14 KNOWLEDGE MODELING OF 'SOFT' DATA IN ARCHITECURAL DESIGN

Sanja Durmisevic, Özer Ciftcioglu and Sevil Sariyildiz
TU Delft, Faculty of Architecture, Technical Design and Informatics
s.durmisevic@bk.tudelft.nl

Abstract

Information technologies are at present used in various disciplines to address issues such as information processing, data mining, knowledge-modeling etc. Its final goal is to provide necessary aid to professionals during decision-making process. This raises already few questions such as, what type of data is considered and are there some new emerging technologies that can improve knowledge modeling and therefore provide better decision support to the professionals? Design professionals are very often confronted with soft data which they somehow need to interpret and finally integrate into design. The architectural design task is one example having linguistic qualities as priory design information. This is especially the case when qualities of certain space are discussed, like for example in post occupancy evaluation of the buildings, where the relationship between spatial characteristics and psychological aspects plays an important role. Expressions such as: bright colour, light room, large space are some of these examples and therefore a special method is needed for representation and processing of such vague expressions and concepts. Referring to the complexity of task in dealing with the soft data as well as dealing with soft computing, the paper first identifies the source of these complexities referring to the architectural design tasks. Following this, data collection and soft computing analysis method based on one case study will be presented, whereby the focus will be on knowledge modeling. Finally, the results of the sensitivity analyses together with the conclusions regarding the observed effectiveness of the approach are presented.

Keywords: *soft data, knowledge modeling, sensitivity analysis*

INTRODUCTION

Architecture is a science where most things can be learned from previous examples and from the successes and failures of previous designs, applying the already known or inventing some new techniques. In other words, design is evolving according to requirements of specific time, available technology, existing knowledge and personal ability of a designer to combine all these features into a new, evolved design. This implies that the designer should be able to derive from each previous design some qualitative values, especially on customer approval regarding building's quality as to assure a successful design. It is interesting to note that on a far larger scale, the success of any customer oriented organization, greatly depends on this particular aspect as well (*figure 1*), meaning that the feedback from a customer is quite important for innovation and learning process, both for organization and individual (EFQM, 1999)..

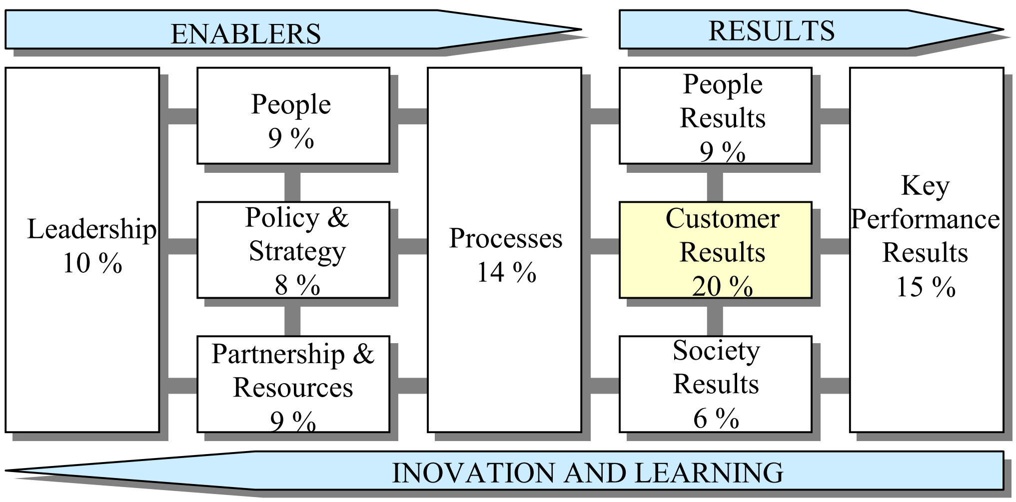


Figure 1: ® The EFQM Excellence Model (EFQM, 1999)

Very simplified, it could be said that a building is a delivered product to a customer who is a final arbiter of a product in sense of building performance and in terms of its functionality and customers comfort while in that building. A customer may be a short-time visitor, dweller or an employee. Then, to deliver the most suitable product, a designer should know much about customer's needs and requirements and especially customer preferences when it comes to specific design issues. Knowing more about such issues would create a base that would lead in first place to systematic learning and thereafter to innovation and in such way improve the overall performance of a designer or an architectural office in general. Such information on buildings is up till now rather poorly recorded or if recorded it is not a general public domain knowledge, meaning that not each designer could acquire it at any time.

From the figure above, we see that the main indicator of an organization's performance depends on customer/user results (20 %). At this point, it is rather interesting to look at the results of Hamel's research who developed a descriptive psychological model of the architectural design process, which shows how architects actually design (Hamel, 1990). He classified five main tasks embraced in architectural design being *gathering information*, *decomposing problems*, *solving partial problems*, *integrating partial solutions* and *shaping the result into a design*. During design process 66.5% of a design activity is related to more technical aspects of design while at the same time rather few attention is paid to the use of a building and the users (7.3%). Even more interesting are the results on type of sources that architects 'consult' in order to extract information related to user issues. When it comes to user oriented aspects, architects relied in first instance on their own knowledge and personal estimations. The information about the use of a building and the users is almost never supplemented with other sources such as literature, communication with the client and communication with the expert. It is still not clear why is that so, and the author himself (Hamel, 1990) gave some possible interpretations. It can be that the architects have confidence in their own knowledge. They can also think that this knowledge is not available on the scientific level. Or it is possible that the knowledge is available but not in a form that is for them accessible. In most cases, the post-occupancy evaluation of previously designed buildings is hardly ever available and yet customer and perhaps even the society approval/disapproval is recorded in such evaluation. The results from post-occupancy evaluation would in time form an extensive knowledge base out of which knowledge could be elicited for future projects. The conclusion is, that one way or another, there is no systematic approach to solving problems when it comes to user/customer issues and there are no evident sources to be consulted in order to obtain the necessary information. Furthermore, most of the design decisions that influence the future of a project are made during the conceptual stage of a design process (Rafiq, 2000). A rather significant amount of the total project cost is committed in this stage and is estimated to be between 70 and 80 percent (Evboumwan and Anumba, 1996). This would mean, that all aspects that are relevant for design should receive fair attention at this stage, especially those related to users, since users are final judges of space and together with technical aspects they determine the success/failure of a design.

In this paper, we show the way in which 'soft' data, such as data obtained from the questionnaire, can be processed by means of soft computing techniques in order to design a firm knowledge model for specific building. It is important to mention the necessity of conducting few experiments before creating a final knowledge model, since there can be some discrepancy in data at hand. Therefore, data should not be blindly processed but carefully considered. These issues are the main topics of this paper. Firstly, we explain the data gathering and questionnaire design. Secondly, the fuzzy logic and radial basis function network (RBFN) for knowledge modeling are briefly explained. Thereafter, the importance of data consideration is explained through experiments which were necessary in order to form a solid knowledge model by using the data at hand.

DATA COLLECTION

The questionnaire is not simply the gathering of facts but more importantly it is an instrument for gathering meaningful data to test the hypotheses. Development of a hypothesis is a starting point of any questionnaire. In other words, the hypotheses are consciously developed and tested through the questionnaire by means of words, questions and specific layout/format. The structure, purpose and the meaning of the researched topic is provided through hypotheses. Having firm hypotheses in mind they actually later on serve to form the questions and to omit the unnecessary questions that are not related to the hypothesis (Labaw, 1980).

In general, the way in which data is measured is called a level of measurement or the scale of measurement for variables. There are four kinds of scales in which variables can be measured: the *nominal* scale, the *ordinal* scale, the *interval* scale and the *ratio* scale (Dalen, Leede, 2000). In this research we have used an *ordinal scale measurement*, where such measurement involves placement of values in a rank order, in order to create an ordinal scale variable. The relationship between observations takes on a form of 'greater than 'and 'less than'. For this questionnaire a five-point scale was used. In such way, respondents have the opportunity to 'strongly agree' or 'strongly disagree' with a question or to strongly express their opinion regarding design issues.

For a case study, Blaak underground station in the Netherlands was chosen for which the questionnaire was designed. This is an important exchange station, which is situated in the center of Rotterdam. It is at the same time a tram, metro and a train station. Tram station is situated on a ground level. Metro platforms are one level below ground (at approximately -7m) and train platforms are two levels below ground (at approximately -14 meters). From 27th May till 30th May 2000, one thousand of questionnaires were handed out to the passengers visiting the station.

The questionnaire covered aspects that are related to *safety* and *comfort* at the station. In total there were 43 aspects in input space each having five possible options and two aspects in the output space again having five possible options. The latter two are design variables being *safety* and *comfort*. The input aspects, which are identified to be related to comfort are given in *table 1* and those related to safety are presented in *table 2* (Ciftcioglu, Durmisevic, et al., 2001). Main purpose of the questionnaire was to provide information on user's perception regarding specific spatial characteristics of that station. The questions covered all aspects given in *table 1* and *table 2* and additional two final questions were related to user's perception of public safety and comfort at Blaak station.

Attractiveness	**Wayfinding**	**Daylight**	**Physiological**
Colour	To the station	Pleasantness	Noise
Material	In station	Orientation	Temperature winter
Spatial proportions	Placement of signs		Temperature summer
Furniture	Number of signs		Draft entrance
Maintenance			Draft platforms
Spaciousness entrance			Draft exchange areas
Spaciousness train platform			Ventilation entrance
Spaciousness metro platform			Ventilation platforms
Platform length			
Platform width			
Platform height			
Pleasantness entrance			
Pleasantness train platform			
Pleasantness metro platform			

Table2: Aspects related to comfort (28 aspects)

Overview	Escape	Lighting	Presence of people	Safety surrounding
Entrance	Possibilities	Entrance	Public control	Safety in surrounding
Train platform	Distances	Train platform	Few people daytime	
Metro platform		Metro platform	Few people night	
Exchange area		Exchange area		
		Dark areas		

Table1: Aspects related to safety (15 aspects)

For the analysis, the linguistic information is firstly converted to terms in fuzzy logic domain and after appropriate treatment, the data analyses are carried out and the results are expressed in most comprehensible form for design assessments. Such conversions are referred to as fuzzification and defuzzification, where the data are expressed in numerical form and therefore become convenient for mathematical treatment. This is explained in the following section.

FUZZY LOGIC IN BRIEF

Fuzzy set theory was introduced through Zadeh (Zadeh, 1965). With fuzzy sets, a numerical value is classified into one or more linguistic labels. These labels may be discrete as well as continuous and they are coined as membership functions that represent the numerical strength of linguistic labels for the domain of classification. Since the membership functions can overlap, this results in multi-value representation of the knowledge. An input value intersects with one or more membership functions of the input classification and therefore it is attached to several linguistic labels.

A fuzzy set A on the universe X is a set defined by a membership function μ_A representing a mapping

$\mu_A : X \rightarrow \{0,1\}$

where the value $\mu_A(x)$ for the fuzzy set A is called the membership value of $x \in X$. The membership value can be interpreted as the degree of x belonging to the fuzzy set A. A typical example of membership function is shown in *figure 2*.

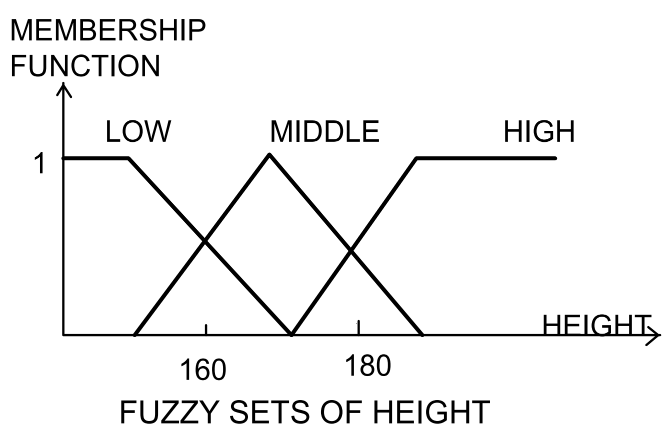


Figure 2: Fuzzy sets of heights

Before entering a fuzzy system, the information at hand are fuzzified. This is done by an input classification, matching the input value against a chosen set of linguistic labels. These labels partly overlap as shown in *figure 2*, so that a numerical value can be classified into more than one label, each with an associated membership value. As it has been earlier mentioned, a five-point scale measurement was used, meaning that there were five fuzzy sets for each question which is actually related to a specific design issue. In *figure 2*, for simplicity reasons, a three-point scale measurement, expressed through three fuzzy sets, is provided.

Inference is carried out with evaluating fuzzy production rules where the propagation of the fuzziness is linear with respect to arithmetic operations. Logical combinations are performed in a systematic way with certain rules known as norms. Since one linguistic value can be attached to several numerical values in the context it is considered, more than one rule might be triggered producing several answers. In general, this multiple answer can be combined to reach an optimal decision or a decision region.

KNOWLEDGE MODELING BY RADIAL BASIS FUNCTION NETWORK

In a complex information environment, to establish the complete fuzzy rules dealing with the knowledge base is a formidable task. To alleviate the problem, the knowledge base can be formed in a distributed and structured form by means of learning so that the structure so formed represents the fuzzy expert system altogether with the consistent rules in any complexity. Here the main task for consistent rules is to carry them out by means of learning process, which should be specially designed for this purpose. The accomplishment of this structure can be achieved by means of a network operating with fuzzy computational units. Such structure can be a radial basis function network (RBFN). The general characteristics of such network are rather diverse and comprehensive with sound mathematical foundations. On one side they can be considered as multivariable multifunctional approximators using basis functions (Broomhead and Lowe, 1988) with functional interpolation and extrapolation capabilities. On the other side they are equivalent to fuzzy logic systems under some lenient conditions (Roger, 1990; Hunt, 1998). Due to these properties, conversely fuzzy logic systems can be used as universal approximators (Kosko, 1994; Wang, 1994). The general structure of a RBFN is shown in *figure 3*.

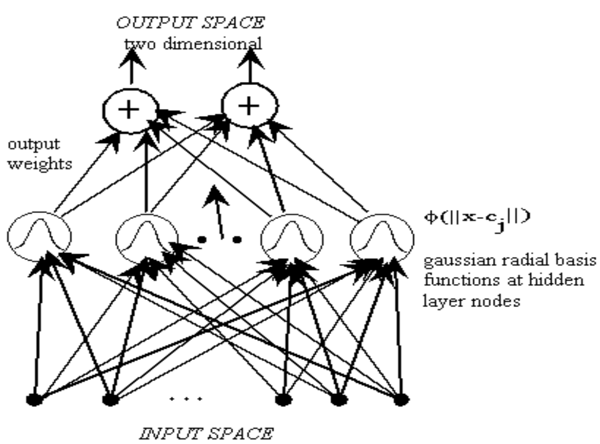


Figure 3: Radial basis function network for knowledge modeling

Without loss of generality, the number of outputs in the network can be extended to a multi-output case. This is a case with an example explained later in this paper, where there are two dimensions at the output space. The network architecture consists of an input layer, a hidden layer and an output layer. The hidden layer consists of a set of radial basis functions as nodes. Each node has a parameter vector **c** defining a cluster center dimension of which is equal to the input vector. The hidden layer node calculates the Euclidean distance between the center and the network's input vector. The distance calculated is used to determine the radial base function output. Conventionally, all the radial basis functions in the hidden layer nodes are the same type and usually gaussian. The response of the output layer node(s) can be seen as a map f: $R^n \rightarrow R$, of the form

$$f(\mathbf{x}) = \mathbf{\Sigma}\, w_i\, \Phi(\|\mathbf{x} - \mathbf{c}_i\|)^2$$

Here the summation is over the number of training data N. $\mathbf{c}_i$ (i=1,2,….N) is the i-th center which may be equal to the input vector $\mathbf{x}_i$ or may be determined in some other way. Once the basis function outputs are determined, the connection weights from hidden layer to the output are determined from a linear set of equations. As a result, accurate functional approximation is obtained. The complexity increases as the size of the training data increases. For a large data set, this may become unpractical, and therefore it is desirable to use limited number of hidden layer nodes in place of having a number equal to N.

The mathematical model employed is a multivariable functional approximation structured in a neuro-fuzzy knowledge representation form. The advantage of such structure is that the knowledge can be effectively modelled by such system with appropriate learning strategy. The neuro-fuzzy structures are well-known and used in engineering and applied sciences since it is especially suitable for data from exact sciences. In contrast to that, the data from soft sciences, including the architectural data as well are mostly linguistic and intuitive rather than exact. In this respect, the approach employed presents novelties in two aspects. Firstly, considering the linguistic nature of the architectural data, fuzzy logic techniques are invoked. Secondly, knowledge is modelled by machine learning methods so that the complex data is structured automatically without use of domain knowledge explicitly. Domain knowledge plays important role in providing the underlying information, which is subject to modelling. Such a model is supposed to be generic and robust enough for design in the domain of concern.

Such form, is compatible with a fuzzy logic structure as the radial basis functions play the role of fuzzy membership functions (Cios et al., 1998) and the output is a fuzzy decision-making based on the soft (fuzzy) architectural design data. In particular, the machine learning method used is orthogonal least squares (OLS) method (Chen et al., 1991), which is the essential requirement to use for machine learning in this particular knowledge modelling research.

After careful deliberation it became evident that for the data at hand, there were some peculiarities which could affect the knowledge model. One such peculiarity was very few data for the 0.1 range in the output space. Our expectation was that if we trained the network on whole range, including the 0.1 range which is rather poorly represented, it would affect the knowledge model in a sense that the estimation error for the unknown cases would become greater since the generalisation error would increase. In order to verify the hypothesis, two separate experiments were conducted.

Experiment 1
Variables explaining comfort and safety were used as an input data (*table 1* and *table 2*). The comfort and safety were used as an output data. Whole data range was from 0.1 till 0.9 (0.1, 0.3, 0.5, 0.7 and 0.9). For comfort aspect only 1.8% of the whole output data had 0.1 value, meaning that only 1.8% of the selected population group did not feel comfortable at Blaak station. Similar case was with the safety aspect, where only 4.1% of the whole output data set had 0.1 value. The OLS training results are given in *figure 4a*, where in total 208 training sets were used for network training. The same training results are represented once more but differently with respect to the hierarchical cluster sequence obtained from the OLS training (*figure 4b*). When acquiring the most representative cases for the whole input/output relation space, these 0.1 cases at the output (which we consider as 'extreme' cases) were all placed in the first 85 priority cases (*figure 4b*).

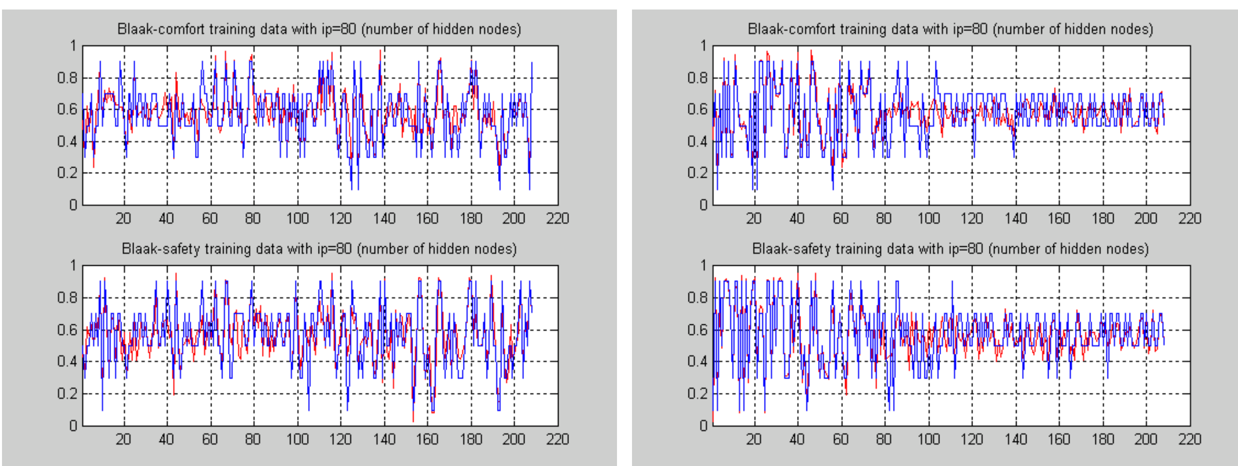


Figure 4: Training results for range 0.1- 0.9 with 80 hidden nodes (4a- figure left); hierarchical cluster sequence obtained from the OLS training (4b-figure right). Broken lines represent the knowledge model response to the training data after training. Continuous lines represent the actual knowledge used for modeling

The network includes 0.1 range and treats that data as important since they are very few cases to learn upon, and in such way the network represented the whole range from 0.1 to 0.9 as good as possible in order to reduce a potential, general error of the model. In the case of Blaak station it turns out to be that 0.1 range is not the most representative of the public opinion at the output space at all. After network training, seven unknown cases were used for testing the network performance and these results are given in *figure 5*.

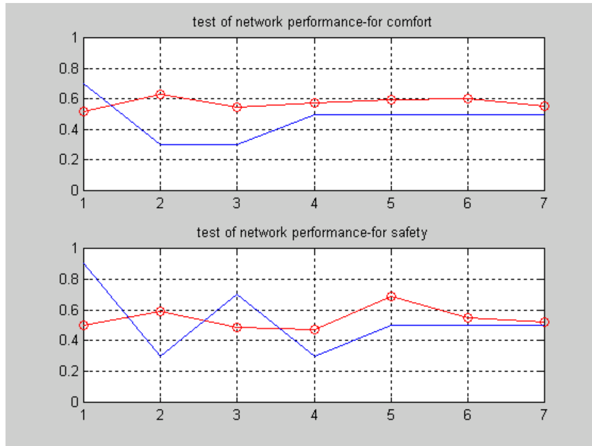


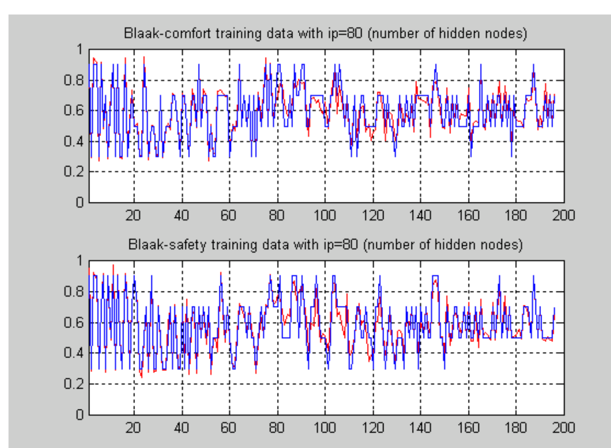
Figure 5: Test results for seven unknown cases (solid line is an actual value and a line with the circles is an estimated value)

If we wanted to make any predictions using the trained neural network, the range below 0.3 would be prone to greater error and at the same time the rest of the model would be affected as well due to high generalisation error. Since the estimation results were not satisfactory, second experiment was done in order to improve the network performance and reduce the generalisation error.

Experiment 2

Being aware of a low-density information present bellow 0.3 range for both safety and comfort, it was decided for this experiment to neglect that range and to move it up to 0.3. In such case, it is expected that the network performance would significantly improve, since the generalisation error for the total model would reduce and better estimations could be obtained, meaning the knowledge model would be more reliable For those reasons 0.1 range at the output space was excluded, together with its associated input values.

Final amount of cases that could be used for the neural network training was 203. A decision was made to use 196 cases for the training and the same seven cases used in previous experiment for testing the network performance were left aside for the test purpose, so that eventually the test results from the two experiments could be compared. In both experiments, number of hidden nodes was set to 80. The OLS training results are given in *figure 6a*, where in total 196 training sets were used for network training. When acquiring the most representative cases for the whole input/output relation space, the 0.3 range was represented through whole model (*figure 6b*).



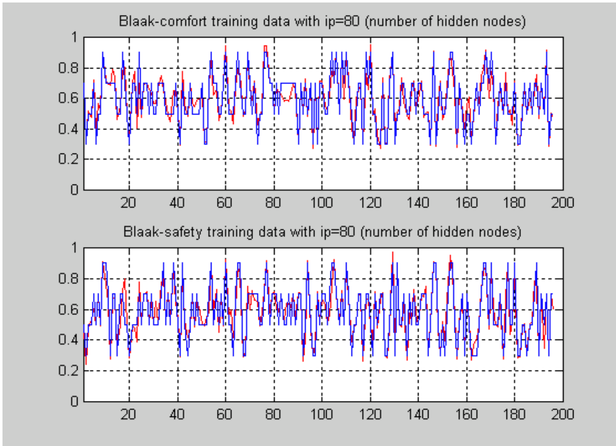


Figure 6: Training results for range 0.3- 0.9 with 80 hidden nodes (6a- figure left); hierarchical cluster sequence obtained from the OLS training (6b-figure right). Broken lines represent the knowledge model response to the training data after training. Continuous lines represent the actual knowledge used for modeling

Having trained the network for the range 0.3 to 0.9, network performance was tested on seven cases and the results are given in *figure 7*.

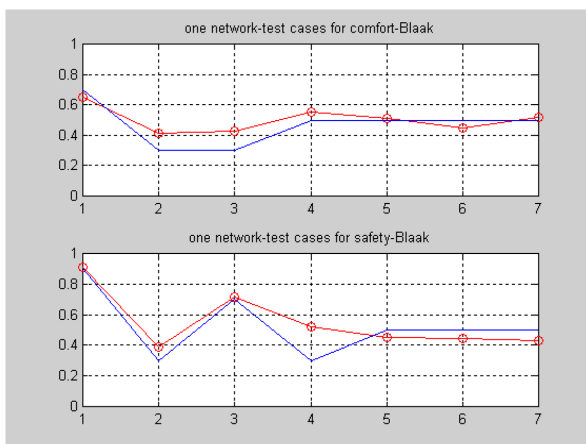


Figure 7: Test results for seven unknown cases (solid line is an actual value and a line with the circles is an estimated value)

Comparing results obtained in *figure 5* and *figure 7*, it becomes evident that network performance is significantly improved once the 0.1 range was excluded, meaning that the later model can be now used as a more reliable knowledge model. It can be concluded that the network tried to cover the whole range during the training, which means that we artificially try to extend the range to 0.1, even though the network has very few examples to learn on that range and in order to reduce its training error, it superficially forces the network to learn on these few examples. The major drawback of this situation is due to the degradation effect on the generalisation capability of the model.

Based on the model obtained from second experiment, the relative dependency of the input variables on comfort and safety is identified by means of sensitivity analysis (Bhatti, 2000) where basically the gradients of comfort and safety with respect to each variable in the input space is computed. The 43 aspects represented in *figure 8* corresponds to aspects listed in *table 1* and *table 2*. The sensitivity analysis is a method used for determining the dependency of the output of a model on the information fed into the model. In other words, sensitivity analysis "studies the relationships between information flowing in and out of the model" (Saltelli, 2000; p.4).

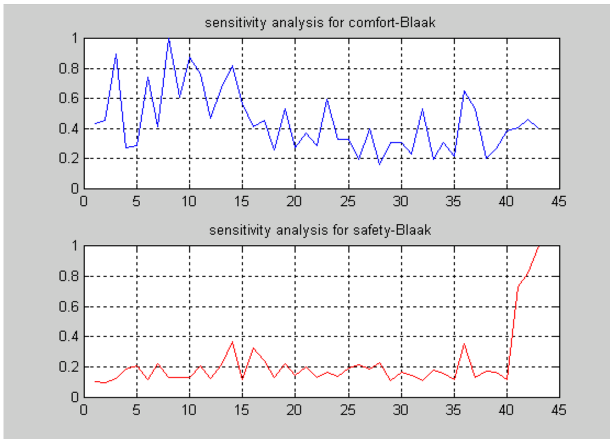


Figure 8: Sensitivity analysis results for comfort (upper picture) and safety (lower picture)

From the results represented in previous figure we can derive the most sensitive aspects in relation to safety and comfort. Exact values and some more insight into these aspects is given in *table 3* and *table 4*. Only few most important aspects are listed in order to illustrate the results obtained by sensitivity analysis.

number	relative importance value	sensitivity to safety
1	1.0000	Safety in surrounding
2	0.8202	Few people present during night
3	0.7266	Few people present during daytime
4	0.3680	Pleasantness of metro platform
5	0.3517	Lighting of train platform
6	0.3239	Wayfinding in station

Table3: most sensitive aspects in relation to safety at Blaak station

number	relative importance value	Sensitivity to comfort
1	1.0000	Spaciousness metro platform
2	0.8972	Spatial proportions
3	0.8723	Platform width
4	0.8171	Pleasantness metro platform
5	0.7647	Platform height
6	0.7403	Spaciousness entrance
7	0.6738	Pleasantness train platform
8	0.6508	Lighting of train platform
9	0.6102	Platform length

Table4: most sensitive aspects in relation to comfort at Blaak station

The outcomes from the knowledge model indicated outstanding potentiality of the model for gaining detailed insight into the information at hand and the effective use of such design information in a structural form as a knowledge base, for enhanced architectural design decisions. In the context of knowledge management, this knowledge model was especially designed to assess the qualitative aspects of design, where at the same time knowledge is captured on user perception of this particular station.

CONCLUSIONS

Situations dealing with the numerical data may occur quite naturally in exact sciences like engineering sciences, life sciences etc. However, the quantities subject to consideration in soft sciences are often qualitative rather than quantitative so that we relate to that type of data as 'soft' data. As an example, in such cases, the quantities may be linguistic so that such quantities have to be somehow expressed in numerical form for treatment by conclusive numerical analysis methods. The analyses are performed by means of soft computing methods. The data subject to analysis and later to knowledge modeling belongs to an underground station that is already being used. For this purpose, the data on psychological aspects are obtained via comprehensive inquiry of the users of underground station. For the analysis, the linguistic information is firstly converted to terms in fuzzy logic domain and after appropriate treatment, the data analyses are carried out and the results are expressed in most comprehensible form for design assessments.
Based on experiments that were conducted in this paper, the performance of the network was significantly improved by excluding 0.1 range. If the 0.1 range was used as well it would mean that the knowledge base 'artificially' extended the output space with rather poor information, which would have an undesirable impact on the generalisation capabilities of the model.

Further on, the sensitivity analysis is conducted, which provided the important results for the designers, ranking in order the aspects that are most sensitive to safety and comfort for Blaak station. With this method, the knowledge is captured on user perception for this particular building. This method can be of great value for an organisational and individual learning and innovation process as argued in the introduction part.

REFERENCES

Bhatti, M. A., (2000). *Practical Optimization Method.* Springer-Telos, Berlin

Broomhead, D.S. and Lowe, D., (1988). *Multivariable Function Interpolation and Adaptive Networks*, Complex Systems, 21, p.321-355

Chen S, C.F.N. Cowan and Grant, P.M., (1991). *Orthogonal Least Squares Algorithm for Radial Basis Function Networks*, IEEE Trans. on Neural networks, Vol.2, No.2, March

Cios K, Pedrycz W and Swiniarski R., (1998). *Data Mining Methods for Knowledge Discovery*, Kluwer Academic Publishers, Bosto/Dordrect/London

Ciftcioglu, Ö., Durmisevic, S. and Sariyildiz, S., (2001)."Multi-Resolutional Knowledge Representation", Conference Proceedings of EuropIA8, 25-27 April 2001, Delft, The Netherlands

Dalen, van J. and Leede, de E., (2000). *Statistisch Onderzoek met SPSS for Windows*, Uitgeverij Lemma BV, Utrecht

EFQM, (1999). *Introducing Excellence*, The European Foundation for Quality Management (EFQM), Brussels

Evboumwan, N. F. O. and Anumba, C. J., (1996). Towards a concurrent engineering model for design-and-build projects, *The structural Engineer*, 74, No. 5, March 1996, pp73-78

Hamel, R., (1990). *Over het Denken van de Architect*. Krips Repro, Meppel, Amsterdam, The Netherlands

Hunt K.J., Haas R. and Murray-Smith R., (1998). Extending the Functional Equivalence of Radial Basis Function Networks and Fuzzy Inference Systems, IEEE Trans. on Neural Networks, Vol.7, No.3, May

Kosko B., (1994). Fuzzy Systems as Universal Approximators, IEEE Trans. on Computers, Vol.43, No 11, November

Labaw, P. J., (1980). *Advanced Questionnaire Design*. Abt Books, Cambridge, Massachusetts Rafiq, M. Y., (2000). A design support tool for optimum building concept generation using a structured genetic algorithm, *International Journal of Computer Integrated Design and Construction*, 2, No. 2, pp92-102

Roger J.S. and Sun C.T., 1990, Functional Equivalence between Radial Basis Function Networks and Fuzzy Inference Systems, IEEE Trans. on Neural Networks, Vol.4, No 1, January Saltelli, A., (2000). *Sensitivity Analysis.* Edited by Saltelli, A., Chan, K. and Scott, E. M. John Wiley & Sons Ltd, West Sussex

Zadeh L.A., (1965). *Fuzzy sets*. Information and Control 8, 338-353

Wang Li-X., (1994). Adaptive Fuzzy Systems and Control, PTR Prentice Hall, Englewood Cliffs, New Jersey 07632