AUTOMATED KNOWLEDGE DISCOVERY AND DATA-DRIVEN SIMULATION MODEL GENERATION OF CONSTRUCTION OPERATIONS

Reza Akhavian
Amir H. Behzadan

University of Central Florida
4000 Central Florida Blvd.
Orlando, FL 32816-2450, USA

ABSTRACT

Computer simulation models help construction engineers evaluate different strategies when planning field operations. Construction jobsites are inherently dynamic and unstructured, and thus developing simulation models that properly represent resource operations and interactions requires meticulous input data modeling. Therefore, unlike existing simulation modeling techniques that mainly target long-term planning and close to steady-state scenarios, a realistic construction simulation model reliable enough for short-term planning and control must be built using factual data obtained from ongoing processes of the real system. This paper presents the latest findings of authors' work in designing an integrated data-driven simulation framework that employs a distributed network of sensors to collect multi-modal data from construction equipment activities. Collected data are fused to create metadata structures and data mining methods are then applied to extract key parameters and discover contextual knowledge necessary to create or refine data-driven simulation models that represent the latest conditions on the ground.

1   INTRODUCTION

For the last few decades, the Construction Engineering and Management (CEM) research community has investigated and promoted a variety of simulation, modeling, and visualization techniques for academic and training purposes. One of the earliest efforts in construction simulation research was made by Halpin (1977) who introduced a discrete event simulation (DES) package called CYCLONE. Later, other simulation systems such as UM-CYCLONE (Ioannou 1989), Micro-CYCLONE (Halpin 1990), MODSIM (Oloufa 1993), STROBOSCOPE (Martinez 1996), and Simphony (Hajjar and AbouRizk 1999) were designed. These systems enabled modelers to generate and work with more sophisticated simulation models through for instance, offering the ability of modeling resources by assigning attributes or using object oriented concepts to provide more robust interactions with the model.

Despite this, the construction industry has remained rather reluctant to fully embrace the widespread use of such advancements to improve the current practices in project planning, design, optimization, operation, and maintenance. While one may argue that the lack of proper training and solid technical background among field practitioners has mainly kept simulation modeling from large scale adaptation by the industry, perhaps an equally important reason is the fact that many existing simulation systems fail to adequately capture the depth and breadth of operations and activities that take place in a dynamic construction jobsite. This inability to model the "realities" of a dynamically changing operational setting has caused the construction industry to be hesitant over placing its full trust on simulation results for operational planning and management (Chang and Hoque 1989; AbouRizk 2010). In the absence of a systematic approach that fills the gap between pure simulation research and realities of the operations that take place in the field, justifying the benefits and added value of simulation-based decision-making and planning beyond the boundaries of academic research is not a trivial task.

From a simulation modeling perspective, it is commonly known that even the most sophisticated computer models fail to provide meaningful and reliable output if the model input is not an accurate representation of the real world phenomenon. Since most conventional simulation frameworks are designed to accept manual input from a human user (e.g. modeler), the majority of the resulting models are based on (at best) expert judgments and opinions, or data from past projects about key operational parameters (e.g. precedence logic, durations, uncertainties). This approach may prove to be adequate for settings such as manufacturing and prefabrication plants where the work environment is very much (if not fully) controlled and ambient factors are kept to a minimum. However, most construction projects especially those that require the use of heavy equipment and multiple crews and work areas take place in unstructured environments that may be hard to comprehend and formulate ahead of time. Given the schedule and cost constraints during project planning, most simulation modelers often use simplifying assumptions and predetermined design parameters to reduce complexities when building simulation models to study workflows and resource demands of a future construction project. Although this approach may streamline the modeling process, it may as well take away from the model flexibility and extensibility for use during the construction phase, negatively impact its accuracy in representing field dynamics as the project makes progress, and ultimately be detrimental to model reliability, verification, and validation (Davis 2007).

In summary, unrealistic input data and the dynamic nature of construction operations are among the most important reasons why existing simulation frameworks mostly fail to provide quality and timely output reliable enough to be readily adopted for decision-making and planning purposes. The lack of a systematic research effort in designing and validating methods and algorithms that can help provide realistic input for construction simulation models is the main motivation behind the presented work. To this end, this paper describes the latest work conducted by the authors in designing and testing a simulation framework that can communicate with the real system through a series of multi-modal sensor-based data collection and fusion methods in order to capture input data and discover contextual knowledge necessary to generate or refine data-driven simulation models of ongoing construction operations.

2 BACKGROUND

2.1 Real Time Simulation

The majority of previous work in creating real time simulation models that are dynamically updated has been conducted in scientific and engineering fields other than CEM. In particular, the paradigm of Dynamic Data-driven Application Simulation (DDDAS) that enables incorporating new data into an existing simulation model, has been used in areas such as emergency management, contaminant tracking, and enhanced chemical progress design (Douglas et al. 2004; Darema 2005; Douglas and Efendiev 2006). In addition to these DDDAS applications, Hunter et al. (2006) developed a simulation model based on inflow data aggregated over a short time interval to create an accurate estimate of the evolving state of transportation systems. In another example, a generic simulation platform for real time simulation modeling in healthcare and manufacturing applications was developed (Tavakoli et al. 2008).

Within the CEM domain, where operational dynamics and the presence of ambient factors can intensify uncertainties, however, there have been limited research to address the problem of dynamically updating simulation models as the project makes progress (Akhavian and Behzadan 2011). For instance, in a special purpose tunneling simulation model, Chung et al. (2006) used Bayesian techniques to update the distributions of input parameters. The framework did not function in real time and data was collected manually on a bi-weekly basis. Hammad and Zhang (2011) proposed a real time simulation system that took into account the required spatio-temporal resolution to enhance safety and productivity. Although they used location tracking sensors to acquire knowledge about duration of activities, using simplifying assumptions and not considering factors other than equipment positions when determining duration values limited the accuracy of the resulting simulation input data. More recently, a real time tracking and adaptive modeling framework was proposed by Song and Eldin (2012) for look-ahead scheduling of heavy construction equipment. Also, Pradhananga and Teizer (2012) suggested that their developed glob-

al positioning system (GPS)-based equipment tracking system has potential to provide a simulation model with more realistic input data. However, both studies drew inferences about duration of activities only based on GPS data logging. Although compared to static input data modeling (where there is no notion of real time data from the system) such single-modal data integration can provide more realistic insight about cycle times, they are still not inclusive enough to cover all the events that trigger start or end point of an activity. Considering these limitations, the authors used the concept of DDDAS to explore the potential of data-driven simulation and visualization for short-term decision making (Akhavian and Behzadan 2012a). Clearly, activity durations are not the only parameters that can impact the output of a simulation model; rather other attributes of the real system such as resource allocations, precedence logics, and service priorities need to be also taken into account and accurately modeled to yield the most realistic simulation output.

2.2 Automated Simulation Model Generation

To the best of authors' knowledge, there has been no systematic attempt within the CEM research community to develop simulation model generators. Previous work on this topic has been mainly limited to manufacturing and industrial engineering applications where product trajectories in a structured network of modules were used to generate adaptive manufacturing simulations (Véjar and Charpentier 2011). Son and Wysk (2001) developed an automated simulation model generator for real time shop floor control. In addition, Yuan et al. (1993) developed a DES generator for operational systems (SGOS) with applications in manufacturing activities such as fabrication, machine set-up, assembly, and part transportation.

3 SYSTEM DESIGN

Considering the unique requirements and constraints in construction simulation modeling, the authors have developed an integrated data-driven simulation modeling framework. Figure 1 shows the system design and main building blocks of this framework in which, a distributed network of sensors collects multi-modal data from field equipment. A reasoning process uses these continuous streams of collected data as well as key project identifiers (e.g. type, number of work areas) to extract model parameters and precedence logic necessary to update the corresponding simulation model. The details of this process and the underlying components will be described in the following Subsections.
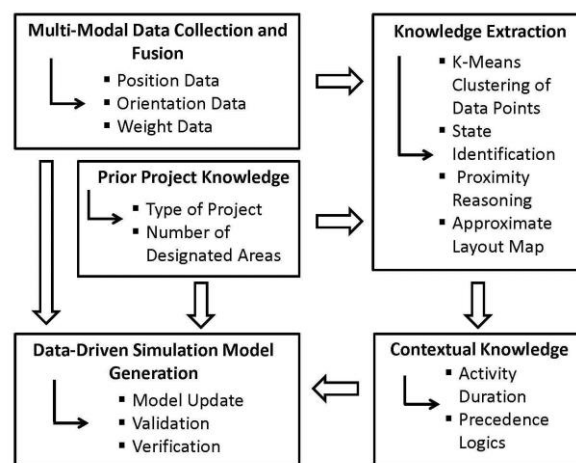

Figure 1: Main building blocks of the developed framework.

3.1 Multi-Modal Data Collection and Fusion

Data collection is a challenging task and thus, automating this process can help reduce uncertainty and improve data accuracy (Martinez 2009; Akhavian and Behzadan 2012b). In the developed framework,

multiple modes of data are captured using sensors mounted on construction equipment. In this stage of the research, a subclass of construction projects that use heavy machinery was targeted. Examples of such operations include earthmoving, paving, excavating, and mining. In all such operations, the state of each equipment can be effectively expressed using a combination of positional, angular, and weight data streaming from that equipment at any given time. Currently, and for indoor applications and proof-of-concept experiments, positional data is collected using ultra-wide band (UWB) tags and receivers. It is also important to fully capture the exact motions of articulated parts of certain construction equipment (e.g. loaders, excavators). This is due to the fact that many such equipment often work in stationary (or close to stationary) positions inside a particular work zone and therefore, positional data alone may not be a good representative of their actual operational mode. Hence, orientation trackers, and in particular, attitude and heading reference system (AHRS) sensors were used in this research to collect more meaningful data about the state of such equipment in various stages of an arbitrary operation. Also, mobile resources (e.g. dump trucks) were equipped with Zigbee-enabled weight sensors to not only facilitate the detection of start and end events of activities that affect the weight of such resources (e.g. loading and dumping), but also to track the amount of transported material. Table 1 shows sensor specifications used in this research.

Table 1: Manufacturers' specifications of sensors used to build the distributed sensor network.

Sensor Type	Key Specifications	
Load Cell	Capacity	5, 10, or 20 Kgs
	Accuracy	± 0.02 %
	Resolution	24 bit
	Update Rate	16 Hz
UWB	Accuracy	15 cm in 3D real time
	Update Rate	0.00225Hz up to 33.75Hz
	Radio Frequencies	Ultra-wideband 6GHz – 8GHz
AHRS	Roll/Pitch Accuracy	0.8° RMS
	Heading Accuracy	0.5° RMS
	Resolution	< 0.5°

3.2   Knowledge Extraction and Reasoning Process

As stated before, the main contribution of this work to the body of knowledge is that it provides means and methods to extract meaningful (contextual) knowledge necessary to automatically generate or refine simulation models that reflect the latest conditions of a real dynamic system. To achieve this, properly identifying the state of resources (herein construction equipment) is the first step. Within this general context, the state of an equipment can be defined using a binary classification of idle or busy. However, for operations-level simulation of construction activities, this classification is considered to be too generic and confusing. Figure 2 shows simple taxonomy of construction resources involved in a typical earthmoving operation. As shown in this Figure, knowing that "a truck is idle (not moving)" one may easily conclude that the truck is out of service (e.g. has a flat tire) and needs mechanical maintenance, whereas another logical conclusion could be that the truck is being loaded by an excavator and thus, is not moving. Therefore, at the operations-level, this high-level binary classification can be further broken down into a hierarchy that involves more meaningful subcategories. Figure 2 also shows subcategories for the busy state of a resource. According to this Figure, if certain types of physical motion (e.g. change in position, change in body configuration, or both) are observed in a truck or a loader, the state of the resource can be categorized as busy. The reasoning algorithm observes the trends of individual data modes and combinations of these modes that correspond to each resource state and relates this information to the knowledge required to describe project activities. Also, as far as generating a DES model is concerned,

once start and end events of an activity are determined using the knowledge extracted from raw data, activity durations can be calculated by comparing the time stamps corresponding to such events.
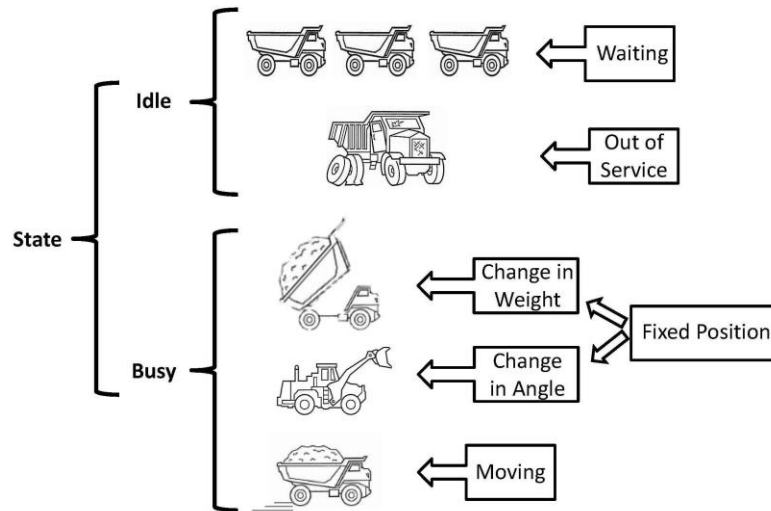

Figure 2: Taxonomy-based state classification of construction fleet.

Another important knowledge that helps in building a precise simulation model is the operational logic. In a real engineering system, there may be instances where multiple types of resources (e.g. small and big trucks) are available to start an activity. However, one resource type (e.g. big truck) may normally receive a higher priority based on constraints, demands, or resource allocation (Martinez 1996). There may be different reasons behind selecting a specific preference logic. For instance, a site engineer may decide that big trucks should be loaded first as they carry more material, and hence the job can be completed faster. When creating simulation models to describe such operations, these operational logics must be taken into account. Therefore, it is critical to develop methods to extract such knowledge and identify any potential precedence logic in the real system from the captured data. The existence of such patterns can be verified through data mining techniques that examine arrival orders of clients in different queues, or departure times and orders from queues that immediately precede active work zones. Other examples of such logics include first in first out (i.e. FIFO), last in first out (i.e. LIFO), or random service.

The third key piece of knowledge about an ongoing construction operation is the layout of the site and the arrangement of resources at any given time. To this end, tracking the position of construction equipment can reveal that the intensity of positional data points in active work areas (e.g. loading and dumping) or waiting queues is higher compared to hauling routes where equipment are transporting and thus data points have less intensity. Thus, clustering algorithms applied to such locally intense data points help detect borders of work areas and waiting queues. One such clustering algorithms is k-means which is a very popular scatter point clustering method due to its simplicity and computational efficiency (Berkhin 2006). Basically, k-means is an iterative process in which first each data point in the n-dimensional space is assigned to an arbitrary point in space representing the centroid of one cluster. Knowing the number of clusters, the goal is to minimize the squared error within each cluster. Following this minimization process, new centroids are calculated to match the means to which part of the data points were assigned. This two-step iterative process continues until readjusting means and assigning data points does not significantly change cluster centroids. In this research, given prior knowledge about the expected number of work zones, k-means is applied to the 3-dimensional (3D) data points in a 3D XYW space representing position (i.e. XY) and weight (i.e. W) data. Therefore, not only a more intense positional data points do represent a more congested area, but also having weight as an attribute of each positional data point helps distinguishing between different work areas. For example, although the loading queue and loading area are ad-

jacent to each other, in the former, the weight data shows constant close-to-zero values, while in the latter, weight is more than zero and is constantly increasing due to the dump truck being loaded.

When a front-end loader is tasked with loading a dump truck, it is already known that in two locations the intensity of points should be higher than anywhere else, and thus, streaming data points are expected to form two clusters for loading and dumping areas. While the XYW dataset is populated, the approximate centroid of these two clusters are computed using the k-means method and they are marked as loading or dumping area considering that weight values are increasing in the former and decreasing in the latter. If, in a more realistic operation, multiple dump trucks and front-end loaders are used, waiting queues are expected to form next to the loading and dumping areas. Also, in case of multiple loading areas, since the front-end loaders or excavators are sought to be equipped with position identifiers (e.g. UWB tags), given the constantly streaming weight data streams and a prior knowledge about the number of loading areas, all such loading areas can be effectively detected. It is also possible to have other areas with high-intensity data points such as a service area for periodic or random maintenance. Such complexities and many other likely scenarios require that a robust and inclusive data mining methodology is designed and used. Taking advantage of multi-modal data helps such data mining methods detect even rare conditions by providing valuable knowledge about the exact state of each equipment.

3.3 Automated Simulation Model Generation and Refinement

A simulation model generator is a tool for translating the real system logic into the simulation language, thus enabling computer to represent the behavior of the model (Mathewson 1984). In order to automatically generate or refine a simulation model, all model components such as resources, activities, variables, and their associated attributes should be either known upfront or properly linked to operational data such that their values can be computed in real time. Most such data are static characteristics that can be manually plugged in to the model. For example, the overall area and coordinates of the jobsite in which a project is taking place and locations of facilities or buildings to be constructed are typically known ahead of time and are not significantly altered during the course of a project. All other input data, that are either subject to regular (predictable) changes while the project is making progress, or cannot be estimated with enough accuracy ahead of time (i.e. uncertainties), need to be dynamically captured in real time.

To facilitate the automated generation and refinement of self-adaptive simulation models, data collected from sensors mounted on construction fleet is mined to extract information and discover operational knowledge corresponding to activity durations, precedence logics, and layout of queues and working areas. Currently, the developed simulation framework can refine pre-generated simulation models. However, this simulation model generator will be modified in the future such that as input datasets are populated, it will automatically search a library of activity cycle diagrams (ACDs) that contains preassembled networks of common fleet operations to gradually generate a simulation model from scratch. If a close match is found that best fits the collected data and extracted operational knowledge, it will be amended to the simulation model. If, however, the search returns no valid results, a new ACD will be automatically created, used for simulation model generation, and then added to the ACD library for future use.

4 PRELIMINARY EXPERIMENTS AND RESULTS

Several laboratory scale experiments were conducted at the Decision Support, Information Modeling, and Automation Laboratory (DESIMAL) at the University of Central Florida (UCF). Given the size constraints of the current test setup, one main challenge was to create layouts that allowed equipment models to freely maneuver so that clean high resolution positional data could be captured. Also, frequency interferences among several remotely controlled (RC) equipment models (due to the limited availability of RC operating frequencies) limited the ability to test certain operationally complicated scenarios. Nevertheless, the experiments were all designed while considering aspects such as generalizability and scalability. In particular, in each experiment large enough datasets were used to resemble real word scenarios. Also, all experiments can be scaled to large scale settings for future implementation, with few modifications to the structure of the algorithms and methods. For example, in a real world setting, GPS units and equipment

onboard instrumentation (OBI) can be readily employed to collect positional and payload data. However, for outdoor workspaces, more robust data cleaning and filtering techniques may still be needed to reduce and potentially eliminate ambient noise. In these experiments, using the developed data mining and reasoning algorithms, operational knowledge were extracted. More details about two such experiments are provided in the following Subsections. The test setup consisted of a 12 m^2 model jobsite and remotely controlled (RC) construction equipment. Positional data of dump trucks were captured using a network of UWB receivers and tags, while loader boom motions were sensed by an AHRS tracker. Zigbee-enabled weight sensors were also used to track the amount of material transported by each dump truck.

4.1 Experiment 1: Extracting Activity Durations

This experiment consisted of multiple dump trucks and thus, it was expected that queues would form in the vicinity of loading and dumping areas. One front-end loader was tasked with loading three (two big and one small) dump trucks, one at a time. In addition, a designated service area was added to the site layout to test if the developed methodology could properly distinguish queues from this service area with no prior information about the location of the service area. As a result, it was expected that five clusters representing loading area, dumping area, loading queue, dumping queue, and service area were detected from the streaming fleet data. Therefore, the k-means algorithm was applied on the collected dataset assuming k = 5. Figure 3 shows the layout of the experiment. The plots of collected positional and weight data in 2D (XY) and 3D (XYW) spaces is shown in Figure 4 where the developed k-means algorithm successfully detected all five clusters and determined their centroids. The results are presented in Table 2.
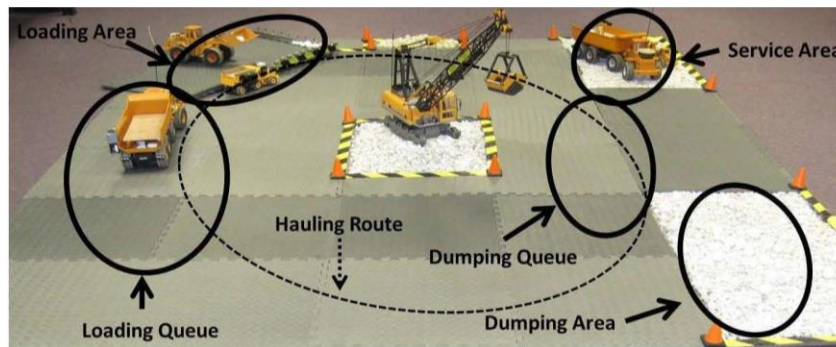

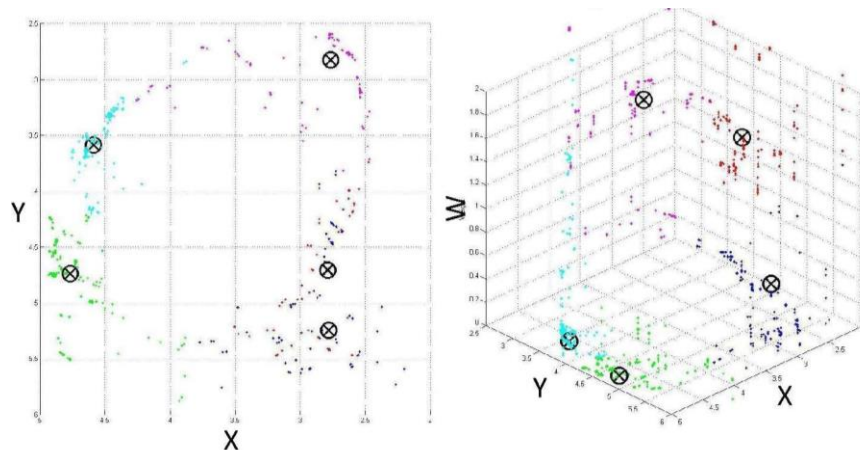Figure 3: Layout of the model construction jobsite used in proof-of-concept experiments.

Figure 4: Results of k-means clustering algorithm applied to the positional and weight data in 2D (XY) and 3D (XYW) spaces.

Table 2: Centroids of the detected clusters using k-Means for experiment 1.

Identified Cluster	X (m)	Y (m)	W (g)
Cluster 1	2.75455	4.42422	1.460313
Cluster 2	4.60887	3.53820	0.000792
Cluster 3	2.92942	5.36794	0.030927
Cluster 4	2.76416	2.77719	1.446068
Cluster 5	4.77365	4.73571	0.001201

Considering the trend of weight data, it was found that cluster 2 (where weight values constantly increased) corresponded to the loading area and cluster 3 (where weight values constantly decreased) corresponded to the dumping area. Taking into account the discussions on Subsection 3.2, the state of each equipment at any given time was then identified. Next, activity durations were calculated using the time-stamped positional and weight data. Further statistical analysis on pools of activity durations provided means and standard deviations corresponding to the duration of each activity. Table 3 compares the observed mean and standard deviation of activity durations (by analyzing the videotape of the experiment) with durations that were approximated based on the overall site layout and resource specifications (e.g. distances of hauling and return routes, equipment speeds), and the extracted mean and standard deviation (calculated using the developed reasoning process and statistical analysis). As presented in Tables 3, the extracted durations were very close to the real observed values, while the approximated (expected) durations were significantly different from observed values. This highlights the advantage of using the developed framework in providing more realistic input data for simulation models.

Table 3: Observed vs. extracted activity duration means and standard deviations for experiment 1.

Activity	Observed Duration (Seconds)		Approximated Duration (Seconds)		Extracted Duration (Seconds)	
	Mean	SD	Mean	SD	Mean	SD
Load	22.3	10.5	10.0	00.0	25.7	11.5
Haul	30.9	8.9	25.0	00.0	34.7	6.3
Dump	10.1	4.8	5.00	00.0	8.1	2.5
Return	28.2	3.6	20.0	00.0	35.0	12.2

4.2 Experiment 2: Discovering Client-Server Precedence Logic

In this experiment, the goal was to discover knowledge about the precedence logic pattern. The layout was identical to experiment 1 (see Figure 3). One front-end loader was tasked with putting soil in dump trucks. However, the time, three big and two small dump trucks operated and the front-end loader operator was asked to serve small dump trucks first (when both small and big dump trucks were available for loading). In other words, the precedence logic of the front-end loader operator in the real system was set to be "*small dump trucks get higher priority for loading*". Similar to experiment 1, positional, orientation, and weight data were collected using a distributed network of sensors mounted on equipment models.

Figure 5 shows a schematic representation of the service pattern extracted from time-stamped positional data streams. As shown in this Figure, the algorithm uses collected positional data from each dump truck to generate two time sequences corresponding to the presence/absence of small and big dump trucks (clients) in the vicinity of the loading area. Then, a third time sequence is generated that reflects which dump truck (client) was picked by the front-end loader operator (server) when both small and big dump trucks were available. This important piece of knowledge can be discovered by observing the trends of data streaming from the front-end loader (positional data) and dump trucks (weight data). In particular, when the front-end loader starts the loading process, there will be only one weight sensor (mounted on a

dump truck) that transmits constantly increasing weight values. The dump truck corresponding to this weight sensor (that can be either small or big) is the client that is being served by the server. If in fact, there is a pattern in the way clients with different attribute (e.g. size) are served, this trend can be revealed through observing several cycles of the operation. Results obtained from constantly streaming data in experiment 2 indicated that out of the 24 loading instances during which both small and big dump trucks were available in the loading area, in 21 instances small trucks received a higher priority and were loaded first while in the remaining 3 instances, big trucks received a higher priority. From the video of the experiment, it was already known that in the real system, small dump trucks received higher priority for loading in all 24 loading instances. Therefore, the reliability of the precedence logic detection algorithm in this experiment was 87.5% with 12.5% false detections. Hence, it was concluded that the operating logic of the front-end loader was "*small dump trucks get higher priority for loading*". This critical knowledge can be used inside the simulation model to yield more accurate forecast of the project performance.
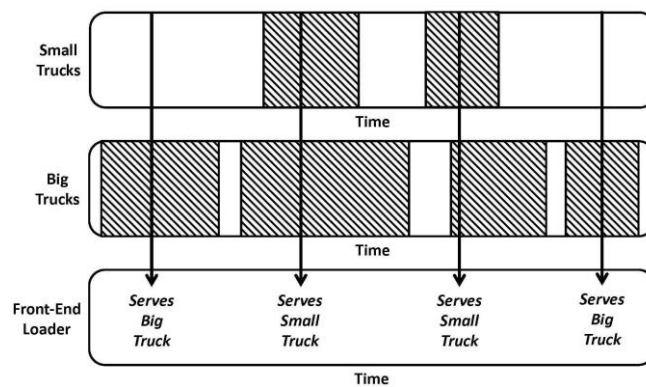

Figure 5: Loading precedence based on the availability of each truck type in the loading area.

In addition to size-based priority that clients may receive from a server, there may be other types of service patterns (a.k.a. disciplines) such as FIFO, LIFO, or random order (Martinez 1996). Such patterns should also be detected from the constantly streaming data. For FIFO or LIFO cases, dump truck arrival times in the loading queue and departure times from the loading queue to the loading area, as well as their weight data can be used to mark the time when a dump truck is selected by the front-end loader for loading. Essentially, right before a dump truck moves from the loading queue to the loading area, the algorithm compares the arrival times of all dump trucks in the loading queue (including the dump truck that is about to leave the queue). If the leaving dump truck has the lowest value of all arrival times, a FIFO pattern may have been used in that loading cycle. With the same token, if the leaving dump truck has the highest value of all arrival times, a LIFO pattern may have been used in that loading cycle. Observing similar (or close to similar) patterns in multiple cycles can lead to a conclusion that a specific service pattern is the underlying discipline for the entire operation. In the absence of a clear statistically significant trend in the service pattern, it can be inferred that the service pattern is random (or close to random).

4.3   Implementation in Data-Driven Simulation

In order to show the value of the extracted knowledge, three DES models were generated in Stroboscope, an open source simulation language designed for construction operations (Martinez 1996). In the first model, observed (i.e. actual) activity durations were used as input of the simulation and no additional input was provided about the existence of a serving precedence. Thus, by default, Stroboscope considered a FIFO discipline when simulating queues. In the second model, activity durations were approximated according to manual measurements taken from the site layout and equipment speeds (expert judgment). Similar to the first model, no additional information from the extracted knowledge about the precedence logic was used in the second model. In the third model, however, the extracted knowledge from streaming multi-modal data was used to generate a more realistic input for the simulation. Such input included sta-

tistical distributions describing activity cycle durations, and also the extracted client-server precedence logic (i.e. higher priority for small trucks), both resulted from the developed data mining algorithms. This precedence logic was coded in the Stroboscope script using the Discipline statement for the Haulers queue (i.e. the queue that modeled the waiting times of dump trucks immediately before they were drawn to the loading area). The predefined Size property of the dump trucks (of Hauler resource type) was used to indicate that the queue must be sorted such that small dump trucks move to the beginning of the queue no matter when they have actually arrived in the queue. Figure 6 contains snapshots of the Stroboscope input script for the third model where activity durations and queue discipline were defined.


```
QUEUE       HQ  Hauler;
DISCIPLINE  HQ  Size;
INIT        HQ  NBT BT;
INIT        HQ  NST ST;
```

```
DURATION    Load    'Normal[25.7,11.5]';
DURATION    Haul    'Normal[34.7,6.3]';
DURATION    Dump    'Normal[8.10,2.5]';
DURATION    Return  'Normal[35.0,12.2]';
```

Figure 6: Queue discipline and activity durations in Stroboscope input script.

To verify simulation results with real world output, two quantifiable measures were used: the Total Time of the operation and Production Rate in terms of the amount of transported soil per unit time. Table 4 shows a summary of verification results for the three Stroboscope models described above. According to Table 4, compared to the model generated based on expert judgments (i.e. model 2), the output obtained from the data-driven simulation model based on the extracted knowledge (i.e. model 3) is in better agreement with the results obtained from the real world operation (i.e. model 1). Hence, the data-driven simulation model provided much reliable output by taking advantage of multi-modal data constantly streaming from the construction equipment and the developed data analysis and mining algorithms. Therefore, if decisions are based upon this simulation model, more realistic results can be expected.

Table 4: Simulation results based on three different input model scripts.

Simulation Output	Input Script		
	1: Observation (Real World)	2: Approximation (Expert Judgment)	3: Data-Driven (Knowledge-Based)
Total Operation Time (min)	63.94	27.83	78.05
Production Rate (Kg/min)	46.92	107.78	38.44

5   CONCLUSIONS AND FUTURE WORK

In order to create more practical simulation models for the construction industry, it is critical to develop means and methods that facilitate the process of "realistic" generation and refinement of simulation models. To achieve this, simulation models must be provided with factual data collected from the real system. The current practice in manually generating simulation models using historical data and engineering assumptions does not take into account all the dynamics and uncertainties inherent in the operations, and thus, is not suitable for short term planning and control purposes. Therefore, the authors were motivated by this need to design and test an integrated framework that uses real time multi-modal operational data from construction equipment and data mining algorithms to translate data to meaningful contextual knowledge. The extracted knowledge (e.g. activity cycle durations, site layout, precedence logics) was used to refine a simulation model frequently based upon the latest conditions of construction resources (i.e. equipment). Some directions for future work in this research include minimizing the amount of col-

lected data so data is collected and transmitted only when a meaningful change in the real system status occurs, and enabling automated generation of construction simulation models from scratch using the extracted knowledge about key model parameters and precedence logic.

REFERENCES

AbouRizk, S. 2010. "Role of Simulation in Construction Engineering and Management." *J. Constr. Eng. Manage.*, 136(10): 1140-1153.

Akhavian, R., and Behzadan, A. H. 2011. "Dynamic Simulation of Construction Activities Using Real Time Field Data Collection." In *18th Workshop of Intelligent Computing in Engineering and Architecture (EG-ICE)*, Edited by 1-10.

Akhavian, R., and Behzadan, A. H. 2012a. "An Integrated Data Collection and Analysis Framework for Remote Monitoring and Planning of Construction Operations." *Advanced Engineering Informatics*, 26(4): 749–761.

Akhavian, R., and Behzadan, A. H. 2012b. "Remote Monitoring of Dynamic Construction Processes Using Automated Equipment Tracking." In *Construction Research Congress 2012*, Edited by 1360-1369. ASCE.

Berkhin, P. 2006. "A Survey of Clustering Data Mining Techniques." *Grouping Multidimensional Data*: 25-71.

Chang, D. Y., and Hoque, M. S. 1989. "A Knowledge-based Simulation System for Construction Process Planning." In *Proceedings of the 6th International Symposium on Automation and Robotics in Construction*, Edited by 348-355.

Chung, T. H., Mohamed, Y., and AbouRizk, S. M. 2006. "Bayesian Updating Application into Simulation in the North Edmonton Sanitary Trunk Tunnel Project." *J. Constr. Eng. Manage.*, 132(8): 882-894.

Darema, F. 2005. "Dynamic Data Driven Applications Systems: New Capabilities for Application Simulations and Measurements." In *Computational Science -- ICCS 2005*, Edited by 610-615.

Davis, W. J. 2007. "On-Line Simulation: Need and Evolving Research Requirements." *Handbook of Simulation*, John Wiley & Sons, Inc., 465-516.

Douglas, C. C., and Efendiev, Y. 2006. "A Dynamic Data-Driven Application Simulation Framework for Contaminant Transport Problems." *Journal of Computers & Mathematics with Applications*, 51(11): 1633-1646.

Douglas, C. C., Shannon, C. E., Efendiev, Y., Ewing, R., Ginting, V., Lazarov, R., Cole, M. J., Jones, G., Johnson, C. R., and Simpson, J. 2004. "A Note on Data-Driven Contaminant Simulation." In *Computational Science - ICCS 2004*, Edited by 701-708.

Hajjar, D., and AbouRizk, S. 1999. "Simphony: An Environment for Building Special Purpose Construction Simulation Tools." In *Proceedings of the 1999 Winter Simulation Conference*, Edited by H. B. N. P. A. Farrington, D. T. Sturrock, and G. W. Evans, 998-1006. ACM.

Halpin, D. W. 1977. "CYCLONE - A Method for Modeling Job Site Processes." *ASCE Journal of Construction Division*, 103(3): 489-499.

Halpin, D. W. 1990. "Microcyclone User's Manual." *Division of Construction Engineering and Management, Purdue University, West Lafayette, Indiana*.

Hammad, A., and Zhang, C. 2011. "Towards Real-Time Simulation of Construction Activities Considering Spatio-Temporal Resolution Requirements for Improving Safety and Productivity." In *Proceedings of the 2011 Winter Simulation Conference*, Edited by R. R. C. S. Jain, J. Himmelspach, K.P. White, and M. Fu, 3533-3544. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Hunter, M. P., Fujimoto, R. M., and Suh, W. 2006. "An Investigation of Real-Time Dynamic Data Driven Transportation Simulation." In *Proceeding of the 2006 Winter Simulation Conference*, Edited by F. P. W. L. F. Perrone, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, 1414-1421. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ioannou, P. 1989. "UM-CYCLONE User's Guide." *Dept. of Civil Engineering, The University of Michigan, Ann Arbor, Mich*.

Martinez, J. C. 1996. "Stroboscope: State and Resource Based Simulation of Construction Processes." Ph.D. thesis, University of Michigan, Ann Arbour, MI.

Martinez, J. C. 2009. "Methodology for Conducting Discrete-Event Simulation Studies in Construction Engineering and Management." *J. Constr. Eng. Manage.*, 136(1): 3-16.

Mathewson, S. C. 1984. "The Application of Program Generator Software and Its Extensions to Discrete Event Simulation Modeling." *IIE Transactions*, 16(1): 3-18.

Oloufa, A. A. 1993. "Modeling Operational Activities in Object-Oriented Simulation." *Journal of Computing in Civil Engineering* 7(1): 94-106.

Pradhananga, N., and Teizer, J. 2012. "GPS-Based Framework Towards More Realistic And Real-Time Construction Equipment Operation Simulation." In *Proceedings of the 2012 Winter Simulation Conference*, Edited by J. H. C. Laroque, R. Pasupathy, O. Rose, and A.M. Uhrmacher, 64. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Son, Y. J., and Wysk, R. A. 2001. "Automatic Simulation Model Generation for Simulation-Based, Real-Time Shop Floor Control." *Computers in Industry*, 45(3): 291-308.

Song, L., and Eldin, N. N. 2012. "Adaptive Real-Time Tracking and Simulation of Heavy Construction Operations for Look-Ahead Scheduling." *Automation in Construction*, 27: 32-39.

Tavakoli, S., Mousavi, A., and Komashie, A. 2008. "A Generic Framework for Real-Time Discrete Event Simulation (DES) Modelling." In *Proceeding of 2008 Winter Simulation Conference*, Edited by R. R. H. S. J. Mason, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler, 1931-1938. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Véjar, A., and Charpentier, P. 2011. "Generation of an Adaptive Simulation Driven By Product Trajectories." *Journal of Intelligent Manufacturing*: 1-13.

Yuan, Y., Dogan, C. A., and Viegelahn, G. L. 1993. "A Flexible Simulation Model Generator." *Computers & Industrial Engineering*, 24(2): 165-175.

AUTHOR BIOGRAPHIES

REZA AKHAVIAN is a Ph.D. student of Construction Engineering at the University of Central Florida (UCF). He received his Master's degree in Civil Engineering in 2012 from the same university. He also holds a B.Sc. degree in Civil Engineering from University of Tehran (Iran). His current research interests include data-driven simulation and visualization, field data collection and fusion, and data mining. He is a student member of the American Society of Civil Engineers (ASCE). His email address is reza@knights.ucf.edu and his web page is https://sites.google.com/site/rezaakhavian/.

AMIR H. BEHZADAN is a Wharton Smith Faculty Fellow and Assistant Professor of Construction Engineering at the University of Central Florida (UCF). He received his Ph.D. in Civil Engineering (Construction Engineering and Management) in 2008 and his Master's degree in Construction Engineering and Management in 2005 both from the University of Michigan, Ann Arbor. He also holds a B.Eng. degree in Civil Engineering from Sharif University of Technology (Tehran, Iran). His current research interests include intelligent and autonomous construction systems, data-driven simulation and visualization, and construction process data mining. He is a member of the American Society of Civil Engineers (ASCE) and serves on the editorial board of the ASCE Journal of Construction Engineering and Management. His email address is amir.behzadan@ucf.edu and his web page is http://pegasus.cc.ucf.edu/~abehzada/.