BELIEF NETWORKS IN CONSTRUCTION SIMULATION

Brenda McCabe
Department of Civil Engineering
University of Toronto
Toronto, Ontario, M5S 1A4, CANADA

ABSTRACT

A method for automatically improving the performance of construction operations was developed by the integration of computer simulation and belief networks. The simulation model is used to represent the operation and to determine the effect that changes in resource configuration have on the model performance. The belief network provides diagnostic analysis of the performance and recommendations for changes to the model. Modifications to the simulation model include selection of alternative resources and resource quantities.

The method is based upon the assumption that ultimate objectives, such as lower costs or shorter project duration, will result from efficient use of resources. Therefore, the improvement process is focused upon resource performance instead of cost or duration. The approach is iterative, and will provide the modeler with results even if user-defined constraints related to performance limits are not met.

1 INTRODUCTION

Simulation is used to model operations such as construction because the operation is too complex to model entirely mathematically, or because there is some uncertainty in the system. Although simulation has been documented as an excellent tool for modeling construction operations (AbouRizk and Halpin 1990, AbouRizk and Dozzi 1993, Smith and Osborne 1995), it has not experienced widespread use in industry (Shi & AbouRizk 1994). One obstacle to the acceptance of simulation by the construction industry is the effort required for experimentation with the model in order to optimize it (McCabe 1997).

A method for automatically improving the performance of a simulated operation was developed through the integration of computer simulation and artificial intelligence, specifically belief networks. The method is not domain specific but requires the interaction of servers and customers in queue-type structures.

2 BACKGROUND

Several simulation languages are available. CYCLONE (Halpin 1976) was developed specifically for modeling construction operations. Many CYCLONE-based systems have been developed to modify or extend the functionality of CYCLONE. These include INSIGHT (Paulson 1978), RESQUE (Chang 1987), UM-CYCLONE (Ioannou 1989), COOPS (Liu and Ioannou 1994), DISCO (Huang, Grigoriadis & Halpin 1994), CIPROS (Tommelein & Odeh 1994), STROBOSCOPE (Martinez & Ioannou 1994), HSM (Sawhney & AbouRizk 1995), and ACPSS (Liu 1996).

More general simulation languages are available, such as Visual SLAM (Pritsker, O'Reilly & LaVal 1997), GPSS/H (Crain & Smith 1994), SIMAN/Cinema (Profozich & Sturrock 1994), and SIMSCRIPT (Russell 1993). These systems are capable of supporting simulation modeling in any domain including manufacturing, industrial engineering and construction. The price paid for increased flexibility, however, is the increased skill level required by the simulation modeler.

General optimization routines and simulation environments are often incompatible because the modeling techniques of simulation and mathematics are so different. Consequently, many optimization methods have been developed for simulation models (Azadivar 1992). The methods that have been developed, and in some cases automated, have been categorized by Azadivar as 1) gradient based search methods, 2) stochastic approximation methods, 3) response surface methods, and, 4) heuristic search methods.

Gradient-based search methods and stochastic approximation methods are focused on continuous movement toward the optimum. These techniques assume unimodal solution functions and contain algorithms to identify the direction of the steepest slope. Riggs (1976) developed an automated sensitivity analysis module for CYCLONE that required the user to provide the upper and lower limits of the resource quantities available for the operation being modeled. Using this method, the user was
able to establish the direction toward which the optimal resource configuration may be found.

The response surface methodology involves fitting regression models to the results of the simulation run evaluated at various states of the problem domain. Azadivar and Talavage (1980) showed that the effectiveness of this method was greatly reduced if the regression function contained sharp ridges or flat surfaces.

Heuristic methods may not guarantee that the solution found is the global optimum because there is often no assumption that the solution function is unimodal. One may be confident that the solution found by the method is very good, but it may not be the optimum. Two formal heuristic methods have been defined by Azadivar (1992): complex search and simulated annealing. Methods that rely upon artificial intelligence, such as genetic algorithms, rule-based systems, and belief networks also fall into this category.

Complex search involves using the results of several simulation runs from different variable parameters to determine the worst point. The worst point is dropped, a new point is generated, and the simulation is rerun. Simulated annealing is a local gradient search method that evaluates the objective function, say, to minimize the cost, at an appropriately chosen point. If the new cost is less than the cost at the previous point, then the new point is accepted and the old one is dropped. To reduce the likelihood of being caught in a local minimum, the method will allow uphill moves based on random variables with controlled probabilities.

Several heuristic techniques have been developed specifically for improving construction operations. Wood and Harris (1980) developed a program that utilized an iterative technique of simulation and manual cost evaluation to optimize concrete delivery truck fleets. Their model was able to analyze various truck and plant capacities.

AbouRizk and Shi (1994) applied heuristics to a DELAY statistic to determine whether the number of resources in a simulation model should be increased or decreased in order to meet project objectives for optimizing cost, production, or resource utilization. The DELAY statistic is equal to the fraction of time a resource is idle relative to its total working time. The limitation of the work, as cited by the authors, is that the system assumed the simulation model itself cannot be modified, and it could not meet multiple objectives, such as optimal cost and production.

Shi and AbouRizk (1995) developed a hybrid simulation and mathematical optimization system for handling large, complex systems. In this model, the large system is broken into smaller sections for separate evaluation of each feasible resource state. The smaller sections are rejoined by mathematical functions and the entire project is optimized mathematically. The method requires significant manipulation by the user to determine the connection types between the smaller simulation model sections, development of the mathematical functions that connect the smaller sections into the entire project, and fine-tuning.

Tompkins and Azadivar (1995) combined genetic algorithms with object-oriented programming in ModSim II to develop a means of optimizing simulation models for manufacturing systems. The system was intended to represent corporate policy for minimizing resource requirements of new operations. Several billion points could be searched resulting in significantly improved solutions over random search methods.

Chan and Chua (1996) developed a hybrid optimization system using genetic algorithms and computer simulation for use in civil engineering applications. Because of the constraints imposed by practical issues of the specific applications, they found that the genetic algorithms were not allowed to fully optimize the solutions.

Most of the techniques developed to "optimize" simulation models are based on modifying resource quantities, but not resource capacities through the selection of alternative resources. An alternative resource is a resource that is able to perform the same function, but has different parameters that affect its performance. For example, an alternative for one truck resource may be another truck of greater capacity but, perhaps, slower acceleration. Other techniques have also focused on a single optimization objective, such as cost. The approach developed in this research focuses on the surrogate objective of improving performance of all resources based on five performance indices. The drive to improve performance instead of cost or project duration leads to the recommendation of alternative resources.

The issue of finding very good vs. optimal solutions is not perceived as a problem by the author, especially as it relates to construction operations. Moreover, if one does not focus on finding the one optimal point, it follows that the modeler could be presented with several near-optimal solutions. Several, equally acceptable solutions may result when the solution function is rather flat near various optimal points or when there are several local optimums that result in similar system performance. Because construction is vulnerable to innumerable external influences that may continually affect its performance, several very good and equally acceptable solutions may be of more value to the construction planner than a single optimal solution.

3 BELIEF NETWORKS

This performance improvement method is based upon the integration of simulation and belief networks. Belief networks may be described as a form of artificial
Belief Networks in Construction Simulation

Intelligence able to incorporate uncertainty and knowledge into their structures (Pearl 1990). Belief networks are directed, acyclic graphs (DAG) with nodes representing the variables in the problem domain, and arcs representing conditional dependence between the variables (Jensen 1996). Directed means that the arcs have an implicit direction represented by an arrow. Acyclic refers to the constraint that when the direction of the arrows are considered, they may not close upon themselves creating a closed circle. The node or variable with an arrow pointing away is the parent node. The node to which the arrow is pointing is considered the child node.

In diagnostic models including the one developed in this research, variables may be categorized as causal variables and effect variables. (Henrion, Breese & Horvitz 1991) Arcs connect the causal variables to the effect variables to depict the conditional relationships i.e. the state of the cause variable (true or false) affects the state of the effect variable. Probabilities are assigned to each variable based on all possible combination states of the parent nodes.

Poole, Mackworth and Goebel (1998) provide guidance for the development of belief networks in four general steps. First, the variables in the domain must be determined. This is directly affected by the scope of the problem definition. Second, the conditional relationships are defined by connecting the nodes with arc. The resulting graph must be acyclic. Third, the states of the nodes are determined. Where possible, the variables are binary to limit the number of probabilities that must be assigned. Finally, the probability related to each conditional relationship is determined. This entails evaluating all combinations of the states of parent nodes and assigning a probability value of the child node states for each parent state combination.

Figure 1 shows the belief network that was developed for diagnosing performance. Note that the nomenclature from queuing theory has been used. The variables SQ, SU, QW, QL, and CD are the performance indices Server Quantity, Server Utilization, Queue Wait time, Queue Length, and Customer Delay. The performance indices are the effect variables in this network and represent the effect or symptoms of poor performance. The variables TooFewCustomers, TooManyServers, etc. are possible causes of poor performance and are the causal variables. Finally, the variables Cost and Duration have been included to provide direction for the search, but are not explicitly incorporated as objectives. Their role will be discussed shortly.

The arcs linking the variables indicate a conditional relationship between the variables. Causal variables that relate to the customer quantity or capacity affect the customer performance and the queuing performance. The causal variables that relate to the server quantity or capacity affect the server performance and the queuing performance. For example, TooManyCustomers may affect the performance of the customers as measured by the indices CD, and the performance of the queues represented by the variables QW, and QL.

The states of the variables are shown in Table 1 and Table 2. All Causal nodes are binary. These values indicate whether the causal variable is affecting the performance of the resources.

![Figure 1: Belief Network for Performance Improvement](image)

The Effect nodes have two or three states, depending upon the limits. In some cases, such as SU, QL and QW, the value of the variable is bounded by two limits, an upper and lower limit. The subscript L indicates the lower limit of the acceptable range of the value of the variable. The subscript U indicates the upper limit of the acceptable range. The variables SQ and CD are bound by only one limit. The value of SQ is evaluated as equal to zero or greater than zero. The variable CD has only an upper limit, as the lower limit is always zero.

Limits are defined by the modeler to provide guidance to the diagnostic processes. This allows the modeler to test various resource management strategies and to impose project or corporate constraints related to the acceptable performance limits for the particular operation being simulated.

The Cost and Duration nodes were added to allow the improvement process to take different approaches to diagnosing the performance. For example, suppose an activity has unacceptably long queues at the servers. If the shortest project duration is the overall objective, then more, larger servers might be an appropriate strategy. However, if cost is the objective, then fewer, smaller customers may better achieve the goal. The iterative process of improvement considers both perspectives.
Table 1: States of Causal Nodes

<table>
<thead>
<tr>
<th>Causal Node</th>
<th>States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Too Many Servers (TMS)</td>
<td>True</td>
</tr>
<tr>
<td>Too Few Servers (TFS)</td>
<td>True</td>
</tr>
<tr>
<td>Too Many Customers (TMC)</td>
<td>True</td>
</tr>
<tr>
<td>Too Few Customers (TFC)</td>
<td>True</td>
</tr>
<tr>
<td>Server Too Big (STB)</td>
<td>True</td>
</tr>
<tr>
<td>Server Too Small (STS)</td>
<td>True</td>
</tr>
<tr>
<td>Customer Too Big (CTB)</td>
<td>True</td>
</tr>
<tr>
<td>Customer Too Small (CTS)</td>
<td>True</td>
</tr>
</tbody>
</table>

Table 2: States of Effect Nodes

<table>
<thead>
<tr>
<th>Effect Node</th>
<th>State</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>QL</td>
<td>QL&lt;QL&lt;QL&lt;QL</td>
<td>QL&lt;QL&lt;QL&lt;QL</td>
</tr>
<tr>
<td>QW</td>
<td>QW&lt;QW&lt;QW&lt;QW</td>
<td>QW&lt;QW&lt;QW&lt;QW</td>
</tr>
<tr>
<td>CD</td>
<td>CD&lt;CD&lt;CD&lt;CD</td>
<td>CD&lt;CD&lt;CD&lt;CD</td>
</tr>
<tr>
<td>SQ</td>
<td>SQ&lt;0</td>
<td>SQ&lt;0</td>
</tr>
<tr>
<td>SU</td>
<td>SU&lt;SU&lt;SU&lt;SU</td>
<td>SU&lt;SU&lt;SU&lt;SU</td>
</tr>
<tr>
<td>Cost</td>
<td>OK</td>
<td>Optimize</td>
</tr>
<tr>
<td>Duratin</td>
<td>OK</td>
<td>Optimize</td>
</tr>
</tbody>
</table>

The probability related to each relationship was determined by the author. The probabilities of the states of CD were evaluated as shown in Table 3. Note that the probability that CD>CD_U is not explicitly shown because the information is redundant. It may be calculated as P(CD>CD_U) = 1 - P(CD ≤ CD_U).

The parents of CD consist of two sets of conflicting variables: CustomerTooBig/CustomerTooSmall, and TooFewCustomers/TooManyCustomers. These states cannot be true at the same time. Therefore, where the combination of the states of the parents indicate that they are both true, the probability assigned to that combination is the same as if they were both false. The strategy results in no clear decision based upon conflicting states of the parents.

At the end of the simulation run, statistics are extracted and the performance indices are calculated for each queuing location in the simulation model. The value of each index is compared to its user-defined limits and then it is entered as evidence to the belief network by setting the appropriate state of the Effect variables to True. In addition, resource constraints, such as having only one of a specific resource or not having any alternative resources, are entered in the same manner. The lack of alternatives is modeled by setting the variables TooBig and TooSmall to False for the appropriate resource. During evaluation of the belief network, the probability of each Causal variable state is calculated using the concepts of Bayes’ Theorem in algorithms designed to solve the networks (McCabe, AbouRizk and Goebel 1998).

Table 3: Probability Assignment for Variable CD

<table>
<thead>
<tr>
<th>CTB</th>
<th>CTS</th>
<th>TFC</th>
<th>TMC</th>
<th>CD≤CD_U</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.99</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>0.10</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>0.90</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>0.99</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.70</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>0.60</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.95</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.70</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>0.20</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>0.00</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>0.60</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>0.20</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>0.99</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>0.10</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>0.90</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0.99</td>
</tr>
</tbody>
</table>

If the probability of a Causal variable = True is greater than 50%, the Causal variable gains a score of one. The evaluation scores are summed over all of the interaction locations in the simulation model. As before, there is concern that conflicting variables will compete for priority. This may occur where one customer is being served by several servers.

For example, suppose that one customer has three servers. The performance of the interaction location of the first server is acceptable in all respects i.e. all performance indices at the location are within their specified limits. Consequently, there is no Causal variable with a probability of True greater than 50%. At the second location, the belief network evaluation indicates the likelihood that CustomerTooBig and ServerTooSmall are greater than 50%. The third server location results in the variables CustomerTooSmall and ServerTooBig having a probability greater than 50%. The customer receives a score of one for each TooBig and TooSmall. Server2 has one score in TooSmall, and Server3 has one score in TooBig. By observation, one may conclude that the customer should not be changed unless there is a similar problem at each of the server locations at which the customer interacts. The best action at this time would be to adjust the parameters of Server2 and Server3, and to rerun the simulation to determine the effects of the change.

McCabe

It might be expected that the belief network should be capable of avoiding conflict recommendations by evaluating the simulation performance as a whole instead
of evaluating each server-customer interaction location individually. However, the separate analysis at each location allows the method to be very general without limiting the number of interaction locations or resources that can be included in the model.

A simple heuristic was developed to handle the problem of conflicting recommendations. If any resource accumulates a score greater than zero for both variables of a conflicting set of variables, then the scores for those two conflicting states are cancelled. In other words, no change will be made to a resource if there is a conflicting evaluation related to that variable. In the above example, the result would be a recommendation to modify the servers.

After the simulation model is modified, the simulation is rerun to determine the effect that the changes have on the performance. The iterative process continues until all performance indices at each queuing location are within their specified limits, or until the process begins to oscillate. The resource configuration and the resulting performance may be stored in a database. At the conclusion of the automated process, the modeler may review the database. Based on the argument that several very good solutions is of greater value to the modeler than one optimal solution, the modeler may more closely examine the resource assignments from the, say, five lowest cost simulation runs.

It is possible that the performance from the iteration that resulted in the lowest cost meet the user-specified limits. In fact, the user-specified limits may be such that there is no resource assignment combination that would meet those limits. Again, this is not of major concern. The limits are used as a guide for the improvement process and should not be considered absolute. Although the performance index limits are used as guidelines for improvement, they will not restrict the model from working toward optimal solutions.

4 PROTOTYPE

A prototype was developed to demonstrate the model. MSBN Microsoft Belief Network® Version 1.001 was the belief network modeling and evaluation environment used for the prototype. The simulation language used was AweSim Version 2.0 by Pritsker Corporation® (Pritsker, O'Reilly & LaVal 1997). Microsoft Access® relational database was used to store the input resource assignment and corresponding output of each simulation run. All of the software was easily integrated using Microsoft Visual Basic®, which also provided the development environment for the user interface.

A unique feature that was added to the prototype is that more than one simulation model may be improved and compared to determine which results in the most effective method. Each model that is entered is considered a scenario. One scenario may differ from another by changes in user-defined performance limits for the same simulation model, changes to the availability of resources, or by development of another simulation model to represent a different construction method. For example, an earthmoving operation may be completed using scrapers and bulldozers or by using trucks and loaders. Each method would be represented by a simulation model and constitute a Scenario.

The automated method was validated by testing the prototype with queuing problems found in literature (Charmichael 1987). A more complicated model was developed to demonstrate the method. The model is an earthmoving operation with loaders, bulldozers, a weigh scale and unloading spaces as servers, and trucks as customers. Alternative resources for the loaders, trucks and bulldozers were identified. The unloading spaces could be modified only in quantity and the weigh scale resource had no alternatives and was limited in quantity to one.

The operation works as follows. A truck is loaded by an available loader and travels to the scale where it is weighed. The truck proceeds to the unloading area where it dumps its load, and returns to the loading area. The material that was left behind holds the unloading space until a bulldozer is available to move and spread the material. A schematic of the model is shown in Figure 2.

![Figure 2: Schematic of Earthmoving Operation](image)

Several constraints were put upon the simulation model itself to ensure the necessary flexibility was available for the prototype to run. For example, all information related to the quantity and capacity of the resources to be used in a simulation run is contained in a text file. The text file was modified between simulation runs and read into the model at the start of the run. The parameters of the resources, including alternative resources, were a standardized database table. These constraints were not considered to be limiting to the simulation modeler, and were specific to the software used to develop the prototype.

At the start of the program, the modeler provides the location of the simulation file(s) on the computer. The program reads the simulation file and extracts information related to resources, the queue locations, and user-defined statistics. The identification of the queuing locations in
which each resource interacts, and the acceptable limits of the performance indices are entered or verified by the user, guided by the information extracted from the simulation file. If more than one scenario is to be compared, the modeler is prompted to add the extra information before the automated improvement process is started.

When the iterative improvement process is complete, the program scans the database containing the input and output parameters from each simulation run. The lowest cost and shortest duration are presented to the modeler including the resource assignment for that iteration. In addition, iterations that did not result in lowest cost or shortest duration but did meet the performance limits are listed. As mentioned, the modeler may review the database itself to determine if any other solution is acceptable.

Figure 3 shows the resource assignments for the demonstration case. Each resource alternative has a different symbol.

![Figure 3: Resource Allocation per Iteration of Demonstration Case](image)

Figure 4 shows the corresponding cost and duration that resulted from the resource assignment. It is evident that the solution function is somewhat flat around the optimum points, and that the construction planner may find more than one of solutions acceptable.

5 CONCLUSIONS

The objective of the research was to develop an automated method of improving the performance of simulated operations. Evaluation and diagnosis of the performance was provided through a belief network. Performance indices were developed to evaluate and measure the resource performance within the simulation run. These indices were input to the belief network as evidence of the current situation. Output of the network includes diagnosis of operation performance, and recommendations for modifications to the model to improve its performance.

![Figure 4: Resulting Cost and Duration per Iteration of Demonstration Case](image)

The model presented here has several limitations. First, the simulation model itself cannot be automatically changed. This would require an understanding of the processes that are being modeled. A method may be developed if the system is limited to a specific process domain. However, this was not within the scope of the research.

The prototype was not optimized for computing time, and it took several minutes to run the demonstration case. While this is significantly less effort than what is required in manual experimentation, increased expectations in automated systems lessens the effectiveness of the prototype. It is the opinion of the author that this method may be effectively applied to process-specific simulation environments where many of the user-defined limits can be standardized to reduce input required by the modeler. The belief network can be integrated with the simulation environment and computation time related to communication between software may be saved.

Finally, the method does not allow combinations of alternatives of a single resource type to be assigned within the same model, such as three units of alternative #1 and one unit of alternative #4 to work together.

More research is required to fully automate model optimization of complex processes. The method discussed here has demonstrated a feasible method for automated improvement without limiting the modeler to a single solution.

6 ACKNOWLEDGEMENTS

The author would like to thank the Decision Theory Group, Microsoft research for the use of their software MSBN®.
REFERENCES


Carmichael, D.G., (1987) Engineering Queues in Construction and Mining, Ellis Horwood Limited, Market Cross House, Cooper Street, Chichester, West Sussex, PO19 1EB, England


Ioannou, P.G., (1989) UM-CYCLONE user's manual, Division of Construction Engineering and Management, Purdue University, West Lafayette, IN


McCabe


AUTHOR BIOGRAPHY

BRENDA MCCABE is an Assistant Professor in the Department of Civil Engineering at University of Toronto, Canada. Her research interests include probabilistic reasoning and simulation as they relate to the field of Construction Engineering and Management. She received a BSc from University of Toronto in 1994 and PhD from University of Alberta in 1997. She has eleven years of industrial experience related to land surveying, oilfield construction, utility engineering and housing development.