Systematic Literature Review of Verification and Validation of Simulation Models in Business and Manufacturing using Topic Modeling

Deepesh Gotherwal[1], Pritam Ranjan[1], Ryan Lekivetz[2]
[1]OMQT Area, Indian Institute of Management Indore
[2]DOE and Reliability, JMP Statistical Discovery LLC

ABSTRACT

Verification and validation (V&V) are integral parts of any simulation study. Validation assesses how accurately conceptual models represent the real system, while verification ensures correct implementation in software. V&V plays a critical role in business and manufacturing, where simulation models imitate complex real-world systems. However, comprehensive statistically grounded literature reviews on V&V of simulation models, particularly from a business management and manufacturing domain standpoint, are scarce. This study addresses that gap by performing *topic modeling* to identify prominent research themes, then reviewing all important research articles to outline the evolution of quantitative methodologies and algorithms on V&V. We also highlight various research gaps and potential directions for future work. For this study, we reviewed the abstracts of more than 6,000 articles indexed in Scopus and Web of Science, along with a comprehensive analysis of 300 research articles.

KEYWORDS

Operations research, statistical validation, design of experiments, metamodels, latent Dirichlet allocation, discrete-event simulation, system dynamics, agent-based modeling.

1. Introduction

Simulation models are commonly used in various application domains to represent *real-world systems or processes* that are typically too expensive to observe or experiment with (Banks, 1999). Law (2024) called *simulation* a method of last resort. Advancements in computing power have facilitated widespread adoption of simulation methods and tools (Mourtzis, 2020). Common simulation approaches include traffic simulation, simulation gaming, system dynamics (SD), discrete event simulation (DES), agent-based simulation (ABS), Monte Carlo simulation, Petri nets, etc. (Jahangirian et al., 2010). In *operations management (OM) and operations research* (OR), simulation is considered to be the second most popular technique, after mathematical modeling (Amoako-Gyampah & Meredith, 1989; Jahangirian et al., 2010; Pannirselvam et al., 1999).

While simulation has become an essential tool for practitioners, their involvement often leads to different levels of awareness of the complexities in developing a simulation model. Building a reliable and efficient simulation model for a complex system is challenging and calls for verification and validation (V&V). In this case, V&V of a simulation model refers to checking its validity, credibility, and usability, which is an important concern (Harper et al., 2021). Numerous scholars have defined the terms *"validation of a simulation model"* and *"verification of a simulation model"* based on their purposes. Sargent (2020) explains that *verification* ensures the conceptual model (abstraction of a real system) is correctly programmed in software, whereas *validation* checks how closely a conceptual model represents a real system. Verification techniques help debug simulation code, while validation

techniques assess the *appropriateness* of assumptions in the conceptual model and alignment with the real system. It is essential to acknowledge that a simulation model is intended to analyze a system for a specific purpose(s); therefore, its validity should be assessed solely for those purpose(s).

Robert G. Sargent, a pioneer in this field, introduced a simplified framework for the development of different phases of a simulation model (Sargent, 1981). Since its inception, this framework has been extensively used and refined by other researchers. The latest version of this framework, presented by Harper et al. (2021), is displayed in Figure 1. According to this framework, a simulation study comprises four stages. The first stage refers to defining and understanding the real-world problem. The second stage builds a conceptual model of the real-world problem. The third stage develops a computer program (i.e., simulation model) of the conceptual model. Finally, in the fourth stage, the simulation model output is analyzed and compared with the real-world data. One must be extremely careful while progressing from one stage to another as even small errors can lead to an unreliable model. To reduce this risk and increase trust, one must perform V&V at each stage of the framework.

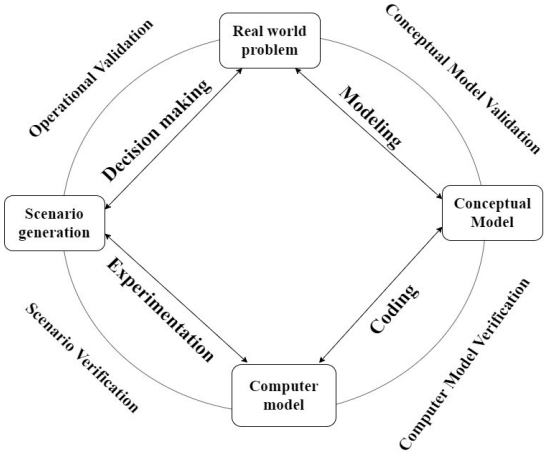

Figure 1. Different stages of the development of a simulation model (Figure 1 of Harper et al. (2021)).

Any study using a simulation model is "practically" incomplete without a proper V&V process (Law, 2022). However, many articles using simulation models either omit V&V or address it insufficiently (Sargent, 2020). Thus, the situation becomes increasingly challenging for more complex models (Brailsford et al., 2019). The limited attention to V&V and its insufficient application advocates the need to raise awareness about V&V among researchers and practitioners (Kabak et al., 2024).

This paper originates from a comprehensive literature review of simulation models in business and manufacturing. Throughout this process, we observed the need to raise awareness of V&V. This paper aims to introduce the concepts of V&V to practitioners who utilize simulations or contribute to model development but may overlook its role in ensuring reliability and accuracy. Even for those familiar with V&V, our systematic approach to literature review through data collection and topic modeling is relevant and raises important research questions.

Research Objectives: The main objectives of this paper are to (1) identify important V&V techniques and methodologies, (2) investigate the status of V&V of simulation models, focusing on those used in business management or manufacturing, and (3) uncover research gaps present in the current literature.

To our knowledge, this is the first systematic review of V&V in simulation models with business and management applications. Nine prominent research themes emerged from thematic analysis via topic modeling. A manual review of all methodological articles identified several important research gaps for future research. This review provides modeling researchers with a unified view of V&V techniques across DES, SD, ABS, and related simulation paradigms, supported by topic-model-based theme identification and a systematic corpus construction process.

Our approach for the systematic literature review can be broken into four steps:

(1) Collect data on published research articles from the *Web of Science* and *Scopus,* apply suitable filters to keep only relevant articles, and screen the remaining abstracts for analysis.

(2) Obtain bibliometric statistics, trends and insights from the database. Then perform a thematic analysis on the text corpus built using the titles, keywords, and abstracts of the relevant articles. The process starts with applying *topic modeling* using Latent Dirichlet Allocation (LDA) to identify an optimal number of prominent themes (or topics) in the research area, followed by a qualitative judgment-based construction of topic labels.

(3) Manually review all methodological articles to outline the evolution of quantitative methods and algorithms for V&V of simulation models. Most of the verification techniques are qualitative in nature, so more discussion has happened on the evolution of validation techniques than on the verification techniques. The validation methodologies have been grouped together based on the data availability of the real system that is being emulated using the simulation model.

(4) Identify potential and intriguing research gaps in literature that can yield crucial future research directions.

The remainder of the paper is organized as follows. Section 2 presents the background on V&V of simulation models with a focus on the business and manufacturing domains, and the research questions that motivate our study. Section 3 discusses the data collection process and a brief descriptive analysis of the data on 300 articles. Section 4 is devoted to topic modeling and the qualitative construction of nine topic labels. Section 5 outlines the evolution of V&V methodologies. Major research gaps are discussed in Section 6, and Section 7 concludes with important remarks and limitations.

2. Background

Research on V&V of simulations began in the 1960s, with Fishman and Kiviat (1968) providing its formal definition. Research on V&V gained momentum when epistemological questions like "Why are so many models built and so few used?" started appearing in the literature (Landry et al., 1983). Dimensions like *credibility* and *acceptability* (Robinson, 2002), *representativeness and usefulness* (Landry et al., 1983), *efficiency* and *effectiveness* (Landry and Oral, 1993), and *trust* (Harper et al., 2021) have been discussed in the literature. In 1993, the *European Journal of Operational Research* (EJOR) published a special issue on "Model Validation in Operational Research", where papers primarily discussed model validation from an epistemological perspective. These papers are summarized by Landry and Oral (1993). The *Winter Simulation Conference* (WSC) has since emerged as a prominent platform for discussion and publication on V&V methods with new approaches introduced almost every year. These techniques vary by the type of simulation method, context, data availability, and several

other factors (Roungas, 2016). The literature also discusses quantitative techniques based on various statistical tests and methods, qualitative approaches such as face validity through experts, and independent verification and validation (IV&V) by third parties (Robinson & Brooks, 2010). In this context, scholars such as Robert G. Sargent, Jack P.C. Kleijnen, and Osman Balci have made significant contributions to literature by proposing a wide spectrum of V&V techniques for different stages of simulation model development (Figure 1). The timeline of key publications is depicted in Figure 2.

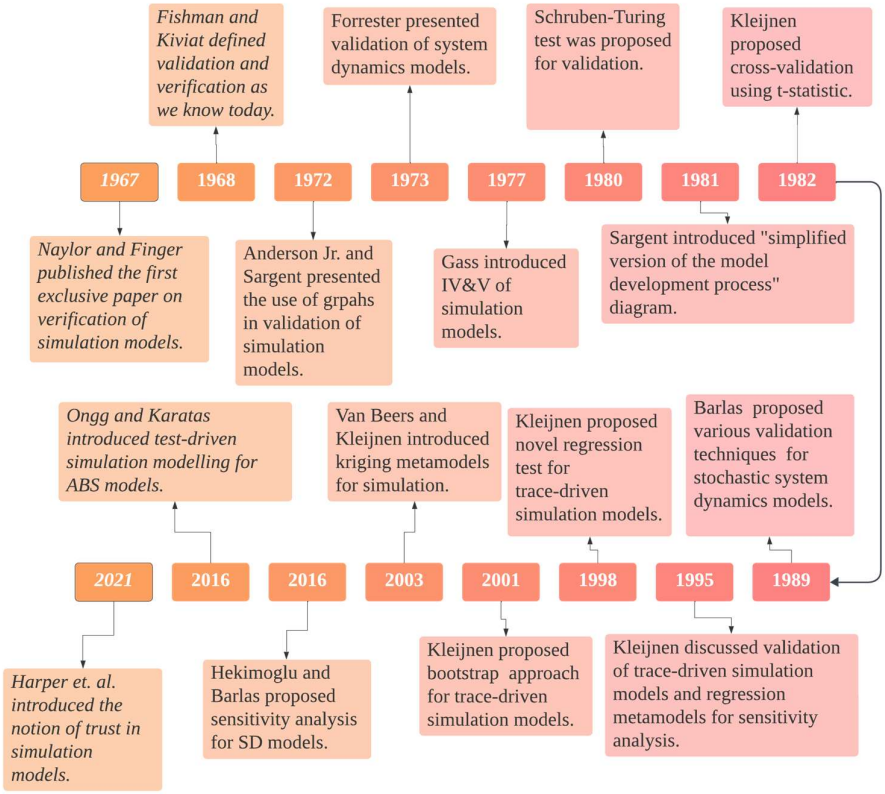

Figure 2. Timeline of a few key publications.

A variety of simulation techniques have been used across domains. According to Jahangirian et al. (2010), frequently implemented techniques for simulation models in business and manufacturing include discrete event simulation (DES), system dynamics (SD), hybrid, agent-based simulation (ABS), Monte Carlo simulation, Petri nets, and intelligent simulation. DES is the most used simulation method (approximately 40%). Balci (1995) categorized various V&V techniques into different segments (Table 1), and this list remains widely used. For a detailed history of V&V of DES simulation models, refer to Sargent and Balci (2017). While many techniques apply to DES, they are also applicable to other simulation methods. SD is the second most used simulation method, and Yaman Barlas is one of the pioneers in proposing various V&V techniques for SD models, including a six-step behavior validation procedure (Barlas, 1989). The hybrid simulation method combines two or more standalone simulation methods (such as DES, SD, ABS, etc.) to model a complex real-world system. However, V&V of hybrid models is rarely reported (Brailsford et al., 2019). ABS is based on DES and object-oriented programming (North & Macal, 2007), so typically use V&V techniques developed for DES.

Table 1. Categorization of V&V techniques (Figure 1 of Balci (1995)).

Informal	Static	Dynamic	Symbolic	Constraint	Formal
Audit	Consistency checking	Black box Testing	Cause-Effect Graphing	Assertion checking	Induction
Desk Checking	Data Flow analysis	Bottom-Up testing	Partition Analysis	Boundary Analysis	Inference
Face Validation	Graph-Based Analysis	Debugging	Path Analysis	Inductive Assertions	Lambda Calculus
Inspections	Semantic Analysis	Execution Monitoring	Symbolic Execution		Logical Deduction
Reviews	Structural Analysis	Execution Profiling			Predicate Calculus
Turing tests	Syntax Analysis	Execution Tracing			Predicate Transformation
Walkthroughs		Field testing			Proof of Correctness
		Graphical Comparisons			
		Predictive Validation			
		Regression Testing			
		Sensitivity Analysis			
		Statistical Techniques			
		Stress Testing			
		Submodel testing			
		Symbolic Debugging			
		Top-Down Testing			
		Visualization			
		White-Box Testing			

To summarize, various simulation methods have been employed across sectors, and the literature offers diverse strategies to enhance the credibility of simulation models. Despite this there is an evident lack of a comprehensive review on V&V of simulation models that can address the following research questions:

- RQ1: What are the prominent research themes in V&V literature?
- RQ2: How have V&V methodologies evolve over time?
- RQ3: What major research gaps remain for future work?

This paper uses a combination of qualitative and quantitative analyses of 300 carefully chosen articles to answer these research questions.

3. Data collection and insights

This section describes the data collection procedure and important descriptive summary of the data on V&V of simulation models for business management and manufacturing applications. Inspired by the studies in Liao et al. (2017), Han et al. (2020), Mustak et al. (2021), Dohale et al. (2022), Naz et al. (2022), Psarommatis and May (2023), we follow a four-step approach for data collection: (1) database search, (2) applying filters, (3) remove duplication, and (4) abstract screening.

3.1. Data collection

Two popular databases, Scopus and Web of Science (WoS), were used to identify relevant research articles. We used three keywords to search for papers in the databases: “Validation,” “Verification,” and “Simulation model”. The Scopus database offers comprehensive coverage of over 4,000 publishers and includes approximately 15,000 reputable peer-reviewed journal articles, monographs, conference proceedings, etc. The WoS database includes documents from over 3,300 publishers and more than

12,000 high-quality research articles. All WoS articles meeting our criteria were also available in Scopus, therefore, we limit our discussion to Scopus for brevity.

Database search: The data collection process in Scopus started with search strings "*Verification* and *Simulation model*" and "*Validation* and *Simulation model*" to find all relevant manuscripts. The search resulted in 4331 and 8314 documents, respectively. We now pass these documents through a few filters to ensure relevance and quality.

Applying filters: We restricted our database to journals and conference proceedings, excluding trade journals, books and reports. Furthermore, we limited document types to only articles, reviews, and conference papers for this study. Next, a subject area filter was applied to include relevant articles and exclude the items that appeared in unrelated areas (e.g., agriculture). The retained articles consist of papers on V&V of simulation models with (a) interesting real-life applications and (b) novel methodologies. For real-life application-oriented papers, we focused on business management and manufacturing domains. In the Scopus database, *Engineering* contains two very relevant domains, i.e., 'operations research' and 'industrial and manufacturing engineering', whereas *Computer Science* includes 'modeling and simulation' and 'software'. The methodologically novel results and innovative algorithms may be published in core *mathematics* journals and not necessarily in the application areas. Simulation models are also commonly used in physics-based applications to study complex processes / phenomena that are otherwise too expensive or even infeasible to observe. Therefore, we explored the allied area '*physics and astronomy'*.

Remove duplication: The search string "*Verification and Simulation model*" produced 2857 articles, whereas search "*Validation and Simulation model*" resulted in 4831 documents. Of the 7688 articles, only 6330 documents were unique, and complete with respect to title, abstract and year of publication.

Abstract screening: We screened the abstracts of 6330 articles and, when necessary, diligently reviewed the full text. Using the practices and criteria used by Naylor and Finger (1967), Landry et al. (1983), Barlas (1989), Kleijnen (1995), and Sargent (2010), we tabulated a set of criteria for further shortlisting the relevant articles (see Table 2).

Table 2. Abstract screening criteria.

SN	Inclusion Criteria	Rationale
1	Proposed a new methodology or improved existing methodology for validation or verification of simulation models.	This study aims to study different V&V techniques.
2	Discuss DOE, sensitivity analysis, or use of metamodels in the context of simulation models.	These methods are frequently used as V&V techniques in the absence of real-world process data.
3	Focus on business and/or manufacturing applications.	The study focuses on V&V techniques in business and management context.
4	Applications in related disciplines like engineering, physics and astronomy, etc.	Want to include any V&V techniques used in the related area.
5	Discuss V&V theory by using the keywords such as quality, trust, accuracy, calibration, confidence, performance, and qualification for a simulation model in the context of business management or manufacturing.	To include those papers that discussed V&V from an epistemological view.

It is clear from Table 2 that Criteria 1 and 2 focus on articles with a methodological contribution on V&V, whereas Criteria 3 and 4 support the articles on business management, manufacturing, and related

area applications. The fifth criterion includes those papers that discussed V&V through epistemological keywords like quality, reliability, calibration, etc. If the abstract of an article satisfied one or more criteria, it was included in the finalized list. Ultimately, a total of 299 papers were selected after the abstract screening process. We manually included Naylor & Finger (1967) which appeared in neither database (WoS nor Scopus) but is an important contribution to the field. Figure 3 presents the flow diagram of the data collection process.

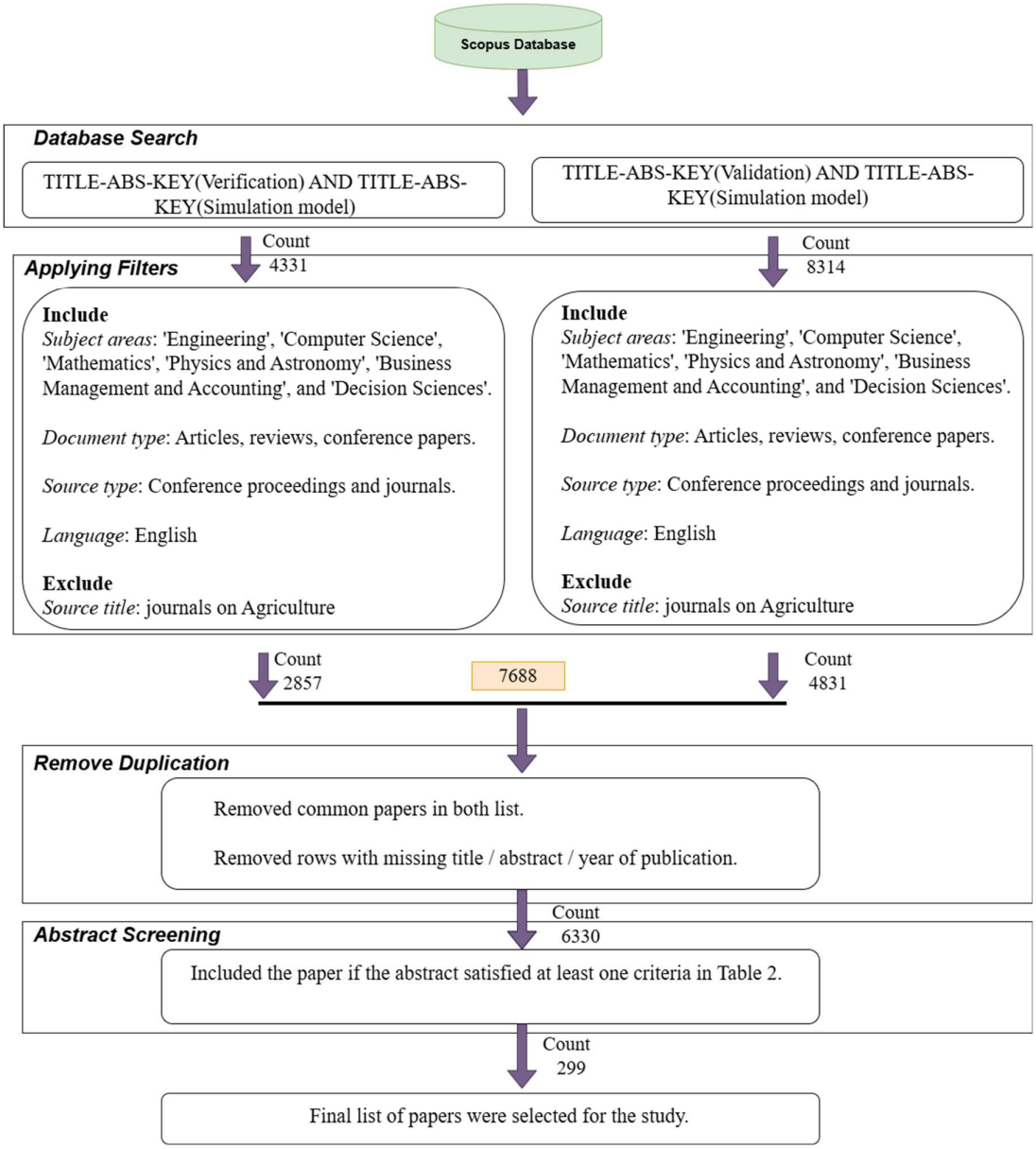

Figure 3. Search and selection process of research articles from the Scopus database.

We manually categorized the final list of 300 papers (299 plus one) into five subject areas. Table 3 presents the distribution of papers from different domains.

Table 3. Subject area-wise distribution of articles in our database.

Subject Areas	Count
Business Management	34
Computer Science (modeling and simulation, software)	164
Industrial and manufacturing engineering	21
Other engineering (including Operations Research)	21
Mathematics	60
Total	**300**

The majority of the articles come from Computer Science, which includes 'modeling and simulation' and software development papers, and Mathematics, which mostly comprise of methodology papers. The application papers fall under Business Management and Manufacturing areas, whereas Engineering area papers contain a mix of both methodology and applications.

3.2. Data insights

We used the Bibliometrix package of *R* (Aria & Cuccurullo, 2017) to summarize the bibliometric records of 300 articles. Some of the notable findings are as follows. Although the first paper on V&V of simulation models was published in 1967 (Naylor & Finger, 1967), there was a significant increase in the average number of research papers published after 1980 and has remained steady. Nearly 45% of the articles have been published in "*Winter Simulation Conference*" proceedings, followed by journals such as the *European Journal of Operational Research, Simulation Series, Journal of the Operational Research Society,* and *International Journal of Production Research*. The same trend has also been observed by dos Santos et al. (2022). Five authors – Robert G. Sargent, Jack P.C. Kleijnen, Osman Balci, Yaman Barlas, and Stewart Robinson, have contributed to approximately 31% of all articles (see Figure 4).

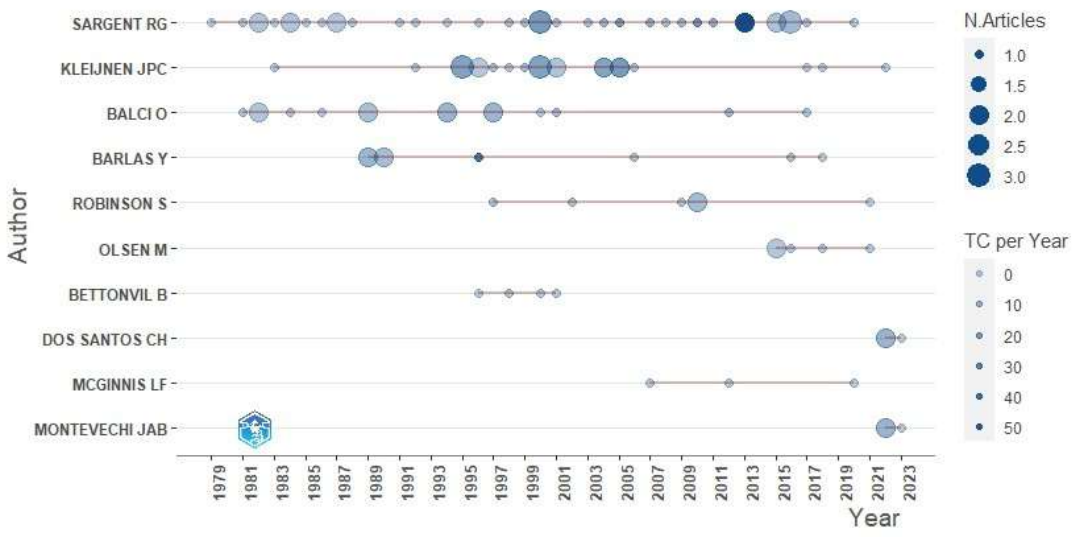

Figure 4. Prominent authors who published on V&V of simulation models (graph generated using bibliometricx package).

VOSviewer (Van Eck & Waltman, 2010) was also used to generate detailed scientometric patterns on co-authorship, co-citation, etc., but instead of discussing such summaries, we focus on topic modeling of the text corpus to identify prominent themes in this area (to answer RQ1), then we manually review the relevant articles to trace the evolution of methodologies on V&V of simulation models (to answer RQ2).

4. Topic Modeling

Topic modeling is an objective approach to analyze a large corpus of text data that can otherwise be extremely challenging to process manually (Jelodar et al., 2019; Maier et al., 2021; Vayansky & Kumar, 2020). Using Latent Dirichlet Allocation (LDA), a widely used natural language processing technique for topic modeling (Vayansky & Kumar, 2020), we extracted the latent and semantic themes (or topics) from the corpus built using 300 selected articles. Alternative methods include Latent Semantic Analysis and Probabilistic Latent Semantic Analysis (Churchill & Singh, 2022).

Our corpus of text for topic modeling includes the titles, abstracts, authors' keywords, and indexed keywords. The corpus of data had to be cleaned before applying LDA for topic modeling. First, all punctuation marks such as @, ",", $, were removed. Second, the entire corpus was converted to lowercase. Next, the corpus was tokenized (i.e., split into individual words), wherein the collection of all tokens (*words*) is known as a bag of words. The final pre-processing step involved the removal of stop words (common English words such as 'a, 'an, 'the', and others). After all the pre-processing, we built a *document-term matrix* (DTM) constructed as a list of unique words per article and a *dictionary* of unique words.

The DTM and the dictionary were passed to *LdaModel* API of the *Gensim* (Řeȟuřek & Sojka, 2010) library in Python for generating important themes (or topics). LDA assumes that the collection of D articles (i.e., corpus) is statistically generated by t distinct themes, and each theme is statistically generated by w distinct words. See Jelodar et al. (2019) for details. The algorithm generates two probability distributions, θ_d – for each document $d \in D$, and ϕ_t – for each of t distinct themes over all distinct words present in the corpus. Typically, Dirichlet distributions are used for the two probability distributions, and the parameters represent topics per document ratio and words per document ratio, respectively (Churchill & Singh, 2022). The number of topics (k) to be extracted from the corpus must also be pre-specified while running the LdaModel. The output of LdaModel is a list of k sets of words, $T_1, T_2, \ldots, T_k$, where T_i consists of the most consistent words defining a particular topic (theme). Section 4.1 deliberates on how to objectively arrive at the optimal number of topics, while Section 4.2 discusses the suitable labeling of the themes based on T_i and requires qualitative input from a subject expert.

4.1. Optimal number of topics

The topic modeling literature discusses various measures to determine an optimal number of topics. Churchill and Singh (2022) broadly classify these evaluative measures into three different categories: "coherence", "coverage", and "qualitative". Interestingly, the quality of topics cannot be solely

measured by quantitative methods, and human intervention is also required to find the optimal number of topics (Blei et al., 2003; Churchill & Singh, 2022).

A set of words is called *coherent* if they fit together meaningfully in an interpretable way. The coherence calculates a topic's score by comparing the semantic similarity of its words. For example, the set of words $\vec{v}_1$ = *{sport, match, players, stamina}* makes more sense than $\vec{v}_2$ = *{sports, ice-cream, fashion, stadium}*. Thus, $\vec{v}_1$ is more coherent than $\vec{v}_2$. The coherence score, C_v, uses word segmentation techniques such as *sliding window* and *Normalized Pointwise Mutual Information* (NPMI) for evaluating the co-occurrence frequencies of words. C_v of a word vector $\vec{v}$ quantifies the likelihood of a set of words in $\vec{v}$ occurring together in the corpus. Mathematically, the coherence score computation evaluates NPMI for every pair of words in the word vector $\vec{v}$ and then finds its average cosine similarity. See Campagnolo et. al. (2022) for details. The average coherence score of the k topics identified by the LDA model quantifies the merits of the set of topics (or themes). The steps of topic modeling, from building a corpus to finding the optimal number of topics, are summarized in Algorithm 1.

Algorithm 1: Identification of Prominent Themes

Input: A set of selected articles.

Output: The set of optimal topics $S_{k_*} = \{T_1, T_2, \dots, T_k\}$.

1. **Define** D as the list of 300 documents created using the title, abstract and keywords of articles.

2. Pre-processing:

a. Remove punctuation from D.

b. Convert all text in D to lowercase.

c. Remove standard stop-words (e.g., "a", "an", "the").

d. Split the cleaned text into individual tokens.

3. **Construct** a Document-Term Matrix (DTM) of size 300 × unique words using the tokens.

4. **Construct** a Dictionary ⇐ list of unique words across all articles.

5. Topic Modeling and Coherence Calculation:

for k = 2 to 30

a. Generate topic set $S_k = \{T_1, T_2, \dots, T_k\} \Leftarrow LDAmodel(Dictionary, DTM)$.

b. Compute the average coherence score $C_v(k)$ for the topic set S_k.

end for

6. Determine Optimal Topics:

a. Find S_{k_o}, where $k_o = argmax\{C_v(k), k = 2, \dots, 30\}$.

b. Refine S_{k_o} based on subjective assessment of resulting topics to obtain S_{k_*} with $k_* \leq k_o$.

7. **Return** S_{k_*}.

We analyzed the distribution of average coherence of the 'best topics found by LDA' versus k, ranging from 2 to 30. The objective would be to find k that maximizes the average coherence score. To obtain

a robust estimate of the optimal number of topics, we repeated the entire procedure of C_V computation ten times with different random seeds and computed the mean C_V curve along with their 95% confidence intervals (see Figure 5). We generated such curves for three other coherence scores C_{NPMI}, C_{UCI}, and C_{UMass}, that measure semantic similarity using different size or type of sliding window and either used NPMI or pointwise mutual information (Campagnolo et. al., 2022). Figure 5 presents the standardized values of these four measures for a fair comparison. All these curves exhibit a similar pattern in terms of maxima, minima, and the change of slope.

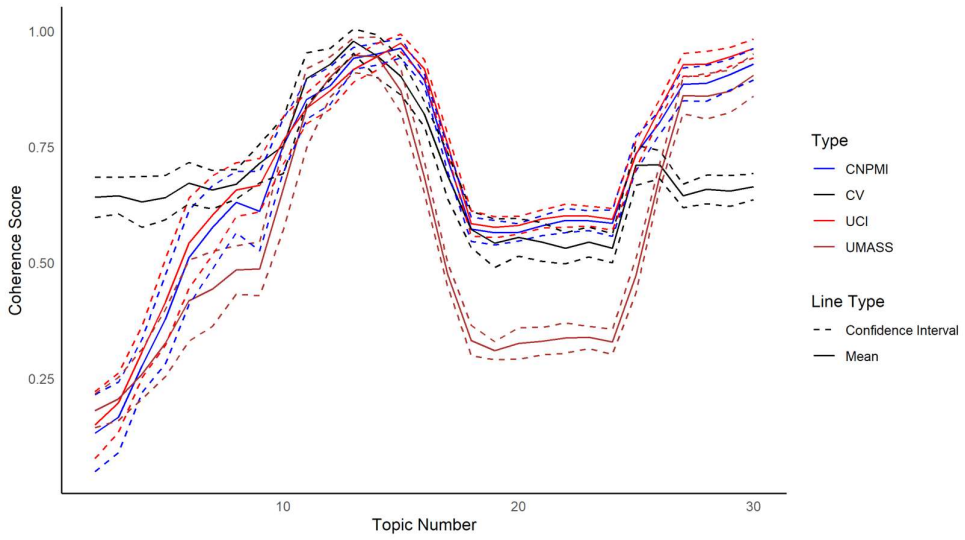

Figure 5. Distribution of C_V, C_{NPMI}, C_{UCI}, and C_{UMass} with respect to the number of topics $k \in \{2,3,\ldots,30\}$.

It is clear from Figure 5 that the average coherence curves exhibit multiple peaks, with the first peak occurring at $k \in \{12,13,14\}$. The coherence then increases to similar maximal values after $k = 27$. The reason for the second peak is perhaps due to the formation of duplicate topics with fewer documents. Thus, simply maximizing the average coherence score with respect to k appears to be inadequate (for our dataset), and subjective human judgment would be required to find the optimal number of topics. The qualitative assessment of the topics obtained by the analysis revealed that the themes start to repeat for $k > 9$. As a result, we used $k = 9$ as the optimal number of topics to properly capture the prominent themes from our database of 300 articles.

The LDA model generates a set of keywords for each topic (Churchill & Singh, 2022). We labeled these topics based on our experience. For instance, the second topic in the final list of nine topics comprised {dynamic, management, development, procedure, structure, credibility, theory, operational, concept, structural}. Keywords like 'dynamic', 'structure', 'structural', and 'credibility' suggested a topic of 'Validation of SD models', as the validation of SD models' literature focuses on the structural nature or structure of SD models. Table 4 presents the keywords generated by the LDA model for each topic, topic labels assigned by us, a list of noteworthy publications, and the count of papers related to each topic found by the *Gensim lda package*. Note that one paper may contribute to two more themes. Section 4.2 discusses these labeled topics in more detail and is a response to the first research question (RQ1). This thematic labeling helped us in categorizing and summarizing the vast literature on V&V on simulation modeling, which in turn facilitated the understanding of the evolution of V&V methods

discussed in Section 5. The model was trained with nine topics using a chunksize of 10 and 10 passes over the corpus, with a symmetric Dirichlet prior of 0.2 for the document–topic distribution (α). The model was further configured with 100 iterations and a fixed random state of 100 to maintain consistency across runs.

Table 4. Prominent topics on V&V of simulation models.

S.No.	Topic label	Keywords	Exemplary studies	Paper count
1	V&V of DES models	Behavior, design, queue, base, mathematical, issue, manufacturing, analysis, function, propose	Balci (1995); Raunak and Olsen (2014); Robinson (2002)	15
2	Validation of SD models	Dynamic, management, develop, procedure, structure, credibility, theory, operational, concept, structural	Barlas (1989, 1996); Hekimoğlu and Barlas (2016)	24
3	General V&V literature	Simulation, validation, verification, process, technique, present, paper, problem, result, research	Balci (2012); Naylor and Finger (1967); Robinson and Brooks (2010); Sargent (2010, 2013, 2020)	170
4	Verification techniques	Validation, validity, datum, database, verification, compare, relate, perspective, information, data	Akbulut et al. (2017); Barad (1998); Yeung (2011)	8
5	Validation using metamodels	Analysis, statistical, simulation, test, output, computer, operation, method, design, experiment	Kleijnen (1995); Kleijnen and Sargent (2000); Kleijnen and Deflandre (2006)	48
6	Epistemological view of V&V	Knowledge, prototype, understand, observe, effort, empirical, multiple, question, simplification, relative	Harper et al. (2021); Landry et al. (1983); Landry and Oral (1993); Oral and Kettani (1993)	4
7	V&V of simulation models in business and manufacturing	Business, improvement, area, petri net, philosophy, traffic, change, mapping, lean, forecast	dos Santos et al. (2024); Friederich et al. (2022); Sarnow and Elbert (2022)	7
8	V&V of simulation models in health or traffic management application	Formal, pattern, simple, check, medical, domain, health, patient, specification, multi	Brügmann et al. (2014); Chou et al. (2001); Swisher et al. (2001)	8
9	V&V of ABS models	Computer, test, software, simulation, base, evaluation, agent, give, engineering, quality	Ongg and Karatas (2016); Troost et al. (2023); Volkmakarewicz and Cleophas (2017)	16

4.2. Thematically labeled topics

Of the nine topics, there are separate themes on V&V of DES models, SD models, and ABS models; two themes for "verification methodology" and "validation using metamodels," respectively; one theme comprises general V&V literature; and a few other themes address interesting health, traffic, business management, and manufacturing applications. Next, we briefly review important citations from each topic.

4.2.1. V&V of DES models

Radiya and Sargent (1987) developed a formalism for DES models based on six different elements to analyze and validate a DES model. Barad (1994) introduced decomposing Timed Petri nets-based methodology for the verification of a DES model. However, the methodology's shortcoming is that it only applies to steady-state and non-terminating DES models. Jacobson and Yücesan (1999) discussed the validation of a DES model structure using the theory of computational complexity to analyze different components of a DES model. Robinson (2002) explored the quality dimension in terms of three components: content, process, and outcome of a DES model for business applications. Chwif et al. (2006) proposed a prescriptive V&V technique for a DES model when there is no data on the real system being studied. Strang (2012) showed how a basic spreadsheet tool can be used to verify the distributional assumptions of arrival and service rate. Can and Heavey (2012) compared genetic algorithms and artificial neural networks for building metamodels for DES models. Raunak and Olsen (2014) developed criteria to measure the validation of a DES model. Montevechi et al. (2022) proposed "Generative Adversarial Networks" to compare the DES model output to the real-system output. Law (2024) discussed various tests, such as the Kolmogorov-Smirnov test, spectrum test, lattice test, etc., for assessing the quality of random number generator – an important component of DES model validation. Rowe and Wright (2001) discuss a method whereby expert analysis can be used to validate simulation models independent of the input and output data. Balci (1995) and Raunak and Olsen (2014) emphasized face validation of input and output data through subject matter experts, goodness-of-fit tests, and sensitivity analysis. Lastly, Lugaresi et al. (2022) presented an innovative online validation technique for evaluating the accuracy of DES models used for manufacturing plants.

4.2.2. Validation of SD models

The validation of SD models focuses on the conceptual modeling part (Tako and Robinson, 2010), as the peculiarities of SD models make ordinary statistical tests unfit for validation (Barlas, 1989). Taylor (1983) is one of the earliest papers to discuss the validation of SD models and introduced the concept of "loop analysis" for this purpose. Barlas (1989) classified validation tests for such models into two categories: structural validity tests and behavior validity tests. Barlas argued that the logical sequence of tests should be as follows: first, structural validity tests; second, structurally oriented behavior tests; and finally, pattern prediction tests. The paper proposed a six-step behavior pattern validation procedure designed to overcome the limitations of standard statistical tests, which often assume normality,

independence, and stationarity of data. See Barlas and Carpenter (1990); Barlas (1996) for more details. Hekimoğlu and Barlas (2016) demonstrated sensitivity analysis of SD model output behavior patterns using a linear regression model. Edali (2022) addressed the issue of manually identifying the parameter space for different output behavior models in sensitivity analysis and proposed a random forest-based metamodel approach.

4.2.3. General V&V literature

This theme consists of papers that discuss the verification and validation (V&V) of simulation models, regardless of the simulation technique or area of application, including seminal contributions by Robert G. Sargent and Osman Balci. The maximum number of papers have been assigned to this theme by the *Gensim lda package*, which is expected given that papers in our database discuss V&V without focusing on any specific simulation technique or application. Naylor and Finger (1967), the oldest work in our database, discussed validation from a philosophical viewpoint, such as "Rationalism", "Empiricism," and "Positive Economics," and various goodness-of-fit tests, such as "Analysis of Variance" and the "Kolmogorov-Smirnov Test," among others, for validating simulation models. Sargent has consistently discussed the inclusion of V&V at every stage of the simulation study, ranging from Sargent (1981) to Sargent (2020). The original framework in Sargent (1981), defined four distinct phases of V&V: "Data Validation", "Conceptual Model Validation," "Model Verification," and "Operational Validation." The data validation process checks both the quality of the original data and the accuracy of any transformation process performed on real data. Synthetic data is tested using an appropriate experimental design. Conceptual model validation evaluates the model's underlying theories, assumptions, mathematical structure, logic, links, and causal relationships using face validation, traces, and distribution fitting to data at the model building stage. The use of advanced simulation software helps reduce model verification efforts; however, static and dynamic testing can be used to check whether the conceptual model has been correctly coded and executed. The final step, operational validation, verifies whether the model output is reasonable over the system's range by comparing the real and simulated outputs graphically, conducting hypothesis tests, and calculating confidence intervals. A more complex framework was developed by Sargent (2013) that included separate structures for "real system" and "simulation model." Robinson and Brooks (2010) discussed the idea of independent V&V (IV&V) of an industrial simulation model developed for nuclear decommissioning and waste management. The IV&V review emphasized ensuring that the right V&V steps were taken during the model building process, rather than doing "after-the-fact V&V". Balci (2012) explored twelve major processes of the lifecycle of large-scale complex simulation models from problem formulation to model building. Fonseca (2023) discussed a taxonomy comprising three categories of assumptions: "Systemic Data assumptions," "Systemic Structural assumptions," and "Simplification assumptions." The paper presented a "W3H testing table" that facilitates the systematic selection of validation tests for each category of assumption. The authors suggested an "assumptions table" as a systematic method for recording and overseeing assumptions, encompassing their validation status and

review criteria. This is especially crucial in the realm of Industry 4.0 and digital twins, where the simulation model requires ongoing validation as the actual system progresses.

4.2.4. Verification techniques

Woodward and Mackulak (1997) proposed event graphs based on reverse engineering to detect logical errors in both static and dynamic properties of a DES model code. Krishnamurthi and Thallikar (1998) proposed an algorithm to detect deadlock in DES models and integrated it into the SIAM commercial simulation language. Barad (1998) proposed Timed Petri nets (TPN) as an analytical verification technique, specifically for steady-state DES or queuing simulation models. Krahl (2005) discussed various techniques to debug a simulation mode, such as "read the manual", animation, model traces, preemptive bug prevention, etc. Yeung (2011) addressed deadlocks for a multi-agent manufacturing system. Roungas (2016) suggested an automated verification technique based on the unit testing method. Yacoub et. al. (2020) developed a DES model-based formalism in *PROMELA* to verify sophisticated software like video games. Sargent (2020) summarized the current verification based on the type of approach used to build the simulation model. The first approach is to use high-level programming languages such as Python, R, etc. to build a complete simulation model. The second one is using "simple" simulation language that provides some basic functionality such as event or time flow routines, or random number generators. The third approach is to use "advanced" simulation packages that provide some advanced functionality, such as model execution or graphic capabilities. The final approach is to utilize full blown simulation packages, such as AnyLogic or Simio, which offer comprehensive functionalities including model building, execution, testing, scenario generation, and animation. It is expected that the verification efforts will reduce as one moves from the first approach to the fourth.

4.2.5. Validation using metamodels

A metamodel is a statistical surrogate of a simulation model that approximates the relationship between the inputs and outputs of a process/system/phenomenon that aim to capture the essential characteristics (Kleijnen, 1995 and Kleijnen, 2009). Metamodels are widely utilized in the simulation literature for four primary purposes: (a) understanding the structure and input-output relationship of a system, (b) prediction, (c) optimization, and (d) validation of a simulation model (Kleijnen & Sargent, 2000). When used in conjunction with the design of simulation experiments (also referred to as computer experiments), metamodels can be used to perform a sensitivity analysis of a simulation model (Kleijnen, 1995). Jack P.C. Kleijnen is a pioneer in this field. Kleijnen (2009) mainly discussed regression and kriging metamodels. Other types of metamodels for validation include Bayesian metamodeling (Pousi et al., 2013), Multiple Regression Integrated K-Means Clustering Algorithm (Irfanoglu et al., 2013), genetic programming-based, ANN-based metamodels for DES models (Can and Heavey, 2012), and Random Forest metamodels (Edali & Yücel, 2019). This category ranks second in paper count as the kriging metamodel is a much-discussed topic in core statistics and machine learning literature.

4.2.6. Epistemological view for V&V

Landry et al. (1983) started a discussion on the dimensions of trust, credibility, and confidence in simulation models. Gass (1977) and Harper et al. (2021) advocated the collaboration between the modeler and different stakeholders (e.g., domain experts, end users, or practitioners) during different stages of simulation studies to help build trust in the simulation models. Landry and Oral (1993) further explored the validation of models from an efficiency (doing things right) versus effectiveness (doing the right thing) standpoint.

4.2.7. V&V of simulation models in business and manufacturing

Production, logistics, and manufacturing applications are dominant in literature, with DES models most commonly used to build simulation models. Cochran (1987) proposed two quantitative V&V techniques, based on the Turing test and mathematical programming, for validating large-scale production simulation models. Rabe et al. (2008) claimed that less attention had been given to V&V of simulation models in production and logistics compared to defense applications. Sarnow and Elbert (2022) developed a qualitative framework for V&V of "generic simulation models" (GMs) - a type of DES model - for solving a logistics problem. This study highlighted the potential of reusable models in operational decision-making processes. dos Santos et al. (2024) used "K-Nearest Neighbors (K-NN)" classification with a "p-control chart" to periodically validate these types of simulation models. Friederich et al. (2022) proposed a data-driven framework for building simulation models as a basis for digital twins for smart factories. Bitencourt et al. (2024) presented a systematic literature review on V&V of digital twin models for manufacturing applications.

4.2.8. V&V of simulation models in health and traffic management application

This topic includes papers that used simulation models in different health or traffic management applications and discussed their V&V. Traffic-highway and health simulation models are common in operations management, decision sciences, information systems, and engineering, and share methodological similarities to our primary applications of interest in business and manufacturing.

For instance, Fitzsimmons (1971) designed a simulation model (SIMSCRIPT) to assess emergency medical systems. Swisher et. al. (2001) built a DES model of a physician's clinic and validated it using various tests from Balci (1998). Bountourelis et al. (2014) modeled an ICU (intensive care unit) that incorporated two key parameters: patient blocking and bed occupancy. They utilized animation for conceptual model validation and graphical comparison of key parameters with real data for operational validation.

Ryan (1979) employed graphical methods to validate an existing bus operation simulation model. Rousseau and Bauer (1996) discussed factor analysis and design of experiments for sensitivity analysis of multivariate output from large-scale transportation simulation models. Afshar and Azadivar (1992) developed a microscopic traffic simulation model to investigate freeway work zones with lane closures and safety hazards, which was validated against field data using graphical tools. Rao et al. (1998) discussed a multistage validation framework for traffic simulation models that consisted of conceptual and operational validations using a *t*-test and a Kolmogorov-Smirnov test. Chou et al. (2001) built a

computer simulation model of pre-timed traffic signals for an urban city in Taiwan to analyze and reduce the average waiting time of vehicles at intersections of the city, validating it via a chi-square goodness-of-fit test. Brügmann et al. (2014) developed a traffic simulation package in Maude, a declarative programming language, utilizing state-of-the-art techniques for the verification and validation (V&V) of the model.

4.2.9. V&V of ABS models

An ABS model is typically a combination of a DES model with object-oriented programming. Thus, many of the V&V techniques for DES are also applied to ABS. However, validating the rules that regulate the agents of a model is the primary concern in ABS models (Ongg and Foramitti, 2021). Verification techniques such as structured code and debugging walk-throughs, unit testing, etc., are popular in this literature (North and Macal, 2007). Arifin et al. (2010) is one of the earliest articles that proposed a V&V methodology specifically for the ABS model. They proposed a *docking* technique that compartmentalizes a simulation model, blocking the flow of errors from one compartment to another. Other notable methodological contributions include parameter sweeping, white-box validation, black-box validation, code debugging (Gerrits et al., 2017), a generic testing framework (Gürcan et al., 2013), a modified metamorphic technique (Olsen & Raunak, 2016), a test-driven approach (Ongg & Karatas, 2016), and a meta-algorithm (Volkmakarewicz and Cleophas, 2017). Gore et al. (2017) introduce a statistical debugging–based verification approach for ABS that systematically analyses execution traces to identify behavioural deviations at the agent level. By linking individual agent interactions to emergent macro-level outcomes, their method enables quantitative trace validation specifically suited to the path-dependent and non-linear dynamics characteristic of ABSs.

5. Evolution of V&V methodologies

We manually reviewed the relevant articles in our database to outline the evolution of quantitative methods developed in the literature for V&V of simulation models. This refers to our research question RQ2.

5.1. Verification methodologies

The objective of verification is to find and correct bugs or errors in simulation code. The majority of the verification techniques in the literature are qualitative in nature or present general debugging guidelines (Balci, 1995 and Roungas, 2016), but none are without issues (Kleijnen, 1995).

From a quantitative approach standpoint, Kleijnen (1995) discussed testing of subroutines that generate pseudorandom and non-uniform distributed random numbers. Krishnamurthi and Thallikar (1998) proposed a deadlock detection algorithm and interfaced it with the commercial simulation

language “SIMAN” for DES models. Yeung (2011) discussed the issue of deadlocks in multi-agent manufacturing systems and proposed a formal verification procedure to identify the same.

Other articles presented interesting verification methodologies that compared the simulation model output with that of a metamodel, a new simulation model, or a mathematical model. For instance, Barad (1998) used Timed Petri nets (TPNs) to decompose a queue simulation model into a Petri net graph and compared its results with the simulation model output. Akbulut et al. (2017) proposed event-oriented model building tools such as “Simulation Graphs” as compared to “process-oriented” models like AnyLogic or Simio, comparing the outputs to verify their simulation models. Henderson and Bryce (2019) used a DES model called the “Force Flow Model” (FFM) to simulate the manpower dynamics of the Canadian Armed Forces and compared the simulation model output with the analytical results obtained from a differential equation model.

5.2. Validation methodologies

Recall that validation refers to assessing how *close* a conceptual model is to a real system (Sargent, 2020). Assuming that the conceptual model is coded (programmed) correctly, the objective of “validation” is to assess the accuracy of the simulation model. The applicability of validation technique depends on context, such as the type of simulation (i.e., DES, SD, hybrid, ABS, etc.) and the availability of inputs and outputs of the real-world system. Data from the real system plays a vital role in validating a simulation model (Kleijnen, 1995). In the categorizing task of topic modeling, Topics 2 and 5 (and partly Topic 1) revealed validation methods vary based on the availability of real-system data. Therefore, we present the evolution of validation methodologies under three categories: (a) both input and output of real-world system data are available, (b) only real-world system outputs are available (and not the inputs), (c) neither input nor output data of the real-world system are available.

5.2.1. Availability of both input and output real data

This scenario, known as a *trace-driven* simulation study, involves running the simulation model on a given set of inputs and comparing the output set with the real-system output.

The relevant literature starts with a basic Student’s *t*-test proposed by Kleijnen (1995) that tests the difference between the average simulated output and the average real-system output. In the same paper, Kleijnen proposed an *F-distribution-based naive* test using a two-dimensional hypothesis, i.e., the means of simulated output and real output are identical, and there is a positive correlation between simulated and real output. Mathematically, the null hypothesis is $H_0: \beta_0 = 0$ and $\beta_1 = 1$ where β_0 and β_1 are coefficients of linear regression model $E(w|v) = \beta_0 + \beta_1 v$, where v and w are real-system and simulation output respectively. Kleijnen et al. (1998) observed that the naive test frequently rejected valid models and proposed a modification referred to as a *novel* test, which tests for the equality of means and variances of the two sets of outputs (i.e., simulation model and the real system) corresponding to the given set of inputs. Mathematically, the null hypothesis is $H_0: \gamma_0 = 0$and $\gamma_1 = 0$ where γ_0 and γ_1 are coefficients of linear regression model $E(D|Q = q) = \gamma_0 + \gamma_1 q$, where D represents the difference between the real and simulation output, and Q is the sum of the two outputs.

Kleijnen et al. (2001) noted that the normality assumption / approximation may not hold for a small number of observations, making the *F*-test invalid. As a result, a Bootstrap methodology was proposed for validating the simulation model. Recent advances include Martens et al. (2006) who introduced a neural network based on a multi-layer perceptron and radial basis function to validate simulation models, while dos Santos et al. (2023) proposed the usage of Digital Twin to validate the model using updated data with machine learning and control charts.

5.2.2. Availability of only output real data

Typically, the input data for the real-world system is unavailable if the objective is to build a simulation model that matches a real system's historical data or when a real-world system is observable but there is no controllable input that can be tuned or experimented with (Kleijnen, 1999). Consequently, the basic notion of pairing the outputs of the simulation model and the real system is violated.

Naylor and Finger (1967) discussed several tests, such as analysis of variance, chi-square tests, factor analysis, Kolmogorov-Smirnov tests, and a few others that could measure the "goodness of fit" of the simulated model output with the real system (or historical) output. Hsu and Hunter (1974) suggested a Bayesian approach to compare real and simulated generated data. Ringuest (1986) developed a chi-square statistic to test whether the difference between simulated and real output falls within a permissible limit. Kleijnen (1999) applied established methods, including the two-sample Student t-test, Johnson's modified Student statistic, and Jackknifing, for the validation of simulation models in this context. Doudareva and Carter (2022) provided a list of popular data-driven validation techniques, including nonparametric tests, factor analysis, and time-series analysis.

5.2.3. Non-availability of real data

There is an abundance of business management and manufacturing applications where real systems cannot be observed or experimented with (Kleijnen, 1999). In such cases, one can only hope to build robust simulation models, and *sensitivity analysis* serves as a popular approach to validate the influence of the inputs and the effect of structural changes on the model (Kleijnen, 1995, 1999).

Design of experiments (DOE) and metamodels are two key tools in the sensitivity analysis methodology. DOE is employed to identify optimal combinations of input parameters for running the simulation model, and an appropriate metamodel helps understand and analyze the relationship between the model's inputs and outputs. Often, the choice of metamodel has an impact on the optimal design. Popular designs for determining the inputs of the simulation model include sequential design, full factorial design, fractional factorial design and orthogonal arrays (Kleijnen, 2017).

The validity of a metamodel must be confirmed before sensitivity analysis. For regression-based metamodels, Kleijnen (1983) proposed cross-validation using the t-statistic, and Kleijnen and Deflandre (2006) proposed a *Bootstrap* approach. Linear regression metamodels (LRMs) assume a linear relationship between simulation model inputs and outputs, which may consist of first-order or second-order polynomials, with optional interaction terms. Kleijnen (1995) recommended using full or fractional factorial designs to achieve high accuracy for LRMs. dos Santos et al. (2006) showed that LRMs often fail to provide an accurate global fit for smooth response functions with arbitrary shapes and proposed the idea of non-linear meta-models (NLMs). Van Beers and Kleijnen (2003) introduced the usage of kriging metamodels (KMs) as compared to low-order polynomial metamodels in

simulation experiments with large input domains. See Kleijnen (2017) for more discussions on LRMs and KMs. Recently, Kleijnen and Van Beers (2022) proposed cross-validation-based tests for KMs.

Over the past decade, the scope of verification and validation in simulation modelling has expanded beyond traditional stand-alone environments to encompass data-driven and real-time simulation systems. Emerging paradigms such as Digital Twins, cloud-based distributed simulations, and AI-assisted model generation using Large Language Models have introduced new validation challenges related to dynamic model updating, code reliability, and consistency across heterogeneous computational platforms.

6. Research gaps and future directions

A comprehensive review and analysis of the V&V of simulation model literature in our paper has revealed several important research gaps that can serve as the entry points for future research, directly addressing research question RQ3.

While validation techniques often involve substantial quantitative methodologies, though expert judgement does play a role. Verification methods remain largely qualitative, with suggestions and guidelines for careful debugging. This is the first important research gap, where more focus is required on developing novel methodologies to objectively verify the correctness of the computer codes that implement conceptual models.

Schruben-Turing test (Kleijnen, 1995) is a popular tool for testing the validity of a simulation model, where the outputs of the real system and the simulation model are presented together to an expert. If the expert can segregate the two sets of outputs, then the simulation model is considered unreliable. Despite the clear nature of the test, this validity testing is typically conducted manually by the expert. However, given the advancements in machine learning and artificial intelligence, one can develop an automated segregation technique for the two sets of outputs requiring methodology more sophisticated than a simple clustering technique, particularly for SD or hybrid models with non-scalar responses.

The current literature on the validation of SD models primarily focuses on behavior validation (Barlas, 1989). There is a scarcity of validation methodology that exploits the structure of the SD simulation model. Future research should focus on building customized validation techniques for SD, ABS, and other hybrid simulation models that can exploit the structural arrangement of the model components, geometry, and other important features of simulation model output.

Additionally, when the output, and not the input, of the real system is available, validation is akin to an inverse problem in computer experiments (Bhattacharjee et al., 2019; Ranjan et al., 2008, 2016). Consideration of relevant conceptual aspects from the larger domain of inverse problem methodologies for the validation of simulation models in this context can be both insightful and helpful.

In the absence of both input and output data of a real system, the reliability of a simulation model is typically assessed via sensitivity analysis. Moreover, Kleijnen has developed a series of methodologies on sensitivity analysis via linear regression and kriging metamodels when simulation model output is scalar in nature (e.g., in a DES model). However, sensitivity analysis for simulation models with non-scalar response (e.g., in SD models) has been inadequately discussed (preliminary work done by Hekimoğlu and Barlas (2016)). Furthermore, most of these methodologies for reliability testing assume continuous inputs and outputs. In this context, sensitivity analysis for simulation models with

categorical outputs and/or mixed inputs (i.e., continuous, discrete, stochastic) is yet to be investigated and provides a promising avenue for future research.

While a plethora of techniques have been developed and discussed for V&V of DES models, several for SD models, and a few for ABS models. However, techniques for V&V of other types of simulation models, such as hybrid models, distributed simulations, and simulation gaming are lacking. Each simulation model brings its unique challenges and presents an opportunity for innovation in developing efficient V&V methods. A generalized framework for determining appropriate V&V techniques across various types of simulation models will be greatly beneficial to practitioners in this area.

The utility of Design of Experiments (DOE) in the context of V&V cannot be overstated. DOE provides a structured and efficient approach to explore the behavior of simulation models across a range of input conditions. While DOE is well-established for investigating input-output relationships and optimizing simulation runs, its broader potential in verification and validation remains largely untapped - especially in the context of System Dynamics (SD) and hybrid simulation models. For instance, DOE could be systematically employed to test model robustness under diverse scenarios (verification) or to structure comparisons between simulated and real-world outputs (validation). However, current literature offers limited guidance on how to adapt DOE frameworks to handle the non-scalar, feedback-driven, and often qualitative nature of SD model outputs. This presents a valuable opportunity to develop tailored DOE methodologies that align with the unique characteristics of complex simulation models and to encourage greater utilization of recent advancements in DOE techniques.

Finally, topic modeling, particularly in the context of a literature review, reveals an additional research gap. The current practice of finding the optimal number of topics involves optimizing the average coherence score (or similar goodness-of-fit criterion) of topics along with expert judgment. It is understandable that forming the topic labels is qualitative in nature, but finding the optimal number of topics can be made less subjective with further research. Dissatisfaction with the current practice is also noted by Baimakhanbetov (2023), Doogan and Buntine (2021), and Krasnov and Sen (2019). While topic modeling has become a popular tool for analyzing large bodies of simulation literature, current approaches typically treat all documents equally, regardless of their scholarly impact. However, citation counts can serve as a proxy for influence and incorporating them into topic modeling either during corpus construction or as a weighting mechanism can highlight influential themes. Despite its potential, this integration remains underexplored in the context of simulation research, presenting a valuable opportunity for methodological innovation.

7. Concluding remarks

Our study was motivated by the absence of a comprehensive systematic review of the vast V&V literature. The combination of topic modeling, biblometric analysis, and a manual review of 300 research articles, provided an in-depth understanding of the literature on V&V of the simulation models commonly used in business management and manufacturing domains. The implementation of LDA for topic modeling resulted in the identification of nine prominent topics and research themes. The biblometric and scientometric analysis helped find significant authors, important keywords, and the edifying chronology of the literature. Moreover, the bibliometric maps and charts highlight the relationships among prominent authors, popular keywords, crucial documents, and significant sources.

This literature review has further enabled us to outline the evolution of popular V&V methodologies and identify the research gaps that can lay the foundation for future researchers in this area.

Through this research endeavor, we learned that the *Winter Simulation Conference, European Journal of Operational Research, Journal of The Operational Research Society, and International Journal of Production Research* are some of the prominent platforms that disseminate research articles on the V&V of simulation models. Furthermore, Robert G. Sargent, Jack P.C. Kleijnen, Osman Balci, and Yaman Barlas are pioneers and have made notable contributions in the field. Upcoming researchers can use this information on the overall domain to garner new insights into the V&V of simulation models.

For business management and manufacturing, DES models are the most popular, followed by SD and ABS models. Although several statistical tests have been developed to validate simulation models, the verification techniques continue to involve qualitative suggestions and guidelines. In this case, the validation tests can be categorized according to the availability of the inputs and outputs of the real system. When both inputs and outputs of the real system are available, standard tests like t-test, chi-square test, etc., can be used to validate the simulation models. However, if we have access to only real-system output and not the inputs, then factor analysis and Kolmogorov-Smirnov test are used. On the other hand, when both input and output are unavailable, sensitivity analysis with linear regression and Kriging metamodels serve to validate the reliability of simulation models.

Although the contribution of this paper stems from the inception of a comprehensive review on this topic, the categorization of articles into nine major themes and the evolution of quantitative methodologies provide valuable insights for researchers. Furthermore, the critical takeaway for future researchers lies in the identified research gaps. The key highlights include the inadequacy of the current practices in finding the optimal number of themes in topic modeling, the lack of quantitative methodologies for the verification of simulation models, and the need to develop a generalized framework for deciding which V&V technique to use for different types of simulation models.

Our research relied on corpus built for this specific study. While extremely useful in helping us come up with our findings, these findings are based on this specific corpus built only on titles, keywords, and abstracts for topic modeling analysis. Employing quantitative analysis of the full text of the articles can potentially yield more comprehensive results in future research.

Data availability statement

The list of 300 articles used for building the text corpus in this systematic literature review are available from the authors upon reasonable request.

References

Afshar, N., & Azadivar, F. (1992). A simulation study of traffic control procedures at highway work zones. *In Proceedings of The 1992 Winter Simulation Conference,* (pp. 1210-1216).

Akbulut, A., Abke, S., & Laroque, C. (2017). Automated model verification using an equivalence test on a reference model. *In Proceedings of The 2017 Winter Simulation Conference*, (pp. 4187-4196).

Amoako-Gyampah, K., & Meredith, J. R. (1989). The operations management research agenda: An update. *Journal of Operations Management*, 8(3), 250–262.

Anderson Jr, H. A., & Sargent, R. G. (1972). A statistical evaluation of the scheduler of an experimental interactive computing system. *Statistical Computer Performance Evaluation,* (pp. 73-98).

Aria, M. & Cuccurullo, C. (2017) bibliometrix: An R-tool for comprehensive science mapping analysis, *Journal of Informetrics*, 11(4), 959-975.

Arifin, S. N., Davis, G. J., Kurtz, S., Gentile, J. E., Zhou, Y., & Madey, G. R. (2010). Divide and conquer: A four-fold docking experience of agent-based models. *In Proceedings of The 2010 Winter Simulation Conference,* (pp. 575-586).

Baimakhanbetov, M. (2023). Determination of the Optimal Number of Topics in the LDA Model When Working with Large Arrays of Text Data. I*n Proceedings of the 2023 IEEE International Conference on Smart Information Systems and Technologies (SIST),* (pp. 332-336).

Balci, O. (1995). Principles and techniques of simulation validation, verification, and testing. *In Proceedings of The 1995 Winter Simulation Conference,* (pp. 147-154).

Balci, O. (1998). Verification, Validation, And Accreditation. *In Proceedings of The 1998 Winter Simulation Conference,* (pp. 41-48).

Balci, O. (2012). A life cycle for modeling and simulation. *Simulation*, 88(7), 870-883.

Banks, J. (1999). Introduction to simulation. *In Proceedings of The 1999 Winter Simulation Conference*, (pp 7–13).

Barad, M. (1994). Decomposing timed petri net models of open queueing networks. *Journal of the Operational Research Society*, 45(12), 1385-1397.

Barad, M. (1998). Timed Petri nets as a verification tool. *In Proceedings of The 1998 Winter Simulation Conference,* (pp. 547-554).

Barlas, Y. (1989). Multiple tests for validation of system dynamics type of simulation models. *European Journal of Operational Research*, 42(1), 59–87.

Barlas, Y., & Carpenter, S. (1990). Philosophical roots of model validation: two paradigms. *System Dynamics Review*, 6(2), 148-166.

Barlas, Y. (1996). Formal aspects of model validity and validation in system dynamics. *System Dynamics Review*, 12(3), 183-210.

Bhattacharjee, N.V., Ranjan, P., Mandal, A. and Tollner, E.W. (2019), A History Matching Approach for Calibrating Hydrological Models. *Environmental and Ecological Statistics*, 26(1), 87-105.

Bitencourt, J., Wooley, A., & Harris, G. (2024). Verification and validation of digital twins: a systematic literature review for manufacturing applications. *International Journal of Production Research*, 1-29.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3 (2003), 993-1022.

Bountourelis, T., Luangkesorn, L., Schaefer, A., Maillart, L., Nabors, S. G., & Clermont, G. (2011). Development and validation of a large-scale ICU simulation model with blocking. *In Proceedings of The 2011 Winter Simulation Conference,* (pp. 1143-1153).

Brailsford, S. C., Eldabi, T., Kunc, M., Mustafee, N., & Osorio, A. F. (2019). Hybrid simulation modelling in operational research: A state-of-the-art review. *European Journal of Operational Research*, 278(3), 721–737.

Brügmann, J., Schreckenberg, M., & Luther, W. (2014). A verifiable simulation model for real-world microscopic traffic simulations. *Simulation Modelling Practice and Theory*, 48, 58-92.

Campagnolo, J. M., Duarte, D., & Dal Bianco, G. (2022). Topic coherence metrics: How sensitive are they? *Journal of Information and Data Management, 13(4)*.

Can, B., & Heavey, C. (2012). A comparison of genetic programming and artificial neural networks in metamodeling of discrete-event simulation models. *Computers and Operations Research*, 39(2), 424–436.

Chou, C. Y., Chen, C. H., & Li, M. H. C. (2001). Application of computer simulation to the design of a traffic signal timer. *Computers & Industrial Engineering*, 39(1-2), 81-94.

Churchill, R., & Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, 54(10s), 1-35.

Chwif, L., Silva, P. S. M., & Shimada, L. M. (2006). A Prescriptive Technique for V&V of Simulation Models When No Real-Life Data are Avaiable. *In Proceedings of The 2006 Winter Simulation Conference* (pp. 911-918).

Cochran, J. K. (1987). Techniques for ascertaining the validity of large-scale production simulation models. *International Journal of Production Research*, 25(2), 233-244.

Dohale, V., Gunasekaran, A., Akarte, M. M., & Verma, P. (2022). 52 Years of manufacturing strategy: an evolutionary review of literature (1969–2021). *International Journal of Production Research*, 60(2), 569-594.

Doogan, C., & Buntine, W. (2021). Topic model or topic twaddle? Re-evaluating demantic interpretability measures. *In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (pp. 3824-3848).*

dos Santos, C. H., Montevechi, J. A. B., de Queiroz, J. A., de Carvalho Miranda, R., & Leal, F. (2022). Decision support in productive processes through DES and ABS in the Digital Twin era: a systematic literature review. *International Journal of Production Research*, 60(8), 2662–2681.

dos Santos, C. H., Campos, A. T., Montevechi, J. A. B., de Carvalho Miranda, R., & Costa, A. F. B. (2023). Digital Twin simulation models: a validation method based on machine learning and control charts. *International Journal of Production Research*, 1-17.

dos Santos, M. I. R., & Nova, A. M. P. (2006). Statistical fitting and validation of non-linear simulation metamodels: A case study. *European Journal of Operational Research*, 171(1), 53-63.

Doudareva, E., & Carter, M. (2022). Discrete event simulation for emergency department modelling: A systematic review of validation methods. *Operations Research for Health Care*, 33, 100340.

Durst, P. J., Anderson, D. T., & Bethel, C. L. (2017). A historical review of the development of verification and validation theories for simulation models. *International Journal of Modeling, Simulation, and Scientific Computing*, 8(02), 1730001.

Edali, M., & Yücel, G. (2019). Exploring the behavior space of agent-based simulation models using random forest metamodels and sequential sampling. *Simulation Modelling Practice and Theory*, 92,62–81.

Edali, M. (2022). Pattern-oriented analysis of system dynamics models via random forests. *System Dynamics Review*, 38(2), 135-166.

Fishman, G. S., & Kiviat, P. J. (1968). The statistics of discrete-event simulation. *Simulation*, 10(4), 185-195.

Fitzsimmons, J. A. (1971). An emergency medical system simulation model. *In Proceedings of The 1971 Winter Simulation Conference,* (pp. 18-25).

Friederich, J., Francis, D. P., Lazarova-Molnar, S., & Mohamed, N. (2022). A framework for data-driven digital twins for smart manufacturing. *Computers in Industry*, 136, 103586.

Fonseca i Casas, P. (2023). A continuous process for validation, verification, and accreditation of simulation models. *Mathematics*, 11(4), 845.

Forrester, J. W. (1973). Confidence in models of social behavior with emphasis on system dynamics models. *System Dynamics Group Working Paper, Sloan School of Management, MIT, Cambridge, MA.*

Gass, S. I. (1977). A procedure for the evaluation of complex models. *In Proceedings of the First International Conference in Mathematical Modeling,* (pp. 247-258).

Gerrits, B., Mes, M., & Schuur, P. (2017). An agent-based simulation model for autonomous trailer docking. *In Proceedings of The 2017 Winter Simulation Conference*, (pp. 1324-1335).

Gore, R., Diallo, S. Y., Padilla, J. J., & Tolk, A. (2017). Applying statistical debugging for enhanced trace validation of agent-based models. *Simulation Modelling Practice and Theory, 77*, 138–154.

Gürcan, ö., Dikenelli, O., & Bernon, C. (2013). A generic testing framework for agent-based simulation models. *Journal of Simulation*, 7(3), 183-201.

Han, Y., Chong, W. K., & Li, D. (2020). A systematic literature review of the capabilities and performance metrics of supply chain resilience. *International Journal of Production Research*, 58(15), 4541-4566.

Harper, A., Mustafee, N., & Yearworth, M. (2021). Facets of trust in simulation studies. *European Journal of Operational Research*, 289(1), 197–213.

Hekimoğlu, M., & Barlas, Y. (2016). Sensitivity analysis for models with multiple behavior modes: a method based on behavior pattern measures. *System Dynamics Review*, 32(3-4), 332-362.

Henderson, J. A., & Bryce, R. M. (2019). Verification methodology for discrete event simulation models of personnel in the Canadian armed forces. *In Proceedings of The 2019 Winter Simulation Conference*, (pp. 2479-2490).

Hsu, D. A., & Hunter, J. S. (1974). Validation of computer simulation models using parametric time series analysis. *In Proceedings of The 1974 Winter Simulation Conference,* (pp. 727-728).

Irfanoglu, E., Akgun, I., & Gunal, M. M. (2013). Metamodeling by using multiple regression integrated K-means clustering algorithm. *In Proceedings of the Emerging M&S Applications in Industry & Academia/Modeling and Humanities Symposium,* (pp. 1-8).

Jacobson, S. H., & Yücesan, E. (1999). On the complexity of verifying structural properties of discrete event simulation models. *Operations Research*, 47(3), 476-481.

Jahangirian, M., Eldabi, T., Naseer, A., Stergioulas, L. K., & Young, T. (2010). Simulation in manufacturing and business: A review. *European Journal of Operational Research*, 203(1), 1–13.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11), 15169-15211.

Kabak, K. E., Hinckeldeyn, J., & Dekkers, R. (2024). A systematic literature review into simulation for building operations management theory: reaching beyond positivism? *Journal of Simulation*, 1-29.

Kleijnen, J. P. (1983). Cross-validation using the t statistic. *European Journal of Operational Research*, 13(2), 133-141.

Kleijnen, J. P. C. (1995). Statistical validation of simulation models. *European Journal of Operational Research*, 87(1), 21-34.

Kleijnen, J. P. C. (1995). Verification and validation of simulation models. *European Journal of Operational Research*, 82(1), 145-162.

Kleijnen, J. P. C., Bettonvil, B., & Van Groenendaal, W. (1998). Validation of trace-driven simulation models: a novel regression test. *Management Science*, 44(6), 812-819.

Kleijnen, J. P. C. (1999). Validation of models: statistical techniques and data availability. *In Proceedings of The 1999 Winter Simulation Conference,* (pp. 647-654).

Kleijnen, J. P. C., & Sargent, R. G. (2000). A methodology for fitting and validating metamodels in simulation. *European Journal of Operational Research*, 120(1), 14–29.

Kleijnen, J. P., Cheng, R. C., & Bettonvil, B. (2001). Validation of trace-driven simulation models: Bootstrap tests. *Management Science*, *47*(11), 1533-1538.

Kleijnen, J. P. C., & Deflandre, D. (2006). Validation of regression metamodels in simulation: Bootstrap approach. *European Journal of Operational Research*, 170(1), 120–131.

Kleijnen, J. P. C. (2009). Kriging metamodeling in simulation: A review. *European Journal of Operational Research*, 192(3), 707-716.

Kleijnen, J. P. (2017). Regression and Kriging metamodels with their experimental designs in simulation: A review. E*uropean Journal of Operational Research*, 256(1), 1-16.

Kleijnen, J. P. C., & van Beers, W. C. M. (2022). Statistical tests for cross-validation of kriging models. *Journal on Computing*, 34(1), 607–621.

Krahl, D. (2005). Debugging simulation models. *In Proceedings of The 2005 Winter Simulation Conference,* (pp. 7-pp).

Krasnov, F., & Sen, A. (2019). The number of topics optimization: Clustering approach. *Machine Learning and Knowledge Extraction*, 1(1), 25.

Krishnamurthi, M., & Thallikar, S. (1998). A deadlock detection interfaces to a commercial simulation language. *Computers & Industrial Engineering*, 34(4), 743-757.

Landry, M., Malouin, J. L., & Oral, M. (1983). Model validation in operations research. *European Journal of Operational Research*, 14(3), 207–220.

Landry, M., & Oral, M. (1993). In search of a valid view of model validation for operations research. *European Journal of Operational Research*, 66(2), 161-167.

Law, A. M. (2022). How to build valid and credible simulation models. *In Proceedings of The 2022 Winter Simulation Conference*, (pp 2013–2015).

Law, A.M. (2024) *Simulation modelling and analysis*. 6th Edition, McGraw-Hill, New York.

Liao, Y., Deschamps, F., Loures, E. D. F. R., & Ramos, L. F. P. (2017). Past, present and future of Industry 4.0-a systematic literature review and research agenda proposal. *International Journal of Production Research*, 55(12), 3609-3629.

Lugaresi, G., Gangemi, S., Gazzoni, G., & Matta, A. (2022). Online validation of simulation-based digital twins exploiting time series analysis. In Proceedings of The 2022 *Winter Simulation Conference*, (pp. 2912-2923).

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., & Adam, S. (2021).

Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *In Computational Methods for Communication Science* (pp. 13-38). Routledge.

Martens, J., Pauwels, K., & Put, F. (2006). A neural network approach to the validation of simulation models. I*n Proceedings of The 2006 Winter Simulation Conference,* (pp. 905-910).

Montevechi, J. A. B., Gabriel, G. T., Campos, A. T., dos Santos, C. H., Leal, F., & Machado, M. E.(2022). Using Generative Adversarial Networks to Validate Discrete Event Simulation Models. *In Proceedings of The 2022 Winter Simulation Conference,* (pp. 2772-2783).

Mourtzis, D. (2020). Simulation in the design and operation of manufacturing systems: state of the art and new trends. *International Journal of Production Research*, 58(7), 1927–1949.

Mustak, M., Salminen, J., Plé, L., & Wirtz, J. (2021). Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *Journal of Business Research*, 124, 389-404.

Naylor, T. H., & Finger, J. M. (1967). Verification of computer simulation models. *Management Science*, 14(2), 92-101.

North M.J. & Macal C. M. (2007). *Managing business complexity discovering strategic solutions with agent-based modeling and simulations*. Oxford University Press.

Naz, F., Agrawal, R., Kumar, A., Gunasekaran, A., Majumdar, A., & Luthra, S. (2022). Reviewing the applications of artificial intelligence in sustainable supply chains: Exploring research propositions for future directions. *Business Strategy and the Environment*, 31(5), 2400-2423.

Olsen, M., & Raunak, M. (2016). Metamorphic validation for agent-based simulation models. *In Proceedings of the Summer Computer Simulation Conference*, 48(9), 234–241.

Onggo, B. S., & Karatas, M. (2016). Test-driven simulation modelling: A case study using agent-based maritime search-operation simulation. *European Journal of Operational Research*, 254(2), 517–531.

Onggo, B. S., & Foramitti, J. (2021). Agent-based modeling and simulation for business and management: a review and tutorial. *In Proceedings of The 2021 Winter Simulation Conference,* (pp. 1-15).

Oral, M., & Kettani, O. (1993). The facets of the modeling and validation process in operations research. *European Journal of Operational Research*, 66(2), 216-234.

Pannirselvam, G. P., Ferguson, L. A., Ash, R. C., & Siferd, S. P. (1999). Operations management research: An update for the 1990s. *Journal of Operations Management*, 18(1), 95–112.

Psarommatis, F., & May, G. (2023). A literature review and design methodology for digital twins in the era of zero-defect manufacturing. *International Journal of Production Research*, 61(16), 5723-5743.

Pousi, J., Poropudas, J., & Virtanen, K. (2013). Simulation metamodelling with Bayesian networks. *Journal of Simulation*, 7(4), 297–311.

Rabe, M., Spieckermann, S., & Wenzel, S. (2008). A new procedure model for verification and validation in production and logistics simulation. *In Proceedings of The 2008 Winter Simulation Conference,* (pp. 1717-1726).

Radiya, A., & Sargent, R. G. (1987). A new formalism for discrete event simulation. *In Proceedings of The 1987 Winter Simulation Conference,* (pp. 554-558).

Ranjan, P., Bingham, D. and Michailidis, G. (2008), Sequential Experiment Design for Contour Estimation from Complex Computer Codes, *Technometrics*, 50, 527-541.

Ranjan, P., Thomas, M., Teismann, H. and Mukhoti, S., (2016), Inverse problem for time-series valued computer model via scalarization, *Open Journal of Statistics*, 6, 528-544.

Rao, L., Owen, L., & Goldsman, D. (1998). Development and application of a validation framework for traffic simulation models. *In Proceedings of The 1998 Winter Simulation Conference,* (pp. 1079-1086).

Raunak, M., & Olsen, M. (2014). Quantifying validation of discrete event simulation models. *In Proceedings of The 2014 Winter Simulation Conference,* (pp. 628-639).

Raunak, M., & Olsen, M. (2015). Simulation validation using metamorphic testing (WIP). *Simulation Series*, 47(10), 520–525.

Radim Řeȟuřek and Petr Sojka, (2010). Software Framework for Topic modelling with large corpora, *In Proceedings of the LREC 2010 Workshop on New Challenges,* (pp. 45-50).

Ringuest, J. L. (1986). A chi-square statistic for validating simulation-generated responses. *Computers & Operations Research*, 13(4), 379-385.

Robinson, S. (2002). General concepts of quality for discrete-event simulation. *European Journal of Operational Research*, 138(1), 103–117.

Robinson, S., & Brooks, R. J. (2010). Independent verification and validation of an industrial simulation model. *Simulation*, 86(7), 405-416.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, (pp. 399-408).

Roungas, B. (2016). Towards the validation of a simulation environment. *In Proceedings of The 2016 Winter Simulation Conference,* (pp. 3676-3677).

Rousseau, G. G., & Bauer Jr, K. W. (1996). Sensitivity analysis of a large-scale transportation simulation using design of experiments and factor analysis. *In Proceedings of The 1996 Winter Simulation Conference,* (pp. 1426-1432).

Ryan, K. T. (1979). Validating a bus operations simulation model. *In Proceedings of The 1979 Winter Simulation Conference,* (pp. 483-495).

Sargent, R. G. (1981). An assessment procedure and a set of criteria for use in the evaluation of computerized models and computer-based modeling tools. *Final Technical Report RADC-TR-80-409*.

Sargent, R. G. (2010). Verification and validation of simulation models. *In Proceedings of The 2010 Winter Simulation Conference,* (pp. 166-183).

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

Sargent, R. G. (2015). An interval statistical procedure for use in validation of simulation models. *Journal of Simulation*, 9, 232-237.

Sargent, R. G., & Balci, O. (2017). History of verification and validation of simulation models. *In Proceedings of The 2017 Winter Simulation Conference,* (pp. 292-307).

Sargent, R. G. (2020). Verification and validation of simulation models: an advanced tutorial. *In Proceedings of The 2020 Winter Simulation Conference,* (pp. 16-29).

Sarjoughian, H. S., Muqsith, M., Huang, D., & Yau, S. S. (2012, March). Validation of service-oriented computing DEVS simulation models. *In Proceedings of the 2012 Symposium on Theory of Modeling and Simulation-DEVS Integrative M&S Symposium,* (pp. 1-8).

Sarnow, T., & Elbert, R. (2022). V&V application in generic simulation models in logistics. *Journal of Simulation*, 1-11.

Saysel, A. K., & Barlas, Y. (2006). Model simplification and validation with indirect structure validity tests. *System Dynamics Review*, 22(3), 241-262.

Schoenenberger, L., Schmid, A., Tanase, R., Beck, M., & Schwaninger, M. (2021). Structural Analysis of System Dynamics Models. *Simulation Modelling Practice and Theory*, 110(2020), 102333.

Schruben, L. W. (1980). Establishing the credibility of simulations. *Simulation*, 34(3), 101-105.

Sinisi, S., Alimguzhin, V., Mancini, T., & Tronci, E. (2021). Reconciling interoperability with efficient verification and validation within open-source simulation environments. *Simulation Modelling Practice and Theory*, 109, 102277.

Strang, K. D. (2012). Importance of verifying queue model assumptions before planning with simulation software. *European Journal of Operational Research*, 218(2), 493-504.

Sterman John D. (2000). *Business Dynamics Systems Thinking and Modeling for a Complex World*. Mc-Graw Hill Higher Education.

Swisher, J. R., Jacobson, S. H., Jun, J. B., & Balci, O. (2001). Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers & Operations Research*, 28(2), 105-125.

Tako, A. A., & Robinson, S. (2010). Model development in discrete-event simulation and system dynamics: An empirical study of expert modellers. *European Journal of Operational Research*, 207(2), 784–794.

Taylor, A. J. (1983). The verification of dynamic simulation models. *Journal of the Operational Research Society*, 34(3), 233-242.

Troost, C., Huber, R., Bell, A. R., van Delden, H., Filatova, T., Le, Q. B., & Berger, T. (2023). How to keep it adequate: A protocol for ensuring validity in agent-based simulation. *Environmental Modelling & Software*, 159, 105559.

Van Beers, W. C., & Kleijnen, J. P. (2003). Kriging for interpolation in random simulation. *Journal of the Operational Research Society*, 54(3), 255-262.

Van Eck, N., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538.

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, 94, 101582.

Volk-Makarewicz, W., & Cleophas, C. (2017). A meta-algorithm for validating agent-based simulation models to support decision making. *In Proceedings of The 2017 Winter Simulation Conference,* (pp. 1348-1359).

Woodward, E. E., & Mackulak, G. T. (1997). Detecting logic errors in discrete-event simulation: reverse engineering through event graphs. *Simulation Practice and Theory*, 5(4), 357-376.

Yacoub, A., Hamri, M. E. A., & Frydman, C. (2020). DEv-PROMELA: modeling, verification, and validation of a video game by combining model-checking and simulation. *Simulation*, 96(11), 881-910.

Yang, S. L., Wang, J. Y., Xin, L. M., & Xu, Z. G. (2022). Verification of intelligent scheduling based on deep reinforcement learning for distributed workshops via discrete event simulation. *Advances in Production Engineering And Management*, 17(4), 401–412.

Yeung, W. L. (2011). Formal verification of negotiation protocols for multi-agent manufacturing systems. *International Journal of Production Research*, 49(12), 3669-3690.