Uncertainty-Aware Forecasting and Inventory Optimization for ATM Cash Management in Vietnam

Huu-Thanh Phan[1,2][0000−0002−2917−3664], Xuan-Bach Le[1,2][0009−0003−6848−5403]*, and Tho Quan[1,2][0000−0003−0467−6254]

[1] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
[2] Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam
{phthanh.sdh231, lexuanbach, qttho}@hcmut.edu.vn

Abstract. ATM networks remain essential cash-distribution infrastructure in cash-intensive economies such as Vietnam, where cash still accounts for most consumer payments. Effective replenishment requires balancing the opportunity cost of idle capital against the risk of stockouts. We evaluate a forecast-then-optimize framework that integrates modern probabilistic deep learning models with a periodic-review base-stock policy. Thirty forecasting configurations are benchmarked, including statistical baselines (SARIMA, Exponential Smoothing, Prophet), tree-based ensembles (LightGBM, Random Forest), and global neural architectures (N-BEATS, Temporal Fusion Transformer), using daily withdrawal data from 84 ATMs over 3.7 years. Performance is assessed not only by forecast accuracy and calibration, but also by downstream inventory cost under a realistic replenishment policy. Global neural quantile models substantially outperform classical alternatives, reducing simulated total cost by roughly half while maintaining near-perfect calibration and 99% fill rates. Importantly, forecast accuracy alone proves to be an unreliable indicator of business outcomes, as models with similar error can yield markedly different costs. ATM-level heterogeneity further shows that no single model dominates across all locations, supporting a monitoring-based deployment strategy that adapts model assignments according to realized cost.

Keywords: ATM cash management · Forecast-then-optimize · Probabilistic forecasting · Base-stock policy · Decision-centric evaluation

1 Introduction

Automated Teller Machine (ATM) networks are central to cash distribution in cash-intensive economies. In Vietnam, cash represents about 70% of consumer transactions [10], making ATM networks primary liquidity channels. Replenishment decisions therefore affect both capital efficiency and service reliability.

* Corresponding author

ATM cash management can be framed as a stochastic inventory problem with asymmetric costs, where excess cash creates holding costs and shortages generate lost transactions as well as reputational or regulatory risks [30]. At scale, overstocking locks up idle capital, whereas understocking increases emergency cash-in-transit operations [28]. Although this setting resembles a high-service-level newsvendor problem under multi-period uncertainty, many systems still rely on static rules that ignore non-stationary demand patterns such as salary cycles and holiday surges [30].

From a research perspective, ATM cash management lies at the intersection of two largely parallel streams: demand forecasting, which estimates future withdrawals and their uncertainty [14,19], and replenishment optimization, which determines refill quantities and schedules to balance cost and service levels [34,7]. Although both areas are mature, they are rarely integrated in practice. Probabilistic multi-step forecasts are typically evaluated using statistical accuracy measures such as MAE or CRPS [12,21]. However, less attention has been given to how distributional properties, especially upper-tail calibration, affect operational costs under fixed inventory policies. As a result, it remains unclear when improvements in probabilistic calibration lead to economically meaningful gains at a given service level.

We address this gap through an end-to-end forecast-then-optimize framework (Figure 1). Multiple models generate predictive distributions of ATM withdrawals, which are then translated into replenishment decisions under a fixed periodic-review base-stock policy, allowing us to isolate the effect of forecast quality. Performance is assessed using both time-series metrics and simulated inventory outcomes, including total cost and fill rate. We find that well-calibrated global neural quantile models reduce replenishment costs relative to classical benchmarks while maintaining target service levels, and that forecast accuracy alone does not reliably indicate operational performance.

Our contributions are fivefold. First, we provide a large-scale empirical benchmark of 30 forecasting configurations spanning statistical, tree-based, and global neural models under consistent training regimes. Second, we adopt a decision-centric evaluation that links forecast distribution quality to downstream inventory cost, revealing systematic gaps between statistical accuracy and operational performance. Third, we introduce a transparent mapping from predictive quantiles to base-stock decisions, enabling controlled comparisons across model families. Fourth, we conduct regime-sensitive analyses across ATMs and peak periods, including Tet, emphasizing the importance of tail behavior at high service levels. Finally, we outline deployment implications through a monitoring-based model selection strategy that adapts per-ATM assignments based on realized cost.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature. Section 3 introduces the formal problem formulation and decision framework. Section 4 describes the experimental design and operational setup. Section 5 presents empirical results and robustness analyses. Section 6 discusses broader insights and practical implications. Finally, Section 7 concludes and outlines directions for future research.

2 Related Work

ATM demand forecasting has progressed from per-machine statistical methods to supervised and neural approaches. Early studies rely on ARIMA, SARIMA, and exponential smoothing [16,24,13]. These models are transparent and easy to implement, but their largely linear and mildly non-stationary assumptions limit their ability to capture regime shifts and holiday effects. To address these limitations, fuzzy and neuro-fuzzy models encode qualitative rules and can produce prediction intervals, but they require manual design and scale poorly in large networks [1].

Machine learning approaches treat ATM forecasting as a regression task with engineered temporal features, including SVR and tree-based ensembles [15]. Suder et al. [26] report that XGBoost outperforms classical and several neural baselines in point-forecast accuracy across 61 ATMs, but their evaluation is limited to error metrics and does not assess inventory impact. More recently, Cedolin et al. [6] compare ARIMA, DNN, and Prophet for ATM cash demand, while Vangala and Vadlamani [30] model chaotic dynamics in Indian ATM withdrawals using LSTM and CNN. Neural models such as LSTMs and encoder-decoder architectures have been applied to multi-week forecasting [2,13,32], typically trained per ATM or cluster and focused on point predictions. Global models such as N-BEATS [23] and the Temporal Fusion Transformer [19], as well as clustering strategies [31,24,25], are designed to improve stability across heterogeneous ATMs. However, most studies still emphasize point-forecast accuracy rather than operational outcomes.

Under asymmetric holding and shortage costs, the full predictive distribution is crucial for inventory control. Ekinci et al. [8] integrate prediction intervals into robust replenishment models to reduce stockout risk at modest holding-cost increases. Rafi et al. [24] apply DeepAR without reporting calibration diagnostics, and Suder et al. [26] use a Bayesian VAR to model regime changes without evaluating policy-level outcomes.

In parallel, replenishment optimization has been studied using heuristic rules, mixed-integer linear programming [7], dynamic programming, and robust optimization [20]. These approaches typically treat forecasts as deterministic inputs and emphasize routing or logistics efficiency [28]. Zeinalkhani et al. [33] recently combine LSTM-based forecasting with multi-objective optimization for branch cash logistics, but do not evaluate calibrated probabilistic forecasts under inventory cost. Reinforcement learning has also been explored in simulation [32], but systematic comparisons with calibrated probabilistic forecasts under a common economic objective remain scarce.

Overall, three gaps remain. First, modern global neural architectures have not been systematically benchmarked for ATM demand. Second, probabilistic forecasts are rarely assessed within a calibrated, decision-focused framework that links distributional quality to inventory cost. Third, forecasting and replenishment optimization are usually examined separately rather than under controlled forecast-and-policy comparisons. In addition, the Vietnamese ATM context is largely underrepresented in existing empirical research.

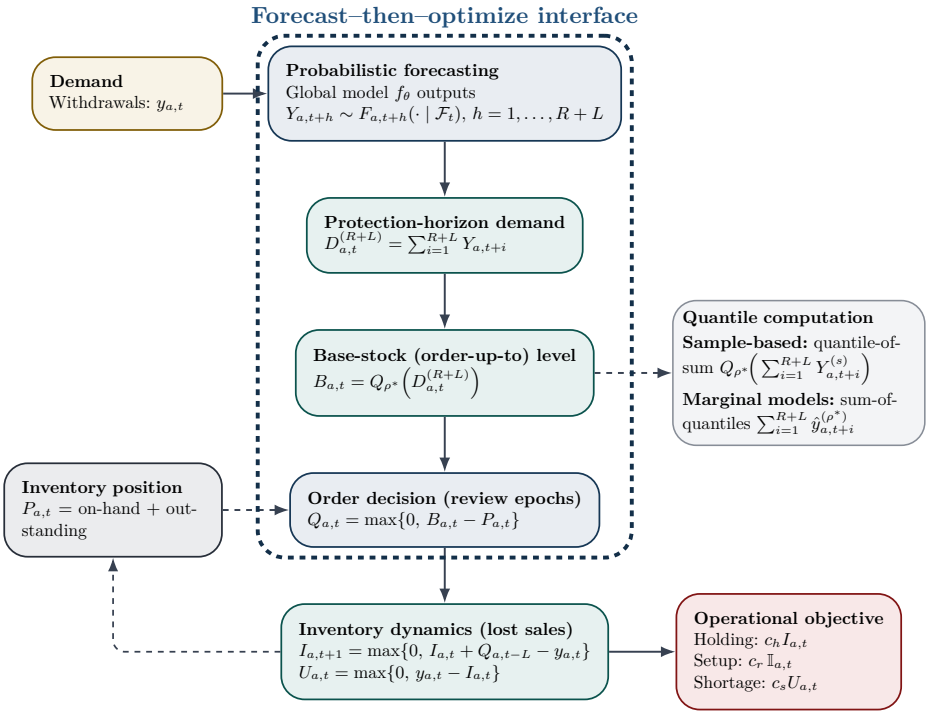

Fig. 1: Forecast-then-optimize framework for ATM replenishment: predictive demand distributions are mapped to periodic-review base-stock decisions and evaluated through inventory dynamics and total cost.

3 Inventory Control under Uncertainty

Figure 1 illustrates our forecast-then-optimize framework. Each ATM is modeled as an independent, single-item periodic-review system, where a probabilistic forecaster generates multi-step demand distributions that are translated into replenishment decisions under a fixed base-stock policy. This separation allows us to isolate how forecast quality, particularly uncertainty calibration, influences downstream operational cost.

Stochastic objective. Let $\mathcal{A}$ denote the set of ATMs and let H be the planning horizon (days). For each ATM $a \in \mathcal{A}$ and each day $t \in \{1, \ldots, H\}$, we define:

- $I_{a,t}$: on-hand inventory (cash available in ATM a at time t),
- $U_{a,t}$: unmet demand (lost withdrawals due to stockout),
- $\mathbb{I}_{a,t} \in \{0, 1\}$: replenishment indicator, where $\mathbb{I}_{a,t} = 1$ if a replenishment occurs and 0 otherwise.

Let $\mathbf{Q} = \{Q_{a,t} : a \in \mathcal{A},\ t = 1, \ldots, H\}$ be the replenishment decisions, where $Q_{a,t}$ is the cash delivered to ATM a at time t. The cost structure consists of a

per-unit holding cost c_h for idle cash, a per-unit shortage penalty c_s for unmet demand, and a fixed replenishment cost c_r incurred each time a delivery is made. The objective is to minimize expected total cost over the horizon H:

$$\min_{\mathbf{Q}} \mathbb{E}_{\mathbf{y}} \left[\sum_{a \in \mathcal{A}} \sum_{t=1}^{H} (c_h I_{a,t} + c_s U_{a,t} + c_r \mathbb{I}_{a,t}) \right], \tag{1}$$

where the expectation is taken over the stochastic demand process $\mathbf{y}$ [7]. This setup follows the standard multi-period lost-sales inventory model with fixed ordering costs commonly used in ATM replenishment [7,8]. The ratio $\rho^* = \frac{c_s}{c_s + c_h}$ defines the classical newsvendor critical fractile and implies a target service level of about 99% in our parameterization [8]. We use this ratio to guide base-stock decisions in the multi-period setting.

Inventory dynamics. The system operates under periodic review with review period R and deterministic lead time L. Let $y_{a,t}$ be realized withdrawals on day t. Inventory evolves as:

$$I_{a,t+1} = \max\{0,\, I_{a,t} + Q_{a,t-L} - y_{a,t}\},$$

where $Q_{a,t-L}$ is the order placed L days earlier. We assume a lost-sales model without backlogging; excess demand is lost and does not carry forward:

$$U_{a,t} = \max\{0,\, y_{a,t} - I_{a,t}\}.$$

Orders satisfy $Q_{a,t} \geq 0$ and can be placed only at scheduled review epochs $t \in \{R, 2R, 3R, \dots\}$.

Probabilistic forecasting. Replenishment decisions require forecasts over the protection horizon, the period during which new orders cannot influence inventory. Under periodic review with lead time L, this horizon spans $R + L$ days.

We use a global forecasting model f_θ to estimate the full conditional predictive distribution of future withdrawals for each ATM. Specifically, for horizon $h = 1, \dots, R + L$, we model:

$$Y_{a,t+h} \sim F_{a,t+h}(\cdot \mid \mathcal{F}_t),$$

where $\mathcal{F}_t$ denotes the information available at time t (for example, past withdrawals and covariates), and $F_{a,t+h}$ is the forecast distribution used for probabilistic decision-making [12,18]. We train models with distributional losses (for example, quantile loss or CRPS) to obtain calibrated uncertainty for high-service-level decisions.

Policy mapping. At each review epoch t, we implement a periodic-review base-stock policy (Figure 1). The order-up-to level is set as the ρ^*-quantile of cumulative demand over the protection horizon:

$$B_{a,t} = Q_{\rho^*} \left(\sum_{i=1}^{R+L} Y_{a,t+i} \right). \tag{2}$$

Table 1: Operational configuration for the ATM replenishment simulation, including review timing, cost structure, service-level target, and policy constraints.

Category	Parameter	Value
Inventory timing	Review period R	7 days
Inventory timing	Lead time L	3 days
Inventory timing	Protection horizon $R + L$	10 days
Cost parameters	Per-order cost c_r	2,000,000 VND
Cost parameters	Holding cost c_h	0.00026 VND/VND-day
Cost parameters	Shortage cost c_s	0.02574 VND/VND
Policy settings	Target service level ρ^*	$\sim$99%
Policy settings	Capacity constraint	Not enforced
Policy settings	Dynamic trigger	Disabled

Here, $\sum_{i=1}^{R+L} Y_{a,t+i}$ is total demand over the review period plus lead time, and $Q_{\rho^*}(\cdot)$ gives the ρ^*-quantile, implying a stockout probability of about $1 - \rho^*$. Given $B_{a,t}$, the replenishment quantity is

$$Q_{a,t} = \max\{0,\, B_{a,t} - P_{a,t}\},$$

where $P_{a,t}$ denotes inventory position at time t, including on-hand cash and outstanding replenishment orders that have been placed but not yet received [9]. Because quantiles are not additive [9], computation of $B_{a,t}$ depends on forecast form: sample-based models use a quantile-of-sum from simulated trajectories, whereas marginal quantile regressors use a sum-of-quantiles approximation (typically conservative under positive dependence). We treat ATMs independently, so costs and decisions are computed separately for each a.

4 Experimental Design and Evaluation Framework

Building on the inventory framework, this section presents the experimental design used for model evaluation. We fix the replenishment environment and dataset, then compare statistical, machine learning, and neural models using both forecasting metrics and simulated inventory outcomes. Policy and cost parameters are held constant so that observed differences reflect forecast quality rather than operational setting changes.

4.1 Operational Setup

Table 1 summarizes the operational configuration used in all experiments. We evaluate every forecasting model under this fixed setting to isolate the effect of predictive quality on inventory decisions.

Table 2: Statistics and filtering criteria for the ATM withdrawal dataset.

Attribute	Value
ATMs retained	84 (from 474 candidates)
Observation period	2022-01-01 – 2025-09-30 (1,369 days)
Test horizon H	30 days
Coverage requirement	$\geq 90\%$ observed (last 1,050 days)
Maximum missing gap	< 10 consecutive days
Data granularity	Daily withdrawals (VND totals)

We simulate replenishment in a periodic-review lost-sales system with review period $R = 7$ days and deterministic lead time $L = 3$ days, yielding a protection horizon of $R + L = 10$ days. Cost parameters are fixed across models: per-order cost $c_r = 2{,}000{,}000$ VND, holding cost $c_h = 0.00026$ VND per VND per day, and shortage penalty $c_s = 0.02574$ VND per VND. These values imply a target service level of $\rho^* \approx 99\%$.

We also standardize policy constraints. We impose no ATM capacity limits and disable dynamic trigger rules to preserve a pure periodic-review base-stock policy. Unless otherwise stated, inventory is initialized at the ρ^*-quantile of lead-time demand to reduce initialization bias.

4.2 Data and Preprocessing

Table 2 summarizes the dataset and filtering criteria. We use daily ATM withdrawal totals in VND from a major Vietnamese commercial bank spanning 2022-01-01 to 2025-09-30 (1,369 days). The unit of analysis is daily withdrawals per ATM, with all identifiers anonymized and no customer-level data used.

From 474 candidate machines, we retain 84 ATMs under strict data-quality criteria: a shared observation window, at least 90% coverage over the final 1,050 pre-test days, no missing gap longer than 10 consecutive days, and complete data over the 30-day test horizon ($H = 30$). We linearly interpolate short remaining gaps but preserve extreme events, such as pre-holiday surges, because these represent genuine tail demand relevant for inventory control.

We apply a deployment-style temporal split consistent with standard time-series practice [14,3]: the final 30 days form the test set, the preceding 20% of observations are used for validation, and the remaining history is used for training. For fairness analysis and clustered model variants, we group ATMs with HDBSCAN [5,22] on dynamic time warping distances [4] computed from z-score-normalized histories. This procedure yields three stable demand cohorts and a set of outliers.

4.3 Forecasting Models

We evaluate 30 forecasting configurations across three dimensions: architecture family, uncertainty formulation, and training scope. This design enables controlled comparison of modeling capacity, information sharing, and calibration.

Statistical baselines. We use three transparent local benchmarks: SARIMA with weekly seasonality, exponential moving average, and Prophet [27]. They provide interpretable classical reference points for per-ATM forecasting performance.

Tree-based models. We implement LightGBM [17] with a direct multi-step strategy using lagged withdrawals, calendar features, and static ATM attributes. We also train a quantile variant to produce conditional quantiles. In addition, we include Random Forest [29] as a point-forecast baseline. Both models are evaluated under local, clustered, and global training schemes.

Neural architectures. We evaluate two neural models. N-BEATS [23] captures trend and seasonality through residual stacks with basis expansions, while the Temporal Fusion Transformer [19] combines recurrent layers, attention, and static embeddings to model cross-ATM heterogeneity. Both architectures are trained in point and quantile variants under local, clustered, and global settings.

Uncertainty modeling. Quantile variants are trained with pinball loss [18]. For neural models, we apply Monte Carlo dropout at inference with 500 stochastic passes to approximate joint demand trajectories [11]. LightGBM quantile models produce marginal quantiles such as q01, q50, and q99. All probabilistic forecasts are translated into replenishment decisions using the common base-stock policy described in Section 3.

Training protocol. We fix hyperparameters in advance and do not perform automated tuning. Neural models are trained with Adam, gradient clipping, and early stopping, while tree-based models use early stopping after 30 to 50 rounds. All configurations share the same temporal splits and validation ratios to ensure comparability. Full settings are archived for reproducibility.

Model inputs. Inputs include static ATM identifiers (embedded for neural models) and calendar features with Vietnamese holidays. Tree-based models additionally use lagged withdrawals at 1 to 7, 14, 30, 60, and 90 days.

4.4 Evaluation Protocol

Table 3 summarizes deterministic and probabilistic metrics used to evaluate predictive performance before forecasts are translated into operational decisions.

Evaluation proceeds in three stages. Each model generates 30-day multi-step forecasts for all 84 ATMs. We compute metrics per ATM and report mean and standard deviation. We then map forecasts to replenishment decisions under the base-stock policy (Eq. 2) and simulate performance over the test horizon.

Table 3: Evaluation metrics used for deterministic and probabilistic assessment.

Metric	Definition	Interpretation
Mean Absolute Error (MAE)	$\frac{1}{H}\sum_{h=1}^{H}\lvert y_{a,t+h}-\hat{y}_{a,t+h}\rvert$	Lower is better
Weighted Absolute Percentage Error (WAPE)	$\frac{\sum_{h=1}^{H}\lvert y_{a,t+h}-\hat{y}_{a,t+h}\rvert}{\sum_{h=1}^{H}\lvert y_{a,t+h}\rvert}$	Scale-free error
Symmetric Absolute Percentage Error (SMAPE)	$\frac{100}{H}\sum_{h=1}^{H}\frac{\lvert y_{a,t+h}-\hat{y}_{a,t+h}\rvert}{(\lvert y_{a,t+h}\rvert+\lvert\hat{y}_{a,t+h}\rvert)/2}$	Symmetric percentage error
Mean Absolute Scaled Error (MASE)	$\frac{\frac{1}{H}\sum_{h=1}^{H}\lvert y_{a,t+h}-\hat{y}_{a,t+h}\rvert}{\frac{1}{T-m}\sum_{t=m+1}^{T}\lvert y_{a,t}-y_{a,t-m}\rvert}$	< 1 outperforms seasonal naive
Continuous Ranked Probability Score (CRPS)	$\int_{-\infty}^{\infty}\left(F(z)-\mathbf{1}\{y\le z\}\right)^2 dz$	Full-distribution score
Coverage	$\frac{1}{H}\sum_{h=1}^{H}\mathbf{1}\{y_{a,t+h}\le Q_\alpha(F_{a,t+h})\}$	Empirical quantile coverage
Calibration error	$\left\lvert\widehat{\mathrm{Cov}}_\alpha-\alpha\right\rvert$	Deviation from nominal coverage

Deterministic accuracy. Point forecasts are evaluated using MAE, SMAPE, WAPE, and MASE. MASE is scaled relative to a seasonal naive baseline with $m = 7$, where values below one indicate improvement.

Probabilistic quality. We assess predictive distributions using CRPS, empirical quantile coverage, and calibration error. CRPS captures overall distribution quality, while coverage and calibration measure uncertainty reliability under high service-level targets.

Business performance. Business metrics are the primary criterion. For each model, we simulate the base-stock policy over the 30-day horizon to compute total cost, fill rate, and cash utilization. With a review period of $R = 7$ days, the 30-day test window contains four scheduled review epochs (at days 7, 14, 21, and 28), each triggering a replenishment decision. Total cost drives model selection, while fill rate and cost decomposition explain performance differences.

5 Experiments and Results

We evaluate 30 configurations across 84 ATMs using accuracy, probabilistic, and business metrics under a fixed base-stock policy. Fleet-level and ATM-level analyses assess performance, calibration, and cost, highlighting both alignment and divergence between statistical accuracy and operational impact.

Table 4: Overall model performance across 84 ATMs, sorted by mean total cost. MAE and Cost are in M VND; SMAPE and WAPE are in %. MASE is relative to a seasonal naive benchmark (< 1 indicates improvement). QL (scaled quantile loss) and CRPS are scaled scores. Cov. is quantile coverage (%), Cal. is calibration error, and Fill is fill rate (%). Arrows indicate optimization direction (↓ lower is better; ↑ higher is better). Best values in each column are bold. Model naming follows `Architecture--Type--Scope`, where `Type` is `Q` (quantile) or `P` (point), and `Scope` is `L` (local), `C` (clustered), or `G` (global). "–" denotes point-only models.

	Deterministic				Probabilistic				Business	
Model	MAE↓	SMAPE↓	WAPE↓	MASE↓	QL↓	CRPS↓	Cov.↑	Cal.↓	Fill↑	Cost↓
TFT-Q-G	**46.9**	**37.2**	**35.5**	**0.506**	**0.019**	0.127	98.5	0.019	99.1	**16.0**
N-BEATS-Q-G	51.6	40.4	37.1	0.537	0.020	0.134	98.1	0.023	99.0	16.1
N-BEATS-Q-C	49.9	39.2	36.3	0.527	0.021	0.131	98.5	0.020	99.2	16.1
TFT-Q-C	48.8	38.2	35.9	0.522	0.022	0.130	98.8	0.018	99.2	16.9
RF-P-L	58.3	41.4	41.6	0.577	–	–	–	–	96.3	17.0
LightGBM-P-L	51.8	39.6	38.6	0.545	–	–	–	–	95.4	17.2
LightGBM-P-C	51.0	39.4	38.4	0.547	–	–	–	–	94.1	18.7
LightGBM-P-G	51.1	39.1	37.9	0.541	–	–	–	–	93.8	18.9
Prophet-Q-L	64.5	42.5	41.0	0.584	0.030	0.166	98.4	0.019	99.2	19.1
N-BEATS-P-C	50.4	39.2	37.8	0.533	–	–	–	–	92.6	19.3
N-BEATS-P-G	51.7	39.6	37.4	0.536	–	–	–	–	93.0	19.7
TFT-P-G	67.2	51.3	49.0	0.691	–	–	–	–	90.0	21.9
RF-P-C	60.0	42.4	42.9	0.602	–	–	–	–	94.7	22.0
RF-P-G	59.5	42.6	43.3	0.605	–	–	–	–	94.6	22.3
LightGBM-Q-C	50.7	38.3	36.4	0.524	0.022	0.097	98.3	0.021	**100.0**	23.4
LightGBM-Q-G	50.2	38.2	36.1	0.522	0.022	**0.096**	98.1	0.022	**100.0**	23.6
TFT-P-C	73.7	53.8	52.4	0.745	–	–	–	–	89.5	23.9
TFT-Q-L	77.4	55.0	52.5	0.726	0.055	0.253	**100.0**	**0.010**	99.9	24.4
LightGBM-Q-L	50.3	38.9	37.4	0.537	0.022	0.100	98.5	0.018	**100.0**	26.2
N-BEATS-P-L	110.7	70.2	81.2	1.045	–	–	–	–	92.2	26.6
EMA-P-L	73.6	46.2	46.3	0.658	–	–	–	–	88.6	30.5
SARIMA-P-L	70.2	47.4	46.6	0.665	–	–	–	–	86.1	34.4
TFT-P-L	193.4	114.5	147.0	1.908	–	–	–	–	65.8	51.2
N-BEATS-Q-L	92.6	60.6	68.3	0.894	0.184	0.370	99.9	0.011	**100.0**	55.2

5.1 Aggregate Performance

Table 4 reports aggregate results across 84 ATMs, ranking 30 configurations by mean total cost alongside deterministic, probabilistic, and business metrics.

Model notation. Model labels follow `Architecture--Type--Scope`. In the suffix `Type--Scope`, `P` denotes point forecasts and `Q` denotes quantile forecasts, while `L`, `C`, and `G` denote local, clustered, and global training, respectively. For example, `P-C` is a clustered point model and `Q-L` is a local quantile model.

A clear hierarchy emerges: a small set of neural quantile models leads. **TFT-Q-G** attains the lowest cost (16.0 M VND), followed closely by N-BEATS-Q-G and N-BEATS-Q-C, combining strong point accuracy, competitive probabilistic scores, and near-perfect fill rates (approximately 99%).

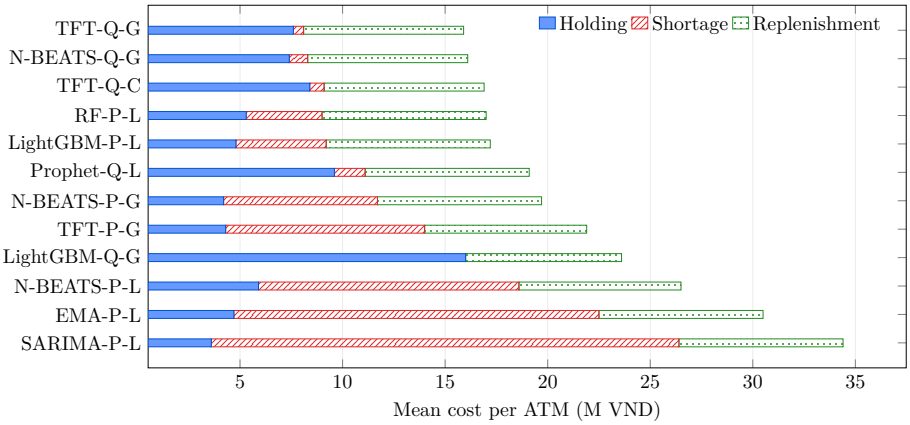

Fig. 2: Mean cost breakdown per ATM by model (holding, shortage, and replenishment costs).

Classical baselines (for example, SARIMA-P-L and EMA-P-L) rank near the bottom with higher costs and weaker service levels, while tree-based models form a middle tier. Compared with SARIMA-P-L (34.4 M VND), the best neural model reduces cost by more than 50% and raises fill rate from about 86% to above 99%, indicating a clear operational improvement.

Cost–accuracy misalignment. The results reveal a gap between point accuracy and business impact. RF-P-L ranks mid-table in MASE but fifth in total cost, showing that deterministic accuracy alone does not ensure cost efficiency. Likewise, strong calibration or coverage by itself is insufficient; effective performance requires balancing inventory and risk rather than optimizing any single metric.

Figure 2 decomposes total cost into holding, shortage, and replenishment components, providing structural insight into the rankings. The leading neural quantile models display balanced profiles: moderate holding and replenishment costs with low shortage cost. In contrast, strong point-forecast baselines reduce holding cost but incur higher shortage cost, indicating weaker protection against demand uncertainty. Tree-based quantile models (for example, LightGBM-Q-G) nearly eliminate shortage cost, but at the expense of substantially higher holding cost, which offsets service gains.

Overall, superior performance stems from calibrated risk allocation across components rather than minimizing any single cost. The top models effectively balance inventory and service, achieving high fill rates and low total cost. This structural evidence reinforces the economic advantage of well-configured probabilistic deep models beyond traditional accuracy metrics.

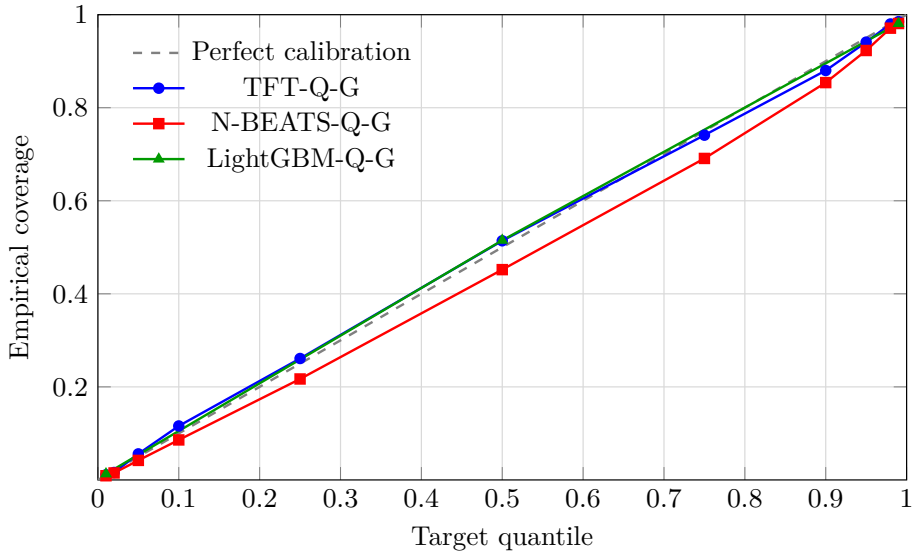

Fig. 3: Calibration curves for leading quantile models; the dashed line denotes perfect calibration (empirical coverage equals target quantile).

Table 5: Per-ATM win-rate analysis (84 ATMs).

Model 1	Model 2	Cost (% win, p)	MAE (% win, p)
TFT-Q-G	N-BEATS-Q-G	40.5% (0.101)	67.9% (<0.001)
TFT-Q-G	TFT-Q-C	**65.5% (0.006)**	57.1% (0.230)
TFT-Q-G	RF-P-L	36.9% (0.021)	**78.6% (<0.001)**
N-BEATS-Q-G	RF-P-L	44.0% (0.326)	69.0% (<0.001)

5.2 Calibration, Efficiency, and Heterogeneity

Figures 3 and 4 summarize calibration and the cost–accuracy trade-off across model families, while Table 5 and Figure 5 provide ATM-level comparisons.

In Figure 3, we observe near-nominal calibration for the leading quantile models. TFT-Q-G closely tracks the diagonal across quantiles; LightGBM-Q-G achieves accurate marginal coverage; N-BEATS-Q-G slightly under-covers in the middle but remains well calibrated at $q = 0.99$. At this extreme level, TFT-Q-G reaches 98.5% coverage, while N-BEATS-Q-G and LightGBM-Q-G both achieve 98.1%. Yet identical tail coverage does not imply similar cost. Despite matching $q = 0.99$ coverage, N-BEATS-Q-G and LightGBM-Q-G differ materially in total inventory cost. We attribute this gap to differences in sharpness and cross-ATM dispersion: safety stock depends on upper-tail geometry, not nominal coverage alone.

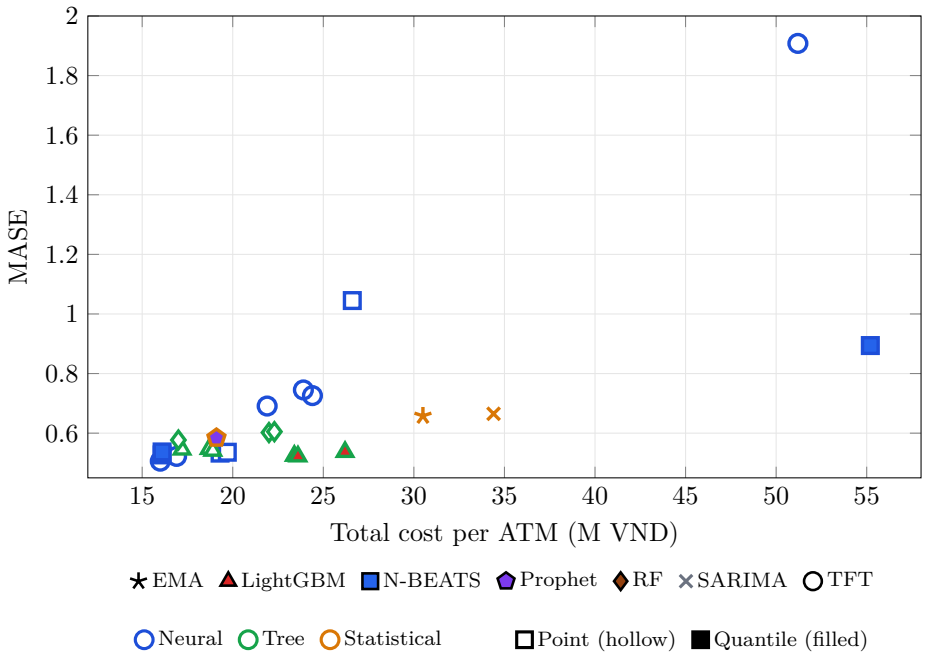

Fig. 4: Cost–accuracy trade-off across all 24 model configurations. Architecture is encoded by marker color and shape, model family by outline color, and mode by fill state (hollow: point; filled: quantile). The lower-left region is preferable.

We next examine the cost–accuracy frontier (Figure 4). Neural quantile models cluster in the lower-left region of the MASE–cost plane, defining the empirical efficiency frontier. TFT-Q-G delivers the strongest joint result, combining the lowest MASE (0.506) with the lowest mean cost (16.0 M VND), with N-BEATS quantile variants close behind. Tree-based quantile models attain competitive MASE but incur higher cost, while statistical baselines remain dominated in both dimensions. This confirms a divergence between accuracy and cost: similar MASE values can yield substantially different operational outcomes.

At the ATM-level, we observe pronounced heterogeneity (Table 5). Although TFT-Q-G ranks first in average cost, it is cheaper than N-BEATS-Q-G on only 40.5% of ATMs, with no significant cost dominance, despite a clear MAE advantage (67.9%, $p < 0.001$). The dispersion around the equal-cost diagonal in Figure 5 indicates that no single model uniformly dominates.

Taken together, we draw three conclusions. First, quantile training moves neural architectures toward the efficiency frontier. Second, architecture remains decisive, with neural quantile models leading, tree-based models intermediate, and statistical baselines lagging. Third, business performance depends jointly on calibration, sharpness, and heterogeneity. Fleet-level averages mask

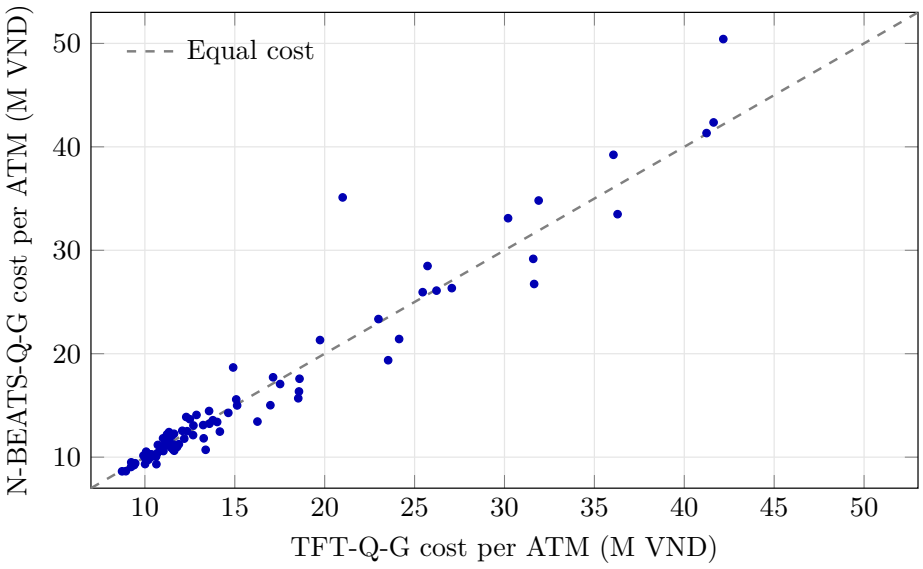

Fig. 5: Per-ATM cost comparison of TFT-Q-G and N-BEATS-Q-G across 84 ATMs. Each point represents one ATM. Points below the diagonal indicate ATMs where N-BEATS-Q-G yields lower cost, while points above the diagonal indicate ATMs where TFT-Q-G is cheaper. The broad spread around the equal-cost line highlights strong ATM-level heterogeneity and confirms that neither model uniformly dominates across locations.

substantial cross-ATM variation, and average cost leadership does not imply universal optimality.

6 Discussion and Limitations

Our empirical results reveal a structural gap between statistical accuracy and operational performance under inventory control. Across 30 configurations, global neural quantile models define the empirical efficiency frontier, delivering the strongest joint trade-off between point accuracy and total cost. TFT-Q-G achieves both the lowest MASE and the lowest mean cost, while N-BEATS-Q-G attains similar cost with slightly weaker central accuracy. These findings indicate that pooled neural architectures effectively exploit cross-location information to stabilize distributional forecasts and improve safety-stock decisions.

More broadly, the results clarify the role of uncertainty modeling in inventory systems. Quantile forecasting improves cost efficiency for neural models in pooled settings by providing actionable upper-tail signals for safety-stock decisions. However, uncertainty estimation alone is not sufficient. Tree-based quantile models such as LightGBM-Q-G achieve competitive calibration and

the lowest CRPS, yet incur substantially higher total cost. Nominal coverage and aggregate probabilistic scores therefore do not directly translate into economic efficiency. Operational performance depends on distributional sharpness and cross-location heterogeneity, specifically how probability mass is allocated in the upper tail, rather than on average calibration alone.

A second structural finding concerns heterogeneity. Although TFT-Q-G achieves the lowest average cost, ATM-level win rates show no universal dominance. Cost advantages vary across locations, and models with weaker point accuracy can outperform stronger ones in specific regimes. This divergence between accuracy and cost indicates that model selection for inventory control should prioritize business outcomes over forecast error alone.

These results motivate an adaptive deployment perspective. Rather than relying on a single global model, practitioners may monitor ATM-level performance and periodically reassign among leading candidates. Such an approach preserves fleet-level robustness while remaining responsive to local demand dynamics.

Sensitivity to cost parameters. Our results are computed under a fixed cost structure (Table 1), where the ratio $c_s/(c_s + c_h) \approx 0.99$ implies a high service-level target. If a bank adopts a more conservative risk appetite by raising the shortage penalty c_s, models with sharper upper-tail forecasts—such as TFT-Q-G—would gain an even larger cost advantage, as the penalty for under-stocking would amplify the benefit of well-calibrated quantile estimates. Conversely, reducing c_s relative to c_h would narrow the gap between quantile and point-forecast models, since holding cost would dominate and conservative safety stock would become more costly. The cost-breakdown analysis in Figure 2 already suggests this pattern: models with near-zero shortage cost (for example, LightGBM-Q-G) pay disproportionately in holding cost, indicating sensitivity to the cost ratio. A formal sensitivity analysis varying these parameters remains an important direction for future work.

Resilience to demand shocks. ATM withdrawals are subject to sudden spikes from local events, holidays, or shifts in consumer behavior. The quantile forecasting approach provides inherent resilience against such demand shocks: by targeting the ρ^*-quantile (approximately the 99th percentile), the base-stock policy maintains a safety buffer calibrated to the upper tail of the forecast distribution. The strong fill rates observed for neural quantile models ($\geq$99%) confirm this protective effect over the test period. However, truly unprecedented events—such as a pandemic-driven withdrawal surge or prolonged ATM malfunction—lie outside the training distribution and may exceed the model's learned tail behavior. Complementary operational safeguards, such as dynamic trigger rules or manual overrides, would be necessary to handle such extreme scenarios.

Sharpness versus calibration. An important distinction emerges between calibration and sharpness of predictive intervals. LightGBM-Q-G achieves the lowest CRPS and near-perfect coverage yet incurs substantially higher total cost than TFT-Q-G. This gap arises because LightGBM-Q-G produces wide prediction

intervals: while marginal coverage is accurate, the 99th-percentile forecasts are overly conservative, leading to excessive safety stock and high holding cost. TFT-Q-G, by contrast, produces sharper intervals that concentrate probability mass more tightly around realized demand, yielding tighter order-up-to levels that reduce holding cost without sacrificing fill rate. This finding underscores that operational efficiency depends on the shape of the forecast distribution—particularly upper-tail sharpness—not merely on aggregate calibration or coverage.

Drivers of heterogeneity. The ATM-level analysis reveals that no single model dominates across all locations, but the current study does not attribute this heterogeneity to specific ATM characteristics. Plausible drivers include differences in demand volatility, seasonality complexity (for example, ATMs near commercial districts versus residential areas), and sensitivity to holiday effects. Neural models may excel at high-volatility ATMs where cross-location information sharing stabilizes forecasts, while simpler models may suffice for ATMs with stable, predictable patterns. Stratifying performance by demand regime—such as volatility quantiles or location type—would provide actionable guidance for model assignment and constitutes a valuable direction for future research.

Limitations. We note several limitations. First, we assume a fixed periodic-review base-stock policy to isolate the effect of forecasting from decision optimization. While this separation clarifies attribution, alternative policies, such as dynamic or service-level-constrained optimization, could alter relative rankings. Second, recorded withdrawals may understate true demand at times when an ATM runs out of cash, since unmet demand is not observed. Without daily on-hand inventory records, this demand censoring cannot be corrected, and both forecast evaluation and simulated costs may differ from outcomes under uncensored demand. This is a common limitation in ATM cash management studies. Third, we evaluate a single institutional dataset of 84 ATMs. We focus on this setting to ensure consistent operational constraints and cost parameters; however, broader geographic, seasonal, or macroeconomic regimes may affect generalizability. Third, we compute downstream cost under fixed hyperparameters and retraining schedules to maintain comparability across models. More frequent updates or adaptive safety-stock rules may interact differently with distributional forecasts and change performance differentials. Finally, we do not conduct controlled ablations to isolate architectural components such as attention, gating, or basis expansions. Given the large design space and computational cost, our analysis remains comparative rather than causal, and architectural interpretations should be viewed as suggestive rather than definitive.

Taken together, the results show that calibrated probabilistic deep models can materially improve operational efficiency. At the same time, calibration, point accuracy, and cost remain distinct dimensions that do not align automatically. Bridging these dimensions, especially under heterogeneity and policy constraints, remains an important direction for future research.

7 Conclusion

We evaluate a forecast-then-optimize framework for ATM cash replenishment in Vietnam, benchmarking 30 configurations across statistical, tree-based, and neural models. Three findings emerge. First, global neural quantile models (TFT-Q-G, N-BEATS-Q-G) deliver the strongest cost–accuracy trade-off, reducing simulated total cost by roughly half relative to classical baselines while maintaining 99% fill rates. Second, forecast accuracy and aggregate calibration are imperfect proxies for business performance: models with strong CRPS can still incur high cost when upper-tail forecasts are poorly shaped, underscoring the need for decision-centric evaluation. Third, pronounced ATM-level heterogeneity argues against uniform deployment and supports monitoring-based reassignment based on realized cost.

Several limitations qualify these results. We evaluate a single 30-day holdout from 84 ATMs of one Vietnamese bank, omit routing and joint replenishment constraints, and keep policy parameters fixed; rolling-origin tests across broader settings would strengthen external validity.

Future work will pursue several directions. First, end-to-end or decision-focused learning—where the forecasting model is trained directly on the inventory cost function rather than prediction loss—offers a promising path to bridge the forecast-optimize gap identified in this study. Second, a systematic sensitivity analysis varying cost parameters and service-level targets would clarify the robustness of model rankings across different bank risk appetites. Additional extensions include integrating routing and network constraints, testing robustness under regime shifts, stratifying performance by ATM demand characteristics, and validating across broader cash networks and markets.

Acknowledgments. We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arora, N., Saini, J.K.R.: Approximating Methodology: Managing Cash in Automated Teller Machines using Fuzzy ARTMAP Network. International Journal of Enhanced Research in Science Technology & Engineering **3**(2), 318–326 (Feb 2014)
2. Asad, M., Shahzaib, M., Abbasi, Y., Rafi, M.: A Long-Short-Term-Memory Based Model for Predicting ATM Replenishment Amount. In: 2020 21st International Arab Conference on Information Technology (ACIT). pp. 1–6. IEEE, Giza, Egypt (Nov 2020). https://doi.org/10.1109/ACIT50332.2020.9300115
3. Bergmeir, C., Hyndman, R.J., Koo, B.: A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis **120**, 70–83 (2018)
4. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: KDD Workshop (1994)

5. Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: PAKDD (2013)
6. CedolĭN, M., Orhan, D., GenevoĭS, M.: Statistical and Artificial Intelligence Based Forecasting Approaches for Cash Demand Problem of Automated Teller Machines. Academic Platform Journal of Engineering and Smart Systems **12**(1), 21–27 (Jan 2024). https://doi.org/10.21541/apjess.1360151
7. Ekinci, Y., Lu, J.C., Duman, E.: Optimization of ATM cash replenishment with group-demand forecasts. Expert Systems with Applications **42**(7), 3480–3490 (May 2015). https://doi.org/10.1016/j.eswa.2014.12.011
8. Ekinci, Y., Serban, N., Duman, E.: Optimal ATM replenishment policies under demand uncertainty. Operational Research **21**(2), 999–1029 (Jun 2021). https://doi.org/10.1007/s12351-019-00466-4
9. Embrechts, P., Nešlehová, J., Wüthrich, M.V.: Additivity properties for Value-at-Risk under Archimedean dependence and heavy-tailedness. Insurance: Mathematics and Economics **44**(2), 164–169 (Apr 2009). https://doi.org/10.1016/j.insmatheco.2008.08.001
10. FOREX Bank AB: Forex travel cash index. https://www.forex.se/en/travel/forex-index/cash-index/ (2025), https://www.forex.se/en/travel/forex-index/cash-index/, approximate shares of consumer payments made in cash by country, compiled from central-bank, payment-survey, and international statistics
11. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (Oct 2016). https://doi.org/10.48550/arXiv.1506.02142, https://arxiv.org/abs/1506.02142
12. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association **102**(477), 359–378 (2007). https://doi.org/10.1198/016214506000001437
13. Gökçay, D.E.: ATM Cash Stock Prediction Using Different Machine Learning Approaches. M.sc. thesis, Sabanci University, Istanbul, Turkey (Dec 2020)
14. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts (2018)
15. Jadwal, P.K., Jain, S., Gupta, U., Khanna, P.: Clustered Support Vector Machine for ATM Cash Repository Prediction. In: Pati, B., Panigrahi, C.R., Misra, S., Pujari, A.K., Bakshi, S. (eds.) Progress in Advanced Computing and Intelligent Engineering, vol. 713, pp. 189–201. Springer Singapore, Singapore (2019). https://doi.org/10.1007/978-981-13-1708-8_18
16. Kamini, V., Ravi, V., Kumar, D.N.: Chaotic time series analysis with neural networks to forecast cash demand in ATMs. In: 2014 IEEE International Conference on Computational Intelligence and Computing Research. pp. 1–5. IEEE, Coimbatore, India (Dec 2014). https://doi.org/10.1109/ICCIC.2014.7238399
17. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems 30 (NeurIPS 2017). pp. 3146–3154. Long Beach, CA, USA (2017)
18. Koenker, R., Bassett, G.: Regression Quantiles. Econometrica **46**(1), 33 (Jan 1978). https://doi.org/10.2307/1913643
19. Lim, B., Arık, S.Ö., Loeff, N., Pfister, T.: Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting **37**(4), 1748–1764 (Oct 2021). https://doi.org/10.1016/j.ijforecast.2021.03.012

20. López Lázaro, J., Barbero Jiménez, Á., Takeda, A.: Improving cash logistics in bank branches by coupling machine learning and robust optimization. Expert Systems with Applications **92**, 236–255 (Feb 2018). https://doi.org/10.1016/j.eswa.2017.09.043
21. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: The m4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting **36**(1), 54–74 (2020)
22. McInnes, L., Healy, J., Astels, S.: Hdbscan: Hierarchical density based clustering. Journal of Open Source Software **2**(11), 205 (2017)
23. Oreshkin, B.N., Carpov, D., Chapados, N., Bengio, Y.: N-BEATS: Neural basis expansion analysis for interpretable time series forecasting (2019). https://doi.org/10.48550/ARXIV.1905.10437, https://arxiv.org/abs/1905.10437
24. Rafi, M., Wahab, M.T., Khan, M.B., Raza, H.: Towards optimal ATM cash replenishment using time series analysis. Journal of Intelligent & Fuzzy Systems **41**(6), 5915–5927 (Dec 2021). https://doi.org/10.3233/JIFS-201953
25. Riabykh, A., Suleimanov, I., Surzhko, D., Konovalikhin, M., Ryazanov, V.: ATM Cash Flow Prediction Using Local and Global Model Approaches in Cash Management Optimization. Pattern Recognition and Image Analysis **32**(4), 803–820 (Dec 2022). https://doi.org/10.1134/S1054661822040113
26. Suder, M., Gurgul, H., Barbosa, B., Machno, A., Lach, Ł.: Effectiveness of ATM withdrawal forecasting methods under different market conditions. Technological Forecasting and Social Change **200**, 123089 (Mar 2024). https://doi.org/10.1016/j.techfore.2023.123089
27. Taylor, S.J., Letham, B.: Forecasting at scale. The American Statistician **72**(1), 37–45 (2018). https://doi.org/10.1080/00031305.2017.1380080
28. Thanh, B.T., Van Tuan, D., Chi, T.A., Van Dai, N., Dinh, N.T.Q., Thuy, N.T., Hoa, N.T.X.: Multiobjective Logistics Optimization for Automated ATM Cash Replenishment Process (Jul 2023), https://arxiv.org/abs/2304.13671
29. Tin Kam Ho: Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. vol. 1, pp. 278–282. IEEE Comput. Soc. Press, Montreal, Que., Canada (1995). https://doi.org/10.1109/ICDAR.1995.598994
30. Vangala, S., Vadlamani, R.: ATM Cash demand forecasting in an Indian Bank with chaos and deep learning. Expert Systems with Applications **211**, 118645 (Jan 2023). https://doi.org/10.1016/j.eswa.2022.118645, https://doi.org/10.1016/j.eswa.2022.118645
31. Venkatesh, K., Ravi, V., Prinzie, A., Poel, D.V.D.: Cash demand forecasting in ATMs by clustering and neural networks. European Journal of Operational Research **232**(2), 383–392 (Jan 2014). https://doi.org/10.1016/j.ejor.2013.07.027
32. Vishwakarma, V., James, H., Bururu, R.K., Matto, J.: ATM Cash Replenishment with Clustering Series (2020), https://www.researchgate.net/publication/341255204
33. Zeinalkhani, M., Ghanbar Tehrani, N., Pasandideh, S.H.R., Pedram, M.M.: Providing a forecasting model and optimization of the cash balance of bank branches and ATMs with the approach of social responsibilities. International Journal of Nonlinear Analysis and Applications **15**(10) (Oct 2024). https://doi.org/10.22075/ijnaa.2023.31335.4613
34. Zipkin, P.H.: Foundations of Inventory Management. McGraw-Hill (2000)