Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

PITFALLS OF THEORETICAL PROBABILITY DISTRIBUTIONS IN SIMULATION STUDIES IN THE ABSENCE OF DATA

Leonardo Rosas Leal

Cenpes
Petrobras
Av. Horácio de Macedo, 950, Ilha do Fundão
Rio de Janeiro-RJ,21941-915, BRAZIL

José Arnaldo Barra Montevechi
Mona Liza Moura de Oliveira
Carlos Henrique dos Santos

Institute of Production and Management Engineering
Federal University of Itajubá
Av. BPS, 1303, Pinheirinho
Itajubá, MG, 37500-000, BRAZIL

Tábata Fernandes Pereira

Institute of Integrated Engineering
Federal University of Itajubá
R. Irmã Ivone Drumond, 200
Itabira, MG, 35903-087, BRAZIL

ABSTRACT

Simulation studies in the absence of data are very common in Operations Research. In this context, the specialized literature frequently recommends the use of certain theoretical probability distributions to model a set of unknown input data. However, scarce and insufficient words of warning are released by the same literature concerning this practice. Several pitfalls may occur and result in immeasurable deviations from the output of the simulation, depending on the real characteristics of the input data. In this sense, the objective of this article is to present these pitfalls, their consequences, and how they can be mitigated. A roadmap is proposed as a mitigation instrument, as well as its application in a real case is shown. Thus, a literature gap on the subject is filled.

1 INTRODUCTION

The motivation for this study comes from the perception that the simulation literature does not expose all possible circumstances on input data modeling for simulation studies. This modeling occurs in two contexts: in the presence or in the absence of data. In other words, with data or with lack of data, respectively (Biller e Gunes 2010). In the absence of data, the matter is especially critical since it is impossible to measure the adherence of the models adopted, considering there is little or no data available for that as well. Even so, no established field of research on the subject is available. Therefore, it will be possible to outline and point out the existence of a literature gap on the subject of simulation studies in the absence of data.

Simulation studies in the presence of data are more frequent in literature than those in the absence of data (Biller and Gunes 2010). In the first case, which is the studies with data, a wide range of literature exists; as for the second case, which is studies with a lack of data, a theoretical development is required as well as publications about it.

Another approach on input data modeling refers to the nature of the theoretical probability distributions adopted constituting two possible branches: the continuous theoretical probability distributions modeling and the discrete theoretical probability distributions modeling (Harrell et al. 2011). This study will discuss the absence of data in the context of continuous distributions, also known as probability density functions.

However, a similar discussion is possible for discrete distributions. From now on, for the sake of simplicity, input data modeling will be designated simply as input modeling.

According to Nelson and Yamnitsky (1998), certain input models adopted in simulation fail for several reasons. Some of these reasons include: the theoretical probability distributions represent delimited and non-flexible ways to model all types of data; the input processes are not inherently independent, either in the form of a time series or concerning other input processes in the simulator; entry processes change over time; and finally, no input data is available for evaluation or selection of a probability distribution.

At this point in the discussion (regarding the absence of data), nothing is mentioned about the possibility that some basic premises in the "input data" may not exist, although the data itself is not available or accessible for analysis. The inexistence of these premises are pitfalls given the current solution that is recommended by the literature for input modeling in the absence of data: the adoption of an estimated theoretical probability distributions without further reservations. For this reason, the object of analysis of this study is dedicated to this context.

Also, the risks of adopting simple input models will be discussed in situations in which these premises do not exist. Considering this context, the objective of this paper is to present the pitfalls of theoretical probability distribution in simulation studies in the absence of data, present their consequences, and show how they can be mitigated. In summary, we will seek to answer what are the impacts of this theoretical gap in the literature, as well as fill it with a theoretical proposition through a roadmap, in addition to showing its applicability in a real case.

This paper is structured in four sections. The first section introduced the focus of this article. The second section presents a literature review on the main topics addressed in this paper. The third section describes the pitfalls, how they can be mitigated through a proposed roadmap, and a short case study to show the applicability of the roadmap. Finally, the last section presents conclusions and suggestions for future work.

2 LITERATURE REVIEW

This literature review is organized in two stages: the presentation of basic concepts in data analysis and a bibliographic review on the input modeling in the absence of data. Understanding the general concepts in data analysis is essential to comprehend the particular situation of input modeling in the absence of data.

2.1 Conceptualization in data analysis

Biller and Gunes (2010) claim that a "true" distribution for a set of data does not exist, but rather a set of possible probability distributions that most reflects the fundamental characteristics of the process it is representing. Also, the authors identify situations in which classical models for input data fail.

The input modeling can use three strategies in simulation studies: the adoption of an empirical probability distribution; the use of the real data sampled for straight use as input data in the simulation and the adoption of theoretical probability distributions in the absence or presence of data (Harrell et al. 2011). As emphasized by the authors, the last case is preferable.

Considered the utilization of the theoretical probability distribution, Biller and Gunes (2010) characterize the input modeling process in three phases: selection of a probability distribution based on physical characteristics of the process and the graphic examination of the data; the definition of the values of the input parameters; and, the verification of the adherence of the data with tests and graphic analysis.

A very common type of data used in simulation projects is time series data. Morettin and Toloi (2006) define time series as a set of observations ordered in time. It can be stationary or non-stationary. The stationary time series implies the data is stable. Thus, both the mean and the variance of a stationary time series have constant or stable values, that is, they develop in time randomly around these values, respectively. Figure 1 illustrates the stationarity (left) and non-stationarity (right) time series over time. One of the most frequent assumptions in time series is the hypothesis that a given series is stationary. However, most of the series that we find in practice show some form of non-stationarities (Morettin and Toloi, 2006).

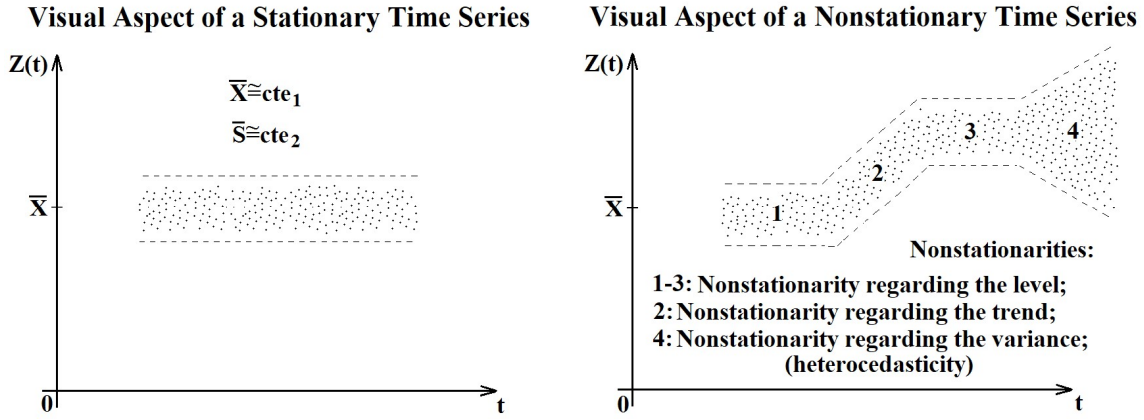


Figure 1: Stationary (left) and non-stationary (right) time series in terms of level, inclination, and variance. Source: Adapted by Morettin and Toloi (2006).

Furthermore, the adoption of theoretical probability distributions requires that the sample data have certain properties, such as independence, homogeneity, and stationarity. Independence is associated with data randomness (weak autocorrelation); homogeneity means that the data come from the same distribution (population) presenting a single mode; and stationarity means that the distribution does not vary over time (Harrell et al. 2011).

As written by Biller and Nelson (2005), the statistical models for input data incorporated in the modeling packages suggest that such data are always independent and identically distributed (iid). However, they warn of the fact that time dependence is very common in real systems studied through Simulation. Biller and Gunes (2010), on the other hand, in a more accurate analysis, identify three types of dependencies: over time, between different processes, or both. Consequently, they classify the input models of the processes that represent them in three types: univariate time series, vector time series, and random vectors.

A univariate time series represents a sequence of random variables, all of which have the same probability distribution, but can exhibit time dependence. This type of dependence is represented by autocorrelation, which is the correlation between observations within a series. For example, the monthly demand for a certain product of a retailer, the time between customer arrivals in a queue over time, etc., can be input models in the form of a univariate time series (Biller and Gunes 2010).

On the other hand, the input model presented by a vector time series represents the dependence on the same time series sequence and between different processes. For example, it can represent the dependence between the monthly demands of a certain product and the demands of a different product interacting synergistically (Biller and Gunes 2010).

Finally, a random vector is used to model each random input in a certain dimension by a different probability distribution allowing the inputs to be dependent on each other. For example, random vectors are often used as input models to analyze the demand for different products within different asset portfolios of a retailer and their annual returns. A multivariate normal distribution is another example of a random vector (Biller and Gunes 2010).

2.2 Input modeling in the absence of data

According to Chwif and Medina (2015), simulation projects in the absence of data appear in three possible contexts: the system to be modeled is new, the cost of collecting data is high, and data cannot be provided for reasons of confidentiality or security.

Historically, some strategies have been devised for input modeling in the absence of data. The most common strategies are the use of engineering and data standards; consultation with specialists (expert opinion); identification of the structural limits of variation of the data values; and the physics or nature involved in the process to be modeled (Banks et al. 2010). Common sense leads one to believe that a little of each of these approaches should be adopted in the absence of data.

As stated by Johnson and Mollaghasemi (1994), the subjective parameter estimation is the basis for input modeling in the absence of data. That is, a subjective estimation of the parameters is done to specify theoretical probability distributions completely. A more comprehensive view of data analysis with a focus on simulation can be seen in Figure 2. It is possible to observe how the input models, data, and applications interact with one another. The gray region in Figure 2 is the contextualization of the subjective parameter estimation in the triumvirate of models, data, and applications by Johnson and Mollaghasemi (1994).

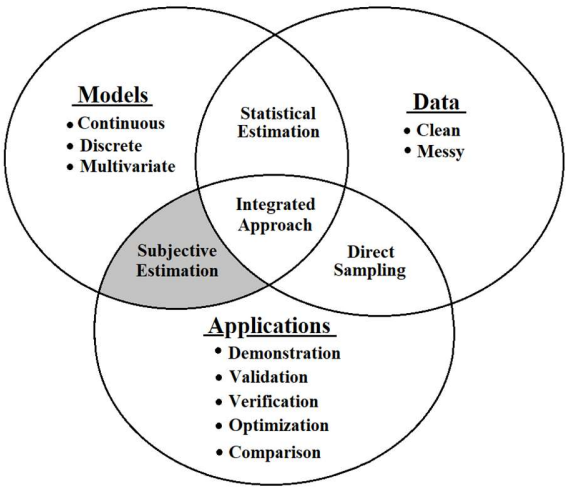


Figure 2: Models, data and applications triumvirate. Source: Adapted by Johnson and Mollaghasemi (1994).

The idea of subjective estimates has been explored for some time. DeBrota et al. (1989) developed an interactive and visual method to subjectively estimate a Limited Johnson Distribution, in cases where there is little or no data available. At this point of this research, it is clear that subjective estimation can be used to directly specify a distribution.

Nelson and Yamnitsky (1998) designed a tutorial for input data modeling when it comes to complex problems. The authors describe models and techniques that are useful when traditional models fail and mention that the absence of data is one of the reasons for these failures, but they do not propose a clear solution for this specific case.

As a result of this discussion, the apparently most widespread strategy for input data modeling in the absence of data is the consultation with specialists (expert opinion). The technique uses subjective estimates. It consists of consulting the specialist about the process that is desired to model so as to obtain quantitative and semi-quantitative information, very much based on the daily experience of the specialist (Biller and Nelson 2002). When using this strategy, the first step is to survey the upper (b) and lower (a) limits of the data set. For this, the subjective estimates of these parameters are raised with the specialists.

Several probability distributions could be specified. Among the most common ones in the absence of data are the uniform, triangular, and beta distributions. Chwif and Medina (2015) extend the possibility of also choosing an exponential or normal distribution and Law (2007) states the use of the lognormal and Weibull distribution (in the form of estimation heuristics for the last one). The uniform distribution is considered a poor choice since extreme values are hardly equiprobable from central values in a real process (Banks et al. 2010). However, if there is a certainty of low process variability, the uniform distribution can be selected according to (Chwif and Medina 2015).

According to Law (2007), if the choice is for a triangular distribution, it would be necessary to estimate another parameter named mode (m), in addition to the a and b already mentioned. If the choice is for a beta distribution, two other parameters would need to be estimated, the shape parameters α_1 e α_2, besides a and b. Some methods are suggested by Law (2007) to estimate these shape parameters, but they are not general methods for any beta distributions whichever it is.

The same author also points out that two problematic questions arise when choosing a triangular distribution: First, would the upper limit (b) of the triangular distribution be fixed indefinitely over time in the real process? Or would it be fixed for a certain period, becoming a (b') later?, see Figure 3; Second, the triangular distribution does not quite represent distributions with infinite tails on the right, as is the case with probability density functions that represent an execution time of a task, therefore limiting the flexibility of the use of this distribution.

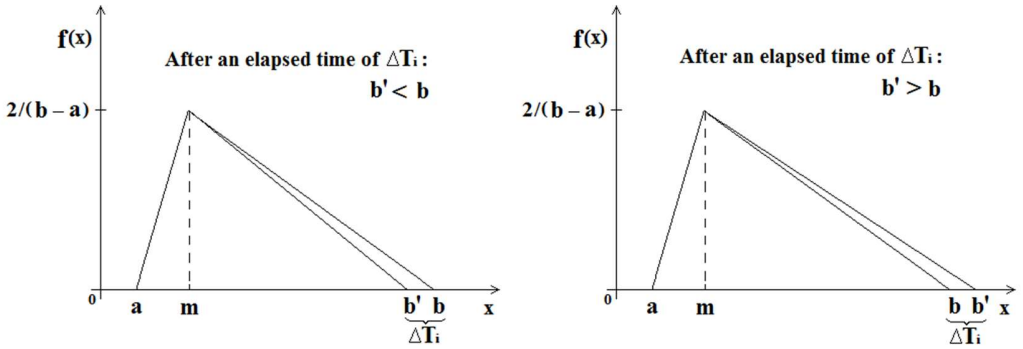


Figure 3: Triangular probability distribution in the absence of data and uncertainty about the constancy of b after an elapsed time of ΔTi, with b '<b, on the left, or b'> b, on the right.

Biller and Nelson (2002) warn of the mistake of adopting probability density functions (PDF) with long tails in cases where the data do not have this amplitude of variation. In other words, one must be careful not to adopt a PDF that draws values impossible to occur in practical terms. On the other hand, Law (2007) also warns about the opposite, which is the mistake of using a probability density function limited on the left or the right for processes that require long tails, such as using a triangular distribution to represent the activity time of a process. Usually, task runtime distributions have long tails to the right to represent delays, unless the task is performed by a machine.

Harrell et al. (2011) state that a normal distribution can be used in the absence of data when only the minimum and maximum value are known. In this case, the standard deviation is estimated to be 1/6 of the amplitude. However, Law (2016) warns of the widespread use of the normal distribution when, in fact, it rarely appears in the real world, with one exception: when the process represents the sum of a large number of other processes, which can be justified by central limit theorem (Upton and Cook 2013).

Alternatively, Cheng (2017) defends the use of a Bayesian approach for estimating input model parameters when input data is not available. The value of a parameter in a simulation run is fixed and sampled using a Bayesian distribution a priori, specified by the analyst, named $\pi(\theta/x)$. This prior distribution of the parameter θ will need to be chosen to reflect the uncertainty about the probable true value of θ. A posterior distribution $\pi(\theta/x)$ is obtained by incorporating new information. Thus, the method can be used repeatedly during an ongoing investigation, allowing additional data or the opinion of specialists to be incorporated in the updating and refinement of $\pi(\theta/x)$. As it involves a learning routine, Cheng (2017) also recommends the use of Machine Learning along with the model.

Goeva and Lam (2014) studied the problem of estimating input distributions when only output results are known. This is called Inverse Problems. In this case, the interest is in the use of simulation to emulate

the output data. According to Tarantola (2005), when using Inverse Problems is necessary to explain all the information of the model's parameters, as well as its uncertainties.

Finally, Chwif and Medina (2015) always recommend carrying out a sensitivity analysis for whatever distribution is adopted. A great variability in the output results, due to the change in the parameters of the adopted distribution, suggests that the estimated distribution should be abandoned and every effort should be made to obtain real data.

3 PITFALLS IN INPUT MODELING IN THE ABSENCE OF DATA AND WAY OF MITIGATION

Biller and Nelson (2005) warn of the fact that simulation specialists tend to ignore the serial dependencies existing in the data because there is no general and widespread method for modeling time series for input data. As a consequence, they use commercial packages to model this data, whose required premises are not present in the same data. Ignoring these dependencies can lead to very poor estimates of the performance measures of a simulation model. The authors suggest that they are debating the subject in the context of simulation studies in the presence of data. However, there is no obstacle to consider the theme in the field of the absence of data, as is the case in this study.

Despite the aforementioned, there is a broad range of literature available on time series that can serve as a guide for the treatment of this type of data. For example, Montegomery et al. (2007) presents a "classic" approach for modeling a time series by decomposing that series into trend, seasonality, and random error. This is what he calls the General Approach to Modeling and Forecasting Time Series. Leal (2018) applied this method in the treatment of data for his case study in simulation, in a set of data that showed tendency and heteroscedasticity, see Figure 4, graphs 1 to 3.

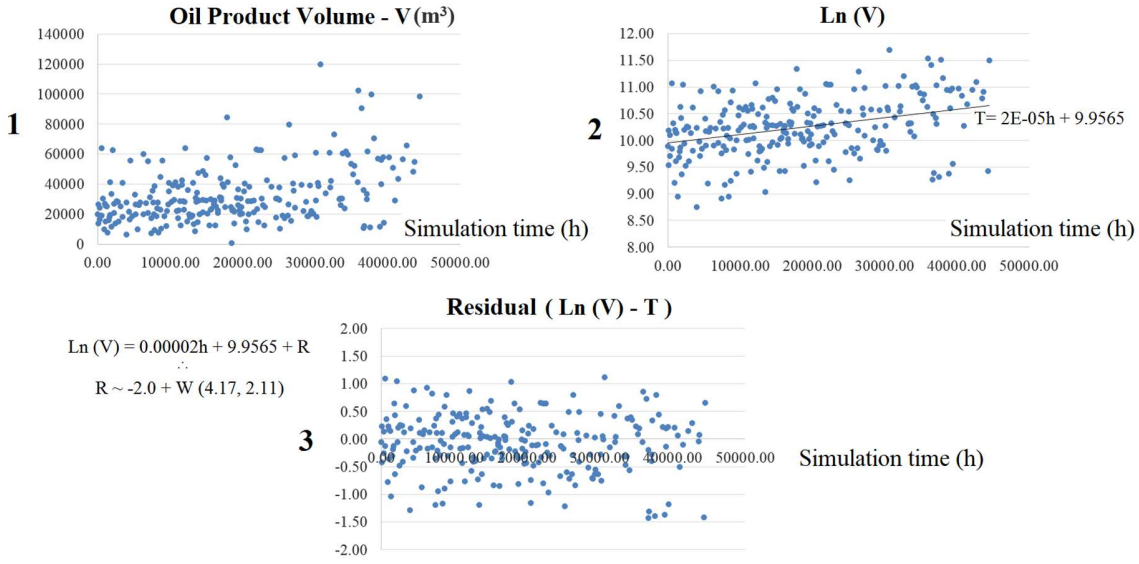


Figure 4: Decomposition method for time series of a movement of an oily product in a marine terminal. Source: Leal (2018).

Nevertheless, this method may not work well in other cases, and thus motivates the use of a more general technique for modeling time series called ARIMA, which admits autocorrelation between random errors (Montegomery et al. 2007). Therefore, the discussions corroborate with the aforementioned view that a general input modeling procedure, in addition to not currently existing, is not also trivial to formulate. It

impacts not only when the simulation study takes place in the presence of data, but also in the absence of data.

As a result of the discussion so far, the pitfalls of theoretical probability distributions in simulation studies in the absence of data can be classified into two groups: revealed pitfalls and unrevealed pitfalls. The designations themselves already show what each one represents. Despite the existence of these two groups, the object of this article is restricted to the second group as known. Nonetheless, a brief presentation of the first group already seen is made to introduce the second group afterward.

3.1 Revealed pitfalls

From a methodological viewpoint, it is interesting to mention the pitfalls listed in the literature before revealing the pitfalls that are not listed in the literature. Among the listed pitfalls, it can be said that it is important to avoid:

- The use of uniform and normal distributions (unless the normal distribution is for a process that represents a sum of several small activities);
- The use of a single probability distribution to represent the time between failures of equipment over its entire life cycle (Elsayed 2012);
- The use of a symmetrical or negative asymmetric distribution to represent the execution time of a manual task;
- The use of limited distributions, on the left or the right, when the nature of the activity requires a distribution with long tails.

3.2 Unrevealed pitfalls

As previously seen, ignoring or disregarding the inherent premises of the input data can lead to very poor estimates of performance measures. As already suggested, this point of view is also valid when referring to simulation studies in the absence of data. In this case, even if there are no data or there are data but are not accessible, one can qualitatively evaluate the set of its properties.

As already discussed, Law (2007) questions whether the adoption of a triangular distribution in a context of lack of data would maintain the same the maximum value in the real process over time. Graphically, the discussion he inaugurates can be illustrated through the already commented Figure 3. Nevertheless, the established debate is limited because of the additional situations that could occur according to the illustration of Figure 5, which is the object of discussion of this article.

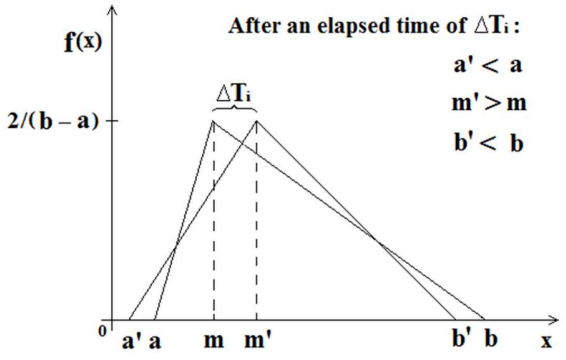


Figure 5: Triangular probability distribution in the absence of data and uncertainties in a, in the mode m and in b, after ΔTi.

By the observation in Figure 5, it is noticed that not only the maximum value b could change, but also the value of the mode m and the minimum value a of the triangular distribution. The letters a', m' and b' would represent a, m and b changed after an elapsed time interval, ΔTi. Therefore, a broader question should be asked about the possibility of not only b changing with an elapsed time (ΔTi), but also a and m as shown in Figure 5.

In consequence, the widest concept of the discussion would be to ask whether it is possible to guarantee the existence of the stationarity in the data before adopting a theoretical probability distribution estimated for simulation studies in the absence of data. The consulted specialist could be asked to give his opinion on the existence of this stationarity, although in a qualitative way.

Furthermore, if the specialist points to the existence of two or more values around which the highest frequencies occur (two or more modes), then the presence of nonhomogeneity is likely to happen. In this case, it is not possible to adopt a single theoretical probability distribution to represent the entry process under analysis. The analysis should be divided into as many samples as the number of different modes.

Despite the classifications that Biller and Gunes (2010) formulate for input models in univariate time series, vector time series, and random vectors, they do not report the possibility of these models occurring in a situation in which data are not available. However, the concepts are useful even in studies in the absence of data. For example, a simulation model of an oil terminal that transfers products with a strong negative correlation between the volumes pumped. When in one month the movement of a product is high, the movement of the other product is low: a typical case of vector time series. Eventually, this modeling could be useful even in a context of lack of data, with prior knowledge about this "correlation", even if in a qualitative way.

At last, it is common to use estimated theoretical distributions even when there is data to be treated, but the urgency of certain simulation projects makes the use of estimated distributions to be adopted, rather than the adjusted ones. Even so, it would be important to verify the premises already mentioned and again even qualitatively and subjectively.

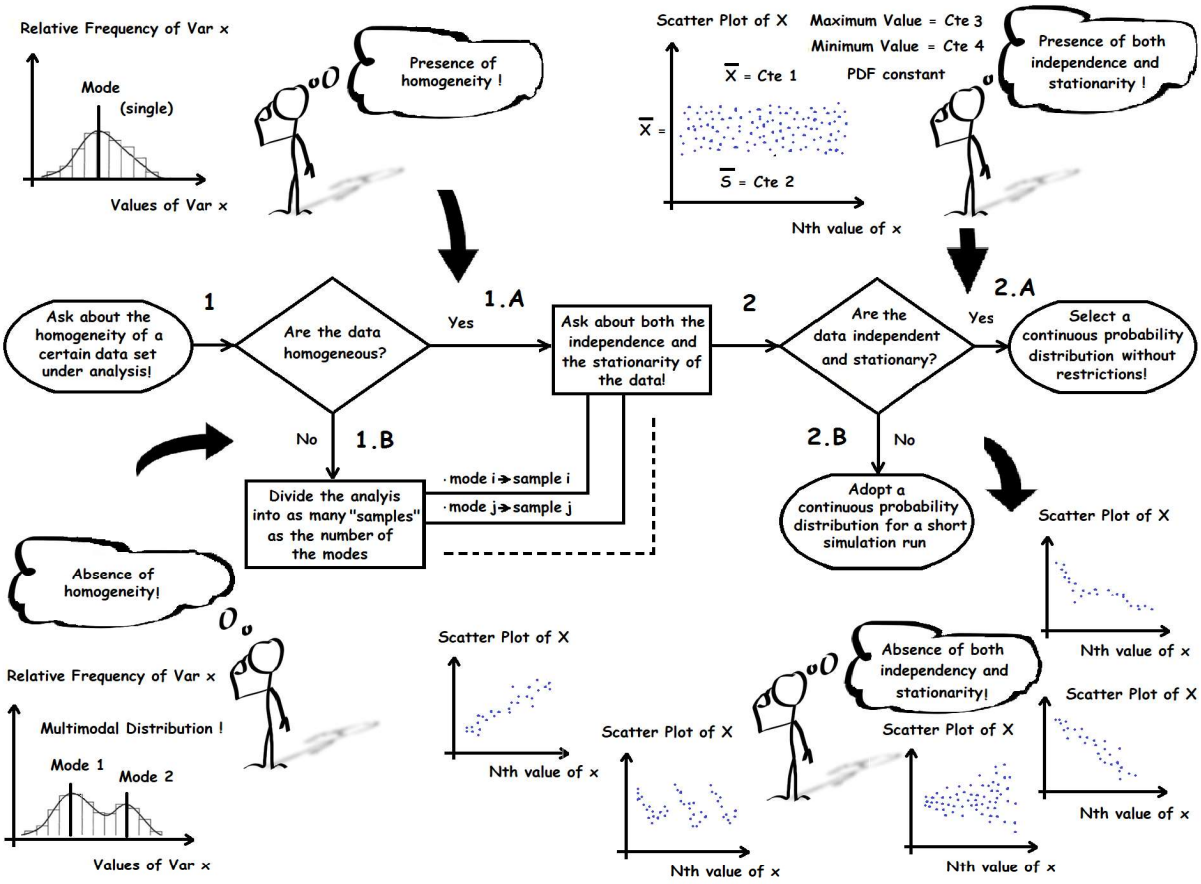
3.3 Consequences

The adoption of theoretical probability distributions in contexts that do not guarantee data independence, homogeneity, and stationarity can generate unpredictable results in the sense that it is not possible to estimate the magnitude of the resulting deviations. As the data are not present, the impacts are likely to be realized only when the studied scenarios are effectively implemented.

Another possible consequence of the lack of these premises is the inconstancy of the distribution parameters in the real process, despite the adoption of input models that assume constant parameters. Large differences may be realized between the variability of the real system and the simulated system.

3.4 Mitigations

As a result of the discussion, this article proposes a roadmap to mitigate the diversity of weaknesses inherent in simulation studies in the absence of data. Its structure is based on the collection of qualitative and subjective perceptions of the specialists that can be used to evaluate both the premises inherent to the "input data" and to estimate the parameters of the distributions to be chosen to represent the process under analysis, see Flowchart 1. It is emphasized that data are not present either because they do not exist or because they are not available for some other reason.



1. Ask the specialist about his perception of the existence of homogeneity for a certain "data set" under analysis. If the perception of homogeneity is confirmed, follow the sequence on the roadmap; if not, the analysis of the "data set" should be divided into as many "samples" as the number of modes;
2. Having treated the question of homogeneity, ask the specialist his perception of the existence of both independence and stationarity of the "data" for each "sample". If so, an estimated continuous probability distribution can be selected; If not, a palliative solution must be adopted, which is to use an estimated continuous probability distribution but only for a short simulation run. In this way, both the effects of the absence of independence and stationarity would be mitigated. However, the notion of what a short simulation run means would depend on the knowledge about the system to be studied;

OBS: The 2.B way is inspired in something already existing in simulation theory, but in the field of discrete probability distributions. A nonstationary arrival process is commonly represented by a group of subintervals in which each subinterval is modeled as a Poisson process characterized by a proper λ.

Flowchart 1: Roadmap for input modeling through qualitative and subjective perceptions of specialists concerning premises, parameters, and distributions.

3.5 Case study

The proposed roadmap is applied in a case study of an industrial facility in the oil and gas industry. It will be possible to verify the achievement of a simulation model behavior very close to reality through a brief validation in a situation in which there is a total absence of data. More detailed case studies will be analyzed in future research given the limited space available in this study.

The analyzed system is composed of a production unit, a pipeline and a road and sea loading terminal (waterway terminal), see Figure 6.A. The pipeline connects the production unit with the terminal. A product x is produced in the production unit and pumped through the pipeline to be stored in the terminal tanks. The product x is shipped from the terminal by trucks and ships.

Certain characteristics of this system are important information for face-to-face validation of the model. All the characteristics are based on the perceptions of the system's operation team. Among these, the main

information is that there is generally one loading truck in line during the business day, but no trucks during the weekends. The amount of time spent by a ship at port (turnaround time of a tanker) is always between thirty-six hours and one hundred and twenty hours. Lastly, the set of loading operations, both for ships and trucks in the terminal, is dimensioned so that there is always space available in the tanks of the terminal (tankfarm of the terminal) in such a way that the production of the product x never needs to be interrupted due to lack of terminal space.

The main processes for entering the simulator are the arrival of trucks at the terminal and the arrival of ships at the terminal. For each of these input processes, an input modeling is generated to feed the simulator of the system through the proposed roadmap. This modeling is the time between truck arrivals and the time between ship arrivals, respectively.

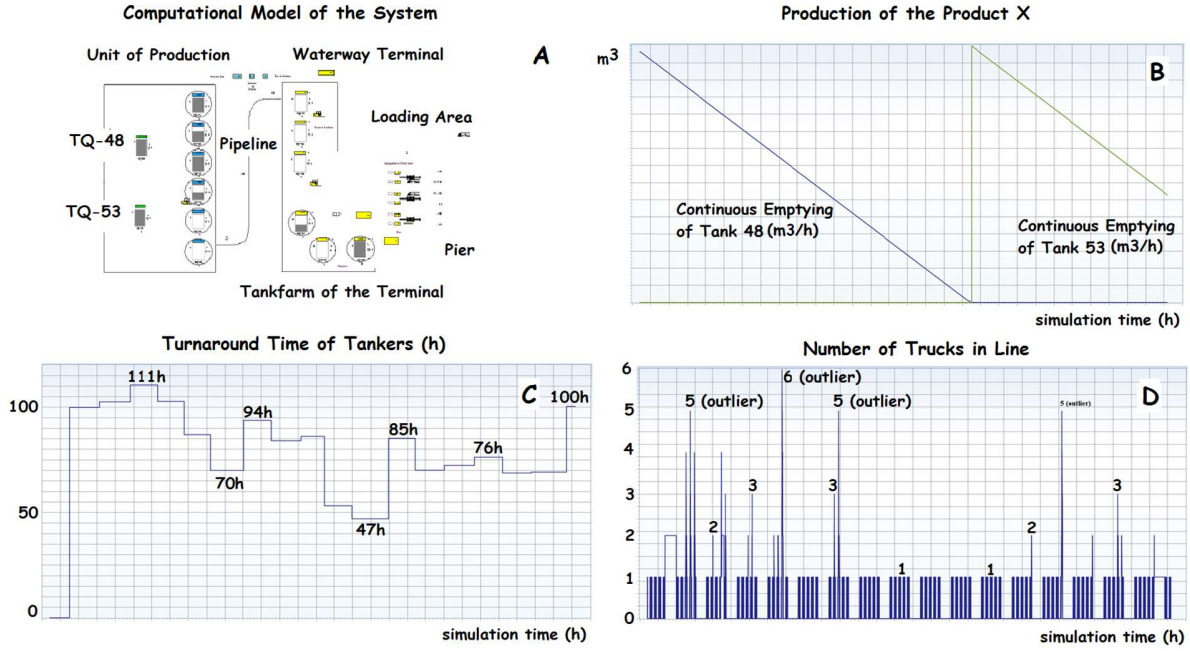


Figure 6: Computational model and graphs of the results of the simulation model output.

For the time between truck arrivals, the specialist was asked under his qualitative and subjective perception about the presence of homogeneity in the data (step 1). The answer was affirmative, with an estimated mode value of (0.35h) for this "data set" (step 1.A). Then, he was asked about his qualitative and subjective perception on the existence of simultaneous independence and stationarity for the same "data set" (step2). The answer was also affirmative. Then, he specified a triangular distribution with the estimated mode value (0.35h), plus the qualitative and subjective estimates for both the minimum (0.33h) and maximum (0.49h) values of the same "data set" (Step 2.A). That is, the estimated continuous probability distribution was a triangular distribution T (0.33, 0.35, 0.49)h.

For the time between ship arrivals, the specialist was asked under his qualitative and subjective perception about the presence of homogeneity in the data (step 1). The answer was affirmative and he defined an estimated mode value of (160h) for this "data set" (step 1.A). Then, he was asked about his qualitative and subjective perception on the existence of simultaneous independence and stationarity for the same "data set" (step2). The answer was negative, as a new class of ships would start operating at the terminal three months ahead, with a loading capacity much higher than the current ships. Consequently, the time between arrivals would probably increase, compromising the stationarity of the data, especially the estimated mode value of 160h (starting stage 2.B). Therefore, a short run of three months was established to be able to use the qualitative and subjective perception of the specialist at the time of this study. A

probability distribution to be specified would have been valid only for three months, as an "expiration date". Thus, a triangular distribution was specified for a short three-month simulation run with the estimated mode value (160h), plus the qualitative and subjective estimates for both the minimum value (155h) and the maximum value (165h) of the same "data set" (step 2.B). It is important to note that the circumstances of this modeling (time between ship arrivals) have limited the forecasting capacity of the simulation model for the long-term: no predictions could be reliable for a period over three months ahead but only for forecasts in a period of up to three months. Within this context, the estimated continuous probability distribution was a triangular T (155,160,165)h.

Finally, we can attest to the adherence of the results of the simulated system with the characteristics already reported from the real system. The production unit does not suffer production interruptions, as can be seen in Figure 6.B, and both the amount of time of ships at the port and the number of trucks in line are in agreement with the perceptions of the system's operation team, as can be seen in Figure 6.C and Figure 6.D, respectively.

4 CONCLUSION

It is possible to state that the simulation literature does not offer a theoretical framework that is sufficient for the input modeling of any system with a widely varied behavioral patterns of input data. This can become a "big trap", not only when the study occurs in the presence of data, but mainly in the absence of data.

The existence of the independence, homogeneity, and stationarity premises of the data are fundamental for the adoption of any continuous theoretical probability distributions in simulation models in the presence of data. Thus, whatever distribution is chosen in the absence of data by the consultation with a specialist, the specialist must also previously analyze the presence of these same premises in the "actual data". Otherwise, some pitfalls may arise, either due to the lack of independence, homogeneity, or stationarity, such as the possibility of the minimum value, mode, and maximum value statistics of a "data" over time not being constant.

In other words, simulation studies in the absence of data may generate unpredictable output if the solution that is currently recommended by the literature is adopted rashly: the immediate selection of an estimated theoretical probability distributions without major concerns. If in the presence of data the verification of the mentioned premises is recommended by the literature, why would it not be in the absence of data, even if it is done qualitatively and subjectively?

In conclusion, some questions should be asked in simulation studies in the absence of data, such as: is there a dependence within the data (autocorrelation)?, are there samples from different populations mixed (non-homogeneity)?, and what if distributions change over time (non-stationarity)? For this reason, a roadmap is proposed to structure the sequence of these questions, the possible answers, and the steps that must be followed to mitigate the weaknesses inherent to simulation studies in the absence of data. With this, the applicability of the roadmap is shown in a real case in the oil and gas industry. In the end, a gap in the literature regarding this subject is filled.

In terms of new researches, the bayesian approach, inverse problems, and consultation with a specialist for multivariate data are issues that deserve attention in future research on the topic.

ACKNOWLEDGMENTS

We thank to Petrobras for its support in the development and publication of this study. The research was also carried out with support from CAPES, FAPEMIG, CNPq and the NEAAD.

REFERENCES

Banks, J., Carson, J.S, Nelson, B.L, Nicol, D.M. 2010. *Discrete Event System Simulation*. 5th ed.Upper Saddle River, New Jersey: Prentice-Hall.

Biller, B. e Gunes, C. 2010. "Introduction to simulation input modeling" In Proceedings of the 2010 Winter Simulation Conference,edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yucesan, 49 – 58. Piscataway, New Jersey: Institute of Electrical and Eletronics Engineers, Inc.

Biller, B. e Nelson, B.L. 2002. "Answers to the Top Ten Input Modeling Questions" In Proceedings of the 2002 Winter Simulation Conference, edited by E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 35 – 40. Piscataway, New Jersey: Institute of Electrical and Eletronics Engineers, Inc.

Biller, B. e Nelson, B.L. 2005. "Fitting Time-Series Input Processes for Simulation", Operations Research,58(3):549 – 559.

Cheng, R. 2017. "History of Input Modeling" In Proceedings of the 2017 Winter Simulation Conference, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 181 – 201. Piscataway, New Jersey: Institute of Electrical and Eletronics Engineers, Inc.

Chwif, L. e Medina, A. 2015. *Modelagem e Simulação de Eventos Discretos*. 4th ed. Rio de Janeiro: Campus.

DeBrota, D.J., Dittus, R.S., Swain, J,J., Roberts, S.D., Wilson, J,R., Venkatraman, S. 1998. "Modeling Input Processes with Johnson Distributions" In Proceedings of the 1989 Winter Simulation Conference, edited by E.A. MacNair, K.J. Musselman, P. Heidelbergerby,308 – 318. Piscataway, New Jersey: Institute of Electrical and Eletronics Engineers, Inc.

Elsayed, A. 2012. *Reliability Engineering*. 2nd ed. New Jersy: Wiley.

Goeva, A. e Lam, H. 2014. "Reconstructing Input Models via Simulation Optimization" In Proceedings of the 2014 Winter Simulation Conference, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 698 – 709. Piscataway, New Jersey: Institute of Electrical and Eletronics Engineers, Inc.

Harrell, C., Ghosh, B. e Bowden, R. 2011. *Simulation Using Promodel*. 3rd ed.New York: McGraw-Hill.

Johnson, M.E. e Mollaghasemi, M. 1994."Simulation input data modeling", Operations Research, 53(1), 47 – 75.

Law, A.M. 2007. *Simulation Modeling & Analysis*, 4th ed. New York. McGraw-Hill.

Law, A.M. 2016. "A Tutorial on How to Select Simulation Input Probability Distributions" In Proceedings of the 2016 Winter Simulation Conference, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 103 – 117. Piscataway, New Jersey: Institute of Electrical and Eletronics Engineers, Inc.

Leal, R. 2018. Stochastic Simulation of an Oil Terminal to Reduce the Turnaround Time of Tankers Through Pipeline Operability Improvements. *Anais* do L Simpósio Brasileiro de Pesquisa Opeacional, Rio de Janeiro (Rio de Janeiro, Brasil).

Montegomery, D.C., Jenninngs, C.L. and Kulahci, M. 2008. *Introduction to Time Series Analysis and Forecasting*. N.J: Wiley.

Morettin, P.A. e Toloi, C.M.C. 2006. *Análise de Séries Temporais*. 2nd. São Paulo:Blücher.

Nelson, B.L. e Yamnitsky, M. 1998. "Input modeling tools for complex problems" In Proceedings of the 1998 Winter Simulation Conference, edited by D.J. Medeiros, E.F. Watson, J.S. Carson and M.S. Manivannan, 105 – 112. Piscataway, New Jersey: Institute of Electrical and Eletronics Engineers, Inc.

Upton, G. e Cook, I. 2013.*A Dictionary of Statistics*.3rd ed.Oxford:Oxford University Press.

Tarantola, A. 2005. *Inverse Problem Theory and Methods for Model Parameter Estimator*, Philadelphia: SIAM.

AUTHOR BIOGRAPHIES

LEONARDO ROSAS LEAL is an industrial engineer at the Petrobras Research and Development Center - Cenpes. He has bachelor's and master's degrees in Industrial Engineering from the Federal University of Rio de Janeiro - UFRJ. His research interest includes Simulation, Machine Learning, Reliability Engineering, Maintenance Engineering and Quality Engineering. His email address is leonardo.leal@petrobras.com.br and leonardorosas@uol.com.br.

JOSÉ ARNALDO BARRA MONTEVECHI is a Titular Professor of Production Engineering and Management Institute at the Federal University of Itajubá, in Brazil. He holds the degrees of Mechanical Engineer from the Federal University of Itajubá, M.Sc. in Mechanical Engineer from the Federal University of Santa Catarina, and Doctorate of Engineering from Polytechnic School of the University of São Paulo. His research interest includes Operational Research, Simulation and Economic Engineering. His e-mail address is montevechi@unifei.edu.br.

MONA LIZA MOURA DE OLIVEIRA is pursuing a Postdoctorate in Industrial Engineering from the Federal University of Itajubá, in Brazil. She has a master's degree in Industrial Engineering and an undergraduate degree in Industrial Engineering which she also received from UNIFEI. Her email address is monaoli@yahoo.com.br.

CARLOS HENRIQUE DOS SANTOS is a Ph.D. Student in Industrial Engineering at the Federal University of Itajubá, in Brazil. His bachelor's and master's degrees in Industrial Engineering from the Federal University of Itajubá. His research interest includes Simulation and Six Sigma. His e-mail address is chenrique.santos@unifei.edu.br.

TÁBATA FERNANDES PEREIRA is a Professor of Business Management at Federal University of Itajubá – Itabira Campus. She is PhD in Industrial Engineering at Federal University of Itajubá with sandwich period at Texas A&M University as a Visiting Scholar. She received her Master degree in Industrial Engineering from the Federal University of Itajubá, and BS in Information Systems from Research and Education Foundation of Itajubá. Her email is tabatafp@unifei.edu.br.