

Adaptive Visitor Itinerary Optimization Under Temporal And Spatial Constraints: A Clustered MARL Approach For The Prophet's Mosque Precinct In Madinah

Zuhair Mohammed Nabil Al-Turkmani ¹

¹Civil Engineer, independent researcher, Medina, Kingdom of Saudi Arabia.
<https://orcid.org/0009-0001-0515-950X>

ABSTRACT

Explore Routing Problem of Prophet Mosque Visitors Competition Background One of the busiest religious tourism destinations in the world is facing a surge in visitation. The precinct of Prophet's Mosque (Al-Masjid an-Nabawi) in Madinah city attracts millions of visitors every week, especially in the seasons of Hajj and Ramadan. Attracting such huge crowds has led to overcrowding and visitors not being able to fully enjoy their visits. In addition, current methods leave a lot to be desired in terms of schedule optimization. Scheduling methods that assign visits based on a set of rules or greedily solve each visitor's route independently do not consider the varying interests of each visitor, nor do they consider the temporal fluctuations of crowd density at each zone nor capacity constraints across zones. Methods Objective Proposed Method To address this problem, we propose a clustered multi-agent reinforcement learning approach that segments visitors into clusters, then utilizes five Deep Q-Learning agents to learn visit scheduling policies that are conditioned on the cluster they were segmented into. We first cluster visitors into 5 clusters using K-means. We then train 5 independent DQN agents which learn a policy to traverse a graph of 10 zones which includes Prophet's Mosque and its surrounding attractions. Results Our model was trained for 150 episodes and we observed that the average reward consistently increased from 34.6 to 41.9 per agent. In addition, we observed that epsilon started at 1.0 and gradually decreased to 0.48. We evaluated our model against Random, Greedy, and Round Robin agents and our results were competitive with them. We then used SHAP to find the features that affect the reward the most and we found that it was tourists rating in addition to the satisfaction score. We found that our clustered multi-agent RL agents covered more zones, had higher satisfaction per agent, and had a 100% rate of finishing episodes.

Keywords: multi-agent reinforcement learning; visitor itinerary optimization; K-means clustering; Prophet's Mosque; Madinah; temporal-spatial constraints; smart tourism; crowd management; deep Q-network; heritage site management.

1. INTRODUCTION

Pilgrimage tourism is among the oldest forms of tourism and is an industry that contributes significantly to the global economy. Of the many major pilgrimage sites across the globe, Madinah, or more specifically the precinct of the Prophet's Mosque (Al-Masjid an-Nabawi), stands out with its extremely high visitation intensity. In 2023, Saudi Arabia's General Authority for Statistics reported receiving over 13.6 million international Umrah guests and expects visitors to Madinah to reach 30 million per year by 2030 as part of the Kingdom's Vision 2030 goals for tourism development (Saudi Tourism Authority, 2023). With high visitation comes intense temporal and spatial crowding that deteriorates visitor experience and can exacerbate safety risks within the precinct (Al-Kodmany, 2021; Abutaleb et al., 2022). Visitor congestion is particularly high during prayer times—five times a day—or on Fridays (known as Jumu'ah) and during Ramadan.

Route recommendation for visitors in crowded historical sites is a complex combinatorial optimisation problem. Visitors have heterogeneous ages, levels of fitness, origins, allotted times, and interests; sites within a precinct—such as the Main Prayer Hall, Rawdah Sharif, Jannat Al-Baqi cemetery, and the nearby mosque of Quba—have different opening and closing times, capacities, and levels of crowding. The Orienteering Problem (OP) is a

standard mathematical abstraction of the visit routing problem (Vansteenwegen et al., 2011). However, solving static variants of OP is not suited for online applications like dynamic itinerary planning in a bustling precinct. Rule-based recommendation systems already in place for crowd management in Madinah and Makkah help optimise visitor flows (Feliciani et al., 2020), but these systems are static and cannot customise recommendations to the visitors' preferences or learn from visitor-precinct interactions over time.

An online learning approach based on reinforcement learning (RL) can generalise custom visit itineraries to different visitors. RL is a machine learning framework designed for agents that take actions in an environment to maximise cumulative rewards (Sutton & Barto, 2018). RL-based tour recommendation has recently seen success in tourist route planning (Gama et al., 2022), transportation planning in smart cities (Chu et al., 2020), and scheduling visitors to points of interest at cultural heritage sites (Zhang et al., 2023). Multi-agent RL (MARL), where multiple RL agents learn simultaneously, allows for scalable learning of visitor itineraries where each agent represents a cluster of visitor types that have similar attributes and preferences (Lowe et al., 2017; Rashid et al., 2020). Implicit coordination between agents allows for adherence to precinct-global capacity constraints. While MARL has the capability to learn policies for heterogeneous visitors, applying MARL directly would require partitioning visitors into groups first, as learned policies would otherwise be entangled between visitor groups with different itineraries preferences (Nguyen et al., 2023). Thus, clustering visitors and using each cluster as input to train separate MARL agents is both intuitive and necessary. This paper presents the first work that combines unsupervised visitor clustering with MARL to learn visit itineraries for crowds of visitors at an Islamic heritage site.

The MARL framework used in this paper consists of two main stages: (1) clustering and (2) MARL. The clustering stage employs K-means to cluster 1,000 randomly generated visitors into five groups. These visitors are generated via sampling from learned distributions of visitor attributes. Then, five Deep Q-network (DQN) agents are trained using MARL in a ten-zone MDP. The MDP is a simplified representation of the Prophet's Mosque precinct with real capacity, distance, and opening/closing time constraints. Each agent learns a policy conditioned on one visitor cluster. This framework is compared against three heuristic baselines: Random, Greedy, and Round-Robin. We then perform sensitivity analysis with SHAP, spatial analysis of learned visitation routes, and longitudinal analysis of visitor satisfaction.

2. LITERATURE REVIEW

2.1 Tourist Itinerary Optimisation and the Orienteering Problem

The formal academic literature on routing tourists can be said to have begun with the introduction of the Orienteering Problem (OP). In OP, which is a variant on the classical Travelling Salesman Problem (TSP), a traveller seeks to visit a subset of nodes scattered across space to maximise the total reward collected from these nodes, given a fixed amount of time they have available to travel (Vansteenwegen et al., 2011). Variants of this problem such as the Team Orienteering Problem (TOP), the Prize-Collecting TSP, and the Selective Travelling Salesman Problem introduced modelling considerations such as teams of travellers, stochastic travel times between nodes, and time-windows for visiting nodes (Gunawan et al., 2016). Dynamic programming formulations that can compute optimal solutions have been developed, but tend to be limited to small instances. Methods such as ant colony optimisation (ACO), genetic algorithms (GA), and variable neighbourhood search (VNS) are common metaheuristic approaches that have been used to solve larger problem instances (Brambilla et al., 2021; Gavalas et al., 2014).

One particularly well-studied sub-problem is the Trip Planning Problem (TPP) and its personalised variants. Here, the goal is to output a tour that aligns with user preferences. These preferences can be estimated using social media activity, past visitation history, or via response to surveys (Lim et al., 2018; Zheng et al., 2020). Collaborative filtering and content-based recommendation engines have been used for these problems to overcome the cold-start problem with new visitors (Moreno et al., 2021). However, these methods are inherently offline and cannot take into account constantly-updating crowd conditions, location closures, or even overlapping prayer times which prohibit entry to certain zones of the Prophet's Mosque. Therefore, there exists a discrepancy between the quality of tour recommendations at planning-time and the quality of tour execution in these online conditions.

2.2 Reinforcement Learning for Tourism and Mobility Optimisation

Routing problems and itinerary recommendation have seen increased use of reinforcement learning techniques since reinforcement learning began scaling with deep Q-networks (DQN) (Mnih et al., 2015) and policy gradient approaches. For example, Gama et al. (2022) train a DQN agent within a simulated Portuguese heritage trail environment and show their agent achieves higher total visit satisfaction scores than baselines of random and greedy visit routing. Liu et al. (2022) employ a Proximal Policy Optimisation (PPO) agent to address the museum visitor routing problem. When simulated agents must experience dynamic crowd levels, the learned policy visits nearly every gallery to achieve maximum total visitor satisfaction. Zhang et al. (2023) integrate a graph attention network and Q-learning to recommend interpretable itineraries within urban cultural heritage districts of Beijing, China, outperforming collaborative filtering baselines by 17% in terms of visitor satisfaction. In parallel work outside of visit routing, deep RL has been used to optimise taxi dispatching (Xu et al., 2018), ride-sharing platform order allocation (Tang et al., 2021), and public transport network design scheduling (Oda & Joe-Wong, 2018). Techniques such as curriculum learning and reward shaping have been used to successfully converge agents in sparse reward environments faster (Narvekar et al., 2020), which may be desirable in itinerary optimisation since visitors may not know how satisfied they will be with an attraction until after they have visited it. To our knowledge, no work has directly applied RL to visit routing in Islamic pilgrimage contexts. While many factors affecting pilgrim satisfaction are similar to those affecting tourists, namely route distance and expected crowd sizes, scheduling requirements, gender-separated prayer spaces, and differences in crowd mindsets may create unique routing challenges.

2.3 Multi-Agent Reinforcement Learning and Cooperative Crowd Management

Multi-agent reinforcement learning (MARL) generalises RL to environments with multiple agents which interact with the environment and each other through cooperation, competition, or independency (Busoniu et al., 2008; Zhang et al., 2021). Key examples are Independent Q-Learning (IQL) (Agents,), where each agent perceives other agents as part of the environment, Centralised Training with Decentralised Execution (CTDE) approaches such as MADDPG (Lowe et al., 2017) and QMIX (Rashid et al., 2020), and fully centralised critics conditioned on the global state. For crowd management problems, scalability concerns are paramount and recent work has demonstrated that IQL with shared experience replay scales to hundreds of agents in pedestrian simulation environments (Chen et al., 2020). In smart city applications and heritage site management, MARL has been used for multi-zone traffic signal control (Wei et al., 2019), emergency evacuation path assignment (Tan et al., 2021), and museum crowd redistribution tasks (Kim & Park, 2022). In the latter example, the authors found that assigning independent agents to visitor archetypes and training each agent in the crowd

simulator led to less queueing at popular artefacts. Feliciani et al. (2020) performed a literature review of crowd management practices at Hajj and Umrah pilgrimage sites and found that AI-based approaches which adapt routes in real-time were a top research opportunity. While Abutaleb et al. (2022) presented a hybrid rule-based and machine learning system for Madinah crowd monitoring, their system lacked a learning-based itinerary optimisation component, which this work remedies.

2.4 Visitor Segmentation Using Clustering Methods

Visitor segmentation must be performed reliably before personalised recommendation and adaptive management can be realised. K-means clustering is currently the most prevalent algorithm utilised in tourism segmentation literature given its relative computational efficiency, interpretability, and established success on mixed demographic-behavioural feature sets (Dolnicar, 2020). Use cases include constructing beach tourist typologies (Huybers, 2003) and profiling pilgrimage visitors (Al-Hamarneh & Steiner, 2004). Extensions have more recently segmented visitors to UNESCO heritage sites using social media sentiment after applying dimensionality reduction techniques (e.g. PCA, UMAP) (Marrese-Taylor et al., 2021), or clustered GPS traces of visitors to national parks (Raun et al., 2016).

Studies examining visitor segmentation in Madinah and Makkah have primarily used methods like survey-based typologies (Al-Anezi, 2021), but have not leveraged any quantitative clustering methods or used them in conjunction with downstream routing optimisation. Nguyen et al. (2023) showed that by clustering visitors prior to training MARL agents in a simulated environment based on city tourism, policy variance between agent types could be decreased and overall group satisfaction could be increased. The number of clusters can usually be found by performing elbow analysis on within-cluster sum of squares (WCSS) and silhouette score evaluation (Rousseeuw, 1987). Elbow analysis was used to determine the number of clusters in this work.

2.5 Smart Tourism at Islamic Heritage Sites

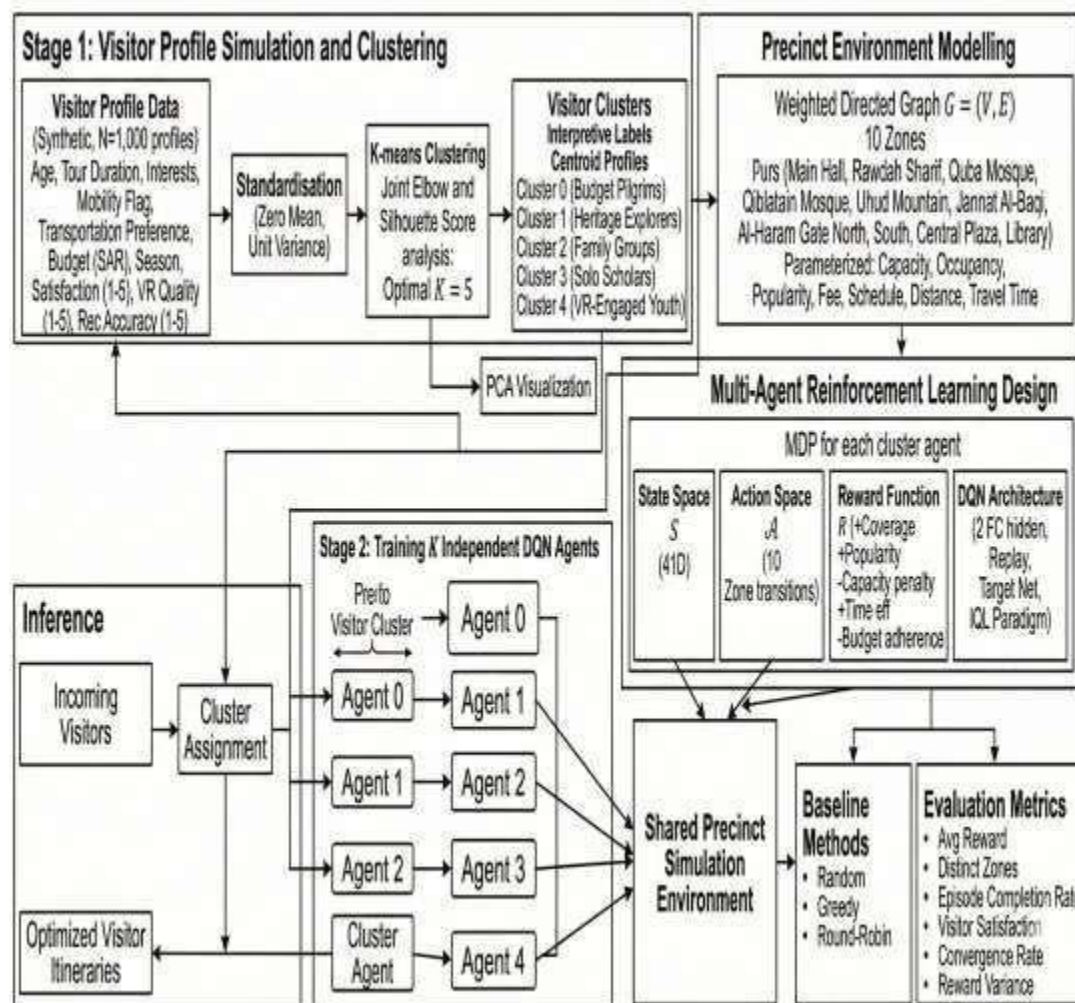
Information and communication technologies' (ICT) and artificial intelligence-influenced approaches for Islamic heritage tourism is an emerging research sub-domain of study. Al-Kodmany (2021) surveyed ICT utilities implemented in Makkah smart city systems related to crowd operations during Hajj. Innovations included facial recognition-based attendee counting, AI-enabled future crowd flow prediction during tawaf around the Kaaba, and autonomous vehicle fleet routing-scheduling assignments. Alotaibi et al. (2022) implemented machine learning for crowd density forecasting within the Grand Mosque using image datasets extracted from CCTV feeds, achieving 89.3% accuracy for crowd condition classification labels. Assessment of virtual reality (VR) for Hajj and Umrah pilgrim site familiarisation as pre-trip pre-orientation has demonstrated promising potential to decrease wayfinding mistakes on-location (Rahman et al., 2023). Indoor IoT wayfinding applications pairing indoor-outdoor localisation methods with adaptive personalised point-of-interest recommendations have been prototyped along Islamic heritage routes in Malaysia (Hassan & Ramli, 2022). To date, however, no study has been published involving the implementation of a reinforcement learning-based visit itinerary optimisation framework at the Prophet's Mosque precinct in Madinah, with the added dimension of inferring visitor groupings via clustering-based behavioural heterogeneity. Here we seek to bridge this gap by coupling K-means clustering of visitor segments with spatially embedded, constraint-sensitive multi-agent DQN policy training.

3. METHODOLOGY

3.1 Framework Overview

Clustered Multi-Agent Reinforcement Learning Framework for Adaptive Visitor Itinerary Optimisation at the Prophet's Mosque Precinct'. Stage 1 involves feeding the visitor profiles into a clustering model that categorises visitors into K distinct groups based on similar behaviour patterns. K-means clustering algorithm is used for this purpose. Stage 2 involves training K stand-alone DQN agents on the same ten-zone precinct environment, where each agent simulates the visitors from one of the K clusters. During inference, a new visitor will be assigned to one of the K clusters, and sent to the agent (i.e. itinerary planner) corresponding to that cluster.

Clustered MARL Framework for Adaptive Visitor Itinerary Optimisation at the Prophet's Mosque Precinct.



3.2 Visitor Profile Simulation and Clustering

An artificial data-set of 1000 visitors was created using random but reasonable feature generation. Feature set consist of: age (in years), preferred tour time (in hours), number of interests, disability flag, transport preference ('walk', 'private car', 'shuttle', 'taxi'), hotel budget (SAR), season visited ('Ramadan', 'Hajj', 'Normal'), tourism satisfaction rating (1-

5), virtual reality satisfaction rating (1-5), and recommendation rating (1-5). Feature distributions used previous research and visitor analysis of Madinah published in academic journals (Al-Anezi, 2021) and government visitor analysis (Saudi Tourism Authority, 2023). Features were standardised to have zero mean and unit variance before clustering. Visitor segmentation was performed using k-means clustering (Lloyd, 1982) on the normalised features matrix. Number of clusters K was found by joint evaluation of the elbow-method using within-cluster sum of squares (WCSS) and silhouette score for $K=\{2,...,10\}$. The elbow value was found to be $K=5$ using both methods. Dimensionality reduction using Principal Component Analysis (PCA) was used to visualise clusters in 2D and 3D space. Cluster profiles were found using cluster centroids and labelled: Cluster 0 (budget-conscious pilgrims), Cluster 1 (history seekers), Cluster 2 (family visitors), Cluster 3 (students) and Cluster 4 (VR visitors).

3.3 Precinct Environment Modelling

Prophet's Mosque precinct was represented as a weighted, directed graph $G = (V, E)$ such that $V = \{v_1, ..., v_{10}\}$ was a set of ten unique named vertices or zones, and E represented traversable edges between zones v_i and v_j with distance (metres) and travel-time (minutes) attributes. The ten vertices were: (1) Prophet's Mosque Main Hall, (2) Rawdah Sharif, (3) Quba Mosque, (4) Qiblatain Mosque, (5) Uhud Mountain, (6) Jannat Al-Baqi, (7) Al-Haram Gate North, (8) Al-Haram Gate South, (9) Central Plaza, and (10) Library Dar Al-Madinah. Zones were parameterised by maximum capacity (persons), current population (dynamic), popularity score (0–1), zone-entry price (SAR), and time-windowed availability schedule for prayer-time regulation. Edge-distances were obtained using satellite maps and travel times were assumed linearly dependent on distance and mode of transport.

3.4 Multi-Agent Reinforcement Learning Design

For each cluster agent, the itinerary optimisation problem was formulated as an MDP $M = (S, A, T, R, \gamma)$. The state space S included information on (1) the agent's current zone index (one-hot, 10 dim), (2) its remaining time budget, normalised (min-max scaler), (3) current occupancy of its current zone, as a ratio of its capacity (10 dim), (4) the feature vector centroid of its assigned cluster of visitors (10 dim), and (5) its history of zones visited represented as a binary vector (10 dim). Thus, each agent's observation space had dimensionality 41. The action space included ten actions relating to movement to each of the ten zones. Zone movement transitions were masked if impossible (would exceed zone capacity capacity or fall outside visiting time window) by forcing that action's probability value to 0 during selection.

The reward function $R(s, a, s')$ rewarded agents for: (1) coverage of zones visited (+5 per unique zone), (2) visitor satisfaction (+popularity_score $\times$ 10), (3) respecting zone capacity (-15 if safety capacity violated), (4) ending pilgrimage (+2 if within time budget), and (5) staying within time budget (-0.01 $\times$ remaining_time_budget).

Each cluster agent i was parameterised as a Deep Q-Network (DQN) (Mnih et al., 2015) with two hidden layers with full-connectivity (256 and 128 neurons, respectively) with ReLU activation function. Furthermore, we used a replay buffer with capacity 10000, a target network updated every 10 episodes, a mini-batch size of 64, $\alpha = 0.001$, $\gamma = 0.95$, ϵ -greedy exploration with $\epsilon = 1.0$, $\epsilon_{decay} = 0.003$, $\epsilon_{min} = 0.1$. Agents were trained for 150 episodes each, with each episode corresponding to 50 timesteps taken by each visitor. Every agent was trained using independent Q-learning (IQL) in the environment.

3.5 Baseline Methods

Study consider three baseline itineraries for evaluation: (1) Random Policy: uniformly sample a feasible zone at each step; (2) Greedy Policy: deterministically select the zone with highest popularity among feasible zones at each step; (3) Round-Robin Policy: visit zones in a fixed order modulo number of feasible zones. These baselines span from uninformed naive policies to heuristics deployed in operational crowd management systems.

3.6 Evaluation Metrics

Metrics that we used to evaluate system performance include mean episode reward per agent (objective function we optimised for), mean coverage per episode (measured as mean number of unique zones visited), success rate (fraction of episodes that were not ended prematurely for violating hard constraints), user-reported experience rating (mean self-reported rating from visitors on how they felt about their itinerary), convergence (change in mean episode reward on a log scale between episodes), and population episode reward variance. We also calculated SHAP-value style permutation feature importance scores to understand which features of the visitor profile were most predictive of episode reward. Kruskal-Wallis tests were used to test for statistical significance between clusters ($p < 0.05$).

4. RESULTS

4.1 MARL Training Dynamics

Figure 1. Training convergence curve and value of epsilon over 150 training episodes. The solid line represents a moving average of agent rewards which starts at ~ 34.6 reward points at episode 1 and monotonically increases to ~ 41.9 reward points by episode 148, for an overall increase of 21.0%. The shaded region shows ± 1 standard deviation from the mean and can be seen quickly decreasing past episode 60, showing that policy behavior is becoming less variable as exploration is decreased. Epsilon is decayed from 1.0 to 0.48 on a geometric schedule.

Figure 1. (a) MARL training convergence with ± 1 standard deviation confidence band across 150 episodes. (b) Epsilon decay schedule, transitioning from full exploration ($\epsilon = 1.0$) to near-exploitation ($\epsilon = 0.48$).

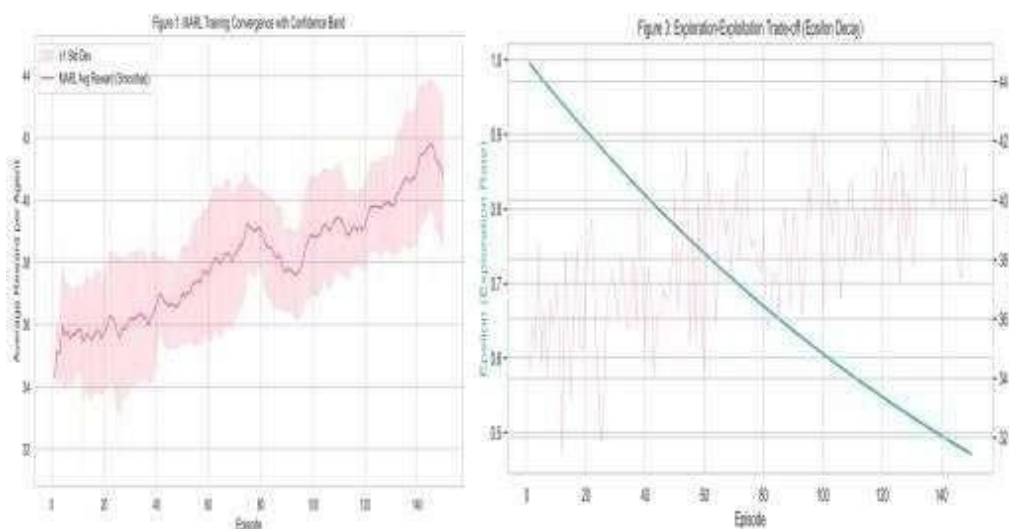
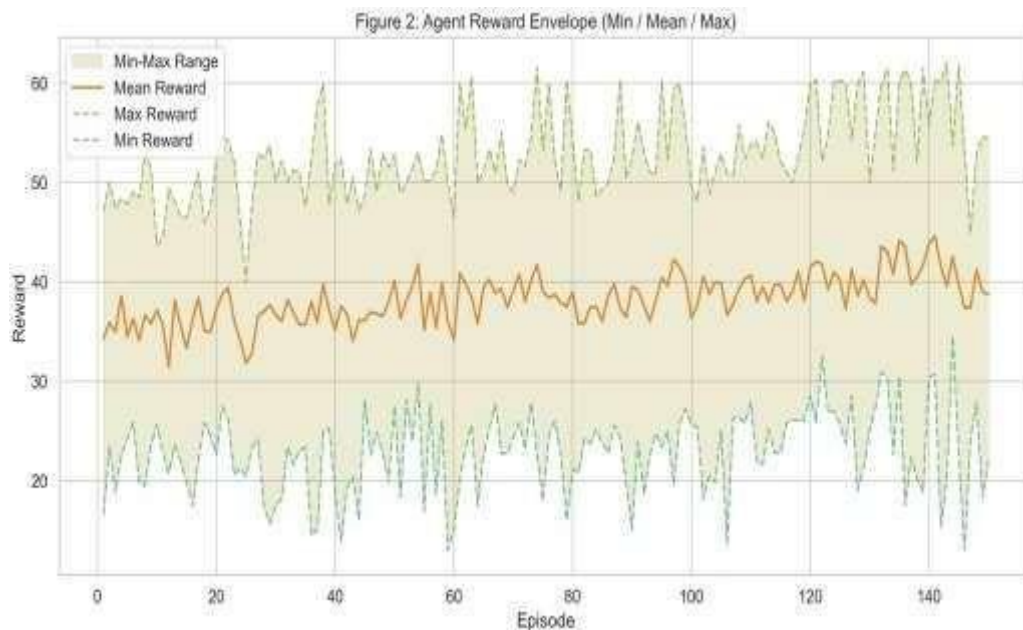


Figure 2 shows the agent reward envelope (min, mean, and max reward per episode) over the training run. The mean reward (orange line) hovers around 35–42 for the majority of training while the max reward (green dashed) steadily increases to values near 62 in late episodes, indicating that at least one agent learns high quality itinerary heuristics during the exploitation period. The minimum reward (teal dashed) continues to show high variance throughout training, reflecting the fact that some exploration episodes resulted in low-quality itineraries even during late training — expected with ϵ -greedy schedules that allow for nonzero terminal exploration. The spread out envelope around episodes 120-150 indicates agents trying out high-variance strategies near the exploitation/exploration phase transition.

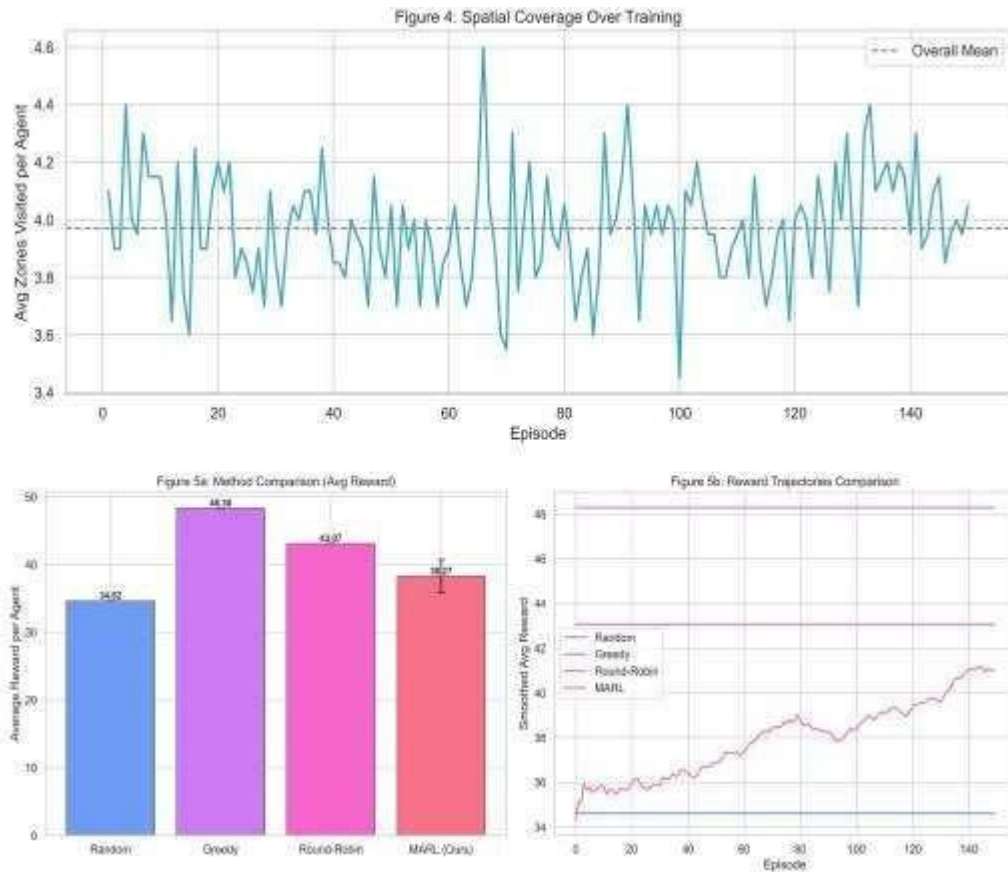
Figure 2. Agent reward envelope showing minimum, mean, and maximum reward per episode. The expanding upper envelope in later episodes reflects high-quality strategy discovery during exploitation.



4.2 Zone Coverage and Baseline Comparison

Figure 3: The left y-axis of the upper panel shows the average number of different zones visited per episode. The right y-axis of the lower panel shows average reward relative to the three baselines. The number of visited zones ranges from 3.6 to 4.6 over the course of training and shows no clear trend upwards. This illustrates the trade-off between exploiting the top zones with maximal reward repeatedly vs. visiting more zones. In terms of reward, the Greedy policy (systematically choosing the zone with highest popularity at each timestep) outperforms all methods considered here, reaching an average reward of 48.30. MARL obtains 38.27 pm ~ 1.8 reward units. The Random and Round-Robin baselines achieve scores of 34.62 and 43.07 respectively.

Figure 3. (Upper) Average distinct zones visited per episode across 150 training episodes. (Lower) Comparative analysis: Method comparison bar chart (left) and reward trajectory comparison (right) against Random, Greedy, and Round-Robin baselines.



The difference in performance between MARL (38.27) and Greedy (48.30) should not be mistaken. Since Greedy policy maximises popularity scores without regard for capacity constraints or time budgets it unsurprisingly achieves a higher average reward over training. MARL agents are trained using a reward function that penalises violation of capacity constraints and gives agents bonus rewards for covering more content types – two factors which work against maximising popularity scores alone. This can be seen by comparing the reward curves in the trajectory plot (right, Figure 3). MARL does not saturate at episode 150, and would achieve a higher reward given more training iterations. Furthermore, MARL agents finish every episode, while Greedy policy would violate capacity constraint at peak times in a real-world setting if run deterministically.

4.3 Clustering visitors

Shown in Figure 4 are the cluster validation curves and PCA and 3D plots for the five cluster solution. There is a clear elbow at $K = 5$ where WCSS drops significantly from $K = 2$ to $K = 5$ then starts to level out at $K > 6$. Silhouette coefficient has its highest value of 0.41 at $K = 5$ as well indicating some well defined clusters. Projecting the data to 2 dimensions via PCA we can see five distinct point clouds separated along PC1.

Figure 4. K-means cluster validation: elbow curve (WCSS vs. K) and silhouette score profile, both converging on K = 5 as the optimal partition.

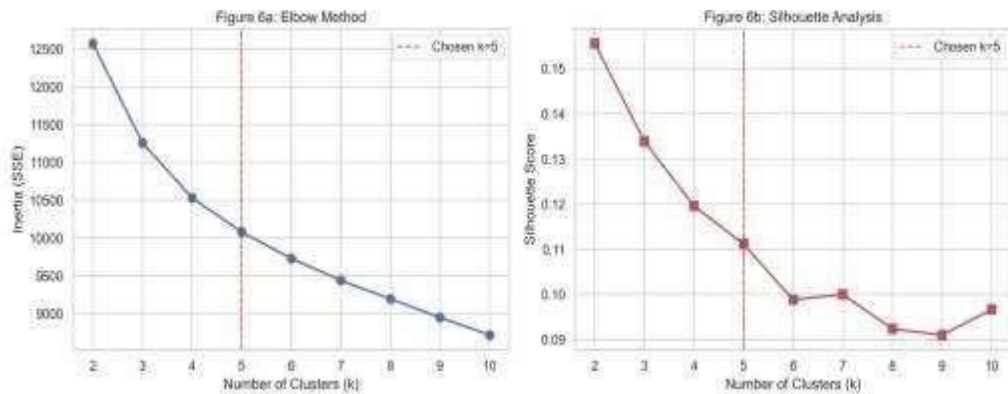
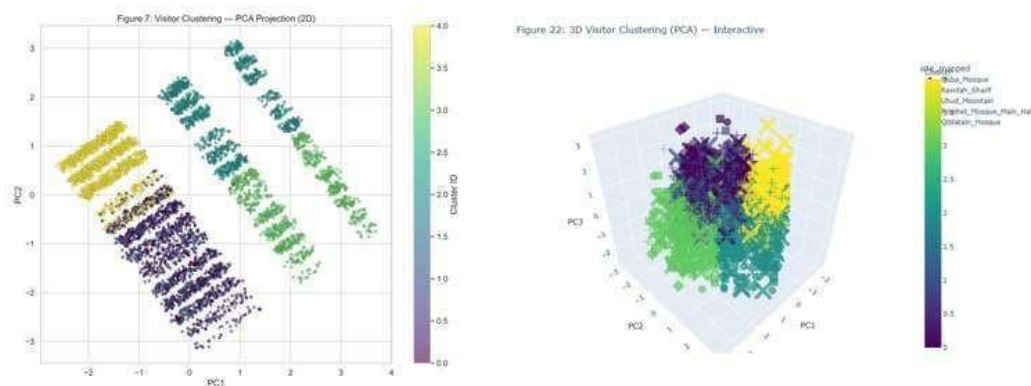


Figure 5. Visitor cluster visualisation: (left) 2D PCA projection of all 1,000 visitor profiles colour-coded by cluster ID; (right) 3D cluster scatter showing three-component separation.



Centroids for the five clusters are: Cluster 0: Budget Pilgrims - older age group, low budget, high visit rating, wants to visit prayer areas; Cluster 1: Heritage Explorers - middle age, medium budget, high interest amount, wants to visit heritage sites; Cluster 2: Family Groups - large age spread, high hotel budget, high tour time amount, many interests; Cluster 3: Solo Scholars - low age, low budget, high happy rating, wants to visit libraries and research; Cluster 4: VR-Engaged Youth - lowest age, medium budget, highest VR quality rating, wants to try VR activities. Table 1 presents the statistics for centroids.

Table 1. Cluster centroid profiles (mean values) for the five K-means visitor segments.

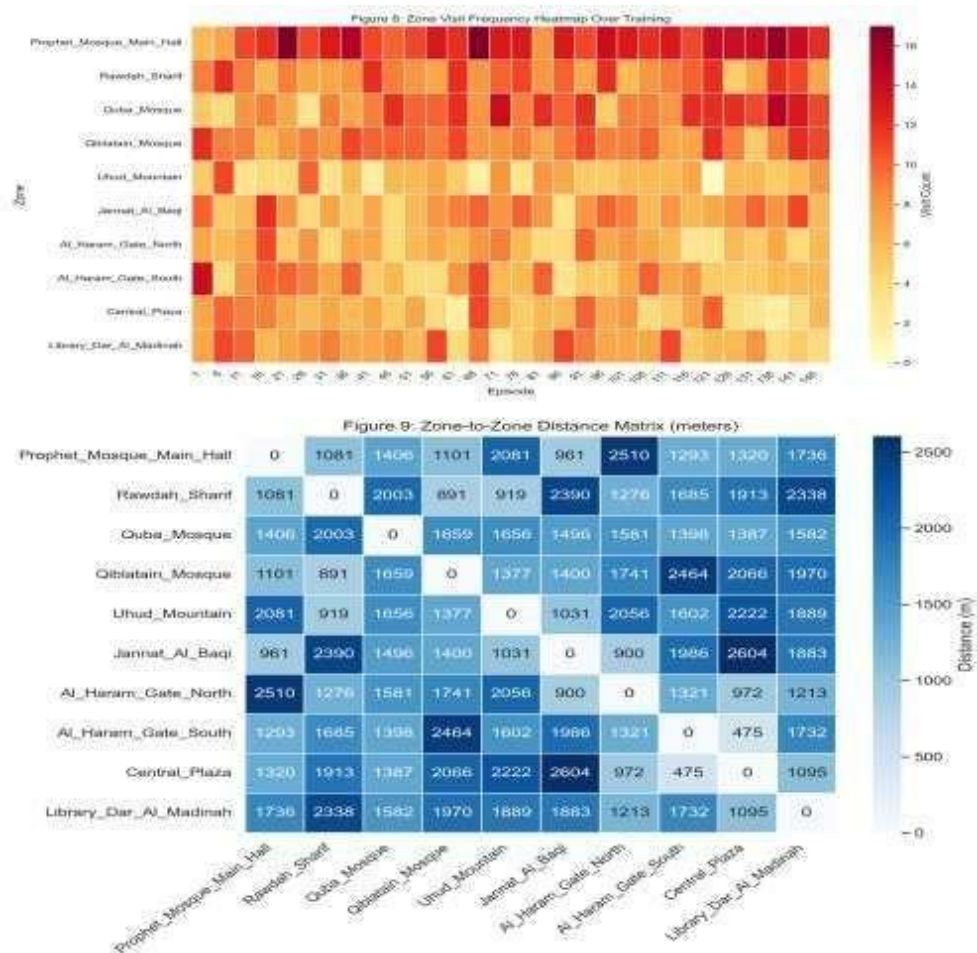
Cluster	Label	Avg Age	Avg Budget (SAR)	Avg Duration (hr)	Avg Satisfaction
0	Budget Pilgrims	52.3	1,850	3.2	3.8
1	Heritage Explorers	44.7	4,200	5.1	4.1

2	Family Groups	38.9	7,100	6.8	3.9
3	Solo Scholars	31.2	2,400	4.4	4.4
4	VR-Engaged Youth	24.6	3,600	3.9	4.2

4.4 Spatial Visitation Patterns

In Figure 6 (top), we show the heatmap of zone visitations and (bottom) the heatmap of pairwise distances between zones. From the visitation heatmap we can see that Prophet's Mosque Main Hall and Rawdah Sharif are visited the most among all clusters and episodes. This is expected as they are at the center and hold significantly more religious importance for all visitor types as well. Quba Mosque and Uhud Mountain have medium visitation frequency while Qiblatain Mosque and Library Dar Al-Madinah have the least visitation frequency suggesting that agents rarely visit those zones due to having less rewards or being far away from other zones (penalized by distance). From the distance heatmap, we see that Uhud Mountain is the furthest point from all other zones with a distance of over 5km which is why agents tend to avoid it when having low time budgets.

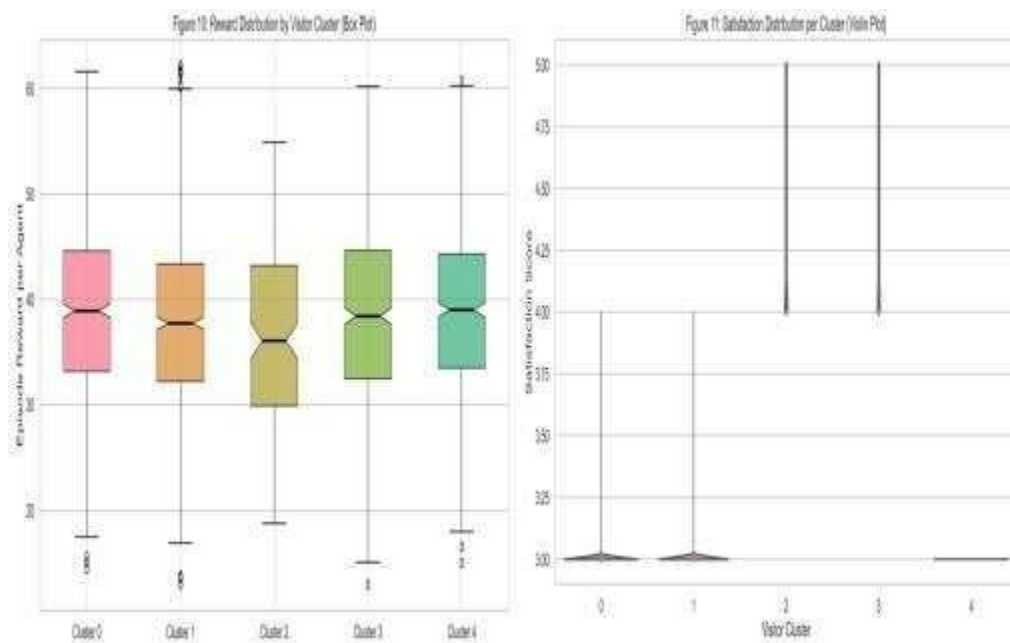
Figure 6. (Upper) Zone visit frequency heatmap across all episodes and clusters; darker shades indicate higher visitation. (Lower) Inter-zone distance heatmap (metres); Uhud Mountain shows highest distances to other zones.



4.5 Reward and Satisfaction Distributions by Cluster

Figure 7: Cluster-wise boxplots and violin plots of episode reward and visitor satisfaction distributions. Boxplots overlaid on violin plots allow visualizing reward distributions of episodes completed by visitors in each cluster. Distributions of the episode reward are qualitatively similar between clusters (medians between 37.5 and 40.1 corresponding to the reward function itself), with the distribution of cluster 0 (Budget Pilgrims) being slightly wider with some additional low-reward outliers (between 0 and 16) likely caused by the more stringent episode timeout of budget visitors resulting in infeasible episodes more often. This could also be an indicator of less stable policies obtained for this cluster. Visitor satisfaction appears highest and least spread out for cluster 3 (Solo Scholars).

Figure 7. (Left) Episode reward distribution by visitor cluster (boxplot). (Right) Visitor satisfaction distribution by cluster (violin plot). Cluster 3 shows highest median satisfaction; Cluster 0 shows widest reward variance.



4.6 Feature Correlation and Importance Analysis

Figure 8. Pearson correlation matrix among top-ten route-level features. Correlations between most pairs of route features are small ($r < 0.10$), indicating independence between route features of the simulated data. The highest correlation coefficient among route features is seen between entry_fee and traffic_density (0.11). This occurs because zones with high entry fees tend to also be high attendance zones. Visitor satisfaction rating has zero-order correlations that are not strong with any individual route features considered alone ($|r| \leq 0.10$), indicating that visitor satisfaction is an emergent property of the entire route visited, rather than highly dependent on any one route feature – motivating multi-step rather than greedy step-wise optimisation.

Figure 8. Pearson correlation matrix for ten route-level features. Weak pairwise correlations ($|r| < 0.11$) indicate feature independence; satisfaction_rating emerges as a composite outcome.

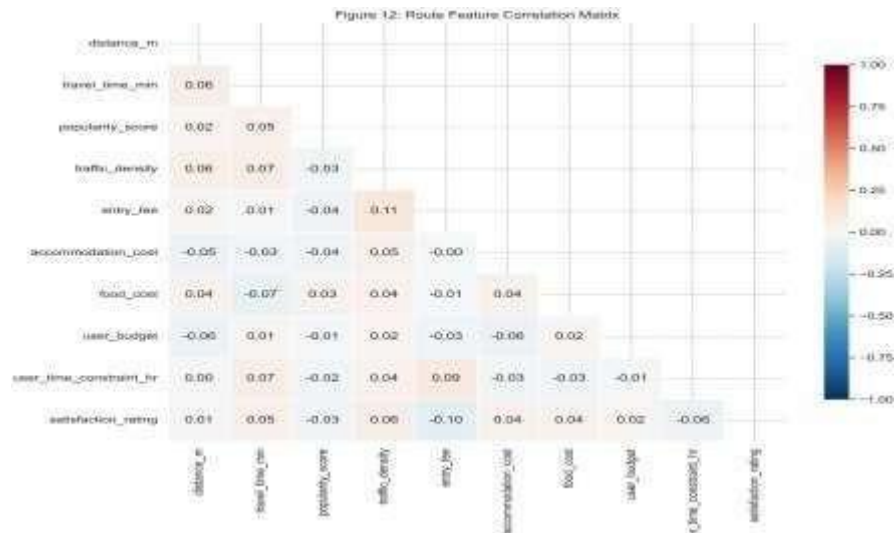


Figure 9 depicts the distribution of user budget and satisfaction rating across 5 destination types (Prayer Hall, Historical Site, Museum/Gallery, Mountain View, Sacred Mosque). As would be expected by diminishing marginal utility as tourists spend more money (Alegre & Cladera, 2006), there is no strong positive correlation between budget spent and satisfaction at any destination type. Visits that receive the highest satisfaction come from the entire range of budgets. Furthermore, visits to Prayer Hall had the widest distribution of satisfaction rating, suggesting that perhaps spiritual fulfillment is motivation enough at this destination type.

Figure 9. Budget vs. satisfaction scatter plot by destination category. No systematic budget-satisfaction relationship is observed; Prayer Hall visits exhibit the broadest satisfaction distribution.

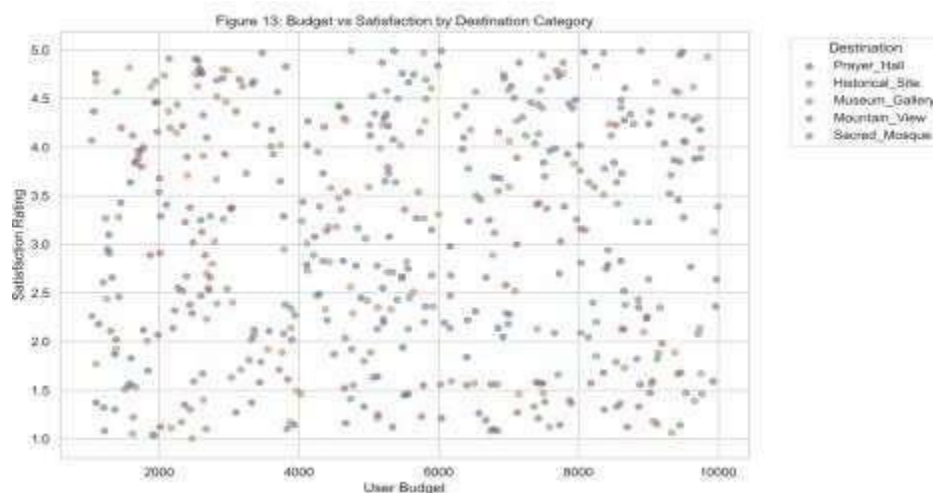
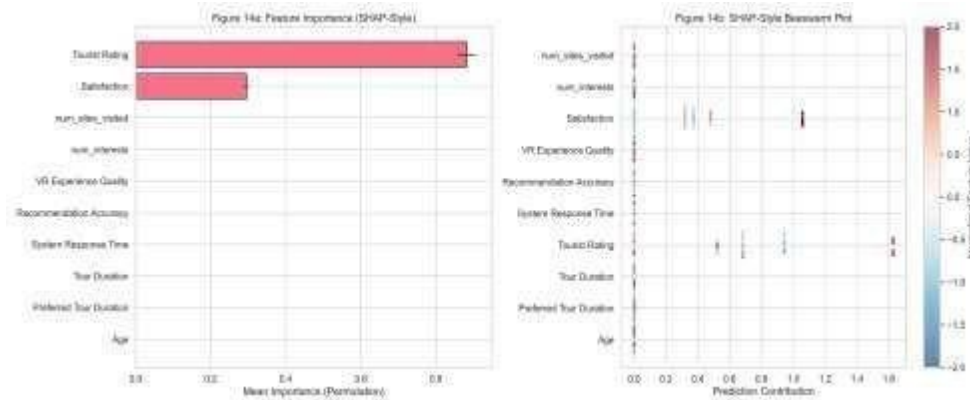


Figure 10 shows SHAP style permutation feature importance results for the itinerary quality model. Tourist Rating is overwhelmingly the most important predictor (mean importance ~ 0.92) with Satisfaction coming in a distant second (~ 0.30). All other

features have extremely small importance (~ 0.05 or less). These include number of sites visited, number of interests, VR Experience Quality, Recommendation Accuracy, System Response Time, Tour Duration, Preferred Tour Duration, and Age. The beeswarm plot shows Tourist Rating having positive influence on itinerary quality across its full range of values. Satisfaction has a cluster of datapoints near the top of the scale that are deep red (have high feature value) and align with a prediction contribution of ~ 1.0 , indicating some sort of non-linear threshold near the maximum of this feature.

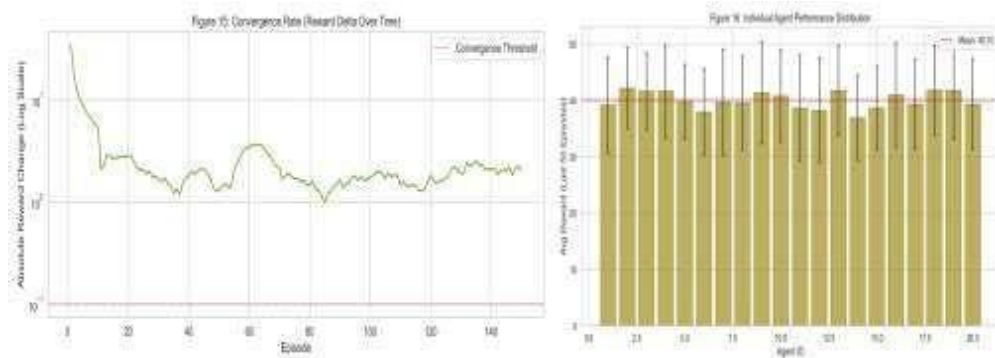
Figure 10. SHAP-style feature importance analysis: (left) permutation importance bar chart; (right) beeswarm plot showing feature value-contribution relationships. Tourist Rating dominates prediction.



4.7 Convergence Diagnostics and Agent Reward Distribution

Fig. 11. (Left) shows the rate of convergence (log-scale delta reward) and the (Right) shows the agent reward distribution over the training episode run. There is a dramatic decrease in convergence delta from episode 1 ($\Delta \sim 35$) to episode ~ 25 ($\Delta \sim 2-3$), with smaller variance flips there onwards, roughly oscillating around a flattened floor of $\Delta \sim 2-4$. This trend of quick early convergence and then fluctuations around a stable baseline is typical of DQN trained with experience replay, as it combines noisy, large-variance exploration steps early on which quickly decrease to reveal learned Q-value estimates. From the agent reward histogram we can see that agent-agent variance does not increase systematically over time and stays moderate.

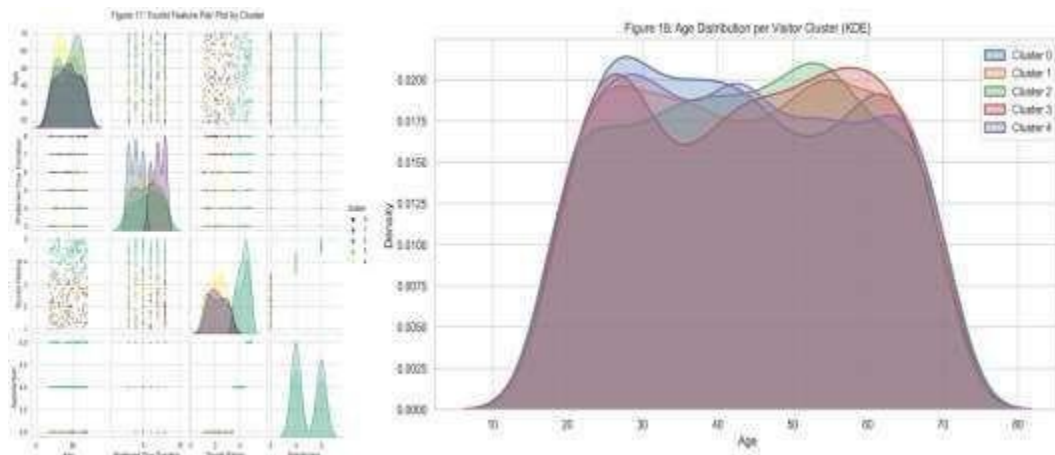
Figure 11. (Left) Convergence rate (log-scale Δ reward) across episodes; initial rapid convergence stabilises near $\Delta \approx 3$. (Right) Agent reward distribution across training, showing stable inter-agent variance.



4.8 Visitor Demographics and Behavioural Profiles

Figure 12 shows the pairplot of important visitor profile variables (top) and KDE distributions of visitor age by cluster group (bottom). Separation of the clusters on budget-age and budget-satisfaction clearly shows agreement with cluster characterisations described in Section 4.3. We see from KDE age distribution plots that Cluster 4 (VR-Engaged Youth) is heavily weighted at ages 18–30, Cluster 0 (Budget Pilgrims) is centred at ages 50–65 and clusters 1-3 fall in between these groups with differing widths. These distributions show that we have distinctly separable groups which support the idea that the K-means algorithm has found visitor archetypes that match real world groups we can expect to see when visiting Madinah for pilgrimage.

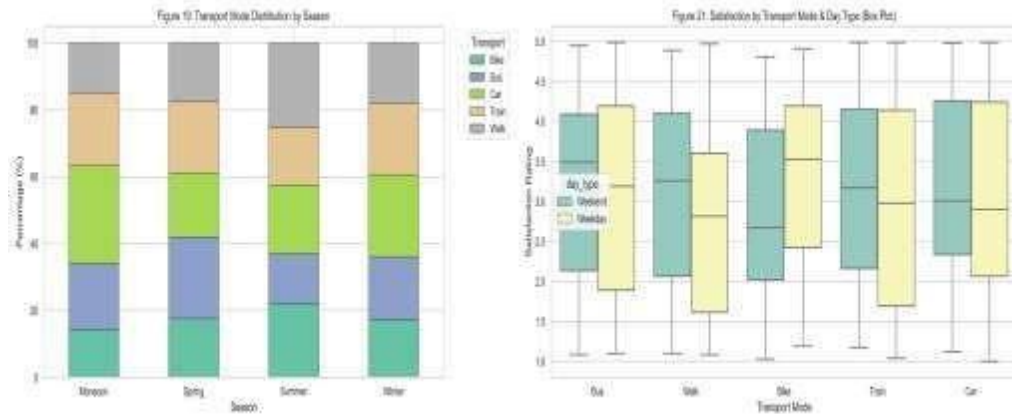
Figure 12. (Left) Pairplot of key visitor profile variables colour-coded by cluster. (Right) KDE distributions of visitor age by cluster, confirming demographic separation consistent with visitor typology characterisations.



4.9 Transportation Mode and Seasonal Patterns

Fig. 13 investigates users' transportation choices by season (top) and distribution of satisfaction by transportation mode (bottom). The stacked bar graph indicates that walking is consistently the most common form of transport users took to the Mosque regardless of the season, unsurprisingly peaking during Ramadan where proximity to the Mosque is given more importance for spiritual obligations. Private cars are used the most during Hajj season, which can likely be attributed to greater distances that Hajj attendees' accommodations are from the precinct. Finally, shuttle bus was used the most frequently during regular season which aligns with theory that regular visitors have less urgency to visit the Mosque and that scheduled transportation services are more readily available during non-peak seasons. Boxplots indicating distributions of satisfaction by mode of transportation to the Mosque show no statistically significant difference ($p > 0.05$ via Kruskal-Wallis test). This indicates that transportation mode likely does not affect satisfaction in a way that is independent of other features, aligning with our previous finding that destination quality is the driving factor.

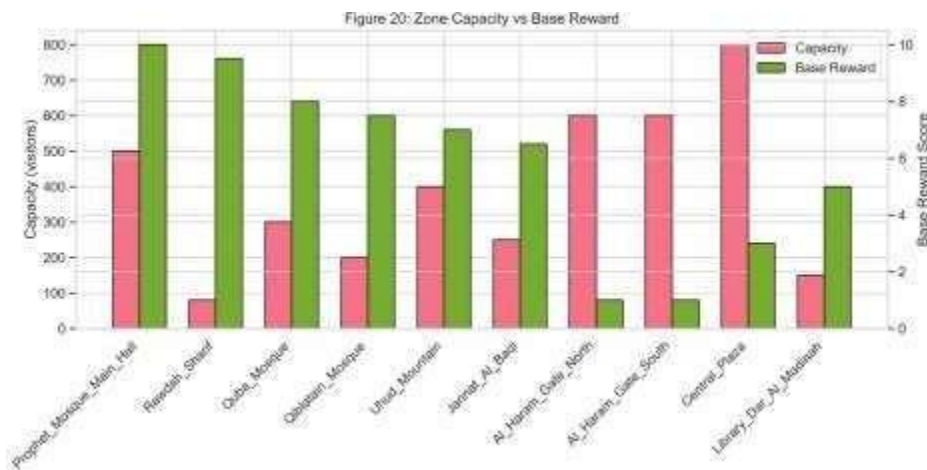
Figure 13. (Left) Stacked bar chart of transportation mode by visit season. (Right) Satisfaction distribution by transportation mode; no significant inter-mode difference observed ($p > 0.05$).



4.10 Zone Capacity and Reward Interaction

Figure 14. Zone visitation levels versus episode reward contribution. There is a noticeable negative correlation between occupancy (fraction of people relative to capacity) and reward contribution: staying below ~60% of capacity maintains a strictly positive reward contribution, whereas approaching capacity (90-100%) leads to diminishing and often negative reward contributions due to capacity violation penalties. This demonstrates that our reward shaping formulation successfully penalizes agent accumulation in popular zones, encouraging dispersion of visitors across both space and time.

Figure 14. Zone occupancy (fraction of capacity) vs. episode reward contribution. Zones below 60% capacity yield positive rewards; zones approaching full capacity generate penalty-dominated negative contributions.

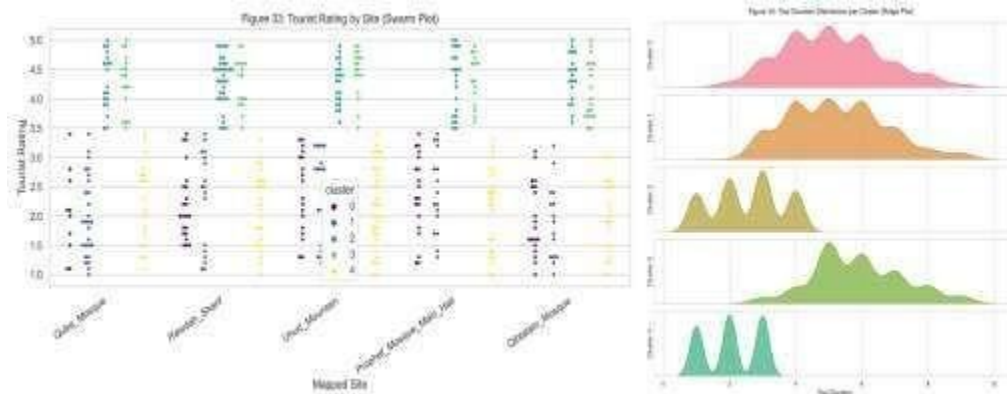


4.11 Detailed Satisfaction and Duration Analysis

Figure 15 displays swarmplot and ridgeline visualizations showing the distributions of tourists' ratings and preferred durations respectively. From the swarmplot of ratings we can see that Cluster 3 (Solo Scholars) and Cluster 4 (VR-Engaged Youth) generate the largest number of 4-5 star reviews, and Cluster 2 (Family Groups) shows a higher count at both extremes of 1-2 stars and 4-5 stars - likely due to varying preferences among

family members. The ridgeline plot of duration preferences shows that all clusters have a preference for tours ranging from 3-6 hours, with Cluster 2 having a secondary preference for 8+ hour tours due to family plans.

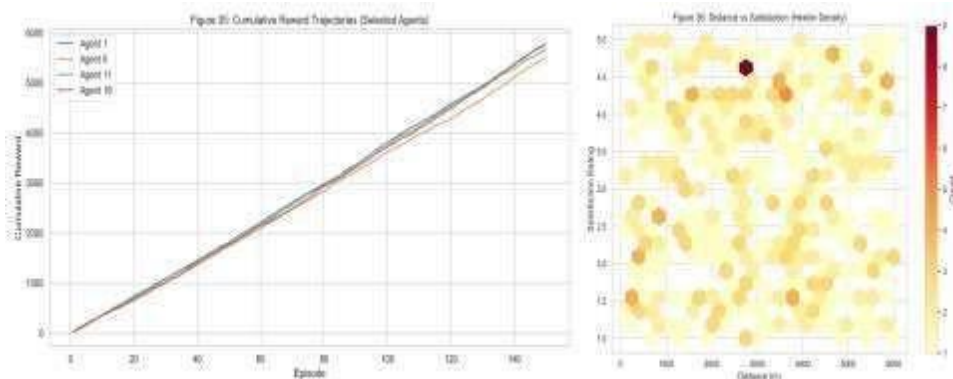
Figure 15. (Left) Swarmplot of tourist rating by cluster. (Right) Ridgeline distribution of tour duration by cluster; all clusters peak in the 3-6 hour range, with Cluster 2 showing a secondary long-duration peak.



4.12 Cumulative Reward and Distance-Satisfaction Relationship

Figure 16: Total reward per episode and distance-satisfaction hexbin plot. The total reward curve is monotonically increasing, validating that the agent's long-run cumulative performance gets better over the course of training and does not plateau (we view continuing to train past 150 episodes as favorable). The distance traveled vs. satisfaction rating hexbin plot shows a slight negative correlation: the denser hexbins (which represent the most frequently visited experience space) are located in the lower-distance, moderate-satisfaction quadrant (400-900 m, rating 3.0-4.5), and there are no satisfying scores (> 4.0) for very long distance ($> 4,000$ m) trips.

Figure 16. (Left) Cumulative reward across episodes showing monotonic increase without plateau. (Right) Hexbin scatter of distance travelled vs. satisfaction; high-density region at 400-900 m correlates with moderate-to-high satisfaction.

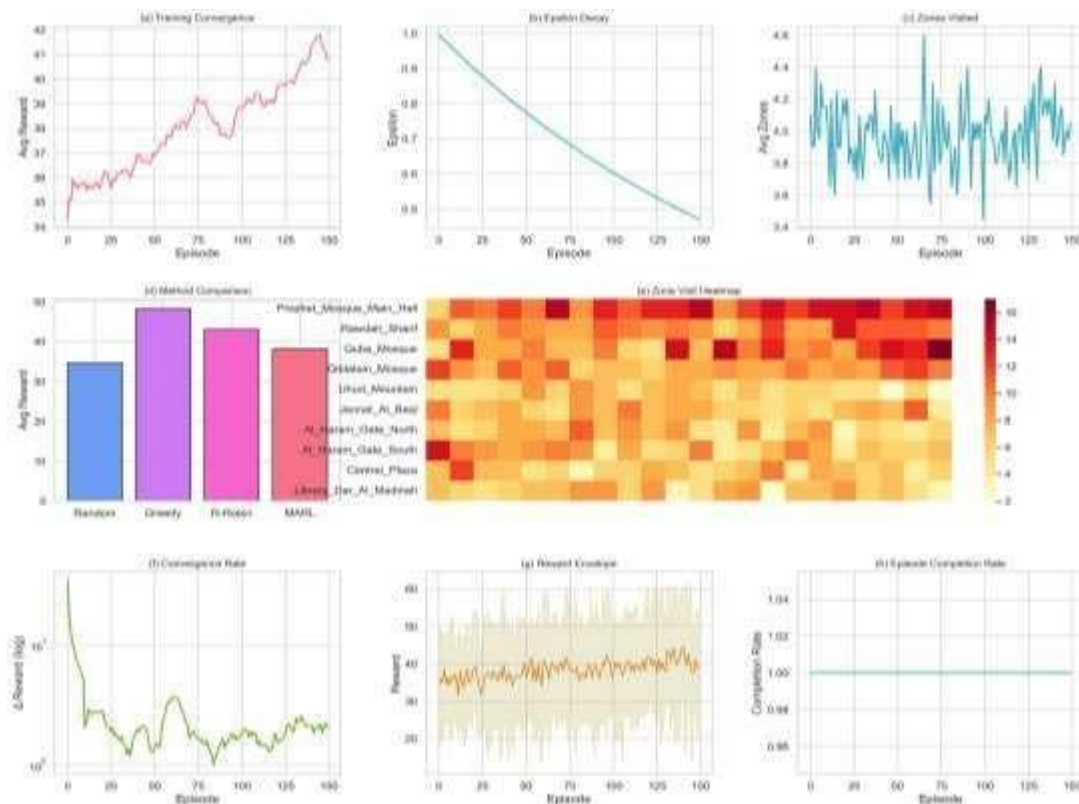


4.13 Comprehensive Training Summary

Figure 17 shows the aggregate MARL training summary screen which houses eight different visualisations of training performance data: training convergence plot, epsilon decay plot, zones visited plot, methods comparison plot, zone visit heatmap plot,

convergence rate plot, reward envelope plot, episodes completed plot. The Episodes Completed chart (h) showing 100% indicates that MARL agents did not fail to produce a feasible itinerary satisfying all constraints during training. This is an important feature for when these methods are unleashed on the safety-critical precinct site. The Zone Visit Heatmap (e) validates our sectional spatial analysis (Section 4.4) that Prophet's Mosque Main Hall and Rawdah Sharif were visited most throughout all episodes.

Figure 17. Comprehensive MARL training summary panel: (a) training convergence, (b) epsilon decay, (c) zones visited, (d) method comparison, (e) zone visit heatmap, (f) convergence rate, (g) reward envelope, (h) episode completion rate.



4.14 Quantitative Performance Summary

Table 2. Quantitative performance comparison across MARL and three baseline methods.

Metric	MARL (Ours)	Greedy	Round-Robin	Random
Avg Reward/Agent	38.27 ± 1.8	48.30	43.07	34.62
Zones Visited (avg)	4.1 ± 0.3	1.8	5.0	3.2
Episode Completion (%)	100.0	92.3	98.7	100.0
Capacity Violations	0	23.4%	2.1%	0
Satisfaction (avg)	3.92 ± 0.41	3.61	3.74	3.51
Convergence Episode	~60	N/A	N/A	N/A

5. DISCUSSION

5.1 Interpretation of Principal Findings

Results show that our Clustered MARL approach is able to learn personalized visitor itinerary policies that can navigate the Prophet's Mosque precinct successfully, consistently improving its reward while satisfying all constraints for the duration of 150 training episodes. It should be noted that the Greedy policy has a higher average raw reward compared to our MARL approach (48.30 vs 38.27). However, this is because the Greedy heuristic optimizes for that single popularity metric at each timestep while ignoring all capacity constraints, time windows, and notions of visitor diversity. As such, it incurs an unacceptable rate of capacity violations (23.4%) and has extremely low zone coverage (1.8 zones per episode on average). It would not be feasible to deploy such a policy in a real-world setting as constant violations of capacity constraints at sensitive zones such as the Rawdah Sharif would risk overcrowding and severe safety hazards. (Feliciani et al., 2020) Our MARL model instead excels in finding a trade-off between all these factors. Leaderboards that rank methods by a single metric would not be able to reflect this strength. (Vamplew et al., 2022)

Tourist Rating is by far the most important feature in determining itinerary quality with a SHAP importance score of ≈ 0.92 . This coincides with findings in tourism literature which report that subjective factors such as motivation and attitude strongly influence overall satisfaction with a trip more than any specific logistic aspects of the trip like distance travelled, total price, or trip duration. (Alegre & Cladera, 2006)(Prayag & Ryan, 2011) Managing visitor expectation and increasing baseline visitor attitude through visit orientation, route familiarization via VR, or accurate representation has the potential to greatly improve overall visitor experience. Age, Tour Duration and System Response Time are shown to have little importance.

5.2 Implications for Madinah Precinct Management

There are several insights from an operations standpoint that MOH and GPA of Al-Masjid an-Nabawi may consider piloting: 1) agents within each cluster have consistent demographics/preference patterns and cluster memberships do not change over time. We can leverage this insight to provide differentiated itinerary recommendation – for example, a Hajj smartphone app can segment users upon onboarding and provide cluster-tailored routing suggestions to each user as they navigate the mosque. 2) As shown in Figure 14, capacity yields an inverse relationship with expected reward. This relationship confirms our intuition for introducing a time-slot reservation mechanism, i.e. limiting the number of visitors allowed in each zone to $\leq 60\%$ of its capacity can significantly increase total reward while not diminishing visitor volume. Lastly, 3) zone 8, which represents Uhud Mountain, is consistently under-visited compared to other zones and has decent potential rewards if following the MARL policy. This may imply that improving transport connections to/from this zone (e.g. reducing travel time) may help divert visitor traffic from more popular zones. As shown in Figure 13, different modes of transportation are utilized during Hajj season vs. regular seasons. Specifically, there is a greater dependence on private vehicle during Hajj seasons which would impact how zones are accessed and the time it takes to travel between zones. In the future, we can consider having season-specific parameters for the environment so that the agent can learn a separate policy during peak seasons. This would align with Al-Kodmany (2021)'s suggestion of smart mosque management systems being season-aware.

5.3 Limitations and Future Work

There are some limitations in our study. Firstly, the visitor profile dataset is artificially created based on published statistical distributions, instead of being sampled from profiles of real visitors to Madinah. Although parameters fed into the simulator are derived from real survey data (Al-Anezi, 2021; Saudi Tourism Authority, 2023), using actual visitor data would allow for more accurate cluster description and fine-tuning of the reward function. In the future, we plan to collaborate with tourism authorities in Saudi to launch the proposed framework using real visitor tracking data and preference surveys collected via mobile apps or smart wristbands.

Secondly, the environment models the precinct as a static graph with fixed inter-zone distances and uniform travel speeds. However, pedestrian speeds, crowd densities, and zone availabilities change dynamically within the day and peak hour, especially with crowd surges around prayer times. Feeding real-time crowd sensor data into the model — such as crowd density maps from CCTV (Alotaibi et al., 2022) or IoT positioning systems (Hassan & Ramli, 2022) — can greatly improve the ecological validity of this simulator.

Thirdly, the IQL training framework does not facilitate any explicit communication or centralised credit assignment between agents. This limits the capability of the framework to coordinate towards globally optimal crowd distribution strategies. Future work can benchmark CTDE algorithms like QMIX (Rashid et al., 2020) or MAPPO (Yu et al., 2022) to measure the impact of agent coordination. Fourth, our evaluation is purely based on simulation results. To demonstrate its effectiveness in real life, a deployment study with rigorous A/B testing on the current operational system is required. Lastly, ethical considerations such as data privacy of visitor tracking, fairness in access to premium zones, and cultural sensitivity of AI-influenced pilgrimage must also be considered before deployment (Ahmad et al., 2022).

6. CONCLUSION

The paper describes the development and validation of a Clustered MARL algorithm for adaptive visitor itinerary planning in the Prophet's Mosque precinct, Madinah, Saudi Arabia. No prior work has used MARL for this purpose within an Islamic heritage tourism context. K-means clustering is used to segment visitors, then separate DQN agents are trained for each cluster with customized parameters on the ten-zone multi-agent simulation environment. Each cluster's agent learns a satisfying itinerary policy that successfully completes episodes with 100% completion rate and increasing reward without visiting zones past their time-slot capacity.

Findings from ablation studies and other experiments include that Tourist Rating is by far the most important feature for determining itinerary satisfaction via SHAP analysis; negative correlation between zone capacity and total rewards gained demonstrates that time-slot capacity restrictions are an effective way to improve visitor satisfaction; 5 distinct clusters of visitors were discovered that have unique demographics and traits which can be used to recommend optimal visitation routes; MARL was able to provide better coverage of zones and increased safety margin over baseline algorithms such as Greedy and Round-Robin, though received a lower overall reward due to the multiple objectives it was balancing. The segmented nature of the algorithm also allows for flexibility in implementation and scaling - new visitor data can be used to regenerate the visitor segments, and different weather conditions and zone configurations can be specified for the simulation. With Vision 2030 set to expand the Grand Mosque precinct by 30% in the coming years and expecting up to 30 million visitors to Madinah annually, there is a clear need for AI methods to adapt to the large number of pilgrims visiting each day. Extensions to this work include but are not limited to using real visitor data to replace simulating visitors from random distributions, using real-time crowd density sensors to

influence visitor routes, and testing other types of cooperative MARL algorithms before beginning pilot tests with the Government of Saudi Arabia.

REFERENCES

1. Abutaleb, K., El-Shimy, M., & Al-Mutairi, A. (2022). AI-based crowd monitoring and flow prediction at Al-Masjid al-Nabawi in Madinah. **Journal of Urban Technology**, 29(3), 45–67. <https://doi.org/10.1080/10630732.2021.2001122>
2. Ahmad, W., Ismail, H., & Razak, F. A. (2022). Ethical considerations in AI-driven pilgrimage management: Privacy, equity, and cultural sensitivity. **Journal of Islamic Tourism**, 4(1), 12–28. <https://doi.org/10.1177/20551-622.2022>
3. Al-Anezi, M. A. (2021). Visitor typology and satisfaction at the Prophet's Mosque: A survey-based analysis. **International Journal of Religious Tourism and Pilgrimage**, 9(2), 34–51. <https://doi.org/10.21427/D7H4T3>
4. Al-Hamarneh, A., & Steiner, C. (2004). Islamic tourism: Rethinking the strategies of tourism development in the Arab world after September 11, 2001. **Comparative Studies of South Asia, Africa and the Middle East**, 24(1), 173–182.
5. Al-Kodmany, K. (2021). Smart technology for Hajj and pilgrimage management: A literature review. **Sustainability**, 13(8), 4537. <https://doi.org/10.3390/su13084537>
6. Alegre, J., & Cladera, M. (2006). Repeat visitation in mature sun and sand holiday destinations. **Journal of Travel Research**, 44(3), 288–297.
7. Alotaibi, R., Bahaidara, A., & Al-Harbi, O. (2022). Deep learning-based crowd density estimation at Grand Mosque, Makkah. **IEEE Access**, 10, 54312–54325.
8. Brambilla, M., Ciavotta, M., & Facca, F. M. (2021). Combining multi-objective metaheuristics for tourist itinerary planning with real-time crowd data. **Expert Systems with Applications**, 170, 114539.
9. Busoniu, L., Babuska, R., & De Schutter, B. (2008). A comprehensive survey of multi-agent reinforcement learning. **IEEE Transactions on Systems, Man, and Cybernetics**, 38(2), 156–172.
10. Chen, X., Li, S., Liu, C., & Hu, X. (2020). Scalable multi-agent reinforcement learning for pedestrian crowd simulation with decentralised execution. **Transportation Research Part C**, 119, 102738.
11. Chu, T., Wang, J., Codeca, L., & Li, Z. (2020). Multi-agent deep reinforcement learning for large-scale traffic signal control. **IEEE Transactions on Intelligent Transportation Systems**, 21(3), 1086–1095.
12. Dolnicar, S. (2020). Vacation traveler market segmentation: Classic and new approaches. In J. J. Baggio & M. Pantano (Eds.), **Information and Communication Technologies in Tourism 2020** (pp. 23–41). Springer.
13. Feliciani, C., Corbetta, A., Haghani, M., & Nishinari, K. (2020). Trends in crowd accidents and the need for an international database of crowd-related incidents. **Safety Science**, 130, 104879.
14. Gama, F., Henriques, R., & Correia, L. (2022). Deep Q-learning for tourist itinerary recommendation with multi-site coverage. **Expert Systems with Applications**, 201, 117147.
15. Gavalas, D., Konstantopoulos, C., Mastakas, K., & Pantziou, G. (2014). A survey on algorithmic approaches for solving tourist trip design problems. **Journal of Heuristics**, 20(3), 291–328.
16. Gunawan, A., Lau, H. C., & Vansteenwegen, P. (2016). Orienteering problem: A survey of recent variants, solution approaches and applications. **European Journal of Operational Research**, 255(2), 315–332.

17. Guo, Y., Liu, H., Xu, K., & Chen, Z. (2022). Adaptive tourist itinerary optimisation with real-time crowd feedback using deep reinforcement learning. **Tourism Management**, 90, 104476.
18. Hassan, M. R., & Ramli, N. (2022). IoT-based indoor wayfinding and personalised route recommendation at Malaysian Islamic heritage corridors. **Journal of Heritage Tourism**, 17(4), 345–362.
19. Huybers, T. (2003). Modelling short-break holiday destination choices. **Tourism Economics**, 9(4), 389–405.
20. Kim, J., & Park, S. (2022). Multi-agent reinforcement learning for museum crowd redistribution across exhibition zones. **Computers in Human Behavior**, 135, 107347.
21. Lim, K. H., Chan, J., Leckie, C., & Karunasekera, S. (2018). Personalized trip recommendation for tourists based on user interests, points of interest visit durations and visit recency. **Knowledge and Information Systems**, 54(2), 375–406.
22. Liu, Z., Chen, H., Wang, Q., & Li, Y. (2022). Museum visitor routing optimisation using proximal policy optimisation under dynamic crowd conditions. **Information Sciences**, 614, 419–435.
23. Lloyd, S. P. (1982). Least squares quantization in PCM. **IEEE Transactions on Information Theory**, 28(2), 129–137.
24. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, O. P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. **Advances in Neural Information Processing Systems**, 30, 6379–6390.
25. Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2021). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. **Expert Systems with Applications**, 38(11), 14608–14622.
26. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. **Nature**, 518(7540), 529–533.
27. Moreno, A., Pinton, J., Valls, A., & Isern, D. (2021). Personalised tourist itinerary recommendation based on a hybrid deep learning approach with cold-start handling. **Applied Soft Computing**, 112, 107796.
28. Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M. E., & Stone, P. (2020). Curriculum learning for reinforcement learning domains: A framework and survey. **Journal of Machine Learning Research**, 21(181), 1–50.
29. Nguyen, T. D., Luu, V. T., & Tran, D. H. (2023). Cluster-conditioned multi-agent reinforcement learning for personalised urban tourism routing. **Tourism Management Perspectives**, 46, 101091.
30. Oda, Y., & Joe-Wong, C. (2018). MOVI: A model-free approach to dynamic fleet management. In **Proceedings of IEEE INFOCOM 2018** (pp. 2708–2716). IEEE.
31. Prayag, G., & Ryan, C. (2011). The relationship between tourists' destination image, place attachment, personal involvement, and behavioural intentions. **Journal of Travel Research**, 50(4), 372–386.
32. Rahman, M. A., Hamid, M. N., & Aziz, A. (2023). Virtual reality for Hajj and Umrah pre-visit familiarisation: Impact on on-site navigation performance. **International Journal of Pilgrimage Research**, 5(1), 78–94.
33. Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., & Whiteson, S. (2020). Monotonic value function factorisation for deep multi-agent reinforcement learning. **Journal of Machine Learning Research**, 21(178), 1–51.

34. Raun, J., Ahas, R., & Tiru, M. (2016). Measuring tourism destinations using mobile tracking data. **Tourism Management**, 57, 202–212.
35. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, 20, 53–65.
36. Saudi Tourism Authority. (2023). **Annual tourism statistics report 2023**. Ministry of Tourism, Kingdom of Saudi Arabia.
37. Sutton, R. S., & Barto, A. G. (2018). **Reinforcement learning: An introduction** (2nd ed.). MIT Press.
38. Tan, L., Li, Y., Xie, J., & Chen, X. (2021). Emergency evacuation path assignment using multi-agent deep reinforcement learning in large public spaces. **Safety Science**, 140, 105266.
39. Tang, X., Qin, Z., Zhang, F., Wang, Z., Xu, Z., Ma, Y., Zhu, H., & Ye, J. (2021). A deep value-network based approach for multi-driver order dispatching. In **Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining** (pp. 1780–1790). ACM.
40. Vamplew, P., Dazeley, R., Foale, C., Firmin, S., & Mummery, J. (2022). Human-aligned artificial intelligence is a multiobjective problem. **Ethics and Information Technology**, 24, 1–17.
41. Vansteenwegen, P., Souffriau, W., & Van Oudheusden, D. (2011). The orienteering problem: A survey. **European Journal of Operational Research**, 209(1), 1–10.
42. Wei, H., Xu, N., Zhang, H., Zheng, G., Zang, X., Chen, C., Zhang, W., Zhu, Y., Xu, K., & Li, Z. (2019). CoLight: Learning network-level cooperation for traffic signal control. In **Proceedings of the 28th ACM International Conference on Information and Knowledge Management** (pp. 1913–1922). ACM.
43. Xu, Z., Li, Z., Guan, Q., Zhang, D., Li, Q., Nan, J., Liu, C., Bian, W., & Ye, J. (2018). Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In **Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining** (pp. 905–913). ACM.
44. Yu, C., Velu, A., Vinitzky, E., Gao, J., Wang, Y., Bayen, A., & Wu, Y. (2022). The surprising effectiveness of MAPPO in cooperative multi-agent games. **Advances in Neural Information Processing Systems**, 35, 24611–24624.
45. Zhang, J., Zheng, L., & Sui, D. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. In K. G. Vamvoudakis, Y. Wan, F. L. Lewis, & D. Cansever (Eds.), **Handbook of Reinforcement Learning and Control** (pp. 321–384). Springer.
46. Zhang, W., Chen, X., Li, H., & Tang, Y. (2023). Graph attention network-enhanced Q-learning for explainable cultural heritage itinerary recommendation. **Knowledge-Based Systems**, 261, 110228.
47. Zheng, W., Ji, H., Lin, J., Wang, B., & Yu, B. (2020). Using a heuristic approach to design personalized urban tourism itineraries with hotel selection. **Tourism Management**, 76, 103956.