

STOCHASTIC SIMULATION IN A LAW FIRM

*A simulation-based analysis of case acceptance under stochastic workload
demand and capacity constraints tailored to law firms*

Rens van Haasteren

2738676

A thesis submitted for the Master's degree of
Econometrics and Operations Research

Supervised by
prof. dr. Bernd Heidergott

PREFACE

This report is submitted in partial fulfillment of the requirements for the Master's programme in Econometrics and Operations Research at Vrije Universiteit Amsterdam. The research is based on data from a mid-sized law firm in the Netherlands and focuses on the simulation of workload and billable capacity. The objective is to develop a tractable yet realistic modeling framework for evaluating different case acceptance policies. In addition, the study investigates the optimal level of allowed monthly overtime that maximizes revenue from overtime while accounting for exponentially increasing costs of overtime.

During the writing process, Writefull and ChatGPT were used to restructure specific sentences and to suggest synonyms. All final interpretations and conclusions remain my own.

I wish to thank my supervisor prof. dr. Heidergott for his valuable guidance throughout the development of my thesis. His conceptual insights were pivotal during the development and with his extended knowledge on stochastic optimization, he answered my specific questions with great enthusiasm.

CONTENTS

0	Abstract	4
1	Introduction	5
1.1	Problem statement	5
1.2	Research questions	5
1.3	Structure	5
2	Related work	6
2.1	Related work	6
2.1.1	Marked point processes	6
2.1.2	Demand forecasting with capacity constraints	6
2.2	Contribution	7
3	Data	8
4	Methodology	9
4.1	Demand forecasting	9
4.1.1	Arrival process	9
4.1.2	Marked point process	10
4.1.3	Workload	10
4.2	Capacity simulation	11
4.2.1	Historical capacity	11
4.2.2	Future capacity	12
4.3	Acceptance policies	13
4.3.1	Policies	13
4.3.2	Stochastic approximation	16
4.4	Randomization	18
5	Exploratory analysis	19
5.1	Demand forecasting	19
5.1.1	Marked point process	19
5.1.2	Workload	20
5.2	Trend and seasonality	21
5.3	Capacity planning	22
5.4	Supplementary methods	23
6	Results	26
6.1	Historical results	26
6.2	Simulation results	27
6.2.1	Policy 0: Accept all	27
6.2.2	Policy I: Hard daily capacity boundary	28
6.2.3	Policy II: Hard monthly capacity boundary	28
6.2.4	Policy III: Soft monthly capacity boundary	29
7	Discussion	33
7.1	Interpretations	33
7.2	Limitations	34
7.3	Challenges	34

8	Conclusion	35
8.1	Summary	35
8.2	Future work	35
9	References	37
10	Appendix	38
10.1	Dispersion	38
10.2	Closed cases	38
10.3	Correlation	38
10.4	Q-Q plots	39
10.5	Monthly distribution	39
10.6	Hours distribution	39
10.7	Daily capacity	40
10.8	Acceptance policies	40
10.8.1	Policy 0. All arrivals	40
10.8.2	Policy I. Hard daily capacity boundary	41
10.8.3	Policy II. Hard monthly capacity boundary	42
10.8.4	Policy III. Soft monthly capacity boundary	43
10.9	Computational intensity	43



ABSTRACT

Law firms operate under stochastic demand while facing capacity constraints, making case acceptance and workload management critical operational decisions. This thesis develops a simulation-based framework to evaluate case acceptance policies using real data from a mid-sized Dutch law firm. Client-level case arrivals are modeled via a Negative Binomial process to capture overdispersion and client heterogeneity, while case workloads and revenues are assigned based on fitted historical distributions. In addition, a billable capacity simulation is constructed based on the historical workforce which takes contractual adjustments into account.

The framework compares acceptance policies under daily and monthly capacity constraints over a multi-year horizon. Results highlight trade-offs between revenue generation and workload stability: unconstrained acceptance maximizes revenue but causes unrealistic workload, whereas capacity-aware policies stabilize workload at the cost of lower revenue. Introducing an allowed monthly overtime improves expected net revenue, but only up to a finite threshold, as exponentially increasing overtime costs eventually outweigh marginal gains. A reliable estimate of the optimal overtime level is determined with a symmetric finite-difference estimator and an asymptotically unbiased confidence interval is provided with the non-overlapping batch means scheme.

Overall, the study demonstrates how simulation-based decision support can quantify long-term trade-offs between revenue and capacity under uncertainty. While abstracting from certain operational complexities, the framework provides a flexible and extensible foundation for evaluating case acceptance policies in legal services and possibly other professional service organizations.

INTRODUCTION

1.1 PROBLEM STATEMENT

Law firms have historically been slow to adopt new technologies and integrate advanced analytical tools into daily operations [18]. Despite evidence that technology can improve efficiency, enhance case management, and support effective client communication, many firms continue to rely on manual and human-intensive workflows. This reluctance is often driven by leadership hesitancy, uncertainty about the effectiveness of new technologies, or concerns regarding the profitability. As a result, law firms are typically not proactive in optimizing internal processes or systematically identifying and capturing new business opportunities.

At the same time, the legal environment is getting increasingly complex. The volume of regulatory instruments in the Netherlands continues to grow [4], European legislative activity is expanding [13], and the European legal services market is projected to increase from \$257.48 billion in 2025 to \$301.69 billion in 2030 [12]. This combination of rising regulatory complexity and market growth places additional pressure on law firms to make timely and well-informed decisions regarding which cases to accept or reject. These developments underscore the need for systematic, data-driven approaches to case acceptance.

Evaluating case acceptance policies requires reliable demand forecasting and realistic capacity modeling. However, to the best of our knowledge, there are no publicly available demand forecasting or capacity models specifically developed to the operations of a law firm. By extension, there are no simulation-based frameworks that assess the performance of different case acceptance policies in this context. The problem addressed in this thesis is characterized by stochastic case arrivals and dynamically evolving capacity constraints. Because acceptance decisions directly affect future workload, capacity utilization, and revenue under uncertainty, the performance of acceptance policies cannot be evaluated analytically. Instead, it must be assessed through simulation-based experiments that explicitly account for stochastic demand and capacity.

1.2 RESEARCH QUESTIONS

To address these challenges, this thesis investigates the following research questions:

- RQ1. How can stochastic case arrivals and their associated workloads be simulated in a realistic yet tractable manner?
- RQ2. How do different case acceptance policies affect the trade-off between revenue generation and capacity utilization over a five-year simulation horizon?
- RQ3. What level of allowed monthly overtime maximizes overtime revenue with an exponentially increasing monthly overtime cost, when simulating over a five-year horizon?

1.3 STRUCTURE

The remainder of this thesis is organized as follows. Section 2 reviews the relevant literature on marked point processes and the different acceptance policies. Section 3 describes the dataset used for the analysis. Section 4 presents the methodological framework, including demand forecasting, capacity simulation, acceptance policies, and stochastic approximation. Section 5 provides an exploratory analysis of intermediate results and introduces additional design choices in Section 5.4. The main findings of the study are reported in Section 6. Sections 7 and 8 contain the discussion and conclusion, respectively. Finally, Section 9 lists all references, and Section 10 consists of the supplementary materials.

RELATED WORK

This section reviews related literature on marked point processes and demand forecasting under capacity constraints. Together this literature forms the methodological foundation for the simulation framework developed in this research. The section closes by mentioning the contribution of this research from an academic and a commercial perspective.

2.1 RELATED WORK

2.1.1 *Marked point processes*

In many real-world applications, arrivals are characterized by more than just their inter-arrival times. For example, inbound calls may be associated with the waiting times, while hospital visits include information on length of stay. Such additional attributes are essential for constructing realistic simulation models and effective decision-support systems. For law firms, cases are more often described by key characteristics such as workload, generated revenue and department, rather than their arrival date. This motivates the use of marked point processes, which extend traditional point processes by associating each arrival with additional information (marks), such as size, type, or value.

Zeller and Scherer [20] apply marked point processes to model cyber risk, where idiosyncratic cyber events (independently affecting individual firms) are represented as separate inhomogeneous Poisson processes, and systematic risk (affecting multiple firms simultaneously) is captured through event marks encoding severity and the set of affected firms. Their work illustrates the flexibility of marked point processes for modeling complex arrival data in which both event dependencies and attributes play a critical role.

Mancini and Paganoni [11] study the application of marked point processes to model hospitalizations of heart failure patients. In their framework, the authors research the temporal dynamics of hospital admissions with a mark representing the length of hospital stay. The constructed marked point process could allow hospitals to model the admissions of patients beforehand, improving capacity planning of staff and beds.

2.1.2 *Demand forecasting with capacity constraints*

In operational environments with limited resources, arrival modeling must be integrated with resource planning to support effective decision-making. This is particularly relevant in professional and emergency services, where both demand and available capacity are subject to stochasticity.

The study of Schuetz and Kolisch [15] is closely related to our problem. They consider a setting in which service providers must accept or reject booking requests from different customer classes. Their objective is to maximize profit while accounting for class-specific revenues, cost of overtime, and refunds for cancellations and no-shows. These revenues and costs are derived from the underlying appointment schedule. The problem is formulated as a continuous-time Markov decision process and solved using simulation-based approximate dynamic programming. An experimental investigation is done on the booking of patients with a fixed capacity to investigate the effectiveness of the proposed model. The optimal policy increased number of examinations by 9% and the average profit by 6.9%.

Barros, Weber, and Reveco [1] demonstrate that demand for emergency healthcare services can be forecasted with high accuracy using data-driven methods, with support vector regression outperforming in terms of predictive accuracy. Based on these forecasts, the authors develop capacity management logic that combines demand predictions with stochastic simulation to evaluate staffing and resource configurations. The resulting framework supports cost-benefit analyses of service quality improvements and resource allocation decisions and has been successfully implemented in a major pediatric hospital in Chile.

Similarly, Solomon et al. [16] examine multi-period stochastic resource planning in professional service organizations. By modeling sequential decision-making under uncertain project demand and supply uncertainty as a Markov Decision Process and applying approximate dynamic programming, they derive policies that dynamically allocate capacity and determine outsourcing needs. Their approach yields improvements in both profitability and resource utilization relative to traditional methods, however they do face computational challenges for larger projects.

These studies contain a methodological foundation for integrating stochastic arrivals, demand forecasting, and capacity constraints within simulation-based decision frameworks.

2.2 CONTRIBUTION

The literature indicates that data-driven and operations-research oriented studies focused specifically on the legal sector remain limited. In line with the historically slow adoption of new technologies, a lot of research is done on the implementation of AI and other technology in the law firms [5, 6, 14]. In addition, current research for law-firms focuses on predicting the judges decision in trials [8, 10].

The primary contribution of this thesis is the application of different acceptance policies to real operational data from a mid-sized Dutch law firm. In particular, this work attempts to develop a tractable simulation framework that models stochastic case arrivals, including associated revenue and workload, alongside a capacity model that explicitly incorporates contractual adjustments as well as the hiring of new employees.

Beyond its academic contribution, the proposed framework offers practical value for law firms and maybe other professional service organizations seeking to evaluate case acceptance policies and explore trade-offs between revenue generation and capacity utilization under uncertainty. As several modeling assumptions are required to ensure tractability and to remain within the scope of this study (see Section 7), the framework is designed as a flexible decision-support tool rather than a fully automated solution.

DATA

The data originate from the day-to-day operations of a mid-sized law firm where this research is conducted. It consists of timekeeping logs, relational database records, and HR data. From these sources, a dataset was constructed where each row represents a time registration, including information on the case, client, and employee. The final dataset contains 1,652,281 rows spanning the period from 2017 to 2025.

Some illustrative rows of the dataset are presented in Table 1: the first three rows of the dataset, the first row corresponding to a vacation activity, and the first row corresponding to a sick leave. To ensure privacy and avoid revealing sensitive information, employee, case, and client (ultimate parent) names have been anonymized. Similarly, the revenue values and hours in the table have been randomized. Section 4.4 offers more on the randomization of the results.

As shown in Table 1, the analysis is conducted at the ultimate parent level, rather than at the individual client level. This aggregation offers several benefits: more cases are grouped together under each parent, which allows distributional parameters per client to be estimated with higher precision. In addition, computation time is slightly reduced because the required variables per client are calculated fewer times. For brevity, the term client will henceforth be used to refer to the ultimate parent entity.

Non-billable activities, such as vacation or sick leave, have revenue equal to zero, as expected. Some rows contain missing values in the CASE CREATED and FTE columns. Table 2 summarizes the number of missing values per variable. Missing CASE CREATED and FTE values are addressed using the techniques introduced in Section 4.

DATE	ACTIVITY	PARENT	CASE CREATED	CASE	EMPLOYEE	FTE	HOURS	REVENUE (€)
2017-01-01	Bill	P1	2016-06-07	C1	E1	1	5.8	1,282.00
2017-01-01	Bill	P2	–	C2	E2	–	0.3	150.50
2017-01-01	Bill	P3	2016-11-18	C3	E3	0.8	0.1	41.50
2017-01-02	Vac	P4	2012-03-11	C4	E4	1	8	0
2017-01-02	Sick	P4	2016-01-15	C5	E5	–	8	0

Table 1: Example rows of the prepared dataset.

Variable	Number of missing values
DATE	0
ACTIVITY	0
PARENT	0
CASE CREATED	20,196
CASE	0
EMPLOYEE	0
FTE	455,644
HOURS	0
REVENUE (€)	0

Table 2: Number of missing values per variable in the dataset.

METHODOLOGY

This methodology section covers the techniques used to construct the simulation model. In Section 4.1, the demand forecasting model is covered, which results in a marked point process where each arrival has an arrival date, a revenue and a size. Section 4.2 introduces several steps and considerations to simulate the monthly billable capacity. In Section 4.3 four different acceptance policies are described by formalizing conditions on which arrivals are accepted based on the billable capacity. To conclude, Section 4.4 discusses a simple technique to ensure sensitive information is not leaked.

4.1 DEMAND FORECASTING

4.1.1 Arrival process

From the related literature, it is evident that a variety of stochastic processes have been employed to model the arrival of events, such as inbound calls of customer service, hospital visits, cyber risks [7, 11, 20]. This can also be applied to the arrival of legal cases. A commonly used framework for modeling such count data is the Negative Binomial distribution. This model accounts for overdispersion, where the variance exceeds the mean, which is observed in the historical arrival of cases of the law-firm (Appendix Figure 24).

Specifically, we have a discrete time model for each client, where each period represents the number of case arrivals in a month. A time-frame of a month is used; a smaller timeframe causes too many months with zero arrivals, while a larger timeframe would aggregate too much arrivals, obscuring meaningful patterns in the arrival behavior. Clients with less than a total of 10 cases are filtered out, due to insufficient data for parameter estimation and hypothesis testing. For each client the Negative Binomial success probability p and dispersion parameter r are fitted. To assess the goodness-of-fit of the model, the non-parametric Monte Carlo Kolmogorov–Smirnov test is used to determine whether the empirical and simulated arrival distributions differ significantly.¹

The probability mass function of the Negative Binomial distribution is given by

$$\mathbb{P}(X = k) = \binom{k+r-1}{k} (1-p)^k p^r, \quad k \in \mathbb{N}_0, \quad (1)$$

where $p \in (0, 1)$ denotes the success probability and $r \in \mathbb{R}^+$ is a dispersion (or shape) parameter.

The dispersion represents the difference between the variance and expectation of a random variable. The parameter r governs the degree of dispersion in the Negative Binomial. From the standard expressions for the mean and variance of the Negative Binomial, the dispersion can be computed:

$$\mathbb{E}[X] = \frac{r(1-p)}{p}, \quad \mathbb{V}\text{ar}[X] = \frac{r(1-p)}{p^2}, \quad \mathbb{V}\text{ar}[X] - \mathbb{E}[X] = \frac{r(1-p)^2}{p^2}. \quad (2)$$

Smaller values of r imply stronger overdispersion.

Empirical evidence (introduced in the exploratory analysis Section 5) from the arriving cases indicates that the total number of arrivals across the firm, as well as the majority of individual clients with more than 10 cases, closely follows a Negative Binomial distribution. This observation is supported by goodness-of-fit tests and visual comparison with empirical histograms. Consequently, the arrival process for each client is modeled using a Negative Binomial distribution with parameters fitted to historical monthly arrival counts.

¹A specific application of the test, where the goodness of fit of a Gamma-Poisson mixture model for a call center is checked, is shown in Jongbloed and Koole [7].

4.1.2 Marked point process

While modeling the number of arrivals provides insight into the frequency of cases, it does not capture any additional characteristics associated with each case. To simulate the law firm in a more realistic setting, we construct a marked point process. The arrival process described in the previous section is a standard point process, which represents a marked point process when auxiliary information is attached to each arrival. In our context, these marks are the revenue and size of the arrived case. The size represents the total number of billable hours of the case. It is important to note that both revenue and size are simulated a priori, even though in practice they are not observable before case completion.

Similar to the arrival process, we fit distributions to each auxiliary variable to formalize the sampling space and the associated probabilities. The marginal distributions are based solely on closed cases, defined as cases with no activity for more than 18 months (see Appendix Figure 25 where is shown that only 3% of the cases have a gap of more than 18 months where nobody worked on the case). Only taking into account the closed cases prevents including cases that are still in progress and therefore have no representative size and revenue. The exploratory analysis shows that both revenue and size are approximately lognormally distributed (Figure 2). In addition, we fit a joint distribution of case size and revenue based on historical data with mean vector $\vec{\mu}$ and covariance matrix Σ . This is done by applying a natural logarithm to both variables and fitting a bivariate Gaussian to these log-transformed variables, showcasing the high correlation between revenue and size. This relationship is intuitive, as larger cases typically generate higher revenues under hourly billing models.

It could be argued that the arrival process and the marks (revenue and size) are correlated: clients with more cases may generate larger and higher-revenue cases. Indeed, Appendix Table 4 shows a significant monotonic positive relationship between the number of cases per client and both revenue and size. To maintain simplicity, however, the marked point process is modeled such that the arrival process is independent of the marks, indicating that the number of arrivals does not influence the size or revenue of the cases. Consequently, this modeling assumption disregards client dynamics, where more cases could lead to more complex, higher-revenue cases.

The fitted bivariate Gaussian is independently sampled, exponentiated, and attached to each simulated arrival from the Negative Binomial model. With this action, the simulated case a now also accommodates a case revenue R_a and case size S_a .

4.1.3 Workload

The arrival process yields a monthly number of arrivals dependent on the client characteristics. However, in practice, demand should be considered in a daily resolution instead of a monthly. To bridge this gap, the monthly arrivals are disaggregated into daily values using an imputation heuristic based on historical weekday patterns. Specifically, the historical distribution of arrivals across weekdays is used to allocate proportions of the monthly total to each day in the corresponding month, thereby preserving the observed weekly structure.²

Let $\mathcal{A}_m$ denote the total number of arrivals in month m , and $\mathcal{D}_m$ the set of all days in that month. Furthermore, let

$$f_m = \{f_{m,d}\}_{d \in \mathcal{D}_m} \quad (3)$$

denote the historical arrival chance on each day d in month m . Then, the arrival dates for the $\mathcal{A}_m$ cases are sampled independently from the categorical distribution defined by f_m . Formally, for each arriving case $a \in \mathcal{A}_m$, the arrival day z_a is drawn as

$$z_a \sim \text{Categorical}(\mathcal{D}_m, \text{weights} = f_m), \quad a \in \mathcal{A}_m. \quad (4)$$

Each simulated arrival a is now represented by a day of arrival z_a , case size S_a , and revenue R_a . In practice, the case is worked on in a stochastic manner starting from the arrival date, depending on factors such as the client relationship, availability of the lawyer, and the nature of the case. For simplicity, the distribution of worked hours per case is assumed to follow the same distribution as the historical workload profile, represented by the function G_{d-z_a} , where $d-z_a \geq 0$ represents the number of days since the start date of the case. It will be shown that G_{d-z_a} approximately follows an inverse square root distribution.

Let $W_{a,m,d}$ denote the workload in hours on case a on day $d \geq z_a$ in month m . This is computed as

$$W_{a,m,d} = G_{d-z_a} \cdot S_a, \quad (5)$$

²A simple averaging of arrivals across years for each month would incorrectly inflate arrivals on weekend days, which typically have few or no arrivals.

where G_{d-z_a} represents the fraction of the case size S_a allocated to day $d - z_a$ since arrival. By construction, these fractions satisfy $\sum_{d \geq z_a} G_{d-z_a} = 1$, ensuring that on the entire case size is worked over all the days since z_a : $\sum_{d \geq z_a} W_{a,m,d} = S_a$.

Similarly, the revenue of each case can be disaggregated across the days it is worked on. The daily revenue for case a on day $d \geq z_a$ in month m is

$$R_{a,m,d} = G_{d-z_a} \cdot R_a. \quad (6)$$

The overall daily workload profile is obtained by summing the individual case workloads:

$$W_{m,d} = \sum_{a \in \mathcal{A}_m} W_{a,m,d}. \quad (7)$$

4.2 CAPACITY SIMULATION

This section outlines the framework used to approximate the historical capacity and to simulate the future capacity of the law firm.

4.2.1 Historical capacity

To quantify historical capacity, preliminary data cleaning is required. Among the 527 employees, only about half (203) have a recorded contractual workload, i.e., full-time equivalent (FTE). Informed imputation is constructed to approximate the FTE of employees with unknown FTE.

This imputation relies on two components: (1) the total number of worked hours of employee i in month m , denoted by $\sum_{d \in \mathcal{D}_m} W_{i,m,d}$, and (2) a random draw $O^{(\text{sample})}$ from the fitted empirical distribution of monthly overtime. The latter accounts for realistic overtime behavior for the employees with an unknown FTE (and therefore no known overtime). The empirical distribution of overtime is estimated from employees with known FTE, taking into account contractual FTE changes within a month. Specifically, an effective FTE is computed via a weighted average over working days before and after any contractual adjustment:

$$\overline{\text{FTE}}_{i,m}^{(\text{known})} = \frac{1}{|\mathcal{Q}_m|} \sum_{d \in \mathcal{Q}_m} \text{FTE}_{i,m,d}^{(\text{known})}, \quad (8)$$

where $\mathcal{Q}_m$ is the set of working days in month m . The overtime of employee i in month m is given by:

$$O_{i,m} = \sum_{d \in \mathcal{D}_m} W_{i,m,d} - H \cdot |\mathcal{Q}_m| \cdot \overline{\text{FTE}}_{i,m}^{(\text{known})}, \quad (9)$$

where $H = 8$ denotes the standard number of working hours per full-time day. Note that the overtime could also be negative, which indicates that the employee worked less than their contractual workload. The overtime formulation aggregates worked hours across all days in the month (including weekends) and remains robust to the allocation of worked hours in a month.³

For employees with unknown FTE, a synthetic FTE value is generated each month by combining the observed worked hours and the sampled overtime:

$$\text{FTE}_{i,m}^{(\text{imputed})} = \max\left(0, \frac{\sum_{d \in \mathcal{D}_m} W_{i,m,d} - O^{(\text{sample})}}{H \cdot |\mathcal{Q}_m|}\right), \quad (10)$$

The clipping at zero ensures that negative FTE values are prevented when the sampled overtime exceeds observed hours. This imputation is recomputed each month to reflect variability in work intensity.

The historical monthly capacity (in hours) is then approximated by summing over each employee in the set of employees $\mathcal{J}$ and multiplying by the standard number of working hours in a day H :

$$C_m^{(\text{hist})} = H \cdot \sum_{i \in \mathcal{J}} \left(\overline{\text{FTE}}_{i,m}^{(\text{known})} + \text{FTE}_{i,m}^{(\text{imputed})} \right), \quad (11)$$

where $\overline{\text{FTE}}_{i,m}^{(\text{known})}$ and $\text{FTE}_{i,m}^{(\text{imputed})}$ represent the observed and imputed FTEs, respectively.

³Theoretically, the employee with 1 FTE could work 24 hours each day in the first week of the month and then take leave for the rest of the month, resulting in excessive daily overtime but zero monthly overtime.

Not the entire capacity can be allocated to case work: employees get sick, go on vacation and undertake non-billable work such as meetings, performance reviews, and coffee breaks. The capacity of employee i in month m can be decomposed as:

$$C_{i,m}^{(\text{hist})} = C_{i,m}^{(\text{hist, BILL})} + C_{i,m}^{(\text{hist, NB})} + C_{i,m}^{(\text{hist, VAC})} + C_{i,m}^{(\text{hist, SICK})}, \quad (12)$$

where the superscripts denote billable work (BILL), non-billable work (NB), vacation (VAC), and sickness (SICK). Aggregating over all employees yields the firm's total monthly capacity:

$$C_m^{(\text{hist})} = \sum_{i \in \mathcal{I}} C_{i,m}^{(\text{hist})} = C_m^{(\text{hist, BILL})} + C_m^{(\text{hist, NB})} + C_m^{(\text{hist, VAC})} + C_m^{(\text{hist, SICK})}. \quad (13)$$

4.2.2 Future capacity

The simulation of future capacity begins by taking the final month of the historical period as the baseline, denoted by m_0 . The corresponding set of employees and their observed ($\overline{\text{FTE}}_{i,m_0}^{(\text{known})}$) and imputed ($\text{FTE}_{i,m_0}^{(\text{imputed})}$) full-time equivalents, serve as the initial information for the simulation.

To emulate contractual adjustments in workload, FTE changes are stochastically generated. Only a relatively small fraction of employees underwent a change in FTE, denoted by ν , consequently only for that fraction FTE changes are simulated. For each employee i , samples are drawn from a fitted empirical distribution of the number of days between historical FTE changes, denoted by $\hat{K}$.

For each simulated FTE change event, a new FTE value is drawn from the empirical distribution of observed FTEs. The draw is performed using weighted sampling, where the employee's previous FTE receives zero weight to ensure that an actual change occurs. Let $\mathcal{T} = \{t_1, \dots, t_T\}$ denote the set of unique observed FTE levels, and let $|\mathcal{T}_j|$ be the number of historical records with FTE equal to t_j . Also, let $\text{FTE}_i^{(\text{prev})}$ denote the current FTE of employee i . The new FTE is sampled according to

$$\mathbb{P}\left(\text{FTE}_i^{(\text{new})} = t_j \mid \text{FTE}_i^{(\text{prev})}\right) = \frac{|\mathcal{T}_j| \mathbb{1}_{\{t_j \neq \text{FTE}_i^{(\text{prev})}\}}}{\sum_{k=1}^T |\mathcal{T}_k| \mathbb{1}_{\{t_k \neq \text{FTE}_i^{(\text{prev})}\}}}. \quad (14)$$

Between two simulated change events, the FTE remains constant. This approach yields a simple yet plausible representation of contractual adjustments.

The simulated capacity of the firm for $m > m_0$ is defined as:

$$C_{m,d}^{(\text{future})} = H \cdot \sum_{i \in \mathcal{I}} \text{FTE}_{i,m,d}. \quad (15)$$

This formulation ensures that future capacity evolves through simulated contractual adjustments while remaining consistent with the empirically observed distribution of FTEs.

Based on the observed ratio of billable to total worked hours, the scalar $\phi^{(\text{BILL})}$ is computed:

$$\phi^{(\text{BILL})} = \frac{C_m^{(\text{hist, BILL})}}{C_m^{(\text{hist})}} \quad (16)$$

The simulated daily billable capacity in month m on day d is computed by $C_{m,d}^{(\text{future, BILL})} = \phi^{(\text{BILL})} \cdot C_{m,d}^{(\text{future})}$. The sequence of the billable capacity boundary is referred to as $C^{(\text{future, BILL})}$. This quantity represents a boundary of the maximum number of billable hours that can be allocated on a given day.

The pseudocode for the capacity simulation is presented in Algorithm 1.

Algorithm 1 Simulation of future capacity

```

1: Initialize set  $C^{(\text{future}, \text{BILL})} \leftarrow \emptyset$   $\triangleright C^{(\text{future}, \text{BILL})}$  stores daily billable capacity  $C_{m,d}^{(\text{future}, \text{BILL})}$ 
2: for each employee  $i \in \mathcal{I}$  do
3:   Initialize  $\text{FTE}_{i,m_0} \leftarrow \overline{\text{FTE}}_{i,m_0}^{(\text{known})}$  or  $\text{FTE}_{i,m_0}^{(\text{imputed})}$  if unknown
4:   if  $\text{Random}(0, 1) < \nu$  then
5:     Sample the number of FTE change events from the empirical distribution  $\hat{K}$ 
6:     for each FTE change event do
7:       Sample  $\text{FTE}_i^{(\text{new})}$  according to Equation (14)
8:     end for
9:     Forward-fill all simulated  $\text{FTE}_i^{(\text{new})}$  values over all days
10:  end if
11: end for
12: for each month  $m \in \mathcal{M}$  with  $m > m_0$  do
13:   for each working day  $d \in \mathcal{D}_m$  do
14:     if  $d \in \mathcal{Q}_m$  then
15:       Compute total daily capacity:  $C_{m,d}^{(\text{future})} \leftarrow H \cdot \sum_{i \in \mathcal{I}} \text{FTE}_{i,m,d}$ 
16:     else
17:        $C_{m,d}^{(\text{future})} \leftarrow 0$ 
18:     end if
19:     Compute billable capacity:  $C_{m,d}^{(\text{future}, \text{BILL})} \leftarrow \phi^{(\text{BILL})} \cdot C_{m,d}^{(\text{future})}$ 
20:      $C^{(\text{future}, \text{BILL})} \leftarrow C^{(\text{future}, \text{BILL})} \cup \{C_{m,d}^{(\text{future}, \text{BILL})}\}$ 
21:   end for
22: end for
23: return  $C^{(\text{future}, \text{BILL})}$ 

```

In the following section, we design acceptance policies that take into account both the daily demand $W_{m,d}$ and the billable capacity constraint $C_{m,d}^{(\text{future}, \text{BILL})}$.

4.3 ACCEPTANCE POLICIES

In the demand forecasting framework, the workload demand is referred to as $W_{m,d}$, and the available billable capacity is simulated to produce a boundary $C^{(\text{future}, \text{BILL})}$. When demand consistently exceeds this capacity, employees face sustained overtime which may result in elevated stress levels and discontent. This suggests that not every arriving cases can or should be accepted. Constructing a suitable acceptance policy is therefore imperative to systematically judge an incoming case. The set of admitted cases is denoted by $\mathcal{A}$.

4.3.1 Policies

This study evaluates four acceptance policies. Policy 0 serves as a theoretical upperbound in which all arrivals are indiscriminately accepted. Policy I enforces a hard daily capacity boundary. Policy II relaxes this assumption by applying a more realistic hard monthly capacity boundary. Finally, Policy III introduces a soft monthly capacity boundary, controlled by an overtime allowance slack variable θ .

Policy 0. Accept all (baseline). Under this policy, all arrivals are accepted. This policy serves exclusively as a baseline for comparison and evaluation of the more restrictive policies below. The results will indicate that this policy is unattainable in practice.

$$\text{Accept case } \tilde{a}. \quad (17)$$

Policy I. Hard daily capacity policy. Cases are accepted as long as the daily billable capacity is not exceeded:

$$\text{Accept case } \tilde{a} \text{ if } \sum_{a' \in \mathcal{A} \cup \{\tilde{a}\}} W_{a',m,d} \leq C_{m,d}^{(\text{future}, \text{BILL})} \quad \forall m, d. \quad (18)$$

From the dynamics of the policy, it is expected that the workload trajectory resembles a sawtooth pattern. The workload should contain intermittent peaks at the daily capacity and troughs just below it as a result of the predetermined hour-distribution mechanism (following Equation (5)).

Policy II. *Hard monthly capacity policy.* Cases are accepted as long as the total workload for the month does not exceed the monthly billable capacity:

$$\text{Accept case } \tilde{a} \text{ if } \sum_{d \in \mathcal{D}_m} \sum_{a' \in \overline{\mathcal{A}} \cup \{\tilde{a}\}} W_{a',m,d} \leq \sum_{d \in \mathcal{D}_m} C_{m,d}^{(\text{future, BILL})} \quad \forall m. \quad (19)$$

This policy permits daily overtime as it is naturally compensated by lower workloads elsewhere in the month. Since cases are accepted until the monthly boundary is reached, acceptations will tend to cluster at the beginning of the month. This is expected to produce a single sawtooth for each month where the peak is above the daily capacity boundary at the start of the month and afterwards the workload steadily decreases until the end of the month where it ends below the boundary.

Policy III. *Soft monthly capacity policy.* This policy introduces a tunable parameter $\theta \geq 0$ that governs the maximum permitted monthly overtime across the entire firm. A case $\tilde{a}$ is accepted when:

$$\text{Accept case } \tilde{a} \text{ if } \sum_{d \in \mathcal{D}_m} \sum_{a' \in \overline{\mathcal{A}}(\theta) \cup \{\tilde{a}\}} W_{a',m,d} \leq \sum_{d \in \mathcal{D}_m} C_{m,d}^{(\text{future, BILL})} + \theta \quad \forall m. \quad (20)$$

Setting $\theta \gg 0$ recovers Policy 0, while $\theta = 0$ reduces Policy III to Policy II.

Whenever a case is accepted and the monthly workload exceeds capacity, the overtime hours and the corresponding overtime revenue is tracked. The monthly overtime for case $\tilde{a}$ in month m is:

$$O_{\tilde{a},m} = \max\left(0, \sum_{d \in \mathcal{D}_m} \sum_{a' \in \overline{\mathcal{A}}(\theta) \cup \{\tilde{a}\}} W_{a',m,d} - \sum_{d \in \mathcal{D}_m} C_{m,d}^{(\text{future, BILL})}\right). \quad (21)$$

The revenue generated by overtime in month m by case $\tilde{a}$ is then:

$$R_{\tilde{a},m}^{(\text{OT})} = \frac{O_{\tilde{a},m}}{\sum_{d \in \mathcal{D}_m} W_{\tilde{a},m,d}} \cdot \sum_{d \in \mathcal{D}_m} R_{\tilde{a},m,d}, \quad (22)$$

with the daily revenue is scaled by the fraction of case $\tilde{a}$'s workload that occurs during overtime. After the simulation is finished, the set of all accepted cases $\overline{\mathcal{A}}(\theta)$ is known, in addition to the quantities $O_{a,m}$ and $R_{a,m}^{(\text{OT})}$ corresponding to each accepted case. The monthly overtime is:

$$O_m(\theta) = \sum_{a' \in \overline{\mathcal{A}}(\theta)} O_{a',m}. \quad (23)$$

This monthly overtime is financially compensated. The marginal discontent and fatigue associated with overtime is unlikely to grow linearly: an extra hour beyond the monthly capacity is far less burdensome than, for instance, the 50th or 100th extra hour. To capture this effect, monthly overtime is penalized as a convex exponential cost function, ensuring that monthly overtime becomes disproportionately expensive with more overtime:

$$B_m(\theta) = B_0(e^{\kappa \cdot O_m(\theta)} - 1), \quad (24)$$

where $B_0 > 0$, and $\kappa > 0$. The subtraction of 1 ensures zero cost when $O_m(\theta) = 0$.

Finally, the net revenue from overtime (also referred to as the net revenue) is summed across all months in the simulation:

$$\text{NR}^{(\text{OT})}(\theta) = \sum_{m > m_0} \left[\sum_{a \in \overline{\mathcal{A}}(\theta)} R_{a,m}^{(\text{OT})} - B_m(\theta) \right], \quad (25)$$

where $\overline{\mathcal{A}}(\theta)$ is the set of accepted cases under condition (20). The parameter θ is optimized to maximize the expected net revenue across simulation runs.⁴

⁴Optimization is necessary due to the convex cost function. A linear overtime penalty would reduce the problem to a trivial rule: accept all cases whose revenue per hour exceeds the hourly financial compensation, and reject the rest.

The pseudocode implementing these four policies is provided in Algorithm 2.

Algorithm 2 Simulation pseudocode

```

1: Select acceptance policy  $P \in \{\text{Policy 0, Policy I, Policy II, Policy III}\}$ 
2: Initialize accepted set  $\mathcal{A} \leftarrow \emptyset$ 
3: Initialize optimization set  $\mathcal{X} \leftarrow \emptyset$  if  $P = \text{Policy III}$ 
4: Simulate future billable capacity using Algorithm 1
5: Define daily capacity  $C_{m,d}^{(\text{future, BILL})}$ 
6: Compute monthly capacity  $C_m^{(\text{future, BILL})} = \sum_{d \in \mathcal{D}_m} C_{m,d}^{(\text{future, BILL})}$ 
7: procedure GENERATECASE( $a, m$ )
8:   Sample log-case size and log-revenue:  $(\ln S_a, \ln R_a) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 
9:   Transform to original scale:  $(S_a, R_a) = (\exp(\ln S_a), \exp(\ln R_a))$ .
10:  Sample arrival day:  $z_a \sim \text{Categorical}(\mathcal{D}_m, \text{weights} = f_m)$ .
11:  Allocate daily workload for case  $a$ :  $W_{a,m,d} = G_{d-z_a} \cdot S_a, \quad \forall d \geq z_a$ .
12:  Allocate daily revenue for case  $a$ :  $R_{a,m,d} = G_{d-z_a} \cdot R_a, \quad \forall d \geq z_a$ .
13:  return  $(W_{a,m,d}, R_{a,m,d})$ .
14: end procedure
15: for each month  $m \in \mathcal{M}$  with  $m > m_0$  do
16:   for each client  $n \in \mathcal{N}$  do
17:    Sample number of arrivals  $A_{n,m} \sim \text{NB}(r_n, p_n)$ 
18:    for  $\tilde{a} = 1$  to  $A_{n,m}$  do
19:       $(W_{a,m,d}, R_{a,m,d}) \leftarrow \text{GENERATECASE}(\tilde{a}, m)$ 
20:      Daily workload  $W_{m,d} \leftarrow \sum_{a' \in \mathcal{A} \cup \tilde{a}} W_{a',m,d}$ 
21:      Compute monthly workload:  $W_m = \sum_{d \in \mathcal{D}_m} W_{m,d}$ .
22:      if  $P = \text{Policy 0}$  then
23:         $\mathcal{A} \leftarrow \mathcal{A} \cup \{\tilde{a}\}$ 
24:      else if  $P = \text{Policy I}$  then
25:        if  $W_{m,d} \leq C_{m,d}^{(\text{future, BILL})} \quad \forall d \in \mathcal{D}_m$  then
26:           $\mathcal{A} \leftarrow \mathcal{A} \cup \{\tilde{a}\}$ 
27:        end if
28:      else if  $P = \text{Policy II}$  then
29:        if  $W_m \leq C_m^{(\text{future, BILL})}$  then
30:           $\mathcal{A} \leftarrow \mathcal{A} \cup \{\tilde{a}\}$ 
31:        end if
32:      else if  $P = \text{Policy III}$  then
33:        if  $W_m \leq C_m^{(\text{future, BILL})} + \theta$  then
34:           $\mathcal{A} \leftarrow \mathcal{A} \cup \{\tilde{a}\}$ 
35:          if  $W_m > C_m^{(\text{future, BILL})}$  then
36:            Compute overtime amount:  $O_{\tilde{a},m} \leftarrow \min(W_{\tilde{a},m}, W_m - C_m^{(\text{future, BILL})})$ 
37:            Compute overtime revenue:  $R_{\tilde{a},m}^{(\text{OT})} \leftarrow \frac{O_{\tilde{a},m}}{W_{\tilde{a},m}} \cdot R_{\tilde{a},m}$ 
38:          else
39:             $O_{\tilde{a},m} \leftarrow 0$ 
40:             $R_{\tilde{a},m}^{(\text{OT})} \leftarrow 0$ 
41:          end if
42:           $\mathcal{X} \leftarrow \mathcal{X} \cup \{(\tilde{a}, O_{\tilde{a},m}, R_{\tilde{a},m}^{(\text{OT})})\}$ 
43:        end if
44:      end if
45:    end for
46:  end for
47: end for
48: return accepted set  $\mathcal{A}$ 
49: return optimization set  $\mathcal{X}$  if  $P = \text{Policy III}$ 

```

The parameter θ from Policy III which maximizes the expected net revenue is found with stochastic approximation (SA). A separate section is dedicated to the mathematical modeling of the SA method, as it is

not trivial.

4.3.2 Stochastic approximation

The net revenue objective $\text{NR}^{(\text{OT})}(\theta)$ is a stochastic quantity, as multiple components of the demand forecasting and capacity simulation rely on random draws from underlying probability distributions, resulting in a complex system. Consequently, $\text{NR}^{(\text{OT})}(\theta)$ admits no closed-form expression and can only be evaluated via simulation. As a result, classical deterministic optimization techniques are not directly applicable.

The parameter θ represents the total monthly allowed overtime that the employees may take. To determine the optimal parameter θ^* that maximizes the expected net revenue across simulations, we employ stochastic approximation. Stochastic approximation methods are specifically designed for optimization problems in which the objective function is observed with noise and its analytical form is unknown. Instead of requiring exact objective evaluations or gradients, SA iteratively updates the parameter estimates using noisy observations obtained from repeated simulations. Each evaluation of $\text{NR}^{(\text{OT})}(\theta)$ requires running the full demand forecasting and capacity simulation, which are computationally heavy. A key reason for selecting SA in this setting is its efficiency under such a limited computational budget. Appendix Section 10.9 reports the computational burden associated with obtaining the results for the different policies.

Mathematical model. The stochastic approximation framework follows the structure and notation of Vázquez-Abad and Heidergott [19]. We consider a maximization problem, where the objective is the expected net revenue and the problem is constrained to a 1-dimensional hypercube Θ :

$$\max_{\theta \in \Theta} \mathbb{E}[\text{NR}^{(\text{OT})}(\theta)], \quad \Theta = [\theta_{\min}, \theta_{\max}], \quad (26)$$

A stochastic approximation algorithm is defined by the recursion projected onto the hypercube:

$$\theta_{n+1} = \Pi_{\Theta}(\theta_n + \varepsilon_n Y_n), \quad n \geq 0, \text{ and } \theta_0 \in \Theta \quad (27)$$

where θ_0 is the starting value at iteration 0. In addition, ε_n denotes the step size at iteration n , and Y_n is an observable noisy estimator of the target vector field $G(\theta_n)$, which corresponds to the gradient of the expected objective. The projection $\Pi_{\Theta}(x)$ ensures that the updated value θ_n remains within the interval $[\theta_{\min}, \theta_{\max}]$.

The feedback depends on an underlying stochastic process that captures all randomness affecting the simulation outcome, denoted by ξ_n . The feedback is written as:

$$Y_n = \phi(\xi_n(\theta_n), \theta_n). \quad (28)$$

The random variables ξ_n and ξ'_n represent independent realizations of the simulation randomness, obtained from separate simulation runs at the same θ_n . In this study, Y_n is constructed using a symmetric finite-difference estimator (also known as the Kiefer-Wolfowitz approach [9]):

$$Y_n = \frac{\text{NR}^{(\text{OT})}(\xi_n, \theta_n + c_n) - \text{NR}^{(\text{OT})}(\xi'_n, \theta_n - c_n)}{2c_n}, \quad (29)$$

where c_n is a decreasing perturbation sequence controlling the finite-difference step size. This estimator provides a noisy approximation of the gradient direction, enabling iterative ascent towards the optimal parameter θ^* . Intuitively, a higher net revenue for the " $+c_n$ "-term compared to the " $-c_n$ "-term suggests that the objective function increases with θ_n , motivating the "+" in front of $\varepsilon_n Y_n$ in Equation (27). Furthermore, with this choice of estimator Y_n , one can see that the perturbation $\theta_n - c_n$ becomes negative if $\theta_n < c_n$. As the overtime cost is only defined for positive values of allowed overtime, the hypercube has domain $\Theta = [c_n, \infty)$.

Well-posedness. For the problem to be well-posed, the objective $\mathbb{E}[\text{NR}^{(\text{OT})}(\theta)]$ must be differentiable with respect to θ . For any fixed realization of the simulation, $\text{NR}^{(\text{OT})}(\theta)$ is generally not differentiable, since the net revenue depends on the indicator in Equation (20). Small perturbations of θ may change the acceptance decision and thereby generate kinks in the realized objective.

Such kinks arise only when the acceptance rule switches between accepting and rejecting an arrival. This occurs when $\text{workload} + \text{arrival workload} = \text{capacity} + \theta$, which is an informal representation of Equation (20). At such realizations, a small change in θ may alter whether the arrival is accepted or rejected. Away from these switching points, perturbations in θ do not affect the acceptance decision and therefore do not change

the realized net revenue. Discontinuities thus occur exclusively at realizations where the acceptance decision changes between $\theta - c_n$ and $\theta + c_n$.

Under the assumption that the joint distribution of workload and arrival workload are continuous distributions, the probability that the above equality holds exactly is zero for any fixed θ . Hence, the set of realizations at which $\text{NR}^{(\text{OT})}(\theta)$ is discontinuous has probability zero. Although the objective is not differentiable for individual realizations, these discontinuities disappear after taking expectations.

So, under the conditions that ensure the objective is locally bounded, the derivative of the expected net revenue $\mathbb{E}[\text{NR}^{(\text{OT})}(\theta)]$ exists and is finite. Consequently, the expected objective is differentiable in θ , and the finite-difference stochastic approximation scheme is justified, since it approximates a well-defined gradient of the smooth expected objective.

Convergence. To establish convergence of the projected stochastic approximation algorithm, regularity conditions following Vázquez-Abad and Heidergott [19] are imposed. First, the hypercube Θ is made convex and compact. This is achieved by introducing an upper bound M and restricting the parameter space to $\Theta = [c_n, M]$. The upper bound M is selected to be sufficiently large.⁵

The step-size sequence $\{\varepsilon_n\}$ is chosen as $\varepsilon_n = \frac{\varepsilon_0}{n+1}$ with ε_0 are constant, which satisfies the condition $\sum_{n=0}^{\infty} \varepsilon_n = \infty$. Furthermore, the condition $\sum_{n=0}^{\infty} \frac{\varepsilon_n^2}{c_n^2} < \infty$ (see Assumption 4.2 in Bhatnagar, Prasad, and Prashanth [2]) must be satisfied. With assumption that c_n is decreasing in n and of the form $\frac{c_0}{(n+1)^\phi}$:

$$\sum_{n=0}^{\infty} \frac{\varepsilon_n^2}{c_n^2} = \sum_{n=0}^{\infty} \left(\frac{\varepsilon_0^2}{n+1} \right)^2 \left(\frac{n+1}{c_0} \right)^{2\phi} = \frac{\varepsilon_0^2}{c_0^{2\phi}} \sum_{n=0}^{\infty} \frac{1}{(n+1)^{2(1-\phi)}} < \infty. \quad (30)$$

The series converges if $2(1-\phi) > 1$, so $\phi < 1/2$. We additionally assume that the target vector field G is continuous on Θ , and that the trajectories of the projected ODE

$$\frac{dx(t)}{dt} = (\Pi_{\Theta} G)(x(t)), \quad x(0) \in \Theta, \quad (31)$$

have a set of asymptotically stable points $S \neq \emptyset$.

Under these assumptions, Theorem 4.1 of Vázquez-Abad and Heidergott [19] states that the shifted interpolated process associated with $\{\theta_n\}$ converges almost surely, as $n \rightarrow \infty$, to a solution of the projected ordinary differential equation (31). In particular, if θ^* is the unique asymptotically stable equilibrium of the projected ODE, then $\theta_n \xrightarrow{\text{a.s.}} \theta^*$.

Common random numbers. In the SA, variance reduction techniques are critical to obtaining stable parameter estimates. One widely used method is the application of common random numbers (CRN). In this study, CRN are employed to reduce the stochastic variability introduced by the capacity and demand simulation components of the model. Specifically, the simulation of FTE changes over time, the FTE associated with new hires, the number of arrivals of big clients and the associated size and revenue are all stochastic variables. If each iteration of the SA algorithm were to use independently generated randomness for these components, the resulting gradient estimates Y_n would blow up, especially due to the large values of $\text{NR}^{(\text{OT})}(\xi_n, \theta_n)$.

By using common random numbers, the same sequence of random draws is applied to the capacity simulation and demand forecasting for $\text{NR}^{(\text{OT})}(\xi_n, \theta_n)$ and $\text{NR}^{(\text{OT})}(\xi'_n, \theta_n)$ within a given iteration. This ensures that the realized trajectory is identical for different perturbations of the acceptance policy, isolating the effect of the policy itself. Consequently, differences in net revenue $\text{NR}^{(\text{OT})}(\xi_n, \theta_n + c_n)$ and $\text{NR}^{(\text{OT})}(\xi'_n, \theta_n - c_n)$ can be attributed almost exclusively to the acceptance policy. This controlled variance allows the SA algorithm to produce more stable gradient estimates, improving convergence behavior.

Output analysis. The sequence $\{\theta_n\}$ exhibits serial dependence, since each θ_n depends on the previous parameter θ_{n-1} . After convergence, the post-convergence iterates are expected to fluctuate around the optimal value θ^* and follow an approximately symmetric distribution, which is frequently well-approximated by a normal distribution. However, naively constructing a confidence interval using the standard deviation of the marginal of $\{\theta_n\}$ under the assumption of independence would underestimate the true uncertainty, because serial correlation inflates the variance of the mean value.

⁵Based on initial findings, M is set to 1,000,000.

To address this issue, we employ the non-overlapping batch means (NOBM) scheme proposed in Steiger and Wilson [17]. In this approach, the post-convergence sequence $\{\theta_n\}_{n=N_c+1}^N$ is divided into k batches, each of size m , such that the total number of post-convergence observations satisfies $N_c = k \cdot m$. The batch means are then computed as

$$\bar{\theta}_i = \frac{1}{m} \sum_{j=(i-1)m+1}^{im} \theta_{N_c+j}, \quad i = 1, \dots, k. \quad (32)$$

To ensure that the batch means can be treated as approximately independent, we test them using a lag-1 correlation criterion; if the lag-1 autocorrelation of the batch means is less than 0.2, the batches pass the test. The batch size m is iteratively increased until the independence test passes, following the recommendations of Steiger and Wilson [17]. Once the batch means satisfy the independence criterion, a standard confidence interval for the mean of θ_n is constructed using the sample variance of the batch means.

4.4 RANDOMIZATION

The data used in this study originates from the day-to-day operations of a law firm and is therefore commercially sensitive. As a result, confidentiality must be ensured throughout the presentation of results. No reporting is performed at the individual client, case, or employee level beyond the anonymization already applied in Table 1. All other results are therefore presented exclusively in numerical or aggregated form.

Controlled disclosure of numerical data can be achieved in several ways. One approach is to create a synthetic data generator that preserves the internal structure and dependencies between variables. While methodologically sound, this approach falls outside the scope of this thesis. A second option is direct randomization of observed values, however, this would distort the structural relationships in the data. A third option is resampling from the estimated underlying distributions, but this would still reveal distributional properties of sensitive variables.

Therefore, a simple randomization strategy is adopted. For each figure, table, or standalone metric that may be considered sensitive, the reported values are multiplied by a randomly drawn scalar. The scalar is drawn from an unspecified distribution with unspecified parameters. This transformation obscures values and underlying distributions while not impacting the relative relationships, trends, and structural patterns in the data. As a result, the analysis remains qualitatively valid, while preventing disclosure of commercially sensitive information.

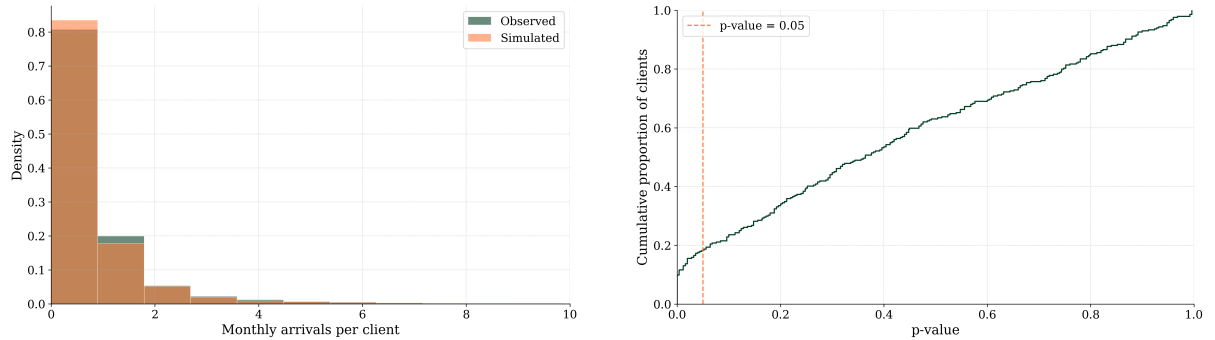
EXPLORATORY ANALYSIS

This section is dedicated to the intermediate results. It includes initial findings of the demand forecasting in Section 5.1, followed by a trend and seasonality analysis in Section 5.2 and concluding with the capacity planning Section 5.3. Based on the preliminary results, supplementary methods are introduced in Section 5.4 to ensure a consistent balance between simplicity and realism in the model.

5.1 DEMAND FORECASTING

5.1.1 Marked point process

The arrival model that we have selected is the Negative Binomial, which can account for overdispersion in the data. Figure 1 illustrates the results of the Negative Binomial simulation and the cumulative distribution function (CDF) of the Monte Carlo KS test p-values to substantiate this modeling choice. In Figure 1a, the observed monthly arrival distribution is shown alongside the simulated arrivals. Simulations were performed by fitting the Negative Binomial parameters for each client with more than 10 total arrivals (referred to as big clients) and generating monthly arrivals equal to the number of months in the observed data. The Negative Binomial distribution across all clients appropriately follows the observed monthly arrivals for low number of monthly arrivals, and the tails of the simulated and observed monthly arrivals are roughly similar. As the Negative Binomial parameters are fitted on each client individually, the Monte Carlo KS test is also performed individually. Each test yields a separate p-value and Figure 1b shows the CDF of p-values, where one can see that roughly 85% of the clients have a p-value larger than 0.05. Because the majority of clients do not reject that the underlying distribution is a Negative Binomial, for simplicity, the clients that reject it are also treated like they have an underlying Negative Binomial distribution.



(a) Observed and Negative Binomial simulated arrivals for each client. Only clients with more than 10 arrivals are shown and fitted.

(b) P-value CDF of Monte Carlo KS test, using 250 bootstrap samples.

Figure 1: Substantiation of the claim to model the arrival process as a Negative Binomial distribution.

The marks that will be attached to the arrival process are the revenue and the size of the case. Figure 2 shows the density plots and fitted curve of revenue (Figure 2a) and size (Figure 2b) for closed cases. Both the revenue and size distributions of closed cases shown in Figure 2 approximately follow a lognormal distribution. Normal Q-Q plots, shown in Appendix Figure 26, are made after applying a natural logarithm to both variables. One can see that both empirical distributions have heavier tails than a normal distribution and both Jacque-Bera tests for normality indicate to reject the hypothesis that the data is normally distributed.

However, the bulk of the data aligns reasonably well with the normal Q–Q line, indicating that a log transformation makes the central part of the distribution approximately normal. The deviations in the tails reflect extreme cases, which can not be modeled well due to their scarcity. Based on these insights, we will assume that the case revenue and size can be modeled as lognormal distributions without significant loss in generality.

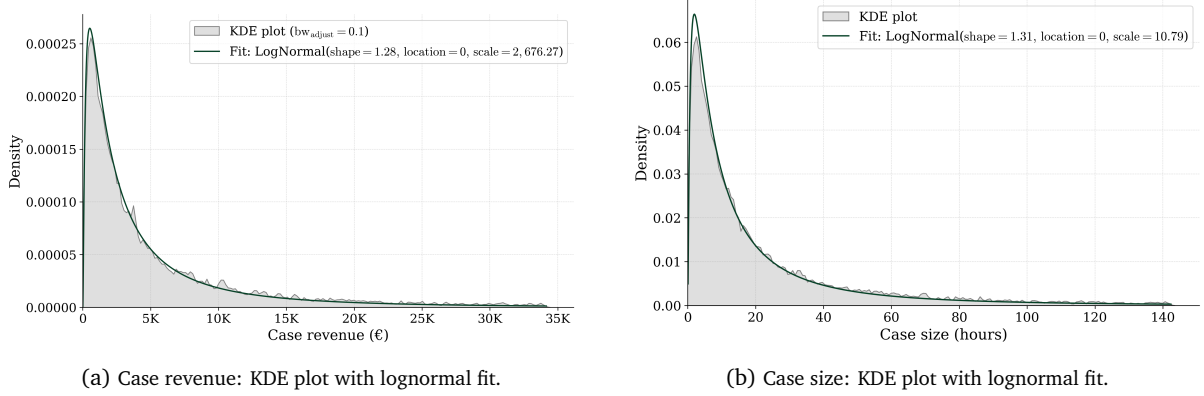


Figure 2: Lognormal fits for revenue and size of closed cases. A 95% quantile cut-off is applied for both plots for clarity.

Under the assumption that both the size and revenue are lognormally distributed, a joint distribution can be built that retains the observed dependence between the two variables. Figure 3 displays the observed combinations of exponentiated case revenue and size for all closed cases, alongside 10,000 samples generated from the fitted bivariate Gaussian.⁶

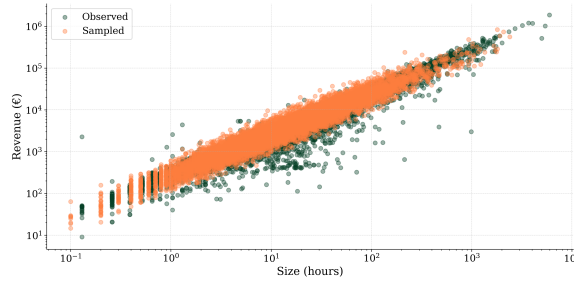


Figure 3: Revenue and size of closed cases, in addition to the sampled cases from the fitted joint distribution. Both axes are shown on a logarithmic scale for clarity.

5.1.2 Workload

The simulated monthly arrivals per client must be allocated to specific days within each month. To achieve this, we recall the historical distribution of case arrivals across weekdays. Figure 4a shows that cases rarely arrive during weekends. To reflect operational realism, any historical arrivals that occur on Saturdays or Sundays are moved to the following Monday, resulting in the adjusted distribution shown in Figure 4b. For each month, this adjusted weekday distribution is replicated across the month to convert the simulated monthly arrivals into daily arrival counts. An example of the resulting daily weights for January 2025 is provided in Appendix Figure 27.

⁶The sample values are also rounded to 1 and 2 decimals for size and revenue, respectively, to match the granularity of the empirical data. This quantization is reflected in the vertical 'stacking' for lower case sizes in Figure 3.

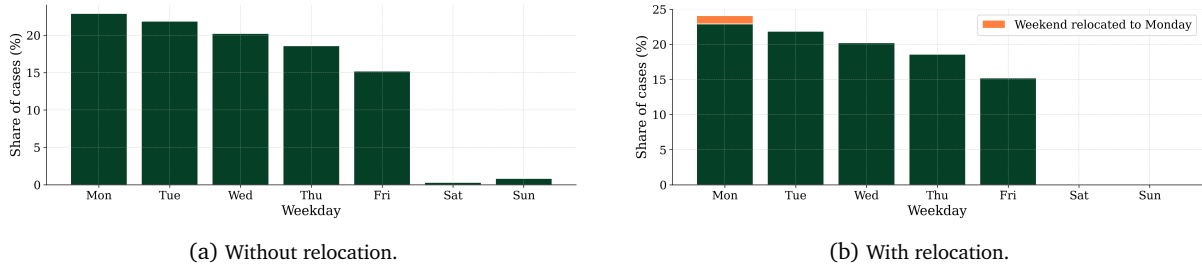


Figure 4: Share of case arrivals across weekdays , where between figure (a) and (b) the cases in the weekend are relocated to Monday.

A new case is generally not solved within the same day, due to the existing workload of the lawyers and the typical size of a case (~ 50 hours). Therefore, the size of the case is spread across multiple days. Figure 5 shows the average proportion of hours worked starting from the arrival case for closed cases. Based on the fit and the size of the arrival, an estimation can be made for the hours worked on the case each day. Appendix Figure 28 shows that for large values of days since case arrival the fit follows the empirical distributions well with a coefficient of determination R^2 of 0.96. There is a slight upwards drift which may be caused due to the minimization being biased by the high variance for small values of days since case arrival.

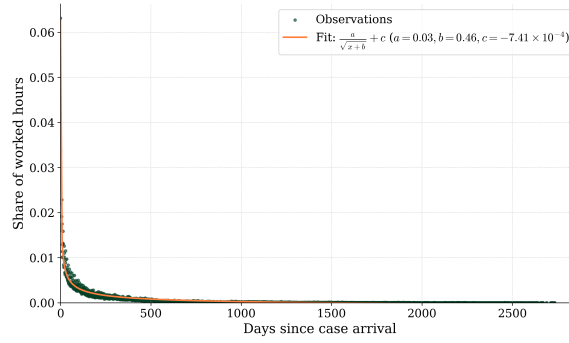


Figure 5: Average weighted hours distribution with fitted shifted inverse square root function. The weighted average is computed by adding the worked hours across all clients for each day from case arrival.

The duration from case start to case close is shown in Figure 6. Most cases last less than 50 days, while some case took more than 5 years to complete. On average, the case duration is roughly 200 days.

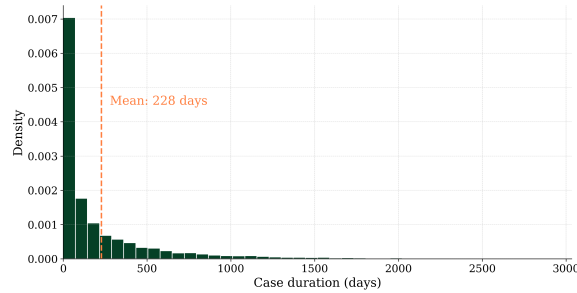


Figure 6: Duration of closed cases.

5.2 TREND AND SEASONALITY

Two clear upward trends are visible in the historical data: one in the revenue per case over time and one in the number of employees over time. The revenue per case, shown in Figure 7, has increased by approximately 60% since 2017, partly due to inflation adjustments and partly due to non-inflationary price increases.⁷ No

⁷The cumulative inflation from 2017 to 2025 in the Netherlands is 34.5% [3]. This indicates that about 25% is due to non-inflationary price increases.

structural seasonality is detected.

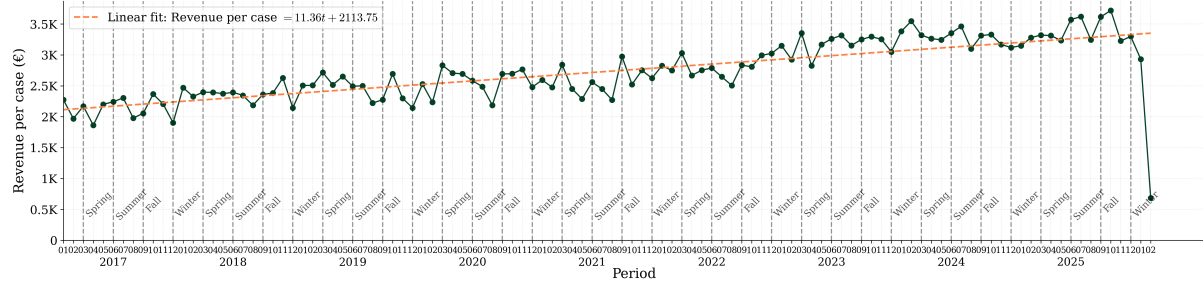


Figure 7: Time series of revenue per case. A linear trend is fitted using ordinary least squares, where t is the month.

The number of employees is represented by the number of individuals with billable hours in a given month. As shown in Figure 8, the employee count exhibits a modest positive trend, corresponding to roughly one additional employee per quarter. Again, no structural seasonality is detected.

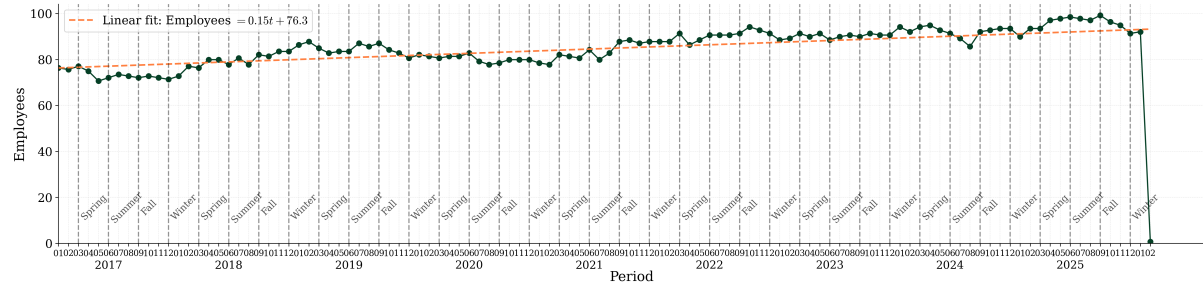


Figure 8: Time series of the number of employees. A linear trend is fitted using ordinary least squares, where t is the month.

5.3 CAPACITY PLANNING

The FTE is the key component in determining the capacity of the law firm. The FTE of more than half of the employees is unknown. These blanks are imputed using Equation (10), where the overtime sample $O^{(sample)}$ is drawn from Figure 9.

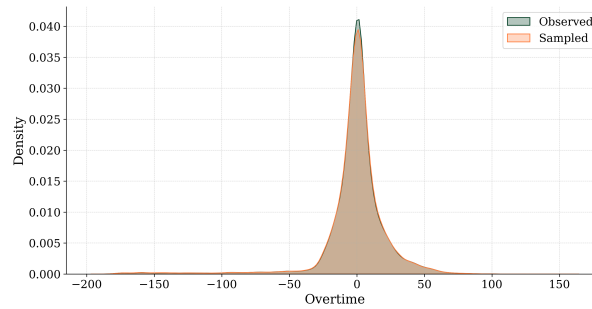


Figure 9: Overtime distribution KDE plot with sampled.

The imputed FTE, together with the known FTE, allows for approximation of the historical capacity. The imputed FTE values, combined with the observed contractual FTE values, enable an estimated reconstruction of the historical capacity. For forward-looking capacity simulations, expected contractual FTE adjustments are explicitly incorporated. Figure 10a displays the empirical distribution of FTE values, while Figure 10b shows the distribution of the number of days between contractual FTE changes. A Gamma distribution is fitted to the latter to support realistic sampling of future contractual adjustment intervals. However, the Gamma fit is imperfect, presumably due to the relatively small sample size (only 216 historical FTE changes). Since the

number of days between adjustments may take any value in practice, realistic sampling is prioritized over obtaining an optimal statistical fit. According to the data, around 16% of the total number of employees in the firm underwent contractual changes.

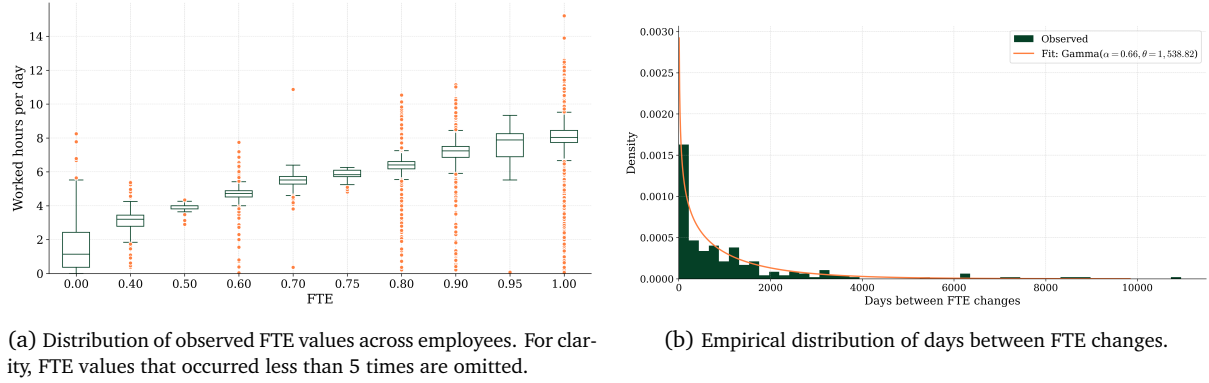


Figure 10: Empirical distributions related to employee workload characteristics.

In practice, only part of the capacity can be allocated to billable work. A simple approximation is made to determine the billable capacity, which entails computing the average proportion billable work relative to all worked hours. In Figure 11b it is shown that the proportion billable work relative to all worked hours remains relatively constant at 40%, with some dips due to the vacation periods shown in Figure 11a. Therefore, $\phi^{(BILL)}$ (from Equation (16)) is set to a fixed 40% of the capacity.

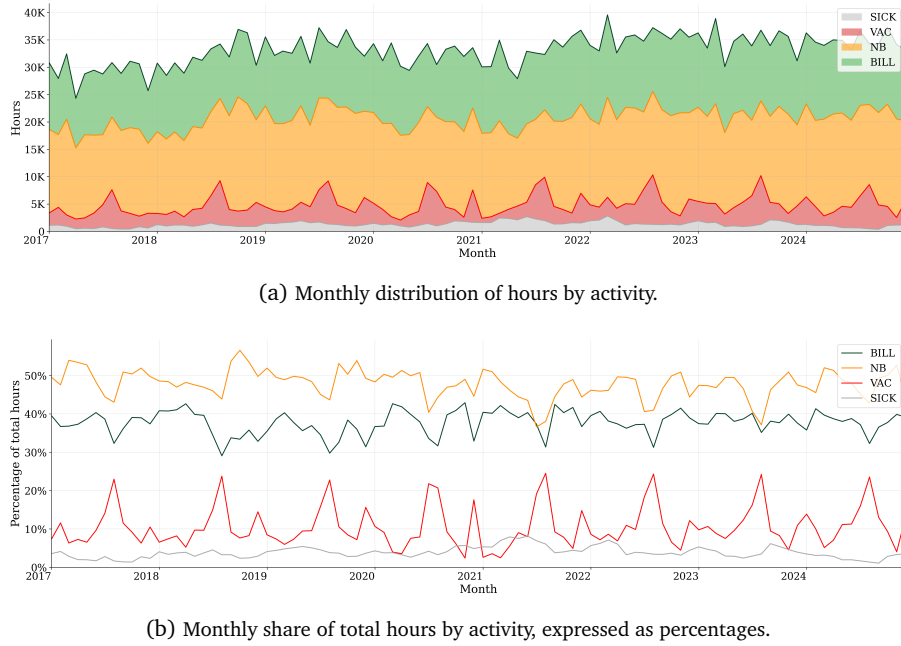


Figure 11: Time series of absolute (top) and relative (bottom) time type composition.

5.4 SUPPLEMENTARY METHODS

Based on the exploratory analysis, several additional steps are taken to accurately simulate the future.

1. *Number of arrivals.* Estimates are only calculated for clients with at least 10 cases. Clients with fewer than 10 cases, referred to as small clients, total about 70 clients; each, on average, generating one new case per month. To account for these small clients, 70 fixed arrivals are added each month alongside the stochastic arrivals of the big clients. Moreover, the case sizes are distributed across days according to Figure 5, whereas the cases in the historical data are worked in a stochastic manner. Thus, the

simulation exhibits lower variance compared to the historical figures. Peaks in simulated workload are further dampened when case sizes are spread across more days. This reduced variance prevents the total (accepted and unaccepted) workload from exceeding the billable capacity. In addition, historical arrival data reflect only accepted cases, as data on rejected cases is unavailable. The simulated monthly arrivals are increased by a substantial 100% to create an environment which allows for exploration of different acceptance policies.

2. *Closed and open cases.* The joint distribution is based on the closed cases. However, due to several general cases with an extreme number of written hours and revenue, the sampled revenue and size does not match with the historical averages. To correct for this mismatch, the sampled revenues and sizes are rescaled so that their average aligns with the historical average of both closed and open cases.
3. *Hours distribution truncation.* The hours distribution of Figure 5 inverse square root fit will be truncated. This reduces computational intensity and increases variance in simulated workload. The truncation is done at 200 days since case arrival, following the average case duration shown in Figure 6. Normalization is performed to maintain that 100% of the case is distributed.
4. *Revenue per case trend.* Due to the upwards trend in the historic revenue per case (as shown in Figure 7), the sampled revenue is scaled by the slope of the trend. As the bivariate Gaussian is based on the revenue per case of the full history (Figure 3), the sampled revenue per case is centered around the average historical revenue per case. Due to the linearity of the trendline, the sampled revenue is based on the average between the starting date and ending date of the history time series, namely January 2021. For each simulated month, the number of months between January 2021 and the simulated month is computed and multiplied by the slope of the trendline to account for the upwards trend in revenue per case.
5. *Number of employees trend.* On average, one employee joins each quarter (Figure 8). Therefore, each quarter in the simulation an employee is also added to the group of base employees.
6. *FTE changes.* The percentage of employees that initiated a contractual change in FTE is 16%. Therefore, for the simulation also the contractual changes are simulated for also only 16% of the employees.
7. *Overtime cost function.* The parameters B_0 and κ in Equation (24) are determined using reference points instead of estimation because the cost of overtime, especially the discontent and fatigue are not quantifiable as a economic cost. The simulation assumes a workforce of approximately 200 employees. Dividing total monthly overtime O_m by this workforce of 200 provides an estimate of overtime per employee.

As a first reference point, we consider a scenario with five hours of overtime per employee per month, corresponding to $O_m = 1,000$. At this level, the marginal cost of overtime is assumed to be approximately €50 EUR per employee, reflecting gross labor costs including salary, employee benefits, and fixed overhead. This implies a total monthly overtime cost of approximately €50,000 EUR.

As a second reference point, we consider ten hours of overtime per employee per month, corresponding to $O_m = 2,000$. At this level, overtime is assumed to have a marginal cost equal to twice the cost at one hour overtime per employee, so €100 EUR, totaling to a monthly overtime cost of approximately €200,000 EUR at $O_m = 2,000$.

The parameters B_0 and κ are selected to provide a close approximation to these reference points. Fixing $\kappa = 0.001$, we can find $B_0 = 30,000$ which provides a reasonable fit. Table 3 shows the overtime cost for different overtimes. Beyond 20 hours overtime per month per employee the cost has significantly outpaced the revenue generated from the overtime.

Est. MO / employee (hours)	Total MO (hours)	Total MO cost (€)	Est. MO cost / employee (€)	Est. MO cost / employee / hour (€ / hour)
1	200	6,642	33	33
2	400	14,755	74	37
5	1,000	51,548	258	52
10	2,000	191,672	958	96
15	3,000	572,567	2,863	191
20	4,000	1,607,945	8,040	402
25	5,000	4,422,395	22,112	884

Table 3: Cost of monthly overtime (MO) with $B_0 = 30,000$ and $\kappa = 0.001$ for several levels of overtime. The estimated overtime per employee is computed assuming a workforce of 200 employees.

RESULTS

This section presents the results of the simulation model, including both historical in Section 6.1 and simulated outcomes in Section 6.2. Historical workforce and arrival data are first described, followed by the results for the four acceptance policies.

6.1 HISTORICAL RESULTS

The base number of employees in the simulation is determined by the workforce in the month preceding the start of the simulation. For each employee, FTE values are sampled from the observed distribution, and contractual adjustments are modeled using a Gamma distribution. To account for the observed upward trend in the workforce over time (Figure 8), one additional employee is added at the start of each quarter.

Figure 12 shows the simulated and historical monthly capacity. The historical capacity display variation in capacity due to inflows and outflows of employees over time. The simulated capacity exhibits a slight upward trend, driven by the net inflow of employees, which becomes noticeable in the later years of the simulation. The baseline workforce remains fixed, and only quarterly additions are modeled, with no outflow. As a result, simulated capacity evolves more smoothly than the historical series. A daily resolution for 2025 is provided in Figure 29 in the Appendix, highlighting differences between weekdays and weekends.

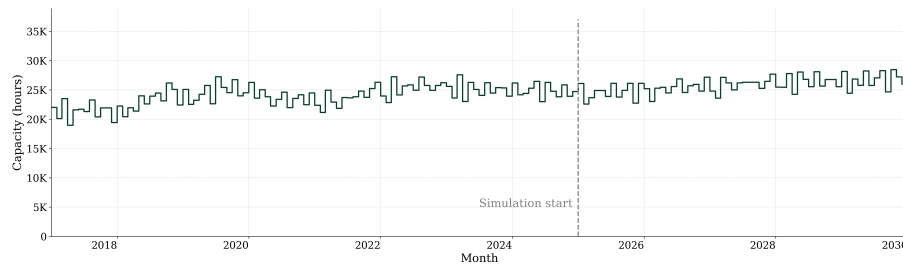


Figure 12: Capacity with a monthly resolution. Simulation starts on January 1, 2025.

Figure 13 illustrates that historical workload rarely exceeds the defined billable capacity. It should be noted that also the historical billable capacity is set at 40% of total capacity and does not reflect actual operational limits for historical periods. In practice, when workload exceeds expected billable capacity, employees can postpone non-billable work in favor of billable tasks.

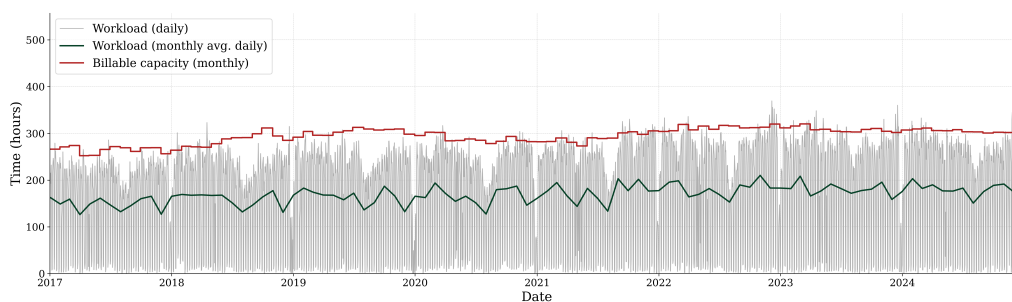


Figure 13: Daily historical workload with a monthly billable capacity boundary from 2017 until the end of 2024.

6.2 SIMULATION RESULTS

Simulation results are presented for each of the four acceptance policies. For Policies 0, I, and II, a realization of the simulation is shown. In addition, histograms of the key metrics (arrivals, workload, revenue) are provided to illustrate the mean and variance induced by the stochastic simulations. Appendix Section 10.8 shows the realization of the simulation together with the historical results for each policy to allow comparison between historical observations and simulated cases.

The coefficient of variation (CV) is used to fairly compare the variability between policies with different means. The CV of policy P and metric m is computed as:

$$CV_{P,m} = \frac{\sigma_{P,m}}{\mu_{P,m}} \quad (33)$$

6.2.1 Policy 0: Accept all

Under Policy 0, all arrivals are accepted. Figure 14 shows that workload frequently exceeds the defined capacity boundary. The simulated monthly arrivals exceed observed arrivals due to the 100% artificial increase applied to the scenario, as described in the supplementary methods Section 5.4. At the start of the simulation, a burn-in period of approximately one year is observed. This occurs because the hours distribution is spread over 200 days, resulting in lower workload in the initial months while arrival workloads accumulate. Evidently, the workload is unmanageable in practice, as the capacity is continuously exceeded, forcing employees to work persistent overtime.

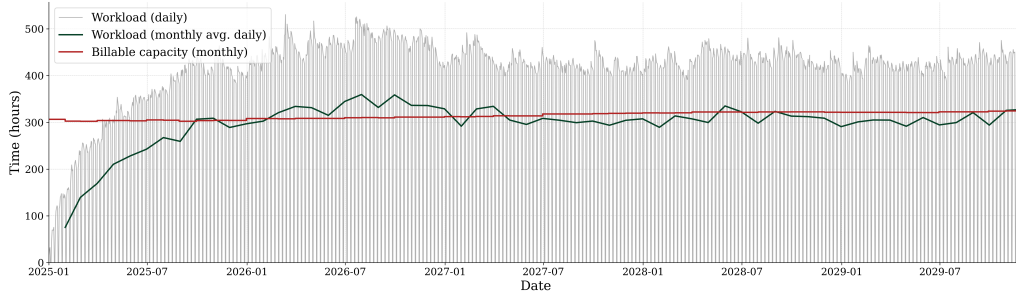


Figure 14: An individual simulation run with Policy 0 for a 5-year period. This simulation contains 21K accepted cases, resulting in 925K hours of workload and € 262M EUR in total revenue. See Appendix Figure 30 for comparison with historical data.

Figure 15 illustrate the mean and the variability of the total simulated arrivals, revenue, and workload induced by the stochastic nature of the simulation. Over a 5-year simulation horizon, on average, the simulation accepts 20.3K arrivals, netting €259M EUR and 0.91M hours of workload. The variance of about $\pm 5\%$ from the mean shows that the variability between runs is low. This is expected since the simulation horizon is relatively long with around 20K accepted arrivals, which reduces the variance in the outcome. The CV of the metrics are: 0.0141, 0.0146, 0.0145, for arrivals, case revenue and size, respectively.

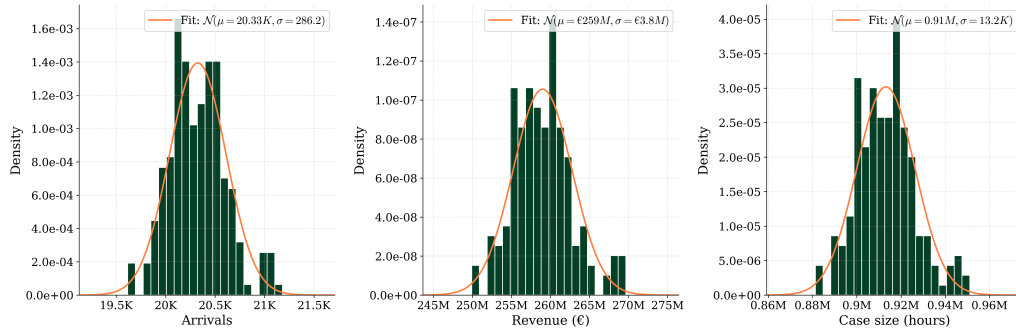


Figure 15: Distribution of arrivals, case revenue, and size under Policy 0 over the 5-year simulation period for 200 runs. A normal distribution is fitted on the data.

6.2.2 Policy I: Hard daily capacity boundary

Policy I enforces a strict daily limit on accepted cases; a new case $\tilde{a}$ is accepted only if adding its workload does not exceed the daily billable capacity (Equation (18)). This ensures that daily workload never exceeds capacity, strictly preventing daily overtime. However, some capacity may remain unused if arrivals are unevenly distributed across the month. As a result, the monthly billable capacity might not be fully utilized. Figure 16 shows one realization of the simulation under Policy I. The realization makes it clear that the monthly billable capacity is amply used; a consequent of the high arrival rate shown in Figure 14.

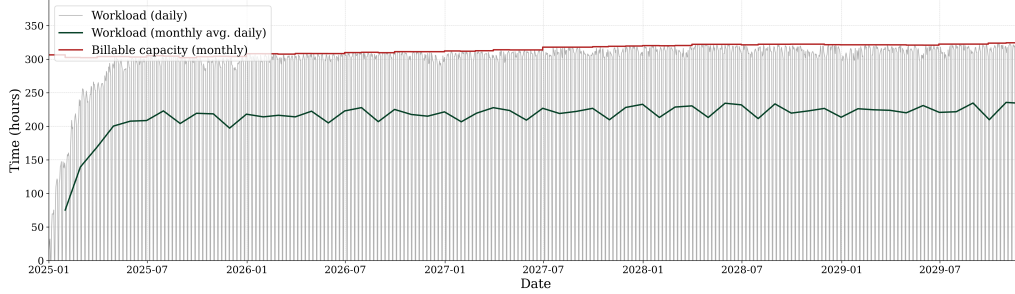


Figure 16: An individual simulation run with Policy I for a 5-year period. This simulation contains 19K arrivals, resulting in 691K hours of workload and € 202M EUR in revenue. See Appendix Figure 31 for comparison with the historical data.

Figure 17 presents the histograms of arrivals, revenue, and size. Policy I imposes a capacity restriction, whereas Policy 0 does not. Consequently, it is expected that all performance metrics under Policy I are weakly lower than under Policy 0. Indeed, the simulations yield on average 19.8K arrivals, €199.5M in revenue, and 0.68M hours of workload.

In addition to lower means, Policy I exhibits lower variance relative to the mean. The coefficients of variation (0.0125, 0.00975, and 0.00913 for arrivals, revenue, and workload, respectively) are all smaller than the corresponding values under Policy 0. This reduction in relative variability is consistent with the imposed capacity restriction. By limiting the number of accepted cases, the policy effectively restricts extreme realizations that would otherwise occur under the unrestricted regime. As a result, both the mean and variance of the outcomes are reduced.

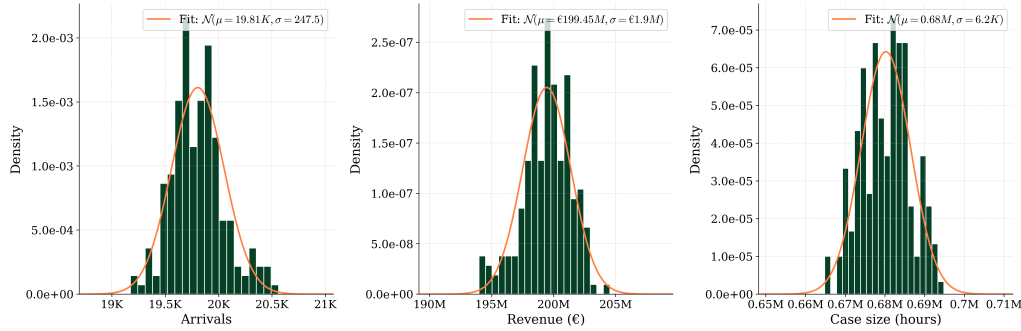


Figure 17: Distribution of arrivals, case revenue, and size under Policy I over the 5-year simulation period for 200 runs. A normal distribution is fitted on the data.

6.2.3 Policy II: Hard monthly capacity boundary

Policy II applies a monthly acceptance limit, accepting new cases only if monthly demand does not exceed the available capacity, as described in Equation (19). This approach ensures that monthly workload remains within capacity while allowing flexibility in daily distribution. Figure 18 illustrates a realization of the simulation under this policy. As expected, the monthly billable capacity is respected in a specific structure; at the start of the month the workload surpasses the billable capacity while at the end of the month the workload dips below the billable capacity.

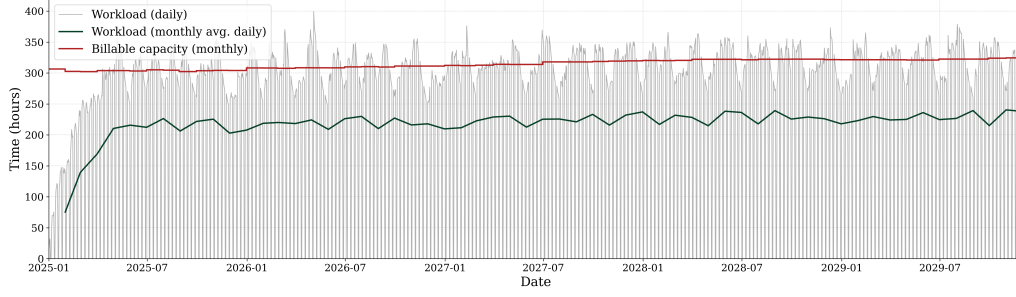


Figure 18: An individual simulation run with Policy II for a 5-year period. This simulation contains 16K arrivals, resulting in 699K hours of workload and €198M EUR in revenue. See Appendix Figure 32 for comparison with historical arrivals.

Figure 19 presents the histograms pertaining to the arrivals, revenue and size of individual simulations across a horizon of 5 years. On average, the simulations produce 16K arrivals, netting €196.7M EUR and 0.69M hours of workload. The number of arrivals and revenue is slightly smaller than Policy I, while the average workload is slightly larger. The CV of the metrics are: 0.0128, 0.0101, 0.00919, which are similar to Policy I.

The metrics of the policy with daily capacity limit (Policy I) and monthly capacity limit (Policy II) are similar, compared to the differences between Policy 0, because both Policy I and Policy II are constrained by the same total capacity over the simulation horizon. In addition, the arrival process and samples from the distributions are time-independent, so a monthly capacity constraint instead of a daily one does not have a significant effect on the outcome of the total number of arrivals, revenue and workload.

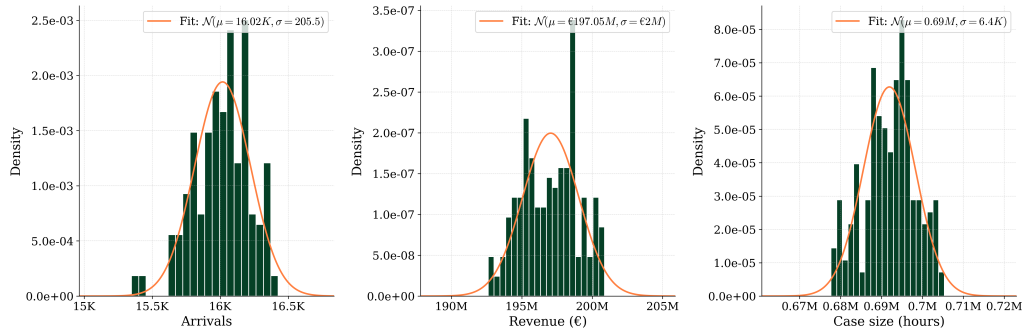


Figure 19: Distribution of arrivals, case revenue, and size under Policy II over the 5-year simulation period for 200 runs. A normal distribution is fitted on the data.

6.2.4 Policy III. Soft monthly capacity boundary

Policy III accepts a new case $\tilde{a}$ only if the resulting monthly workload does not exceed the billable capacity by more than a parameter θ , as formalized in Equation (20). Any workload exceeding the monthly billable capacity incurs an overtime cost according to Equation (24). The objective is to determine the optimal allowed monthly overtime θ^* that maximizes the expected net revenue from overtime. The optimization is done with a stochastic approximation heuristic, in particular a gradient descent where the gradient is estimated by the symmetric finite-difference estimator (see Equation (29)).

Parameters. The stochastic approximation is done for two configurations to validate the convergence. To ensure almost sure convergence of the stochastic approximation algorithm (see Section 4.3.2), we set the finite-difference decay parameter to $\phi \in \{\frac{1}{10}, \frac{1}{5}\}$. Empirical pilot runs indicated that smaller values of ϕ reduce the perturbation size c_n too slowly relative to the number of iterations executed, whereas larger values cause c_n to decay too rapidly, leading to unstable gradient estimates.

Careful selection of the initial values θ_0 , c_0 , and ε_0 is required. If θ_n becomes excessively large, both perturbed simulations $(\theta_n \pm c_n)$ may accept all incoming arrivals. While this would not be problematic under independent simulation runs, the use of common random numbers implies that both perturbations experience

identical arrivals, revenues, and workloads. In this case, the finite-difference estimator Y_n collapses to zero, preventing meaningful gradient updates.

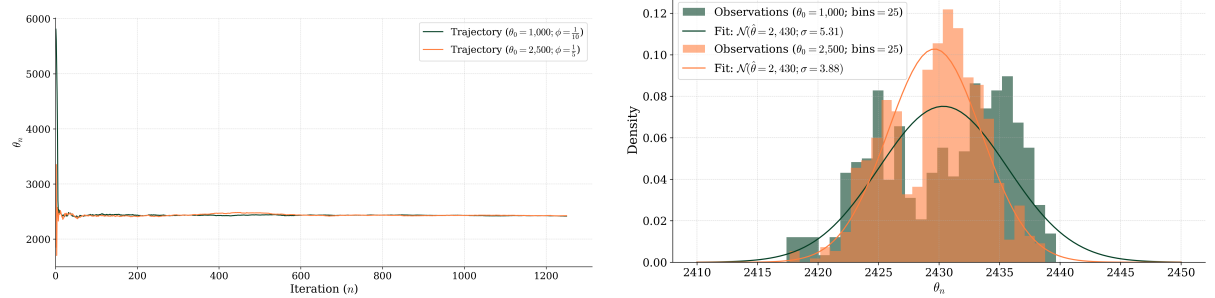
Based on the pilot runs, we select $c_0 = 100$, $\varepsilon_0 = 0.8$, which yield stable updates. The two configurations are evaluated with $(\theta_0, \phi) \in \{(1,000; \frac{1}{5}), (2,500; \frac{1}{10})\}$. Following Section 5.4, the cost function parameters are $B_0 = 30,000$ and $\kappa = 0.001$.

Convergence. We now evaluate the convergence behavior and stability properties of the stochastic approximation procedure. Figure 20a illustrates the evolution of the parameter sequence $\{\theta_n\}$ over the course of the stochastic approximation. The trajectories show rapid convergence toward a stable solution, corresponding to an estimate of the optimal allowed monthly overtime level $\hat{\theta}$. The sequences of both trajectories stabilize to around $\hat{\theta} = 2,430$ after approximately 600 iterations. After stabilization, θ_n fluctuates weakly around this value, with the algorithm continuously correcting decreasing stochastic deviations from $\hat{\theta}$.

To present the mean and variability of the converged solution, Figure 20b displays the empirical distribution of θ_n for $n \geq 600$. A normal distribution is fitted to these post-convergence observations, yielding an estimated mean $\hat{\theta} = 2,430$ and standard deviation of $\sigma = 5.31$ for $\theta_0 = 1,000$ and $\sigma = 3.88$ for $\theta_0 = 2,500$. The normal fit is used for descriptive purposes rather than as a formal asymptotic approximation. Recalling that both standard deviations are too optimistic due to the serial correlation of $\{\theta_n\}$, we determine the confidence interval based on non-overlapping batched means (NOBM). The trajectory with $\theta_0 = 2,500$ passes the independence test for $m = 14$, while the trajectory with $\theta_0 = 1,000$ does not pass the test for any m . Therefore, a two-sided 95% NOBM confidence for the optimal value of θ_n is only set up for the trajectory that passes the test of independence:

$$CI_{0.95}(\theta|\theta_0 = 2,500) = [2,427.94; 2,431.41]. \quad (34)$$

The narrow confidence interval indicates that the stochastic approximation procedure yields a statistically precise estimate of the optimal policy parameter, with $\hat{\theta} = 2,430$ lying in a tight neighborhood of the true optimum.



(a) Stochastic approximation sequence of the allowed monthly overtime parameter θ_n over 1,250 iterations for both configurations.

(b) Empirical distribution of θ_n after convergence ($n \geq 600$) for both configurations, with fitted normal density.

Figure 20: Convergence behavior of the stochastic approximation algorithm for Policy III under the parameter combinations $(\theta_0 = 1,000; \phi = \frac{1}{10})$ and $(\theta_0 = 2,500; \phi = \frac{1}{5})$, with $c_0 = 100$ and $\varepsilon_0 = 0.8$.

Appendix Figure 33 illustrates a representative simulation path for $\theta = 2,430$. Comparing the workload trajectory of Policy III (Appendix Figure 33b) with that of Policy II (Appendix Figure 32b), the overall temporal pattern remains structurally similar, reflecting the similar acceptance logic. However, Policy III exhibits systematically higher workload realizations due to the allowed monthly overtime.

To further examine the post-convergence regime, Figure 21 reports the individual trajectories. Overtime realizations remain bounded by the monthly overtime limit θ_n , and the average trajectory moves within the range $[2,000; 2,500]$ hours.

The corresponding revenue and cost dynamics are shown in Figure 21b. As the trajectories after convergence are shown, we can see that the vertical separation between overtime revenue and overtime cost is maximized, consistent with the maximization of the underlying objective function.

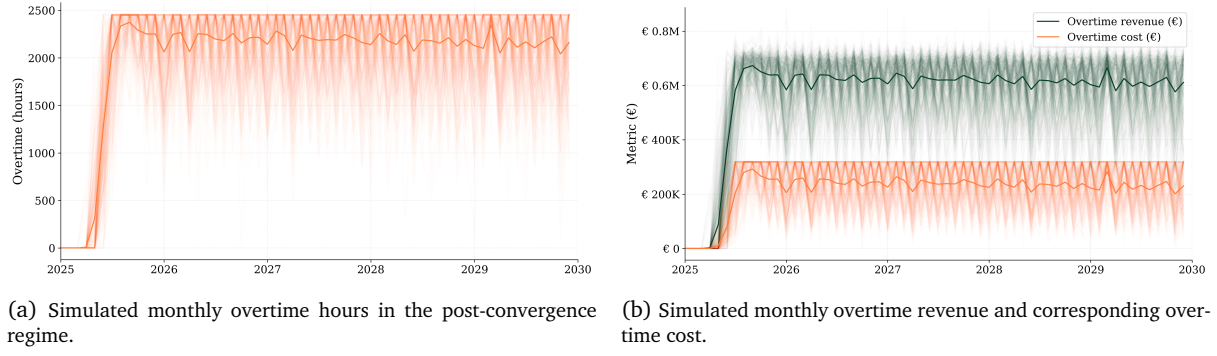


Figure 21: Trajectories in the post-convergence phase ($n \geq 600$, $\theta_0 = 2,500$). Faint lines denote individual simulation paths; solid lines represent the average trajectory across runs.

Taken together, these results confirm that the stochastic approximation procedure converges to a stable solution. The algorithm operates within a narrow neighborhood of the optimal policy without exhibiting structural drift or instability. Importantly, the post-convergence dynamics demonstrate that the optimization problem is well-behaved in the vicinity of $\hat{\theta}$. The revenue and cost trajectories present no evidence of explosive behavior. This indicates that the cost specification and stochastic simulation produce a stable maximization problem.

Local sensitivity analysis. To assess whether the converged value $\hat{\theta}$ corresponds to a local maximizer of the expected net revenue, we evaluate $\mathbb{E}[\text{NR}^{(\text{OT})}(\theta)]$ on a grid of θ values in a neighborhood of $\hat{\theta}$. Each evaluation is approximated using ten independent simulation replications to reduce the stochasticity of the outcome.

Figure 22 depicts the resulting estimated objective values over the interval $[1,000; 4,000]$. The curve exhibits a clear maximum in the vicinity of $\hat{\theta}$, with decreasing revenue for both smaller and larger values of θ . This confirms that the solution identified by the stochastic approximation algorithm corresponds to a locally optimal trade-off between increased case acceptance and higher overtime costs. The economic structure of the optimization problem suggests unimodality, as marginal net revenue is positive for small θ and negative for sufficiently large θ .

Furthermore, the smooth shape of the estimated objective function in this region provides empirical support for the differentiability of $\mathbb{E}[\text{NR}^{(\text{OT})}(\theta)]$ established in the well-posedness analysis. Although individual simulation realizations involve non-smooth acceptance rules, averaging over the stochastic environment yields a smooth expected objective. This effect is amplified with more realizations.

From a managerial perspective, the results imply that moderate stochastic variation in realized overtime does not undermine the optimality of the chosen policy parameter. The firm can therefore interpret $\hat{\theta}$ as a robust operational guideline.

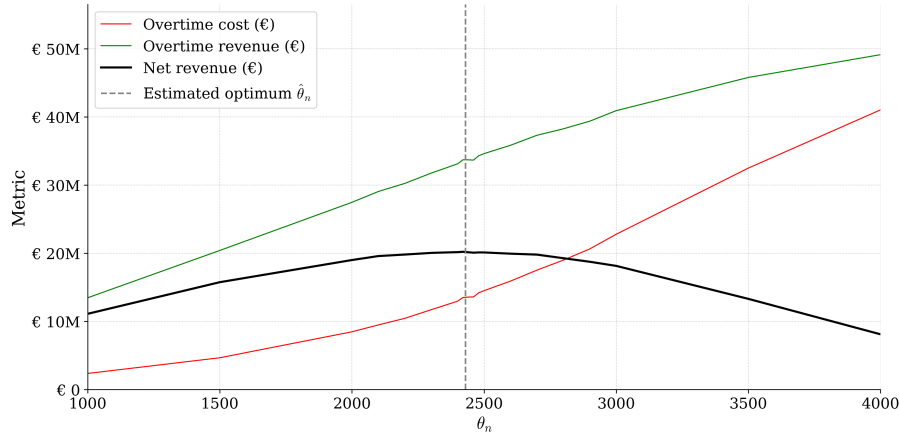


Figure 22: Local sensitivity of the expected net revenue with respect to the allowed overtime parameter θ . The objective is evaluated in a neighborhood of the converged value $\hat{\theta}$, using 10 replications for each θ_n . The parameter values considered are 1,000 and 4,000; values between 2,000 and 2,400 in increments of 100; between 2,400 and 2,500 in increments of 20; and between 2,500 and 3,000 in increments of 100. The maximum attained near $\hat{\theta}$ indicates local optimality.

Estimated optimum. Figure 23 illustrates the empirical distribution of performance outcomes under Policy III evaluated at the optimized parameter value $\theta = 2,430$. The histograms summarize the mean and variance of arrivals, case revenue, and workload over the 5-year simulation horizon across 200 independent replications. The distributions appear approximately symmetric and well-behaved, with no evidence of heavy tails or pronounced skewness, similar to the other policies. The fitted normal densities provide a reasonable descriptive approximation of the empirical variation. This suggests that, at the optimized allowed monthly overtime level, the model operates in a stable regime where stochastic fluctuations remain bounded and centered around their respective means.

On average, Policy III accepts approximately 18K arrivals over the simulation horizon, generating about €232M in total revenue and 0.8M hours of workload. Interestingly, the average number of accepted arrivals is less than Policy I and more than Policy II. However the Policy III corresponds to an increase in revenue and workload of around 10% - 15% relative to Policy I and Policy II. The increase arises because the marginal net revenue of allowing additional overtime remains positive for small values of θ , making additional accepted cases profitable up to the optimized threshold. The coefficients of variation (CV) for arrivals, revenue, and workload are 0.00913, 0.00746, and 0.00698, respectively. The coefficients of variation are lower than under Policy I and Policy II. This indicates that, relative to scale, outcomes are more predictable at the optimized overtime level. The optimization not only raises expected net revenue, but also improves stability in relative terms.

Overall, the histogram analysis indicates that the optimized parameter $\theta = 2,430$ not only maximizes expected net revenue locally, but also delivers predictable and statistically stable long-run performance.

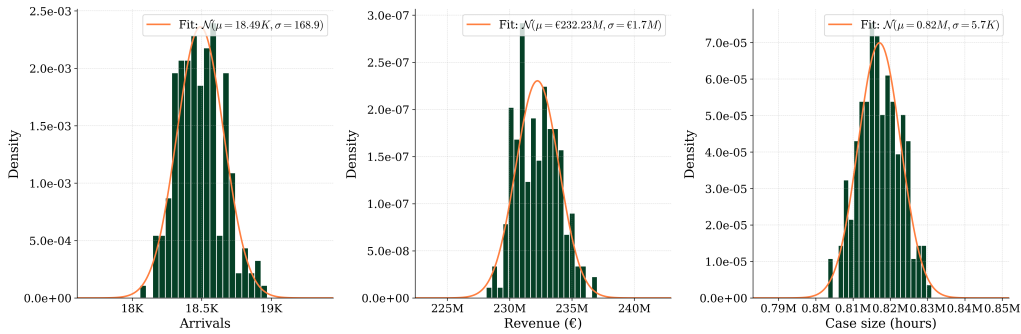


Figure 23: Distribution of arrivals, case revenue, and size under Policy III with $\theta = 2,430$ over the 5-year simulation period for 200 runs. A normal distribution is fitted on the data.

DISCUSSION

The discussion begins with an interpretation of the results in relation to the research questions in Section 7.1. Subsequently, the main limitations of the modeling approach are discussed in Section 7.2. Finally, broader challenges related to applying simulation-based decision support in legal services are addressed in Section 7.3.

7.1 INTERPRETATIONS

RQ1. How can stochastic case arrivals and their associated workloads be simulated in a realistic yet tractable manner?

The results demonstrate that stochastic case arrivals and their associated workloads can be simulated in a manner that balances empirical realism with tractability by separating the demand process into three components: (i) a client-level arrival process, (ii) stochastic case-level workload and revenue marks, and (iii) a conversion from monthly arrivals to a daily resolution. Modeling arrivals at the monthly level for each client using a Negative Binomial distribution captures the overdispersion and client-specific heterogeneity. Based on the empirical case workload distribution (Figure 5) and the average intra-month arrival pattern derived from the weekly pattern (Figure 4), monthly arrivals can be allocated to revenues, workloads, specific dates and subsequently distributed over time according to the predetermined workload profile.

RQ2. How do different case acceptance policies affect the trade-off between revenue generation and capacity utilization over a five-year simulation horizon?

The simulation results in Section 6.2 reveal systematic trade-offs between revenue generation and capacity utilization across the different acceptance policies. Policy 0, which accepts all arrivals, establishes an unconstrained benchmark. It yields the highest accepted volume, revenue, and workload, but does so at the expense of uncontrolled capacity overruns. As such, it represents a theoretical upper bound rather than a realistic policy.

Policy I imposes a strict daily capacity constraint. As a consequence, accepted cases, revenue, and workload are approximately 20% lower than under Policy 0. Policy I keeps the workload consistently below the daily capacity boundary. Policy II relaxes the constraint from daily to monthly capacity, permitting intra-month overtime. Because both Policy I and Policy II are ultimately governed by the same monthly billable capacity boundary, their accepted arrivals and total revenue are similar. The principal difference lies in the temporal structure of the workload. Under Policy II, acceptance decisions depend on the monthly capacity rather than daily limits. Consequently, workload exceeds the daily capacity boundary at the start of the month by accepting all cases. As cumulative workload approaches the monthly boundary, additional arrivals are rejected, causing workload to decline toward the end of the month and resulting in a workload below the capacity boundary.

Policy III extends Policy II by explicitly allowing monthly overtime up to some parameter θ . Consequently, it preserves the same intra-month temporal structure as Policy II but operates at a systematically higher workload level due to the expanded allowed capacity. This increases revenue relative to Policy II, but at an exponentially increasing cost of overtime.

Overall, the results highlight that loosening the temporal rigidity of capacity constraints, i.e. from daily to monthly, and from fixed to flexible capacity, shifts the temporal structure of the workload.

RQ3. What level of allowed monthly overtime maximizes the overtime revenue with an exponentially increasing monthly overtime cost, when simulating over a five-year horizon?

Allowing monthly overtime expands the feasible acceptance region beyond the fixed billable capacity, thereby increasing expected revenue. However, the exponential overtime cost function implies convex marginal costs, so that the marginal net revenue from additional overtime eventually becomes negative. Consequently, the expected net overtime revenue is convex in the allowed monthly overtime threshold θ , implying the existence of an interior optimum; both $\theta = 0$ (reflects Policy II) and $\theta \gg 0$ (reflects Policy O) are suboptimal.

The optimal threshold equates the expected marginal revenue from overtime to the expected marginal overtime cost. For small values of θ , the marginal benefit of relaxing the constraint is high because additional cases can be accepted at relatively low incremental cost. As θ increases, however, the exponential cost component amplifies the marginal cost of overtime. This interaction produces a unique level of allowed monthly overtime that maximizes the expected net revenue from overtime.

Under the specified cost parameters (Equation (??) with $B_0 = 30,000$ and $\kappa = 0.001$), the stochastic approximation procedure identifies an optimal threshold of $\hat{\theta} = 2,430$ hours. The local sensitivity analysis confirms that this value corresponds to a local maximizer of the expected net overtime revenue over the five-year simulation horizon. A comparison of the coefficients of variation across policies further shows that Policy III, evaluated at the optimized threshold, exhibits the lowest relative variability in arrivals, revenue, and workload. Hence, at $\hat{\theta}$, the policy not only maximizes expected net revenue locally, but also yields the most stable performance in relative terms among all considered policies.

7.2 LIMITATIONS

Several limitations of the proposed framework should be acknowledged. First, case workloads are assumed to be identically distributed over a fixed horizon of 200 working days according to the fitted distribution in Figure 5. This abstraction ignores substantial heterogeneity in case duration, intensity, and temporal clustering observed in legal practice. While this assumption improves tractability, it restricts the model's applicability to case-level analysis and has a large impact on the set of accepted arrivals.

Second, case size and revenue are simulated a priori at the time of arrival. In the legal sector, both quantities are unobservable at arrival. As a result, case acceptance decisions in the model rely on information that would not be available to decision-makers at the time of arrival. Therefore, the decision process of the acceptance policies are not replicable in a real law firm.

Third, billable capacity is modeled in a generalized manner and omits daily scheduling constraints, employee specialization, and dynamics in the non-billable activities. These factors are known to play a significant role in legal service production and may affect short-term billable capacity.

Expanding on the previous limitation; the workforce dynamics are simplified. The model assumes a fixed workforce size with a net inflow of one employee per quarter, based on the historical trend. It does not capture the in- and outflow of employees over time, which may affect variability in capacity that is seen in practice.

Finally, the overtime cost function is assumed rather than empirically estimated. Although the exponential specification captures increasing marginal costs of overtime, alternative functional forms and parameters could lead to different optimal allowed overtimes. The estimated value of $\hat{\theta}$ should therefore be interpreted as conditional on the maintained cost specification.

7.3 CHALLENGES

Beyond model-specific limitations, several structural challenges arise when applying simulation-based decision support in legal services. A first challenge concerns the nature of acceptance decisions. In practice, acceptance choices are shaped not only by expected revenue and capacity considerations, but also by strategic, relational, and reputational factors that are difficult to observe and quantify into formal objective functions. These considerations introduce multi-objective decision criteria that are difficult to formalize within a single quantitative objective function. As a result, model-based recommendations that optimize financial and capacity metrics may systematically diverge from observed behavior.

Furthermore, organizational adoption poses a non-trivial challenge. Even well-specified models may face resistance if their implications conflict with traditional heuristics, or partner incentives. This is particularly relevant in professional service firms, where individual autonomy is a central feature. Consequently, tractability, interpretability, and scenario-based communication are integral for effective use.

Despite these challenges, the results illustrate that simulation-based approaches can generate valuable insights into the long-run consequences of acceptance and capacity policies under uncertainty. Similar frameworks, such as the one presented in this thesis, might complement managerial judgment and support systematic decision-making in legal services and other professional service organizations.

CONCLUSION

The conclusion begins with a summary of the research in Section 8.1. Directions for future research and model extensions are discussed in Section 8.2.

8.1 SUMMARY

This thesis developed a simulation-based framework on real data for evaluating case acceptance policies in law firms operating under stochastic demand and capacity. By modeling case arrivals, workloads, and revenues probabilistically, the framework captures fundamental sources of uncertainty inherent to legal service operations. The simulation environment enables the evaluation of policies over multi-year horizons.

The results demonstrate that explicitly incorporating capacity constraints fundamentally shapes the temporal structure of the workload. While unrestricted acceptance maximizes expected revenue, it persistently exceeds capacity and has a higher relative variability. Hard capacity constraints reduce both mean revenue and variance, resulting in a more stable and realistic workload. Allowing controlled monthly overtime introduces an economically meaningful interior optimum: overtime increases expected net revenue, but convex overtime costs incur diminishing marginal gains. The stochastic approximation procedure converges to an optimal overtime threshold, with a relatively narrow confidence interval and stable post-convergence dynamics. Local sensitivity analysis confirms that the identified solution corresponds to a local maximizer of expected net revenue. Moreover, histogram-based analysis indicates that performance under the optimized policy is stable, with bounded variance driven by the stochasticity of demand and capacity.

Although the model abstracts from several operational complexities, it demonstrates that simulation-based optimization can quantify trade-offs between revenue generation and capacity utilization. The framework therefore provides both conceptual insight and a methodological foundation for more advanced decision-support systems. In conclusion, this thesis demonstrates the feasibility of constructing a stochastic simulation that jointly models demand and capacity in a law firm context. Evaluating multiple acceptance policies conditioned on workload and capacity highlights how different rules affect long-run performance. While the current model is not yet suitable for organizational deployment, the extensions outlined in Section 8.2 suggest a feasible path toward practical application in professional service settings.

8.2 FUTURE WORK

Several extensions could enhance the realism of the proposed framework. First, the model could allow workload shifting over time, for example by postponing non-billable activities or deferring portions of case work. This would better reflect short-term prioritization and day-to-day trade-offs faced by lawyers.

Second, future research could replace the assumption of a priori knowledge of case size and revenue with estimates available at arrival. Acceptance decisions would then be based on predicted workloads and revenues, with realized outcomes deviating from these forecasts. Incorporating forecast uncertainty would allow for the assessment of policy robustness and the design of fallback mechanisms.

Third, introducing heterogeneity among lawyers, such as differences in experience, specialization, and team composition, would allow for more sophisticated capacity simulation. Furthermore, cases could include requirements regarding specialization, seniority, and team composition, for a more accurate matching between case demands and available capacity.

Fourth, and most critically, the model assumes that case hours are worked according to a predetermined schedule. This could be relaxed by modeling case work as a stochastic arrival process where the arrivals represent the number of hours in a day that a case is worked on. Such a process could be inhomogeneous

and depend on expected case size, current capacity, and team availability, substantially increasing behavioral realism at the cost of additional complexity.

Taken together, these extensions move the framework toward the broader objective of constructing a digital twin of a law firm. While ambitious, this direction is worth the effort, as a development-ready simulation model could meaningfully support policy evaluation and scenario analysis under stochastic demand and capacity for professional services.

REFERENCES

- [1] Oscar Barros, Richard Weber, and Carlos Reveco. “Demand analysis and capacity management for hospital emergencies using advanced forecasting models and stochastic simulation”. In: *Operations Research Perspectives* 8 (2021), p. 100208. DOI: 10.1016/j.orp.2021.100208.
- [2] S. Bhatnagar, H. Prasad, and L. Prashanth. “Kiefer-Wolfowitz Algorithm”. In: *Stochastic Recursive Algorithms for Optimization*. Vol. 434. Lecture Notes in Control and Information Sciences. Springer, London, 2013, pp. 73–97. DOI: 10.1007/978-1-4471-4285-0_4. URL: https://doi.org/10.1007/978-1-4471-4285-0_4.
- [3] Centraal Bureau voor de Statistiek. *Persons: Educational attainment, sex, age and year*. 2025. URL: <https://opendata.cbs.nl/#/CBS/en/dataset/70936eng/table> (visited on 02/16/2026).
- [4] Eerste Kamer der Staten-Generaal. *Kabinetsplan aanpak administratieve lasten: Verslag van een schriftelijk overleg*. Eerste Kamer der Staten-Generaal, 2019.
- [5] Erwin Fielt et al. “Business model innovation and strategic transformation when confronting digital disruption: The case of data-driven business models for professional services”. In: *Australasian conference on information systems 2018*. UTS ePRESS. 2018.
- [6] Kenneth A. Grady. “Mining legal data: Collecting and analyzing 21st century gold”. In: *Journal of Internet Law* 20 (2017), pp. 1–12.
- [7] Geurt Jongbloed and Ger Koole. “Managing uncertainty in call centres using Poisson mixtures”. In: *Applied Stochastic Models in Business and Industry* 17.4 (2001), pp. 307–318. DOI: 10.1002/asmb.446.
- [8] Daniel Martin Katz. “Quantitative legal prediction — or how I learned to stop worrying and start preparing for the data-driven future of the legal services industry”. In: *Emory Law Journal* 62 (2012), pp. 909–966.
- [9] Jack Kiefer and Jacob Wolfowitz. “Stochastic estimation of the maximum of a regression function”. In: *Annals of Mathematical Statistics* 23.3 (1952), pp. 462–466. DOI: 10.1214/aoms/1177729392.
- [10] Moa Lidén. “Can criminal justice be predicted? Using regression analysis to predict judges’ decisions on petitions for new criminal trials”. In: *Science & Justice* 64.1 (2024), pp. 43–49. DOI: 10.1016/j.scijus.2023.10.002.
- [11] Luca Mancini and Anna Maria Paganoni. “Marked point process models for the admissions of heart failure patients”. In: *Statistical Analysis and Data Mining* 12.2 (2019), pp. 125–135. DOI: 10.1002/sam.11421.
- [12] Mordor Intelligence. *Europe legal services market size & share analysis: Growth trends & forecasts (2025–2030)*. Mordor Intelligence, 2025. URL: <https://www.mordorintelligence.com/industry-reports/europe-legal-services-market>.
- [13] Michal Ovádek. *Legislative output of the European Union: An automatically updated overview of the EU’s legislative activity*. 2025. URL: <https://michalovadek.github.io/eulaw/> (visited on 02/16/2026).
- [14] Frida Pemer, Charlotte Walker, and Tale Skjølsvik. “Digital business model innovation in tech-based professional service firms”. In: *Handbook of Services and Artificial Intelligence*. Edward Elgar Publishing, 2024, pp. 82–97.
- [15] Hans-Joerg Schuetz and Rainer Kolisch. “Approximate dynamic programming for capacity allocation in the service industry”. In: *European Journal of Operational Research* 218.1 (2012), pp. 239–250. DOI: 10.1016/j.ejor.2011.10.021.
- [16] Stanislaus Solomon et al. “Multiperiod stochastic resource planning in professional services organizations”. In: *Decision Sciences* 50.6 (2019), pp. 1281–1318. DOI: 10.1111/deci.12370.
- [17] Natalie M. Steiger and James R. Wilson. “An improved batch means procedure for simulation output analysis”. In: *Management Science* 48.12 (2002), pp. 1569–1586. DOI: 10.1287/mnsc.48.12.1569.110.
- [18] Thomson Reuters. *How technology drives law firm modernization*. White paper. Thomson Reuters, July 2021. URL: <https://legal.thomsonreuters.com/en/insights/white-papers/tech-drives-law-firm-modernization> (visited on 02/16/2026).
- [19] Felisa Vázquez-Abad and Bernd Heidergott. *Optimization and learning via stochastic gradient search*. Princeton University Press, 2025. ISBN: 9780691245867.
- [20] Georg Zeller and Matthias Scherer. “A comprehensive model for cyber risk based on marked point processes and its application to insurance”. In: *European Actuarial Journal* 12.1 (2022), pp. 33–85. DOI: 10.1007/s13385-021-00289-1.

APPENDIX

10.1 DISPERSION

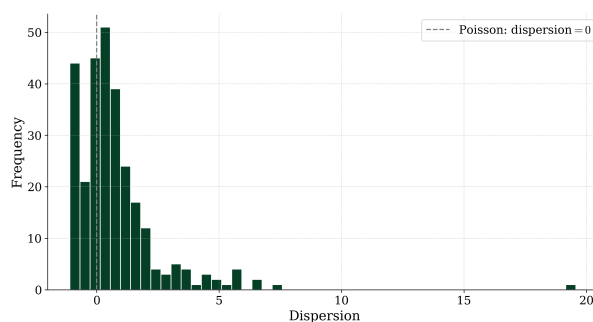


Figure 24: Dispersion based on observed arrivals on client level. A vertical line at zero dispersion is drawn to indicate the dispersion of the Poisson distribution.

10.2 CLOSED CASES

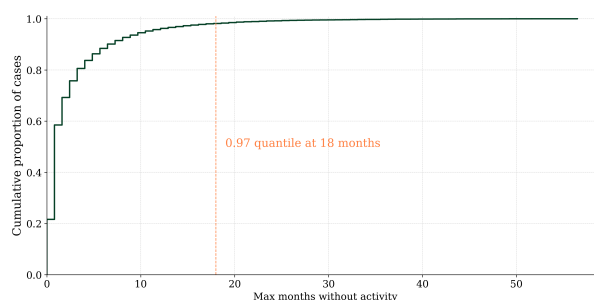


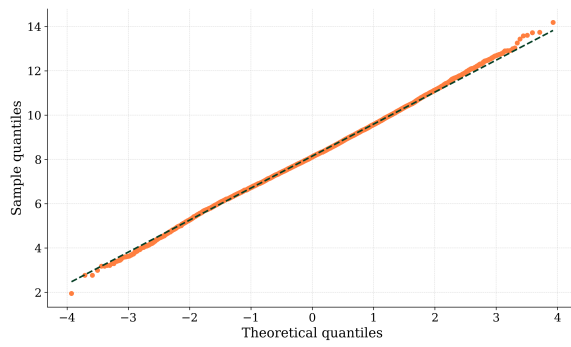
Figure 25: CDF of maximum inactive months per case. A vertical line is drawn at 18 months where only 3% of the cases have more months of inactivity.

10.3 CORRELATION

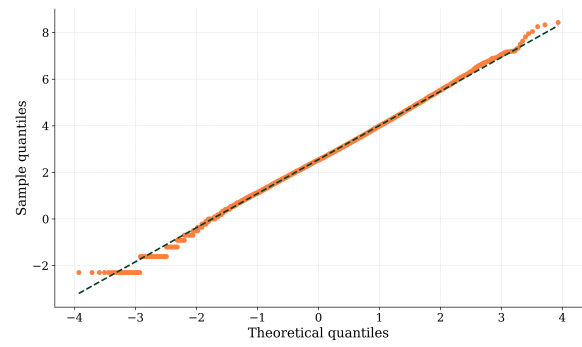
Variable	Spearman correlation	P-value
Average revenue per case	0.359	1.86×10^{-129}
Average size per case	0.365	1.92×10^{-134}

Table 4: Correlation between number of cases and average case characteristics per client.

10.4 Q-Q PLOTS



(a) Case revenue. Jacque-Bera test for normality statistic is 41.2045, with a p-value of 1.1286e-09.



(b) Case size. Jacque-Bera test for normality statistic is 41.0125, with a p-value of 1.2424e-09.

Figure 26: Normal Q-Q plots of $\ln(\text{revenue})$ and $\ln(\text{case size})$.

10.5 MONTHLY DISTRIBUTION

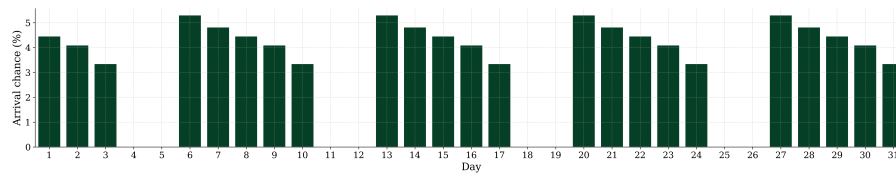


Figure 27: Arrival chance of cases for example month January 2025.

10.6 HOURS DISTRIBUTION

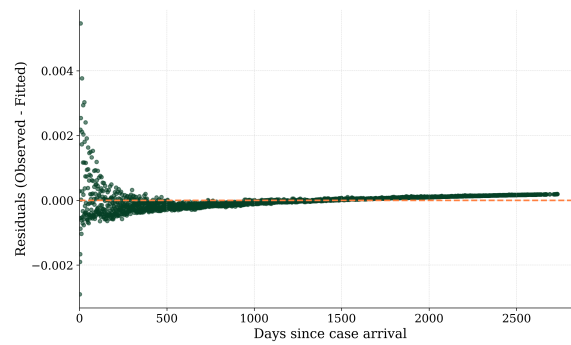


Figure 28: Residuals of hours distribution with shifted inverse square root fit. The coefficient of determination $R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 0.96$, where SS_{res} is the sum of squares of residuals and SS_{tot} is total sum of squares.

10.7 DAILY CAPACITY

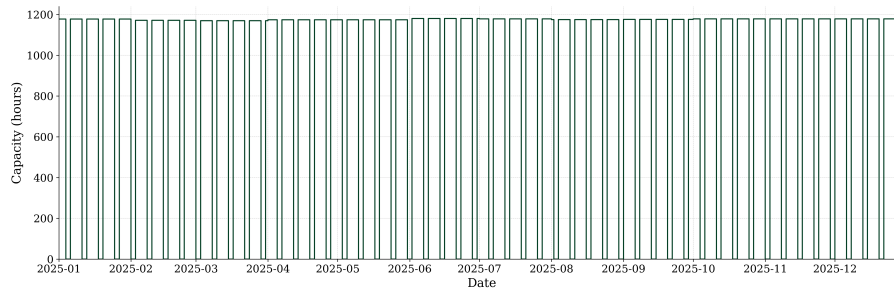
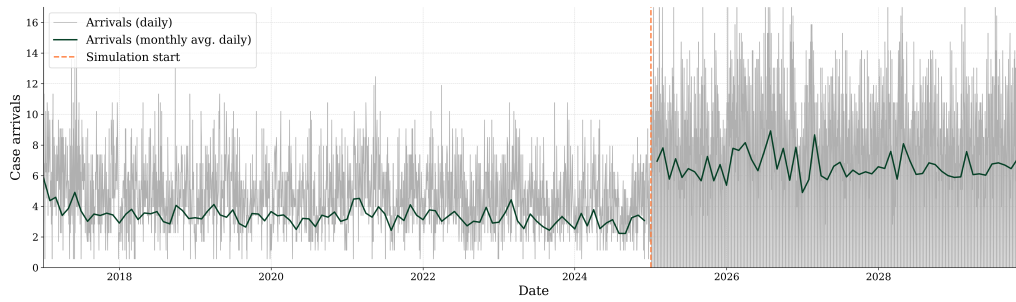


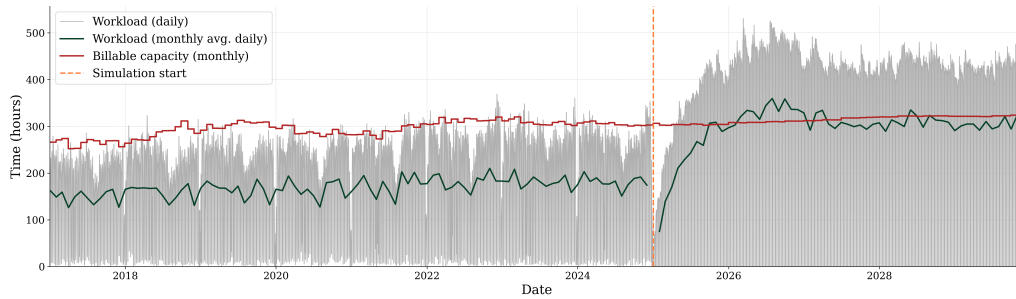
Figure 29: Simulated capacity with a daily resolution in 2025.

10.8 ACCEPTANCE POLICIES

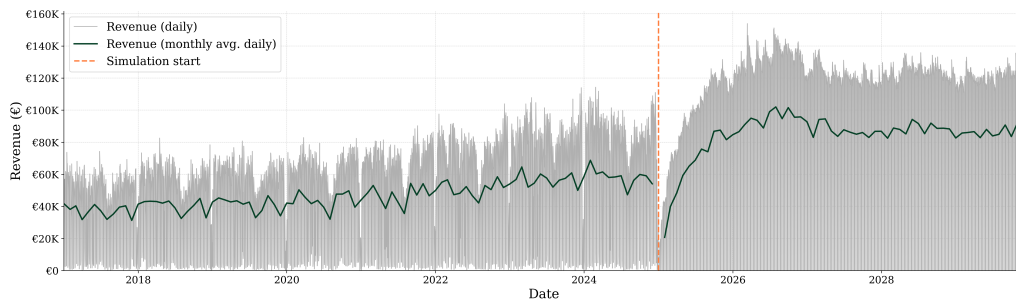
10.8.1 Policy 0. All arrivals



(a) Historical and simulated number of arrivals.



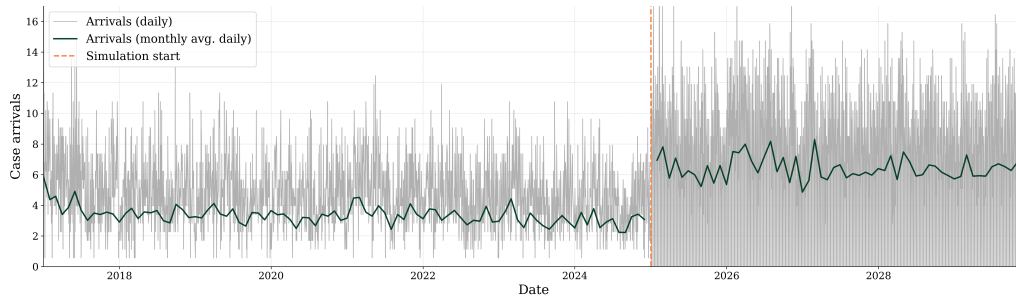
(b) Historical and simulated workload. The billable capacity is shown to indicate a soft boundary for the monthly workload.



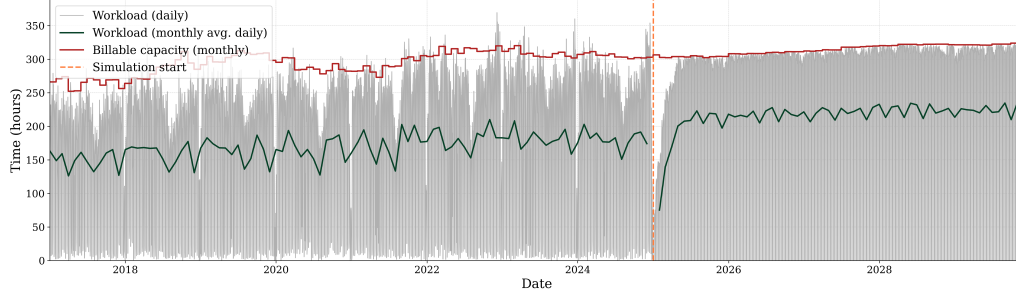
(c) Historical and simulated revenue.

Figure 30: Time series of workload and revenue for Policy 0.

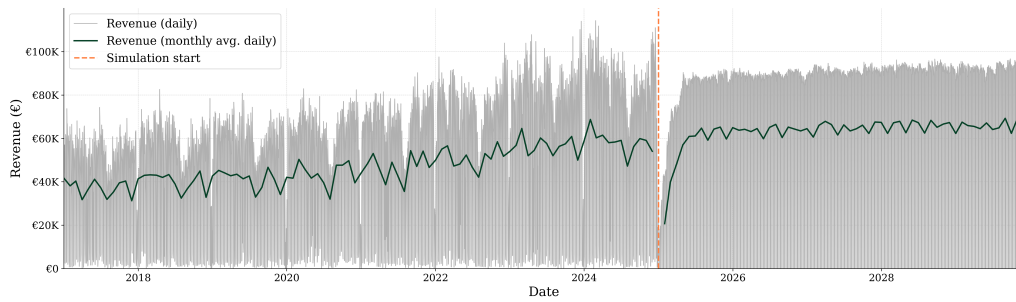
10.8.2 Policy I. Hard daily capacity boundary



(a) Historical and simulated number of arrivals.



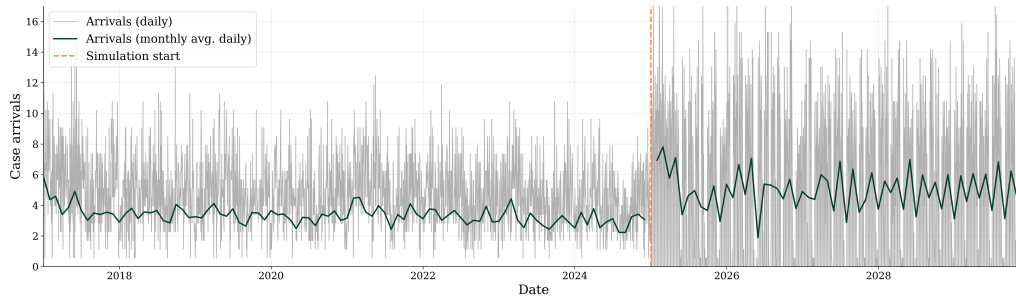
(b) Historical and simulated workload. The billable capacity is shown to indicate a soft boundary for the monthly workload.



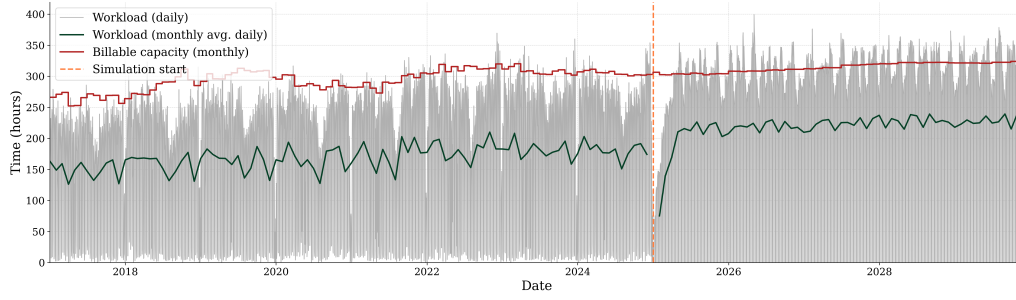
(c) Historical and simulated revenue.

Figure 31: Time series of workload and revenue for Policy I.

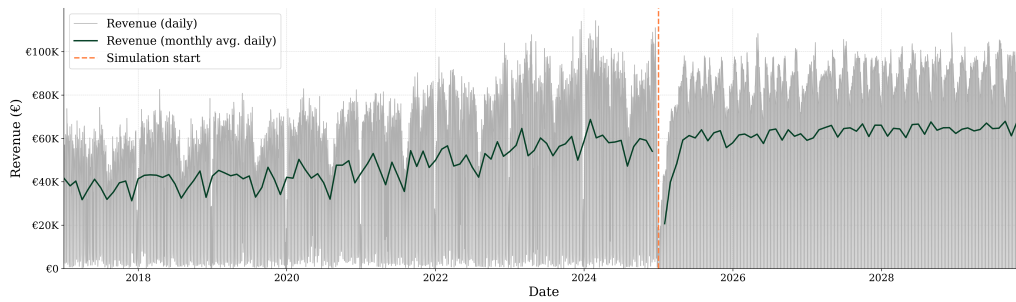
10.8.3 Policy II. Hard monthly capacity boundary



(a) Historical and simulated number of arrivals.



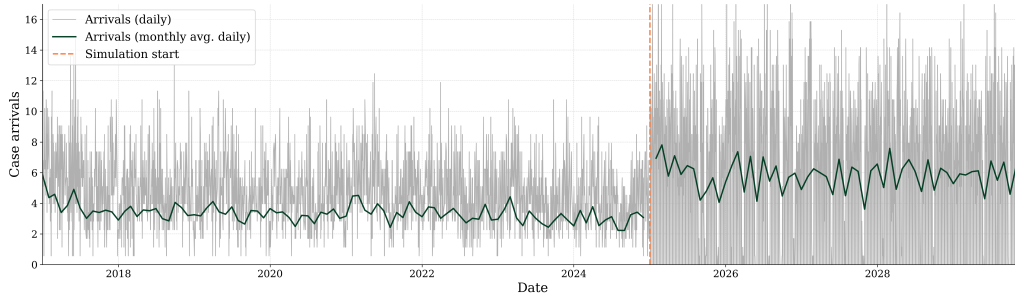
(b) Historical and simulated workload. The billable capacity is shown to indicate a soft boundary for the monthly workload.



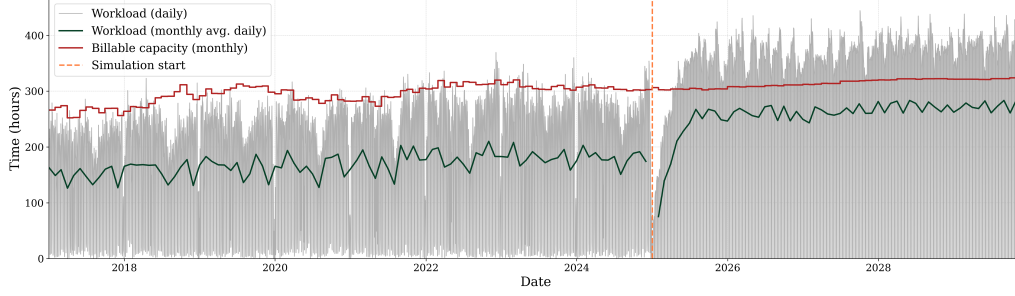
(c) Historical and simulated revenue.

Figure 32: Time series of workload and revenue for Policy II.

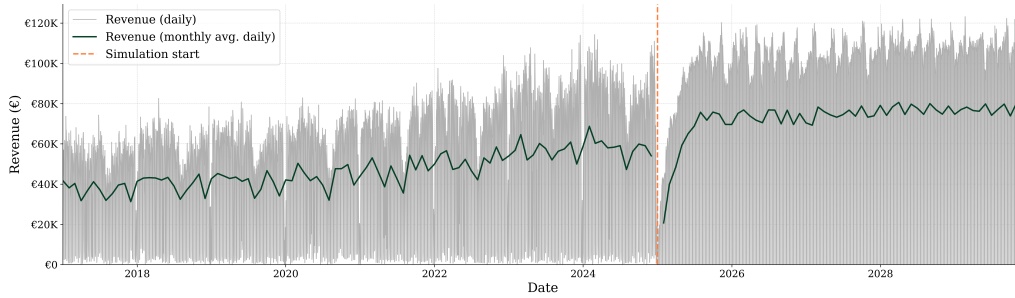
10.8.4 Policy III. Soft monthly capacity boundary



(a) Historical and simulated number of arrivals.



(b) Historical and simulated workload. The billable capacity is shown to indicate a soft boundary for the monthly workload.



(c) Historical and simulated revenue.

Figure 33: Time series of arrivals, workload and revenue for Policy III with $\theta = 2,430$.

10.9 COMPUTATIONAL INTENSITY

Running the model with a simulation horizon of five years requires between 60 and 90 seconds per replication, depending on the realized number of arrivals. The 200 replications used to construct the histograms (Figures 15, 17, 19, and 23) required approximately eight hours of total computation time.

Each stochastic approximation run took approximately three minutes to complete. For both initializations ($\theta_0 \in \{1,000; 2,500\}$), the algorithm was executed for 1,250 iterations. Since each iteration evaluates two perturbed parameter values, this resulted in a total of 5,000 simulation runs. The complete stochastic approximation procedure therefore required approximately 250 hours of computational time.