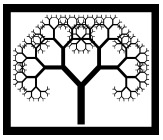
Paper 52



Proceedings of the Second International Conference on
Soft Computing Technology in
Civil, Structural and Environmental Engineering,
Y. Tsompanakis and B.H.V. Topping, (Editors),
Civil-Comp Press, Stirlingshire, Scotland

Data Mining Techniques for the Assessment of Factors Contributing to the Damage of Residential Houses in Australia

N.Y. Osman-Schlegel[1], Z.A. Krezel[1] and K.J. McManus[2]
[1] School of Architecture and Building
Deakin University, Victoria, Australia
[2] Faculty of Engineering and Industrial Sciences
Swinburne University of Technology, Victoria, Australia

Abstract

This paper reports on the preparation and management processes of inconsistent data on damage on residential houses in Victoria, Australia. There are no existing specific and fully relevant databases readily available except for the incomplete paper-based and electronic-based reports. Therefore, the extracting of information from the reports is complicated and time consuming in order to extract and include all the necessary information needed for analysis of damage on residential houses founded on expansive soils. Data mining is adopted to develop a database. Statistical methods and Artificial Intelligence methods are used to quantify the quality of data. The paper concludes that the development of such database could enable BHC to evaluate the usefulness of the reports prepared on the reported damage properties for further analysis.

Keywords: data mining, chi-square test, categorical regression, artificial intelligence, databases.

1 Introduction

Data mining is the process of discovering information which is not obvious in a large collection of data. It enables solving problems by analysing data already present in database [1]. Generally, ill-defined and low quality data leads to a low quality of data mining results. Data mining begins with examining a massive quantity of data which is then selected and pre-processed in a database known as Data Warehouse. It is then transformed into a smaller database known as Data Mart which focuses on the main subject of the analysis. The next step of data mining is using data mining algorithm; statistical methods and Artificial Intelligence methods. These methods are used to quantify the quality of data and the preferred factors

This paper reports on the preparation and management process of inconsistent data on damage on residential houses in Victoria, Australia obtained from the

Building Housing Commission (BHC) of Victoria, Australia. The BHC owns and manages over 73,000 properties across Victoria with an annual budget of $1Million for maintenance and repairs of over 200 structural damages reported annually [2]. The data used in this paper is based on information extracted from 600 reports provided by the BHC [3]. The extraction and addition of information from the BHC reports is extremely complicated with many shortcomings noticeable. Therefore, the management of extracting the information from the reports had to be done thoroughly in order to include the necessary information needed for the development of a BHC's database. The selections of the relevant information in the reports are based on the studies and investigations done by other researchers on ground movement of expansive soils which is the main course of damage on residential houses [3]. However, not all information needed for such analysis is available in the BHC reports. There is no existing specific and fully relevant database readily available from BHC, except for the incomplete paper-based and electronic-based reports. The aim is to develop a database that can enable BHC in judging the quality of the reporting on their properties. It can also assist them in re-evaluating the information from the reporting firms. This could also enable the BHC to identify the need for additional information and the relevance of the data use in the analysis of damage to residential houses.

The data in the Data Mart can be used to undertake different analysis such as, ranking of the importance of the factors causing particular types of damage, predicting the future development of damage, generating detailed reports with substantial filtering options and many more. These analyses can assist in the asset management of the housing stock that needs maintenance, rehabilitation, demolition or reconstruction. Significant financial and other resources could optimise the decision-making process, as the database will not only be more efficient and easy to use but also will be readily available for any kind of data analysis using variety of programs or software.

2 Data Preparation

Real world data tend to be "dirty", incomplete and inconsistent. Detecting data anomalies, rectifying them, and reducing the data to be analysed can lead to better decision making [1]. Therefore, data preparation is vital before the analysis to improve the quality of the data thus helping to improve the accuracy and efficiency of the required outcome. The first step towards the development of a database is to identify and define the objectives of the analysis. The objective of the analysis reported in this paper is to predict structural damage conditions. The specific focus of the paper is on the preparation and management process of BHC data to develop a comprehensive and consistent database.

The selection of data from the report is crucial for the analysis of the database. It provides the fundamental input for the subsequent data analysis [4]. Two databases (Data Warehouse and Data Mart) are used to assist in the selection process of the significant data from the BHC reports and to quantify the quality of data.

2.1 Data Warehouse

A Data Warehouse is a storage of data that has been extracted from operational data [5]. The key concept of a Data Warehouse is to create a critical volume of information available that can be used for further analytical processing and decision making. The Data Warehouse contains very large data sets since the information in it is subject-oriented, non-volatile, and of an historic nature [6]. The available raw data, from approximately 600 reports from BHC are transformed into a common data format. This is done by identifying and extracting common information plus other information that might be of importance from all the available reports from BHC. This included information such as climate and geological condition that might be useful for the analysis. Since the reports are not uniform, it is useful to include the most common and useful information extracted from the reports in the database even though not all the information will be used for the development or the analysis. This data from different sources is then integrated into a central database. The Data Warehouse stores all information that can be extracted from the reports.

2.2 Data Mart

Data Mart is a subset of Data Warehouse and focused on a particular subject. Data extracted from the Data Warehouse often results additional transformation to produce a uniform and standard database for preliminary analysis. Only relevant categories from the Data Warehouse are included in the Data Mart. The categories for the analysis are chosen based on studies of relevant and related work performed on expansive soils [3]. New factors had to be included since the analysis is dealing with damage to residential houses with regards to soil moisture; the factors influencing the damage had to be considered. This included climate (Thornthwaite Moisture Index), structural system (Wall Construction), foundation system, age of residential houses, soil characteristics (Geology), vegetation, site leakages and pre and post construction. These factors proved to be the most common damage factor potential based on the studies and investigations done by other researchers on ground movement of expansive soils which is the main course of damage on residential houses [3] as shown in Table 1 are used in the Data Mart.

Category	Factors
Property Information	Geographical region (R)
Building Information	Construction Footing (CF) Construction Wall (CW) Age of residential houses (Age)
Site Information	Climate (TMI) Geology (G) Vegetation (V)
Consultant's diagnosis	Damage Classification (DC)

Table 1: Categories and Factors for the Data Mart

2.3 Data Pre-Processing

Real world databases are highly susceptible to problems such as noisy, missing and inconsistent data [1]. In this paper, the issues of the BHC reports are that the reports are (i) not scientific and non-uniform, and, (ii) not consistent where some of the reports had missing attributes. The BHC reports are recorded by different engineering companies based mainly on the tenant's complaints concerning damage and subsequent site investigations of the residential houses. This resulted in a series of disparate reports with dissimilar information. The damage as inspected by internal BHC inspectors who reported on the deterioration in terms of which building trades will be involved in the repair works and the extent of their involvement. A thorough diagnosis of the damage and geological site investigation was then conducted by consulting engineers using their own paper-based report templates. Therefore, it is expected that the information in the reports is different and some lack important information about structural systems, footings and foundations, climate, soil classification or geological site, vegetation, leakages and many more. This makes it difficult to analyse the data because of many factors and different level of accuracy.

2.3.1 Missing Data

Many practical issues result from unreliable data. Some factors measured by humans could be subjected to an observational error. They might be corrupted by noise, or values might be missing altogether. Some of the reports obtained from BHC contained missing or inaccurate data. Approximately one third of the BHC data are missing demonstrated by incomplete lines of data in the reports which in such format cannot be incorporated in the database. For instance, some reports did not take into account the type of wall construction or the age of the building. Since there are only about 600 lines of data available in the database, it is a crucial decision to delete or omit the entire lines as this reduces the amount of data available for analysis. This influences the accuracy of the results from the analysis if insufficient volume of data is used.

Some data mining algorithms have a minimum and maximum number of data required for the analysis in order to get accurate and unbiased results. However, more data do not guarantee more accurate or better results. The approach to dealing with missing values or incomplete data is to go through a "cleansing" process; either to drop them from the analysis or substitute typical values for them [7, 8]. The lines in which there are missing data might be useful for the prediction of damage. Therefore, it might be useful to substitute the missing values instead of omitting the entire lines which consist of some missing values or parameters in the data set. However, most programs for analysis do not manage missing values very well as a missing value cannot be multiplied or compared to other values [5]. There should be a rule of thumb when dealing with missing values as it is important to judge whether the complete set of data with missing data is relevant or not in the analysis. Adriaans and Zantinge [9] state that a general rule for any deletion of data must be a conscious decision, after a thorough analysis of a possible consequences.

There are two ways to deal with missing data which are to omit or to replace it with a default value. It is best to generate multiple solutions with and without missing values according to Weiss and Indurkhya [5] which is adopted in the paper to avoid any bias or misleading results. For the replacement of missing values, the numeric value "0" is used because such replacement has minimal affect the outcome of the analysis. The Data Marts with (600 lines of data) and without (350 lines of data) missing data are called "original" and "complete" respectively. The Data Mart which gives higher performance value is then chosen and used for the subsequent analysis.

2.3.2 Data Coding

Unified information system for BHC is recommended since it has the potential of assisting in the organisation of data into ordered and high-quality database. Qualitative and quantitative variables are developed using selected information from the reports. Qualitative variables are classified into levels, sometimes known as categories while quantitative variables are linked to intrinsically numerical quantities [4]. The selected factors in the Data Mart are further refined to suit the input format requirements of a particular data mining process.

Since most of the information in the Data Mart is in qualitative (text) form it is essential to transform it into numerical quantities. The text form is not preferred as it can be inconsistent with a specific standard form and create significant difficulties for analysis. Also, most of the factors in the categories have different formats such as text, numbers and units. It is recommended to code the factors accordingly into a uniform format. This is to ensure that there is no bias in the outcome resulting from different range of values in the factors.

Normalisation is chosen for data coding. Normalisation is the process of efficiently organising data in a database where it eliminates redundant data and ensures data dependencies make sense [10]. The first step is to set the quantitative and qualitative variables in the Data Mart in numeric format. This transformation applies to the sub-variables from each category. The numerical setting is then coded into [1,0] interval so that all the values fall between the range between 0 and 1. This is to ensure uniformity in the values of all the categories. Since the values are between 0 and 1, the discrepancy or bias towards larger values can be eliminated. Consequently, it is to avoid bias in learning errors to weights of numerically large valued parameters. Weiss and Indurkhya [5] indicate that experience has shown that normalised numbers scaled similar to the output values, lead to a better training.

A simple mathematical formula is used to code the numeric variables for all sub-variables in each category as in equation (1) where In is the number of indicator and Imax is the maximum value of numeric indicator. Table 2 shows a sample of data coding for one of the factors in the Data Mart.

$$\text{Code} = I_n / I_{max} \quad (1)$$

Sub Variable (Year)	Indicator	Code
Missing Data	0	0.00
1 to 10	1	0.17
11 to 20	2	0.33
21 to 30	3	0.50
31 to 40	4	0.67
41 to 50	5	0.83
>50	6	1.00

Table 2 : Data Coding for Age

3 Data Mining Algorithm

To enhance the quality of the Data Mart, it is essential to quantify whether the selection of factors or the input parameters made are precise and to determine the performance value of the "original" and "complete" Data Marts. Statistical analysis and Artificial Intelligence are adopted.

3.1 Statistical Methods

A **categorical regression (CATREG)** using the commercial software package, SPSS version 14 for Windows is selected to check the significance of the factors in Table 1. CATREG is chosen because it can handle nominal independent variables and is used to find the best-fitting model. A CATREG is used to predict a dependent variable (the Damage Conditions) from a set of independent variables (the selected parameters) [11]. Table 3 shows the significance of the factors chosen using categorical regression analysis. The chosen factors except for Age are significant where the significant values are less than 0.05 ($P<0.05$). This indicated that the factors are correctly chosen for the purpose of predicting the damage conditions (without considering other possible factors). The Age can be eliminated. However, it will be still used together with the other factors in this stage.

Factors	Standardised Coefficients		dF	F	Sig.
	Beta	*Std. Error*			
Geographical Region (R)	-0.125	0.045	2	7.848	0.000
Construction Wall (CW)	0.133	0.042	2	9.938	0.000
Construction Footing (CF)	0.116	0.042	3	7.570	0.000
Age of residential houses (Age)	-0.039	0.040	2	0.935	0.393
Geology (G)	0.107	0.048	4	5.055	0.001
Climate (TMI)	0.202	0.064	2	10.034	0.000
Vegetation (V)	-0.091	0.045	2	4.017	0.019

Table 3 : Coefficients in Regression of Categorical Data

Another statistical method; **Pseudo R-Square test** is applied to identify how the model improves prediction capability. It takes values between 0 and 1, becomes larger as the model "fits better", and provides a simple and clear interpretation of the data [12]. In terms of the capability of predicting damage conditions, the "complete" Data Mart is more capable as shown in Table 4. This can be seen from the higher values of Cox and Snell, Nagelkerke and McFadden Pseudo R-Square values. In McFadden Pseudo R-Square, the value for "complete" Data Mart is nearly doubled. This indicates that the "complete" Data Mart is capable of predicting damage condition accurately compared to the "original" Data Mart. The values in Cox and Snell; and Nagelkerke also show the value for "complete" Data Mart is much higher compared to the "original" Data Mart. Thus it can be concluded that "complete" Data Mart is the better choice in maximising the capability of predicting damage conditions to residential houses on expansive soils.

Data Mart	Cox and Snell	Nagelkerke	McFadden
"original"	0.123	0.131	0.047
"complete"	0.177	0.195	0.082

Table 4 : Pseudo R-Square Values

3.2 Artificial Intelligence Technique

To confirm the accuracy of the performance of the Data Marts using statistical methods, a hybrid technique using Neural Network and Genetic Algorithm is

chosen. A Neural Network is a computing paradigm inspired by the human brain. It consists of an interconnected group of simple processing elements, called neurons that work together to generate an output function. The goal of the network is to learn some association between the input and output patterns, or to analyse, or to find the structure of the input patterns [13]. Figure 1 shows the comparison between biological and artificial neurons.

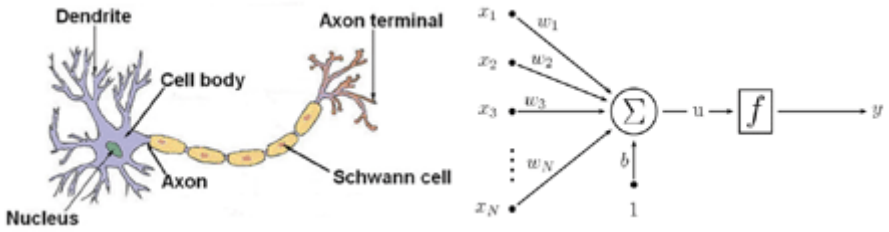


Figure 1: A Comparison of a Biological and an Artificial Neuron ; A Real Neuron (left) and an Artificial Neuron (right) [14]

Genetic Algorithm which is based on a Darwinian-type survival of the fittest strategy [15] easily handle functions that are highly non-linear, complex, and noisy; in such cases the traditional gradient-based methods are found to be inefficient. It can also perform global search as against the local one performed by the gradient-based methods. Thus, Genetic Algorithm is most likely to arrive at the global optimum of the objective function.

Neural Network and Genetic Algorithm have been established as two major research and application areas in Artificial Intelligence. Neural Network is of particular interest because of its robustness, parallelism, and its abilities [5]. Genetic Algorithm on the other hand is robust and can deal with a wide range of problem areas including those which are difficult for other methods to solve as a global search method [16]. The use of hybrid technique is the most promising approach for the task of analysing the data dependencies and function approximation in order to provide a reliable outcome. It is originally motivated by the astonishing success of these concepts in their biological counterparts. Despite their totally different approaches, both can merely be seen as optimisation methods which are used in a wide range of applications, where traditional methods often proved to be unsatisfactory [17].

This paper adopts the use of Genetic Algorithms for the training of Neural Networks as shown in Figure 2. The general intention is to gain an optimal parameter set which improves the learning performance of conventional neural learning process which is Backpropagation. It is proven [18] that in almost all cases, the evolved network (Genetic Algorithm and Neural Network) shows a significantly improved learning behaviour compared with using Neural Network alone. It has also been shown [19] that the hybrid technique is by far superior, both in terms of development time and performance.

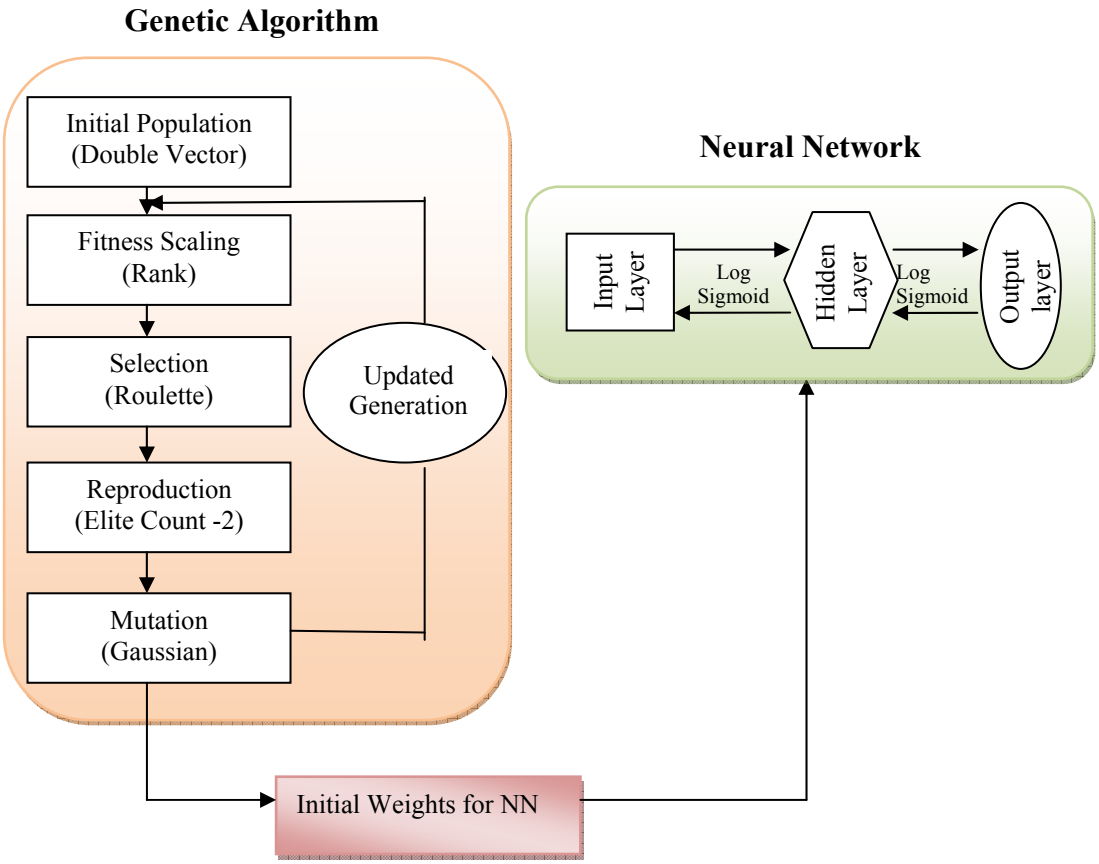


Figure 2: Hybrid Artificial Intelligence (GANN)

Figure 3 shows the performance of Neural Network compared to the performance of Hybrid Artificial Intelligence technique; Neural Network trained by Genetic Algorithm (GANN). Analysis of the selection of the options and variables for both algorithms are needed to determine the best performing network. However, it is beyond the scope of this paper. A detailed analysis can be seen in the first author's PhD thesis [3]. The goal is to have a mean squared error (performance) equals to zero. However, this is impossible unless the data used is perfect without "noise". It is shown in Figure 3 that the performance of the Data Mart using hybrid technique performs better than the Data Mart which only uses Neural Network where the mean squared error for Data Mart using hybrid technique has lower value. It is also shown that the "complete" Data Mart perform better compared to the "original" Data Mart. "Complete" Data Mart performed 30% better than the "original" Data Mart. "complete" Data Mart shows the best results by producing the smallest value of mean square error, which implied that "complete" Data Mart is the most reliable Data Mart. Even though the Neural Network approach is capable of handling data with missing values, it is essential to use "complete" Data Mart as it gives better performed network thus more accurate results. This also coincides with findings using statistical methods mentioned earlier in this paper.

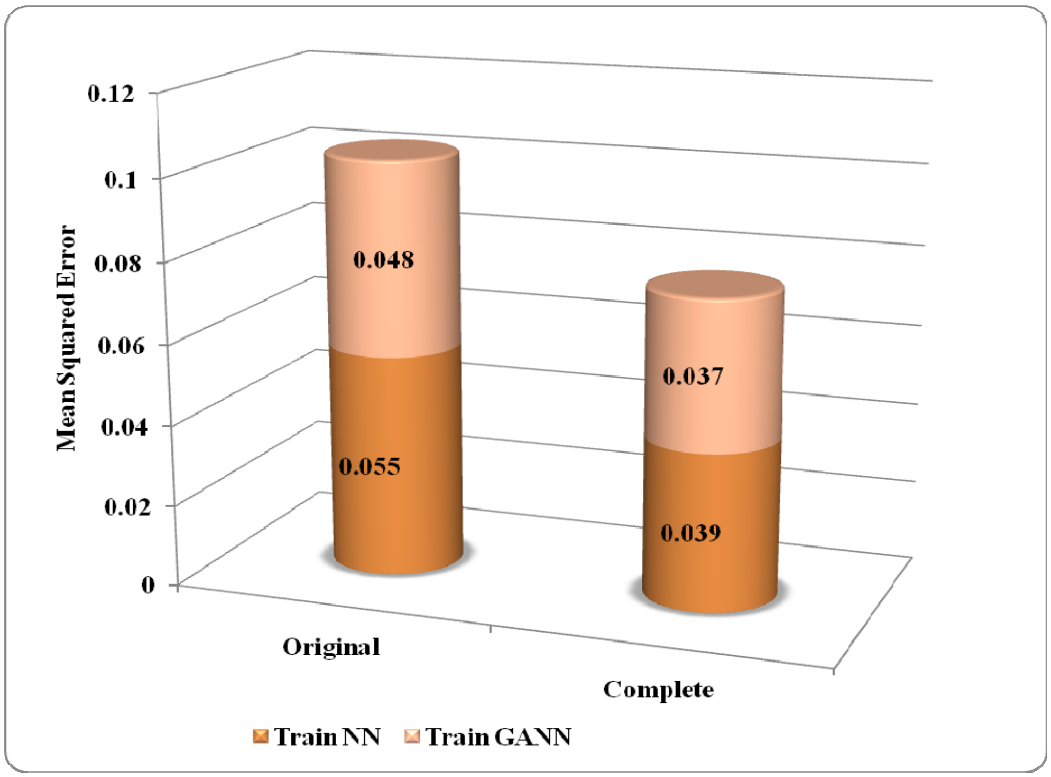


Figure 3: Performance of Neural Network (NN) vs. Neural Network Trained with Genetic Algorithm (GANN)

4 Conclusions

Data mining is vital in order to discover information which is not obvious in a large collection of data and solving problems by analysing data already present in database. The solution for an accurate and reliable analysis is to develop a consistent data collection method for engineering firms in a form of a uniform database to ease analysis [20]. This paper describes the process of data mining to develop a uniform database for BHC. Two database; Data Warehouse and Data Mart were developed. The development of a Data Warehouse hence the Data Mart enable BHC to evaluate the usefulness of the reports prepared on the reported damage to residential houses. It can also assist BHC in re-evaluating the information given by the engineering firms. This can enable BHC to distinguish between any additional or relevant and non-relevant information needed in analysing damage to residential houses on expansive soils in Victoria. The Data Mart can be used to undertake different analysis such as analysing the important factors causing particular types of damage, predicting development of damage in the future and generating detailed reports with substantial filtering options. This analysis can assist in the asset management of the housing stock that needs maintenance, reconstruction or demolition. Financial and other resources can be saved as the database will not only be easy to use but also be readily available for variety of data analysis using various programs or software.

References

[1] S. Chakrabarti, *Data mining : know it all*. Burlington, MA ; London: Elsevier/Morgan Kaufmann Publishers, 2009.

[2] DepartmentofHumanServices;, "Sustaining our housing : Asset management strategy 2004-2009," Department of Human Services, Melbourne, Australia.2004.

[3] N. Y. Osman, "The development of a predictive damage condition model of light structures on expansive soils using Hybrid Artificial Intelligence techniques," PhD, Faculty of Engineering and Industrial Sciences, Swinburne University of technology, Hawthorn, 2007.

[4] P. Giudici, *Applied Data Mining : Statistical Methods for Business and Industry*. England: John Wiley & Sons, 2003.

[5] S. M. Weiss and N. Indurkhya, *Predictive Data Mining - A Practical guide*. San Fransisco, USA: Morgan Kaufmann Publishers, Inc., 1998.

[6] P. Adriaans and D. Zantinge, *Data Mining,*. Harlow, UK: Addison Wesley Longman Ltd, 1996.

[7] J. P. Bigus, *Data Mining with Neural Networks - Solving Business Problems from Application Development to Decision Support*. New York, USA: McGraw-Hill, 1996.

[8] D. Tamraparni, *Exploratory data mining and data cleaning*. USA: John Wiley & Sons, 2003.

[9] P. Adriaans and D. Zantinge, *Data Mining*. Harlow, UK: Addison Wesley Longman Ltd., 1996.

[10] M. Nardo, M. Saisana, A. Saltelli, and S. Tarantola, "Tools for Composite Indi-cators Building," ed, 2005.

[11] J. J. Meulman and W. J. Heiser, *SPSS Categories 13.0*. Chicago: SPSS Inc, 2004.

[12] E. S. Shtatland, S. Moore, and M. B. Barton, "Why we need an R2 measures of fit (and not only one) in Proc Logistic and Proc Genmod," presented at the Proceedings of the Twenty-Fifth Annual SAS® Users Group International Conference, Indianapolis, Indiana 2000.

[13] H. Abdi, "Neural Networks,' Encyclopaedia of Social Sciences Research Methods " *Quantitative Applications in the Social Sciences,* vol. 124, 2003.

[14] S. Poles;. (9 May 2011). *Meta-modeling with modeFRONTIER: Advantages and Perspectives*. Available: http://www.enginsoft.com/software/modefrontier/documentation/metamodelling.html

[15] J. H. Holland, "Genetic Algorithms : Computer programs that "evolve" in ways that resemble natural selection can solve complex problems even their creators do not fully understand," *Scientific American,* vol. July,1992, pp. 44-55, 1992.

[16] D. Beasley, D. R. Bull, and R. R. Martin, "An Overview of Genetic Algorithms: Part 1, Fundamentals," *University Computing,* vol. 15, pp. 58-69, 1993.

[17] B. Omer, *Genetic Algorithms for Neural Network Training on Transputers*. University of Newcastle upon Tyne: Department of Computing Science, 1995.
[18] G. Weiß, "Neural networks and evolutionary computation. Part I: Hybrid approaches in artificial intelligence," presented at the IEEE International Conference on Evolutionary Computation Nagoya University, Japan, 1994.
[19] D. Curran and C. O'Riordan, "Applying Evolutionary Computation to Designing Neural Networks: A Study of a State of the Art," 2002.
[20] N. Y. Osman, K. J. McManus, and A. W. M. Ng, "Management And Analysis Of Data For Damage Of Light Structures On Expansive Soils In Victoria, Australia," in *Proceedings of the 1st International conference on Structural condition assessment, monitoring and improvement*, Perth, Australia, 2005, pp. 283-290.