Investigation of the case-based reasoning retrieval process to estimate resources in construction projects

Conference Paper

Author(s):
García de Soto, Borja; Adey, Bryan T.

Publication date:
2015

Permanent link:
https://doi.org/10.3929/ethz-b-000109674

Rights / license:
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

Originally published in:
Procedia Engineering 123, https://doi.org/10.1016/j.proeng.2015.10.074
Investigation of the case-based reasoning retrieval process to estimate resources in construction projects

Borja GARCÍA DE SOTOa, Bryan T. ADEYA

a Institute of Construction and Infrastructure Management, ETH Zurich, Stefano-Franscini-Platz 5, Zurich, 8093, Switzerland

Abstract

Case-based reasoning (CBR) is a methodology that is seeing increasing use to make predictions during the early phases of a project. It allows estimators to exploit existing knowledge to make predictions that are considerably better than without its use. All CBR, however, is not identical, and variations in how CBR is done can affect the accuracy of the predictions. One particular area of sensitivity is the retrieval phase, i.e. the way in which the CBR determines the closeness between the new and the existing cases. In this paper, CBR is used to make estimates of resources for construction projects, and the use of the nearest neighbor technique to identify the similarity for the retrieval phase to predict the construction material quantities (CMQs) in concrete structures is investigated. Two types of distances, i.e. 1) the City-block distance and 2) the Euclidean distance, and four different types of weights, based on regression analysis and feature counting, to account for the relative importance of the different parameters, are investigated. The four different types of weights used were 1) the adjusted unstandardized coefficients from the regression models, 2) the unadjusted unstandardized coefficients from the regression models, 3) the standardized coefficients from the regression models, and 4) equal weights (i.e., feature counting), in which the weights applied are 1/k, and k is the number of parameter being compared to determine the distance.

The mean absolute percentage error (MAPE) was used to evaluate each combination investigated. It was found that for a similarity threshold of 90%, the CBR methodology using the City-block distance with the adjusted unstandardized coefficients from the regression analysis models using the transformed (LN) dataset as weights, gave the best results, with a MAPE of 8.16%. The worst results were obtained from the CBR methodology using the Euclidean distance with feature counting weights, with a MAPE of 28.40%.

Keywords: artificial intelligence; case-based reasoning; preliminary estimates; resource estimates; retrieval process.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of the Creative Construction Conference 2015

*Corresponding author. Tel.: +41 44 633 92 30; fax: +41 44 633 10 88.
E-mail address: garcia.de.soto@ibi.baug.ethz.ch
1. Introduction

Case-based reasoning (CBR) has been used to make estimates in construction projects. [1] compared CBR with multiple regression and neural networks and found that the NN model made more accurate estimates than either the models using regression or CBR; however, the latter performed better than the NN model with respect to long-term use and maintenance, and resulted in a better balance between the time to develop the model and the accuracy of the estimates. The CBR system was developed based on ESTEEM, a CBR development tool, produced by ESTEEM Software Inc. Other researchers have used CBR in combination with other techniques. For example, [2] evaluated different ways to determine the weight of attributes in a CBR model for the estimation of construction cost of residential buildings in Korea. They used three methods: assumed equal weights (EW), the gradient descent method (GDM), and the analytic hierarchy process (AHP). The results showed that the AHP–CBR model was more accurate, reliable, and explanatory than the EW–CBR model, which applied equal weights for the attributes, and the GDM–CBR model, which determined the weights of attributes using the GDM. Similarly, [3] developed a hybrid analytic hierarchy process (AHP) using CBR to estimate the cost of highway projects in South Korea. The AHP method was used to assign the weights to the different cost factors. [4] developed a CBR hybrid model for the prediction of duration and costs made during the early stages of multi-family housing projects in Korea. The CBR used the nearest-neighbor retrieval method for the similarity function. [5] worked on a CBR model that used the standardized coefficients from multiple regression analysis as attribute weights to determine the case similarity and the unstandardized coefficients for the revision phase. [6] developed a conceptual cost prediction model that combined rough set theory (RST), CBR and genetic algorithms (GA) for cost estimates during the conceptual planning phase of public road projects in Korea. When comparing the mean absolute percentage errors (MAPEs) from the CBR model with the traditional cost per mile method, they found that the CBR model performed better than the traditional method. [7] developed a GA-based CBR system to predict the construction cost of high-rise buildings in South Korea during the preliminary design stage. They found that the errors from the GA-based CBR system were lower than feature counting-CBR system (in which all factors were assigned equal weights).

Although most of these CBR systems use the nearest-neighbor retrieval method, they do not explore different similarity functions. In addition, with few exceptions ([8, 9]), CBR has been mostly used to estimate construction costs directly, as opposed to estimate construction material quantities (CMQs) during the early phases of a project. In addition, To address these limitations, this paper concentrates on CMQ estimates. This type of estimates are beneficial because they allow for a clear separation between technical estimates (quantities) and market fluctuations (cost of materials and labor) and they can be easily coupled with cost data (i.e., the corresponding unit cost for each estimated CMQ) to develop cost estimates ([9]). For example, the CMQs or their unit costs can be updated separately during the different phases of the project. In addition it puts managers in a better position to make decisions and keep track of the project by controlling the changes in quantities and costs independently ([9]). The use of CBR allows the exploitation of existing knowledge to significantly improve such estimates. Variations in how CBR is implemented, however, can affect the accuracy of the estimates, especially in the retrieval processes, i.e. the way in which the closeness (or similarity) between the target and the existing cases is determined.

In addition, different variations of the retrieval process in CBR are investigated. The effect is evaluated by using CBR to estimate the CMQs to be used in concrete structures. All investigated retrieval processes use of the nearest neighbor technique to identify existing structures that are similar to the target structures. The differences in the retrieval processes investigated are the types of distances and the types of weights used to account for the relative importance of the different parameters. The two types of distances are, 1) the City-block distance and 2) the Euclidean distance. The four types of weights are 1) the adjusted unstandardized coefficients from regression analysis models, 2) the unadjusted unstandardized coefficients from regression models, 3) the standardized coefficients from regression models, and 4) equal weights (i.e., feature counting), in which the weights applied are \(1/k\), and \(k\) is the number of parameters being compared to determine the distance. The variations in the retrieval processes are evaluated by comparing the MAPE of each.
2. The retrieval process

The retrieval process is the first step in CBR. It requires determining the key parameters to be used to match the target cases with the similar existing cases, determining the values of the key parameters of the target, and determining which of the existing cases have values of the key parameters that are similar to the target case. There are different methods to determine the distance between the existing cases and the target, e.g. the nearest neighbor method, the induction method, the knowledge based induction method, and the template retrieval method ([10]). The most common, however, is the nearest neighbor method ([11]). Within the nearest neighbor method, it is possible to use different similarity functions ([12, 13]), which essentially varies the range of the values of the key parameters considered to be similar.

2.1. Distances

The calculation of the distance using the nearest-neighbor method uses a form of the power, or Minkowski, distance [Equation (1)], in which the user defined variables, \( p \) and \( m \), can be modified to achieve the desired distance function. For example, when \( p = m = 1 \), the distances will be calculated in accordance with the City-block distance; when \( p = m = 2 \), the distances will be calculated in accordance with the Euclidean distance ([12]). The Euclidean distance is the most common ([13, 14]).

\[
Dist(X_o, X_j) = \left( \sum_{i=1}^{n} \left( |x_{oi} - x_{ji}| \right)^{p/m} \right)^{1/m}
\]  

Equation (1)

Where,
- \( X_o \): existing case
- \( X_j \): target (i.e., new) case
- \( x_{oi} \): scaled value of the \( i^{th} \) parameter for the existing case (\( X_o \))
- \( x_{ji} \): scaled value of the \( i^{th} \) parameter for the target case (\( X_j \))
- \( n \): number of parameters, from \( i = 1 \) to \( n \)
- \( p \): user-defined variable related to the importance of the differences of individual parameters
- \( m \): user-defined variable related to the importance of large differences between the cases being compared

In these calculations if the parameters have different ranges then the parameters with the large ranges can overpower the ones with smaller ranges. For example, if two parameters, \( \chi \) and \( \psi \) were used to determine the Euclidean distance, and \( \chi \) can have values ranging between 1 and 1,000, while \( \psi \) can have values ranging between 1 and 10, then the effect of \( \psi \) on the distance function will be over-shadowed by \( \chi \). To avoid this problem, the parameters used to calculate distances should be normalized ([14, 15]). One way to do this normalization is to scale the values of all parameters to be between 0 and 1 [using Equation (2)].

\[
X_{i,norm} = \begin{cases} 
\frac{X_i - X_{i,min}}{X_{i,max} - X_{i,min}} & ; \forall X_{i,max} > X_{i,min} \\
0.5 & ; \forall X_{i,max} = X_{i,min}
\end{cases}
\]

Equation (2)

Where,
- \( X_{i,norm} \): normalized value between 0 and 1
- \( X_i \): raw parameter to be normalized
\( X_i^{\text{min}} \): minimum value for parameter \( X_i \) (minimum of input\(^1\) or existing)

\( X_i^{\text{max}} \): maximum value for parameter \( X_i \) (maximum of input or existing)

This scaling is convenient because it defines the maximum and minimum value for the distance and makes it possible to use the values of different parameters even if their natural values are on different scales. Using the 0-1 range, the basic concepts of bounded ranges, reflexivity, and symmetry ([15, 16]), summarized below, are met.

- Bounded ranges: \( \text{Sim}(x,y) \leq 1 \rightarrow \text{Dist}(x,y) \geq 0 \) (non-negativity)
- Reflexivity: for \( x=y \), when \( \text{Sim}(x,y) = 1 \rightarrow \text{Dist}(x,y) = 0 \), and vice-versa
- Symmetry: \( \text{Sim}(x,y) = \text{Sim}(y,x) \rightarrow \text{Dist}(x,y) = \text{Dist}(y,x) \)

### 2.2. Weights

To account for the relative importance of each parameter, the distances are weighted based on the relative importance of the different key parameters ([11]) [Equation (3)]. For example, in the determination of the value of an estimate, an increase of one unit in the value of variable \( \chi \) may be results in a much larger increase in the estimate than an increase of one unit in the value of the variable \( \psi \). Therefore, when determining the similarity between the target and the existing structures, more importance should be given to variable \( \chi \) than variable \( \psi \). Weights can be determined in different ways. Examples include weights based on the coefficients used in regression models, weights based on the weights of the connections in neural network models, equal importance, or simply based on expert opinion ([02, 03, 04, 05, 06, 07]).

\[
\text{Dist}(X_a, X_f) = \sum_{i=1}^{n} \left( \left| x_{a_i} - x_{f_i} \right|^m \right)^{1/m} \times w_i
\]

Where,

- \( w_i \) : weight corresponding to the \( i^{\text{th}} \) parameter

The weights should be adjusted so that the sum of the adjusted weights equal to one (i.e., \( \sum w_{i, \text{adj}} = 1 \)). This ensures satisfaction of the reflexivity principle (for \( x=y \), when \( \text{Sim}(x,y) = 1 \rightarrow \text{Dist}(x,y) = 0 \)) and accommodates the scaling between 0 and 1 of the key parameters (i.e., continuous independent variables from the regression model). It is also harmonious with the application of the 0 to 1 bounded ranges (i.e., \( \text{Sim}(x,y) \leq 1 \rightarrow \text{Dist}(x,y) \geq 0 \)) and hence to keeping the required scale when determining the similarity ([13, 15]) [Equation (4)].

\[
\text{Dist}(X_a, X_f)_{\text{adj}} = \frac{\sum_{i=1}^{n} \left( \left| x_{a_i} - x_{f_i} \right|^m \right)^{1/m} \times w_{i, \text{adj}}}{\sum_{i=1}^{n} w_i} = \sum_{i=1}^{n} \left( \left| x_{a_i} - x_{f_i} \right|^m \right)^{1/m} \times w_{i, \text{adj}}
\]

Where,

- \( w_i \) : weight corresponding to the \( i^{\text{th}} \) parameter
- \( w_{i, \text{adj}} \) : adjusted weight corresponding to the \( i^{\text{th}} \) parameter [Equation (5)]

\(^1\) This is done to avoid computational problems in the case that the input from the new structure being estimated is outside the range of the existing data by adjusting the range to accommodate the new value (as either a maximum or a minimum, whatever the case might be) and ensure that the scale between 0 and 1 is done properly.
\[ \frac{w_i}{\sum_{j=1}^{n} w_j} \]  

(5)

The weighted distance is used in the similarity function [Equation (6)] during the retrieval process.

\[ \text{Sim}(X_o, X_j) = 1 - \text{Dist}(X_o, X_j)_{adj} \]  

(6)

2.3. Similarity threshold

Once the similarities between the target and existing cases have been calculated, one has to decide which existing cases to use. This is done by setting a similarity threshold. One can think of the similarity threshold as a filter. For example, if the similarity threshold is set to 100%, then only the existing cases with the same values for all key parameters as those of the target case will be considered “similar”. If the similarity threshold is set to 50%, then only the existing cases where Equation (6) has a value of greater than 50% will be considered “similar”.

The exact values of the similarity threshold vary from field to field and situation to situation. [17], for example, chose a similarity threshold of 75% to predict the outcome of construction litigation. [18] used a 70% similarity threshold as sufficient for the final prediction of an international market situation. [13] used an 80% similarity threshold in the estimation of the costs of military and public constructions projects in Korea. [19] suggested a similarity threshold of 80% to estimate the CMQs of structures in manufacturing plants. However, in those studies no specific information was given about how the similarity threshold value was determined. Other researchers did not use a similarity threshold at all and just used the case (or a number of cases) with the highest similarity with the target case ([2, 4, 20]).

In this paper, the similarity threshold used to compare the performance of the different distance-weight combinations (see Table 1) was set to 90%. The selection of this similarity threshold is based on the concept that the higher the similarity threshold the higher the similarity between the new and existing cases; hence, the basis for the estimate of CMQs.

3. Investigated variations in the retrieval process

In all investigated variations in the retrieval process, the distances between the target and the existing structures were determined using the nearest neighbor method [Equation (4)]. The distances were adjusted using four different types of weights. Once the weighted distance was determined, Equation (6) (was used to determine the similarity between the target and existing structures. The distance functions used were 1) the City-block distance \((p=m=1)\), and 2) the Euclidean distance \((p=m=2)\). The weights used were 1) the adjusted unstandardized coefficients (to account for the scaling of the data between 0 and 1) from the regression models using the transformed (LN) data set, 2) the unadjusted unstandardized coefficients (not taking into account the scaling of the data between 0 and 1) from the Regression models using the transformed (LN) data set, 3) the standardized coefficients from the regression models using the transformed (LN) data set, and 4) the equal weights (i.e., feature counting), in which the weights applied are

\[ 2 \] In the development of the regression models, the linear regression equation form was used by taking the natural logarithms of both CMQs and the parameters. It was then transformed to a nonlinear equation. Therefore to be compatible with that transformation, and for the ease of computation, the natural log values of the variables are used for the scaling process.
1/k, and k is the number of parameter being compared to determine the distance. The eight variations are summarized in Table 1. They were investigated for a range of similarity thresholds in the following example.

Table 1. Summary of investigated variations

| Distance (used to determine Similarity) | Weights | Adjusted unstandardized coefficients (wt1) | Unstandardized coefficients (wt2) | Standardized coefficients (wt3) | Equal weights (wt4) |
|----------------------------------------|---------|---------------------------------------------|----------------------------------|---------------------------------|--------------------|
| City-block (SF1)                        | 1       | 2                                           | 3                                | 4                               |                    |
| Euclidean (SF2)                         | 5       | 6                                           | 7                                | 8                               |                    |

4. Example

The effect on the variations in the retrieval process summarized in Table 1 were investigated by estimating the CMQs to be used in storage structures.

4.1. Data

The initial data consisted of CMQs from 58 storage structures from 8 plants located around the world. The structures were randomly split, so that 80% of the structures (46) were considered to be existing cases and 20% of the structures (12) were used to evaluate the variations in the retrieval process. An 80%:20% data split is typically recommended for training and validating model ([21]) and it should be done in a random way to avoid bias in a given set. The data included in both data sets are given in Table 3 (existing structures) and Table 4 (target structures), respectively.

4.2. Weights

The weights used are derived from the selected regression models developed in the study by [22]. The unstandardized regression coefficients were adjusted using Equation (7) to account for the normalization (i.e., scaling between 0 and 1) of the parameters. The unstandardized, adjusted unstandardized and standardized coefficients, from selected regression models, and feature counting weights, are shown in Table 2.

\[
\beta_{adjusted} = \bar{\beta}_i = \begin{cases} 
\frac{\beta_i \left( \ln \left( X_i^{max} \right) - \ln \left( X_i^{min} \right) \right)}{\sum |\beta_i| \left( \ln \left( X_i^{max} \right) - \ln \left( X_i^{min} \right) \right)} & ; \forall X_i^{max} > X_i^{min} \\
0 & ; \forall X_i^{max} = X_i^{min} 
\end{cases}
\] (7)

The regression models are those developed to give the best-fit relationships between the key parameters and the estimated CMQs.
Table 2. Weights

| Model and variables | Unstandardized Coefficients | Adjusted Unstandardized Coefficients (storage A) | Adjusted Unstandardized Coefficients (storage B) | Adjusted Unstandardized Coefficients (storage C) | Standardized Coefficients (1/k; \(k=4\)) |
|---------------------|----------------------------|------------------------------------------------|------------------------------------------------|------------------------------------------------|------------------------------------------|
| CO* ln_Cap          | 0.43                       | 0.52                                           | 0.49                                           | 0.50                                           | 1.05                                     | 0.25                                    |
|                    | ln_Diam                    | 0.77                                           | 0.22                                           | 0.33                                           | 0.30                                     | 0.88                                     | 0.25                                    |
|                    | ln_Height                  | 0.10                                           | 0.11                                           | 0.07                                           | 0.09                                     | 0.10                                     | 0.25                                    |
|                    | ln_Ground acc              | 0.57                                           | 0.16                                           | 0.10                                           | 0.12                                     | 0.13                                     | 0.25                                    |
| RE* ln_Cap         | 0.40                       | 0.48                                           | 0.48                                           | 0.47                                           | 1.00                                     | 0.25                                    |
|                    | ln_Diam                    | 0.65                                           | 0.19                                           | 0.30                                           | 0.26                                     | 0.77                                     | 0.25                                    |
|                    | ln_Height                  | 0.14                                           | 0.14                                           | 0.09                                           | 0.12                                     | 0.14                                     | 0.25                                    |
|                    | ln_Ground acc              | 0.70                                           | 0.19                                           | 0.13                                           | 0.14                                     | 0.16                                     | 0.25                                    |

*CO: Concrete; RE: Reinforcement

4.3. Evaluation measure

The variations in the retrieval process were evaluated by measuring the mean absolute percentage errors (MAPE) [Equation (8)] of the estimated CMQs with the actual CMQs for the target structures.

\[
MAPE(\%) = \frac{100}{n} \times \sum \left( \frac{CMQ_{est} - CMQ_{act}}{CMQ_{act}} \right) \tag{8}
\]

Where,
- \(n\): sample size
- \(CMQ_{est}\): estimated amount
- \(CMQ_{act}\): actual amount
### Table 3. Data for (existing) storage structures

| Struct. ID | Subtype | Parameters | CMQs |
|------------|---------|------------|------|
|            |         | Storage capacity (in metric tons) |     |
|            |         | Interior diameter of storage structure (in meters) |     |
|            |         | Height from top of foundation to top of concrete structure (in meters) |     |
|            |         | Design wind speed in accordance with the Eurocode 1, EN 1991 1-4 (2010) (wind design) (in meters per second) |     |
|            |         | Spectral response acceleration for 1.0 sec. period (2% probability of exceedance in 50 years) in accordance with the 2009 IBC, which are based on the 2002 USGS National Seismic Hazard Maps (expressed in decimal form as a function of g) |     |
|            |         | Soil bearing capacity (in tons per m²) |     |
|            |         | Soil factor / soil coefficient in accordance with the Eurocode 8, EN 1998 1-6 (2006) |     |
|            |         | Total amount of concrete for upper structure and foundation (in m³) |     |
|            |         | Total amount of reinforcement for upper structure and foundation (in metric tons) |     |

1 A 7'500 18 25 42 0.07 45 1.35 3'555 583
2 A 10'000 18 30 39 0.02 36 1.40 3'358 625
3 A 10'000 18 30 33 0.01 41 1.38 3'431 628
4 A 12'000 18 40 44 0.02 41 1.38 4'579 682
5 A 11'500 20 30 38 0.28 20 1.55 3'866 758
6 A 10'000 22 60 39 0.21 10 1.43 3'358 625
7 A 18'700 22 70 42 0.01 19 1.60 5'585 788
8 A 13'500 19 45 37 0.03 32 1.42 4'310 892
9 A 12'900 19 50 34 0.20 14 1.37 4'622 981
10 A 25'700 18 50 37 0.24 35 1.36 6'341 996
11 A 18'900 22 59 41 0.12 9 1.47 6'005 1'099
12 A 14'600 24 57 40 0.25 15 1.52 6'160 1'103
13 A 16'000 20 65 31 0.24 30 1.40 6'002 1'206
14 B 22'500 28 20 54 0.05 25 1.50 2'707 482
15 B 30'400 44 30 33 0.08 83 1.25 5'647 853
16 B 39'700 44 25 25 0.04 49 1.55 6'248 950
17 B 45'000 38 20 42 0.07 45 1.35 5'930 858
18 B 60'000 39 30 33 0.03 92 1.00 6'251 1'113
19 B 67'200 31 28 27 0.15 65 1.40 6'494 1'117
20 B 95'300 33 30 38 0.04 57 1.20 7'488 1'131
21 B 96'500 29 40 24 0.23 17 1.38 7'307 1'147
22 B 60'000 45 25 39 0.02 37 1.40 7'483 1'152
23 B 60'000 38 45 20 0.22 37 1.40 7'469 1'173
24 B 61'200 44 48 51 0.21 28 1.42 8'476 1'204
25 B 70'000 50 25 47 0.08 46 1.35 10'200 1'469
26 B 79'400 45 32 21 0.13 40 1.50 9'168 1'604
27 B 111'900 51 30 33 0.01 41 1.38 9'867 1'648
28 B 106'800 46 26 52 0.08 45 1.26 10'530 1'664
29 B 98'300 50 35 41 0.22 19 1.19 11'578 1'795
30 B 90'000 45 50 51 0.24 7.5 1.60 12'209 2'161
31 C 9'000 16 30 40 0.04 10 1.37 3'825 535
32 C 8'000 15 35 31 0.24 30 1.40 3'690 542
33 C 7'000 15 30 36 0.05 25 1.50 3'516 674
34 C 8'000 18 40 20 0.02 37 1.40 4'390 733
35 C 8'000 17 40 35 0.16 31 1.59 4'098 799
36 C 12'800 18 65 42 0.19 37 1.58 5'446 801
37 C 19'100 19 65 42 0.02 36 1.39 6'185 829
38 C 12'000 18 60 31 0.22 7.5 1.60 4'969 855
39 C 18'000 20 70 44 0.02 41 1.38 5'968 877
40 C 13'500 17 65 22 0.03 38 1.51 4'810 933
41 C 20'700 20 75 46 0.20 31 1.59 7'440 1'339
42 C 22'800 22 68 47 0.08 46 1.35 7'290 1'341

---

5 Storage capacity (in metric tons)
6 Interior diameter of storage structure (in meters)
7 Height from top of foundation to top of concrete structure (in meters)
8 Design wind speed in accordance with the Eurocode 1, EN 1991 1-4 (2010) (wind design) (in meters per second)
9 Spectral response acceleration for 1.0 sec. period (2% probability of exceedance in 50 years) in accordance with the 2009 IBC, which are based on the 2002 USGS National Seismic Hazard Maps (expressed in decimal form as a function of g)
10 Soil bearing capacity (in tons per m²)
11 Soil factor / soil coefficient in accordance with the Eurocode 8, EN 1998 1-6 (2006)
12 Total amount of concrete for upper structure and foundation (in m³)
13 Total amount of reinforcement for upper structure and foundation (in metric tons)
Table 4. Data for (target) storage structures

| Struct. ID | Subtype | Parameters                  | CMQs                  |
|------------|---------|-----------------------------|-----------------------|
|            |         | Cap (t)                     | Diam (m)              |
|            |         | Height (m)                  | Wind sp (m/s)         |
|            |         |                             | Ground acc (S1xg)     |
|            |         |                             | Soil BC (t/m²)        |
|            |         |                             | Soil factor           |
|            |         |                             | Concrete (m³)         |
|            |         |                             | Reinf (t)             |
| T1         | A       | 10’000                       | 18                    |
| T2         | A       | 10’700                       | 28                    |
| T3         | A       | 14’000                       | 30                    |
| T4         | A       | 14’800                       | 34                    |
| T5         | B       | 59’000                       | 32                    |
| T6         | B       | 60’000                       | 32                    |
| T7         | B       | 75’000                       | 32                    |
| T8         | B       | 107’900                      | 32                    |
| T9         | C       | 8’000                        | 32                    |
| T10        | C       | 13’300                       | 32                    |
| T11        | C       | 15’500                       | 32                    |
| T12        | C       | 21’800                       | 32                    |

4.4. Evaluation process

The evaluation process is shown in Figure 1, which was followed for every structure subtype and CMQ combination. To generalize, the results for each structure subtype and CMQs were combined. The validation cases for the different structure subtypes, and corresponding CMQs, were used as target structures (from j to k) and not included as existing cases (i.e., the case where Sim = 100% was not applicable). The structures used as existing cases (case-based) were the structures used to develop the RA (from i to n). Therefore, this process assumes that there is n number of existing structures of the same subtype as the target structure.

4.5. Results

In total 576 events (6 structure subtype-CMQ scenarios x 8 cases x 1 similarity threshold value x 12 target structures) were used to determine which case (i.e., distance and weight combination used in similarity function) produced the best results. For generalization purposes, the results for each structure subtype for the different CMQs evaluated were combined and summarized in Table 5. They show the MAPE for all the CMQs in all structure subtypes with a similarity threshold of 90% for the different distance-weight combinations (cases) evaluated.

Table 5. Overall MAPE (%) for different distance-weight combinations (cases 1-8) for a similarity threshold of 90%

| Case No.* | 1 (SF1-wt1) | 2 (SF1-wt2) | 3 (SF1-wt3) | 4 (SF1-wt4) | 5 (SF2-wt1) | 6 (SF2-wt2) | 7 (SF2-wt3) | 8 (SF2-wt4) |
|-----------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| MAPE (%)  | 8.16        | 9.21        | 8.99        | 11.87       | 23.68       | 24.53       | 21.94       | 28.4        |

13 Three structure subtypes (storages A, B, and C) and two CMQs (total amount of concrete and total amount of reinforcement).
Figure 1: Evaluation process

4.6. Accuracy of investigated variations

As shown in Table 5, the retrieval process using the City-block distance (cases 1-4) provides better results than those obtained using the Euclidean distance (cases 5-8). This is because the use of the transformed and scaled parameters in the Euclidean distances, in which the differences between the different parameters are squared, dampens their effect (especially for small differences). This is true independent of the weights used. When the retrieval process uses the City-block distance, which uses the sum of distances along each dimension, the estimation of the distances is
more sensitive to differences among the different key parameters\textsuperscript{14}. This can be seen in the small example in the following section.

The results from Table 5 also show that the type of weight used did not affect the results significantly (e.g., the largest difference between maximum and minimum MAPE values for a given similarity threshold for cases 1-4 was 3.71\% (11.87\% - 8.16\%) when the similarity threshold was 90\%.

Although not large, the retrieval process using the City-block distance (Cases 1-4) and the adjusted unstandardized coefficients from the RA models performed slightly better (i.e., with a MAPE of 8.16\% vs. 8.99\% for the next closest case (Table 5)) than the others. Case 1 is optimal.

4.7. Sensitivity of distance measures

The sensitivity of the City-block distance and the Euclidean distance to differences in the values of parameters is shown using the transformed and normalized data for storage structure subtype A (Table 6) to determine the distances between a target structure (ID T3) and selected existing storage structures (IDs 7, 8, 11, 12, 13). The results are summarized in Table 7.

Table 6. Transformed and normalized parameters used for comparison of distance calculation

| Structure | Transformed (ln) and normalized (0-1) parameters |
|-----------|-------------------------------------------------|
| ID        | Type | Subtype | Cap   | Diam  | Height | Wind sp | Ground acc | Soil BC | Soil factor |
|-----------|------|---------|-------|-------|--------|---------|-------------|---------|-------------|
| T3        | Storage | A      | 0.50  | 0.37  | 0.85   | 0.17    | 1.00        | 0.55    | 0.72        |
| 8         | Storage | A      | 0.47  | 0.19  | 0.57   | 0.55    | 0.06        | 0.58    | 0.75        |
| 12        | Storage | A      | 0.54  | 1.00  | 0.80   | 0.75    | 0.78        | 0.28    | 0.90        |
| 13        | Storage | A      | 0.61  | 0.37  | 0.93   | 0.09    | 0.75        | 0.55    | 0.72        |
| 7         | Storage | A      | 0.73  | 0.70  | 1.00   | 0.88    | 0.01        | 0.37    | 1.00        |
| 11        | Storage | A      | 0.74  | 0.70  | 0.83   | 0.82    | 0.37        | 0.07    | 0.81        |

Table 7. Distances between target and existing structures using Euclidean and City-block equations

| Euclidean       | Distance between | Cap | Diam | Height | Wind sp | Ground acc | Soil BC | Soil factor | Distance |
|-----------------|------------------|-----|------|--------|---------|-------------|---------|-------------|----------|
| T3 and 8        | 0.00             | 0.03| 0.08 | 0.15   | 0.88    | 0.00        | 0.00    | 1.07        |
| T3 and 12       | 0.00             | 0.40| 0.00 | 0.34   | 0.05    | 0.07        | 0.03    | 0.95        |
| T3 and 13       | 0.01             | 0.00| 0.01 | 0.01   | 0.06    | 0.00        | 0.00    | 0.29        |
| T3 and 7        | 0.05             | 0.11| 0.02 | 0.51   | 0.98    | 0.03        | 0.08    | 1.34        |
| T3 and 11       | 0.06             | 0.11| 0.00 | 0.42   | 0.40    | 0.23        | 0.01    | 1.11        |

| City-block     | Distance between | Cap | Diam | Height | Wind sp | Ground acc | Soil BC | Soil factor | Distance |
|----------------|------------------|-----|------|--------|---------|-------------|---------|-------------|----------|
| T3 and 8       | 0.03             | 0.18| 0.28 | 0.38   | 0.94    | 0.03        | 0.03    | 1.87        |
| T3 and 12       | 0.04             | 0.63| 0.05 | 0.58   | 0.22    | 0.27        | 0.18    | 1.98        |
| T3 and 13       | 0.11             | 0.00| 0.08 | 0.08   | 0.25    | 0.00        | 0.00    | 0.53        |
| T3 and 7        | 0.23             | 0.33| 0.15 | 0.71   | 0.99    | 0.18        | 0.28    | 2.88        |
| T3 and 11       | 0.24             | 0.33| 0.02 | 0.65   | 0.63    | 0.48        | 0.09    | 2.45        |

As can be seen in Table 7, the differences between each parameter are damped in the Euclidean distances when compared to the City-block distances (e.g., for the diameter between T3 and 8 the differences are 0.03 and 0.18 for the Euclidean and City-block distances, respectively). Therefore, the City-block distance is more sensitive to the differences between the target and existing structures. This sensitivity is transferred to the ranking of the existing

\textsuperscript{14} As indicated, this is true when the transformed and scaled parameters are used, hence the difference for a given parameter between two structures (diff.) belong to an open interval (0, 1), so that 0 < diff. < 1 and diff.\textsuperscript{2} < diff.
structures (e.g., from more to less similar), their selection during the CBR retrieval phase, and ultimately affecting the basis for the estimation of the target structure.

5. Conclusion

In this paper, different variations of the retrieval process in CBR were investigated by using CBR to estimate the CMQs to be used in storage structures. All investigated retrieval processes use of the nearest neighbor technique to identify existing structures that are similar to the target structures. The variations in the retrieval processes are evaluated by comparing the mean absolute percentage errors (MAPE) of each.

Of the eight retrieval processes evaluated it was found that the retrieval process that used the City-block distance with the adjusted unstandardized coefficients yielded the most accurate results, with a (MAPE of 8.16%). The worst results were obtained when the Euclidean distance with equal weights were used (MAPE = 28.40%). All using a similarity threshold of 90%.

The use of CBR to make estimates of CMQs, as presented in this paper, has some limitations. CMQ-based estimates can become very demanding and time consuming. A model has to be developed and tested for every structure and its corresponding CMQs. In addition, CBR, as most estimations models, are based on historical data (i.e., stored cases). When the stored cases are not similar to the new cases (e.g., do not meet the similarity threshold), the estimations cannot be done using CBR. In addition, the current process does not include an important phase of CBR: the revision phase. To address these limitations, the authors are currently working on a revision or adaptation phase using regression to accounts for the differences in the values of the parameters between the target and the existing similar structure using the selected regression model, modified to account for the % error from the regression model. For the cases where similar structures are not available, the estimation of the CMQs can be done using the developed regression models.

References

[01] Kim, G. H., An, S. H., & Kang, K. I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. Building and environment, 39(10), 1235-1242.
[02] An, S. H., Kim, G. H., & Kang, K. I. (2007). A case-based reasoning cost estimating model using experience by analytic hierarchy process. Building and Environment, 42(7), 2573-2579.
[03] Kim, S. (2013). Hybrid forecasting system based on case-based reasoning and analytic hierarchy process for cost estimation. Journal of Civil Engineering and Management, 19(1), 86-96.
[04] Koo, C., Hong, T., Hyun, C., & Koo, K. (2010). A CBR-based hybrid model for predicting a construction duration and cost based on project characteristics in multi-family housing projects. Canadian Journal of Civil Engineering, 37(5), 739-752.
[05] Jin, R., Cho, K., Hyun, C., & Son, M. (2012). MRA-based revised CBR model for cost prediction in the early stage of construction projects. Expert Systems with Applications, 39(5), 5214-5222.
[06] Choi, S., Kim, D. Y., Han, S. H., & Kwak, Y. H. (2013). Conceptual Cost Prediction Model for Public Road Planning via Rough Set Theory and Case-Based Reasoning. Journal of Construction Engineering and Management, 140(1).
[07] Kim, S., & Shim, J. H. (2013). Combining case-based reasoning with genetic algorithm optimization for preliminary cost estimation in construction industry. Canadian Journal of Civil Engineering, 41(1), 65-73.
[08] Du, J. and Bormann, J. (2014). Improved Similarity Measure in Case-Based Reasoning with Global Sensitivity Analysis: An Example of Construction Quantity Estimating. Journal of Computing in Civil Engineering, 28(6), 04014020.
[09] García de Soto, B. (2014). A methodology to make accurate preliminary estimates of construction material quantities for construction projects (Doctoral dissertation). Swiss Federal Institute of Technology in Zurich (ETH Zürich). Diss. ETH No. 22313, Zurich, Switzerland.
[10] Watson, I., & Marir, F. (1994). Case-based reasoning: A review. Knowledge Engineering Review, 9(4), 327-354.
[11] Watson, I. (1998). Applying case-based reasoning: techniques for enterprise systems. San Francisco, CA: Morgan Kaufmann Publishers Inc.
[12] Núñez, H., Sánchez-Marre, M., Cortés, U., Comas, J., Martínez, M., Rodríguez-Roda, I., & Poch, M. (2004). A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations. Environmental Modelling & Software, 19(9), 809-819.
[13] Ji, S. H., Park, M., & Lee, H. S. (2011). Cost estimation model for building projects using case-based reasoning. Canadian Journal of Civil Engineering, 38(5), 570-581.
[14] Wilson, D. R., & Martinez, T. R. (1997). Improved Heterogeneous Distance Functions. Journal of Artificial Intelligence Research, 6, 1-34.
[15] López, B. (2013). Case-based reasoning: a concise introduction. Synthesis Lectures on Artificial Intelligence and Machine Learning, 7(1), 1-103.

[16] Burkhard, H. D., & Richter, M. M. (2001). On the notion of similarity in case based reasoning and fuzzy theory. In Soft computing in case based reasoning (pp. 29-45). London, UK: Springer.

[17] Arditi, D., & Tokdemir, O. B. (1999). Using Case-Based Reasoning to Predict the Outcome of Construction Litigation. Computer–Aided Civil and Infrastructure Engineering, 14(6), 385-393.

[18] Ozorhon, B., Dikmen, I., & Birgonul, M. T. (2006). Case-based reasoning model for international market selection. Journal of Construction Engineering and Management, 132(9), 940-948.

[19] García de Soto, B., Fernando, D., & Adey, B. T. (2013). Methodology to accurately estimate the quantities of construction materials used in cement plant construction projects. MIP Project. Internal HGRS Report. Holderbank, Switzerland.

[20] Jin, R., Han, S., Hyun, C., & Kim, J. (2014). Improving Accuracy of Early Stage Cost Estimation by Revising Categorical Variables in a Case-Based Reasoning Model. Journal of Construction Engineering and Management. 140(7), 04014025.

[21] Hegazy, T., & Ayed, A. (1998). Neural network model for parametric cost estimation of highway projects. Journal of Construction Engineering and Management, 124(3), 210-218.

[22] García de Soto, B., Adey, B. T., & Fernando, D. (2014). A process for the development and evaluation of preliminary construction material quantity estimation models using backward-elimination-regression and neural networks. Journal of Cost Analysis and Parametrics. 7(3), 180-218, DOI: 10.1080/1941658X.2014.984880.