a hyperplane, a subspace that is one dimension less than its surrounding space, will be identified in this feature space according to the projected dataset. The aim is to find the optimal hyperplane that separates the data points in the classes, while simultaneously maximizing the margin (i.e., the distance between the hyperplane and the closest points of the training set) for linearly separable patterns (Leskovec et al. 2014). The hyperplane $f(x,w)$ is represented by a linear function in the feature space:

$$f(x,w) = \sum_{j=1}^{m} w_j \, g_j(x) + b \tag{1}$$

Where $g_j(x)_{,j=1,\ldots,m}$ denotes a set of nonlinear transformations, and b is the "bias" term. For SVM regression purposes, Cortes and Vapnik (1995) suggested to use a so called $\mathcal{E}$, the insensitive loss function that penalizes error only if it is greater than $\mathcal{E}$ (Shevade et al. 2000). So the $|\xi|_{\mathcal{E}}$ is represented as:

$$|\xi|_{\varepsilon} = \begin{cases} 0 & if\ |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases} \tag{2}$$

Using (non-negative) slack variables ξ_i and ${\xi_i}^*$, the final optimization problem to be solved can be formulated as:

$$\text{Minimize} \quad \frac{1}{2}\left||w|\right|^2 + C\sum_{i=1}^{l}(\xi_i + {\xi_i}^*) \tag{3}$$

Subjected to:

$$\begin{cases} y_i - f(x_i, w) \leq \varepsilon - {\xi_i}^* \\ f(x_i, w) - y_i \leq \varepsilon - \xi_i \\ \xi_i, {\xi_i}^* \geq 0, i = 1, \dots, n \end{cases} \tag{4}$$

SVM regression finds the linear regression in the high-dimension feature space using ε while reducing the model complexity by minimizing $\left||w|\right|^2$.

Sequential minimal optimization (SMO), an algorithm introduced by Platt (1998), is used to solve the very large quadratic programming (QP) optimization problems in SVM through breaking them into a series of smallest possible QP problems. In this way problems can be solved analytically, eliminating the need for numerical optimization algorithms (Platt 1998).

5.3.4 Gaussian Processes

Gaussian process is a powerful non-linear prediction tool, which can be used for Bayesian regression as well as in the learning process of both supervised and unsupervised learning frameworks (Bishop 2006). It is a non-parametric stochastic process that generalizes the Gaussian probability distribution. A Gaussian process sometimes is described as a distribution over functions ($P(f)$), where f is a function that projects input space (vector **x**) to feature space (vector **r**) and for any finite subset of X the marginal distribution over that subset $P(f)$ has a Gaussian distribution. The f could be an infinite-dimensional quantity. As a result, Gaussian process extends multivariate Gaussian distributions to infinite dimensionality (Rasmussen and Williams 2006). One of the advantages of a Gaussian process model is that its formulation is probabilistic. This is especially useful for probabilistic prediction and also enables the model parameters inference for kernel shape and noise level control (Chu and Ghahramani 2006).

5.3.5 Ensemble Methods Used in This Research

According to Rokach (2010), the idea of ensemble learning models started with Tukey (1977) at late 1970s by simply combining two linear regression models using residual of the first model for the second modeling process. This effort was then followed by many other attempts, such as partitioning the input space and using two or more classifiers (Dasarathy and Sheela 1979) or using the AdaBoost algorithm (Freund and Schapire 1996). The purpose for ensemble modeling is to achieve better prediction performance by combining multiple learning algorithms.

5.3.5.1 Additive Regression (Gradient Boosting)

Regression trees are well known for many advantages such as flexibility of input variables (e.g., numeric, ordinal, binary, and categorical variables) and immunity to the effects of extreme outliers. However, these methods usually suffer from the lack of accuracy. Gradient boosting, first introduced by Friedman (2001), is an additive regression tree model that can overcome this drawback through the application of a boosting technique (Friedman and Meulman 2003). According to Friedman (2001), additive regression is a metadata learner that improves the performance of weak prediction models (e.g., regression tree models) by applying the stochastic gradient boosting technique. The technique mainly involves fitting sequence of models: The first model in the sequence is trained based on the original dataset, and each of the next models is trained on a new dataset containing the residual errors remained from fitting the previous model.

5.3.5.2 Bagging

Bagging is short for Bootstrap Aggregating. Breiman (1994) defines bagging as a way to generate multiple versions of a predictor, through which a more robust predictor can be generated. It is an ensemble meta-algorithm that improves the accuracy and stability of the prediction. The algorithm is based on generating bootstrap replications of a dataset and using these different versions of the dataset as new training sets to generate multiple models. The final prediction is achieved through combining the outcomes of these models (i.e., averaging the results for the regression problem and using plurality voting for the classification problem). Previous studies have shown that bagging can significantly improve the results of unstable models (e.g., models sensitive to small changes in the training dataset), models with high dimensional dataset problems, and classification and regression tree models (Breiman 1994; Buhlmann and Yu 2002).

5.3.6 Methodology and Experimental Settings

5.3.6.1 Concrete Experimental Design and Data Collection

In this study, 36 different batches of concrete were designed and prepared. Each batch contained different replacement percentages of FA Class F (0%, 20%, 30% or 40%) and Haydite LWA (0%, 33%, 67% or 100%) besides the use of either Portland cement (PC) Type I/II or PLC Type GUL. In this way, the effects of alternative materials on the compressive strength of concrete can be examined more accurately. The FA Class F replaced part of PC or PLC by different percentages of weight, and Haydite LWA substituted pea gravel by different percentages of volume. Their numerical values were used as inputs for the tested models. In addition to the above three variables, the actual water content, the amounts of sand, pea gravel and Micro Air®, as well as the concrete curing age were selected as the other influential variables for the models. Table 5.1 shows the range, mean, and SD of those variables in this experimental study.

Table 5. 1. Parameters and values for concrete mix design (per cubic meter of concrete)

Parameter	Min.	Max.	Mean	SD
Age (day)	3	90	35.12	35.37
Water (kg)	210.61	210.61	210.61	0
PC or PLC (kg)	226.63	528.02	346.18	102.07
FA (kg)	0	211.21	79.80	72.37
Sand (kg)	741.60	901.78	768.29	59.91
Pea gravel (kg)	0	750.49	483.40	229.54
Haydite (kg)	0	368.42	131.13	113.03
Micro Air (ml)	112.17	135.38	123.78	11.64

All the concrete mixed in the experiment was assumed to be air-entrained (considered to be used outdoors in cold climate) by adding Micro Air, an air entraining agent, to the mixtures. The intended slump was 12.70 - 15.24 cm and the air content was 6-7%. Concrete was mixed in a laboratory mixer and the whole processes of making, pouring and curing concrete were performed based on ASTM C 31/C 31 M – 06 guideline. Three 10.16 cm by 20.32 cm cylinders from each batch of concrete mixture were tested in each of four different curing ages of 3, 7, 28 and 90 days for compressive strength. The average test result of each three cylinders formed a data point in the database. All the details for the experiments can be found in Jin (2013).

5.3.6.2 Parameter Setting of Data Mining Models

In this study, the Weka workbench toolbox (Waikato 2015) was used to generate the examined machine learning models for predicting compressive strength of the environmentally friendly concrete. Since one of the original goals for experimental testing was to compare the compressive strength of PC and PLC concrete, this research performed a simple paired t-test on the PC and PLC concrete datasets, which confirmed a statistical difference between these two groups. To evaluate the potential impact of the statistically different datasets on the prediction accuracy of data mining models, this research took the following three-step approach: The first was to test the selected data mining models based on the PC or PLC dataset only. In such cases, eight variables were used to generate the models. The second step was to examine the selected models based on the whole dataset including all PC and PLC concrete samples. In the modeling process, nine variables including a new binary variable “cement type” were used. Thirdly, the prediction performance of data mining models based on different datasets was compared to learn whether simpler models with eight variables and individual datasets will lead to better prediction accuracy, or the prediction accuracy can be improved by a larger sample size though additional variable(s) may be needed, leading to more complex models.

Many input parameters need to be set up for most data mining algorithms. The setting of input parameters could affect the accuracy and/or reliability of generated models. In this research, extensive hand-tuning was performed on each model to identify the parameter setting that could lead to the highest prediction accuracy among all the examined model settings while avoiding over-fitting issues. Specifically, before hand-tuning, literature and past experiments related to each of the tested models were carefully reviewed to identify influential model parameters and their commonly used values. During parameter tuning, all possible combinations of these parameter values were tested. For example, suppose there are N influential parameters for a studied data mining model, indexed by n=1,…N. For parameter n, there are k_n values. Then all $\prod_{n=1}^{N} k_n$ possible parameter combinations were tested, and the parameter setting associated with the best performance of this model was determined. In this study, all the R, R^2, RMSE, and MAE reported for the tested data mining models were associated with the best performance achieved through the parameter tuning process.

5.3.6.3 Performance Measures

The models were trained with different parameter settings. Their prediction accuracy was evaluated and compared based on four frequently used performance measurements in previous studies: R, R^2, RMSE, and MAE. R, RMSE, and MAE are formulated as:

$$\mathrm{R} = \frac{\sum_{i=1}^{n}(P_i - \mu_P)(A_i - \mu_A)}{\sqrt{\sum_{i=1}^{n}(P_i - \mu_P)^2 \sum_{i=1}^{n}(A_i - \mu_A)^2}} \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - A_i)^2}{n}} \tag{6}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|P_i - A_i|}{n} \tag{7}$$

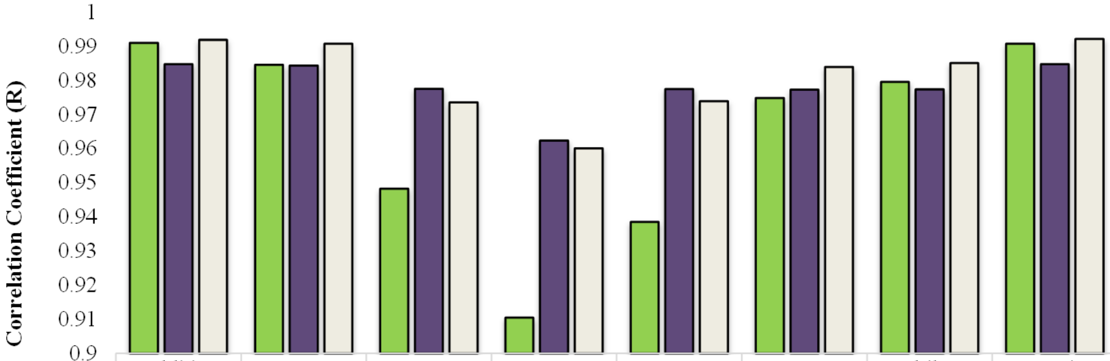
Where A_i and P_i represent the actual and predicted compressive strength of concrete samples related to data point *i*, respectively, *n* is the total number of data points in the validation set(s), μ_A is the mean value of observations, and μ_P is the mean value of predictions. For simulations performed by the Weka toolbox, R^2 is equal to the square of R.

A 10-fold cross-validation was used in this study to minimize the bias associated with the random sampling of the training and holdout data samples in regular validation methods. Analytical Results and Discussion

5.3.6.4 Comparison Results for the Data Mining Models Tested

In the following, comparison results for the data mining models tested in this research are presented. Due to its poor prediction accuracy (e.g., R values at 0.5226, 0.6001 and 0.6208 for the PLC, PC and combined datasets, respectively), the decision stump model is excluded from most of the tables and figures presented below with the exception of the results related to ensemble models. This is because this study found that when decision stump was used as the base classifier for the ensemble models the prediction accuracy was acceptable, which is consistent with the result presented in Chou et al. (2011).

Figure 5.1 shows the R values achieved by each of the eight data mining models based on the PC, PLC, and whole datasets. It was found that the prediction accuracy increased in five of the tested models when combining the two datasets (PC and PLC) and using the cement type as an additional binary input. Exceptions are the three regression tree models (i.e., M5P, REPTree and M5-Rules), in which the accuracy of prediction based on the PC concrete dataset was slightly better than the whole dataset. The bolded R-values listed for additive regression and Gaussian processes regression are the highest among all the models tested for individual datasets.



	Additive Regression	Bagging	M5P	REPTree	M5Rules	SMOreg	Multilayer Perceptron	Gaussian Processes
🟩 PLC	**0.9909**	0.9845	0.9482	0.9105	0.9385	0.9748	0.9795	0.9907
🟪 PC	**0.9847**	0.9843	0.9775	0.9623	0.9774	0.9772	0.9773	**0.9847**
□ PLC & PC	0.9918	0.9907	0.9735	0.9601	0.9738	0.9839	0.985	**0.9921**

Figure 5. 1. R-values for each of the tested models based on different datasets

The performance of prediction models in terms of RMSE and MAE is presented in Tables 5.2 and 5.3. The bolded value in each row represents the highest prediction accuracy achieved in this study when different datasets were used for testing individual and ensemble models. These results confirm that according to both criteria (MAE and RMSE), additive regression obtained the highest prediction accuracy for the comprehensive strength of PLC samples while the individual Gaussian processes regression model achieved the highest prediction accuracy for both the PC and whole datasets.

Table 5. 2. MAE calculated for the tested models based on different datasets

Dataset	Additive Regression	Bagging	M5P	REPTree	M5-Rules	SMOreg	Multilayer Perceptron	Gaussian Processes
PLC	**1.52**	2.1038	3.4854	4.9203	3.9587	2.4839	1.946	1.6343
PC	1.8992	1.9536	2.4113	3.0505	2.3633	2.36	2.1796	**1.8784**
PLC & PC	1.3976	1.5662	2.4536	3.3953	2.4793	2.072	1.9625	**1.3756**

Table 5. 3. RMSE calculated for the tested models based on different datasets

Dataset	Additive Regression	Bagging	M5P	REPTree	M5-Rules	SMOreg	Multilayer Perceptron	Gaussian Processes
PLC	**2.0309**	2.6724	4.7615	6.2041	5.2028	3.3491	3.1178	2.2236
PC	2.4223	2.4563	2.9852	3.8477	2.9705	2.9571	2.9439	**2.4154**
PLC & PC	1.8624	1.9902	3.3367	4.1663	3.3169	2.6104	2.5473	**1.837**

The information presented above shows that the listed models all had acceptable prediction performance after extensive parameter tuning. Further, the Gaussian processes regression model achieved the best prediction accuracy based on all the three performance measures while REPTree had the lowest. Table 5.4 below lists the parameter settings used for these models to achieve their highest prediction accuracy. In particular, the option of “polykernel” was selected for all of the four models that need a kernel as their covariance matrix. These include additive regression, bagging, Gaussian processes, and SMOreg. From this point forward, the analysis and results are solely presented for the whole (PC & PLC) dataset, which was proven to have improved the prediction accuracy for most models tested in this study.

Table 5. 4. Selected parameter settings for achieving the highest accuracy of the tested models

Data mining model	R	Name of parameter	Selected value
Additive regression	0.9918	Base classifier	Gaussian process
		Number (no.) of iteration	10
		Shrinkage rate	1
		Level of Gaussian noise	0.002
		Kernel of the choice	polykernel
		Exponent value	3
Bagging	0.9907	Base classifier	Gaussian process
		No. of iteration	80
		Bagging size percentage	100
		Level of Gaussian noise	0.007
		Kernel of the choice	polykernel
		Exponent value	3
M5P	0.9735	Min. no. of instances	5
M5-Rules	0.9738	Min. no. of instances	4
REPTree	0.9601	Min. total weight of instances	1
		Min. proportion of the variance	0.0001
SMOreg	0.9839	Kernel of the choice	polykernel
		Exponent value	3
Multilayer perceptron	0.9849	Node No. for first hidden layer	15
		Node No. for second hidden layer	8
		Learning rate	0.1
		Momentum	0.25
		Training time	10000
		Validation threshold	20
Gaussian processes regression	0.9921	Kernel of the choice	polykernel
		Exponent value	3
		Level of Gaussian noise	0.0005

Figure 5.2 illustrates the relationship between the predicted and actual compressive strength of the studied concrete samples for each of the eight predictive models. All the plots show fairly linear relationships between predicted and actual values. Apparently, the Gaussian processes regression model is the best representative of actual experimental data with the highest R^2 at 0.9842, closely followed by additive regression and bagging.

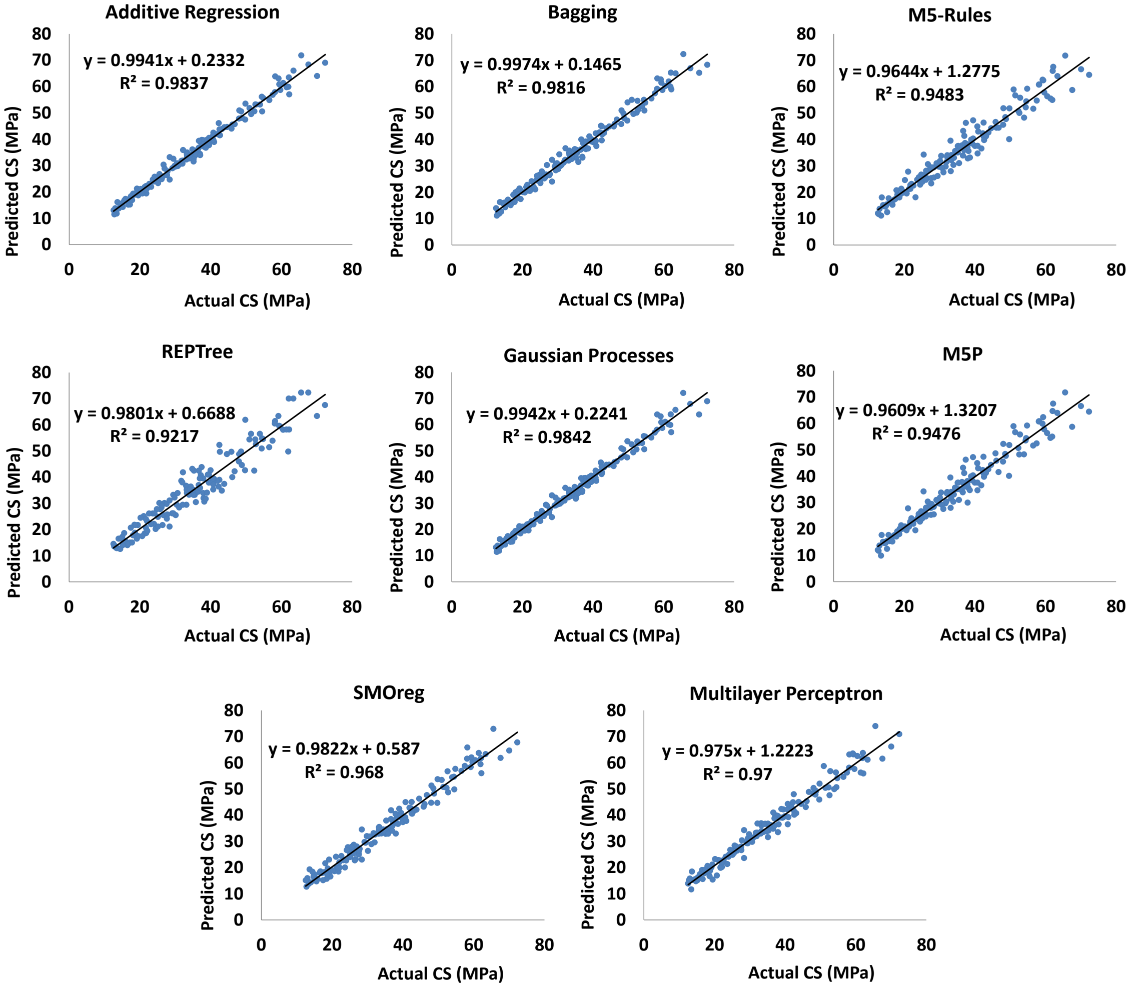


Figure 5. 2. Predicted vs. actual compressive strength (abbreviated as CS in the figure)

Figure 5.3 displays the distribution of residuals and percentage error for the tested models. It is observed that in all these plots when the actual compressive strength of concrete samples increased, residuals became larger but the associated percentage errors decreased. Similar to the early findings, Gaussian processes regression, bagging, and additive regressions are the models with prediction results being the closest to the actual experimental values.

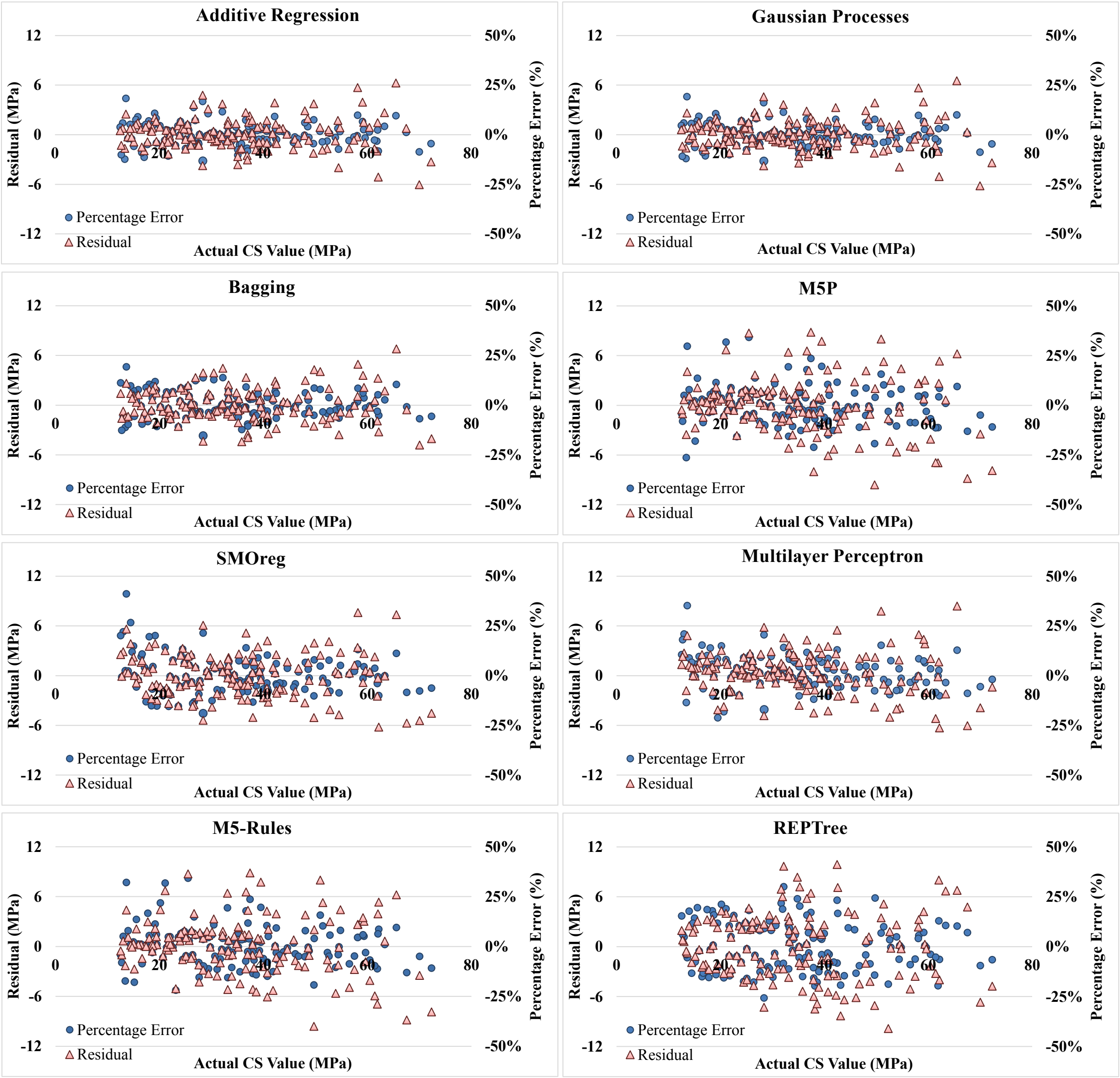


Figure 5. 3. Residuals and percentage errors vs. actual compressive strength values

Table 5.5 compares R values achieved by the seven individual data mining models as well as two ensemble methods with each of individual data mining models used as base classifier. The comparison results show that both the additive regression and bagging algorithms using regression tree models as the base classifier achieved better prediction accuracy than individual regression tree models. On the other hand, when SMOreg, Gaussian processes, and multilayer perceptron were used as the base classifier, mixed results were generated. The bolded R values highlight the strong performance of Gaussian processes regression: It achieved the highest prediction accuracy in individual model comparison; the highest accuracy of prediction for additive regression and bagging was also achieved when the Gaussian processes was used as their base classification model. This finding is particularly important since Gaussian processes regression has rarely been applied in existing research to predict concrete properties.

Table 5. 5. R for individual models and ensemble models using different classifiers

Method	**REPTree**	**M5-Rules**	**M5P**	**Decision Stump**	**SMOreg**	**Gaussian Processes**	**Multilayer Perceptron**
Individual model	0.9601	0.9738	0.9735	0.6208	0.9839	**0.9921**	0.985
Additive regression	0.9822	0.9778	0.9917	0.9712	0.9845	**0.9918**	0.9793
Bagging	0.9701	0.9765	0.9786	0.9421	0.9823	**0.9907**	0.9899

Note: The bolded R values are the highest among the compared models.

Table 5.6 lists the average time spent for building each of the tested models. These times were associated with the parameter settings for these models to achieve the highest prediction accuracy in parameter tuning. Due to the use of 10-fold cross validation, the training time for each of these models was much longer than the time used to build the model. Although many variables could affect the length of the training time, the total time was mostly proportional to the time used to build the model. The results indicate that even though the three more advanced predictive models achieved higher prediction accuracy, in general, they are far more time-consuming compared to individual regression tree models as well as ensemble models with regression tree as base classifier. The individual Gaussian processes model was somewhat an exception with relatively fast building and training time.

Table 5. 6. Time (in second) for building each data mining model

Method	REPTree	M5Rules	M5P	Decision Stump	SMOreg	Gaussian Processes	Multilayer Perceptron
Individual model	0.02	0.14	0.05	0	10.19	0.33	42.46
Additive regression	0.03	0.42	1.92	0.17	43.82	3.26	167.36
Bagging	0.28	1.09	3.71	0.03	127.02	27.89	419.42

5.3.6.5 Comparison with Previous Work

Table 5.7 compares this study with some of the primary previous work, which used data mining models to predict the compressive strength of concrete, for consistencies and differences. It can be observed that the majority of previous work was specifically focused on high-performance concrete (HPC) with the added blast-furnace slag (BFS), FA, and superplasticizer. Hand-tuning was frequently used in previous work, but mostly for ANN models. The methods for selecting input variables are very similar; i.e., using the major variables associated with concrete mix design and lab testing.

The comparison of R^2 values obtained by different studies shows that eight of the data mining models examined in this research offered fairly high prediction accuracy with R^2 ranging from 0.9217 to 0.9842. Moreover, compared with the same types of models examined in previous research, i.e., M5P, SVM, bagging, and additive regression, this study achieved relatively better prediction performance. This could be due to the extensive parameter tuning process performed for each model, the input variables, parameter values and unique datasets used, and the testing of different base classifiers for ensemble models. In general, the performance of data mining models can be improved by a thorough parameter tuning procedure. This research applied the cross-validation method for evaluating the accuracy of predictions, which was not the case in most of the previous studies listed in Table 5.7 except for Chou et al. (2011) and Deepa et al. (2010). Compared to the traditional validation method, cross-validation usually lowers the R^2 values of tested models but improves the generalization and reliability of the assessment.

Table 5. 7. Comparison of model prediction accuracy with previous studies

Previous work	Sample size	Data mining technique	R^2	Concrete type	Parameter tuning method	Input variable(s)
Yeh 1998[a]	727	ANN	0.914	HPC	Hand-tuning (i.e., trial-and-error) for ANN	Cement, FA, blast-furnace slag (BFS), water, superplasticizer, coarse and fine aggregates, and curing age
		Linear regression	0.574			
Gupta et al. 2006	864	Neural-expert system	0.5776	HPC	No tuning	Concrete mix grade, size and shape of specimen, curing technique and period, maximum temperature, relative humidity and wind velocity, and period of strength
Fazel Zarandi et al. 2008	458	Fuzzy polynomial neural networks	0.8209	HPC	Hand-tuning	Coarse and fine aggregates, superplasticizer, silica fume, water, and cement
Yeh and Lien 2009	1196	Genetic operation trees	0.8669	HPC	No tuning	Cement, FA, BFS, water, superplasticizer, coarse and fine aggregates, and curing age
		ANN	0.9338			
Chou et al. 2011	1030	ANN	0.9091	HPC	Hand-tuning	Cement, FA, BFS, water, superplasticizer, coarse and fine aggregate, and curing age
		Multiple regression	0.6112			
		SVM	0.8858			
		Multiple additive regression trees	0.9108			
		Bagging regression trees	0.8904			
Deepa et al. 2010	300	Multilayer perceptron (ANN)	0.625	HPC	Hand-tuning for ANN	Cement, BFS, FA, water, superplasticizer, coarse and fine aggregates, and curing age
		Linear regression	0.491			
		M5P model tree	0.787			

Continue

Table 5.7. Continued

Atici 2011	135	ANN	0.9801	Concrete contains BFS and FA	Hand-tuning for ANN	Cement, BFS, curing age, ultrasonic pulse velocity, rebound number, and FA
		Multiple regression	0.899			
Erdal et al. 2013	1030	ANN	0.9088	HPC	Hand-tuning	Cement, FA, BFS, water, superplasticizer, coarse and fine aggregates, and curing age
		Bagged ANN	0.9278			
		Gradient boosted ANN	0.927			
		Wavelet bagged ANN	0.9397			
		Wavelet gradient boosted ANN	0.9528			
This research[b]	144	M5P model tree	0.9476	Concrete contains FA, Haydite LWA, and PLC	Hand-tuning	Cement type, curing age, water, cementitious material, FA, sand, pea gravel, Haydite LWA, and Micro Air
		M5-Rules	0.9482			
		REPTree	0.9217			
		Multilayer perceptron (ANN)	0.97			
		SMOreg (SVM)	0.968			
		Gaussian processes regression	0.9843			
		Additive regression	0.9837			
		Bagging	0.9816			

[a] In Yeh (1998), the database was divided into four different sets. Each time one set was used for testing and the other three sets were used for training. The listed R^2 value is the average for the four testing datasets.

[b] All the R^2 values listed for this study were the square of the R values achieved in Weka based on the whole dataset. These values can also be seen in Figure. 5.2.

According to Table 5.7, ANN, in most cases, led to higher prediction accuracy than traditional modeling approaches such as linear regression or regression tree models. Also, the Gaussian processes regression model studied in this research provided the highest prediction accuracy (R^2 = 0.9837) among all the data mining models compared, while having a relatively fast modeling speed. Based on the extent of literature review performed by the authors, this research seemed to be the first work that examined Gaussian processes regression for predicting concrete properties. Its strong performance confirmed by this research suggests a great need for further investigation of this method.

In this research, the additive regression model would rank first in prediction accuracy when without the presence of Gaussian processes regression, which is consistent with the results from Chou et al. (2011). However, Chou et al. used decision stump as base classifier; this research found that additive regression based on decision stump had the lowest accuracy and the other six tested base classifiers could improve the prediction performance of additive regression. Also, in Chou et al. (2011), the prediction performance of bagging with the base fast decision tree learner was not as good as the ANN model. In contrast, this study found that bagging could provide better prediction accuracy than the ANN model when using the advanced methods (i.e., Gaussian processes regression and multilayer perceptron) as base classifiers. Since this comparison was performed without fully evaluating the impact of variations between or among the compared studies (e.g., how extensively the tuning was performed and the difference between/among datasets) on model performance, the comparison results have to be cautiously interpreted.

5.3.7 Correlation between or among Input Variables

In this research, all the variables used in concrete mix design and lab testing were adopted as input variables for the tested data mining models. The strong correlation between or among input variables, if existing, could cause a problem called multicollinearity (Atici 2011). According to Alin (2010), multicollinearity is commonly defined as the linear relationship among two or more independent variables, which adds difficulty in determining the individual role of each independent variable and affects the reliability of model parameter estimates. The problem is related to the nature of the data. This research adopted the Weka attribute selector and JMP stepwise regression analysis to evaluate the effects of multicollinearity on the tested data mining models. The results suggested that the input variables could be reduced to a subset of four, including cementitious material (kg), concrete age (day), Micro Air (ml), and Haydite (kg), to give the best merit for this modeling problem.

Table 5.8 shows the correlation matrix for dependent and independent variables used in the tested models. It can be observed that three out of the four aforementioned input variables (i.e., cementitious material, concrete age, and Micro Air) have very high correlation with compressive strength. In contrast, there was almost no correlation between water and compressive strength. This is because the quantity of water was kept unchanged for all the batches in the mix design. Correlation analysis for the nine input variables reveals that except for cement type, concrete age, and water, other variables have some non-negligible correlation with each other.

Table 5. 8. Correlation matrix for dependent and independent variables

Variable	1	2	3	4	5	6	7	8	9
1. Cement type	1								
2. Concrete age	0	1							
3. Water	0	0	1						
4. PC/PLC	0.000	0	-0.000	1					
5. FA	-0.000	0.000	-0.000	-0.350	1				
6. Sand	0	0	0.000	-0.098	-0.495	1			
7. Pea gravel	0.017	-0.000	-0.000	0.017	0.083	-0.169	1		
8. Haydite	-0.016	-0.000	-0.000	-0.017	-0.084	0.169	-1.000	1	
9. Micro Air	0	0	-0.000	0.749	0.358	-0.447	0.075	-0.076	1
10. Compressive strength	0.122	0.584	-0.000	0.676	-0.004	-0.210	0.187	-0.187	0.671

This research reevaluated each data mining model's performance based on the four input variables suggested by the JMP stepwise regression analysis and the parameter setting listed in Table 5.4. The results presented in Table 5.10 show that eliminating five input variables only caused a very small performance loss for each tested model. However, since the analyses conducted in this research were based on the results of experiments performed to determine the effects of alternative materials on the compressive strength of environmentally friendly concrete, eliminating some variable(s) could lose information and reduce the generalizability of the model. Considering that the multicollinearity problem has little effect on the overall fit of a model and generally does not affect predictions (Studenmund 2000; Kutner et al. 2004), this research does not recommend dropping any variable(s).

Table 5. 9. Performance of models based on four input variables

Data mining model	R	Change in R	RMSE	MAE
Additive Regression	0.9744	-0.0174	3.2822	2.5442
Bagging	0.9744	-0.0163	3.281	2.5241
M5P	0.9646	-0.0089	3.8386	2.8782
REPTree	0.9437	-0.0164	4.9233	3.8702
M5Rules	0.9614	-0.0124	4.0111	3.062
SMOreg	0.9708	-0.0131	3.4967	2.6994
Multilayer Perceptron	0.9639	-0.0211	3.895	2.9032
Gaussian Processes	0.9744	-0.0177	3.2822	2.5442

5.3.8 Conclusions

This research aimed to evaluate the potential of using data mining techniques for predicting the compressive strength of environmentally friendly concrete containing FA, Haydite LWA, and/or PLC.

The obtained analytical results suggest that all of the tested models, except for decision stump, can provide acceptable prediction accuracy with R^2 ranging from 0.9217 (for REPTree) to 0.9842 (for Gaussian processes regression). The Gaussian processes regression model showed the best prediction accuracy as an individual data mining model. Also, when used as base classifier, it helped the two ensemble models achieve the best prediction performance. This observation is important since the Gaussian processes regression model is rarely investigated in previous work in this field.

The results of this research also indicate that in most cases, except for M5P, REPTree, and M5-Rules, training the models with the whole dataset containing PC and PLC concrete samples provided better prediction accuracy than using only the PC or PLC dataset.

Furthermore, although the three advanced data mining models achieved higher prediction accuracy than the four regression tree models, the time required for building and training these advanced models was significantly longer.

Chapter 6. Conclusions

6.1 Research Findings

As discussed before, the knowledge gap between the industry and research community in construction is what hindered the companies to explore the full value of their data assets. Furthermore, many researchers in the construction industry are still using simple basic statistical analytics to perform their studies. This research aimed to address and fill these gaps by performing a comprehensive analysis of existing application of big data analytics techniques for construction-related subjects and investigating the application of selected predictive models in predicting concrete strength and soil erosion. The findings of this research are summarized below.

6.1.1 Trend on the Implementation of Big Data Analytical Techniques in Construction Research

This dissertation first investigated the applications of selected big data analytical techniques in construction-related research from 2000 to 2015 (up to September) and provided a literature-driven analysis of the trends, directions, and status. For this purpose, the application of 26 popular big data analysis techniques in six different construction research areas (represented by 30 prestigious journals) was reviewed and analyzed.

This research identified 10,329 different papers in these six research areas. Several patterns, trends, and relationships were found as the results of this investigation. A data-driven list of 178 subcategories and 84 categories of construction related research subjects was created. The results were tabulated, mapped, visualized and explained. Most importantly, an application map of big data analytics techniques vs. construction related subjects was produced and the significant patterns were analyzed. Some of the main findings of this research include:

- The three areas of building energy and performance (50.6%), computation and analytics in construction (45.8%) and infrastructure (43.4%) had the higher application rates of the selected big data techniques compared to other areas.
- The area of computation and analytics in construction had the highest share of applications for the majority of big data techniques (17 out of 26).
- Percentages of papers in the selected journals that have applied big data analytics techniques slightly fluctuated between years and have become more stable (around 30%) since 2004.
- Simulation, predictive modeling, optimization, statistics, and regression were the five most frequently used techniques that had many overlap with other techniques (1/3 of the total papers).
- The six least frequently used techniques were A/B testing, sentimental analysis, crowdsourcing, unsupervised learning, supervised learning, association rule learning.
- Construction research field has a strong history of the application of techniques such as neural networks, classification, genetic algorithm, visualization, time series, data mining and network analysis.

- There is a great potential for applying techniques, including pattern recognition, signal processing, cluster analysis, data fusion, machine learning, ensemble learning, spatial analysis and natural language processing, in construction research field.

The results of this research can not only provide a better understanding of the application of these techniques in existing construction-related studies but also help practitioners and researchers to identify suitable analytic techniques for their specific research topics/problems.

6.1.2 Application of ANN in Construction Research – Case Studies

This research examined the capability and accuracy of ANN for predictions of two common construction research subjects. Two different datasets were applied. The first dataset includes the information and test results for 144 samples of concrete mixture design with different settings or testing ages. The results of analysis for this case study show that:

- MLP is an appropriate tool for predicting the CS of environmentally friendly concrete.
- Both input methods (numerical and relative) are accurate enough although the numerical method has a small advantage for PC-concrete and the relative method is slightly better for PLC-concrete.
- One hidden layer MLP models provide better prediction accuracy than the Two hidden layer MLP models

- The MLP with four independent input variables and the proper number of neurons in the hidden layer eliminates the multicollinearity problem and is still accurate enough for prediction, even though it is not recommended.
- The highest correlation coefficient (R) achieved in this case study was 0.9678, resulting from the model for PC-Number with (8-12-1) structure.

The second case study was related to a dataset containing test results for 442 settings of highway soil erosion amounts with different test sections, rainfall events, and vegetation communities. The results of the analysis suggest that:

- ANN has an acceptable accuracy for prediction of soil erosion in highway slopes.
- Tuning can improve the prediction performance of ANN for soil erosion.
- The highest R achieved in this case study was 0.9776, resulting from the model with (13-8-4-1) structure.
- The highest R observed for models with 10 input variables was 0.9727 resulting from the model with (10-17-1-1) structure, showing a very small performance loss compared to models with 13 input variables.

6.1.3 Comparison of Different Data Mining Techniques for Predicting Compressive Strength of Environmentally Friendly Concrete

The potential and accuracy performance of using data mining techniques for predicting the compressive strength of environmentally friendly concrete containing FA, Haydite LWA, and/or PLC was evaluated. In particular, four common regression tree models (M5P, REPTree, M5-Rules, and decision stump) and three more advanced predictive models (ANN based on multilayer perceptron, SMOreg-based SVM regression, and Gaussian processes regression) were generated and tested individually. Then they were used as base classifiers in two ensemble models (additive regression and bagging) to evaluate the effects of boosting. The results of this research indicate that:

- All of the tested models, except for decision stump, provided acceptable prediction accuracy with R^2 ranging from 0.9217 to 0.9842 with proper tuning.
- The Gaussian processes regression model, which was rarely investigated in previous work in this field, showed promising results:
    - Achieved the best prediction accuracy as an individual data mining model
    - When used as base classifier, helped the two ensemble models achieve the best prediction performance
    - Showing great potential for further study
- Although the three advanced data mining models achieved higher prediction accuracy, the time required for building and training these models was significantly longer than other models. This should be considered a factor in choosing an appropriate data mining model in practice. Particularly, when dealing with a very large dataset, using an ensemble method with a regression tree base classifier seems to be a more practical alternative.
- In most cases, training the models with the whole dataset containing PC and PLC concrete samples provided better prediction accuracy than using only the PC or PLC dataset.

With the demonstrated potential of using data mining models to predict concrete comprehensive strength, future research can adopt this approach to study other properties of concrete such as tensile strength, durability, or concrete slump.

6.2 Future Research

The original purpose of this dissertation was to propose a framework for automating the generation of an application map of big data analytics technique vs. construction research subjects for various research areas. The plan was to first manually analyze the publication and generate an accurate dataset and then to use supervised learning and natural language processing techniques to automate this process of information extraction. However, the primary result of this automation did not achieve satisfactory accuracy and the time limitation did not allow the author to further investigate this potential. This approach will be a topic for future research.

References

Abdeen, M. A. M., and Hodhod, H. (2010). "Experimental investigation and development of artificial neural network model for the properties of locally produced light weight aggregate concrete." *Sci. Res. Org. Eng.*, 2(6), 408-419.

Abdollahzadeh, A, M Mukhlisin, and A.E Shafie. "Predict Soil Erosion with Artificial Neural Network in Tanakami (japan)." *Wseas Transactions on Computers*. 10.2 (2011): 51-60. Print.

ACI. (1997). *Standard practice for selecting proportions for structural lightweight concrete*. ACI Manual of Concrete Practice, 211 pg.

ADAT, (2015). <http://adat.crl.edu/platforms/about/engineering_village> (Oct. 05, 2015).

AECBIGDATA (2013). "Next major disruptive force" *AEC BIG DATA*, <http://www.aecbigdata.com/2013/02/09/next-major-disruptive-force/> (Apr.23, 2013).

Afan, H. A., El-Shafie, A., Yaseen, Z. M., Hameed, M. M., Wan, M. W. H. M., and Hussain, A. (March 01, 2015). ANN Based Sediment Prediction Model Utilizing Different Input Scenarios. *Water Resources Management: An International Journal - Published for the European Water Resources Association (ewra), 29,* 4, 1231-1245.

Aiyer, B. G., Kim, D., Karingattikkal, N., Samui, P., and Rao, P. R. (2014). "Prediction of compressive strength of self-compacting concrete using least square support vector machine and relevance vector machine." *KSCE J. Civ. Eng.*, 18(6), 1753-1758.

Akande, O. K., Owolabi, O. T., Twaha, S., and Olatunji, S. O. (2014). "Performance comparison of SVM and ANN in predicting compressive strength of concrete." *IOSR J. Comput. Eng.*, 16(5), 88-94.

Akhavian, R., and Behzadan, A. H. (2015). "Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers." *Adv. Eng. Info,* 3.

Alin, A. (2010). "Multicollinearity." *WIREs Comp Stat*, 2(May/Jun), 370-374.

Alshihri, M. M., Azmy, A. M., and El-Bisy, M. S. (2009). "Neural networks for predicting compressive strength of structural light weight concrete." *Constr. Build. Mater*, 23(6), 2214-2219.

Atici, U. (2011). “Prediction of the strength of mineral admixture concrete using multivariable regression analysis and an artificial neural network.” *Expert Syst. Appl.,* 38(8), 9609-9618.

ASTM C 31/C 31M-06, *Standard practice for making and curing concrete test specimens in the field*. ASTM International, West Conshohocken, PA, 2007, 3rd Edition, 1-5.

Banerjee, U. (2012).” What is the Definition of Big Data?” *Technology Trend Analysis*, <http://setandbma.wordpress.com/2012/12/21/definition-of-big-data/> (Apr.23, 2013).

Basri, H. B., Mannan, M. A., and Zain, M. F. M. (1999). “Concrete using waste oil palm shells as aggregate.” *Cem. Concr. Res.*, 29(4), 619–622.

Berry, M., Stephens, J., and Cross, D. (2011). “Performance of 100% fly ash concrete with recycled glass aggregate.” *ACI Mater. J.,* 108(4), 378-384.

Big data fact sheet. (2012), <https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf>

Bishop, C. M. (2006). *Pattern recognition and machine learning*, Springer, New York.

Bloom, E., and Gohn, B. (2012). *Smart Buildings: Top Trends to Watch in 2012 and Beyond.*, Pike Research LLC, Boulder, USA.

Breiman, L. (1994). “Bagging predictors.” *Mach. Learn.,* 24(2), 123-140.

Brilakis, I., and Soibelman, L. (2005). "Content-Based Search Engines for Construction Image Databases." *Autom. Constr.,* 14(4).

Buhlmann, P., and Yu, B. (2002). “Analyzing bagging.” *Ann. Statist.,* 30, 927-961.

Burrell, G., and Morgan, G. (1979). *Sociological Paradigms and Organizational Analysis: Elements of the Sociology of Corporate Life,* Heinemann, London.

Byers, A. (2011). “Big Data: The next frontier for innovation, competition, and productivity.” New York, NY: McKinsey and Company

Caudill, M. (1987). “Neural networks primer.” *J. AI Expert*, 2(12), 46-52.

Carter, P. (2011). *Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO,* International Data Corporation, Framingham, Massachusetts, USA

Chartered Institute of Building (2003). *Code of Practice for Project Management for Construction and Development*. Blackwell Pub., Oxford, UK; Malden, MA.

Concrete Monthly. (2004). *Cement formulation change to lower emissions* [WWW Document]. <URL http://www.concretemonthly.com/monthly/art.php/424>

Chan, A. P. C., Scott, D., and Lam, E. W. M. (2002). "Framework of Success Criteria for Design/Build Projects" *Journal of Management in Engineering,* 18(3), 120-128.

Chan, A. P. C., Chan, D. W. M., Chiang, Y. H., Tang, B. S., Chan, E. H. W., and Ho, K. S. K. (2004). "Exploring Critical Success Factors for Partnering in Construction Projects" *Journal of Construction Engineering and Management,* 130(2), 188-198.

Constructech, (2012). "BIM, Big Data, and Construction" *Constructech*, <http://www.constructech.com/news/articles/article.aspx?article_id=9503> (Apr.23, 2013)

Cortes, C., and Vapnik, V. (1995). "Support-vector networks." *Mach. Learn.,* 20(3), 273-297.

Cheng, M. Y., Chou, J. S., Roy, A. F., and Wu, Y. W. (2012). "High-performance concrete compressive strength prediction using time-weighted evolutionary fuzzy support vector machines inference model." *Autom. Constr.,* 28, 106-115.

Philip Chen C, Zhang C-Y (2014), "Data-intensive applications, challenges, techniques and technologies: a survey on big data." *Inf. Sci.,* 275:314–347

Chou, J. S., Chiu C. K., Farfoura M., and Al-Taharwa I. (2011). "Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques." *J. Comput. Civ. Eng.,* 25(3), 242-253.

Chu, W., and Ghahramani, Z. (2006). "Gaussian processes for ordinal regression." *J. Mach. Learn. Res.,* 6(7), 1019-1042.

Davies, R., and Harty, C. (2013). "Implementing Site BIM: A case study of ICT innovation on a large hospital project." *Automation in Construction*, 30, 15–24. doi:10.1016/j.autcon.2012.11.024.

Dasarathy, B. V., and Sheela, B. V. (1979). "A composite classifier system design: Concepts and methodology." *Proc. IEEE.*, 67(5), 708-713.

Deepa, C., Sathiyakumari, K., and Sudha, V. (2010). "Prediction of the compressive strength of high performance concrete mix using tree based modeling." *Int. J. Comput. Appl. T.,* 6(5), 18-24.

D'Oca, S., and Hong, T. (2015). "Occupancy schedules learning process through a data mining framework." *Energy Build. 88,* 395-408.

Dumbill, D. (2012). “An introduction to the big data landscape” *What is big data*, <http://strata.oreilly.com/2012/01/what-is-big-data.html> (Apr.23, 2013)

Economist Intelligence Unit (EIU). (2011), *Big data: Harnessing a game-changing asset*, The Economist Group, London, England.

Elghamrawy, T., Boukamp, F., The role of VR and BIM to manage the construction and design processes. (2010). "Managing Construction Information using RFID-Based Semantic Contexts." *Autom. Constr.,* 19(8), 1056-1066.

Erdal, H. I., Karakurt, O., and Namli, E. (2013). “High performance concrete compressive strength forecasting using ensemble models based on discrete wavelet transform.” *Eng. App. Artif. Intell.,* 26(4), 1246-1254.

Etxeberria, M., Vázquez, E., Marí, A., and Barra, M. (2007). “Influence of amount of recycled coarse aggregates and production process on properties of recycled aggregate concrete.” *Cem. Concr. Res.*, 37(5), 735-742.

Fazel Zarandi, M. H., Türksen, I. B., Sobhani, J., and Ramezanianpour, A. A. (2008). “Fuzzy polynomial neural networks for approximation of the compressive strength of concrete.” *Appl. Soft Comput.,* 8(1), 488–498.

Expanded Shale, Clay & Slate Institute (ESCSI). (2007). *Chapter 1. Overview and history of the expanded shale, clay and slate industry*. Salt Lake City, Utah. 1-2.

Forsyth, A. R., Bubb, K. A., and Cox, M. E. (January 01, 2006). Runoff, sediment loss and water quality from forest roads in a southeast Queensland coastal plain Pinus plantation.(Report). *Forest Ecology and Management, 221,* 1-3.

Flanagan, D C, J E. Gilley, and T G. Franti. "Water Erosion Prediction Project (wepp): Development History, Model Capabilities, and Future Enhancements." *Transactions of the Asabe*. 50.5 (2007). Print.

Freund, Y., and Schapire, R. E. (1996). “Experiments with a new boosting algorithm.” *Mach. Learn.: Proc., 13th Int. Conf.*, Bari, Italy, 148-156.

Friedman, J. H. (2001). “Greedy function approximation: A gradient boosting machine.” *Ann. Statist.,* 29(5), 1189-1232.

Friedman, J. H., and Meulman, J. J. (2003). “Multiple additive regression trees with application in epidemiology.” *Stat. Med.,* 22(9), 1365-1381.

Gandomi, A., and Haider, M. (2014). “Beyond the hype: Big data concepts, methods, and analytics”. *Inter. J. of Info. Manage., 35(*2), 137-144.

Gantz, J. and Reinsel, D. (2012). *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*, International Data Corporation, Global Headquarters, Framingham, MA, USA.

Garcia, J., Fdez-Valdivia, J., Rodriguez-Sanchez, R., and Fdez-Vidal, X. (2002). "Performance of the Kullback-Leibler information gain for predicting image fidelity." *Proc., 16th Int. Conf. Patt. Recog.,* Vol. 3, IEEE Computer Society Press, Washington, D.C., 843-848.

Gartner, 2012. Gartner IT Glossary - Big Data. Accessed 28th April 2016. <http://www.gartner.com/it-glossary/big-data.>

Gupta, R., Kewalramani, M. A., and Goel, A. (2006). "Prediction of concrete strength using neural-expert system." *J. Mater. Civ. Eng.*, 18(3), 462-466.

Gupta, S. M. (2007). "Support vector machines based modelling of concrete strength." *World Acad. Sci.: Eng. Technol.,* 3(1), 12-18.

Green, K. and Tashman, L. (2009). "Percentage error: What denominator?" *Foresight: The International Journal of Applied Forecasting,* 12(Winter), 36–40.

Hand, D., Mannila, H., and Smyth, P. 2001. *Principles of Data Mining*. MIT Press

Harris, T. M., and Boardman, J. 1998. "Alternative Approaches to Soil Erosion Prediction and Conservation Using Expert Systems and Neural Networks". *NATO ASI SERIES I GLOBAL ENVIRONMENTAL CHANGE.* 55, 461-480.

Haydite Hydraulic Press Brick Company. (2013). *Haydite lightweight concrete* [WWW Document]. URL <http://www.hpbhaydite.com/concrete/concrete.html>

Holmes, G., Hall, M., and Frank, E. (1999). "Generating rule sets from model trees." *Lect. Notes Comput. Sci.,* 1747, 1-12.

Howard, P. (2012). "*Big Data Analytics with Hadoop and Sybase IQ*", White Paper by Bloor Research, April 2012, London, United Kingdom

Hilbert, M. (2016), "Big Data for Development: A Review of Promises and Challenges", *Development Policy Review*, Vol. 34 No. 1, pp. 135–174

Iba, W., and Langley, P. (1992). "Induction of one-level decision trees." *Proc., 9th Int. Conf. Mach. Learn.,* Morgan Kaufmann, San Francisco, CA, 233–240.

International Data Corporation (IDC), (2012). *Worldwide Big Data technology and Services 2012-2015 Forecast*, IDC Global Headquarters, MA, USA.

Jin, R. (2013). *A statistical modeling approach to studying the effects of alternative and* *waste materials on green concrete properties*, Ph.D. Dissertation, The Ohio State University, Columbus, OH. <http://rave.ohiolink.edu/etdc/view.cgi?acc%5Fnum=osu1372854071>.

John, J. St. (2012). “Big Data for Big Buildings: SPARC Greens the VA.” *Greentechmedia*, <http://www.greentechmedia.com/articles/read/big-data-for-big-buildings-sparc-greens-the-va> (Apr.23, 2013).

Kenai, S., Soboyejo, W., & Soboyejo, A. (2004). “Some engineering properties of limestone concrete.” *Mater. Manuf. Process.,* 19(5), 949-962.

Kevern, J. T., Schaefer, V. R., and Wang, K. (2011). “Mixture proportion development and performance evaluation of pervious concrete for overlay applications.” *ACI Mater. J.,* 108(4), 439-448.

Khalaf, F. M., and Devenny, A. S. (2004). “Recycling of demolished masonry rubbles as coarse aggregate in concrete: Review.” *J. Mater. Civ. Eng.*, 16(4), 331–340.

Khan, S. U., Ayub, T. F. A., and Rafeeqi, S. (2013). “Prediction of compressive strength of plain concrete confined with ferrocement using artificial neural network (ANN) and comparison with existing mathematical models.” *Amer. J. Civ. Eng. Arch.*, 1(1), 7-14.

Kim, H., Soibelman, L., Grobler, F. (2008). "Factor Selection for Delay Analysis using Knowledge Discovery in Databases." *Automation in Construction,* 17(5), 550-560.

Kotsiantis, S., Kanellopoulos, D., and Zaharakis, I. (2006). “Bagged averaging of regression models.” *Int. Feder. Inf. Process. Publications (Ifip),* 204, 53-60.

Kutner, M., Nachtsheim, C., and Neter, J. (2004). *Applied linear statistical models*, 5th Ed., McGraw-Hill, New York, NY.

Laflen, J. M., Flanagan, D. C., and Engel, B. A. (2004). “Soil erosion and sediment yield prediction accuracy using wepp.” *Jawra Journal of the American Water Resources Association*, 40(2), 289-297.

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. META Group Research Note 6, 70.

Lee, S. (February 01, 2004). “Soil erosion assessment and its verification using the Universal Soil Loss Equation and Geographic Information System: a case study at Boun, Korea.” *Environmental Geology: International Journal of Geosciences, 45,* 4, 457-465.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of massive datasets,* Cambridge University Press, U.K. <http://proquest.safaribooksonline.com/9781316147047> (Jun. 6, 2015).

Li, H., Cheng, E. W., Love, P. E. (2000). "Partnering Research in Construction" *Engineering Construction and Architectural Management,* 7, 76-92.

Lin, K. Y., Soibelman, L. (2006). "Promoting Transactions for A/E/C Product Information." *Autom. Constr.,* Knowledge Enabled Information System Applications in Construction, 15(6), 746-757.

Lin, Y. (2008). "Developing Construction Assistant Experience Management System using People-Based Maps." *Automation in Construction Automation in Construction,* 17(8), 975-982.

Licznar, P., and Nearing, M. A. (2003). "Artificial neural networks of soil erosion and runoff prediction at the plot scale." *Catena,* 51(2), 89-114.

Limbachiya, M., Meddah, M. S., and Ouchagour, Y. (2012). "Performance of Portland/silica fume cement concrete produced with recycled concrete aggregate." *ACI Mater. J.,* 109(1), 91-100.

Lopes, Vicente Lucio. 1987. *A numerical model of watershed erosion and sediment yield*. Ann Arbor, Mich: University Microfilms International.

Lubeck, A., Gastaldini, A., Barin, D., and Siqueira, H. (2012). "Compressive strength and electrical properties of concrete with white Portland cement and blast-furnace slag." *Cem. Concr. Compos.,* 34(3), 392-399.

Lumpkin, G. (2013). "Integrated for Insight" *Oracle White Paper*, <http://www.oracle.com/us/technologies/big-data/big-data-strategy-guide-1536569.pdf> (Apr.23, 2013)

Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C. and Hung

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5(4), 115-133.

Madlool, N., Saidur, R., Rahim, N., and Kamalisarvestani, M. (2013). "An overview of energy savings measures for cement industries." *Renew. Sust. Energ. Rev*., 19, 18-29.

McKinsey Global Institute (MGI). (2011)*, Big Data: The Next Frontier for Innovation, Competition, and Productivity.* MGI, Lexington, KY. KY.

Menandas, J.J., Joshi, J.J. (2014). "Data Mining with Parallel Processing Technique for Complexity Reduction and Characterization of Big Data", *Glob J. of Adv. Res.* Vol-1, Issue-1 PP. 69-80.

Mount, N. J., and Abrahart, R. J. (2011). “Load or concentration, logged or unlogged? Addressing ten years of uncertainty in neural network suspended sediment prediction.” *Hydrological Processes, 25,* 20, 3144-3157.

Mortensen, C. (2012), *Leveraging Analytics and Big Data for Business Growth: 7 Steps to Kick start your Big Data implementation*, International Data Corporation, IDC Asia/Pacific, 80 Anson Road, Singapore.

Mosley, M., Brackett, M. H., Earley, S., Henderson, D. (2009). "DAMA guide to the data management body of knowledge (DAMA-DMBOK guide)." *Data Administration Management Association International,* Town and Country, MO.

Nataraja, M. C., Jayaram, M. A., and Ravikumar, C. N. (2006). “A fuzzy-neuro model for normal concrete mix design.” *Eng. Letters*, 13(3), 98-107.

Omran, B. A., Chen, Q., and Jin, R. (2016). "Comparison of Data Mining Techniques for Predicting Compressive Strength of Environmentally Friendly Concrete." J. Comput. Civ. Eng., 10.1061/(ASCE)CP.1943-5487.0000596, 04016029.

Omran, B. A., Chen, Q., and Jin, R. (2014). “Prediction of Compressive Strength of ‘Green’ Concrete Using Artificial Neural Networks.” *Proc., 50th ASC Ann. Int. Conf.,* Associated Schools of Construction (ASC), Windsor, CO.

Poole, D. L., Mackworth, A. K., Goebel, R. (1998). *Computational Intelligence: A Logical Approach*, Oxford University Press, New York.

Platt, J. C. (1998). *Sequential minimal optimizer: A fast algorithm for training support vector machines*, Technical Report MSR-TR-98-14, Microsoft Research, Redmond, WA.

Quinlan, J. R. (1992). “Learning with continuous classes.” *Proc., 5th Australian Joint Conf. Artif. Intell.*, Sydney, Australia, 343–348.

Rasmussen, C. E., and Williams, C. K. I. (2006). *Gaussian processes for machine learning*, MIT Press, Cambridge, MA.

Research Trends. (2012). *Special Issue on Big Data*, Elsevier, Amsterdam, Netherlands, Issue 30.

Rokach, L. (2010). “Ensemble-based classifiers.” *Artif. Intell. Rev.: Int. Sci. Eng. J.*, 33(1-2), 1-39.

Rosenblatt, F. (1958). “The perceptron: A probabilistic model for information storage and organization in the brain.” *Psychol. Rev*., 65(6), 386-408.

Rumelhart, D. E., Widrow, B., and Lehr, M. A. (1994). “The basic ideas in neural networks.” *Commun. ACM*, 37(3), 87-92.

Russell, S. J., and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*, Prentice Hall/Pearson Education, Upper Saddle River, N.J. Sacks, R., Navon, R., Shapira, A., Brodetsky, I. (2003). "Monitoring Construction Equipment for Automated Project Performance Control." *NIST Special Publication* /, (989), 161-166.

Sage. (2014). *Sage 2014 construction IT survey*, <http://www.sage.com/na/~/media/sage-job ready/assets/2014_construction_it-survey_infographic-1> (Oct. 05, 2015)

Saridemir M., Ozcan F., Severcan M. H., and Topçu I. B. (2009). "Prediction of long-term effects of GGBFS on compressive strength of concrete by artificial neural networks and fuzzy logic." *Constr. Build. Mater.*, 23(3), 1279-1286.

Shahria Alam, M., Slater, E., and Muntasir billah, A. H. M. (2013). "Green concrete made with RCA and FRP scrap aggregate: Fresh and hardened properties." *J. Mater. Civ. Eng.,* 25(12), 1783-1794.

Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., and Murthy, k. R. k. (2000). "Improvements to the SMO algorithm for SVM regression." *IEEE Trans. Neural Net.,* 11(5), 1188-93.

Studenmund, A. H. (2000). *Using econometrics: A practical guide, 4th Ed.*, Addison-Wesley, Reading, MA.

Teicholz, P. (2004). "Labor Productivity Declines in the Construction Industry: Causes and Remedies" *AECbytes*, <http://www.aecbytes.com/viewpoint/2013/issue_67.html> (Apr.23, 2013)

Topçu, I. B., and Saridemir, M. (2007). "Prediction of properties of waste AAC aggregate concrete using artificial neural network." *Comp. Mater. Sci.,* 41(1), 117-125.

The white house. (2012). <https://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal>

Thomas, M. D. A., Cail, K., Blair, B., Delagrave, A., and Barcelo, L. (2010). Equivalent performance with half the clinker content using PLC and SCM. *Proc., NRMCA Concrete Sustainability Conference.*

Tserng, H. P., Yin, S. Y. L., Dzeng, R. J., Wou, B., Tsai, M. D., Chen, W. Y. (2009). "A Study of Ontology-Based Risk Management Framework of Construction Projects through Project Life Cycle." *Autom. Constr.,* 18(7), 994-1008.

Tu, J. V. (1996). "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes." *Journal of Clinical Epidemiology,* 49*(*11), 1225-1231.

Tukey, J. W. (1977). *Exploratory data analysis*, Addison-Wesley, Reading, MA.

United States Census Bureau (2013). “Annual Value of Construction Put in Place 2002-2012” *U.S. Census Bureau*, <http://www.census.gov/const/C30/total.pdf> (Apr.23, 2013)

USDA. 2016. *Revised Universal Soil Loss Equation (RUSLE),* Welcome to RUSLE 1 and RUSLE 2", Agricultural Research Service < https://www.ars.usda.gov/southeast-area/oxford-ms/national-sedimentation-laboratory/watershed-physical-processes-research/docs/revised-universal-soil-loss-equation-rusle-welcome-to-rusle-1-and-rusle-2/>

Venkatraman, S., and Yoong, P. (2009). “Role of mobile technology in the construction industry - a case study.” *Inter. J. of Business Info. Syst.,* 4*(*2), 195-209.

Vilalta, R., and Drissi, Y. (2002). “A perspective view and survey of meta-learning.” *Artif. Intell. Rev.,* 18(2), 77-95.

Vilalta, R., Giraud-Carrier, C., Brazdil, P., and Soares, C. (2004). “Using meta-learning to support data-mining.” *Int. J. Comput. Sci. Appl.*, 1(1), 31-45.

Violino, B. (2013). The 'Internet of things' will mean really, really big data, <http://www.infoworld.com/article/2611319/computer-hardware/the--internet-of-things--will-mean-really--really-big-data.html>

Wactlar, H. (2012). ”Big data R&D initiative.” CISE Directorate, National Science Foundation, NIST Big Data Meeting.

Waikato, (2015). <http://www.cs.waikato.ac.nz/ml/index.html> (Jun. 6, 2015).

Wang C., Chen M.H., Schifano, E., Wu, J., and Yan, J. (2015). “Survey of Statistical Methods and Computing for Big Data”, arXiv: 1502.07989v1 [stat.CO].

Wang, Y., and Witten, I. H. (1997). “Introduction of model trees for predicting continuous classes.” *Proc., 1997 Euro. Conf. Mach. Learn.*, University of Economics, Faculty of Informatics and Statistics, Prague.

WBCSD (World Business Council for Sustainable Development). (2009). Cement technology roadmap 2009 - Carbon emissions reductions up to 2050. <http://www.wbcsd.org/Pages/EDocument/EDocumentDetails.aspx?ID=11423&NoSearchContextKey=true> (Jun. 6, 2014).

Williams, J. R. (1975). Sediment – yield prediction with universal equation using runoff energy factor. Proceedings of the sediment Yield Workshop, USDA Sedimentation Laboratory, Oxford, Mississippi.

Wipro (2013). “Spotlight: Big Data” *Analytics & Information Management*, <http://www.wipro.com/services/analytics-information-management/ > (Apr.23, 2013)

Wing, C.K. (1997)."The Ranking of Construction Management Journals" *Construction Management and Economics*, 15(4), 387-398.

Wischmeier, Walter H., and Dwight David Smith. 1978. *Predicting rainfall erosion losses: a guide to conservation planning*. [Washington, D.C.]: U.S. Department of Agriculture, Science, and Education Administration.

Witten, I. H., and Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*, 2nd Ed., Morgan Kaufman, San Francisco, CA.

Woo, J., Clayton, M. J., Johnson, R. E., Flores, B. E., Ellis, C. (2004). "Dynamic Knowledge Map: Reusing Experts' Tacit Knowledge in the AEC Industry." *Autom. Constr.,* 13(2). 203-207.

Woolhiser, D. A., Smith, R. E., Goodrich, D. C., (1990). *KINEROS, a kinematic runoff and erosion model: documentation and user manual*. U.S. Department of Agriculture, Agricultural Research Service, p. 130.

Wu, X., Wu, G.-Q., Zhu, X., and Ding, W. (2014). Data mining with big data. *IEEE Trans. on Know. Data Eng.* 26, 97-107.

Xinyu Geng, Hualong Dai, and Li Yang. 2015. "Prediction of Soil Erosion Induced Sediment Production using Fuzzy Neural Network Model." *Metallurgical & Mining Industry.* 2015 (8).

Xue, X., Shen Q., Fan H., Li H., Fan S. (2012). "IT Supported Collaborative Work in A/E/C Projects: A Ten-Year Review." *Automation in Construction,* 21(1), 1-9.

Yan, K., Xu, H., Shen G., and Liu P. (2013). "Prediction of splitting tensile strength from cylinder compressive strength of concrete by support vector machine." Adv. Mater. Sci. Eng., 10.1155/2013/597257.

Yang, E. I., Yi, S. T., and Leem, Y. M. (2005). "Effect of oyster shell substituted for fine aggregate on concrete characteristics: Part I. Fundamental properties." *Cem. Concr. Res.*, 35(11), 2175-2182.

Yazdi, J. S., Kalantary, F., and Yazdi, H. S. (2013). "Prediction of elastic modulus of concrete using support vector committee method." *J. Mater. Civ. Eng.*, 25(1), 9-20.

Yeh, I. C. (1998). "Modeling of strength of high performance concrete using artificial neural networks." *Cem. Concr. Res.,* 28(12), 1797–1808.

Yeh, I. C., and Lien, L.-C. (2009). "Knowledge discovery of concrete material using genetic operation trees." *Expert Syst. Appl.*, 36(3), 5807–5812.

Ylijoki O, Porras J. 2016. “Perspectives to definition of big data: a mapping study and discussion.” *Journal of Innovation Management* **4**(1): 69–91.

Yu, W. D. and Hsu, J.-Y. (2013). "Content-Based Text Mining Technique for Retrieval of CAD Documents." *Automation in Construction,* 31, 65-74.

Zhai, D., Goodrum P.M., Haas C.T., Caldas C.H. (2009). "Relationship between Automation and Integration of Construction Information Systems and Labor Productivity" *Journal of Construction Engineering and Management,* 135(8), 746-753.

Zhang, G. (1998). “Forecasting with artificial neural networks: The state of the art.” *Int. J. Forecasting,* 14(1), 35-62.

Zheng, S., You, Y., Li, F., and Liu, G. (2012). "The Remote Video Monitoring System Design and Development for Underground Substation Construction Process" *Communications in Computer and Information Science*, (345), 601-605.

Zhu, B.F., Jian, G. (2014). “Discussion of the evaluation method and value of green building's POE in the era of large data.” *J. of Harbin Inst. of Tech. (1005-9113)*, 21(5), 10.

Appendix A: Categories and Subcategories

Category	Subcategory	Category	Subcategory	Category	Subcategory
Economic Analysis	Cost Analysis	**Layout Planning**	Construction Site Layout	**Positioning System**	Positioning System
	Time Cost Analysis		Layout Planning		Location Sensing
	Investment Analysis		Urban Layout Planning		Sensor Placement
	Asset Management	**Sustainable Development and Waste Management**	Sustainable Development	**Building Information Modeling BIM**	Building Information Modeling BIM
	Social Capital		Construction Waste Management	**Non-Construction Related Paper**	Non-Construction Related Paper
	Cost Benefit Analysis		Waste Water Management		Optimal Control
Knowledge Discovery and Information Management	Knowledge and Information Management		Sustainable Development and Waste Management	**Project Management**	Project Management
	Knowledge Discovery Techniques	**Predicting Concrete Properties**	Predicting Compressive Strength Of Concrete	**Construction Simulation**	Construction Simulation
	Data Analysis		Predicting Concrete Properties	**Document Identification and Management**	Document Management
	Data Visualization		Concrete Design Optimization		Document Identification
	Data Mining		Concrete Mixture	**Education**	Education
	Data Integration	**Project Planning**	Project Planning	**Process Management**	Process Management
	Database Systems	**Seismology**	Seismology	**Review**	Review
	Information Technology	**Project Execution and**	Project Execution and Operation	**Excavation**	Excavation Analysis

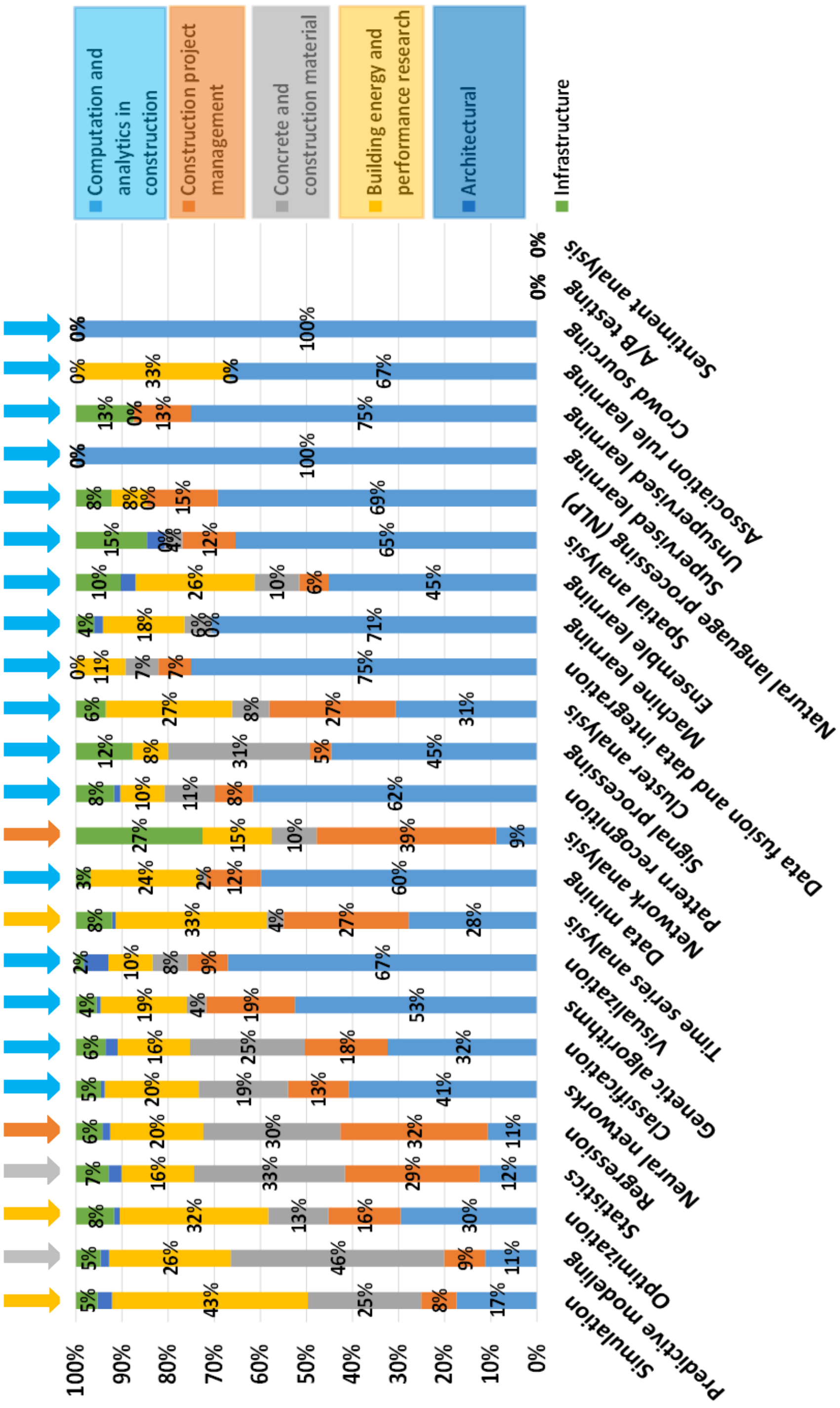
		Operation			
	Noisy Data Classification	**Damage Detection**	Damage Detection		Excavation Automation
	LIDAR Data Analysis	**System Identification and Analysis**	System Identification	**Organizational Issue**	Organizational Issue
	Data Collection		System Analysis	**Life Cycle Analysis**	Life Cycle Analysis
	Information Management	**Accident Prevention and Road Safety**	Road -Safety and Accident Prevention	**Procurement Management**	Procurement Management
	Knowledge Management Information Technology		Incident Detection	**Tower Crane Operation**	Tower Crane Operation
	Knowledge Discovery	**Infrastructure Assessment**	Infrastructure Condition Assessment		Predicting Hoisting Time
Design	Design Analysis		Infrastructure Analysis		Tower Crane Tracking
	Structural Design		Infrastructure Assessment		Temporary Hoist Planning
	Highway Design	**Action/Object Recognition and Image Processing**	Action Recognition	**Environmental Issue (Air and Water Pollution)**	Biological Activity Prediction
	Architectural Design		Image Processing		Water Pollution
	Collaborative Design and Construction		Object Recognition		Air Pollution
	Green Building Design	**Maintenance and Facilities Management**	Maintenance Management		Emission Control
Traffic and Transportation Management	Traffic Flow Evaluation		Facility Management		Environmental Engineering
	Transportation Management		Maintenance and Facilities Management		Environmental Impact
	Vehicle Classification	**Material Analysis**	Material Analysis	**Market/Customer Analysis**	Customer Analyses

	Freeway Operating Condition	**Resource Management**	Resource Management		Market Analysis
	Traffic Signals Timing	**Text Processing**	Text Processing	**Soil Typological Classification**	Soil Typological Classification
Monitoring and Controlling Projects	Structural Health Monitoring	**Building Energy Performance**	Building Energy Performance	**Conflict Management**	Conflict Management
	Project Monitoring			**Disasters Management and Emergency Response**	Emergency Response
	Project Progress Monitoring	**Miscellaneous Topics**	Benchmarking		Disasters Management
	Nonintrusive Load Monitoring		Office Classification	**Construction Equipment/Machinery Management**	Construction Machinery
	Crowd Monitoring System		Prediction Of Flashover		Operational Efficiency Of Construction Equipment
	Environmental Monitoring Network		Speech Recognition		Heavy Construction Equipment Value
	Monitoring and Controlling Projects		Websites		Construction Equipment Management
Algorithms	Algorithms		Other Topics	**Corrosion Detection**	Corrosion Detection
Project Scheduling	Project Scheduling	**Network Modeling and Optimization**	Network Modeling and Optimization	**Innovation Assessment**	Innovation Assessment
Infrastructure Management	Infrastructure Management	**Other Modeling Research**	Temporal Modeling	**Spatial Analysis**	Spatial Analysis
Safety Management and Occupational Health	Safety Management and Occupational Health		Feature-Based Modeling		Spatial Query Language
Project Performance Analysis	Performance Analysis		Model Generation	**Thermal Analysis and HVAC Analysis**	HVAC Analysis
	Project Performance Analysis		Evaporation Modeling		Thermal Analysis

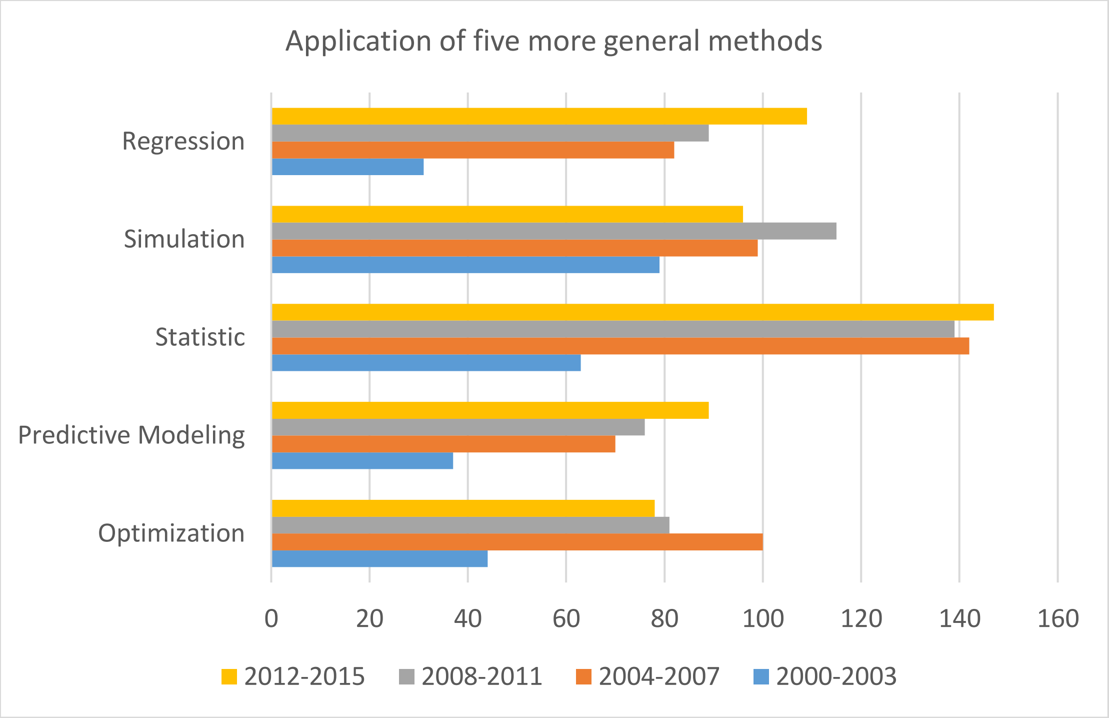
	Power Plant Analysis		Modeling Blast Wave Propagation	**Fault Detection**	Fault Detection
Risk Analysis and Management	Risk Analysis and Management		Enterprise Modeling	**Robotics**	Robotics
Defect Detection	Defect Detection		Industrial Simulation Model	**Contract Management**	Contract Management
Project Collaboration and Communication	Project Communication	**Production Management**	Production Management	**Stakeholder Management**	Stakeholder Management
	Collaborative Construction		Product Evaluation	**Construction Litigation**	Construction Litigation
	Project Collaboration and Communication		Product Data Models	**Slope Stability**	Slope Stability
Pavement Evaluation	Pavement Evaluation	**Bidding Strategy**	Bidding Strategy	**Space Use Analysis**	Space Use Analysis
Decision Making	Decision Making	**Human Resource Management**	Human Resource Management	**Aggregate Classification**	Aggregate Classification
	Strategic Analysis and Decision Making		Human Resource Planning	**Deterioration Detection**	Deterioration Detection
Artificial Reality	Augmented Reality	**Crack Detection**	Crack Detection	**Geotechnical Engineering**	Geotechnical Engineering
	Virtual Reality	**Water Distribution Systems**	Water Distribution Systems	**Lean Construction**	Lean Construction
	Mixed Reality	**Contractor Qualification**	Contractor Qualification	**Rework Analysis**	Rework Analysis
Structural Analysis	Structural Analysis	**Dispute Management**	Dispute Prediction	**Tunneling**	Tunneling
	Building Dynamic Properties		Dispute Forecasting	**Delay Analysis**	Delay Analysis
Productivity Analysis	Productivity Analysis		Dispute Resolution	**Problem Solving**	Problem Solving
		Construction	Construction	**Transaction**	Transaction

	Industry	Industry	**Deletion**	Deletion

Appendix B: Distribution of Papers Using Each Big Data Technique in Six Construction-Related Research Areas



Appendix C: Frequency of Application of Top Five Most Used Methods in Four Periods of Time

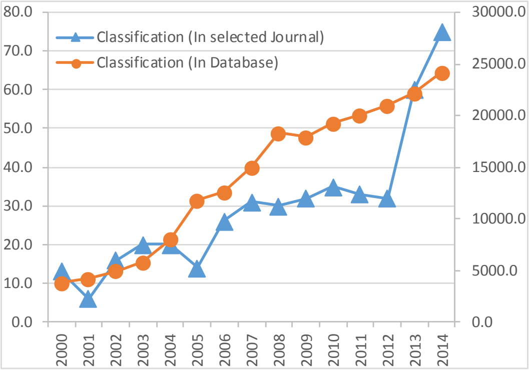


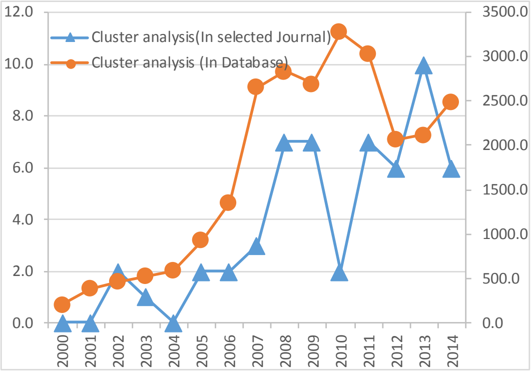
Appendix D: Data and Documents in the Construction Industry

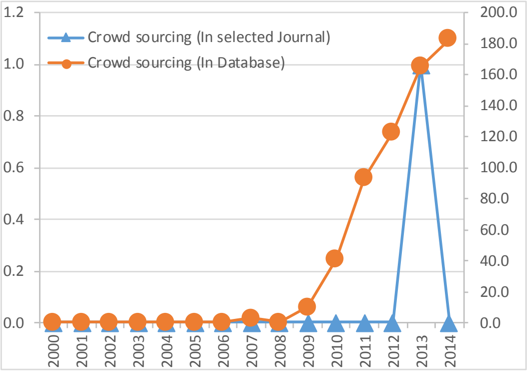
Phase	Documents Type
Pre-Contract Phase	Project feasibility documents
	Request for Qualification and its response Documents
	Request for Proposal and its response Documents
Contractual Agreement	Specifications
	Pre-Construction Agreements
	Other Pre-Construction Documentation
	Contractual Agreements
	Contract Change Orders/Amendments
	Contract Drawings and Revisions
Planning and Design Documents	Structural Design Document
	Mechanical Design Documents
	Electrical Design Documents
	Architect’s Bulletins
	As build Documents
	Project Schedule
Communication Documents	Requests for Information
	Correspondence (external)
	Memoranda (internal)
	E-mails
Progress Reports	Daily Reports (Field Reports)
	Weekly Reports
	Request for monthly progress payments (cost plus fee contract)
	Safety exposure reports
	Meeting minutes,
	Equipment's repair and maintenances data
Accounting and Financing	Accounting Documents
	Financial Documents
Remote Sensor	Site Images
	Site Videos
	Sensor's Data

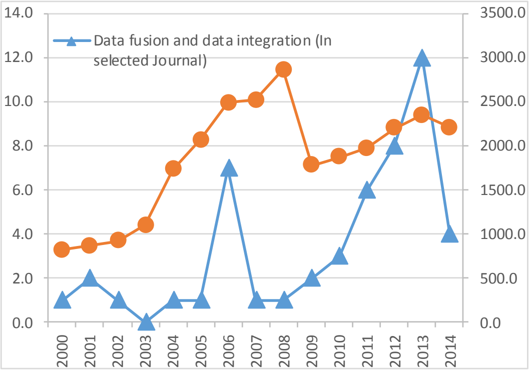
Appendix E: Trend Analysis for Each of 26 Selected Analytics Techniques

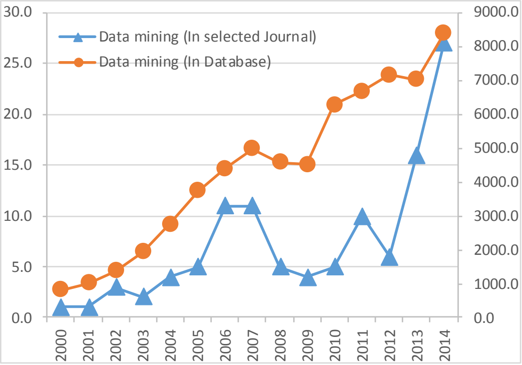
Left vertical axis shows the number of papers in the selected journal. Right vertical axis presents the number of papers in total engineering database. Horizontal axis presents the publication years.

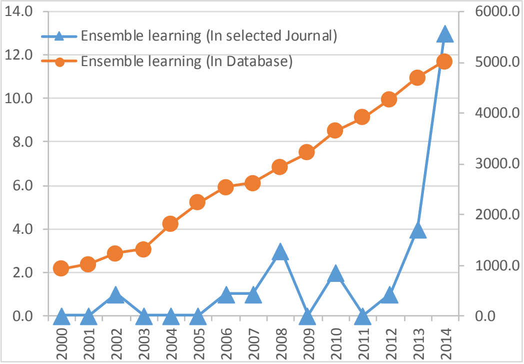


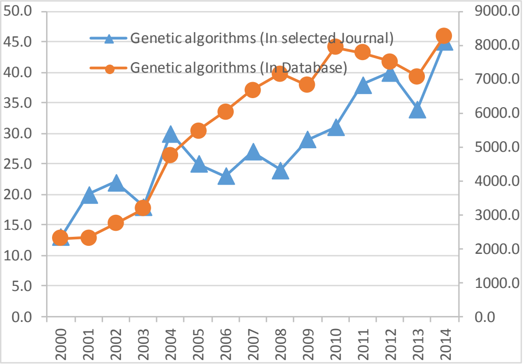


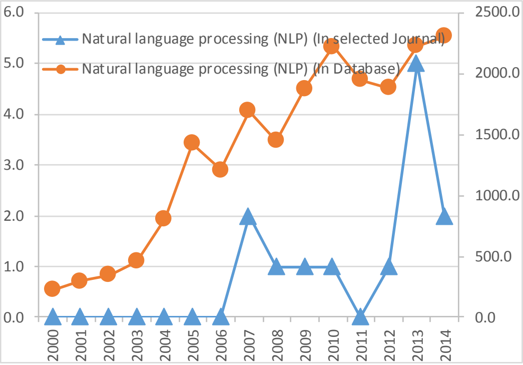


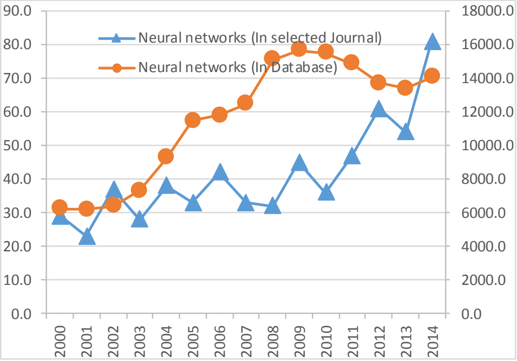


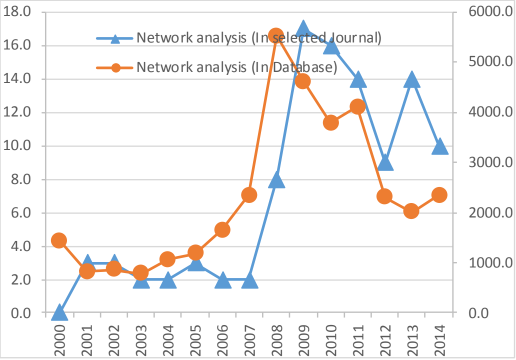


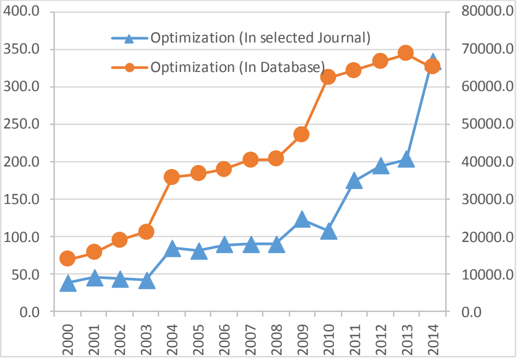


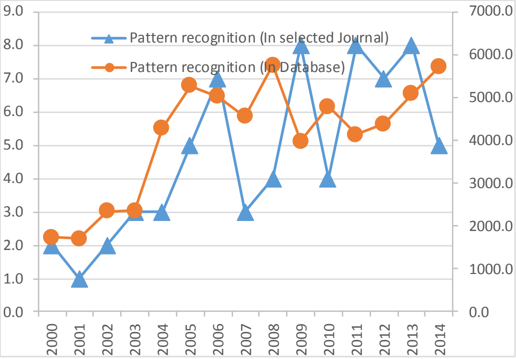

Genetic algorithms (In selected Journal)
Genetic algorithms (In Database)
0.0 5.0 10.0 15.0 20.0 25.0 30.0 35.0 40.0 45.0 50.0
0.0 1000.0 2000.0 3000.0 4000.0 5000.0 6000.0 7000.0 8000.0 9000.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

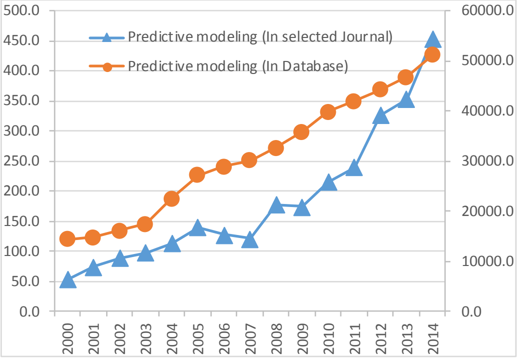

Natural language processing (NLP) (In selected Journal)
Natural language processing (NLP) (In Database)
0.0 1.0 2.0 3.0 4.0 5.0 6.0
0.0 500.0 1000.0 1500.0 2000.0 2500.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

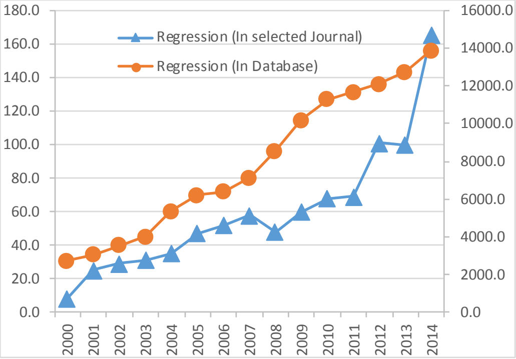

Neural networks (In selected Journal)
Neural networks (In Database)
0.0 10.0 20.0 30.0 40.0 50.0 60.0 70.0 80.0 90.0
0.0 2000.0 4000.0 6000.0 8000.0 10000.0 12000.0 14000.0 16000.0 18000.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

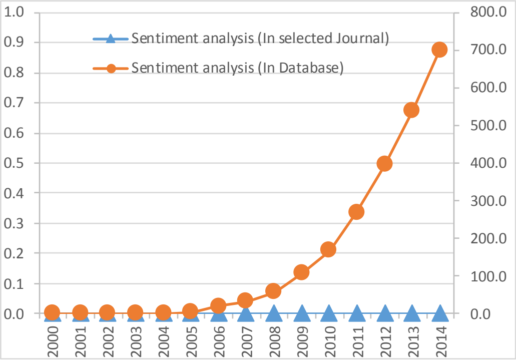

Network analysis (In selected Journal)
Network analysis (In Database)
0.0 2.0 4.0 6.0 8.0 10.0 12.0 14.0 16.0 18.0
0.0 1000.0 2000.0 3000.0 4000.0 5000.0 6000.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

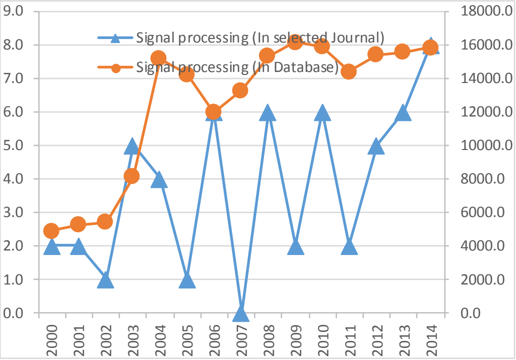

Optimization (In selected Journal)
Optimization (In Database)
0.0 50.0 100.0 150.0 200.0 250.0 300.0 350.0 400.0
0.0 10000.0 20000.0 30000.0 40000.0 50000.0 60000.0 70000.0 80000.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

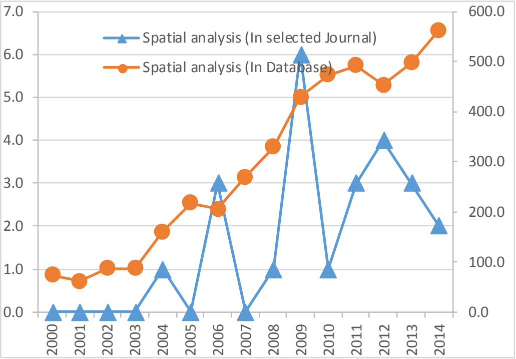

Pattern recognition (In selected Journal)
Pattern recognition (In Database)
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0
0.0 1000.0 2000.0 3000.0 4000.0 5000.0 6000.0 7000.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

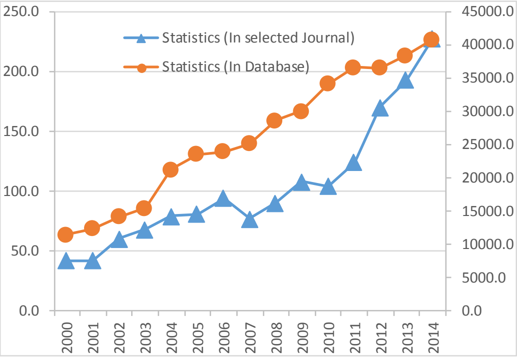

Predictive modeling (In selected Journal)
Predictive modeling (In Database)
0.0 50.0 100.0 150.0 200.0 250.0 300.0 350.0 400.0 450.0 500.0
0.0 10000.0 20000.0 30000.0 40000.0 50000.0 60000.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

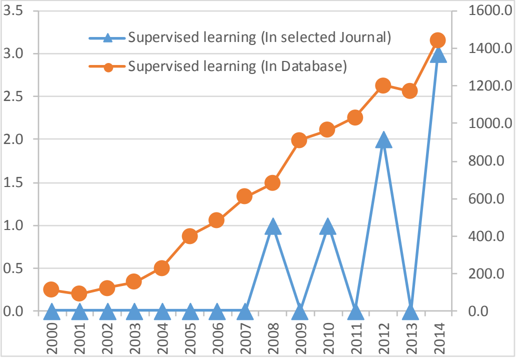

Regression (In selected Journal)
Regression (In Database)
0.0 20.0 40.0 60.0 80.0 100.0 120.0 140.0 160.0 180.0
0.0 2000.0 4000.0 6000.0 8000.0 10000.0 12000.0 14000.0 16000.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

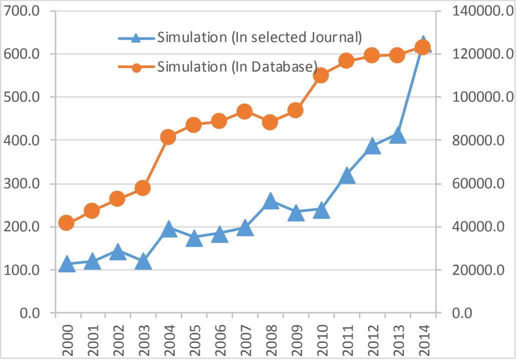

Sentiment analysis (In selected Journal)
Sentiment analysis (In Database)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
800.0
700.0
600.0
500.0
400.0
300.0
200.0
100.0
0.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

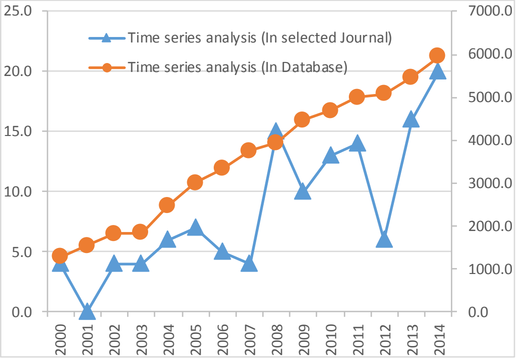

Signal processing (In selected Journal)
Signal processing (In Database)
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
18000.0
16000.0
14000.0
12000.0
10000.0
8000.0
6000.0
4000.0
2000.0
0.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

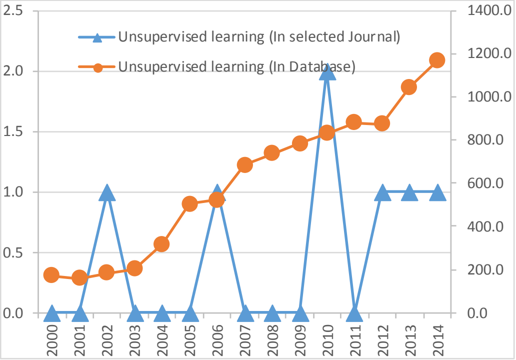

Spatial analysis (In selected Journal)
Spatial analysis (In Database)
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
600.0
500.0
400.0
300.0
200.0
100.0
0.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

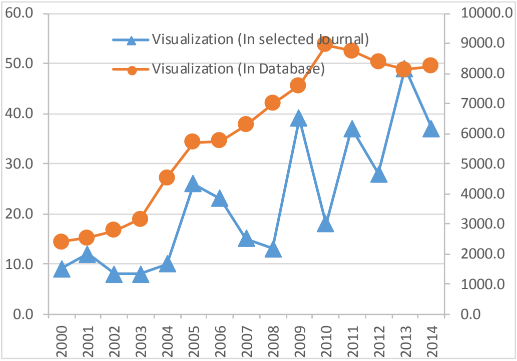

Statistics (In selected Journal)
Statistics (In Database)
250.0
200.0
150.0
100.0
50.0
0.0
45000.0
40000.0
35000.0
30000.0
25000.0
20000.0
15000.0
10000.0
5000.0
0.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014

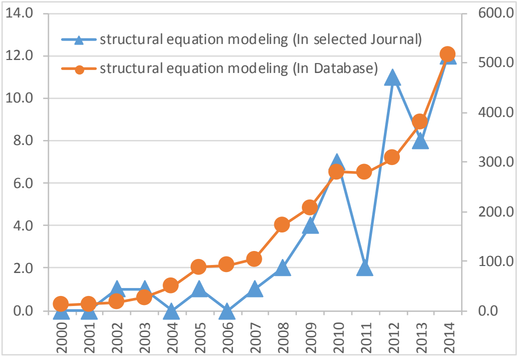

Supervised learning (In selected Journal)
Supervised learning (In Database)
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1600.0
1400.0
1200.0
1000.0
800.0
600.0
400.0
200.0
0.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014


Simulation (In selected Journal)
Simulation (In Database)
700.0
600.0
500.0
400.0
300.0
200.0
100.0
0.0
140000.0
120000.0
100000.0
80000.0
60000.0
40000.0
20000.0
0.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014


Time series analysis (In selected Journal)
Time series analysis (In Database)
25.0
20.0
15.0
10.0
5.0
0.0
7000.0
6000.0
5000.0
4000.0
3000.0
2000.0
1000.0
0.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014


Unsupervised learning (In selected Journal)
Unsupervised learning (In Database)
2.5
2.0
1.5
1.0
0.5
0.0
1400.0
1200.0
1000.0
800.0
600.0
400.0
200.0
0.0
2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014


Visualization (In selected Journal)
Visualization (In Database)
60.0
50.0
40.0
30.0
20.0
10.0
0.0
10000.0
9000.0
8000.0
7000.0
6000.0
5000.0
4000.0
3000.0
2000.0
1000.0
0.0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014


structural equation modeling (In selected Journal)
structural equation modeling (In Database)
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
600.0
500.0
400.0
300.0
200.0
100.0
0.0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014