ISSN 1407-7493 **DATORZINĀTNE** **2009-7493**
COMPUTER SCIENCE

INFORMATION TECHNOLOGY AND MANAGEMENT SCIENCE
INFORMĀCIJAS TEHNOLOĢIJA UN VADĪBAS ZINĀTNE

STATISTICAL INPUT DATA ANALYSIS FOR SUPPLY CHAIN SIMULATION

STATISTISKĀ IEEJAS DATU ANALĪZE PIEGĀDES ĶĒŽU IMITĀCIJAS MODELĒŠANĀ

Galina Merkuryeva, Dr. habil., Professor, Department of Modelling and Simulation, Institute of Information Technology, Riga Technical University, 1 Kalku Str., Riga LV-1658, Latvia, e-mail: galina.merkurjeva@rtu.lv

Olesya Vecherinska, Ph.D. student, Department of Modelling and Simulation, Institute of Information Technology, Riga Technical University, 1 Kalku Str., Riga LV-1658, Latvia, e-mail: olesja.vecerinska@rtu.lv

Jonas Hatem, Master Handelsingenieur, DES, Project manager and business consultant at MÖBIUS Ltd. Kortrijksesteenweg 152, 9830 Sint-Martens-Latem, Belgium, e-mail: Jonas.Hatem@mobius.be

Keywords: simulation model, input data, normal distribution, statistical analysis, truncated distribution

***Abstract** – Stochastic simulation models utilize probability distributions to represent a multitude of randomly occurring events. Theoretical distributions are used to represent empirical data because they help smooth data irregularities that may exist due to values missed during the data collection period. The incompatibility between specific characteristics of the theoretical distribution and assumptions of simulation and mathematical calculus present an actual problem in supply chains. The paper is based on the analysis of mentioned contradictions. Different approaches to deal with theoretical probability distributions in supply chains are described in the paper.*

Introduction

Supply chain is the dynamic system where one or some parameters (lead times, customer arrivals, processing times or market demand) are varying over time. Variability of parameters is the reason why simulation is an appropriate technology in evaluating performances of supply chains. When designing the simulation model, input data containing elements of random behaviour can provide a good measure of model veracity. Simulation output is generally a function of the model input data. That is why it is essential that input data are relevant and accurate. Specific properties of input data in the model are dependent on simulation goals and requirements to calculations of system parameters.

Analysis of simulation input data presented in this paper is a part of a wider simulation-based research [1], where an approach for comparing efficiency of replenishment policies is developed. Specific objectives of this research were to estimate an optimality gap between two replenishment policies, i.e. cyclic and non-cyclic ones, as well as to analyse the influence of the different parameters of replenishment policies (coefficient of demand variation, lead-time variation, etc.) to the gap behaviour. Both theoretical background in the research area and available simulation environment are analysed. This offers a challenge to analyse properties of input data used in a supply chain.

The problem and methods of representation variability in simulation models are described. In particular, generation of the normally distributed demand with a high coefficient of demand variation could lead to 'negative demand'. To avoid the 'negative demand' problem, possible solutions are discussed in the paper and illustrated with numerical examples.

Problem Description

Within this research the simulation-based analysis is used in order to investigate the gap between performances of cyclic and non-cyclic replenishment policies in conditions of demand variability and uncertainty. The investigation is based on the multi-echelon simulation model with variable lead time of processes and stochastic demand.

Conceptually the model works with decentralized information and independent and normally distributed demand in the last echelon. The assumption about normally distributed customer demand is introduced by the following reasons:

- determination of initial values of replenishment policies' parameters (e.g., reorder point and order quantity for a non-cyclic policy, and cycle length and order up to level for cyclic one) are based on analytical calculus, where the customer demand is supposed to be normally distributed [2, 3]; and
- output data analysis based on estimation of a confidence interval requires the normal demand distribution.

The demand variability is determined as an important factor influencing a choice of replenishment policies and efficiency of their utilization [4]. That is why the coefficient of the demand variation (*CODVAR*) measured in units is the main parameter of the interest while comparing replenishment policies:

$$CODVAR = \frac{\sigma}{\mu}, \qquad (1)$$

where

σ is a standard deviation of an item's demand;

μ is an average demand for item i.

Here, μ and σ are determined based on the historical data of the real production company.

If demand standard deviation is quite large, 'negative demand' could be randomly generated in the

model. In this case, the theoretical Normal distribution cannot be directly used for random demand generation. Possible problem solutions and demand distribution alternative forms are analysed in paragraph 4.

Theoretical Background

Input Modelling

Almost all real-world systems contain one or more sources of randomness. Stochastic simulation models utilize probability distributions to represent a multitude of randomly occurring events. There are many sources that we can use to acquire input data like historical records, manufacturer specifications, vendor claims, operator estimates, management estimates, automatic data capture, and direct observation. Data collection and input data generation in the model influencing an accuracy of simulation output.

In a simulation project, the ultimate use of input data is to drive simulation. Basically, this process involves collection of input data, their analysis and generation in the simulation model. The fundamental approach to describe input data is to identify the theoretical distribution that represents input data. The possible weakness of this approach is that random numbers generated from the theoretical distribution might be unusual or incorrect in the context of the real system. This problem also occurs within presented research.

Distribution Forms

Commonly distributions are classified as *discrete* that use a finite or countable number of different values and as *continuous* with uncountable number of different values. Both classes of probability distributions used in simulation input modelling can be divided into:

- Theoretical distributions (e.g., normal, gamma);
- Empirical distributions;
- Flexible families of distributions (e.g. Johnson or Pearson distribution) [5].

Continuous distributions can be further classified according to the range of values that they can produce:

1. *Nonnegative continuous distributions* take on values in the range (a, ∞), where a is typically 0 or any positive value.
2. *Bounded continuous distributions* take on value in the range (a, b), where $a<b$ and a, b are typically positive values.
3. *Unbounded continuous distributions* take on values in the range $(-\infty, \infty)$, that is unbounded.

Theoretical distributions typically have locations and scale parameters, and zero, one or two shape parameters and are used to represent empirical data distributions because they help smooth data irregularities that may exist due to values missed during the data collection phase. A review of commonly used theoretical probability distributions in simulation is given in [5].

Theoretical distributions maintain some advantages, namely, limited use of a computer memory and an ability to change the random streams in order to perform multiple replications. Further, a sensitivity analysis can be performed easily for theoretical distributions. The assumption that a theoretical distribution gives the correct range of variability depends upon how good is an approximation of the population data. Another problem related to the range of variability is that many theoretical distributions have long tails, and there are a few occasions on which extreme values might be sampled.

Empirical distributions show the frequency with which data values or their ranges occur. They are represented by histograms or frequency charts built on the historical data. Empirical distributions can be constructed by summarizing simulation tracing data. During simulation run the data are sampled from empirical distributions by using random numbers. Empirical distributions have a number of drawbacks: (1) they can represent only bounded distributions, (2) the quality of representation is completely dependent on the quality of a sample data available, (3) the upper tail of the distribution can be unreliable for the small sample size, and (4) the probability that history repeats itself exactly is zero. For these reasons, theoretical distributions or flexible families are preferred over empirical distributions in a simulation context.

Flexible families can be considered as continuous distribution forms that are mathematically related, i.e. one distribution can be derived from another through application of mathematical transformations and taking into account context constraints. For example, the Johnson distribution family is based on transformations of the normal distribution (that is unbounded) to both bounded and nonnegative ranges of variability [5].

Selecting a Distribution Based on Sample Size

Techniques for modelling randomness in simulation are depended on a sample size of observed data [6]. For less then 20 data points a sample mean or certain theoretical distribution can be used. Larger sample sizes allow fitting the observed data to a theoretical distribution or constructing an empirical distribution (see, Table 1).

Table 1

Input modelling techniques depending on the sample size

Sample size	*Suggested input modelling methods*
Less than 20	Use a sample mean, or exponential, triangular, normal or uniform distributions
20-200	Fit theoretical distribution
More than 200	Construct empirical distribution

The Trace-Driven Simulation

A *trace* is a stream of data that describes a sequence of events. For example, the trace that describe occurrence of events in the simulation model is read during simulation runs. Typically, tracing data are stored in a data file or a spreadsheet. Traces are normally obtained by collecting data from the real system.

The trace-driven simulation is particularly beneficial for validating a model. Here, the simulation results based on the historical input data are compared with the performance measures obtained from the real system. The major drawbacks of this method are: (1) prevention of the 'what-if' simulation analysis, (2) an obvious need of the real system existence, and (3) the utilization of the unique data set, which can be out of a range of a sample data to be used in the future.

Comparison of the Input Modelling Methods

The following input modelling methods are used in a simulation:

1. Fitting a theoretical distribution;
2. Constructing an empirical distribution;
3. Using historical data;
4. Expert estimations.

Advantages and disadvantages of these methods [5] are summarised in Table 2. Ideally, a set of data points for modelling variability would be available to fitting a theoretical distribution or construction an empirical one. Otherwise, expert estimations could be used for input data modelling. In practice, the choice of a method is depend on a simulation project context and input data availability. Within mentioned project, for modelling of input data theoretical distributions are introduced.

Table 2

Comparison of the input modelling methods

Method	*Advantages*	*Disadvantages*
Fitting theoretical distribution	Smooth sample data. Generate values outside a sample range. Represent data compactly. Easy scale for a sensitivity analysis.	No theoretical distribution may fit to a sample data. Could generate inappropriate values outside a sample range.
Constructing empirical distribution	Used when no theoretical distribution fits to data.	Irregular distribution may be formed for small data samples. Cannot usually generate values outside of range of data. Difficult to scale for a sensitivity analysis. Inconvenient to incorporate large data set in simulation.
Using trace data	Efficient in model validation.	Reproduce only historical behaviour.
Expert estimations	Used when input data points are not available.	Lack of accuracy.

Problem Alternative Solutions

To deal with normally distributed demand within the described project, the following techniques are investigated, i.e.

1. Transformation of normally distributed demand,
2. Introducing truncated normal distribution, and
3. Using an alternative distribution.

From analysis of historical data, the product demand is defined by the normal distribution with the mean value $\mu = 9333$. To determine the sensitivity of the simulation output to demand variation, *CODVAR* parameter has to be changed in the range between 0.1 and 1. The initial value of the standard deviation within analysis is set to $\sigma = 4666.5$, so that $CODVAR = 0.5$.

'Normal Demand' Transformation

The iterative procedure for modelling normally distributed demand with a large coefficient of demand variation is developed within ECLIPS project by a

consortium partner Mobius Ltd [7]. It is based on the generation of normally distributed demand and its iterative transformation in Ms Excel environment.

The main steps of the procedure are following:

Step 1. Generate the normal distribution of the demand with *n* observations and known parameters μ, σ, and *CODVAR*:

$$NORMINV(RAND(); \mu\ ; \sigma\ ). \qquad (2)$$

Step 2. Replace the negative demand in the distribution generated with zero demand, i.e. define

$$MAX(x_i;0) \qquad (3)$$

Step 3. Determine estimates $\bar{x}_i, s_i, CODVAR_i$ of the distribution modified in Step 2, where $\bar{x}_i, s_i$ are an average demand and standard deviation in iteration *i*. Go to Step 7, if received estimates are approximately equal to μ and σ, else

Step 4. Perform demand calibration in iteration *i*:

$$x_i = \frac{\mu * x_{i-1}}{\bar{x}_{i-1}}. \qquad (4)$$

Step 5. Calculate estimates $\bar{x}_i, s_i, CODVAR_i$ of the distribution generated in Step 4, and stop procedure if these estimates are approximately equal to μ,σ, and *CODVAR*, else

Step 6. Modify demand distribution taking into account *CODVAR* value (see, formula (5)) and return to Step 2:

$$x_i = x_{i-2} + x_{i-2} * \frac{x_{i-2}}{\mu} * (SIGN(x_{i-2} - \bar{x}_{i-2}) * (\frac{CODVAR}{CODVAR_{i-2}} - 1)). \qquad (5)$$

Step 7. Transformed distribution is checked on normality using chi-square test, and the probability density function based on sample observations of the customer demand is built (see, Figure 1).

Steps 2-6 are repeated until a predefined number of iterations are completed. In example, the final distribution fit to the normal distribution with parameters μ and σ as $\chi^2_{\text{de facto}} = 18.97$, $\chi^2_{\text{Crit}} = 19.68$ and $\chi^2_{\text{de facto}} < \chi^2_{\text{Crit}}$,

Let note that the 'negative demand' area of the normal distribution grows with *CODVAR* tends to 1. Let's check that the transformed distribution fit to the normal distribution if *CODVAR = 1*. In example, the sample set received after the normal demand transformation has the same parameters as initial normal distribution of the demand, but $\chi^2_{\text{de facto}} > \chi^2_{\text{Crit}}$, where $\chi^2_{\text{de facto}} = 7433317187$ and $\chi^2_{\text{Crit}} = 19.68$. In this case the transformed distribution does not fit to the normal probability distribution (see, also Figure 2). Thus, it could not be used for modelling normally distributed demand.

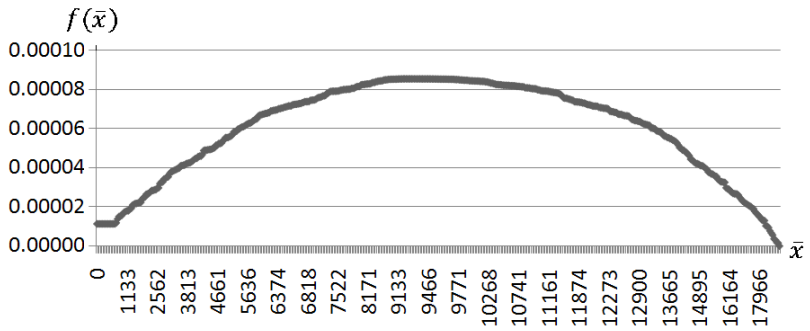

Fig.1. Probability density function of the transformed normal distribution with CODVAR = 0.5

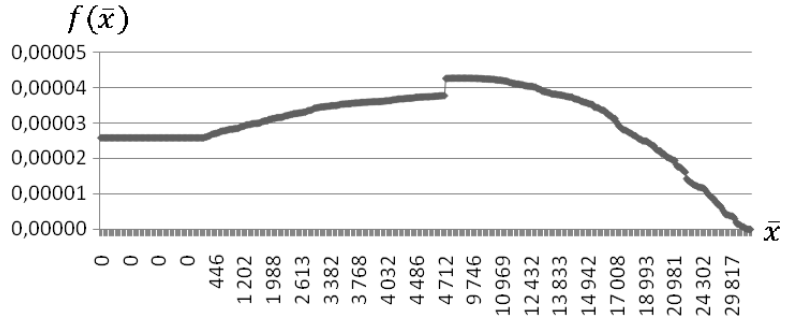

Fig. 2. Probability density function of the transformed normal distribution with CODVAR = 1

Truncated Normal Distribution

If it is known that a random variable can never take on values larger than known value *a*, it might be desirable to truncate the fitted theoretical distribution at *a* [8].

To avoid the negative values generated by the normal distribution demand, the truncation in the point zero is possible. The recalculations to determine μ and σ values of truncated in zero normal distribution are defined by (6) – (8).

$$\mu_e = \mu + c\sigma\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\mu}{\sigma}\right)^2\right), \quad (6)$$

$$\sigma_e^{\,2} = \sigma^2 - \mu_e\left(\mu_e - \mu\right), \quad (7)$$

$$c = \frac{1}{1-\Phi\left(-\frac{\mu}{\sigma}\right)}, \quad (8)$$

where

c – rationing multiplier;

μ – sample mean of truncated normal distribution;

μ_e – sample mean of initial normal distribution;

σ^2 – sample variance of truncated normal distribution;

$\sigma_e^{\,2}$ – sample variance of initial normal distribution.

Parameters μ, σ are calculated in the *MathCad 13* software. For example, if initial values of the normal distribution parameters are μ_e = *9333 and* σ_e= *4666.5*, the recalculated values of the truncated normal distribution are μ = *8855 and* σ = *5120*. As the result we receive the different values of the parameters for initial normally distributed demand and truncated one. In this case, we receive CODVAR ≠ CODVAR_e. In general, truncated normal distribution mathematically defined from initial normal distribution doesn't allow achieving the same values of μ, σ and *CODVAR*.

Comparison of probability density functions of two normal distributions with *CODVAR equal to 0.5 and 1*, and truncated distribution is illustrated in Figure 3. To determine parameters of the truncated normal distribution in *MathCad* software, initial parameters μ and σ were normalised by decreasing their values in 5000 times.

In Figure 3, stripped and doted lines represent density function of empirical distributions with patarameters (μ_e,σ_e) and *CODVAR* values equal to 0.5 and 1, correspondingly; and straight line represents probability density function of truncated in zero distribution with parameters (μ,σ).

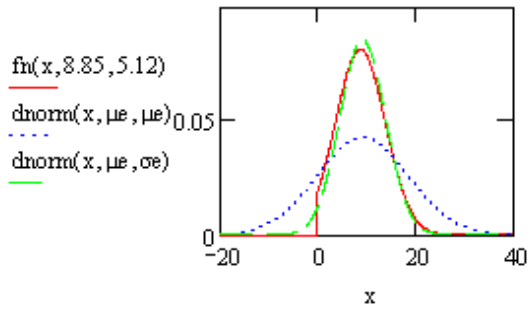

Fig. 3. Probability density functions of different normal distribution representations

Let note, wested area in the left from zero (i.e. truncated point) grows as the negative demand area of normal distribution grows when the *CODVAR* tends towards 1 and more (see, also paragraph 4.1).

Utilization of the Alternative Distribution

In practice, the lognormal distribution [8] is used when the normal distribution is not suitable because of its negative demand area. At the same time, the demand generated by the lognormal distribution is always positive. Formulas (9) and (10) provide recalculation of estimates μ_L and σ_L for the lognormal distribution if parameters μ and σ of the normal distribution are known, i.e.:

$$\mu_L = \ln\mu - \frac{\sigma_L^2}{2}, \quad (9)$$

$$\sigma_L = \sqrt{\ln(1+(\sigma/\mu)^2)}. \quad (10)$$

As a result, we receive $\text{CODVAR}_L \neq \text{CODVAR}$, where CODVAR_L is coefficient of demand variation in the lognormal distribution.

The utilization of lognormal distribution gives also the opportunity to analyse influence of a wide range of *CODVAR* values to the costs of replenishment policies in supply chains. By this, the lognormal distribution was chosen as the most appropriate to perform a sensitivity analysis in the simulation project.

Conclusion

The incompatibility between specific characteristics of the theoretical distribution and assumptions of simulation and mathematical calculus present an actual problem in supply chains. The research performed is based on the analysis of the practical problem related to the selecting an appropriate input modelling method to analyse efficiency of replenishment policies in supply chains.

The choice of a method is depend on a simulation project context and input data availability. The demand transformation procedure does not provide the

normality of transformed distribution for large CODVAR values. The distribution truncation leads to a different CODVAR value as well as its utilization in simulation environment is complicated. For the described research the normal distribution was replaced with lognormal one, and necessary mathematical calculus is provided.

Acknowledgements

The authors would like to thank Prof. A. Andronov for support and helpful advices in preparing this paper.

References

1. Merkuryeva G., Vecherinska O. Simulation-Based Approach for Comparison of (s, Q) and (R, S) Replenishment Policies Utilization Efficiency in Multi-echelon Supply Chains. UKSIM 10th International Conference on Computer Modelling and Simulation, 2008. April. - Cambridge, 2008, P. 434-440.
2. Chopra S., Meindl P. Supply Chain Management, 3rd edition. - New Jersey: Prentice Hall, 2006, 536 p.
3. Simchi-Levi D., Kaminsky P. Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies, 2nd edition. - New York: McGraw-Hill Irwin, 2003, 354 p.
4. Merkuryeva G., Timmermans S., Vecherinska O. Evaluating the 'optimality gap' between cyclic and non-cyclic planning policies in supply chains. 6th International Conference on Production Engineering, 2006. December. - Wroclaw, 2006, P.155-162.
5. Banks J. Handbook of simulation. - Hoboken: John Wiley & Sons Ltd, 1998, 841 p.
6. Greasley A. Simulation Modelling for Business. - Hampshire: Ashgate Publishing Limited, 2004.
7. Eclips Project - http://www.eclipsproject.com. - Visit date 2009. August.
8. Andronov A. M., Kopitov E.A., Gringlaz L.J. Probability Theory and Mathematical Statistics. – St.-Petersburg: Piter, 2004, 460 p. (published in Russian)

Gaļina Merkurjeva, Oļesja Večerinska, Jonas Hatem. Statistiskā ieejas datu analīze piegādes ķēžu imitācijas modelēšanā

Piegādes ķēde ir dinamiskā sistēma, kur viens vai vairāki parametri, tādi kā, piemēram, piegādes laiks, klientu pieprasījums ir mainīgi laikā. Sistēmas mainīgums ir par iemeslu, kāpēc imitācijas modelēšana ir piemērota tehnoloģija piegādes ķēžu izpētei un analīzei. Stohastiskie imitācijas modeļi izmanto varbūtiskos sadalījumus lai reprezentētu stohastiski notikušo gadījumu kopu. Bieži teorētiskie sadalījumi tiek izmantoti lai attēlotu empīriskos datus, jo tie palīdz nolīdzināt datu neregularitāti, kas var kļūst par vērtību izlaišanas sekām datu vākšanas periodā. Starpība starp teorētisko sadalījumu īpašībām un ierobežojumiem imitācijā un analītiskajos rēķinos var kļūst par aktuālo problēmu piegādes ķēžu pētījumos. Šis raksts balstās uz mīnēto pretrunu analīzi konkrētajā izpētes gadījumā. Raksta ietvaros aprakstītas iespējamas alternatīvas darbam ar teorētisko sadalījumu īpašībām.

Галина Меркурьева, Олеся Вечеринская, Йонас Хатэм. Статистический анализ входных данных в имитационном моделировании цепей поставок

Цепь поставок – это динамическая система, где один или несколько параметров, таких как, например, время доставки, спрос клиента изменяются во времени. Непостоянство системы является причиной того, почему иммитационное моделирование является эфективной технологией в исследовании и анализе цепей поставок. Стохастические иммитационные модели используют вероятностные распределения для описания множества случайно происходящих событий. Обычно, для описания эмпирических данных используются теоретические распределения, так как они способствуют сглаживанию нерегулярности данных, к чему приводит потеря определенных значений в период сбора данных. Несоответствие между свойствами теоретического распределения и ограничениями в иммитации и аналитических расчетах может стать актуальной проблемой при исследовании цепи поставок. Данная статья основывается на анализе упомянутых несоответствий в конкретном случае использования. В рамках статьи описываются возможные альтернативы при работе со свойствами теоретического распределения.