A Bandit-Based Reinforcement Learning Simulated Annealing for the Omnichannel Last-Mile Distribution

Vincent F. Yu[1], Renan S. Maglasang[1], Aldy Gunawan[2], Ching-Hsuan Lin[1], Kuan-Fu Chen[3], and Shih-Wei Lin[4]

[1] Department of Industrial Management, National Taiwan University of Science and Technology, Taipei 106, Taiwan, {vincent,renan,m10901001}@ntust.edu.tw
[2] School of Computing and Information Systems, Singapore Management University, Singapore, Singapore, aldygunawan@smu.edu.sg
[3] College of Intelligent Computing, Chang Gung University, Taoyuan 333, Taiwan, kfchen@mail.cgu.edu.tw
[4] Department of Information Management, Chang Gung University, Taoyuan 333, Taiwan, swlin@mail.cgu.edu.tw

Abstract. Omnichannel last-mile distribution requires vehicles to manage deliveries, returns, and inventory across multiple fulfillment sources. In omnichannel retail, a central warehouse replenishes stores, customer orders are fulfilled from stores or warehouses based on inventory, and returns may be collected during delivery routes. This study introduces the Vehicle Routing Problem in Omnichannel Distribution with Simultaneous Pickup and Delivery (VRPOSPD), covering delivery-only, pickup-only, and simultaneous pickup-and-delivery operations under store-only and hybrid store–depot fulfillment scenarios. A Bandit-Based Reinforcement Learning Simulated Annealing (BBRLSA) algorithm with adaptive neighborhood selection and reward-based learning is proposed to solve the problem. Computational results show that BBRLSA outperforms conventional simulated annealing in terms of solution quality, robustness, scalability, and CPU time particularly under hybrid fulfillment settings.

Keywords: Vehicle routing problem · Omnichannel last-mile distribution· Reinforcement learning · Simulated annealing.

1 Introduction

E-commerce growth has increased the complexity of last-mile omnichannel distribution beyond traditional systems [2]. Vehicles must handle deliveries, returns, and dynamic loads from stores or depots, creating routing–inventory interdependencies that challenge classical VRPs [4]. While pickup-and-delivery models provide a foundation [3], most assume paired requests [7] and overlook shared fulfillment nodes. Recent studies incorporate store-based fulfillment and returns [6], but routing and inventory decisions are still often treated separately.

We introduce the VRP in Omnichannel Distribution with Simultaneous Pickup and Delivery (VRPOSPD), involving a depot, stores, and customers with heterogeneous requests under pooled inventory. Two scenarios are studied: store-only (M1) and hybrid store–depot (M2). To solve this tightly coupled problem, we propose a Bandit-Based Reinforcement Learning Simulated Annealing (BBRLSA) framework that improves adaptive neighborhood selection using magnitude-based rewards, recent-biased learning, and ε-greedy exploration [5].

Building on the omnichannel last-mile distribution setting of Abdulkader et al. [1], where warehouses replenish stores and online orders are fulfilled based on inventory availability, the VRPOSPD extends the setting by incorporating customer returns and simultaneous pickup-and-delivery operations. Consider a network $\mathcal{N} = \{0\} \cup \mathcal{R} \cup \mathcal{C}$, where node 0 denotes a central depot, $\mathcal{R}$ the set of retail stores, and $\mathcal{C}$ the set of online customers. A homogeneous fleet of vehicles k is dispatched from the depot to carry out delivery and pickup operations within the same route. Routing decisions are whether vehicle k travels directly from node i to node j or not. Retail stores—and sometimes the depot—hold pooled inventories of product types $p \in \mathcal{P}$ and act as fulfillment nodes. Inventory is homogeneous within each type and not preassigned to customers; pickups represent replenishment rather than specific requests. Customers are divided into pickup-only $\mathcal{C}^{(1)}$, delivery-only $\mathcal{C}^{(2)}$, and simultaneous pickup-and-delivery $\mathcal{C}^{(3)}$ sets. Deliveries specify product type and quantity, while returns increase vehicle load. For each delivery customer, a fulfillment node f is chosen. Two scenarios are considered: Model M1 restricts fulfillment to retail stores ($f \in \mathcal{R}$), whereas Model M2 allows fulfillment from stores or the depot ($f \in \mathcal{R} \cup 0$). Inventory feasibility at fulfillment nodes and consistency between routing and sourcing decisions are enforced. Vehicles may replenish multiple times along a route, but each store can be visited at most once; when store s is visited, the vehicle collects the aggregated delivery quantities for all customers assigned to s and served later on that route. In both scenarios, routes start and end at the depot, each customer is visited exactly once, and flow conservation and fulfillment–customer precedence are maintained. The VRPOSPD aims to minimize total routing cost, measured by total vehicle travel distance, while satisfying routing, fulfillment, inventory, precedence, and capacity constraints.

2 The Proposed Algorithm

BBRLSA extends simulated annealing by incorporating a bandit-based adaptive operator selection mechanism. To generate neighboring solutions, BBRLSA uses three operators: insertion, swap, and inversion. Insertion relocates a randomly selected customer within or across routes, swap exchanges two customer nodes, and inversion reverses a subsequence within a route. Due to store replenishment and fulfillment precedence constraints in VRPOSPD, all generated solutions are checked for fulfillment consistency, store precedence, inventory feasibility, and vehicle capacity. Infeasible solutions are handled via penalization within the simulated annealing framework. During the search, one neighborhood operator is selected at each iteration using an ε-greedy reinforcement learning mechanism.

All operators are initially assigned equal probabilities, which are then updated adaptively based on historical rewards from improvements they generate.

The algorithm starts with an initial solution, which is treated as both the current solution and the best-known solution, together with an initial temperature. Each neighborhood operator is initialized with a performance estimate and usage count to facilitate adaptive learning throughout the search process. At each temperature level, a fixed number of neighborhood evaluations is performed. During each iteration, a neighborhood operator is selected using an ε-greedy policy. With probability 1-ε, the operator with the highest estimated performance is selected (exploitation), whereas with probability ε, a neighborhood operator is selected randomly (exploration). This mechanism balances search intensification and diversification while enabling the algorithm to progressively favor operators that historically produce higher-quality improvements.

The selected operator generates a candidate solution, which is evaluated using a penalized objective function. Acceptance follows the simulated annealing criterion: improving solutions are accepted deterministically, while worsening solutions may still be accepted with a probability that depends on the temperature and the magnitude of deterioration. This allows the algorithm to escape local optima. After generating a neighboring solution, a reward is assigned based on the relative improvement in the penalized objective function. Operators that achieve larger reductions receive higher rewards, while non-improving moves receive zero reward. Operator quality is updated via incremental averaging, allowing the algorithm to favor consistently effective operators while maintaining exploration through the ε-greedy mechanism. The temperature is reduced geometrically at each level, and the algorithm terminates after a fixed number of consecutive non-improving temperature levels. Overall, BBRLSA combines adaptive learning and simulated annealing to efficiently guide the search toward high-quality routing solutions.

3 Preliminary Results

Table 1 compares the performance of ESA/SA and BBRLSA on large-scale store-fulfillment (M1) and hybrid-fulfillment (M2) VRPOSPD instances. The instances are based on large VRPO instances [1], which include a single warehouse, customers, demand quantities, store stock levels per product, and the x–y coordinates of each node. These instances were modified to incorporate simultaneous pickup-and-delivery and return operations, with 10–25 stores and 25–150 customers. Customer demand is generated randomly as pickup-and-delivery, pickup-only, or delivery-only. Each scenario contains 60 instances classified into low-, medium-, and high-inventory levels. The Average Percent Gap ($\%Gap_{Ave}$) measures the solution quality relative to the best-known solution (BKS) values obtained from ten runs per instance:

$$\%Gap_{Ave} = \frac{1}{n}\sum_{i=1}^{n} \%Gap_i, \quad \%Gap_i = \frac{Obj_i^{ave} - Obj_i^{BKS}}{Obj_i^{BKS}} \times 100$$

The Success Rate represents the percentage of runs that exactly attain the BKS:

$$SR = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(Z_i^{best} = Z_i^{BKS}\right) \times 100$$

where n is the total number of instances, Z_i^{best} is the best objective value obtained by the algorithm for instance i, Z_i^{BKS} is the best-known solution value for instance i, and $\mathbb{I}(\cdot)$ is an indicator function equal to 1 if $Z_i^{best} = Z_i^{BKS}$ and 0 otherwise. Both SA and BBRLSA were implemented in Microsoft Visual C++ 2019 and executed on a desktop equipped with an Intel(R) Xeon(R) CPU E3-1245 v6 @ 3.70 GHz and 64 GB of RAM under Windows 11. The reported $\%TimeDiff$ measures the relative computational time difference between BBRLSA and SA under the same computing environment, where a negative value indicates that BBRLSA requires less computational time than SA.

Results show that M2 instances achieve lower average percentage gaps and higher success rates than M1 for both algorithms, suggesting that hybrid fulfillment increases routing flexibility. BBRLSA also shows stronger improvements over SA in M2. On average, BBRLSA reduces the gap by 0.50 percentage points in M2 but slightly worsens it by 0.05 in M1. Improvements in M1 are small and instance-dependent, whereas M2 shows more consistent gains, especially in low- and medium-inventory groups, with gap reductions of up to 0.5 percentage points and success rate increases up to 20%. In terms of CPU time, BBRLSA is 8.967% times faster than SA in M1 and 7.351% times faster than SA in M2 as indicated by the negative $\%TimeDiff$ values.

	%Gap-Ave		Success Rates (%)		CPU Time (BBRLSA, s)			
	SA	BBRLSA	SA	BBRLSA	Min	Max	Ave	%TimeDiff
M1 instances								
Overall Performance	**3.39**	**3.44**	**53.33**	**60.00**	**14.2**	**420.9**	**150.4**	**-8.967**
M1:01–20 instances (low)	5.12	5.35	50.00	55.00	13.6	424.3	153.5	-8.790
M1:21–40 instances (medium)	3.03	3.02	50.00	70.00	14.3	407.4	150.2	-8.934
M1:41–60 instances (high)	2.02	1.96	60.00	55.00	14.7	431.0	147.3	-9.176
M2 instances								
Overall Performance	**2.45**	**1.95**	**60.00**	**71.67**	**21.9**	**543.6**	**201.4**	**-7.351**
M2:01–20 instances (low)	2.29	1.65	60.00	80.00	17.9	551.7	198.4	-8.108
M2:21–40 instances (medium)	2.64	1.98	55.00	70.00	19.3	578.6	200.9	-7.132
M2:41–60 instances (high)	2.41	2.21	65.00	65.00	28.4	500.6	205.0	-6.813

Table 1. Performance comparison of SA and BBRLSA.

4 Conclusion

This paper introduces the Vehicle Routing Problem in Omnichannel Distribution with Simultaneous Pickup and Delivery (VRPOSPD), integrating routing, fulfillment, and loading decisions beyond classical models. We examine store-only and hybrid store–depot scenarios to assess sourcing flexibility, and propose a Bandit-Based Reinforcement Learning Simulated Annealing (BBRLSA) algorithm with adaptive operator selection. Results show BBRLSA outperforms standard SA in solution quality, robustness, and scalability, and CPU time especially in hybrid settings, with statistical tests confirming consistent improvements.

References

1. Abdulkader, M., Gajpal, Y., ElMekkawy, T.Y.: Vehicle routing problem in omnichannel retailing distribution systems. International Journal of Production Economics **196**, 43–55 (2018)
2. Arslan, B., Schrotenboer, A.H., Atan, Z.: Optimal fulfillment and replenishment for omnichannel retailers with standard shipping contracts. European Journal of Operational Research **323**(2), 642–656 (2025)
3. Berbeglia, G., Cordeau, J.F., Gribkovskaia, I., Laporte, G.: Static pickup and delivery problems: A classification scheme and survey. TOP **15**(1), 1–31 (2007)
4. Hübner, A.H., Kuhn, H., Wollenburg, J., Towers, N., Kotzab, H.: Last mile fulfilment and distribution in omni-channel grocery retailing: a strategic planning framework. International Journal of Retail Distribution Management **44**(3) (03 2016)
5. Kahalimoghadam, M., Thompson, R.G., Rajabifard, A.: Self-adaptive metaheuristic-based emissions reduction in a collaborative vehicle routing problem. Sustainable Cities and Society **110**, 105577 (2024)
6. Mandal, P., Basu, P., Saha, K.: Forays into omnichannel: An online retailer's strategies for managing product returns. European Journal of Operational Research **292**(2), 633–651 (2021)
7. Parragh, S.N., Doerner, K.F., Hartl, R.F.: A survey on pickup and delivery problems. Journal für Betriebswirtschaft **58**(1), 21–51 (2008)