Multi-Level Reinforcement Learning for Dynamic Assortment Control in a Cosmetics Retail Network

Robertas Jurkus, Vitalijus Denisovas, Darius Drungilas, Mindaugas Kurmis

Klaipėda University, Faculty of Marine Technologies and Natural Sciences
Corresponding author: robertas.jurkus@ku.lt

Klaipeda University

Research Context and Questions

Large retail networks must continuously adapt product assortments to changing demand, limited shelf capacity, and substantial differences between stores. With tens of thousands of products, training a single reinforcement learning model across all store–SKU combinations becomes computationally expensive, while independent store models may lose knowledge shared across the retail network.

We therefore investigate a **multi-level reinforcement learning approach based on Proximal Policy Optimization (PPO)**, where a global policy is progressively adapted to store-size groups and individual stores.

Research questions:

- **RQ1** How can reinforcement learning be scaled to large retail assortment spaces?
- **RQ2** How does **global → group → store** policy adaptation affect assortment performance?
- **RQ3** Can the learned policies adapt to local demand while maintaining **stable assortment decisions**?

Method and Multi-Level RL Workflow

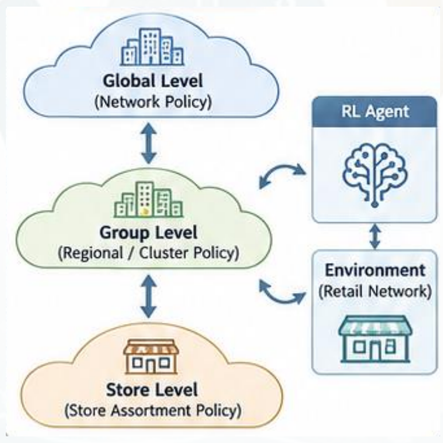
Progressive specialization: A global PPO policy is trained on all stores and the full network-wide SKU universe, then transferred and fine-tuned for store-size groups and individual stores.

Global PPO → Store-Size Group PPO → Store PPO
$\theta^{(o)} \rightarrow \theta^{(g)} \rightarrow \theta^{(l)}$

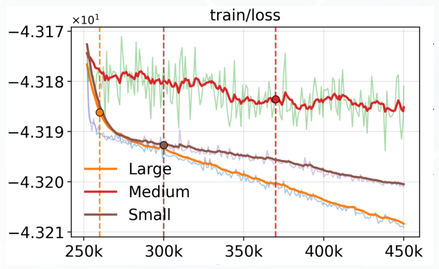
This preserves shared retail knowledge while adapting the policy to increasingly local conditions.

Scalable decision space: Store-level decisions reduce the per-step space from $O(|K| \times |L|)$ to $O(|K|)$.

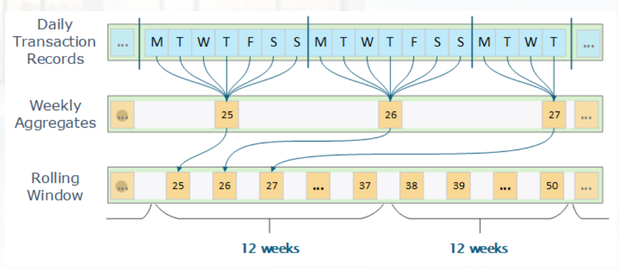
PPO ranks SKUs by priority; the highest-ranked products are selected subject to store capacity and operational constraints, including mandatory products, assortment persistence, and the 12-week lock.



$\theta^{(o)}$, $\theta^{(g)}$, $\theta^{(l)}$: global, group, store PPO parameters;
K: SKUs;
L: stores.



PPO Training Loss (TensorBoard)



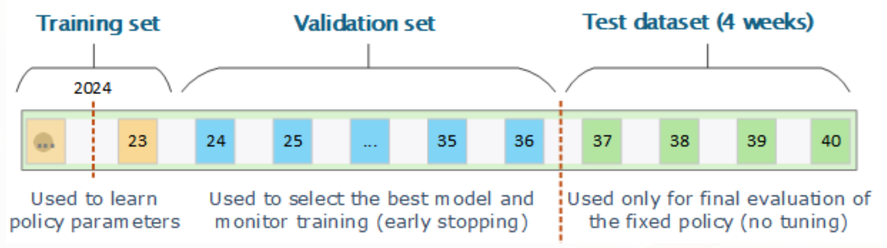
Data Aggregation

Real-world ERP Data

- Anonymized transaction and inventory data from a Lithuanian cosmetics and personal care retail network.
- 26 stores: 4 large, 12 medium, 10 small; 35,786 Stock Keeping Unit (SKUs) across multiple product categories.
- Historical period: January 2022 – September 2024.
- Daily ERP records included sales, prices, stock levels, product, store, and date information.
- Inactive records with zero stock and zero sales were removed.
- Daily data were aggregated into weekly store–SKU observations, combining weekly sales, average prices, and beginning/end-of-week stock.
- A 12-week rolling window was used to derive recent demand, profit, margin, and stock features.

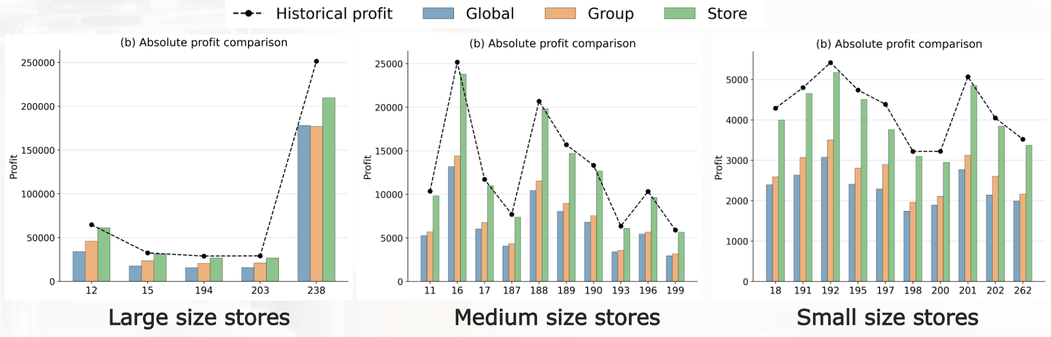
Discussion/Conclusions

- [RQ1] Scalability: Global → group → store specialization reduces model complexity.
- [RQ2] Adaptation: Store-level PPO achieved the closest historical profit recovery.
- [RQ3] Stability: Policies remained adaptive while limiting unnecessary assortment changes.
- Overall: Multi-level PPO combines **scalability, local adaptation, and stable assortment control.**



Results

- Evaluation: Profit recovery improved with policy specialization (test dataset):
  Global 55.53% → Group 63.73% → Store 93.94%
- Store-level PPO provided the closest approximation to historical profit across store-size groups.
- Repeated training: Four PPO runs with different seeds tested whether the observed improvement was consistent.
  Global 54.4 ± 1.0% → Group 71.7 ± 1.1% → Store 94.9 ± 0.5%
- The global → group → store ordering remained consistent, with low store-level variability.



Large size stores | Medium size stores | Small size stores

Profit recovery across store-size groups



Repeated-training robustness

Acknowledgment

This work was supported by the R&D project "Developing a coherent system for promoting innovative activities" (No. 02-020-K-0040), implemented by UAB "Inherent Simplicity Baltic" in partnership with Klaipeda University (KU) during 2024-2026.

References

[1] Meisheri et al., "Scalable Multi-Product Inventory Control with Lead Time Constraints," Neural Comput. Appl., 2022.
[2] van Hezewijk et al., "Using PPO for the Stochastic Capacitated Lot Sizing Problem," Int. J. Prod. Res., 2023.
[3] Ming et al., "Cooperative Modular RL for Large Discrete Action Space Problems," Neural Networks, 2023.
[4] Li et al., "Deep RL for Online Assortment Customization: A Data-Driven Approach," Prod. Oper. Manag., 2026.
[5] R. Jurkus, D. Drungilas, V. Denisov, Ž. Lukošius, S. Gucaga, and M. Kurmis, "A Multi-Level Deep Reinforcement Learning Approach for Dynamic Assortment Control in Large-Scale Retail Networks," IEEE Access, 2026.