DDPS: A Project Methodology for Data-Driven Process Simulation

Completed Research Full Paper

Laura Johanna Oberle
Chair of Service Operations Management,
University of Mannheim
laura.oberle@uni-mannheim.de

Han van der Aa
Data and Web Science Group,
University of Mannheim
han.van.der.aa@uni-mannheim.de

Abstract

Business Process Simulation (BPS) is commonly used by decision makers to evaluate the expected impact of process changes, without interfering in running systems. Though powerful, the benefits of BPS are highly dependent on the quality of the employed simulation model. Given that both manual and automated approaches have their limitations, we propose that simulation models should rather be created using a project-based approach, in which data-driven analysis techniques are applied in a manner tailored to the characteristics of the project at hand. In this work, we provide guidance for this by proposing DDPS, a project methodology for Data-Driven Process Simulation. The methodology guides users through project preparation and parameter estimation to the creation and validation of the simulation model itself, i.e., a digital twin derived from execution data. Overall, DDPS thus supports practitioners by making the proper execution of BPS projects more feasible.

Keywords

Digital twins, process simulation, data-driven analysis, process mining.

Introduction

Business Process Simulation (BPS) is a popular technique for the quantitative analysis of business processes. BPS uses a simulator to generate hypothetical instances of a process (Dumas et al., 2018, p.279), providing insights into the expected performance of a process design through measures such as the overall cycle time, resource utilization, or waiting time for a given activity (Camargo et al., 2020). In this manner, BPS enables decision makers to analyze and compare alternative process designs—so-called *what-if analysis*—without having to actually implement process changes. However, the potential of BPS hinges on the availability of a simulation model that accurately reflects the dynamics of a process.

Establishing such simulation models is known to be highly complex (Camargo et al., 2020); the broader task of building digital twins of a process is even known to be one of the most challenging concerns in business process management (Dumas, 2021). Creating simulation models by hand can be regarded as a manual process discovery task. Therefore, it suffers from the same pitfalls as manual discovery in general, such as its time-consuming nature, as well as issues of objectivity and richness (Dumas et al., 2018, p.176). However, through the advent of Industry 4.0 and other widespread digitization initiatives, simulation models can also be created on the basis of real process execution data, captured in so-called event logs, which avoids some of the biases associated with manual establishment of simulation models. Although various automated approaches have been proposed for this purpose (Abohamad et al., 2017; Camargo et al., 2021; Liu et al., 2012; Măruşter and van Beest, 2009; Rozinat et al., 2009), they struggle with their own challenges, such as the completeness and correctness of recorded process behavior (Bose et al., 2013), missing process steps (Fahland et al., 2022), and obscured information, such as work schedules and patterns (Estrada-Torres et al., 2021). Such issues, in turn, can have a considerable impact on the reliability of the obtained simulation results.

Given that the exact manifestation and impact of such issues differ per simulation project, no single automated discovery approach will provide a silver bullet for the end-to-end discovery of accurate simulation models. Therefore, we argue for a balanced approach for the creation of simulation models, achieved by combining the benefits of data-driven process mining techniques with project-specific manual interventions. As such, rather than depending on a single discovery technique to generate a simulation model at the press of a button, we propose to apply different process mining techniques to tackle specific parts of the discovery task when they are relevant for a given simulation project. Examples include the use of dedicated techniques to resolve data quality issues, event abstraction when dealing with low-level data, or techniques that detect schedules and batching behavior of resources.

However, the flexibility of such a hybrid approach also increases the complexity of a simulation project, adding various degrees of freedom to the BPS discovery task. To properly use data-driven BPS, the creation of a simulation model should thus be carefully considered in light of the characteristics of a given process, the available event data and domain knowledge, and purpose of the simulation. Therefore, our work targets the following research objective: *Develop an artifact that guides users through the various facets involved in the data-driven discovery of business process simulation models.*

To achieve this objective, we developed the DDPS methodology, a project methodology for Data-Driven Process Simulation. DDPS allows organizations to approach data-driven BPS in a structured manner, benefiting from the potential of a broad range of process mining techniques for various facets of the model discovery task. DDPS consists of seven stages that contribute to the establishment, verification, and validation of a simulation model on the basis of an available event log. For each stage, we discuss its goals, the main challenges that one may face in a given simulation project, and highlight automated as well as manual approaches that can help to overcome them. We validated and improved the DDPS methodology through interviews with six experts from both industry and research.

Related Work

Discovery of simulation models. Various approaches support the end-to-end discovery of simulation models for BPS on the basis of event logs. Early works in this regard (Măruşter and van Beest, 2009; Rozinat et al. 2009) go back to the first stages of process mining research, focusing on the discovery of Colored Petri Nets as a foundation for simulation. More recent and, thus, more advanced approaches build on broader process mining developments, employing, e.g., more accurate discovery techniques (Abohamad et al., 2017), causal analysis (Liu et al., 2012), and log repair to overcome data quality issues (Camargo et al., 2020). Recent work (Camargo et al., 2021) has also turned to using deep learning, which better captures the temporal dynamics as compared to data-driven simulation methods purely based on automated discovery. Next to these end-to-end solutions, various techniques target specific issues in the discovery of simulation models, such as the discovery of work schedules (Estrada-Torres et al., 2021) and the inference of missing start times (Fracca et al., 2022) or entire activities (Andrews and Wynn, 2017).

Overall, process mining is thus commonly used to calibrate simulation models. Yet, as mentioned before, automated end-to-end discovery approaches have their downsides, so that none of them provides a clear solution for all simulation tasks (Dumas, 2021). Furthermore, techniques that target specific issues are highly useful, but lack concrete guidance on when to use which, and how to combine them. Our DDPS methodology provides exactly such guidance, so that users can employ and combine suitable, data-driven techniques for purpose-driven, end-to-end creation of simulation models.

Related project methodologies. Various project methodologies have been proposed that differ from DDPS in terms of scope, but that also strive to guide users during analysis projects. In the context of process mining, the primary methodology is the PM² methodology (van Eck et al., 2015) guides users through the execution of process mining projects by describing various stages and their respective inputs, outputs and tasks. As such, it is positioned as the process mining alternative of prior data mining methodologies, such as CRISP-DM (Wirth and Hipp, 2000) and SEMMA (Mariscal et al., 2010). In the context of simulation, the book by Law (2015) provides an extensive discussion of general methodology for simulation projects, complemented by more concrete step-by-step instructions (Law, 2019). With respect to data-driven BPS, work by Martin et al. (2016) provides an overview of the relation between process mining and BPS, focusing on the various discovery tasks and modeling components, though without any procedural guidance.

Compared to these works, our DDPS methodology fulfills the methodological requirements established by Law, while providing a specialized procedural approach for the execution of data-driven BPS projects, so that users can tackle challenges specific to BPS in an appropriate manner, using the right techniques.

Research Method

We followed the Design Science Research Methodology (DSRM) by Peffers et al. (2007) in our work. Our research has an *objective-centered solution* as its entry point, with the objective to provide users with guidance to establish accurate business process simulation models. To achieve this, we established the DDPS project methodology, which splits up the complex task of simulation model creation into seven key stages (as visualized in Figure 1) that each provide guidance for a specific part of the model-creation task.

To establish the high-level structure of the methodology, we started from the main building blocks of a process simulation model, which is the need to determine parameters for the so-called control-flow, time, and resource perspectives of a process (Camargo et al., 2020). We dedicate a single stage to each of these perspectives (Stages 3 to 5), followed by Stage 6, which combines these building blocks into a first version of a simulation model. Then, we followed best practices from existing project methodologies (Law, 2019; van Eck et al., 2015) to cover key tasks of simulation projects, such as that a project should start with planning (Stage 1) and data preparation (Stage 2), and include both verification (part of Stage 6) and validation (Stage 7) towards the end. Then, given this high-level structure, we filled in the individual stages by determining which intermediary outcomes are required per stage and, by considering state-of-the-art literature, which challenges users may face and which techniques can help overcome these.

We evaluated the version of DDPS obtained in this manner by conducting expert interviews as a formative evaluation (Venable et al., 2016) to assess and improve our methodology. The interview insights to small adaptations, like adding an explicit calibration loop, resulting in the version of DDPS presented herein.

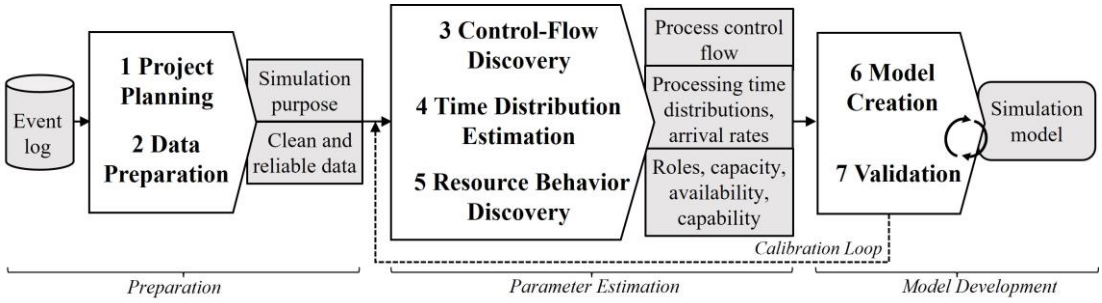

Figure 1: The seven stages of the DDPS project methodology

DDPS Project Methodology

This section introduces the DDPS project methodology, consisting of the seven stages visualized in Figure 1. The first two stages form the groundwork for a simulation project, involving project planning (Stage 1) and data preparation (Stage 2). The next three jointly establish the parameters for a simulation model, targeting the control-flow (Stage 3), time (Stage 4), and resource (Stage 5) perspectives of a process. The outcomes of these stages are then used to build and verify an initial version of the simulation model (Stage 6). Finally, the model is validated to ensure its proper functioning and representativeness (Stage 7), which may result in iterative calibration loops. Afterwards, the simulation model is ready to aid decision making by allowing users to evaluate alternative process changes.

Note that although verification and validation of the entire model are conducted at the end to evaluate the interplay of the various building blocks in a simulation model, we advocate for the immediate validation of intermediary outcomes, as detailed in the respective stages below. Furthermore, Figure 1 only shows a calibration loop after Stage 7 for clarity, even though users may go back to earlier stages at any time if a certain stage provides a reason for this. For example, gaining insights into the available data (Stage 2) may lead to a revision of the defined simulation purpose (Stage 1), whereas insights gained into resource behavior (Stage 5) may lead to a refinement of the defined control-flow activities (Stage 3).

Stage 1: Project Planning

To start a simulation project, the purpose must be defined, i.e., the later-on performed scenario analysis, from which the corresponding desired level of detail of the simulation and the scope regarding process, time frame, resources, and relevant performance measures for evaluating process alternatives are derived. Thinking about the final use of the simulation model before creating it ensures that the built simulation model is capable of evaluating the desired scenarios, as the necessary components and details are included adequately. A concrete purpose thus determines how closely the simulation model should reflect the real process with regard to different factors. BPS projects generally have the same high-level purpose, i.e., the evaluation of some kind of process change(s), yet the specific scenarios to be tested and factors to be changed differ and, thus, determine how the following stages should be executed. For instance, if the purpose is to compare costs of different process alternatives, proper cost information must be included, whereas if the purpose is to compare different resource allocation possibilities to improve the waiting time of customers, costs are negligible but accurate resource information is crucial.

From the purpose and scope, the required level of detail is derived, including decisions regarding the simplifications that are acceptable without diminishing the simulation result's validity. These decisions shape the degree of precision required in the subsequent stages. Concrete decisions in this regard may include the choice to whether or not model an IT system as a resource with unconstrained availability, or whether to differentiate the incoming cases by their type instead of treating them all as equal.

Stage 2: Data Preparation

The goal of this stage is to obtain clean and reliable data for the subsequent stages, which is achieved by obtaining an event log itself, removing noise and, where possible, resolving data quality issues. Aside from an event log, useful simulation input may be artifacts such as process models or work schedules, which, when available, can be used to confirm or adapt intermediary outcomes in the remainder. Having a high-quality event log is a key requirement for any process mining project, thus also when performing data-driven BPS. Obtaining data requires an organization to properly record what happens in its processes, whereas the subsequent extraction is a complex and time-consuming endeavor, which needs to be adapted to the specifics of a process and its underlying IT infrastructure (Murillas et al., 2015). Once obtained, it is important to assess an event log with respect to data quality issues. Such issues can involve incomplete, noisy, or imprecise data (Bose et al., 2013) which can reduce the validity of analytical results obtained on their basis (Solti, 2019) and require data preparation to improve the efficiency of any subsequent analysis (Zhang et al., 2003).

While the challenges of obtaining event logs and dealing with quality issues are largely the same for BPS as for other process mining projects, two aspects that are particularly crucial in the context of process simulation are life-cycle transitions and resource information. Whereas most process mining tasks only use completion events, BPS particularly benefits from the availability of accurate timestamps that capture both the start and completion of an activity instance, since this allows one to accurately determine the exact processing time of a step. If, instead, an activity duration is calculated based on its and the next activity's start time, any waiting time would be included in there as well, which causes undesirable inflation of processing times, as well as resource utilization and, thus, the overall process throughput time. Furthermore, having accurate data on the resource perspective of a process is also of crucial importance for BPS. To this end, it is not just important to capture which resource instances perform each event, but also to avoid polluting data on this perspective. Although these aspects are best tackled by obtaining highly expressive event logs, we later (Stages 4 and 5) cover specific methods to deal with data quality issues in cases where they are unavoidable.

In light of these challenges and to increase overall data quality, a key preprocessing task is to filter out incomplete cases from a log. Further, excluding less frequent data reduces noise in the data (Solti, 2019) and allows for more rigid role assignments, availability checks, and processing time calculations. To achieve this, it is, e.g., advised to omit instances with 75% or more irregular transitions (Măruşter and van Beest, 2009). For simplicity, all automated tasks performed by unconstrained IT systems can be omitted here, since they do not impact the simulation. Finally, it is important to recognize that any applied preprocessing step should be carefully considered when interpreting the obtained simulation results.

Stage 3: Control-Flow Discovery

Stage 3 involves the creation of a process model, which forms the structural (i.e., control-flow) basis for a simulation model. This task must precede the other parameter-estimation stages, as it determines the activities whose processing times will be estimated (Stage 4) and to which resources are allocated (Stage 5). Stage 3 itself contains three parts: event abstraction, model discovery, and decision-point annotation.

Event abstraction. Before discovering a process model, it is important to recognize that process simulation requires a model that captures business activities, which are performed by resources and take a certain amount of time. Therefore, it may be necessary to obtain a more appropriate level of granularity for the event data, achieved by grouping together related events. Such abstraction can be performed in different manners, depending on the characteristics of the available data. If events with different life-cycle stages are available (cf. Stage 2), corresponding start and completion events need to be associated with each other. This event correlation task can be performed using established techniques (Diba et al., 2020). Furthermore, if the event classes themselves are at a too low-level of granularity, or if life-cycle information is not available, it may be beneficial to group together related event classes in order to obtain higher-level activities. If a set of desired activities is already known (e.g., through domain knowledge or a normative process model), supervised abstraction techniques may be employed for this purpose, whereas otherwise unsupervised abstraction is required (Diba et al., 2020). In either case, it is important to ensure that grouped events are performed by the same resource, in order to establish activities that can be simulated. For this, constraint-based abstraction may be employed (Rebmann et al., 2022).

Model discovery. Once the appropriate level of granularity has been achieved, a process model can be discovered from the (abstracted) event data. Depending on the intended simulation tool, this model should be a Petri net or BPMN model. Various techniques can be used for discovery, e.g., the inductive miner infrequent (Leemans et al., 2014) and the Split miner 2.0 (Augusto et al., 2021). To ensure that the process model is fit for simulation, manual post-processing of the obtained model may be useful, which can involve the removal of incorrect dependencies. This discovery task may be performed in an iterative manner with event abstraction, until a suitable model has been obtained.

A discovered process model can be validated by comparing it to a normative process model if one exists. In case of discrepancies, the cause of the differences should be investigated. If differences are due to data quality issues, such as faulty recordings or missing events, the normative model can be used to repair the discovered process model. However, if differences occur because reality deviates from expectations, the model discovered from unbiased data should be favored.

Decision points. Finally, for BPS, special attention must be drawn to the modeling of decision points in a process. For choices in a process, this involves the derivation of branching probabilities from an event log or, if relevant attributes are available, the use of decision mining to represent data-driven decisions (Bazhenova and Weske, 2016). Finally, for loops, it is advisable to limit the maximum number of loops being executed to prevent cases from getting stuck during simulation.

Stage 4: Time Distribution Estimation

In stage 4, the necessary time parameters are estimated or fitted to statistical distributions, covering the inter-arrival rate of cases and the processing times of activities. Based on the individual project's scope and level of detail established in Stage 1, these estimations may need to be done for different case types. When fitting any distribution, the fundamental choice between a theoretical and an empirical distribution arises. To determine the suitability and fit of a theoretical distribution, hypothesis tests like the Chi Square or Kolmogorov-Smirnov test can be used. Empirical distributions should be used for p-values of 0.1 or higher, indicating a poor fit of the theoretical distribution (Kelton et al., 2015, p.187). Graphical plots or histograms help to visually determine the most representative theoretical distribution, and in case of discrete histograms the duration can be rounded appropriately (see, e.g., Watson et al., 1998).

Inter-arrival rate. Instead of asking for a specific number of cases to be created, simulation tools commonly require the inter-arrival rate of the cases. Therefore, a distribution must be fitted to the data. Unless specified otherwise, it is common to assume a random and hence Poisson process, which leads to exponentially distributed inter-arrival rates, if the arrivals are independent from each other (Cachon and Terwiesch, 2018, pp.129-135). However, it may be necessary to avoid such an exponential distribution due

to, for example, batch arrivals or appointment schedules. Here, it is also important to recognize that any empirical or theoretical distribution of time parameters induces some degree of variability in the simulation model. Depending on the temporal scope of the simulation project, analysts must decide if further variability should be added by, e.g., accounting for seasonal differences. In essence, more realistic modelling often adds complexity to the model. The trade-off between simplicity and a close resemblance of reality must be managed appropriately, depending on the purpose of the simulation project.

Processing times. As every activity demands a certain time from available resources, the processing time has a great impact on the simulation outcomes and should hence be calculated carefully (cf. Stage 2). Generally, the availability of life-cycle transition data significantly eases the calculation. If these are not available, the calculation should be done carefully to only include processing times, ignoring any waits in between. Alternatively, dedicated techniques (Fracca et al., 2022) may be used to estimate start timestamps when only end timestamps are available. The obtained processing times should then be fitted to a probabilistic distribution, leading to the evaluation of a theoretical distribution's suitability. Obvious outliers with unrealistically long duration should be excluded by applying an upper bound cut-off.

Stage 5: Resource Behavior Discovery

Finally, we consider the resource perspective of a simulation model by deriving the roles, capacities, capabilities, work schedules, and availability of the resources in a process, i.e., which resources there are and how those behave and perform. Given the importance of resources in simulation, the discovery of the resource perspective must be conducted in a rigid manner to get satisfactory results (Martin et al., 2016).

Recent work proposes approaches to handle more complex behavior in human resource modeling. For instance, multi-tasking and availability constraints can be included in the simulation model formulation by adjusting processing times of multi-tasked tasks and using an algorithm to infer work schedules from calendar expressions (Estrada-Torres et al., 2021). Other problems encountered in this area include part-time or batch work, different processing speed for the same task and resource, or accounting for resources involved in more than just the simulated process (van der Aalst et al., 2009).

Another approach to extract service speed and workload of the employees involves a regression analysis to determine and include the relationship in the model formulation (Nakatumba and van der Aalst, 2010). As event logs do not directly show important information on the total available time of a resource, its capabilities or its working schedule, these crucial input parameters must be determined otherwise. First, the availability can rudimentarily be checked by determining the time interval from the first to the last event. Of course, this interval would include any breaks in between and exclude any manual work not recorded in the event log. From the individual resource, an overall capacity per role can be derived by looking at all resources in the same role. Any part-time resources, relevant work habits, or other tasks that resources are involved in should be accounted for by reducing the capacity (Estrada-Torres et al., 2021; van der Aalst et al., 2009). Knowing the role and work schedule, a resource can only be assigned to an activity if it is both capable and available at the start time. The correlation of a resource performing an activity and its successor could also be included in the assignment rules. Finally, for non-human resources, such as machines, relevant failure rates and maintenance periods should be accounted for when modeling their availability, in cases where this affects the analysis purpose of a BPS project.

Stage 6: Model Creation

Based on the parameters estimated in the previous stages, a first version of the simulation model is built and internally verified. Different tools are available to construct a BPS model, like Arena or Anylogic for a manual creation, as well as integrated BPS functions offered by certain commercial process mining tools. Drawing on the results from Stages 1 to 5, the discovered process model, preferably a BPMN model, can be enriched by assigning roles depending on capability and availability and attaching the right distribution to the model's activities. During simulation, these resources are then deployed according to the chosen settings. These include, e.g., the handling of breaks, during which resources either disrupt the current activity to continue afterwards, or not.

It is crucial to decide on a warm-up period (WP). If the process has no terminating event and work in progress is passed on, a WP allows to reach a steady-state behavior of the system. This WP must be chosen sufficiently large to analyze the system's performance, independent of the simulation's initial

conditions (Law, 2015). The steady-state simulation performance can be checked using two-moment approximations from queuing theory for single-stage processes (Whitt, 1993) like a or hotline.

Finally, the model verification ensures that the model behaves as intended if input variables are changed. Thereby, the model syntax can be debugged iteratively by observing the model animation during the simulation run to check if, e.g., loops are executed correctly (Law, 2015, pp.251-255). Input parameters can be purposefully changed to verify that other aspects, like the throughput time, respond accordingly.

Stage 7: Validation

Before relying on the simulation results for decision making, the created model's validity must be confirmed. For this task, the simulation model as well as historical data on the process results derived from process mining is used. As outcome, the simulation model is ready to be put to use in scenario analysis.

The validation aims at ensuring the simulated model's adequate representativeness of the real system by checking if the results are reasonable. Therefore, the model output should be compared to the historical data of the real system, using some statistics and 95% confidence intervals, as in Watson et al. (1998). Running the model for a number of replications, i.e., 100, increases the confidence in the results. Useful measures to be compared include the number of cases created or completed, throughput and waiting times, and the occurrence frequency of events, like cancellations or approvals in an application process. Internal validity can be ensured by comparing the replications, which should show similar results.

This comparison can also help to iteratively improve any input parameters that could not be precisely estimated before. If, e.g., processing times and wait times were not clearly differentiated, a comparison of the simulation output and the real data allows for a refinement of the processing time distributions. In this manner, the simulation model can be iteratively improved until suitable. One special issue to be considered in this stage are invisible events, i.e., process steps that are not captured in an event log, such as manual tasks or extraneous waiting times, which create a discrepancy between predicted and actual throughput times. While calibrating the simulation model, such discrepancies can be discovered and resolved, e.g., by quantifying shelf time (Andrews and Wynn, 2017).

Expert Interviews

To evaluate the usefulness and completeness of our DDPS methodology, we interviewed six experts from consulting (2 interviewees), software providers of a data-driven simulation tool (3), and research (1). The experts represent either potential users of the methodology with limited knowledge about BPS projects but prior experience with other methodologies, or experts who are developing a semi of fully automated simulation tool for industry users. In individual, semi-structured interviews, they were asked to name the major challenges of a simulation project, to assess the general procedure of DDPS, to check for missing content, and to evaluate DDPS's usability, efficiency, and ease of application. Details on the interviews can be found in the supplementary material.[1]

High-level insights. In terms of the need for support, the experts recognized the value of the DDPS methodology in how it provides structure to the complex and time-intensive task of process simulation. When considering DDPS in light of automated approaches, an industry user mentioned that "*tools don't explain anything, but only ask for input, yet one does not know how this input is used and processed*". The expert from academia clarified that "*research very much focuses on tools and automation, but tools will never replace feedback loops or explanations from domain experts*". Another interviewee concluded that DDPS is "*very helpful for inexperienced users, who need to gain an understanding of the functionalities and need to know what to pay attention to*". Thus, a user's understanding of the underlying mechanisms is extremely valuable, and the DDPS methodology was confirmed to provide this.

When it comes to the project methodology itself, the interviewees overall found that DDPS offers an intuitive and generalized methodology for creating data-driven simulation models. In terms of the specifics, all experts confirmed the seven stages as necessary tasks for creating a simulation model from data. The interviewees appreciated the input that DDPS provides to users per stage and how it supports

[1] https://gitlab.uni-mannheim.de/loberle/ddps-supplementary-files

different considerations to be made. In that regard, DDPS is perceived to provide the right level of abstraction, so that it provides enough guidance to users, while leaving enough flexibility for project-specific adaptations. For example, an interviewee mentioned that "*for some processes, corresponding simulation models are easier to derive. Yet, the variance in processes and data contradicts a one-size-fits-all approach.*" Furthermore, DDPS is judged to "*give structure to this simulation model creation process, which is generally a difficult, time-intensive, and often even trial-and-error task*".

Improvement suggestions. The interviewees also provided various kinds of input that we subsequently integrated into the DDPS methodology or that we regard as valuable directions for future work. First, interviewees confirmed the need to define a clear simulation purpose (Stage 1): "*Knowing up front what the simulation model should address*" forms the basis for choices made throughout the remainder of a simulation project, primarily resulting in the use of a consistent level of detail throughout each stage. For example, if the processing times and arrival rates differ for each case type, the process flow, the conditional probabilities, and even task allocations to certain roles may differ. In this regard, valuable additional feedback was the advice to "*outline high-level scenarios [to be tested using simulation] to be even more precise early on*", which is an aspect that we incorporated in Stage 1.

Second, an interviewee mentioned the user's challenge on "*how to make realistic parameter changes for the individual scenarios*". When adapting parameter settings during validation or when creating new scenarios to test redesigns, it is important to ensure that parameters are not blindly adjusted until the data gives in, i.e., until the simulation yields the desired outcome. Instead, parameter settings should be adapted with care to ensure that they actually reflect realistic assumptions about the system, e.g., by avoiding that employees are modeled to work 24-hour shifts without breaks. Thus, when adjusting the settings of a BPS model during validation (Stage 7), we now explicitly recommend a careful assessment of the (expected) impact parameters have on target indicators, either in a top-down or bottom-up manner.

A third point relates to treating validation as first-class citizen: "*validation is especially important, because otherwise the further analyses do not make sense or give wrong impressions*". No matter how the first six DDPS stages are performed, the validation stage should not be underestimated. Only in-depth comparison of the simulated behavior to the recorded behavior or, if possible, to the real system can ensure the validity of the simulation models and its underlying design choices. As mentioned by our interviewees, business users appreciate being supported during planning and validation, which at first glance seem less relevant, but actually are challenges requiring time and effort (see, e.g., Watson et al., 1998). To further stress the relevance of validation, we now explicitly captured a calibration loop in the main overview of DDPS (Figure 1) and recommend performing validation in per stage, where possible.

Finally, interviewees pointed out further possibilities on how to guide users during the execution of the individual stages of DDPS, such as by developing decision trees that provide specific recommendations on which aspects to consider or techniques to apply in a particular context or given the characteristics of the available data. We consider this as a great suggestion for future research.

Implications

Implications for research. The DDPS project methodology provides a structure on how to approach the discovery of accurate simulations models in a step-by-step manner, breaking up the larger task into more manageable chunks. Our work can thus streamline the discourse on relevant techniques, as opposed to the current, highly fragmented state. In addition, the stage-based nature of DDPS implies that it becomes less important for process mining techniques that relate to BPS to always end up yielding a simulation model themselves, since their contribution in the form of supporting a specific (sub-)stage in DDPS can already be highly valuable in itself. Currently, this is often not the case, resulting in various BPS works that propose techniques that yield simulation models tailored to a highly specific issue, such as inference of extraneous delays (Chapela-Campa and Dumas, 2022) or missing start times (Fracca et al., 2022). As a result, the specific solutions proposed do not receive the full attention that they may warrant, being only part of a larger approach. Furthermore, it also becomes harder to combine such individual techniques, given that their output is an entire simulation model, rather than being optimized to provide a specific intermediary result, which may be combined with other techniques that tackle different issues.

Implications for practice. DDPS makes BPS more manageable and accessible for business users. By providing step-by-step guidelines, DDPS makes the combination of process mining and simulation easy to

use for organizations, allowing them to embrace new forms of value generation. The DDPS project methodology thereby guides user through the BPS models creation by making use of both the increasing abundance of data collected in the execution of business processes, as well as the latest technological advancements in process mining. In this regard, practitioners get a clear idea about the minimal data requirements, how to deal with data quality issues, and challenges endangering successful model creation.

Conclusion

In this paper, we presented the DDPS methodology to support users in their execution of data-driven process simulation projects. Rather than enforcing a one-size-fits-all approach, DDPS helps analysts to make the right decisions when developing a simulation model, tailored to the specifics of the project at hand, such as its purpose and the characteristics of the available data. As such, the different stages can be performed using a range of automated, semi-automated, or manual analysis techniques, depending on the project's individual challenges. Interviews conducted with experts from industry and academia highlight the perceived value added by the DDPS methodology.

We acknowledge two limitations with respect to the current evaluation of our work. First, it must be noted that the interviewed experts did not directly apply DDPS, but rather gave their opinion after being presented its key stages and their tasks. Second, for a summative and naturalistic evaluation DDPS needs to be applied on real-life data in an application scenario, which ensures a match between the outcome and the expectations. Such a scenario will confirm the relevance of hybrid discovery in comparison to a fully automated approach. In future work, we hence aim to address these limitations by evaluating the DDPS methodology in realistic contexts, i.e., by having industry users apply it for their processes. Furthermore, we aim to provide users with additional support during simulation projects by establishing methods, such as decision trees or even automated techniques, that tell users which specific techniques should be employed in their particular context. Overall, while we expect future approaches to complement the current state-of-the-art techniques for data-driven process simulation, our methodology remains valuable, as the stages and their underlying tasks inherently need to be performed in all BPS projects.

REFERENCES

Abohamad, W., Ramy, A., and Arisha, A. 2017. "A hybrid process-mining approach for simulation modeling," in *2019 WSC*, W. K. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer and E. H. Page (eds.), Las Vegas, NV, Piscataway, NJ: IEEE, pp. 1527-1538.

Andrews, R., and Wynn, M. 2017. "Shelf Time Analysis in CTP Insurance Claims Processing," in *PAKDD Workshops*, U. Kang, E.-P. Lim, J. X. Yu and Y.-S. Moon (eds.), Cham: Springer, pp. 151-162.

Augusto, A., Dumas, M., and La Rosa, M. 2021. "Automated Discovery of Process Models with True Concurrency and Inclusive Choices," in *Process Mining Workshops*, S. Leemans and H. Leopold (eds.), Cham: Springer International Publishing, pp. 43-56.

Bazhenova, E., and Weske, M. 2016. "Deriving Decision Models from Process Models by Enhanced Decision Mining," in *BPM Workshops*, M. Reichert and H. A. Reijers (eds.), Springer, pp. 444-457.

Bose, R. J. C., Mans, R., and van der Aalst, W. M. P. 2013. "Wanna improve process mining results?" in *2013 IEEE Symposium on CIDM*, Singapore, Singapore, IEEE, pp. 127-134.

Cachon, G., and Terwiesch, C. 2018. Matching Supply with Demand, McGraw-Hill Publishing.

Camargo, M., Dumas, M., and González-Rojas, O. 2020. "Automated discovery of business process simulation models from event logs," Decision Support Systems (134:3), p. 113284.

Camargo, M., Dumas, M., and González-Rojas, O. 2021. "Learning Accurate Business Process Simulation Models from Event Logs via Automated Process Discovery and Deep Learning," arXiv preprint arXiv:2103.11944.

Chapela-Campa, D., and Dumas, M. 2022. "Modeling Extraneous Activity Delays in Business Process Simulation"

Diba, K., Batoulis, K., Weidlich, M., and Weske, M. 2020. "Extraction, correlation, and abstraction of event data for process mining," WIREs Data Mining and Knowledge Discovery (10:3), p. e1346.

Dumas, M. 2021. "Constructing Digital Twins for Accurate and Reliable What-If Business Process Analysis," in *BPM Workshops*, pp. 6-10.

Dumas, M., La Rosa, M., Mendling, J., and Reijers, H. A. 2018. Fundamentals of Business Process Management, Berlin, Heidelberg: Springer.

Estrada-Torres, B., Camargo, M., Dumas, M., García-Bañuelos, L., Mahdy, I., and Yerokhin, M. 2021. "Discovering business process simulation models in the presence of multitasking and availability constraints," Data & Knowledge Engineering (134:1), p. 101896.

Fahland, D., Denisov, V., and van der Aalst, W. M. 2022. "Inferring Unobserved Events in Systems with Shared Resources and Queues," Fundamenta Informaticae (183:3-4), pp. 203-242.

Fracca, C., Leoni, M. de, Asnicar, F., and Turco, A. 2022. "Estimating Activity Start Timestamps in the Presence of Waiting Times via Process Simulation," in *CAiSE conference*, X. Franch, G. Poels, F. Gailly and M. Snoeck (eds.), Cham: Springer, pp. 287-303.

Kelton, W. D., Sadowski, R. P., and Zupick, N. B. 2015. Simulation with Arena, New York, NY: McGraw-Hill Education.

Law, A. M. 2015. Simulation modeling and analysis, New York, NY: McGraw-Hill Education.

Law, A. M. 2019. "How to Build Valid and Credible Simulation Models," in *2019 WSC*, National Harbor, MD, USA, IEEE, pp. 1402-1414.

Leemans, S. J. J., Fahland, D., and van der Aalst, W. M. P. 2014. "Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour," in *BPM Workshops*, N. Lohmann, M. Song and P. Wohed (eds.), Cham: Springer, pp. 66-78.

Liu, Y., Zhang, H., Li, C., and Jiao, R. J. 2012. "Workflow simulation for operational decision support using event graph through process mining," Decision Support Systems (52:3), pp. 685-697.

Mariscal, G., Marbán, Ó., and Fernández, C. 2010. "A survey of data mining and knowledge discovery process models and methodologies," The Knowledge Engineering Review (25:2), pp. 137-166.

Martin, N., Depaire, B., and Caris, A. 2016. "The Use of Process Mining in Business Process Simulation Model Construction," Business & Information Systems Engineering (58:1), pp. 73-87.

Măruşter, L., and van Beest, N. R. T. P. 2009. "Redesigning business processes: a methodology based on simulation and process mining techniques," Knowledge and Information Systems (21:3), pp. 267-297.

Murillas, E. G. L. de, van der Aalst, W. M. P., and Reijers, H. A. 2015. "Process Mining on Databases: Unearthing Historical Data from Redo Logs," in *Business Process Management*, H. R. Motahari-Nezhad, J. Recker and M. Weidlich (eds.), Cham: Springer, pp. 367-385.

Nakatumba, J., and van der Aalst, W. M. P. 2010. "Analyzing Resource Behavior Using Process Mining," in *BPM Workshops*, W. M. P. van der Aalst, J. Mylopoulos, N. M. Sadeh, M. J. Shaw, C. Szyperski, S. Rinderle-Ma, S. Sadiq and F. Leymann (eds.), Berlin, Heidelberg: Springer, pp. 69-80.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," JMIS (24:3), pp. 45-77.

Rebmann, A., Weidlich, M., and van der Aa, H. 2022. "GECCO: Constraint-driven Abstraction of Low-level Event Logs," in *2022 IEEE 38th ICDE*, Kuala Lumpur, Malaysia. IEEE, pp. 150-163.

Rozinat, A., Mans, R. S., Song, M., and van der Aalst, W. M. P. 2009. "Discovering simulation models," Information Systems (34:3), pp. 305-327.

Sargent, R. G. 2013. "Verification and validation of simulation models," Journal of Simulation (7:1), pp. 12-24.

Solti, A. 2019. "Event Log Cleaning for Business Process Analytics," in *Encyclopedia of Big Data Technologies*, S. Sakr and A. Y. Zomaya (eds.), Cham: Springer.

van der Aalst, W. M. P. 2016. Process Mining, Berlin, Heidelberg: Springer.

van der Aalst, W. M. P., Nakatumba, J., Rozinat, A., and Russell, N. 2009. "Business Process Simulation: How to get it right?" in *International Handbook on Business Process Management*, J. Vom Brocke and M. Rosemann (eds.), Springer.

van Eck, M. L., Lu, X., Leemans, S. J. J., and van der Aalst, W. M. P. 2015. "PM²: A Process Mining Project Methodology," in *CAiSE Conference*, J. Zdravkovic, M. Kirikova and P. Johannesson (eds.), Cham: Springer, pp. 297-313.

Venable, J., Pries-Heje, J., and Baskerville, R. 2016. "FEDS: a Framework for Evaluation in Design Science Research," European Journal of Information Systems (25:1), pp. 77-89.

Watson, E. F., Chawda, P. P., McCarthy, B., Drevna, M. J., and Sadowski, R. P. 1998. "A Simulation Metamodel for Response-Time Planning," Decision Sciences (29:1), pp. 217-241.

Whitt, W. 1993. "Approximations for the GI/G/m queue," Production and Operations Management (2:2), pp. 114-161.

Wirth, R., and Hipp, J. 2000. "CRISP-DM: Towards a standard process model for data mining," in Conference on the Practical Applications of Knowledge Discovery and Data Mining, pp. 29-40.

Zhang, S., Zhang, C., and Yang, Q. 2003. "Data preparation for data mining," Applied Artificial Intelligence (17:5-6), pp. 375-381.