Proceedings of the 2011 Winter Simulation Conference
S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, eds.

MULTIPLE INPUT AND MULTIPLE OUTPUT SIMULATION METAMODELING USING BAYESIAN NETWORKS

Jirka Poropudas
Jouni Pousi
Kai Virtanen

Systems Analysis Laboratory
Aalto University - School of Science
P.O.Box 11100, FIN - 00076 Aalto, Finland

ABSTRACT

This paper proposes a novel approach to multiple input and multiple output (MIMO) simulation metamodeling using Bayesian networks (BNs). A BN is a probabilistic model that represents the joint probability distribution of a set of random variables and enables the efficient calculation of their marginal and conditional distributions. A BN metamodel gives a non-parametric description for the joint probability distribution of random variables representing simulation inputs and outputs by combining MIMO data provided by stochastic simulation with available background knowledge about the system under consideration. The BN metamodel allows various what-if analyses that are used for studying the marginal probability distributions of the outputs, the input uncertainty, the dependence between the inputs and the outputs, and the dependence between the outputs as well as for inverse reasoning. The construction and utilization of BN metamodels in simulation studies are illustrated with an example involving a queueing model.

1 INTRODUCTION

Simulation is an analysis methodology where the operation of a real-world or conceptual system is imitated by generating a sample of artificial histories of the system (Law 2006). Simulation models are applied to the study of systems whose analysis would otherwise be overly difficult, expensive, or dangerous. This paper focuses on stochastic simulation where systems with internal uncertainties and random factors are studied. In particular, discrete event simulation (DES, see, e.g., Law 2006) is discussed. In a simulation model, inputs describe the factors affecting the system, i.e., system settings and configurations as well as operating environments. Simulation outputs, on the other hand, are artificial observations of the system produced by the simulation model. In a simulation study, simulations are performed with alternative values of the inputs and the observed outputs are recorded. The data are then used to draw inferences concerning the operating characteristics of the system.

A simulation model, although simpler than a real-world system, can still be complex. The repetition of simulations may be time consuming and the sheer size of simulation data sets can make them unwieldy. To avoid this inconvenience, simulation metamodels (see, e.g., Friedman 1996, Barton 1998, Kleijnen 2008) are used to represent the dependence between simulation inputs and outputs. The most commonly used metamodels are input-output mappings that project the values of simulation inputs to the expected values of outputs. They include, e.g., regression models (Kleijnen 2008), spline models (Barton 1998), neural networks (Fonseca, Navaresse, and Moynihan 2003), Kriging models (Ankenman, Nelson, and Staum 2010), response surfaces (Kleijnen and Sargent 2000), and game theoretic models (Poropudas and Virtanen 2010a, Pousi, Poropudas, and Virtanen 2010). There also exists other metamodeling approaches such as dynamic Bayesian networks describing the time evolution of DES models (Poropudas and Virtanen 2007,

Poropudas and Virtanen 2011) and influence diagrams used for studying the consequences of decision alternatives in simulation based decision making and optimization (Poropudas and Virtanen 2009).

In this paper, Bayesian networks (BNs, Pearl 1991) are used as probabilistic multiple input and multiple output (MIMO) simulation metamodels. A BN metamodel is a representation for the joint probability distribution of random variables associated with simulation inputs and outputs. The BN metamodel is used to calculate marginal and conditional probability distributions as well as expected values and other descriptive statistics related to the inputs and the outputs. That is, the complete probability distributions are modeled – not just expected values as is the case with the existing metamodels. The BN metamodel enables various what-if analyses that are used for studying the marginal probability distributions of the outputs, the input uncertainty, and the dependence between the inputs and the outputs. Additionally, the BN metamodel is used to examine the dependence between the outputs and perform inverse reasoning. As far as the authors know, such analyses are beyond the scope of the existing MIMO simulation metamodels.

Unlike many existing metamodels, BNs are non-parametric, i.e., they do not involve assumptions about the forms of the probability distributions. On the other hand, the accurate construction of the BN metamodels necessitates large data sets, which limits their utilization in the context of expensive simulations. In addition, the BN metamodels discussed involve only discrete random variables and, therefore, they cannot be used for prediction of outputs similarly to, e.g., regression metamodels. Yet, the BNs allow effective what-if analyses which could be time consuming if conducted based on raw simulation data. Overall, the BNs offer a flexible approach to MIMO metamodeling. Furthermore, the construction and utilization of the BNs are aided by numerous available BN software (e.g., Andersen, Olesen, and Jensen 1990, Decision Systems Laboratory 2010).

As stated above, BN metamodels enable the modeling of uncertainty related to simulation inputs. Here, the input uncertainty (see, e.g., Henderson 2003, Biller and Gomez 2010) refers to imperfect information regarding the values of the simulation inputs that represent, e.g., alternative structures of simulation models, different functional forms of probability distributions included in the simulation models, or unknown values of the parameters of these distributions. Such sources of variability are taken into account by including the input uncertainty in, e.g., the confidence intervals of simulation outputs. There are various approaches to the analysis of the input uncertainty (see, e.g., Barton, Nelson, and Xie 2010) such as the Bayesian method (Chick 1997, Zouaoui and Wilson 2004), bootstrapping (Barton and Schruben 1993, Cheng and Holland 1997), the delta method (Cheng and Holland 2004), and the interval based method (Batarseh and Wang 2008). One alternative is to conduct a sensitivity analysis for the simulation inputs (see, e.g., Law 2006). In this paper, the input uncertainty is analyzed by performing simulations with different values of the inputs and assigning prior probability distributions to the values of the inputs. The joint probability distribution of the inputs and the outputs is then represented with a BN metamodel. This enables the calculation of the probability distributions of the outputs for both fixed and uncertain values of the inputs which allows one to distinguish the effect of the input uncertainty from the inherent randomness of the simulation model (see, e.g., Barton, Nelson, and Xie 2010).

The paper is structured as follows. A short introduction to BNs is given in Section 2. The process for constructing BN metamodels based on simulation data is also presented. Section 3 introduces the utilization of such metamodels in simulation studies. The analysis capabilities are further explored and illustrated in Section 4 by presenting an example analysis of a queueing model. Finally, the paper is summarized and conclusions are given in Section 5.

2 CONSTRUCTION OF BAYESIAN NETWORK METAMODELS

A Bayesian network (BN, Pearl 1991) is a probabilistic model that presents the joint distribution of a set of discrete random variables on three levels: relational, functional, and numerical. On the relational level, the BN is a graphical representation of dependencies between the random variables using nodes and arcs. Conditional probability distributions and the chain rule (Pearl 1991) are used for the functional representation for the joint distribution of the variables. Finally, efficient algorithms (e.g., Jensen 2001,

Neapolitan 2004), are used to calculate marginal and conditional probability distributions of the variables on the numerical level. The definition of a BN is based on a directed acyclical graph consisting of chance nodes that represent the random variables. Each chance node of the BN has a discrete set of values and a probability table that contains the conditional probabilities of these values when the values of the parents of the node, i.e., the nodes with an arc pointing to the node of interest, are known.

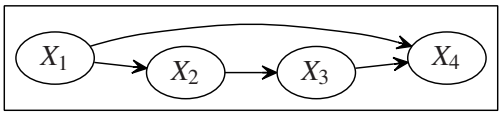


Figure 1: Example BN representing the joint probability distribution of four random variables X_1, X_2, X_3, and X_4.

For example, Figure 1 illustrates a BN representing the joint probability distribution of four random variables X_1, X_2, X_3, and X_4, i.e., the probabilities $P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$ for all possible combinations of values x_1, x_2, x_3, and x_4. The arcs of the BN show the dependencies between the variables. The BN also includes conditional probability tables consisting of the conditional probabilities $P(X_1 = x_1)$, $P(X_2 = x_2|X_1 = x_1)$, $P(X_3 = x_3|X_2 = x_2)$, and $P(X_4 = x_4|X_1 = x_1, X_3 = x_3)$ for all possible combinations of values x_1, x_2, x_3, and x_4. Note that there is no direct dependence between X_2 and X_4.

Next, the construction of a BN metamodel representing the joint probability distribution of simulation inputs and outputs is discussed. The construction consists of collection of simulation data, determination of the network structure, estimation of conditional probability tables, and validation. The determination of the structure and the estimation of the conditional probability tables are assisted by available BN software, such as HUGIN (Andersen, Olesen, and Jensen 1990) and GeNIe (Decision Systems Laboratory 2010), that include graphical user interfaces for inclusion of expert knowledge as well as readily implemented algorithms for analysis of data.

Simulation data are collected by performing simulations according to a suitable experimental design (Kleijnen 2008). BN metamodels represent discrete random variables and, therefore, continuous input variables are discretized. The uniform discretization is the most straightforward but also other discretization techniques, e.g., emphasizing the most interesting values of the variables, can be employed. On the other hand, some variables may be worthy of more detailed description and a finer discretization grid. Then, all combinations of the values of the input variables with positive probabilities are simulated and the corresponding values of outputs are recorded into the simulation data set.

The structure of a BN metamodel is determined in two phases. First, the initial structure of the BN is obtained by including the apparent dependencies into the network, i.e., by connecting the interdependent nodes with arcs. The initial structure determines dependencies between the inputs and obvious dependencies between the inputs and the outputs. Then, the simulation data are searched for dependencies between the variables using statistical tests and various algorithms (e.g., Spirtes, Glymour, and Scheines 2001) that are available in BN software. If significant dependencies are found, the initial structure is refined with additional arcs (Heckerman 1997). The structure of the BN can also be determined entirely based on the simulation data without specifying any dependencies initially but the inclusion of the known dependencies simplifies the construction. If there are numerous potential dependencies to be tested, the multiple comparisons needed for testing them all are time consuming and may lead to incorrect findings. Furthermore, the structure of a BN representing a given set of dependencies is not necessarily unique and more easily interpretable structures are obtained by designing the network using expert knowledge. On the other hand, it may be informative to initially omit all the arcs between the outputs and augment the network with additional arcs if such dependencies are evident in the data.

Once the structure of the BN is determined, conditional probability tables are defined. The input uncertainty is described by the probability distributions of the inputs. These distributions cannot be estimated from the simulation data and they are defined using other means. Any means for simulation input modeling are applicable as the distributions can be based on real-world data, expert assessment,

or combination of the two. For independent inputs, marginal probability distributions are defined. If an input depends on other inputs, a conditional probability distribution is defined for all combinations of the values of other inputs. If there are any excluded combinations of inputs, i.e., combinations with zero probability, the corresponding simulation replications need not be performed in the collection of data. For the simulation outputs, conditional probability tables are calculated from the simulation data using maximum likelihood (ML) estimators, i.e., the relative frequencies of the observed combinations of the values (Poropudas and Virtanen 2011).

Finally, the constructed BN metamodel is validated in order to ascertain that it gives an adequate representation for the simulation model. The validation is performed by comparing the probability distributions given by the BN with those estimated from an independent simulation data set. The comparison is based on the χ^2-test for goodness of fit as well as on the confidence intervals of the probability estimates. In practice, the validation follows the same principles as the validation of dynamic BN metamodels detailed in Poropudas and Virtanen 2011.

3 UTILIZATION OF BAYESIAN NETWORK METAMODELS

Now, the utilization of BN metamodels is introduced by discussing a BN representing the joint probability distribution $P(X_1 = x_1, \ldots, X_n = x_n, Y_1 = y_1, \ldots, Y_m = y_m)$ where $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ denote simulation input variables and output variables, respectively, and $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ are their values. The BN is used for probabilistic inference regarding marginal and conditional probability distributions of the variables as well as joint distributions of subsets of the variables. This inference is based on what-if analyses concerning the marginal probability distributions of the outputs, the input uncertainty, the dependence between the inputs and the outputs, and the dependence between the outputs as well as inverse reasoning.

The marginal distribution of an output given by the BN, e.g., the probabilities $P(Y_j = y_j)$, gives a complete description of the output of the simulation model. It is used to calculate, e.g., the expected value of the output, i.e., $E(Y_j) = \sum_j y_j P(Y_j = y_j)$. The BN metamodel can also be used to calculate descriptive statistics other than expected values. These statistics include, e.g., variances, quantiles, and medians of the probability distributions. The accuracy of the estimates of probabilities and expected values obtained with the BN is assessed using confidence intervals that are based on the asymptotic normality of ML estimators (Poropudas and Virtanen 2011).

The joint probability distribution of the inputs, i.e., $P(X_1 = x_1, \ldots, X_n = x_n)$, reflects the input uncertainty. The effects of this uncertainty are studied by comparing the distribution of the outputs with fixed and uncertain values of the inputs. In practice, the impact of the input uncertainty is revealed by the conditional distribution where the value of the input is fixed, e.g., $P(Y_j = y_j | X_i = x_i)$, and the marginal distribution of the output, e.g., $P(Y_j = y_j)$. Now, the difference between these distributions is an explicit representation of the consequences of the input uncertainty.

The joint distributions of subsets of variables, e.g., $P(X_i = x_i, Y_j = y_j)$, include information about the dependencies between the inputs and the outputs. These dependencies are studied by calculating the conditional distributions of the outputs, e.g., $P(Y_j = y_j | X_i = x_i)$ or $P(Y_j = y_j | X_i = x_i, X_k = x_k)$, for different values of the inputs. Also, one can calculate conditional expected values, e.g., $E(Y_j | X_i) = \sum_h y_h P(Y_j = y_h | X_i = x_i)$ or $E(Y_j | X_i, X_k) = \sum_h y_h P(Y_j = y_h | X_i = x_i, X_k = x_k)$, and present them as a function of the inputs. This inspection is congruent with regression metamodels (e.g., Kleijnen 2008) that map the values of the inputs to the expected values of the outputs.

In addition to these analyses, interdependencies between the simulation outputs are examined by calculating the conditional probability distributions of the outputs, e.g., $P(Y_j = y_j | Y_\ell = y_\ell)$ and $P(Y_j = y_j | X_i = x_i, Y_\ell = y_\ell)$. Similarly, conditional expected values such $E(Y_j | Y_\ell) = \sum_h y_h P(Y_j = y_h | Y_\ell = y_\ell)$ and $E(Y_j | X_i, Y_\ell) = \sum_h y_h P(Y_j = y_h | X_i = x_i, Y_\ell = y_\ell)$ can be calculated in order to display the average dependence between the outputs.

With BN metamodels, one can also employ an inverse approach where the conditional probability distributions of the inputs, e.g., $P(X_i = x_i | Y_j = y_j)$, are studied. In this reasoning, the probabilities $P(X_i = x_i)$

correspond to the prior distribution of an input. Then, the conditional probabilities $P(X_i = x_i|Y_j = y_j)$ yield its posterior distribution, i.e., its updated probability distribution after the value of Y_j has been observed.

4 EXAMPLE ANALYSIS

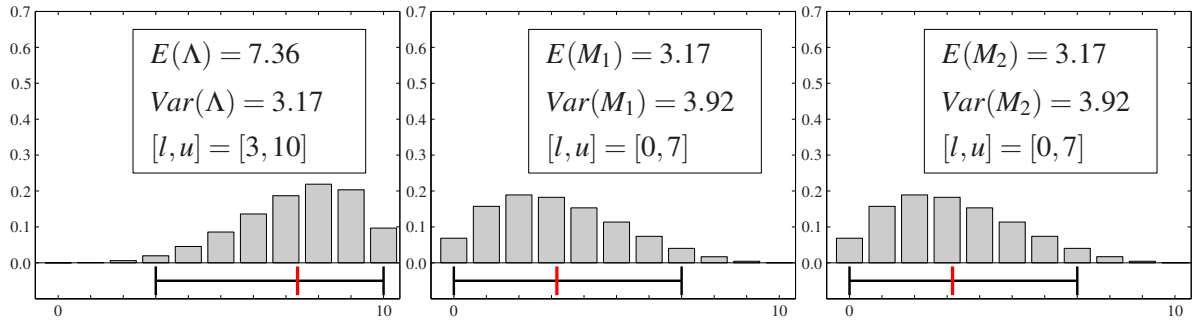
The construction and utilization of BN metamodels are illustrated with an example related to a queueing model (e.g., Law 2006). A BN metamodel is constructed based on simulation data and analyzed following the principles presented in Sections 2 and 3 using the GeNIe software (Decision Systems Laboratory 2010).

The simulation model under consideration represents a single queue with Poisson arrivals (intensity Λ) and two servers with exponential service times (service intensities M_1 and M_2). The intensities related to the arrivals and the service times are the inputs of the simulation model. Now, they are treated as random variables whose values are determined at the beginning of a day, i.e., the intensities change from day to day but remain constant during any given day. In the example, three alternative cases are studied:

- Case 1: The inputs have known values, i.e., $\Lambda = 7$, $M_1 = 3$, and $M_2 = 3$.
- Case 2: The inputs are independent random variables with the marginal distributions given in Figure 2.
- Case 3: The inputs are dependent random variables with same marginal distributions as in Case 2 and the conditional distributions illustrated in Figure 3.

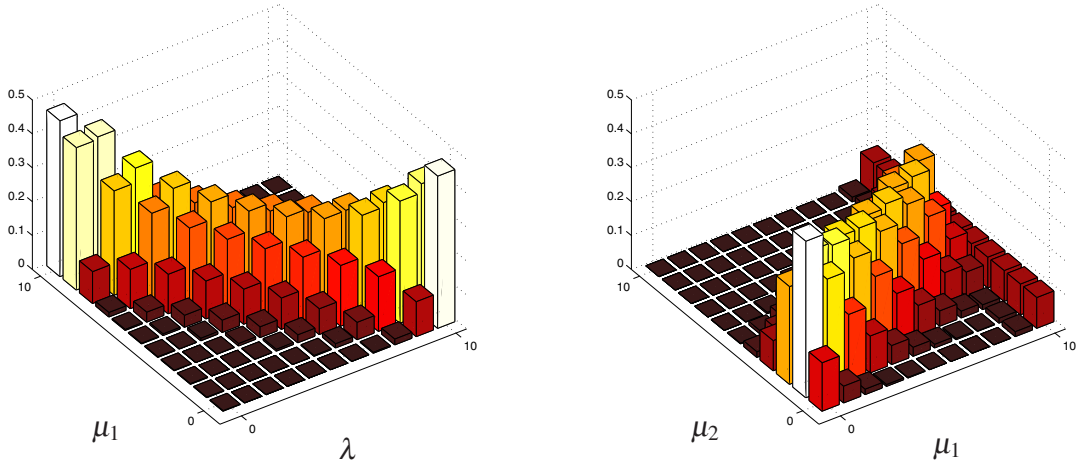
Note that Case 1 is a special case of Case 2 where the probability associated with a single combination of values of the inputs is equal to one.

In the example, the ranges of the inputs are set as $[0,10]$ and they are discretized uniformly so that $\lambda, \mu_1, \mu_2 \in \{0,1,\ldots,10\}$. For each combination of the values of the inputs, 10000 simulation replications are performed. The suitable number of replications is detailed in Poropudas and Virtanen 2011. In each replication, the queue is simulated for $T = 1$ time unit and the values of the simulation outputs are recorded. The outputs are the average number of customers in the system, the maximum number of customers in the system, and the number of customers in the system at the end of the simulation, denoted by $\bar{Y}$, $Y_{\max}$, and Y_T, respectively. The simulation outputs $Y_{\max}$ and Y_T are discrete and, after performing the simulations, their ranges are set as $y_{\max} \in \{0,1,\ldots,15,17.8\}$ and $y_T \in \{0,1,\ldots,15,17.9\}$. The output $\bar{Y}$, on the other hand, is a continuous variable. For the construction of the BN metamodel, it is discretized so that $\bar{y} \in \{0,0.5,1.5,\ldots,12.5,13.4\}$ where the values refer to the middle points of the uniformly spaced discretization bins. Note that the bins containing the highest values are represented by the mean of the observations falling into the bin.



(a) $P(\Lambda = \lambda)$. (b) $P(M_1 = \mu_1)$. (c) $P(M_2 = \mu_2)$.

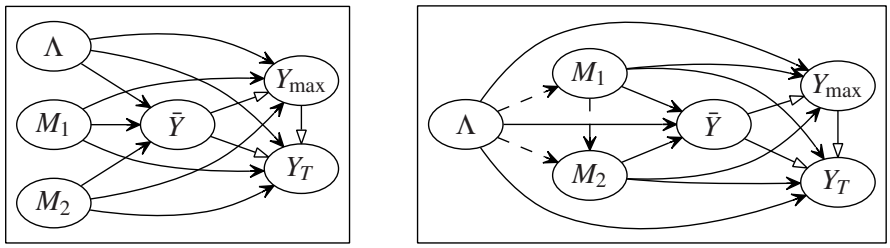
Figure 2: Marginal probability distributions of the inputs Λ, M_1, and M_2 in Cases 2 and 3. The red markers denote the expected values of the variables and the horizontal error bars denote the range $[l,u]$ between the distributions' 2.5% and 97.5% quantiles.



(a) $P(M_1 = \mu_1 | \Lambda = \lambda)$.

(b) $P(M_2 = \mu_2 | \Lambda = 8, M_1 = \mu_1)$.

Figure 3: Conditional probability distributions of the inputs M_1 and M_2 in Case 3.



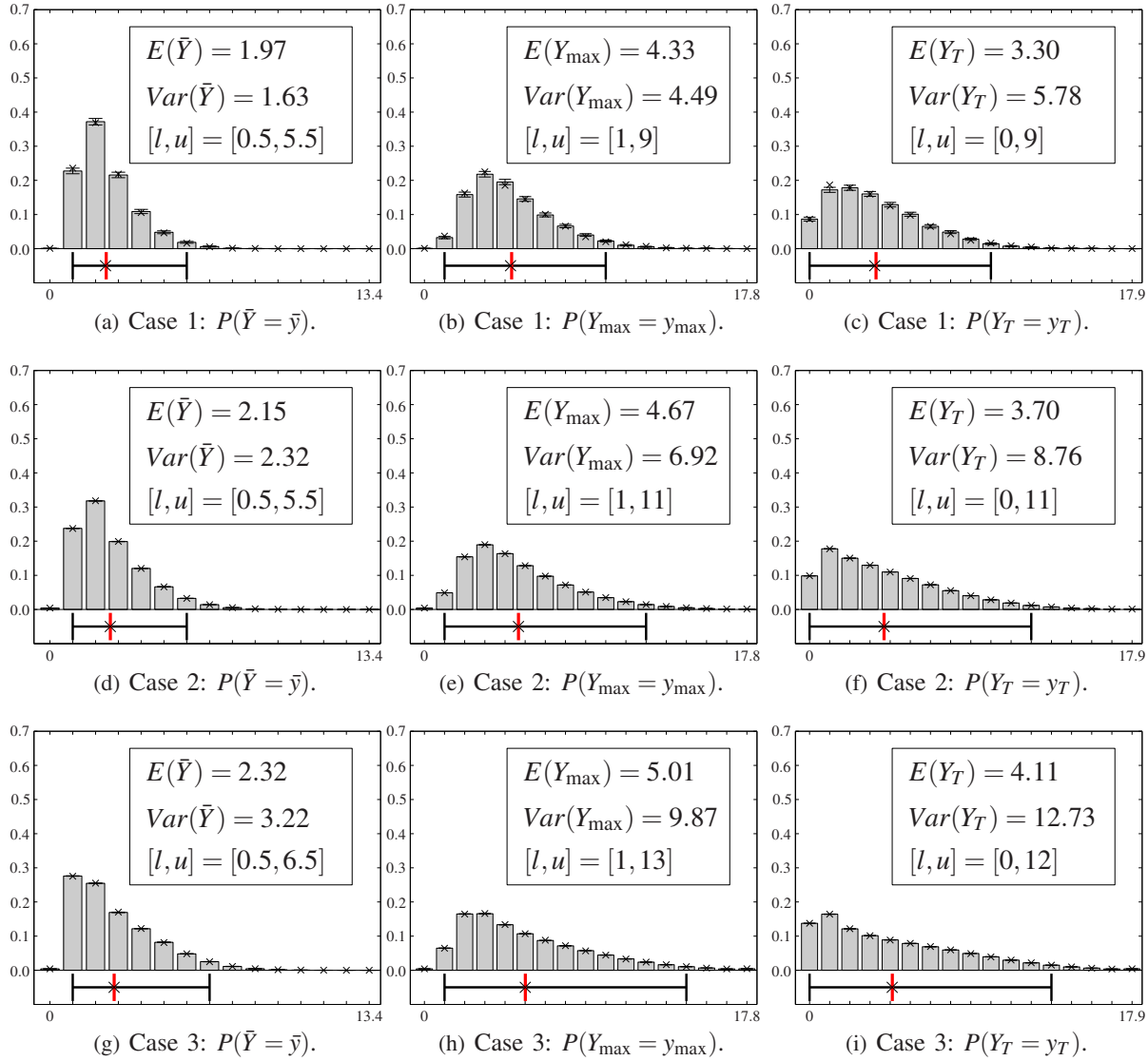
(a) Cases 1 and 2: Independent inputs.

(b) Case 3: Dependent inputs.

Figure 4: BN metamodel representing the joint probability distribution of the simulation inputs (Λ, M_1, and M_2) and outputs ($\bar{Y}$, $Y_{\max}$, and Y_T).

In the construction of the BN metamodel, the initial structure of the network is first determined by making all the outputs $\bar{Y}$, $Y_{\max}$, and Y_T dependent on all the inputs Λ, M_1, and M_2. In Figure 4, this structure is depicted by black arcs. In Case 3, there are also dependencies between the inputs illustrated by dashed arcs in Figure 4(b). The structure is finalized based on additional dependencies found in the simulation data. Statistically significant dependencies between all the outputs are discovered in the simulation data in both cases which is pointed out by white arcs in Figures 4(a) and 4(b).

After the structure of the BN is determined, conditional probability tables are defined. Now, for illustration purposes, the marginal probability distributions for the inputs (Figure 2) are assessed by the authors but they could also be estimated from real-world data. In Case 3, the conditional probability distributions of the inputs (Figure 3) are defined so that there is a positive correlation between the service intensities M_1 and M_2 while both of these variables are negatively correlated with the arrival intensity Λ. Note that conditional probability distributions similar to Figure 3(b) are determined for all values of Λ. For the outputs, conditional probability distributions are estimated from the simulation data. The validation of the BN metamodel is conducted by comparing the probability estimates given by the model with an independent simulation data set. The independent estimates are denoted by crosses in all the following figures and they match the confidence intervals obtained using the BN which is a positive finding related to the validity of the metamodel. Here, a more detailed discussion of the validation is omitted. For details of validation analysis, see Poropudas and Virtanen 2011.



(a) Case 1: $P(\bar{Y} = \bar{y})$. (b) Case 1: $P(Y_{\max} = y_{\max})$. (c) Case 1: $P(Y_T = y_T)$.

(d) Case 2: $P(\bar{Y} = \bar{y})$. (e) Case 2: $P(Y_{\max} = y_{\max})$. (f) Case 2: $P(Y_T = y_T)$.

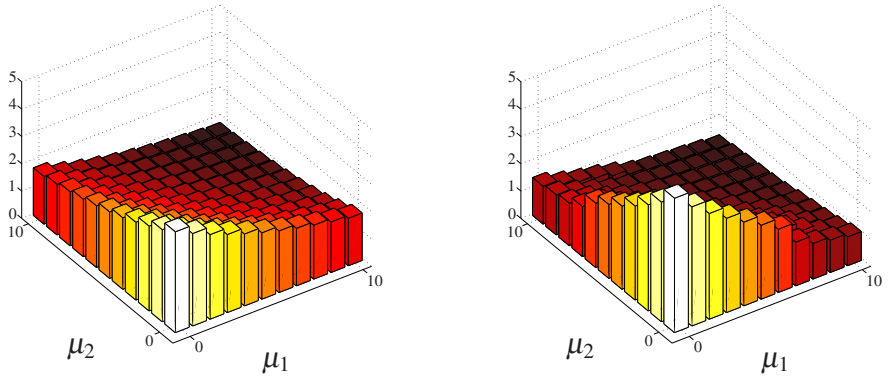
(g) Case 3: $P(\bar{Y} = \bar{y})$. (h) Case 3: $P(Y_{\max} = y_{\max})$. (i) Case 3: $P(Y_T = y_T)$.

Figure 5: Marginal probability distributions of the outputs $\bar{Y}$, $Y_{\max}$, and Y_T in the three cases. The error bars refer to the 95% confidence intervals of the individual probability estimates. The red markers denote the expected values of the variables and the horizontal error bars denote the range $[l,u]$ between the distributions' 2.5% and 97.5% quantiles. The crosses denote the estimates obtained from the independent data set in order to validate the metamodel.

Once the BN metamodel is completed, it is used to conduct what-if analyses regarding the properties of the simulated queue. Figure 5 presents the marginal probability distributions of the output variables for all the three cases. The BN retains all the available information about the distributions of the outputs which enables the calculation of any descriptive statistics. For example, the expected values, variances, 2.5% and 97.5% quantiles of the output distributions are displayed in Figure 5.

The effect of the input uncertainty is studied by comparing the probability distributions of the outputs in Case 1 with the other cases. The expected values of the distributions are smaller in Case 1 and the distributions in Figures 5(a)-(c) have smaller variances than those in Figures 5(d)-(f), i.e., the input uncertainty increases the variances of the outputs. The effect is amplified in Case 3 where the dependence

between the outputs increases the variances even further, see Figures 5(g)-(i). Additionally, the correlations between the inputs increase the expected values of the outputs slightly.



(a) Case 2: Independent inputs.

(b) Case 3: Dependent inputs.

Figure 6: Conditional expected values $E(\bar{Y}|M_1 = \mu_1, M_2 = \mu_2)$.

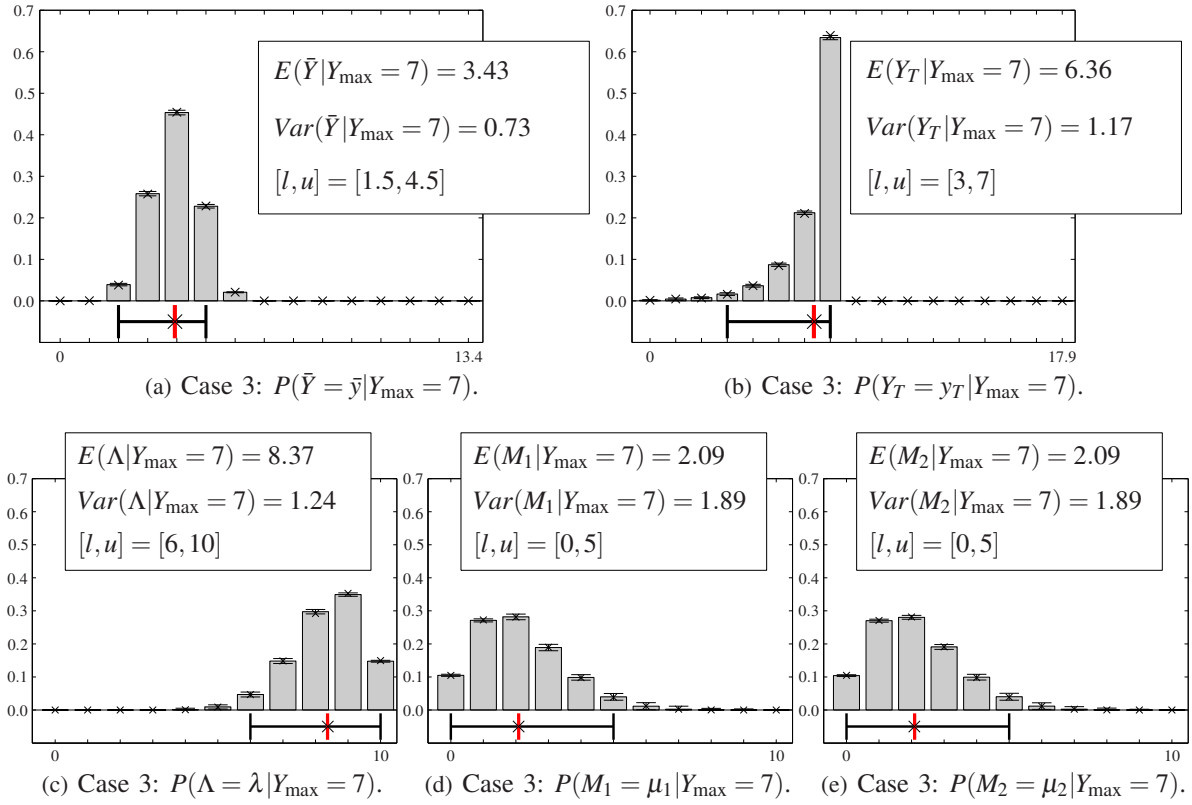
The dependence between the inputs and the outputs is further studied by calculating the conditional expected value of the output $\bar{Y}$ as a function of the inputs M_1 and M_2 for Cases 2 and 3, see Figure 6. The conditional expected values imply that the average number of customers in the system decreases with the increase in the service intensities. In Case 3, the arrival intensity Λ is dependent on the service intensities M_1 and M_2. Therefore, the dependence between $\bar{Y}$ and the service intensities is stronger, i.e., the large values of M_1 and M_2 imply smaller values of Λ which results in more drastic effects on the average number of customers compared to Case 2.

With BN metamodels, one can also examine the dependence between simulation outputs. In Figures 7(a)-(b), the value of $Y_{\max}$ is fixed at 7 and the conditional probability distributions are updated for the other output variables. The consequences of the observed maximum number of customers are studied by comparing Figures 7(a)-(b) with the unconditional distributions presented in Figures 5(g) and 5(i). The conditional distributions of $\bar{Y}$ and Y_T have both shifted to the right and their conditional expected values are larger. On the other hand, the probabilities $P(Y_T = y_t|Y_{\max} = 7)$ are equal to zero, if $y_t > 7$. Notably, the conditional variances of both outputs are smaller, i.e., the observed value of one output decreases the uncertainty about the other outputs' values.

The dependence between the outputs is further studied by calculating the expected value of $\bar{Y}$ as a function of $Y_{\max}$ and Y_T which is described in Figure 8(a). The conditional expected value $E(\bar{Y}|Y_{\max} = y_{\max}, Y_T = y_T)$ increases as the value of either one of the other outputs increases. Note also that the combinations of the output values where $Y_T > Y_{\max}$ have zero probability and they are excluded. Figure 8(b) shows a cross section of Figure 8(a) with the confidence intervals of the estimates of the conditional expected value.

The BN metamodel is also employed for inverse reasoning where the value of one or more outputs is fixed and the probability distributions of the inputs are updated. The probability distributions of the inputs conditional on the observation that $Y_{\max} = 7$ are shown in Figures 7(c)-(e). The conditional expected value of the arrival intensity is larger and the conditional expected values of the service intensities are smaller than their unconditional counterparts in Figures 2(a)-(c). Furthermore, the conditional variances of the inputs are smaller than their unconditional variances, which means that the observed value of the output reduces the uncertainty related to the inputs.

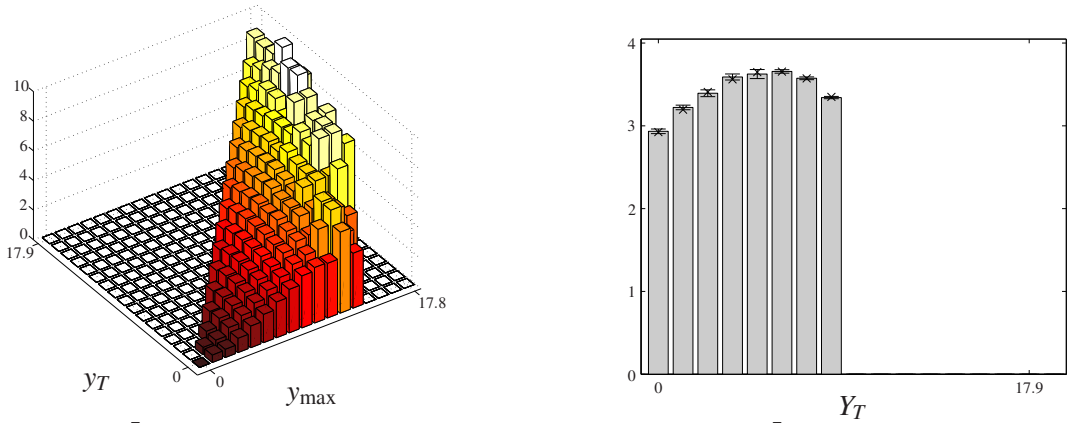
In this example, a BN metamodel is used as a MIMO metamodel for a queueing model and its analysis capabilities are demonstrated. The example illustrates that the BN can be used in variety of ways to study the queueing system using the joint probability distribution of the inputs and the outputs. Note that only a fraction of potential analyses are presented and, most importantly, with the aid of the constructed BN and the available BN software further analyses require little additional effort. For example, various



(a) Case 3: $P(\bar{Y}=\bar{y}|Y_{\max}=7)$. (b) Case 3: $P(Y_T=y_T|Y_{\max}=7)$.

(c) Case 3: $P(\Lambda=\lambda|Y_{\max}=7)$. (d) Case 3: $P(M_1=\mu_1|Y_{\max}=7)$. (e) Case 3: $P(M_2=\mu_2|Y_{\max}=7)$.

Figure 7: Conditional probability distributions of the inputs M_1, M_2 and Λ as well as the outputs $\bar{Y}$ and Y_T in Case 3 when $Y_{\max}=7$. The error bars refer to the 95% confidence intervals of the individual probability estimates. The red markers denote the conditional expected values of the variables and the horizontal error bars denote the range $[l,u]$ between the distributions' 2.5% and 97.5% quantiles. The crosses denote the estimates obtained from the independent data set in order to validate the metamodel.



(a) $E(\bar{Y}|Y_{\max}=y_{\max},Y_T=y_T)$. (b) $E(\bar{Y}|Y_{\max}=7,Y_T=y_T)$.

Figure 8: Conditional expected values of $\bar{Y}$ in Case 3. The error bars denote the 95% confidence intervals of the estimates of the conditional expected values. The crosses denote the estimates obtained from the independent data set in order to validate the metamodel.

what-if analyses can be conducted or the effects of alternative probability distributions of the inputs can be examined.

5 CONCLUSIONS

This paper presented the construction of BN metamodels based on simulation data and their utilization in simulation studies. The BNs serve as MIMO simulation metamodels that provide a non-parametric representation for the joint probability distribution of the simulation inputs and outputs. The presented example demonstrated that BN metamodels can be used for the efficient calculation of marginal and conditional probability distributions which allows analyses regarding the marginal probability distributions of the outputs, the effect of the input uncertainty, the dependence between the inputs and the outputs, and the dependence between the outputs as well as inverse reasoning. The BN retains complete information about the distributions of the simulation outputs and their interdependencies which enables flexible analyses and calculation of any descriptive statistics for the outputs. These analyses are beyond the scope of existing simulation metamodels and, without the help of BNs, they would necessitate the repeated re-screening of the simulation data requiring more computational effort than the use of the BNs.

BN metamodels allow the probabilistic modeling and analysis of uncertainty related to simulation inputs. The inputs are represented by random variables whose distributions are estimated using real-world data and/or expert assessment. Furthermore, the effects of possible dependencies between the inputs are easily quantified with BN metamodels. Note that if the joint probability distribution of the inputs is to be estimated from real-world data, all the inputs have to be observed simultaneously in order to capture their dependencies. Alternatively, independent data sets can be employed to estimate the marginal distributions of the individual inputs. Then, the dependencies between the inputs can be defined using expert knowledge and, e.g., copula models (Nelsen 2006).

Considering the input uncertainty, BN metamodels fulfill three of the four main requirements posed in (Henderson 2003), i.e., they are transparent, valid, and implementable. Unfortunately, their construction requires large data sets which may limit their applicability to expensive simulations. Therefore, one of the main topics of future research is the reduction of the number of required simulation replications by developing more advanced sampling schemes and/or discretization methods. For example, the simulation effort could be better concentrated by discretizing the most interesting inputs in full detail while using only a few alternative values for the less interesting inputs. On the other hand, sequential sampling where additional simulations are performed for some of the value combinations of inputs could be employed. The number of necessary simulations could also be reduced by using a similar approximative interpolation scheme as in Poropudas and Virtanen 2010b that would allow the prediction of conditional probability distributions of outputs related to values of inputs that have not been simulated.

Influence diagrams are an extension of BNs that have been used as metamodels in the context of simulation based decision making and optimization (Poropudas and Virtanen 2009). In future, MIMO metamodels presented in this paper could be used as a basis for multi-criteria influence diagrams utilized in simulation studies dealing with decision problems with multiple objectives. The BNs could also be used as tools for model selection and parameter estimation. In such tools, the inputs included in a BN would correspond to alternative models or values of parameters with uninformative prior distributions. Then, simulations are performed and the BN is built for some outputs that are observable in real-world. Finally, the real-world observations are fed to the BN and the distributions of the inputs are updated. The posterior distributions can then be used for comparing the likelihood of the alternative models or the values of the parameters.

To summarize, BN metamodels enable analyses that are beyond the capabilities of the existing simulation metamodeling techniques. They remove the need for the repetitive re-screening of simulation data in the estimation of conditional probabilities which expedites simulation studies. The BNs allow various alternative what-if analyses that provide additional insight to the behavior of the simulation model. Finally, the construction and use of the BN metamodels are made easy by the available BN software.

REFERENCES

Andersen, S. K., K. G. Olesen, and F. V. Jensen. 1990. *HUGIN – A shell for building Bayesian belief universes for expert systems*. San Francisco, CA: Morgan Kaufmann.

Ankenman, B., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling". *Operations Research* 58 (2): 371–382.

Barton, R. R. 1998, December. "Simulation Metamodels". In *Proceedings of the 1998 Winter Simulation Conference*, edited by D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 167–174. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barton, R. R., B. L. Nelson, and W. Xie. 2010, December. "A framework for input uncertainty analysis". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 1189–1198. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barton, R. R., and L. W. Schruben. 1993, December. "Uniform and bootstrap resampling of empirical distributions". In *Proceedings of the 1993 Winter Simulation Conference*, edited by G. W. Evans, M. Mollaghasemi, E. C. Russell, and W. E. Biles, 503–508. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Batarseh, O. G., and Y. Wang. 2008, December. "Reliable simulation with input uncertainties using an interval-based approach". In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. J. Mason, R. R. Hill, L. Moench, O. Rose, T. Jefferson, and J. W. Fowler, 344–352. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Biller, B., and C. Gomez. 2010, December. "Capturing parameter uncertainty in simulations with correlated inputs". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 1167–1177. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Cheng, R. C., and W. Holland. 1997. "Sensitivity of computer simulation experiments to errors in input data". *Journal of Statistical Computation and Simulation* 57 (1): 219–241.

Cheng, R. C., and W. Holland. 2004, Oct.. "Calculation of confidence intervals for simulation output". *ACM Transactions on Modeling and Computer Simulation* 14 (4): 344–362.

Chick, S. E. 1997, December. "Bayesian analysis for simulation input and output". In *Proceedings of the 1997 Winter Simulation Conference*, edited by S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson, 253–260. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Decision Systems Laboratory 2010. "GeNIe (Graphical Network Interface)". Available via `http://genie.sis.pitt.edu/` [accessed March 16, 2011].

Fonseca, D. J., D. O. Navaresse, and G. P. Moynihan. 2003. "Simulation metamodeling through artificial neural networks". *Engineering Applications of Artificial Intelligence* 16 (3): 177–183.

Friedman, L. W. 1996. *The simulation metamodel*. Norwell, MA: Kluwer Academic Publishers.

Heckerman, D. 1997. "Bayesian Networks for Data Mining". *Data Mining and Knowledge Discovery* 1 (1): 79–119.

Henderson, S. G. 2003, December. "Input model uncertainty: why do we care and what should we do about it?". In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 90–100. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Jensen, F. V. 2001. *Bayesian Networks and Decision Graphs (Information Science and Statistics)*. New York, NY: Springer-Verlag.

Kleijnen, J. P. C. 2008. *Design and Analysis of Simulation Experiments*. New York, NY: Springer Science+Business Media.

Kleijnen, J. P. C., and R. G. Sargent. 2000. "A Methodology for Fitting and Validating Metamodels in Simulation". *European Journal of Operational Research* 120 (1): 14–29.

Law, A. 2006. *Simulation Modeling and Analysis*. New York, NY: McGraw-Hill Science/Engineering/Math.

Neapolitan, R. E. 2004. *Learning Bayesian networks*. Upper Saddle River, NJ: Prentice Hall.

Nelsen, R. B. 2006. *An introduction to copulas*. New York, NY: Springer Science+Business Media.

Pearl, J. 1991. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Poropudas, J., and K. Virtanen. 2007, December. "Analyzing Air Combat Simulation Results with Dynamic Bayesian Networks". In *Proceedings of the 2007 Winter Simulation Conference*, edited by S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 1370–1377. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Poropudas, J., and K. Virtanen. 2009, December. "Influence Diagrams in Analysis of Discrete Event Simulation Data". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 696–708. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Poropudas, J., and K. Virtanen. 2010a. "Game Theoretic Validation and Analysis of Air Combat Simulation Models". *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans.* 40 (5): 1057–1070.

Poropudas, J., and K. Virtanen. 2010b, December. "Simulation Metamodeling in Continuous Time Using Dynamic Bayesian Networks". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 935–946. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Poropudas, J., and K. Virtanen. 2011, Nov.. "Simulation Metamodeling with Dynamic Bayesian Networks". *European Journal on Operational Research* 214 (3): 644–655.

Pousi, J., J. Poropudas, and K. Virtanen. 2010, December. "Game Theoretic Simulation Metamodeling with Stochastic Kriging". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 1456–1467. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Spirtes, P., C. Glymour, and R. Scheines. 2001. *Causation, Prediction, and Search (Adaptive Computation and Machine Learning)*. Cambridge, MA, USA: The MIT Press.

Zouaoui, F., and J. R. Wilson. 2004. "Accounting for input-model and input-parameter uncertainties in simulation". *IIE Transactions* 36 (11): 1135–1151.

AUTHOR BIOGRAPHIES

JIRKA POROPUDAS received his M.Sc. degree in systems and operations research from the Helsinki University of Technology, Espoo, Finland, in 2005, and M.Soc.Sc. degree in statistics from University of Helsinki, Helsinki, Finland, in 2011. He is scheduled to defend his doctoral dissertation in 2011 at the Aalto University School of Science, Espoo, Finland. His research interests include simulation, simulation metamodeling, and statistical analysis of basketball. His e-mail address is Jirka.Poropudas@tkk.fi.

JOUNI POUSI received his M.Sc. degree in computational engineering from the Helsinki University of Technology, Espoo, Finland, in 2009. He is currently working on his doctoral thesis at the Systems Analysis Laboratory in the Aalto University School of Science, Espoo, Finland. His research interests include multi-criteria decision analysis, simulation, and game theory. His email address is Jouni.Pousi@tkk.fi.

KAI VIRTANEN received the M.Sc. and Dr.Tech. degrees in systems and operations research from the Helsinki University of Technology, Espoo, Finland, in 1996 and 2005, respectively. He is currently Adjunct Professor at the Systems Analysis Laboratory in the Aalto University School of Science, Espoo, Finland. His research interests include optimization, decision and game theory with particular attention to aerospace applications as well as discrete-event simulation. He is the author of about 40 publications in scientific journals and conferences on these fields. His e-mail address is Kai.Virtanen@tkk.fi.