Queueing Models and
Service Management
Vol. 9, No. 3, page 25-56, 2026

QMSM

Analytical, Simulation, and Learning-Based Approaches for Decision Support in Stochastic Service Systems: A Comparative Study

Zhe George Zhang[1,2,*] and Jiexun Li[1]

[1]Department of Decision Sciences
Western Washington University, Bellingham, WA98255, USA

[2]Beedie School of Business,
Simon Fraser University, Burnaby, BC, V1M 2J2, Canada

(*Received June 2026; accepted August 2026*)

Abstract. This paper presents a comparative analysis of three major methodologies for modeling stochastic service systems: analytical modeling (AM), simulation modeling (SM), and learning-based modeling (LM). While these approaches have been widely studied in isolation, relatively little research has systematically examined their respective roles and complementarities in supporting decision making for complex service systems. Using representative applications from international border-crossing operations and healthcare service systems, this study evaluates the strengths and limitations of the three methodologies in terms of modeling assumptions, interpretability, computational requirements, and predictive capability. The results show that analytical models provide valuable structural insights and prescriptive guidance under simplifying assumptions. Simulation models offer flexible and realistic representations of system dynamics for policy evaluation, and learning-based models deliver strong predictive performance in data-rich and highly dynamic environments. In addition to comparing these approaches, the paper proposes an integrated decision-support framework in which learning-based models generate short-term forecasts, analytical models design operational policies, and simulation models evaluate policy performance under realistic conditions. The findings highlight the complementary nature of these methodologies and provide practical guidance for selecting and integrating modeling approaches to support decision making in stochastic service systems.

Keywords: Analytical modeling, border-crossing systems, learning-based modeling, simulation modeling, stochastic service systems.

1. Introduction

Stochastic modeling is an essential tool for understanding and predicting the operational behavior of service systems characterized by uncertainty. Researchers and practitioners study such systems by a range of methodologies, including analytical models, simulation

* Corresponding author
Email: george.zhang@wwu.edu

models, and learning-based models. This research is aimed at identifying the contexts in which each approach that is most suitable for studying service systems.

In this paper, we adopt a broad definition of the terms “model” and “modeling” as any systematic approach that abstracts the relationships between input variables (e.g., decision variables and policy parameters) and output variables (e.g.. performance measures) in a stochastic service system. This definition differs from the narrower interpretation often used in operations research or econometrics, where a model typically refers to a purely mathematical formulation expressed through equations or analytical expressions. In our research, we distinguish three major paradigms: analytical modeling (AM), simulation modeling (SM), and learning-based modeling (LM). Analytical models rely on explicit mathematical formulations, simulation models represent system dynamics through computational experiments, and learning-based models infer relationships directly from data.

Analytical models have traditionally been the foundation of stochastic system analysis. By imposing tractable assumptions, they yield closed-form or approximate expressions that provide structural insights and support policy design. However, these models often require assumptions—such as stationarity or Markovian dynamics—that may not hold in practice. When system complexity or realism violates these assumptions, simulation models offer a flexible alternative by explicitly representing stochastic processes and operational rules. Through repeated sampling, simulation can evaluate system performance under a wide range of scenarios. More recently, the increasing availability of operational data has enabled learning-based approaches, which can capture complex and nonlinear relationships without requiring explicit structural assumptions. These models often achieve strong predictive performance but may lack interpretability and direct prescriptive capability.

Despite extensive research on each of these approaches, prior studies have largely examined them in isolation. As a result, there is limited understanding of how these methodologies compare and how they can be jointly used to support decision making in complex service systems. This gap is particularly important where systems are both complex and data-rich. To address this gap, this study provides a comparative analysis of analytical, simulation, and learning-based models in the context of stochastic service systems. Using real-world systems such as international border-crossing stations and healthcare service systems, we evaluate the three methodologies from several key dimensions, including modeling assumptions, interpretability, computational requirements, and predictive capability. Our key contribution is an integrated decision-support framework that clarifies the circumstances under which each methodology is most suitable: analytical models guide policy design, simulation models evaluate system performance under realistic conditions, and learning-based models generate short-term forecasts.

The remainder of this paper is organized as follows. Section 2 reviews relevant literature on stochastic service systems, waiting-time prediction, and customer decision behavior. Section 3 presents a comparative analysis of the three modeling approaches through representative applications. Section 4 discusses implications and introduces the integrated decision-support framework. Finally, Section 5 concludes and outlines future directions.

2. Literature Review

This section reviews the literature on the three modeling methodologies for analyzing stochastic service systems. Rather than providing an exhaustive survey, we focus on key representative studies that illustrate how each methodology has been applied and what insights it provides.

2.1. Analytical models

Analytical queueing models have long been developed as a mathematical framework for analyzing the dynamics of waiting lines. They have been widely applied to service systems characterized by stochastic arrivals and service processes. These models typically characterize system performance by factors such as arrival rates, service rates, and the number of servers. These models allow researchers to obtain closed-form expressions for key performance indicators such as expected queue length, expected waiting time, and server utilization. Such analytical results provide useful guidance for system design and operational policy decisions. In the context of border-crossing systems, Zhang [26] developed an analytical model to examine waiting lines under a congestion-based staffing policy. The study derived closed-form expressions for expected queue length and waiting times under specific operational assumptions and demonstrated how staffing policies influence system performance. This work was extended in [27] by incorporating more realistic inspection processes through Coxian-distributed service times and addressing security-check considerations. Although analytical models provide important theoretical insights, they often rely on simplifying assumptions to maintain mathematical tractability. Common assumptions include steady-state conditions and Markovian arrival and service processes. Under these assumptions, the system can be analyzed using stationary Markov chains, while heavy-traffic approximations allow the application of fluid or diffusion models to derive approximate system behavior. While these assumptions facilitate analytical derivations, they may not always hold in real-world border-crossing operations, where arrival rates, service processes, and operational policies can vary significantly over time. Consequently, validating and calibrating analytical models using real operational data is essential to ensure their reliability and practical applicability. In many cases, analytical models serve as useful approximations that provide insights into system behavior, even when the underlying assumptions are only partially satisfied.

Beyond border-crossing applications, analytical queueing models have been widely applied to other service systems, particularly call centers and healthcare operations. Foundational studies in call-center analysis demonstrate how queueing models can be used to derive staffing rules and performance approximations for large-scale service systems, including environments with customer abandonment and time-varying demand [8, 9]. In particular, the Erlang-A model and many-server approximations provide important insights into the trade-offs between service quality and system efficiency. In addition, approximation methods such as the stationary independent period-by-period (SIPP) approach have been developed to handle nonstationary arrival processes [11]. In healthcare systems, analytical queueing models have also been used to study patient flow, appointment scheduling, and congestion management, illustrating the continued relevance of analytical modeling when

system structure is sufficiently well understood [10].

Despite their strengths, analytical models typically rely on simplifying assumptions—such as steady-state conditions or Markovian dynamics—to maintain tractability. These assumptions may limit their applicability in complex and time-varying real-world systems. Therefore, analytical models are often best viewed as providing structural insights and approximations rather than fully realistic representations of system behavior.

2.2. Simulation models

Simulation models have been widely used to analyze stochastic service systems, particularly when system dynamics are too complex to be captured by tractable analytical models. In contrast to analytical approaches, simulation models provide a flexible computational framework that explicitly represents system operations, stochastic variability, and complex interactions among system components [17]. Through generating multiple realizations of system behavior over time, simulation models can estimate key performance measures such as waiting times, queue lengths, and resource utilization under a variety of operational scenarios [22].

In the context of border-crossing systems, simulation models have been applied to evaluate inspection processes, staffing policies, and system performance under different operational configurations. For example, Khoshons et al. [16] developed a discrete-event simulation framework to evaluate international border-crossing clearance systems, focusing on commercial vehicle inspection and pre-screening programs. Their model was used to analyze alternative operational scenarios and assess system performance using multiple measures of effectiveness. Similarly, micro-simulation approaches have been used to evaluate priority crossing programs and border inspection strategies, capturing detailed interactions among vehicles, inspection processes, and system control policies [3, 4]. These studies demonstrate that the ability of simulation in modeling complex interactions among vehicle arrivals, service processes, and operational policies.

Beyond border-crossing applications, simulation modeling has been extensively applied in transportation systems to analyze traffic flow, congestion dynamics, and system performance. Simulation models can represent system behavior at different levels of detail, ranging from microscopic models that capture individual vehicle behavior to macroscopic models that describe aggregate traffic flows. These models enable analysts to evaluate infrastructure design and traffic control policies under varying demand conditions and operational scenarios.

Simulation modeling is also widely used in healthcare service systems. Numerous studies have applied discrete-event simulation to analyze patient flow, appointment scheduling, and resource allocation in healthcare facilities. For example, Jun et al. [14] provided a comprehensive survey of simulation applications in healthcare clinics, demonstrating how simulation can be used to evaluate system performance under realistic assumptions such as stochastic service times and variable patient arrivals. Similarly, Cayirli and Veral [6] reviewed outpatient scheduling models and highlighted the role of simulation in analyzing appointment systems with no-shows, patient heterogeneity, and time-dependent demand. These studies illustrate that simulation modeling serves as an important decision-support tool for analyzing complex service systems where structure is known but

mathematical tractability is limited.

Despite its flexibility and modeling power, simulation modeling also presents several challenges. Developing a reliable simulation model requires detailed knowledge of system operations and sufficient data for parameter calibration. In addition, simulation experiments can be computationally intensive, particularly when evaluating large-scale systems or optimizing multiple decision variables. Nevertheless, simulation remains one of the most practical approaches for analyzing stochastic service systems, especially when analytical models are infeasible and data-driven models alone cannot fully capture system dynamics.

2.3. Learning-based models

Recent advances in data collection and information systems have made learning-based approaches increasingly important for analyzing and predicting performance in stochastic service systems. Large volumes of operational data, e.g., arrivals, staffing levels, traffic conditions, waiting times, weather, and other contextual effects, allow studying complex relationships among system variables without entirely relying on traceable analytical models. In the literature, an important distinction is between purely data-driven predictors and hybrid methods that combine statistical learning with queueing-based structure [1].

For border crossings, learning-based methods have shown their prediction capability. Khan [15] proposed a framework for predicting and displaying delays at road border crossings. Lin et al. [18] developed transient multi-server queueing models for border-crossing delay prediction, showing how traffic flow information and queue dynamics can be used for short-term forecasting. Using data of U.S.-Canada border crossings, Yu et al. [25] adopted a data-driven approach to manpower planning and showed that both arrival rates and service rates vary systematically with time of day and operating conditions. For commercial trucks, Moniruzzaman et al., [19] used artificial neural networks (ANNs) to predict short-term crossing times and traffic volumes. More recently, Sharma et al. [20] applied gradient boosting and random forest regression to predict commercial vehicle wait times at a U.S.–Mexico border crossing. These studies collectively show that learning-based approaches can significantly improve prediction performance in data-rich service environments.

Related work in other service settings points in the same direction. In emergency departments, Sun et al. [21] used quantile regression for real-time waiting-time prediction. Ang et al. [1] proposed Q-Lasso, a hybrid method that combines statistical learning with fluid-model estimators and outperformed rolling averages and standalone fluid estimators across several hospitals. Arora et al. [2] extended this line by using quantile regression forests to generate probabilistic, rather than only point, forecasts of emergency-department waiting times. More generally, Chocron et al. [7] compared queueing-theoretic and machine-learning approaches for multiclass service systems and showed that machine learning is not uniformly superior. In certain settings, queueing-based models can perform comparably or even better than ML models while requiring less computation.

Overall, the literature suggests that learning-based models are especially valuable when service systems are data rich, highly time varying, and characterized by nonlinear interactions that are difficult to specify analytically. However, these approaches also have important limitations. They typically require substantial amounts of high-quality data, may

degrade when operating conditions shift, and often provide less interpretability than structural queueing models. Therefore, learning-based methods may provide limited prescriptive insight for designing operational policies in stochastic service systems. The most promising direction may be not to replace analytical or simulation models, but to integrate the three approaches in a unified framework that combines domain structure with prediction capabilities.

2.4. Our contributions

Although a substantial body of literature exists on analytical queueing models, simulation models, and learning-based models for stochastic service systems, these research streams have largely developed independently. Analytical models focused on deriving structural performance measures under simplifying assumptions, simulation models emphasize realistic system representation and policy evaluation, and learning-based approaches prioritize predictive accuracy using empirical data. However, relatively little research has systematically examined how these three modeling paradigms compare when applied to complex service systems. This study makes three contributions.

First, we provide a unified comparative framework for understanding analytical modeling (AM), simulation modeling (SM), and learning-based modeling (LM) in stochastic service systems and support operational decision making. By comparing these methodologies in terms of interpretability, modeling realism, computational requirements, and predictive capability, we clarify the contexts in which each approach is most appropriate.

Second, the paper illustrates the comparative strengths and limitations of these approaches through representative applications, including international border-crossing operations and healthcare service systems. These examples demonstrate how the suitability of each approach depends on the system characteristics, such as structural knowledge, system complexity, and data availability.

Third, the study proposes a practical decision-support framework that can guide researchers and practitioners in selecting appropriate modeling methodologies under different circumstances. In this framework, analytical models are shown to be particularly valuable for deriving structural insights and optimizing policy parameters when system assumptions are reasonably satisfied; simulation models are useful for evaluating operational policies and exploring system behavior under complex and uncertain conditions; and learning-based models are most effective for forecasting system performance when large datasets are available and the relationships among variables are difficult to specify analytically.

Finally, the paper emphasizes the complementary roles of these methodologies rather than treating them as competing alternatives. By integrating analytical insights, simulation-based experimentation, and data-driven prediction techniques, decision makers can develop more comprehensive insights into managing complex stochastic service systems.

3. Analysis of Three Modeling Approaches

In this section, we compare three major modeling methodologies for stochastic service systems: analytical modeling (AM), simulation modeling (SM), and learning-based modeling (LM). Each methodology provides a distinct perspective for understanding system

behavior and supporting operational decision making. Analytical models emphasize mathematical tractability and structural insight, simulation models focus on realistic representation of system dynamics, and learning-based models leverage empirical data to generate predictive models.

To illustrate the characteristics of these approaches, we examine two representative service systems characterized by congestion and operational uncertainty. The first example is an international border-crossing system between Washington State in the United States and the province of British Columbia in Canada. This system represents a typical stochastic service system in which travelers arrive randomly, service times vary across inspection booths, and congestion levels fluctuate throughout the day. The second example is a specialist medical clinic, where patients are scheduled for consultation with a physician and service durations are highly variable. Both systems exhibit stochastic arrivals, uncertain service times, and operational decision variables that influence system performance. These examples allow us to examine how the three modeling approaches perform under practical conditions and to highlight their respective strengths and limitations. We evaluate the approaches along several key dimensions, including modeling transparency, realism of system representation, data requirements, computational complexity, and predictive capability. By analyzing the same types of service systems using different methodologies, we aim to clarify when each approach is most effective and how they can complement one another in supporting decision making for complex service systems.

3.1. Analytical model approach

Analytical modeling has long been one of the primary methodologies for studying stochastic service systems. In analytical models, system behavior is represented through mathematical equations that describe the relationships between input variables and system performance measures. By solving these equations, researchers can derive explicit expressions for key performance indicators such as waiting times, queue lengths, and server utilization. Several classical analytical results in queueing theory illustrate the power of this approach. One well-known example is the family of Erlang formulas, including Erlang B, Erlang C, and Erlang A models, which were originally developed to analyze telephone traffic systems. These models provide closed-form expressions for important performance measures such as blocking probabilities and waiting-time distributions in multi-server systems. Despite being developed more than a century ago, Erlang-type models continue to serve as fundamental tools for analyzing modern service systems such as call centers, telecommunications networks, and customer service operations. Another influential analytical result is Whitt's square-root staffing rule, which provides a simple approximation for determining the appropriate number of servers required in large-scale service systems [12]. Under this rule, the number of required servers is approximately given by $s = a + \beta\sqrt{a}$, where a is offered load, a ratio of arrival rate to service rate, and β represents a parameter reflecting the desired service quality level. This formula reveals an important structural insight: the number of servers required to maintain acceptable service levels grows proportionally to the square root of the system load rather than linearly with demand. Such insights are difficult to obtain without analytical modeling. This result has been widely used and further refined in the literature [13, 24]. These classical examples demonstrate that

analytical models can produce closed-form expressions that reveal the underlying structure of service systems. Such structural insights are particularly valuable for designing operational policies and optimizing decision variables, making analytical modeling an important tool for prescriptive analytics.

To illustrate how analytical modeling can be applied to border-crossing systems, we refer to the work by Zhang [26]. In this model, vehicles arrive randomly at a border-crossing inspection station and are served by multiple inspection booths. To capture characteristics of the system operations, Zhang [26] introduced a congestion-based staffing (CBS) policy, in which the number of open inspection booths changes according to predefined congestion thresholds. By analyzing the steady-state behavior of this system, Zhang [26] derived closed-form expressions for several important performance measures, including the expected queue length, the expected waiting time of vehicles, and the probability distribution of the number of vehicles in the system. Here is a summary of assumptions made in Zhang [26]:

(1) The customers (cars) arrive at the border-crossing station according to a Poisson process with a time-dependent arrival rate λ.
(2) The service time of each server (inspection booth) follows the same exponential distribution with a rate μ.
(3) The number of servers on duty can be either at a high level c, maximum number of servers available, or at a low level $c - e$, depending on the queue length. Whenever the number of waiting customers is greater than a threshold N and the system is at a low staffing level (i.e. $c - e$ servers on duty), the e servers will resume serving the queue or the number of servers increases from $c - e$ to c. Whenever the number of idle servers reaches e, these e servers are shut down and will resume service until the queue length reaches N again. Such a service-capacity-change cycle repeats.
(4) The whole system has been operating for a long time and the steady-state (stationary state) has been reached.
(5) There is no set-up time to add additional servers when the queue exceeds N.

Under these assumptions, we can treat the border-crossing station as an *M/M/s* queue with s switches between c and $c - e$. The system performance measures such as expected waiting time (or queue length) and operating cost can be computed by some closed-form formulas derived by Zhang [26]. For example, the expected queue length (including customers in service) under the CBS policy can be expressed as the function of the policy parameter N, a nice formula as follows:

$$E(L) \approx (kN^2 + lN + q) / 2(kN + r) \tag{1}$$

where k, l, q, and r can be computed from the system parameters such as arrival rate (λ), service rate (μ), maximum number of inspection booths (c), and the number of idle inspection booths to be shut down (e). The details of computing $k, l, q,$ and r can be found in Zhang [26]. With these formulas, it is very easy to develop a cost or profit function to optimize the CBS policy. For example, under a given cost structure such as the unit customer (vehicle) waiting cost rate h and staffing level changeover cost S, a cost minimization threshold can be expressed by a nice square root formula:

$$N = -\frac{r}{k} + \sqrt{\left(\frac{r}{k}\right)^2 + \psi},$$

where

$$\psi = \frac{2S}{kh} + \frac{q}{k} - \frac{rl}{k^2}.$$

Another advantage of this analytical approach is that there are not many parameters involved (only l, μ, c and e). Thus, we do not need a lot of empirical data for parameter calibration. These formulas clearly illustrate how system performance depends on key operational parameters such as the arrival rate, the service rate, the number of inspection booths, and the congestion threshold used in the staffing policy. Such explicit relationships are one of the main advantages of analytical modeling, because they provide direct insights into how operational decisions influence system performance.

However, these advantages rely on simplifying assumptions that ensure mathematical tractability. These assumptions may not hold in many practical scenarios. First, Assumption 1 may be violated if the arrival rate varies over time. In practice, arrival rates often depend on the time of day and may be represented as a time-dependent function, denoted by $\lambda(t)$, particularly when the planning horizon spans several hours (e.g., 8 or 24 hours). Second, the service time (inspection time) may not follow an exponential distribution. In reality, inspection durations may follow a phase-type distribution, as illustrated in [27], or even a distribution with time-dependent parameters. Furthermore, inspection booths may be operated by officers with different experience levels, leading to heterogeneous service rates across servers. Under such circumstances, Assumption 2 may also be violated. Third, the actual staffing policy may involve multiple staffing levels that change dynamically or irregularly over time due to various reasons, rather than only two fixed levels as assumed in the analytical model. Because of the violation of the first three assumptions, Assumption 4, which typically relies on steady-state conditions, may also fail to hold in practice. Finally, Assumption 5 may not be realistic because opening an inspection booth often requires a significant setup time, such as opening the gate and activating the computer system used for processing travelers. Unfortunately, incorporating non-zero setup times (or changeover times) into analytical models can greatly increase the mathematical complexity of the model and may even render the problem analytically intractable. In summary, a key limitation of the analytical modeling approach is that obtaining tractable closed-form results often requires a number of simplifying assumptions, some of which may not be fully satisfied in real operational environments.

Despite these limitations, analytical models, despite moderate violations of the assumptions, can still remain reasonably accurate to predict system performance and guide the design of operational policies. For example, when the arrival rate varies over time (a violation of Assumption 1), the planning horizon can be divided into a number of shorter time intervals, such as one- or two-hour periods. Within each interval, the system can be approximated as a stationary system with a constant arrival rate. This approach is commonly referred to as the stationary independent period-by-period (SIPP) approximation, which was proposed by Green et al. [11]. The SIPP approximation works well when the stationary

system formulas are relatively robust with respect to moderate variations in arrival rates within each time interval. The effectiveness of this approximation has been confirmed in several real-world studies, including Zhang [26]. Readers are therefore referred to these studies for examples in which analytical models can still provide reasonably accurate approximations for real operational systems. A similar robustness analysis can be conducted when the assumption of exponentially distributed service times (Assumption 2) is violated. In the context of border-crossing stations, inspection times may follow empirically observed distributions rather than the exponential distribution assumed in the analytical model. For example, Lin et al. [18] suggested that inspection times can be reasonably approximated by an Erlang-2 distribution. In such cases, the robustness of the analytical model can be evaluated by simulating the system using empirically observed service-time distributions and comparing the simulation results with the performance measures predicted by the analytical model developed in Zhang [26], such as the formulas presented in Equation (1). Simulation experiments indicate that, for mean-based performance measures such as expected queue length or expected operating costs, the analytical model can remain reasonably robust even when the service-time distribution deviates from the exponential assumption. In particular, Zhang's analytical model provides fairly accurate predictions when service times follow Erlang-distributed inspection times as proposed by Lin et al. [18]. However, for distribution-based performance measures, such as the risk profile or variability of system performance, the robustness of analytical models may become more questionable. In such situations, simulation models provide a useful complementary tool for evaluating system performance under more realistic assumptions. For systems in which Assumption 3 is not satisfied, the analytical model may no longer be applicable because the resulting system becomes significantly more complex and closed-form formulas are generally not obtainable. In practice, however, staffing levels at border-crossing stations during peak periods are often adjusted infrequently, and the use of two staffing levels (low and high) can serve as a reasonable approximation during these busy periods. It should also be noted that changing staffing levels requires a certain amount of setup time, and such temporary service interruptions may increase congestion levels significantly during rush hours. In general operational environments, however, staffing policies may involve multiple staffing levels that change dynamically depending on the queue length or congestion level. When such multi-level dynamic staffing policies are introduced, the resulting analytical model becomes considerably more complicated, and deriving closed-form formulas for system performance measures becomes extremely difficult or even infeasible. In such cases, numerical methods or simulation approaches are typically required to evaluate system performance. Similarly, when the system does not reach a steady state or when Assumption 4 is violated, analytical models may become difficult to apply because time-dependent system dynamics can make the analysis mathematically intractable. Nevertheless, if the primary interest lies in mean-based performance measures, the steady-state formulas derived from analytical models may still provide useful approximations for transient systems. Another important factor is the setup time required for staffing level changes. In many analytical models of stochastic service systems, setup times are assumed to be zero because incorporating non-zero setup times often leads to substantial analytical complexity and may render the model intractable. This assumption explains the use of Assumption 5 in many analytical formulations. In reality,

however, setup times are typically non-zero and may significantly affect system performance. Ignoring such setup times may therefore introduce bias into the analytical results.

It is worth noting that in practice - particularly during non-peak periods - the staffing level (i.e., the number of open inspection booths) may not follow a deterministic rule but may change randomly depending on operational circumstances. Under such conditions, both analytical models may face difficulties in representing the system accurately. In such situations, alternative approaches such as learning-based models that rely on historical operational data may be more appropriate.

In summary, analytical modeling provides a powerful framework for studying stochastic service systems and can yield valuable structural insights through closed-form formulas. These analytical results help reveal key system parameters—such as arrival rates, service rates, and staffing level—affect system performance and therefore provide important guidance for policy design and decision making. However, its applicability often depends on the validity of simplifying assumptions. When these assumptions are not fully satisfied, analytical formulas may still provide useful approximations for certain performance measures, but their accuracy and applicability may become limited. In situations where system dynamics are too complex to be represented by tractable mathematical models, we should consider alternative modeling approaches when analyzing system behavior.

3.2. Simulation model approach

Simulation modeling provides a flexible computational framework for analyzing stochastic service systems whose dynamics are too complex to be represented by tractable analytical models. Unlike analytical modeling, which relies on mathematical equations and closed-form formulas, simulation models represent system operations explicitly that mimic the evolution of the system over time. By repeatedly generating random realizations of system events such as arrivals, service completions, and routing decisions, simulation models can estimate key performance measures including waiting times, queue lengths, and resource utilization.

Under the SM approach, the system dynamics is modeled as the interaction of random variables, which are classified into independent and dependent variables. Values of each independent random variable (also called input variable) are randomly generated (or sampled) based on its probability distributions. Values of each dependent random variable (also called intermediate or output variable) are computed based on the relations with the independent variables. The goal of the simulation approach is to find system statistical behaviors by repeating a large number of replications.

A key advantage of simulation modeling is its ability to incorporate realistic system features that are often difficult to capture in analytical models. For example, simulation models can accommodate time-dependent arrival patterns, general service-time distributions, heterogeneous servers, dynamic staffing policies, and non-zero setup times when opening or closing service stations. These features allow simulation models to represent operational complexities commonly observed in real service systems with fewer structural assumptions.

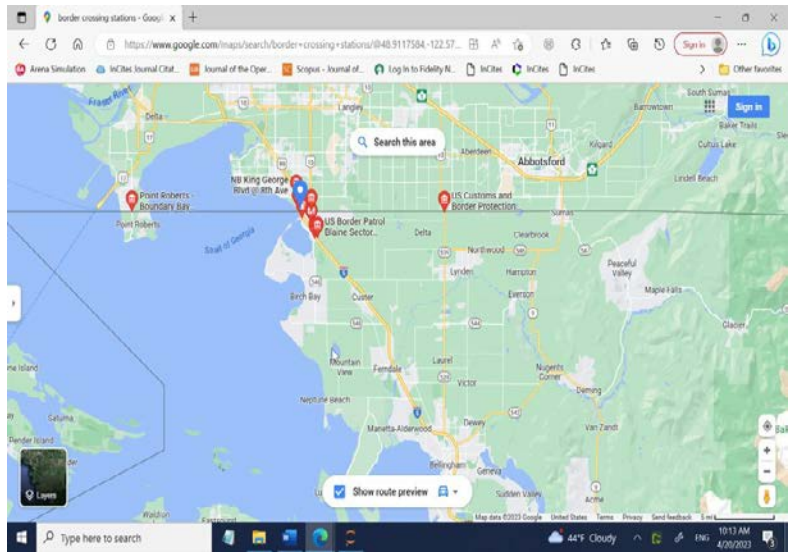


Figure 1. Four Border-Crossing Stations between Blaine and Sumas, WA (Surrey and Abbotsford, BC).

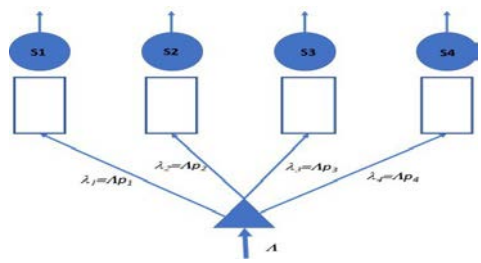


Figure 2. A parallel queueing system (supermarket model).

To illustrate the simulation modeling approach, consider again the border-crossing system introduced earlier. In practice, vehicles arriving at a border region may choose among multiple border-crossing stations based on factors such as travel distance, expected waiting times, or publicly available congestion information, as illustrated in Figure 1. Such traveler routing behavior can significantly affect congestion levels at different crossing locations. Incorporating these behavioral interactions into an analytical model can be extremely difficult. This kind of multi-queue configuration is called parallel queueing system as illustrated in Figure 2 and can be extremely difficult to analyze using the AM approach with customer choice behavior. However, a simulation model can be developed to include these complicating factors explicitly. For example, a discrete-event simulation model based on Arena can generate vehicle arrivals, simulate inspection processes at multiple booths, incorporate traveler routing decisions among multiple border-crossing stations, and evaluate the impact of alternative staffing policies on system performance, as illustrated in Figures 3 to 7.

Simulation modeling is also widely used in healthcare service systems, such as specialist medical clinics where patients are scheduled to receive consultation from a physician. A representative example, based on one of the authors' consulting projects, involves scheduling patients in a specialist physician's clinic where AM approach fails.

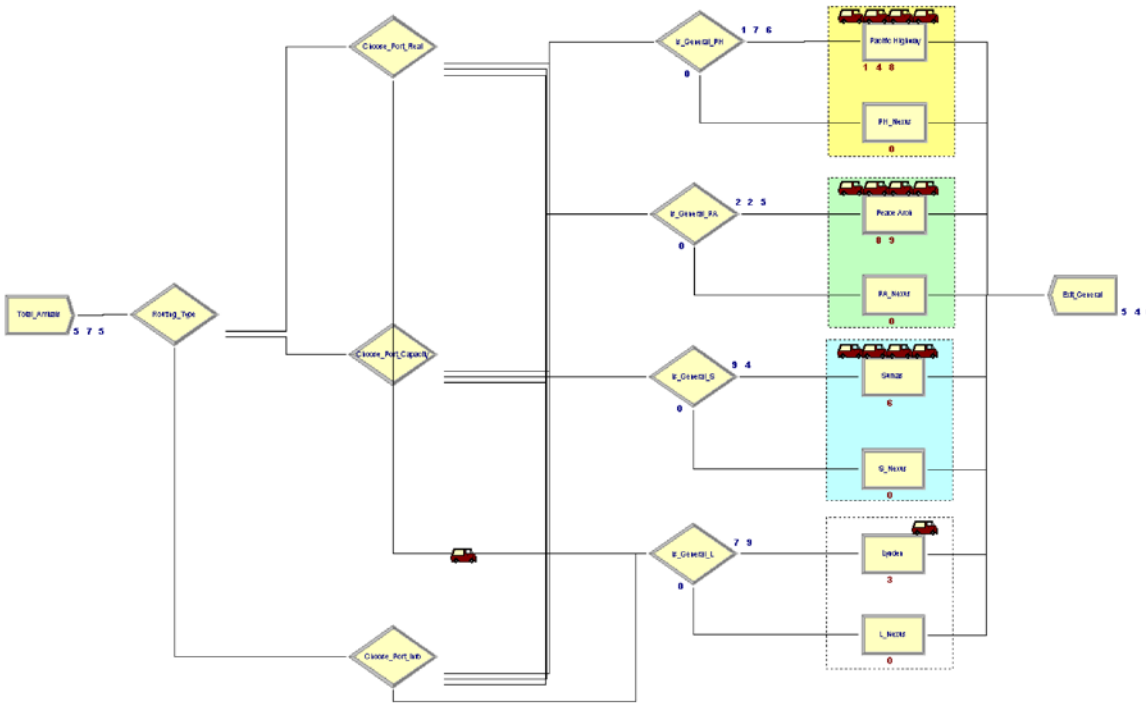


Figure 3: Inter-Dependent Systems with Information Sharing.

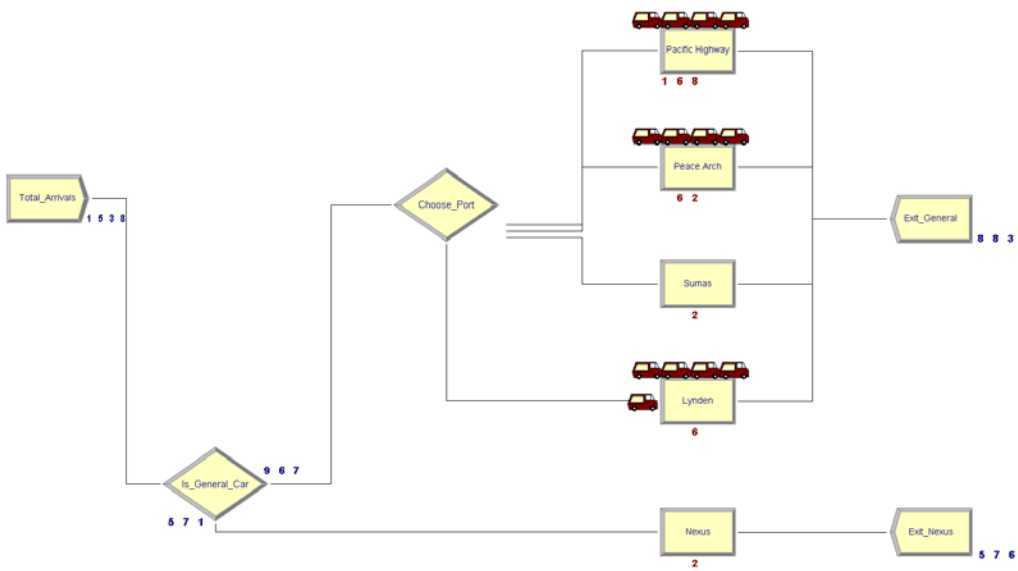
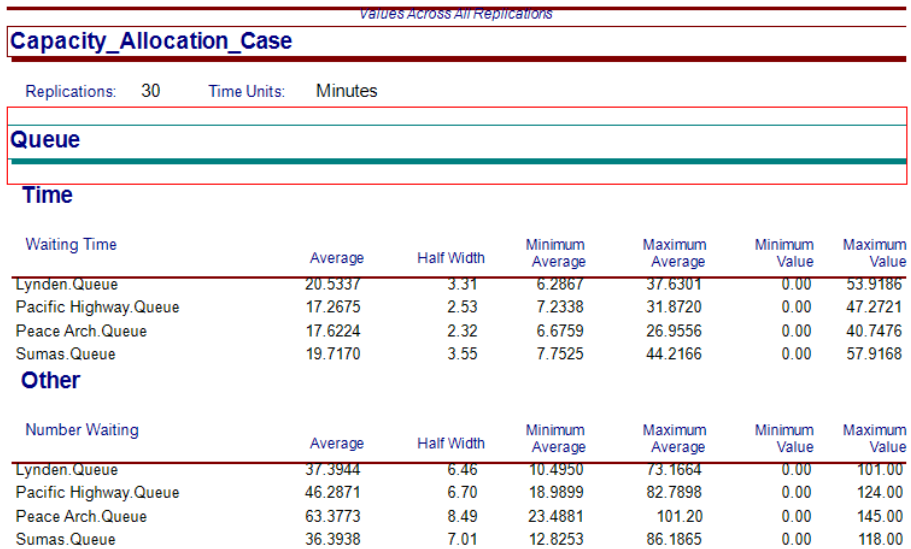


Figure 4. Independent Systems without Information Sharing.



Values Across All Replications

Capacity_Allocation_Case

Replications: 30 Time Units: Minutes

Queue

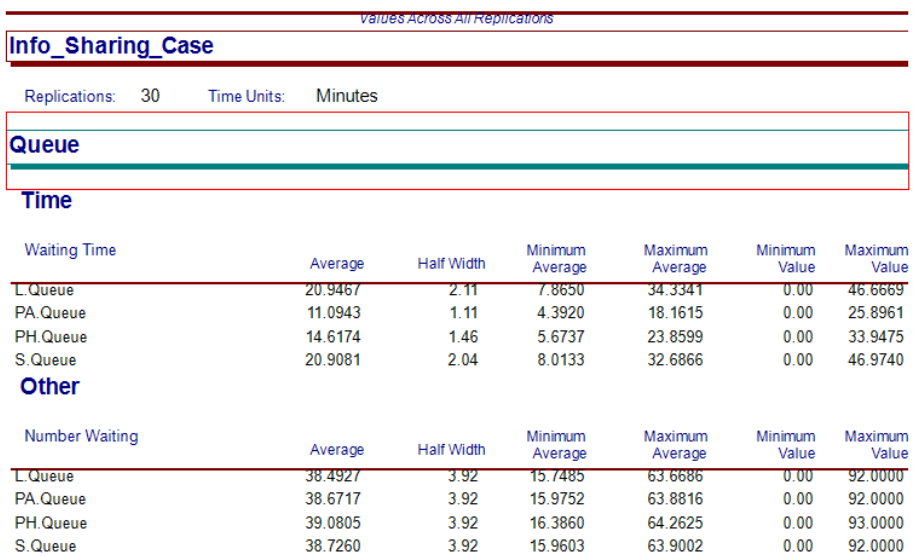
Time

Waiting Time	Average	Half Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Lynden.Queue	20.5337	3.31	6.2867	37.6301	0.00	53.9186
Pacific Highway.Queue	17.2675	2.53	7.2338	31.8720	0.00	47.2721
Peace Arch.Queue	17.6224	2.32	6.6759	26.9556	0.00	40.7476
Sumas.Queue	19.7170	3.55	7.7525	44.2166	0.00	57.9168

Other

Number Waiting	Average	Half Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Lynden.Queue	37.3944	6.46	10.4950	73.1664	0.00	101.00
Pacific Highway.Queue	46.2871	6.70	18.9899	82.7898	0.00	124.00
Peace Arch.Queue	63.3773	8.49	23.4881	101.20	0.00	145.00
Sumas.Queue	36.3938	7.01	12.8253	86.1865	0.00	118.00

Figure 5. Some results for flow allocation based on service capacities of ports of entry.



Values Across All Replications

Info_Sharing_Case

Replications: 30 Time Units: Minutes

Queue

Time

Waiting Time	Average	Half Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
L.Queue	20.9467	2.11	7.8650	34.3341	0.00	46.6669
PA.Queue	11.0943	1.11	4.3920	18.1615	0.00	25.8961
PH.Queue	14.6174	1.46	5.6737	23.8599	0.00	33.9475
S.Queue	20.9081	2.04	8.0133	32.6866	0.00	46.9740

Other

Number Waiting	Average	Half Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
L.Queue	38.4927	3.92	15.7485	63.6686	0.00	92.0000
PA.Queue	38.6717	3.92	15.9752	63.8816	0.00	92.0000
PH.Queue	39.0805	3.92	16.3860	64.2625	0.00	93.0000
S.Queue	38.7260	3.92	15.9603	63.9002	0.00	92.0000

Figure 6. Some results for flow allocation based on wait information disclosure.

Values Across All Replications

Real_Flow_Case

Replications: 30 Time Units: Minutes

Queue

Time

Waiting Time	Average	Half Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Lynden.Queue	4.7197	1.44	1.0590	15.3661	0.00	27.0415
Pacific Highway.Queue	46.2267	2.82	30.2755	67.4845	14.8560	83.8958
Peace Arch.Queue	20.5977	2.45	6.1438	35.9859	0.00	44.4140
Sumas.Queue	3.3663	0.92	0.7358	12.3796	0.00	20.9000

Other

Number Waiting	Average	Half Width	Minimum Average	Maximum Average	Minimum Value	Maximum Value
Lynden.Queue	7.6137	2.44	1.5554	25.0078	0.00	46.0000
Pacific Highway.Queue	146.76	9.55	95.3645	210.15	48.0000	264.00
Peace Arch.Queue	74.2696	8.88	22.5533	130.76	0.00	166.00
Sumas.Queue	5.2183	1.51	1.0056	19.6706	0.00	37.0000

Figure 7. Some results based on real flows to the p.

The example concerns a urology clinic located in British Columbia, Canada. This clinic maintains a long waiting list of patients who are referred by family physicians for consultation. An important operational task for the clinic is to determine how many patients should be scheduled for consultation on each working day. The consultation time required for each patient is random and follows a probability distribution whose parameters (e.g., mean and standard deviation) can be estimated from historical operational data. A major challenge facing the scheduler is to determine the appropriate number of patients to schedule each day while balancing several operational objectives. These objectives include minimizing the waiting time experienced by patients who arrive for their appointments, minimizing overtime working hours for the physician and supporting staff, and maximizing the number of patients served per day in order to reduce the backlog of patients on the waiting list and/or maximizing the clinic's profit. For example, if the clinic operates for eight hours per day (i.e., 480 minutes) and the mean consultation time for a patient is 30 minutes, scheduling 16 patients might initially appear to be reasonable based on a simple mean-value calculation. However, because consultation times are stochastic rather than deterministic, such a mean-based scheduling rule can perform poorly in practice. The variability in service times may lead to substantial patient waiting times during the day or significant overtime working hours for the physician and clinic staff. From a queueing modeling perspective, the clinic service process can be represented as a D/G/1 model with a finite operating horizon each day, where patients arrive according to a predetermined appointment schedule and service times follow a general probability distribution. Additionally, the system never reaches steady state. In this setting, key performance measures—such as average patient waiting time and staff overtime—cannot be obtained in closed form using classical analytical queueing models such as the M/M/1 queue, which typically assume Markovian dynamics and steady-state operating conditions. The failure of these assumptions makes traditional analytical modeling approaches difficult to apply in this context. In contrast, a simple simulation model, such as a spreadsheet-based simulation

implemented in Excel, can be constructed relatively easily to evaluate system performance under different scheduling policies. By simulating patient arrivals and consultation times, the model can estimate important performance measures such as average patient waiting time and total overtime working hours. Decision variables such as the number of patients scheduled per day or the spacing between appointment times can then be adjusted to evaluate different scheduling policies. For example, under a specified cost structure that accounts for patient waiting time and staff overtime costs, the simulation model can be used to identify an appropriate number of patients to schedule each day. Figures 8 and 9 illustrate such a simulation model and its application to policy optimization. Using this approach, the simulation analysis indicates that scheduling approximately 35 patients per day provides a reasonable balance between reducing patient waiting times, controlling staff overtime, and improving the utilization of clinic resources.

In summary, simulation modeling provides a flexible and practical approach for analyzing stochastic service systems when analytical models become mathematically intractable or when system assumptions are difficult to justify. By explicitly representing system operations and generating stochastic realizations of system events, simulation models allow for evaluating system performance under realistic operating conditions and for conducting what-if analyses by adjusting decision variables under alternative operational policies. Despite these advantages, simulation modeling also presents several challenges. Developing a reliable simulation model typically requires detailed knowledge of the system structure and the relationships among key variables. In addition, simulation experiments can be computationally intensive, especially when evaluating large-scale systems or performing optimization studies.

A Clinic with Scheduled Arrival Patients				Built by Z.G. Zhang ©									M=	1.005809				
Issue: Tradeoff between the idleness of doctor and the overtime of the business hours												AWT	LL of CI	15.36476	M=			1
Assumptions: Service time for each patient is normally distributed with mean of 30 min and standard deviation of 5 minites (estimated based on real data)													UL of CI	17.37638	# of repli			556.4086
Options: Schedule every 30 min, 40 min, 50 min					ST				R	600	9600	Profit	6773.962				VaR(p)	40.90843
					mean	30	stdev	10	h	0						27.5	522.5	stdev
Simulation Section		Interval between scheduled customers				hr	min		I	100		Averages	28.26038	261.9292	16.37057			12.03485
		30		Total business hours		8	480					Replication	OT	TW	AWT	Index	Sorted	
Patient #	Arrival	Sevice start	Service Duration	Service end	Waiting								72.40647	469.4878	29.34299	1	62.98453	
1	0	0	33.39426052	33.39426052	0				Output Metrics			1	23.61983	237.0742	14.81714	2	62.38058	
2	30	33.3942605	16.50241468	49.8966752	3.394261				Overtime	72.40647		2	32.13285	320.7284	20.04553	3	59.2621	
3	60	60	32.64124939	92.64124939	0				total waits	469.4878		3	2.758231	236.5946	14.78716	4	58.70124	
4	90	92.6412494	18.31948321	110.9607326	2.641249				AVG WT	29.34299		4	44.0772	182.9682	11.43551	5	56.01733	
5	120	120	44.54172917	164.5417292	0							5	33.00796	236.8457	14.80286	6	53.60471	
6	150	164.541729	26.14909626	190.6908254	14.54173							6	0	129.7421	8.108882	7	52.47504	
7	180	190.690825	48.78635639	239.4771818	10.69083							7	41.44617	507.8249	31.73906	8	51.0106	
8	210	239.477182	49.99450293	289.4716847	29.47718							8	0	82.32543	5.145339	9	50.98202	
9	240	289.471685	29.56336303	319.0350478	49.47168							9	55.3564	262.5714	16.41072	10	50.72685	
10	270	319.035048	30.23757894	349.2726267	49.03505							10	58.12706	360.8573	22.55358	11	50.22041	
11	300	349.272627	36.54873607	385.8213628	49.27263							11	39.54885	305.9692	19.12308	12	50.09454	
12	330	385.821363	33.13508458	418.9564473	55.82136							12	36.63517	518.3156	32.39472	13	49.08282	
13	360	418.956447	28.70716944	447.6636168	58.95645							13	11.98783	320.5264	20.0329	14	49.0177	
14	390	447.663617	9.685592957	457.3492097	57.66362							14	38.5788	351.8028	21.98767	15	48.62989	
15	420	457.34921	43.82339809	501.1726078	37.34921							15	11.355	93.1202	5.820012	16	47.5206	
16	450	501.172608	51.23386479	552.4064726	51.17261							16	16.29133	207.0256	12.9391	17	45.93545	
17	overtime	552.406473	44.02713029	596.4336029	#VALUE!							17	0	171.5018	10.71886	18	45.20088	
18	#VALUE!	#VALUE!	24.20447879	#VALUE!	#VALUE!							18	56.3362	816.1696	51.0106	19	45.13073	
												19	0	98.33759	6.146099	20	44.05529	
												20	23.04262	254.9825	15.93641	21	43.80739	
												21	50.60919	92.71035	5.794397	22	43.01072	
												22	38.09267	209.131	13.07069	23	42.74697	
												23	0	33.22188	2.076368	24	42.13961	
												24	13.56054	107.915	6.744687	25	41.68095	
												25	38.53285	362.7294	22.67058	26	41.07584	
												26	53.98525	466.7481	29.17176	27	40.94642	
												27	68.51666	488.0209	30.50131	28	40.90843	
												28	5.281806	46.43552	2.90222	29	40.544	
												29	16.87374	244.6454	15.29034	30	40.2762	
												30	25.72773	193.1615	12.07259	31	40.0587	
												31	60.36717	573.8919	35.86824	32	39.80997	
												32	2.695345	398.4734	24.90458	33	39.67767	
												33	21.26197	156.4783	9.779894	34	38.98689	
												34	0	215.4636	13.46585	35	38.97318	
												35	42.29426	315.1864	19.69915	36	38.82845	
												36	23.60827	196.4352	12.2772	37	38.73276	
												549	56.89695	399.5788	24.97368	550	0.748701	
												550	0	97.37099	6.085687	551	0	

In this case, analytical model is not available. Due to the general service distribution and finite time horizon. Traditional Markov chain approach does not work.

Figure 8. A Spreadsheet-based Simulation Model for a specialist clinic.

Schedule Time Every		profit
in	30	6961.184
minutes	35	7352.688
	40	7126.671

Profit=rN-TW*h-OT*l
N=number of patients scheduled per day
r=revenue per patiemy
h=waiting cost per patient per minute
l=overtime cost for running the clinic per minute

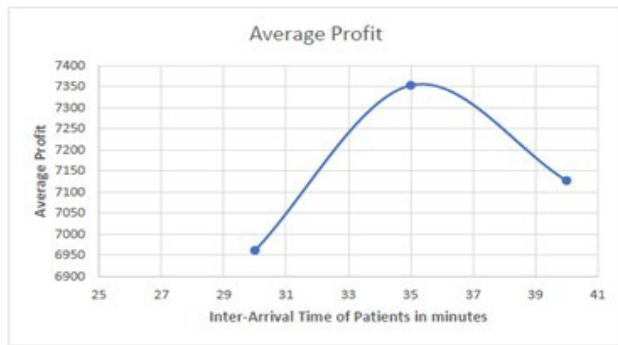


Figure 9. Average Profit of Clinic.

In many modern service systems, particularly those supported by advanced information technologies, the underlying relationships among variables may be unknown or be too complex to model analytically or through detailed simulation structures. However, large amounts of operational data are available. In such situations, data-driven approaches, such as machine or deep learning, provide an alternative modeling framework for predicting system performance.

3.3. Learning-based model approach

Learning-based models provides a data-driven approach for modeling stochastic service systems when the relationships between input variables and system performance measures are difficult to specify analytically or through explicit simulation logic. By learning patterns directly from historical data, these models can capture complex relationships among variables and generate forecasts of system performance. Recent advances in information technology have enabled service systems to collect and store large amounts of operational data. For example, modern border-crossing stations routinely record information such as vehicle arrival times, waiting times, traffic volumes, weather conditions, and other contextual variables (Cascade Gateway, https://www.cascadegatewaydata.com/CustomQuery; U.S. Customs and Border Protection, https://bwt.cbp.gov/). These datasets provide valuable opportunities to apply machine learning techniques to develop predictive models of border-crossing congestion. To illustrate the learning-based models, consider again the border-crossing system discussed earlier. The main reason for utilizing this system is due to the system complexity and availability of data. The objective in this case is to predict future waiting times at a particular border-crossing station based on historical observations and other relevant explanatory variables. A data-driven model can be trained using past data. Once trained, the model can generate predictions of future waiting times, which can then be used to provide information to travelers or to assist authorities in managing border-crossing operations. In practice, such prediction models can be formulated either as a regression or a classification problem. In a regression framework, the model directly predicts the expected waiting time as a continuous variable. Alternatively, in a classification framework, the model is to predict the waiting time in several predefined levels (for example, short, medium,

or long waiting times), along with their corresponding probabilities. Such probabilistic predictions can provide useful information for travelers and decision makers when evaluating alternative travel routes or border-crossing locations.

Data Collection*:* The border-crossing stations between Washington State in the United States and the province of British Columbia in Canada are among the busiest land border crossings in North America, handling millions of vehicles each year. During peak travel periods and holidays, heavy traffic can lead to significant delays and congestion. In recent years, technologies such as automated border kiosks and RFID-based systems have enabled the automatic collection of operational data at these facilities. Waiting-time and traffic data for several border-crossing stations in northwestern Washington and the Lower Mainland of British Columbia are publicly available from an official border-crossing information website. In this study, we use data from this source, and the key variables are summarized in Table 1.

Table 1. Characteristics of Data Set.

Time period: 2025-07-01-2025-12-31	Direction: southbound (entering US from Canada)	Vehicle type: cars
Window size: 5 minutes	Number of rows (observations) : 309,804	Number of columns (feature and label variables): 47
Ports of entry: Peace Arch (PA), Pacific Highway (PH), Lyn- den/Aldergrove (LA), Sumas/Huntingdon (SH)		

To develop machine learning models, it is necessary to define the input variables (features) and output variables used for prediction. The selected variables are summarized in Table 2. The input variables include time-related variables (e.g., month, day, hour, and minute), indicators for weekdays and holidays, and operational variables describing the system state at each port of entry, such as delay, queue length, number of vehicles in queue, service rate, traffic volume, and lane availability. The output variables correspond to the predicted waiting times at the four ports of entry (LA, PH, PA, and SH) for several prediction horizons (5, 15, 30, and 60 minutes ahead).

The variable EffectiveStart, which represents the start time of each five-minute observation window, was converted to a date-time format to facilitate time-based analysis. The dataset was then divided into training, validation, and testing subsets. The border-crossing dataset is automatically recorded 24 hours per day and 7 days per week by the border information system (Cascade Gateway, n.d.). A Python program was developed to retrieve the XML data and extract the number of open lanes at each port of entry. Based on these observations, the most frequent value (mode) of open lanes for each port and time period was used to construct four additional input variables: OpenLanes-LA, OpenLanes-PH, OpenLanes-PA, and OpenLanes-SH. These variables were then incorporated into the machine learning models to improve the prediction of border-crossing delays.

Table 2. Input and output variables for LM models.

Input Variables (Selected columns of data set):	**Output Variables** (Variables to be predicted): the delays at the four ports of entry in 5, 15, 30, and 60 minutes later:
EffectiveStart: start time of each 5-min window; Month: month 1-4; Day: day of the month 1-31; Hour: hour of the day 0-23; Minute of hour: 0, 5, 10, ..., 55; Weekday: whether it is a weekday, 0 or 1; Holiday: whether it is a US/Canada holiday; Delay-LA, Delay-PH, Delay-PA, Delay-SH: the average delay at the four ports of entry; QueueLength-LA, QueueLength-PH, QueueLength-PA, QueueLength-SH: the average queue length at the four ports of entry; VehiclesInQueue-LA, VehiclesInQueue-PH, VehiclesInQueue-PA, VehiclesInQueue-SH: the average vehicles in queue at the four ports of entry; ServiceRate-LA, ServiceRate-PH, ServiceRate-PA, ServiceRate-SH: the average service rate at the four ports of entry; Volume-LA, Volume-PH, Volume-PA, Volume-SH: the average volume at the four ports of entry; Availability-LA, Availability-PH, Availability-PA, Availability-SH: the average availability at the four ports of entry.	"Delay-LA-5min": Wait time in 5 min at LA; "Delay-PH-5min": Wait time in 5 min at PH; "Delay-PA-5min": Wait time in 5 min at PA; "Delay-SH-5min": Wait time in 5 min at SH; "Delay-LA-15min": Wait time in 15 min at LA; "Delay-PH-15min": Wait time in 15 min at PH; "Delay-PA-15min": Wait time in 15 min at PA; "Delay-SH-15min": Wait time in 15 min at SH; "Delay-LA-30min": Wait time in 30 min at LA; "Delay-PH-30min": Wait time in 30 min at PH; "Delay-PA-30min": Wait time in 30 min at PA; "Delay-SH-30min": Wait time in 30 min at SH; "Delay-LA-60min": Wait time in 60 min at LA; "Delay-PH-60min": Wait time in 60 min at PH; "Delay-PA-60min": Wait time in 60 min at PA; "Delay-SH-60min": Wait time in 60 min at SH;

Model Selection: To forecast future border waiting times, we consider six models from three classes: (1) machine-learning models: Random Forest and XGBoost; (2) deep-learning models: Long Short-Term Memory (LSTM) networks with dropout and Temporal Convolutional Network (TCN); and (3) traditional time-series models Moving Average (MA) and Exponential Smoothing (ES). These models capture different characteristics and complexities of the dataset. Because the analytical modeling (AM) and simulation modeling (SM) approaches are difficult to apply in this context due to non-stationary traffic patterns and complex relationships among many variables, our goal is to examine whether these data-driven models can outperform a naive forecasting method currently used in border-crossing operations. The naive forecasting approach estimates future waiting time by using the current queue length or current delay as the prediction for the next time period. For example, if the current queue length is 50 vehicles and the average service time per vehicle is assumed to be one minute, the predicted waiting time would be 50 minutes. This estimate is updated periodically as new queue-length information becomes available. However, this approach implicitly assumes that system conditions remain stationary within the forecasting horizon, which is often unrealistic for border-crossing systems. Another limitation of the naive approach is that providing only the average waiting time may not offer sufficiently useful information for travelers. In practice, it may be more informative to provide a range or an interval for the expected waiting time rather than a single point estimate. One way to achieve this is to categorize waiting times into several discrete delay levels. Based on historical data, we can produce boxplots over a 12-hour daytime period to show the wait time distribution as shown in Figure 10. Based on the wait time distribution, we notice the proportions in the following three intervals: 1-10: 69.6% , 10-30: 16.1%, and Over 30: 14.4%. We utilize these three intervals as three wait time classes representing three

congestion levels, as shown in Table 3. This approach turns the machine learning model into a classification type, labeled as 10/30 3-class model, although more than three category classification or the regression can be applied to produce multi-class delay or average wait forecasts (reported in online Appendix – website is available upon request).

Table 3. Waiting time classes.

Waiting time interval	**Waiting time class interpretation**
0-10 minutes:	This range represents very short wait times, indicating minimal delays
10 to 30 minutes:	This range captures medium delays
Over 30 minutes:	This range represents long delays

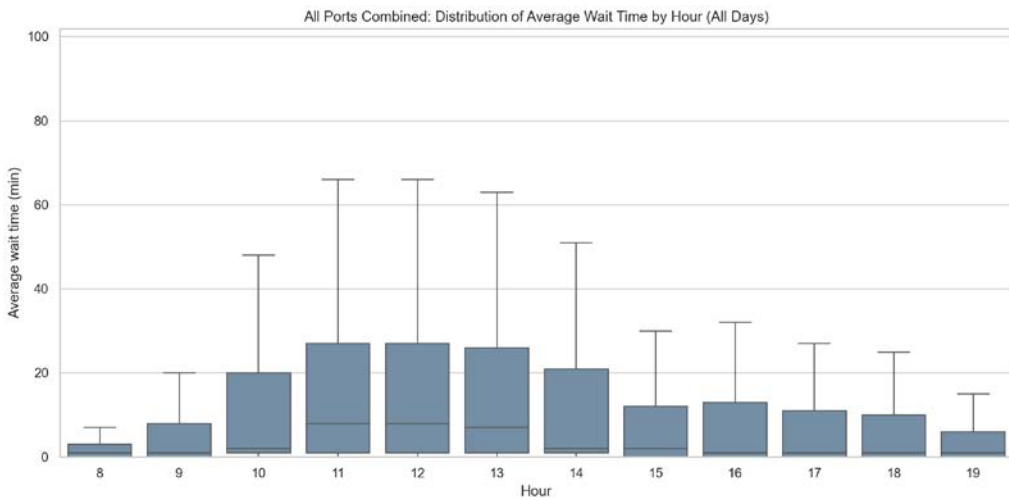


Figure 10. Waiting time distribution over 12-hour daytime period.

Table 4 summarizes model parameter selections, the optimal window size for the moving average (MA) model is empirically determined and provides a reasonable balance between smoothing historical observations and maintaining responsiveness to short-term fluctuations. Similarly, exponential smoothing constant for ES is appropriately chosen based on trial and error. For the machine learning models, key parameters for Random Forest and XGBoost, such as the number of trees, maximum tree depth, and learning rate, were tuned to improve predictive performance. For the deep learning model LSTM and TCN, parameters including the number of layers, number of hidden units, learning rate, and dropout rate were carefully selected to balance model complexity and prediction accuracy.

Evaluation Metrics: Evaluating classification models involves assessing their performance using a variety of metrics. Each metric offers insights into different aspects of the model's predictive capabilities. We also conducted other multi-class models such as 6-class model reported in Appendix, which has more imbalanced classes (i.e. some long-wait classes have fewer data points than short-wait classes). Compared with more heavily imbalanced class definitions, the 10/30 split should make the middle and upper classes easier to learn. Macro precision, recall, and F1 are especially important here because they

Table 4. Model parameter settings used in the 10/30 3-class experiment.

Model	**Configuration**
Moving Average	Horizon-aware rolling mean window: 3 (5 min), 6 (15 min), 12 (30 min), 24 (60 min); point forecasts converted to 3-class probabilities using a Gaussian residual model
Exponential Smoothing	Horizon-aware smoothing parameter (a): 0.40 (5 min), 0.30 (15 min), 0.20 (30 min), 0.15 (60 min); point forecasts converted to 3-class probabilities using a Gaussian residual model
Random Forest	600 trees; max depth 18; min samples split 20; min samples leaf 5; max features = $\sqrt{p}$; bootstrap = TRUE; balanced subsampling; random state = 42
XGBoost	Multi-class soft probability objective; 800 trees; learning rate 0.03; max depth 6; min child weight 5; subsample 0.8; column sample 0.8; (L_2) regularization 1.0; (L_1) regularization 0.1; evaluation metric = multiclass log-loss
TCN	Sequence length 36; 4 hidden layers with 64 channels each; kernel size 3; dropout 0.10; batch size 256; 60 epochs; patience 10; learning rate (5×10^{-4}); weight decay (10^{-5}); gradient clipping 1.0
LSTM	Sequence length 60; hidden size 64; 1 recurrent layer; head hidden size 128; dropout 0.0; batch size 64; 60 epochs; patience 10; learning rate (3×10^{-4}); gradient clipping 2.0; class-weighted cross-entropy loss

reveal whether models are performing well across all three classes rather than only the dominant low-delay class. Table 5 reports the overall prediction performance of the competing models for the three-class delay classification problem in terms of these major metrics. Among all models, XGBoost achieves the highest overall accuracy (0.8989) and strong performance across the macro-averaged precision, recall, and F1 metrics. The Random Forest model also performs competitively, with slightly lower accuracy but comparable F1 score. These results indicate that tree-based ensemble learning models are particularly effective for this type of structured operational dataset. The deep learning models, Temporal Convolutional Network (TCN) and LSTM, achieve reasonable predictive performance but do not outperform the tree-based models in this setting. This result is consistent with previous studies showing that deep learning models may not always provide significant advantages over ensemble tree methods when applied to tabular datasets with limited temporal structure. In contrast, the traditional statistical models, Moving Average and Exponential Smoothing, perform noticeably worse than the machine learning models, although they still provide reasonable baseline forecasts. Overall, these results demonstrate that learning-based models can significantly improve delay prediction accuracy compared with traditional time-series forecasting approaches, suggesting their potential usefulness for performance evaluation of stochastic service systems such as creating traveler information systems at border-crossing stations.

Table 5. Overall performance metrics of forecasting models (means across all prediction targets).

Model	**Accuracy**	**Macro Precision**	**Macro Recall**	**Macro F1**
XGBoost	0.8989	0.7732	0.7206	0.7346
Random Forest	0.8915	0.7373	0.7698	0.7468
Temporal Convolutional Network (TCN)	0.8794	0.7468	0.6839	0.7010
Exponential Smoothing	0.8544	0.6631	0.6544	0.6576
Moving Average	0.8480	0.6445	0.6352	0.6364
LSTM	0.8453	0.6615	0.5983	0.6025

Figures 11 and 12 illustrate the hourly prediction accuracy for the Peace Arch (PA) crossing under weekday and weekend traffic conditions (more results for other performance metrics at different ports are provided in Appendix). The dashed curve represents the average observed waiting time, allowing us to examine how forecasting performance varies with congestion levels. Several interesting patterns can be observed. First, prediction accuracy tends to decline during the period when congestion begins to increase rapidly (transient behavior of the stochastic process). In both figures, the drop in accuracy occurs slightly before the peak waiting time, indicating a lagging relationship between prediction performance and the evolution of congestion. This pattern suggests that forecasting models encounter the greatest difficulty during the transition from low congestion to rapidly increasing demand. Once the congestion level stabilizes (steady-state behavior of the stochastic process), prediction accuracy generally improves again even though the waiting time remains relatively high. Second, prediction performance on weekends appears more volatile than on weekdays. Weekend traffic exhibits sharper increases in congestion and larger fluctuations in accuracy, indicating that weekend travel patterns may be less regular (or more non-stationary) and therefore more difficult to predict. Overall, the results suggest that prediction accuracy is influenced not only by the level of congestion but also by the rate at which traffic conditions change.

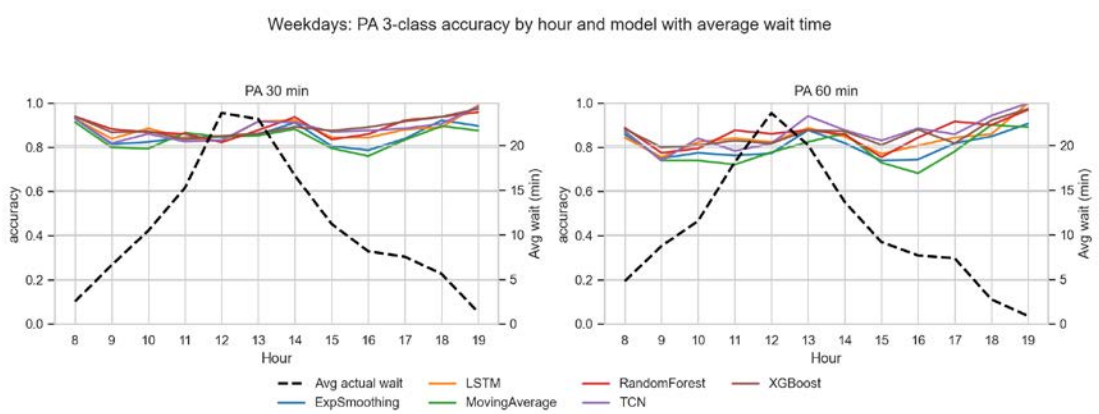


Figure 11. Weekday prediction accuracy at the Peace Arch (PA) crossing- southbound for cars.

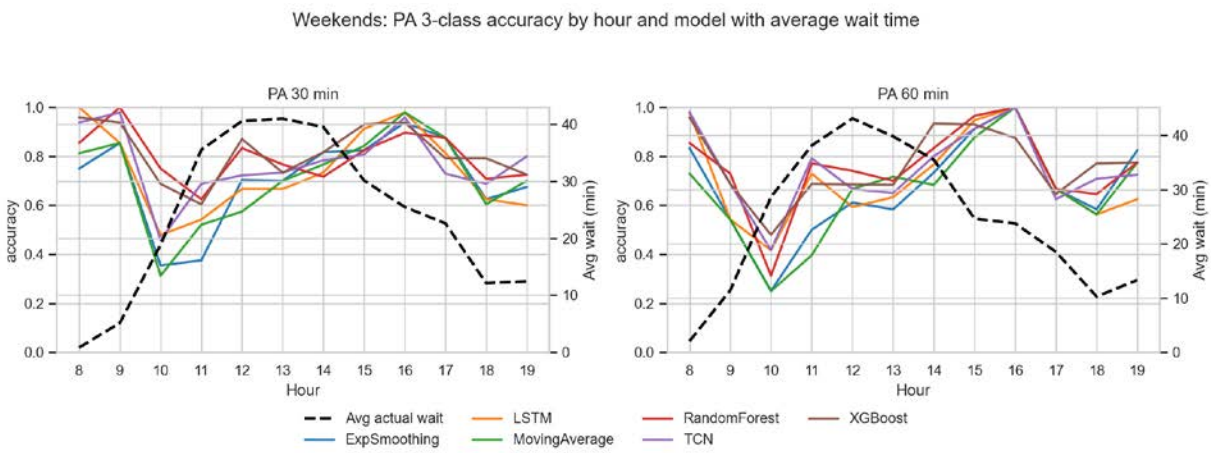


Figure 12. Weekend prediction accuracy at the Peace Arch (PA) crossing – southbound for cars.

Based on extensive experiments, it is observed that when system dynamics become highly time-dependent and non-stationary, both analytical and learning-based approaches may face substantial difficulty in providing accurate and reliable performance predictions. In such situations, simulation modeling often becomes the most dependable methodology, because it can explicitly represent complex operational dynamics without requiring closed-form assumptions or stable statistical relationships.

Figure 13 further illustrates the probabilistic forecasts generated by the XGBoost classification model for the Peace Arch crossing during weekend mornings. The stacked areas represent the predicted probabilities for the three classes of waits. As the actual waiting time increases, the model gradually reallocates prediction confidence from the low-delay class to the medium-delay class and eventually to the high-delay class. Compared with the weekday case, the weekend patterns appear less smooth and more irregular, reflecting the greater variability and less predictable nature of weekend traffic conditions. This result is consistent with the earlier accuracy figures, which showed that prediction performance on weekends is generally more volatile than on weekdays. Nevertheless, the classification-based approach still captures the overall progression of congestion levels and provides probabilistic information about future delays, which may be more informative for travelers and border management authorities than a single point estimate (produced by regression and presented in Appendix).

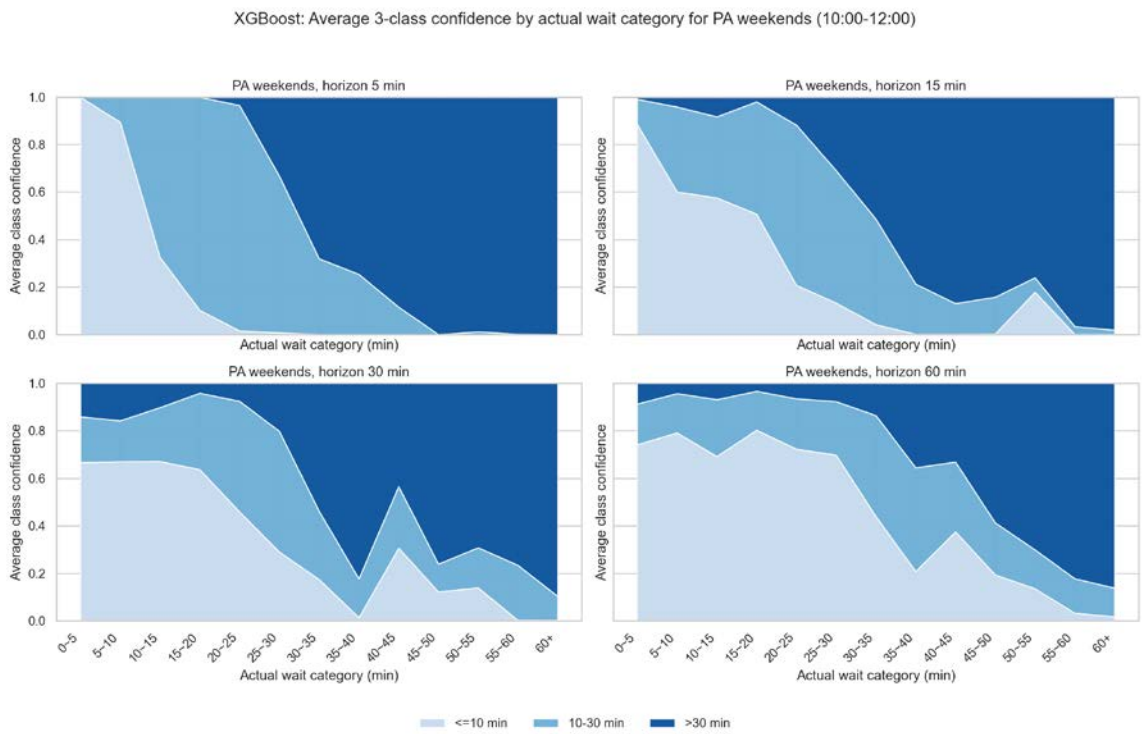


Figure 13. Weekend probabilistic forecasts based on the 3-class of waits at the Peace Arch crossing.

To further understand how the learning-based model makes its predictions, we examine the relative importance of the input variables used in the classification model. Figure 14 presents the feature importance ranking obtained from the Random Forest model for the three-class delay prediction problem. An interesting observation is that the most influential variables in the learning-based model differ from those typically emphasized in analytical queueing models. In classical analytical models of queueing systems, the key parameters are usually the arrival rate (traffic volume), service rate of servers, and the number of servers (e.g., open lanes), which determine the steady-state behavior of the system. However, in the learning-based model, variables related to the current system state, such as the number of vehicles in queue and the current delay levels at nearby ports of entry, appear to be much more important predictors. One possible reason for this difference is that analytical models are typically designed to analyze steady-state performance measures, whereas the machine learning models in this study aim to forecast future waiting times over short horizons. Such forecasts depend heavily on the current congestion level and the time-dependent evolution of the system rather than solely on structural system parameters. As a result, the learning-based models rely more strongly on observable real-time system-state variables, while variables such as service rates, traffic volume, and the number of open lanes play a relatively smaller role in the prediction task. This contrast highlights a fundamental difference between the two approaches: analytical models focus on theoretical system structure and steady-state relationships, whereas learning-based models exploit time-dependent data patterns and real-time system conditions to generate

forecasts.

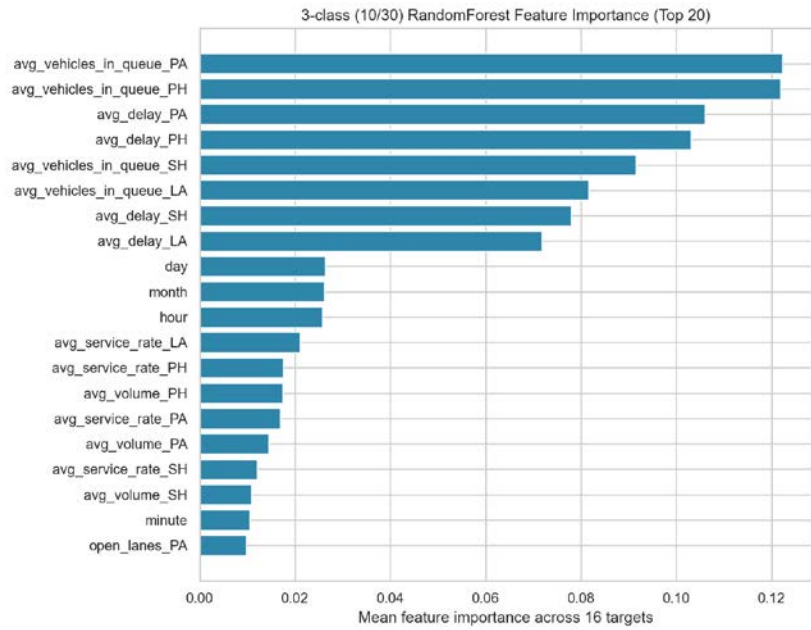


Figure 14. Feature importance ranking for the Random Forest delay classification model.

Overall, learning-based models offer several important advantages. First, they can capture complex nonlinear relationships among system variables without requiring explicit assumptions about the underlying system structure. Second, they can adapt to changing system conditions as new data becomes available. Third, they can generate predictive models even when the operational mechanisms of the system are only partially understood. However, these models also have several limitations. Unlike analytical models that provide closed-form expressions to reveal system behavior, the black-box nature of LM limits interpretability and prescriptive insight. Furthermore, training such models requires large and reliable data, and their prediction performance depend heavily on the availability and quality of data.

These observations highlight the complementary roles of the three modeling approaches to studying stochastic service systems. Analytical models provide valuable structural insights into system behavior through mathematically derived relationships among key parameters, such as arrival rates, service rates, and the number of servers. However, these models often rely on simplifying assumptions and are primarily designed for steady-state analysis. Simulation models, in contrast, offer greater flexibility in representing complex operational processes and dynamic policies, allowing analysts to evaluate system performance under realistic conditions when analytical solutions are not tractable. Learning-based models provide a different perspective by exploiting large datasets to generate predictions of future system performance.

4. Implications and Discussion

Our comparative analysis highlights important trade-offs and complementarities among

analytical modeling (AM), simulation modeling (SM), and machine learning (ML) approaches for analyzing stochastic service systems. Using examples from international border-crossing operations and healthcare service systems, we demonstrate that each methodology possesses distinct strengths and limitations. They can serve complementary roles in decision support within different operational environments.

4.1. Theoretical insights

Our analysis reveals several theoretical insights regarding the roles of the three modeling approaches.

Prescriptive versus predictive capabilities. Analytical models are well suited for prescriptive analysis, enabling decision makers to identify optimal control policies and system configurations. In contrast, learning-based models are primarily predictive, focusing on forecasting future system performance based on historical data. Simulation models occupy an intermediate position, supporting policy evaluation through experimentation rather than direct optimization.

Structural modeling versus data-driven learning. Analytical and simulation approaches rely on an explicit representation of system structure and operational mechanisms. In contrast, machine learning methods learn patterns directly from data without requiring explicit specification of the underlying system relationships. As a result, LM can be particularly useful in environments where system dynamics are highly complex or poorly understood.

Steady-state versus time-dependent analysis. Analytical models often focus on steady-state system behavior, whereas machine learning models emphasize time-dependent prediction based on evolving system states. Simulation modeling can accommodate both perspectives, enabling analysis of transient dynamics as well as long-run system performance.

These differences explain why no single approach dominates across all settings. Instead, the effectiveness of each methodology depends on the characteristics of the system and the objectives of the analysis.

4.2. Practical implications

Our results also provide several practical guidelines for selecting modeling approaches in stochastic service systems.

Analytical Models (AM): Analytical modeling approaches are particularly valuable for identifying optimal decision variables and policy parameters. Because AM models explicitly describe relationships among system variables, they can provide strong prescriptive insights for designing efficient operational policies. However, their applicability often depends on simplifying assumptions such as stationarity, known functional relationships, and tractable probability distributions. When these assumptions are violated, analytical solutions may become difficult or impossible to obtain.

Simulation Models (SM): Simulation modeling provides a flexible framework for analyzing complex stochastic systems whose dynamics cannot be captured by tractable analytical models. By explicitly representing system operations, simulation models allow practitioners to conduct what-if and sensitivity analyses, evaluate alternative policies, and

observe system behavior under different scenarios. Although simulation models may require significant computational effort and careful model construction, their flexibility makes them particularly useful for studying systems with complicated operational rules or dynamic policies.

Learning-based Models (LM): Learning-based approaches provide powerful data-driven predictive tools that can leverage large datasets to uncover complex and nonlinear relationships among system variables. These models include classical statistical forecasting methods, machine learning algorithms, and deep learning models. Because LM approaches learn patterns directly from observed data, they can generate accurate forecasts even when the underlying system dynamics are not fully understood. However, unlike analytical or simulation models, learning-based models generally do not provide direct prescriptions for optimal decision variables or operational policies. Consequently, LM approaches are particularly valuable in data-rich environments where prediction accuracy is the primary objective.

4.3. Policy and operational recommendations

From a practical perspective, the choice of modeling approach should depend on the characteristics of the system and the objectives of the analysis. Use analytical modeling (AM) when system relationships are well understood and the goal is to determine optimal policies or control variables under clearly defined assumptions. Use simulation modeling (SM) when system dynamics are complex but can still be reasonably described through operational rules, allowing analysts to evaluate alternative policies through scenario experimentation. Use learning-based models (LM) when large datasets are available but the relationships among variables are highly complex, dynamic, or poorly understood, and the primary objective is accurate prediction of system performance. In practice, these approaches should be viewed as complementary tools rather than competing methodologies. Combining structural insights from analytical models, operational experimentation from simulation models, and predictive capabilities from machine learning can provide a more comprehensive understanding of complex stochastic service systems.

4.4. An integrated decision-support framework: A border-crossing example

To take advantage of the distinct strengths of AM, SM, and LM approaches, this study effectively combine these approaches within a unified decision-support framework.

Figure 15 illustrates such a framework in the context of border-crossing operations to support congestion management and staffing decisions. In this framework, learning-based models primarily support demand-side (travelers) decision making by predicting congestion patterns; analytical models support supply-side (service provider) decision making by designing staffing and operational policies; and simulation models evaluate the combined effects of these policies on both system performance and customer behavior.

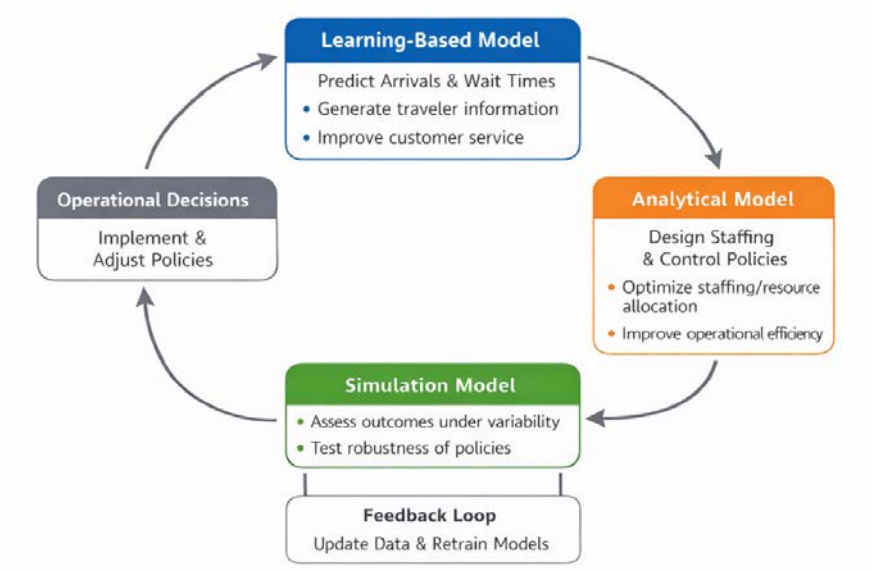


Figure 15. An Integrated AM–SM–LM decision-support framework for border-crossing management

Consider the multiple border-crossing stations connecting Washington State in the United States and the province of British Columbia in Canada, as discussed in previous sections. In such an environment, learning-based models can first be used to generate short-term forecasts of congestion conditions. Using historical and real-time operational data—such as traffic volumes, queue lengths, weather conditions, and time-of-day patterns—learning-based models can predict future waiting times or congestion levels at different ports of entry over short horizons (e.g., 5, 15, 30, or 60 minutes ahead). These predictions provide valuable real-time information about expected system conditions and can help detect early signs of congestion buildup. Once congestion forecasts are available, analytical models can be applied to design operational policies. For example, queueing-based analytical models can be used to determine optimal staffing thresholds or congestion-based staffing rules that specify when additional inspection booths should be opened or closed. Because analytical models provide explicit relationships between system parameters and performance measures, they can offer prescriptive guidance for adjusting staffing policies or control parameters to achieve desired operational objectives. Finally, simulation models can be used to evaluate the robustness of these policies under more realistic conditions. A simulation model can incorporate additional complexities that are difficult to capture analytically, such as driver route-choice behavior among multiple crossings, variability in inspection times, dynamic information sharing, and feedback effects between traveler decisions and congestion levels. Through simulation experiments, analysts can test how proposed staffing policies perform under different demand scenarios and assess the distributional performance of the system, including extreme congestion events.

This example illustrates how the three modeling approaches can play complementary roles in the analysis of complex stochastic service systems. Learning-based models provide predictive insights based on real operational data, analytical models offer structural

understanding and policy design, and simulation models enable detailed evaluation of operational policies under realistic system dynamics. Integrating these methodologies can therefore create a more comprehensive decision-support framework for managing congestion and improving operational performance in complex service systems.

5. Conclusion

This study presents a comparative analysis of three major methodologies for modeling stochastic service systems: analytical modeling (AM), simulation modeling (SM), and learning-based modeling (LM). Using representative examples from international border-crossing operations and healthcare service systems, we examine how these approaches differ in terms of modeling assumptions, analytical capabilities, computational requirements, and practical applicability. The results highlight the complementary roles of these methodologies in supporting decision making for complex service systems. Analytical models provide valuable structural insights into system behavior through closed-form relationships among key parameters such as arrival rates, service rates, and staffing levels. These insights are particularly useful for prescriptive analysis, allowing decision makers to determine optimal policies and operational control parameters. However, the usefulness of analytical models is often constrained by simplifying assumptions required for mathematical tractability. Simulation modeling offers a flexible alternative when system dynamics become too complex for analytical treatment. By explicitly representing operational processes and stochastic variability, simulation models allow analysts to conduct scenario analysis and evaluate alternative operational policies under realistic conditions. Simulation therefore serves as an important bridge between theoretical analysis and real-world operational experimentation. Learning-based models introduce a fundamentally different paradigm by leveraging large datasets to identify patterns and generate predictions of system performance. In environments where relationships among variables are highly complex, time-dependent, or poorly understood, learning-based approaches can provide powerful predictive capabilities. However, these models typically offer limited interpretability and do not directly provide prescriptive guidance for policy design.

The empirical analysis of border-crossing data illustrates several important observations. While learning-based models can achieve strong predictive performance, their accuracy tends to decline when system dynamics become highly non-stationary or when congestion conditions change rapidly. In contrast, simulation models remain robust in such environments because they explicitly represent the underlying operational mechanisms of the system. These findings suggest that these three approaches should be viewed as complementary tools, rather than competing alternatives, for understanding stochastic service systems.

From a decision-support perspective, the choice of modeling approach should depend on the characteristics of the system and the objectives of the analysis. Analytical models are most suitable when system relationships are well understood and the goal is to derive optimal policies or structural insights. Simulation models are appropriate when operational mechanisms are known but system dynamics are too complex for tractable analytical

solutions. Learning-based models are particularly valuable when large datasets are available but the relationships among variables are difficult to specify explicitly.

From a managerial perspective, the results of this study highlight the importance of selecting modeling tools that align with the characteristics of the operational environment and the decision objectives. In complex service systems, decision makers must balance multiple objectives including service efficiency, resource utilization, and customer waiting times. This study proposes an integrated framework that leverages the strengths of all three methodologies. In this framework, learning-based models generate short-term forecasts, analytical models guide policy design, and simulation models evaluate policy performance under realistic conditions. Understanding the complementary strengths of these methodologies can help managers design more effective decision-support systems for stochastic service operations.

Future research may further explore the framework that more tightly integrate these approaches. For example, analytical models can provide structural insights and theoretical guidance, simulation models can evaluate operational policies under realistic conditions, and learning-based models can provide accurate predictions based on real operational data. Such integrated approaches have the potential to significantly enhance decision-support capabilities for managing complex stochastic service systems. As service systems become increasingly data-rich and operationally complex, understanding how these modeling paradigms interact will be critical for advancing research and practice in decision support.

References

[1] Ang, E., Kwasnick, S., Bayati, M., Plambeck, E. L., & Aratow, M. (2015). Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management*, 18(1), 141–156. https://doi.org/10.1287/MSOM.2015.0560

[2] Arora, S., Taylor, J. W., & Mak, H. Y. (2023). Probabilistic forecasting of patient waiting times in an emergency department. *Manufacturing & Service Operations Management*, 25(4), 1489–1508. https://doi.org/10.1287/MSOM.2023.1210

[3] Brijmohan, A., & Khan, A. (2011). Development of methodology for quantifying the effectiveness of priority crossing programs at land border crossings. *Canadian Journal of Civil Engineering*, 38(8), 932-943.

[4] Brijmohan, A., & Khan, A. (2013). Adaptive road border crossing system management: the role of priority programs. *Journal of Advanced Transportation*, 47(2), 225–238. https://doi.org/10.1002/ATR.161;PAGEGROUP:STRING:PUBLICATION

[5] Cascade Gateway. Border wait times. Retrieved April 13, 2026, from https://cascadegatewaydata.com/Dashboard

[6] Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, 12(4), 519–549. https://doi.org/10.1111/J.1937-5956.2003.TB00218.X;WGROUP:STRING:PUBLICATION

[7] Chocron, E., Cohen, I., & Feigin, P. (2024). Delay prediction for managing multiclass service systems: An investigation of queueing theory and machine learning approaches. *IEEE Transactions on Engineering Management*, 71, 4469–4479. https://doi.org/10.1109/TEM.2022.3222094

[8] Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2), 79-141. https://doi.org/10.1287/MSOM.5.2.79.16071

[9] Garnett, O., Mandelbaum, A., & Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3), 208-227. https://doi.org/10.1287/MSOM.4.3.208.7753

[10] Green, L. (2006). Queueing analysis in healthcare. In R. W. Hall (Ed.), *Patient flow: reducing delay in healthcare delivery* (pp. 281–307). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-33636-7_10

[11] Green, L. V, Kolesar, P. J., & Soares, J. (2001). Improving the sipp approach for staffing service systems that have cyclic demands. *Operations Research*, 49(4), 549–564.

[12] Halfin, S., & Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3), 567–588. https://doi.org/10.1287/OPRE.29.3.567

[13] Janssen, A. J. E. M., Van Leeuwaarden, J. S. H., & Zwart, B. (2011). Refining square-root safety staffing by expanding Erlang C. *Operations Research*, 59(6), 1512–1522. https://doi.org/10.1287/OPRE.1110.0991;PAGE:STRING:ARTICLE/CHAPTER

[14] Jun, J. B., Jacobson, S. H., & Swisher, J. R. (1999). Application of discrete-event simulation in health care clinics: A survey. *The Journal of the Operational Research Society*, 50(2), 109. https://doi.org/10.2307/3010560

[15] Khan, A. M. (2010). Prediction and display of delay at road border crossings. *The Open Transportation Journal*, 04(1), 09–22. https://doi.org/10.2174/1874447801004010009

[16] Khoshons, M. K., Lim, C. C., & Sayed, T. (2006). Simulation and evaluation of international border crossing clearance systems. *Transportation Research Record*, 1966(1), 1–9. https://doi.org/10.1177/0361198106196600101

[17] Law, A. M. (2015). Simulation Modeling and Analysis (5th ed.). McGraw-Hill Education.

[18] Lin, L., Wang, Q., & Sadek, A. W. (2014). Border crossing delay prediction using transient multi-server queueing models. *Transportation Research Part A: Policy and Practice*, 64, 65–91. https://doi.org/10.1016/J.TRA.2014.03.013

[19] Moniruzzaman, M., Maoh, H., & Anderson, W. (2016). Short-term prediction of border crossing time and traffic volume for commercial trucks: A case study for the Ambassador Bridge. *Transportation Research Part C: Emerging Technologies*, 63, 182–194. https://doi.org/10.1016/J.TRC.2015.12.004

[20] Sharma, S., Kang, D. H., Montes de Oca, J. R., & Mudgal, A. (2021). Machine learning methods for commercial vehicle wait time prediction at a border crossing. *Research in Transportation Economics*, 89, 101034. https://doi.org/10.1016/J.RETREC.2021.101034

[21] Sun, Y., Teow, K. L., Heng, B. H., Ooi, C. K., & Tay, S. Y. (2012). Real-time prediction of waiting time in the emergency department, using quantile regression. *Annals of Emergency Medicine*, 60(3), 299–308. https://doi.org/10.1016/J.ANNEMERGMED.2012.03.011

[22] Tsai, S. C., Chen, H., Wang, H., & Zhang, Z. G. (2021). Simulation optimization in security screening systems subject to budget and waiting time constraints. *Naval Research Logistics*, 68(7), 920–936. https://doi.org/10.1002/NAV.21976;WGROUP:STRING:PUBLICATION

[23] U.S. Customs and Border Protection. Border wait times. Retrieved April 13, 2026, from https://bwt.cbp.gov/

[24] Whitt, W. (2007). What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics*, 54(5), 476–484. https://doi.org/10.1002/NAV.20243;WGROUP:STRING:PUBLICATION

[25] Yu, M., Ding, Y., Lindsey, R., & Shi, C. (2016). A data-driven approach to manpower planning at U.S.–Canada border crossings. *Transportation Research Part A: Policy and Practice*, *91*, 34–47. https://doi.org/10.1016/J.TRA.2016.06.006

[26] Zhang, Z. G. (2008). Performance analysis of a queue with congestion-based staffing policy. *Management Science*, 55(2), 240–251. https://doi.org/10.1287/MNSC.1080.0914

[27] Zhang, Z. G., Luh, H. P., & Wang, C. H. (2011). Modeling security-check queues. *Management Science*, 57(11), 1979–1995. https://doi.org/10.1287/MNSC.1110.1399