

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

A Reinforcement Learning Approach to Leverage Urban Mobility Carbon Efficiency

Bruna Brasil Leão Marques



Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Rosaldo Rossetti

Co-Supervisor: Prof. António Ribeiro da Costa

June 30, 2025

A Reinforcement Learning Approach to Leverage Urban Mobility Carbon Efficiency

Bruna Brasil Leão Marques

Mestrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

President: Prof. Ana Paula Rocha

Examiner: Prof. Pedro Campos

Supervisor: Prof. Rosaldo Rossetti

June 30, 2025

Resumo

Dada a crescente urgência em reduzir as emissões de gases com efeito de estufa, a sustentabilidade na mobilidade urbana tornou-se um desafio crucial. O setor dos transportes, especialmente em áreas metropolitanas densamente povoadas, é um dos principais responsáveis pelas emissões de carbono.

Esta dissertação propõe um sistema de coordenação baseado em aprendizagem por reforço, com o objetivo de reduzir as emissões de carbono em redes de transporte multimodal. A solução modela o ambiente urbano como uma rede de zonas interligadas e uma população de viajantes, cada um com um par origem-destino e preferências individuais sobre os modos de transporte. Um agente coordenador central é responsável por determinar a ordem de prioridade com que os viajantes são atribuídos aos percursos disponíveis. O problema de aprendizagem é formulado como um cenário do tipo Multi-Armed Bandit (MAB), onde o agente explora sequências para identificar aquelas que minimizam as emissões acumuladas. Ao contrário dos sistemas tradicionais de atribuição, esta abordagem não exige cooperação total.

O coordenador utiliza uma estratégia ϵ -greedy com decaimento exponencial para equilibrar a exploração de novas sequências com a exploração de sequências promissoras. A aprendizagem é guiada por uma função de recompensa baseada em poupança de emissões normalizada, que incorpora as preferências dos viajantes, a estrutura dos percursos e restrições reais, como a capacidade dos modos de transporte. O sistema também considera a variabilidade diária através da introdução de ruído estocástico nas emissões.

Foram realizados experimentos abrangendo diferentes cenários sintéticos de cidade, variando o tamanho da população, a conectividade entre zonas e a disponibilidade de modos de transporte. Os resultados demonstram que a coordenação baseada em aprendizagem reduz consistentemente as emissões totais e promove a mudança para modos de transporte com menores emissões, especialmente em redes estruturalmente limitadas, onde a flexibilidade comportamental é mais relevante. Foram ainda simulados cenários com políticas de incentivos externos, como transporte público gratuito e aumento do custo dos combustíveis, através da modificação das preferências dos viajantes. A eficácia de cada estratégia de incentivo revelou-se dependente da estrutura da rede e da escala populacional, evidenciando o potencial do modelo para simulação e planeamento de políticas públicas. Adicionalmente, foi realizada uma análise de sensibilidade que comparou diferentes estratégias de decaimento da exploração (linear, exponencial e *simulated annealing*) com vários valores de ϵ . A estratégia de decaimento exponencial revelou-se a mais robusta e adaptável.

Em suma, este trabalho oferece um modelo para coordenar deslocações urbanas de forma mais eficiente em termos de carbono. Ao integrar aprendizagem por reforço com restrições reais de mobilidade e preferências individuais, esta abordagem representa um avanço no domínio dos sistemas inteligentes de transporte. O modelo proposto pode funcionar como ferramenta de apoio à decisão para planeadores urbanos que pretendam desenvolver soluções de mobilidade mais sustentáveis e inteligentes.

Abstract

As cities face growing pressure to reduce greenhouse gas emissions, sustainable urban mobility has become a critical challenge. Transportation, especially in densely populated metropolitan areas, is a major contributor to carbon emissions. Despite the availability of multiple travel modes, such as metro, bus, private car, and walking, individual mobility decisions are often inefficient from an environmental point of view.

This dissertation introduces a reinforcement learning-based coordination framework aimed at reducing carbon emissions in multimodal transport systems. The proposed solution models the urban environment as a network of interconnected zones and a population as travellers with origin-destination pairs and individual preferences over transport modes. A central coordinator agent is tasked with determining the priority order in which travellers are assigned to available paths. The learning problem is framed as a Multi-Armed Bandit (MAB) scenario, where the agent sequentially explores and exploits traveller orderings to identify those that lead to the lowest cumulative emissions. Unlike traditional scheduling systems, this approach does not rely on full cooperation but rather influences outcomes through intelligent prioritization.

The coordinator employs an ϵ -greedy strategy with exponential decay to balance exploration of new traveller sequences and exploitation of high-performing ones. A reward function based on normalized emission savings guides the learning process, incorporating traveller preferences, path structure, and real-time constraints such as transport capacity. The system also accounts for daily variability through stochastic perturbation of emissions.

A comprehensive experimental evaluation was conducted using synthetic city scenarios with varying population sizes, zone connectivity, and transport mode availability. The results demonstrate that the proposed learning-based coordination consistently reduces total emissions and encourages shifts away from high-emission modes, particularly in structurally constrained networks where behavioural flexibility is more critical. Simulations further evaluated the impact of external policy incentives, like free public transport and increased fuel prices, by adjusting traveller preferences to reflect real-world interventions. The effectiveness of each incentive strategy was shown to depend on network structure and population scale, highlighting the suitability of the model for policy simulation and planning. Furthermore, an ablation study compared multiple exploration decay strategies (linear, exponential, and *simulated annealing*) across different ϵ values. Exponential ϵ -decay proved to be the most robust and adaptive strategy. These findings underscore the importance of well-calibrated exploration in sequential learning tasks involving structured, interdependent decisions.

In conclusion, this work offers a framework for coordinating urban travel in a carbon-efficient way. By integrating reinforcement learning with real-world mobility constraints and preferences, it advances the state of the art in intelligent transportation systems. The proposed model can serve as a decision-support tool for urban planners and policymakers seeking to design smarter and greener mobility solutions.

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) offer a global framework for promoting peace and prosperity while protecting the planet. This dissertation contributes to these goals by presenting a reinforcement learning approach to reduce carbon emissions in urban transportation systems. The proposed model promotes behavioural adaptation through intelligent path assignment, aligning individual preferences with system-wide sustainability objectives.

This work directly contributes to the following Sustainable Development Goals:

SDG 11 Make cities and human settlements inclusive, safe, resilient and sustainable

SDG 13 Take urgent action to combat climate change and its impacts

SDG 9 Make cities and human settlements inclusive, safe, resilient and sustainable

SDG	Target	Contribution	Performance Indicators and Metrics
11	11.2	Supports the development of efficient, inclusive transport systems by promoting low-emission alternatives customized to individual traveller preferences.	Reduction in CO ₂ emissions per passenger-kilometre. Modal shift from car to metro, bus, and walking.
13	13.2	Enables the integration of emissions-aware coordination mechanisms in transport policy and planning simulations.	Emission savings achieved with FreePT and ExpensiveFuel scenarios. Total emission reduction over simulation horizon.
9	9.1	Applies advanced AI methods to create adaptable digital infrastructure for transport, capable of responding to real-time conditions and user preferences.	Convergence rate toward sustainable assignment sequences. Adaptability to network structure and incentive changes.

Acknowledgements

As I reach the end of this journey, I am very thankful for the people who supported me, believed in me and helped me grow.

Firstly, I would like to express my gratitude to my supervisor, Professor Rosaldo Rossetti, for his guidance throughout the development of this thesis. Additionally, I would like to thank Professors António da Costa and Zafeiris Kokkinogenis, for the insightful feedback and input that were essential to the completion of this work.

To my family, my parents, brother, grandmother and grandfather, thank you for always listening, offering advice, and believing in me unconditionally. Thank you for a lifetime of unconditional love and support.

To my boyfriend, your support and care have meant everything to me. You have always been there for every challenge, every stressful moment, offering your help and patience no matter the situation. You listened, encouraged, lightened my load when I felt overwhelmed, and stood by me every step of the way.

To all my friends who have been part of this journey, thank you for making these last five years so memorable. Thank you for the laughter, the study nights, the tears, the trips and the celebrations.

Bruna Marques

*"When life gets you down, you know what you gotta do?
Just keep swimming."*

Dory

Contents

1	Introduction	1
1.1	Scope and Motivation	1
1.2	Aim and Goals	2
1.3	Document Structure	3
2	Preliminaries	4
2.1	Reinforcement Learning	4
2.1.1	Markov Decision Processes	5
2.1.2	The Exploration–Exploitation Dilemma	5
2.2	Multi-Armed Bandits	6
3	State of the Art	9
3.1	Review Protocol	9
3.2	Research Questions	9
3.3	Search Queries	10
3.4	Primary Results	11
3.5	Filtering	12
3.6	Comparative Analysis of Reviewed Studies	14
3.7	Query 1 Main Results	16
3.7.1	Dynamic Task Allocation: The DC-MATA Framework	16
3.7.2	A Contextual Multi-armed Bandit Approach to Personalised Trip Itinerary Planning	18
3.7.3	Quality-Aware User Scheduling for Federated Mobile Crowdsensing	18
3.7.4	Bandit Learning Based Stable Matching for Decentralized Task Offloading in Dynamic Fog Computing Networks	19
3.7.5	Neural Constrained Combinatorial Bandits	20
3.7.6	Online Worker Selection Towards High Quality Map Collection for Autonomous Driving	21
3.7.7	Risk-Averse Combinatorial Semi-Bandits	21
3.7.8	EWS: Towards Cost-Effective Job Scheduling via Combinatorial Multi-Armed Bandit Learning	22
3.8	Query 2 Main Results	23
3.8.1	Multi-Agent Reinforcement Learning for Electric Vehicles Joint Routing and Scheduling	23
3.8.2	Dual-Objective Reinforcement Learning-Based Adaptive Traffic Signal Control for Decarbonization and Efficiency Optimisation	23
3.8.3	Graph Reinforcement Learning for Carbon-Aware Electric Vehicles in Power-Transport Networks	24

3.8.4	Multi-Agent Reinforcement Learning for Electric Vehicle Decarbonized Routing and Scheduling	24
3.9	Multimodal Mobility Systems and Infrastructure Planning	25
4	Methodology	29
4.1	Markov Decision Process	29
4.2	Proposed Approach	30
4.2.1	Traveller Path Utility Score	32
4.2.2	Reward Function	33
4.2.3	Environmental Dynamics	34
4.2.4	Epsilon-Decay Sampling	35
5	Empirical Demonstration	36
5.1	Experimental Setup	36
5.2	Evaluation Objectives	38
5.3	Experiment 1: Evaluating the Learning-Based Coordination Strategy for Emission Reduction	39
5.4	Experiment 2: Influence of Behavioural Incentives on Transport Mode Choices .	43
5.5	Experiment 3: Ablation Study on the Exploration Parameter ϵ and Decay Strategies	45
5.5.1	By Population	47
5.5.2	By Decay Strategy	49
5.5.3	Epsilon-Based Combinations	50
5.6	Summary	50
6	Conclusions	52
6.1	Main Contributions	52
6.2	Addressing the Research Questions	53
6.3	Future Work	53
	References	55
A	Experiment 1: Extended Results	59
A.0.1	Population Size of 1540	59
A.0.2	Population Size of 1080	61
A.0.3	Population Size of 840	62
A.0.4	Population Size of 560	63
A.0.5	Population Size of 360	64
B	Experiment 2: Extended Results	66
B.0.1	Population Size of 1540	66
B.0.2	Population Size of 1080	68
B.0.3	Population Size of 840	71
B.0.4	Population Size of 560	72
B.0.5	Population Size of 360	74
C	Experiment 3: Extended Results	76
C.0.1	Population Size of 1080	76
C.0.2	Population Size of 840	80
C.0.3	Population Size of 560	82
C.0.4	Population Size of 360	84

List of Figures

2.1	The agent–environment interaction in a Markov decision process [38].	5
2.2	Visual representation of the Multi-Armed Bandit problem ¹	6
3.1	Query 1 PRISMA Flowchart	12
5.1	Illustration of the Connection Types	38
5.2	Moving average of rewards in different population sizes.	44
5.3	Moving average of reward in different population sizes under three incentive conditions.	46
5.4	Moving average of rewards for 1080 people in four-zone under different ϵ -decay strategies.	48
5.5	Moving average of rewards for 840 people in four-zone under different ϵ -decay strategies.	48
5.6	Moving average of rewards for 360 people under different ϵ -decay strategies. . .	49
A.1	Transport mode usage evolution in six-zone for scenarios with 1540 individuals. .	60
A.2	Transport mode usage evolution for scenarios with 1080 individuals.	62
A.3	Transport mode usage evolution for scenarios with 840 individuals.	63
A.4	Transport mode usage evolution for scenarios with 560 individuals.	64
A.5	Transport mode usage evolution for scenarios with 360 individuals.	64
B.1	Moving average of reward for scenarios with 1540 individuals in six-zone under different incentive strategies.	66
B.2	Emission levels for scenarios with 1540 individuals in six-zone under different incentive strategies.	67
B.3	Transport mode usage evolution for 1540 individuals in six-zone scenarios under different incentive strategies.	67
B.4	Moving average of reward for scenarios with 1080 individuals in six-zone under different incentive strategies.	68
B.5	Moving average of reward for scenarios with 1080 individuals in four-zone under different incentive strategies.	68
B.6	Emission levels for scenarios with 1080 individuals in six-zone under different incentive strategies.	69
B.7	Emission levels for scenarios with 1080 individuals in four-zone under different incentive strategies.	69
B.8	Transport mode usage evolution for 1080 individuals in six-zone scenarios under different incentive strategies.	70
B.9	Transport mode usage evolution over time for 1080 individuals in four-zone scenarios under different incentive strategies.	70

B.10	Moving average of reward for scenarios with 840 individuals in four-zone under different incentive strategies.	71
B.11	Emission levels for scenarios with 840 individuals in four-zone under different incentive strategies.	71
B.12	Transport mode usage evolution over time for 840 individuals in four-zone scenarios under different incentive strategies.	72
B.13	Moving average of reward over time for scenarios with 560 individuals in four-zone under different incentive strategies.	72
B.14	Emission levels for scenarios with 560 individuals in four-zone under different incentive strategies.	73
B.15	Transport mode usage evolution over time for 560 individuals in four-zone scenarios under different incentive strategies.	73
B.16	Moving average of reward for scenarios with 360 individuals in four-zone under different incentive strategies.	74
B.17	Emission levels for scenarios with 360 individuals in four-zone under different incentive strategies.	74
B.18	Transport mode usage evolution for 360 people in four-zone under different incentive strategies.	75
C.1	Moving average of rewards for 1080 individuals in six-zone (no diagonals) under different ϵ -decay strategies.	76
C.2	Moving average of rewards for 1080 individuals in six-zone (PT Limited) under different ϵ -decay strategies.	77
C.3	Moving average of rewards for 1080 individuals in four-zone (no diagonals) under different ϵ -decay strategies.	77
C.4	Moving average of rewards for 1080 individuals in four-zone (PT Limited) under different ϵ -decay strategies.	78
C.5	Emission levels for scenarios with 1080 individuals in six-zone (no diagonals) under different ϵ -decay strategies.	78
C.6	Emission levels for scenarios with 1080 individuals in six-zone (PT Limited) under different ϵ -decay strategies.	79
C.7	Emission levels for scenarios with 1080 individuals in four-zone (no diagonals). .	79
C.8	Emission levels for scenarios with 1080 individuals in four-zone (PT Limited) under different ϵ -decay strategies.	80
C.9	Moving average of rewards for 840 people in four-zone (no diagonals) under different ϵ -decay strategies.	80
C.10	Moving average of rewards for 840 people in four-zone (PT Limited) under different ϵ -decay strategies.	81
C.11	Emission levels for scenarios with 840 individuals in four-zone (no diagonals) under different ϵ -decay strategies.	81
C.12	Emission levels for scenarios with 840 individuals in four-zone (PT Limited) under different ϵ -decay strategies.	82
C.13	Moving average of rewards for 560 people in four-zone (no diagonals) under different ϵ -decay strategies.	82
C.14	Moving average of rewards for 560 people in four-zone (PT Limited) under different ϵ -decay strategies.	83
C.15	Emission levels for scenarios with 560 individuals in four-zone (no diagonals) under different ϵ -decay strategies.	83

C.16 Emission levels for scenarios with 560 individuals in four-zone (PT Limited) under different ε -decay strategies.	84
C.17 Moving average of rewards for 360 people in four-zone (no diagonals) under different ε -decay strategies.	84
C.18 Emission levels for scenarios with 360 individuals in four-zone (no diagonals). .	85

List of Tables

3.1	Number of results in each database.	11
3.2	Query 1 Inclusion Criteria for Literature Selection.	13
3.3	Query 1 Exclusion Criteria for Literature Selection.	13
3.4	Query 2 Inclusion Criteria for Literature Selection.	14
3.5	Query 2 Exclusion Criteria for Literature Selection.	14
3.6	Comparison of most relevant articles from query 1 across relevant research dimensions.	15
3.7	Comparison of most relevant articles from query 2 across relevant research dimensions.	15
4.1	Emissions by transport mode.	33
5.1	Transport Mode Capacity Limits.	37
5.2	Transport Mode Capacity Limits.	37
5.3	Performance summary for all experimental scenarios.	39
5.4	Initial temperatures values derived from epsilon.	47
A.1	Difference in transport mode usage for scenarios with 1540 people.	59
A.2	Difference in transport mode usage for scenarios with 1080 people.	61
A.3	Difference in transport mode usage for scenarios with 840 people.	62
A.4	Difference in transport mode usage for scenarios with 560 people.	63
A.5	Difference in transport mode usage for scenarios with 360 people.	65

Acronyms and Abbreviations

CMAB	Contextual Multi-Armed Bandit
EV	Electric Vehicle
GHG	Greenhouse Gases
MA	Moving Average
MAB	Multi-Armed Bandit
MARL	Multi-Agent Reinforcement Learning
MDP	Markov Decision Process
RL	Reinforcement Learning
SA	Simulated Annealing
UCB	Upper Confidence Bound

Chapter 1

Introduction

1.1 Scope and Motivation

Urban areas face increasing challenges related to pollution and carbon emissions, much of which stems from inefficient and carbon-intensive transportation systems. Approximately 25% of global CO₂ emissions originate from the transport sector, and in 2023 alone, transport emissions rose by 1.1%, driven by increasing energy demands across multiple sectors [22]. In the European Union, road transport accounted for 76% of the transport sector’s total greenhouse gas (GHG) emissions in 2021, making it the largest contributor among all transport modes [17].

This context underscores the urgency of rethinking urban mobility to reduce greenhouse gas emissions and promote sustainability. In dense urban environments, travellers typically have access to multiple transport modes, such as car, metro, bus, and walking, and their daily decisions significantly impact both individual experience and system-wide environmental performance. These choices are shaped by personal preferences, infrastructure availability, policy-driven incentives, and real-time environmental or social conditions.

As urban populations continue to grow, the need for intelligent and low-emission transport solutions becomes more pressing. Cities must find effective ways to reduce their environmental impact while maintaining convenience and accessibility. Optimising transport mode assignments and route coordination can lead to substantial reductions in traffic congestion, air pollution, and carbon emissions.

However, encouraging commuters to adopt more sustainable travel behaviours remains a complex challenge. Solutions must balance individual preferences with collective outcomes, and must be able to adapt in real time to dynamic and heterogeneous urban contexts. This opens an opportunity for intelligent coordination algorithms that can learn from traveller behaviour, infrastructure constraints, and environmental variability to promote greener decisions without imposing rigid rules.

To explore this challenge, consider a scenario in which a city launches a coordinated mobility program aimed at reducing emissions. Participants agree to have their daily travel assignments managed by a central system that learns over time how to assign travellers in a way that optimises

overall efficiency and sustainability. Rather than forcing choices, the system prioritises travellers based on their potential to influence emissions, making earlier assignments more impactful. This turns urban transport into a sequential decision-making problem, where coordination becomes an adaptive learning task that evolves with the system.

1.2 Aim and Goals

Many existing transport management systems lack mechanisms to dynamically incorporate individual travel preferences or adapt to evolving environmental conditions. In highly dynamic multimodal networks, this absence of adaptability and personalization can lead to inefficient resource allocation and increased environmental impact.

Therefore, the central aim of this dissertation is to explore how reinforcement learning, particularly through the Multi-Armed Bandit (MAB) paradigm, can be used to minimise carbon emissions in urban multimodal transport systems. The proposed methodology is built around a coordination strategy that learns how to assign travellers to transport modes and paths in a way that respects individual preferences while reducing system-wide emissions.

The work proposes a modelling abstraction where each traveller is represented as an arm in a MAB problem, and the agent learns over time which priority orders are more effective for collective emission reduction. The learning process is guided by a reward function based on emissions saved and updated through regret signals. A personalised utility model determines which paths are more attractive for each traveller based on their individual preferences and the current availability of transport modes.

This framework is tested in a simulation environment that models a city as a network of zones and a population of heterogeneous travellers. The simulation includes several experimental configurations, each reflecting different urban layouts and transport infrastructures. The evaluation is structured around three types of experiments. The first assesses whether the coordination strategy effectively reduces emissions across scenarios with varying population sizes and network structures. The second examines how changes in traveller preferences, prompted by hypothetical incentives like free public transport or rising fuel costs, affect behaviour and emissions. The third is an ablation study, evaluating how different exploration strategies and decay parameters influence the agent's learning performance and stability.

The dissertation introduces a novel application of regret-based MAB coordination to the domain of sustainable transport planning. It develops a simulation-based evaluation framework that allows testing under diverse urban constraints. It demonstrates the practical feasibility of using learning methods to support complex decision-making in urban mobility systems. Finally, it provides empirical evidence on how coordination pressure, behavioural diversity, and environmental variability affect system performance and sustainability outcomes.

This problem has relevance in multiple domains. For policymakers, it offers a tool to evaluate how different incentive schemes or infrastructure changes might influence commuter behaviour. For transport planners, it provides a flexible strategy for managing transport resources in response

to varying demand conditions. For urban residents affected by pollution and climate change. Additionally, for the research community, it contributes to the advancement of work in multi-agent systems, reinforcement learning, and sustainable urban mobility.

1.3 Document Structure

This document is structured as follows. Chapter 2 introduces the Multi-Armed Bandit problem and its extensions, including ϵ -greedy and contextual strategies, and contextualizes their relevance to urban mobility scenarios. Chapter 3 reviews recent literature on the application of MAB and reinforcement learning to task allocation and mobility coordination, identifying key challenges and opportunities. Chapter 4 formalizes the problem setting and presents the proposed coordination framework, including the learning model, reward structure, and exploration strategies. Chapter 5 describes the experimental setup and evaluates system performance under varying urban configurations and policy scenarios. Chapter 6 concludes the dissertation by summarizing the main contributions and outlining potential directions for future work. Finally, Appendix A, B and C includes extended experimental results and additional visualizations.

Chapter 2

Preliminaries

Given that the core objective of this dissertation is to apply Multi-Armed Bandits (MAB) to support emission-aware decision-making in urban mobility, this chapter introduces the foundational concepts that support the proposed methodology. It begins with an overview of Reinforcement Learning (RL), including its formulation via Markov Decision Processes (MDPs) and the central exploration–exploitation dilemma. RL provides the broader conceptual framework for learning from interaction and feedback. Then, it focuses on the MAB problem, a simplified form of RL. The formal definition, key algorithms (e.g., ϵ -greedy, UCB, Thompson Sampling), and relevant extensions to structured decision spaces are presented as the basis for the coordination approach developed in later chapters.

2.1 Reinforcement Learning

Reinforcement Learning (RL) is a computational framework for learning how to act optimally in dynamic and uncertain environments through trial and error. The agent interacts with an environment over time by taking actions, receiving feedback in the form of rewards, and adjusting its behaviour to maximize long-term performance. Unlike supervised learning, where labelled data is available, reinforcement learning must discover successful strategies by exploring the environment [38].

Mathematically, RL problems are typically formulated as Markov Decision Processes (MDPs), where decisions are sequential and outcomes are partly random and partly under the agent’s control. The agent’s goal is to learn a policy $\pi(a|s)$ that maps each state $s \in S$ to an action $a \in A$ in a way that maximizes the expected cumulative reward over time [38].

RL is particularly well-suited for transport coordination scenarios, such as the one studied in this dissertation. In dynamic mobility systems, the environment is non-deterministic and preferences vary across agents. Traditional optimization techniques struggle in such contexts due to the lack of stationarity and full observability [30]. RL, by contrast, can adapt to such changes through experience, continuously refining its decision strategy. In this work, a simplified variant of RL,

the Multi-Armed Bandit, is used to model the coordination problem in a scalable and efficient way [4].

2.1.1 Markov Decision Processes

A Markov Decision Process (MDP) provides the standard mathematical model for reinforcement learning problems. It is defined by a tuple (S, A, P, R, γ) , where S is a set of possible states, A is a set of available actions, $P(s'|s, a)$ defines the transition probability from state s to s' after action a , $R(s, a)$ is the expected reward for taking action a in state s , and $\gamma \in [0, 1]$ is a discount factor that weights future rewards [38].

The agent–environment interaction is illustrated in Figure 2.1, which follows the standard formulation introduced by Sutton and Barto [38].

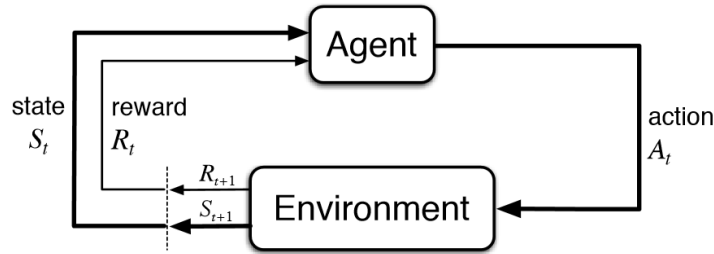


Figure 2.1: The agent–environment interaction in a Markov decision process [38].

At each time step t , the agent observes a state S_t , selects an action A_t , and receives a scalar reward R_{t+1} and the next state S_{t+1} . The process then repeats. A key property of MDPs is the *Markov property*, which asserts that the future is conditionally independent of the past, given the present state [38].

In this dissertation, the coordination problem is cast as a stateless bandit process, **a special case of an MDP**, the Multi-Armed Bandit, with a single state and no state transitions. This abstraction retains the core RL concepts (learning from reward feedback) while avoiding the need to model complex transitions. The details of this formulation and its application to the coordination problem are discussed later in Section 4.1 of Chapter 4.

2.1.2 The Exploration–Exploitation Dilemma

One of the central challenges in reinforcement learning is the *exploration–exploitation dilemma*. To achieve long-term reward maximization, the agent must strike a balance between two competing goals: *exploitation*, which involves selecting the best-known action to maximize immediate reward based on past experience; and *exploration*, which involves trying out new actions to gather information that might lead to better rewards in the future [4, 38].

This tension is critical in domains where the reward function is not known a priori, such as emission-aware mobility planning. A greedy strategy that always exploits current knowledge may

converge prematurely to a suboptimal solution, while excessive exploration may delay convergence and degrade performance [30].

A common strategy to balance these two forces is the ϵ -greedy algorithm. At each time step t , the agent selects the action with the highest estimated reward with probability $1 - \epsilon$, and a random action with probability ϵ . The ϵ parameter can be constant or decrease over time, known as ϵ -decay [38].

2.2 Multi-Armed Bandits

The name "**Multi-Armed Bandit**" comes from a gambling analogy in which a player must decide which of several slot machines (the "one-armed bandits") to play, each with an unknown but fixed payout distribution.

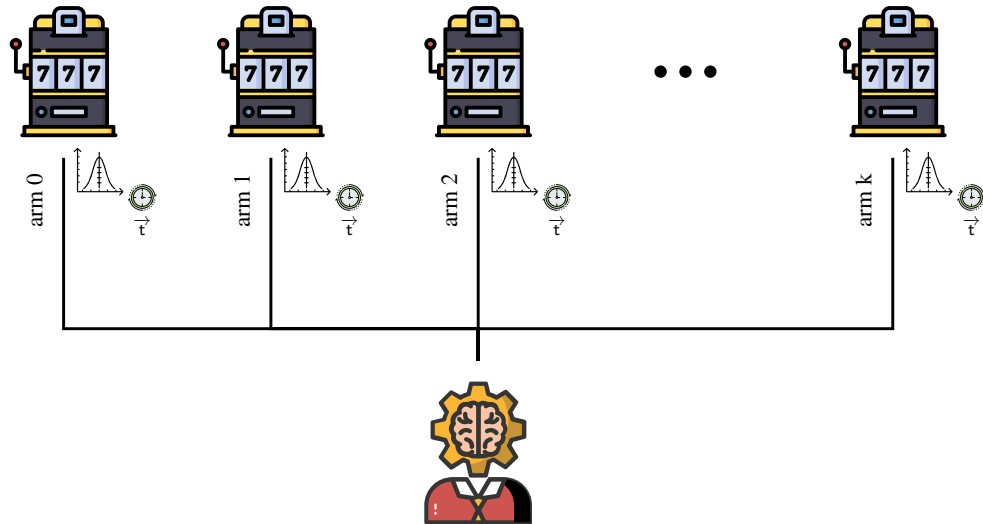


Figure 2.2: Visual representation of the Multi-Armed Bandit problem¹.

Multi-Armed Bandits are foundational tools for sequential decision-making under uncertainty, especially in environments where prior knowledge is unavailable and decisions must be continuously adapted based on partial feedback [28, 38]. While MABs have been widely used in large-scale production systems such as recommendation systems [1], their scope has expanded to more complex, structured problems involving adaptive control and coordination.

Recent work has extended MAB models to address real-world complexities such as concept drift [29], evolving item availability [31], and multi-objective recommendation [32]. These adaptations are particularly relevant to dynamic systems where the underlying reward structure changes over time or where multiple competing objectives must be balanced.

More closely aligned with our approach are the applications of MABs in online algorithm selection [20, 24] and in scenarios where decisions must be made over combinatorial or structured

¹Icons sourced from <https://www.flaticon.com/>.

action spaces. In these cases, MAB algorithms are used not merely to select among static alternatives, but to learn high-reward configurations or sequences of actions. For example, in portfolio optimisation, AutoML pipelines, or adaptive scheduling, the agent must learn optimal sequences of decisions under uncertainty.

The Multi-Armed Bandit problem is a classical formulation in online learning and decision theory. It models sequential decision-making under uncertainty, where an agent repeatedly chooses from a finite set of actions (or “arms”) in order to maximize the cumulative reward over time. The central dilemma is the balance between **exploration** (gathering new information) and **exploitation** (selecting the best-known action so far) [11, 25].

While the classical MAB framework assumes **independent arms** and scalar rewards, many real-world applications involve more complex decision spaces. In structured environments, combinatorial semi-bandits extend the MAB model to settings where actions consist of selecting subsets or combinations of arms, such as paths in a network, slates of recommendations, or bundles of resources. These models allow for partial feedback on individual components of the chosen action, enabling more efficient learning in high-dimensional or structured domains [37]. Such extensions are particularly relevant in applications like online routing, adaptive scheduling, and portfolio optimisation, where the action space is exponentially large but exhibits combinatorial structure.

Formally, let $A = \{a_1, a_2, \dots, a_N\}$ represent a set of N arms. Each arm $a_i \in A$ is associated with an unknown reward distribution D_{a_i} , from which rewards are drawn whenever that arm is selected. The mean of the distribution for each arm is denoted by $\mu_{a_i} \in [0, 1]$, representing the long-term average reward that the arm would yield if chosen repeatedly.

The decision-making process unfolds over a sequence of time steps $t = 1, 2, \dots, T$. At each step t , the agent chooses one arm $a_t \in A$ according to a decision policy π , observes a reward $R_{a_t, t} \sim D_{a_t}$, and updates the policy using the observed feedback. Over time, the agent’s objective is to identify a sequence of arm selections that maximizes the **cumulative reward** collected over the time horizon:

$$\sum_{t=1}^T R_{a_t, t}, \quad (2.1)$$

or, equivalently, to minimize the cumulative regret, which measures the difference between the expected reward of always playing the best arm and that of the selected sequence:

$$\rho_T = T\mu^* - \sum_{t=1}^T R_{a_t, t}, \quad (2.2)$$

where $\mu^* = \max_{a_i \in A} \mu_{a_i}$ is the expected reward of the optimal arm.

Letard et al. [25] emphasize that MAB algorithms are particularly well suited for dynamic online decision-making contexts such as recommendation systems and model portfolios, due to their ability to adaptively refine action-value estimates from sparse feedback without full supervision or training data.

One of the simplest strategies for the MAB is the ε -greedy algorithm [38]. In each round, with probability $1 - \varepsilon$, the agent selects the arm with the highest empirical reward (“greedy” choice), and with probability ε it explores by picking an arm uniformly at random. Thus a fixed parameter ε controls the exploration rate. If ε is constant, the agent continues to explore with the same frequency, which can lead to high regret in long runs.

At each time step t , the action a_t is selected as:

$$a_t = \begin{cases} \arg \max_{a_i \in A} \hat{\mu}_{a_i, t}, & \text{with probability } 1 - \varepsilon, \\ \text{random arm from } A, & \text{with probability } \varepsilon, \end{cases} \quad (2.3)$$

where $\hat{\mu}_{a_i, t}$ denotes the estimated average reward value of arm a_i at time t . Thus, this approach includes random exploration of the solution space according to a constant or decreasing probability $\varepsilon \in [0, 1]$. For a known horizon T , the number of iterations in the exploitation phase will therefore be approximately $T \cdot (1 - \varepsilon)$.

Due to its simplicity and adaptability, ε -Greedy and its variants are frequently employed in meta-learning and adaptive optimisation tasks, including algorithm selection in dynamic systems [25].

A common refinement is to use ε -decay, where ε decreases over time according to a given decay function. Early on the algorithm explores frequently (large ε) to gather information, then gradually reduces ε to focus on the best-performing arm [35].

Beyond ε -greedy, other two important examples are Upper Confidence Bound (UCB) methods and Thompson Sampling (TS). UCB1 is a family of algorithms that uses optimism under uncertainty. At each step the agent selects the arm maximizing $\hat{\mu}_a + c\sqrt{\frac{\ln t}{n_a}}$, where $\hat{\mu}_a$ is the empirical mean reward, n_a is the count of pulls of arm a , and t is the current round. The $\sqrt{\ln t / n_a}$ term is a confidence bonus that shrinks as an arm is sampled more. Intuitively, UCB1 chooses arms that either have high estimated reward or have been sampled few times [7]. Thompson Sampling is a Bayesian approach. At each round, it samples a possible reward value for each arm from its current belief (posterior distribution), then selects the arm with the highest sampled value. This way, it naturally balances exploration and exploitation: it explores more when uncertainty is high and exploits more when it is confident about the best arm [3].

Chapter 3

State of the Art

This section presents a systematic literature review focused on the application of Multi-Armed Bandit to dynamic task allocation and decision-making problems in mobility contexts. The aim is to identify state-of-the-art techniques, practical implementations, and research gaps related to how MAB frameworks can be applied to sequential decision-making in transportation systems, with a particular emphasis on carbon efficiency and emission reduction.

3.1 Review Protocol

To conduct the literature review, a Systematic Literature Review (SLR) methodology was adopted to ensure a transparent, structured, and replicable process. The review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework, which standardizes the stages of identifying, screening, and selecting relevant studies.

Searches were conducted on three scientific databases: IEEE Xplore [21], Scopus [36], and Web of Science [45]. Two queries were designed to reflect the research objectives, one focused on sequential decision-making and dynamic task coordination using MAB, and the other on the use of Reinforcement Learning (RL) to reduce environmental impacts in mobility systems.

After collecting the initial set of publications, a multi-step filtering process was applied to refine the results. Duplicate entries were removed and the remaining records were screened manually using the Rayyan platform [34]. This process resulted in a relevant set of studies.

The following sections describe the research questions, search queries, and filtering criteria in more detail, followed by an analysis of the selected articles for each query and their relevance to this work.

3.2 Research Questions

The goal of this review is to support the development of a learning-based, centralized decision-making framework that coordinates multimodal urban mobility in a way that minimizes system-wide emissions while respecting user preferences. Based on this objective, the following research

questions were defined:

RQ1: How can Multi-Armed Bandit algorithms be applied in urban environments through personalised and dynamic mobility coordination strategies?

RQ2: How can Reinforcement Learning be applied to reduce pollution (GHG, carbon emissions, sound pollution, etc.) in urban environments through different mobility policies?

To evaluate the research questions using the PICO framework, the questions were devised upon four important pieces of information: Population (who is the study about?), Intervention (What is being tested or introduced?), Comparison (What is the intervention compared against?) and Outcome (What are the expected results?).

Research Question 1:

1. **Population (P):** Studies and works that use MAB in dynamic task allocation and sequential decision-making.
2. **Intervention (I):** Application of MAB to dynamically coordinate and prioritize travel decisions in order to mitigate pollution.
3. **Comparison (C):** Alternative heuristic, static, or non-learning-based scheduling and mobility assignment methods.
4. **Outcome (O):** Improved carbon efficiency and reduced emissions through adaptive transport coordination.

Research Question 2:

1. **Population (P):** Studies and works that use RL in urban transport systems.
2. **Intervention (I):** Utilization of RL in various transportation modes and mobility policies to mitigate pollution, including carbon emissions and other environmental impacts.
3. **Comparison (C):** Alternative approaches using RL, or other machine learning methods, to achieve reduction of pollution in transportation.
4. **Outcome (O):** Improved carbon efficiency, reduced GHG and other pollutants, and optimised mobility systems in urban environments through RL.

3.3 Search Queries

The search process began by identifying the key concepts and defining appropriate keywords relevant to both the algorithmic framework (e.g., multi-armed bandits, reinforcement learning) and the application domain (e.g., sustainable mobility, emissions, carbon markets). Two distinct queries were formulated, each corresponding to one of the research questions guiding this study.

For the first research question (Q1), which investigates intelligent task allocation and routing using learning-based methods, the following query was used:

```
("multi-armed bandit" OR "combinatorial bandit") AND
( "dynamic scheduling" OR "sequential decision-making" OR "task
selection" OR "task allocation" OR "task assignment" OR "traffic
assignment" OR "trip assignment" OR "route selection" OR
"sequence-aware")
```

This query captures studies involving bandit algorithms applied to dynamic or sequential allocation tasks, such as route or trip planning.

For the second research question (Q2), which focuses on emissions-aware transport coordination using reinforcement learning, the following query was applied:

```
("multi-agent reinforcement learning") AND
("mobility" OR "urban transport*") AND
("carbon" OR "greenhouse gas" OR "carbon emission" OR "pollution
reduction") AND ("schedule" OR "route")
```

This query targets literature on multi-agent reinforcement learning applied to mobility systems with a focus on emissions reduction and sustainable transport policy enforcement.

3.4 Primary Results

The results of the search conducted on the three selected databases, IEEE Xplore [21], Scopus [36], and Web of Science [45], are presented in Table 3.1. A total of 984 results were retrieved across both queries.

Table 3.1: Number of results in each database.

Query	IEEE Xplore	Scopus	Web of Science
Q1	355	237	111
Q2	164	114	3

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework was followed. This approach provides a standardized and comprehensive structure for identifying, screening, and selecting studies for inclusion. The review process is illustrated in Figure 3.1.

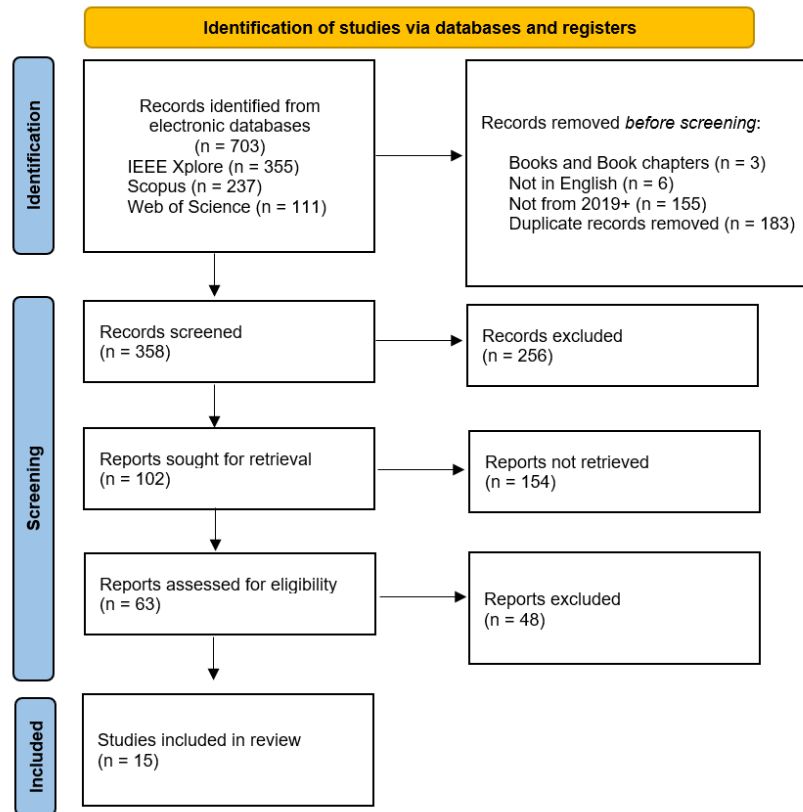


Figure 3.1: Query 1 PRISMA Flowchart

3.5 Filtering

To support the systematic identification of relevant literature, Rayyan [34], a web-based platform for managing and screening records in systematic reviews, was employed.

For query 1, an initial filter by publication year (from 2019, inclusive) and by language (English) resulted in 544 articles. After removing 183 duplicates, a total of 358 unique entries remained. For Query 2, the same filters returned 251 articles. After removing duplicates, 229 unique entries were retained.

These entries were then subjected to manual screening in multiple stages. First, abstracts were reviewed to determine thematic relevance. Subsequently, the introductions and conclusions were examined to assess methodological alignment. The refined inclusion and exclusion criteria used in the screening process are presented in Tables 3.2 and 3.3 for query 1, and 3.4 and 3.5 for query 2.

Table 3.2: Query 1 Inclusion Criteria for Literature Selection.

ID	Inclusion Criteria
IC1	Date of the publication since 2019 (inclusive)
IC2	Full-text access
IC3	Written in English
IC4	It is an article, conference paper, or review/survey paper, editorials
IC5	Applies Multi-Armed Bandit or Reinforcement Learning methodologies
IC6	Addresses sequential decision-making under uncertainty
IC7	Applies MAB/RL in dynamic task allocation, agent coordination, or routing scenarios
IC8	Mentions centralized or coordinator-based learning frameworks
IC9	Focuses on urban mobility, intelligent transport systems, or task assignment under environmental constraints
IC10	Considers emission reduction, energy efficiency, or sustainability goals in transport or mobility systems
IC11	Describes real case studies

Table 3.3: Query 1 Exclusion Criteria for Literature Selection.

ID	Exclusion Criteria.
EC1	It is a work-in-progress publication, extended abstracts, posters, opinion papers, white papers
EC2	It is duplicated
EC3	It is a summary of a workshop, seminar, lecture, or tutorial, book chapter
EC4	It is a shallow or superficial discussion
EC5	Focused solely on unrelated domains (e.g., agriculture, healthcare, security) without transferable methodologies
EC6	Uses unrelated learning techniques (e.g., supervised learning only, without MAB/RL relevance)

Table 3.4: Query 2 Inclusion Criteria for Literature Selection.

ID	Inclusion Criterias
IC1	Date of the publication since 2019 (inclusive)
IC2	Full-text access
IC3	Published in the English language
IC4	It is an article, conference paper, or review/survey paper, editorials
IC5	Applies Multi-Agent Reinforcement Learning, Multi-Agent Systems, or Reinforcement Learning methodologies
IC6	Focuses on urban mobility
IC7	Aims to reduce greenhouse gases, carbon, particles pollutants emissions
IC8	Mentions carbon market or related concepts
IC9	Describes real case studies

Table 3.5: Query 2 Exclusion Criteria for Literature Selection.

ID	Exclusion Criterias
EC1	It is a work-in-progress publication, extended abstracts, posters, opinion papers, white papers
EC2	It is duplicated
EC3	It is a summary of a workshop, seminar, lecture, or tutorial, book chapter
EC4	It is a shallow or superficial discussion
EC5	Applies to other areas than mobility

3.6 Comparative Analysis of Reviewed Studies

This section presents a comparative analysis of the most relevant literature aligned with the objectives and methodology of this dissertation. The purpose of this comparison is to systematically evaluate how recent research contributions relate to the key dimensions of this work: personalised, uncertainty-aware, reinforcement learning for low-emission, multimodal urban transport coordination.

The analysis is divided into two tables, Table 3.6 and Table 3.7. Each table includes the top 10 most relevant articles retrieved from two distinct literature review queries.

The columns in the tables represent key research characteristics:

- **RL:** Whether the article explicitly uses reinforcement learning techniques.
- **MAB:** Whether the article applies multi-armed bandits.
- **Emission Reduction:** Whether the work explicitly aims to reduce pollution or CO₂ or GHG.

- **Urban Mobility:** Whether the setting of the study is situated in urban transport systems.
- **Multimodal Transport:** Whether multiple transportation modes are considered.
- **Preferences:** Whether the approach incorporates individual preferences or personalised decision-making.
- **Uncertainty:** Whether is designed to handle dynamic, stochastic, or incomplete information.

This evaluation highlights both the diversity and the limitations of current research in this domain. While several studies explore emission-aware routing, RL coordination, or MAB-based decision frameworks, to the best of our knowledge, no single work in the literature combines all the critical features proposed in this dissertation, namely, uncertainty-aware, reinforcement learning, driven prioritization of traveller assignment in a multimodal, preference-sensitive, and emissions-optimised urban transport network. This dissertation therefore aims to bridge this gap.

Table 3.6: Comparison of most relevant articles from query 1 across relevant research dimensions.

Reference	RL	MAB	Emission Reduction	Urban Mobility	Multimodal Transport	Preferences	Uncertainty
[14]	Yes	Yes	No	No	No	Yes	Yes
[26]	Yes	Yes	No	Yes	Yes	Yes	Yes
[40]	Yes	Yes	No	No	No	Yes	Yes
[49]	Yes	Yes	No	No	No	Yes	Yes
[41]	Yes	Yes	No	No	No	No	Yes
[8]	Yes	Yes	No	No	No	No	Yes
[47]	Yes	Yes	No	No	No	No	Yes
[13]	Yes	Yes	No	Yes	No	Yes	Yes
[10]	Yes	Yes	No	No	No	No	Yes
[12]	Yes	Yes	No	No	No	No	Yes

Table 3.7: Comparison of most relevant articles from query 2 across relevant research dimensions.

Reference	RL	MAB	Emission Reduction	Urban Mobility	Multimodal Transport	Preferences	Uncertainty
[43]	Yes	No	Yes	Yes	No	No	Yes
[33]	Yes	No	Yes	Yes	Yes	No	Yes
[15]	Yes	No	Yes	Yes	No	No	Yes
[19]	Yes	No	Yes	Yes	No	No	Yes
[9]	Yes	No	Yes	Yes	No	Yes	Yes
[50]	Yes	No	Yes	Yes	No	No	Yes
[39]	Yes	No	No	Yes	No	No	Yes
[42]	Yes	No	Yes	Yes	No	No	Yes
[48]	Yes	No	Yes	Yes	No	No	Yes
[44]	Yes	No	Yes	Yes	Yes	Yes	Yes

3.7 Query 1 Main Results

This section presents a selection of the most relevant works recovered through the first query (Q1) of the systematic literature review.

3.7.1 Dynamic Task Allocation: The DC-MATA Framework

The key inspiration for this work is the **Dynamic and Cooperative Multi-Agent Task Allocation (DC-MATA)** framework [14], which addresses the challenge of allocating dynamically changing tasks among self-interested agents with limited information sharing. Traditional centralized methods such as the Assignment Problem (AP) assume full cooperation and complete access to agent preferences, conditions that are often unrealistic in decentralized or competitive environments. On the other hand, purely game-theoretic models account for individual preferences but often lack coordination, leading to inefficient allocations.

DC-MATA proposes a hybrid framework that introduces a *coordinating agent* responsible for defining a dynamic *priority order*. This order determines the sequence in which agents select tasks, allowing each agent to choose the task that maximizes its private utility without fully revealing its internal valuation. The only shared information with the coordinator is the task chosen and its associated benefit, maintaining privacy while supporting cooperative behaviour.

A key element in this setup is the **benefit matrix**, a table representing the utility each agent would gain from being assigned to each available task. This matrix evolves over time to reflect environmental changes such as resource availability, agent performance, and task complexity. Agents privately consult this matrix to decide their optimal task during their turn.

Formally, if we denote:

- I : the set of agents,
- J^t : the set of tasks available at time t ,
- b_{ij}^t : the benefit of assigning agent i to task j at time t ,
- $x_{ij}^t \in \{0, 1\}$: a binary decision variable indicating assignment,

The objective is to maximize total system reward over a time horizon H :

$$R = \max \sum_{t \in H} \sum_{i \in I} \sum_{j \in J^t} b_{ij}^t \cdot x_{ij}^t \quad (3.1)$$

The task allocation process at each time step $t \in H$ unfolds over N sequential steps, where N denotes the number of agents and, equivalently, the number of tasks available at that time. Agents follow a greedy strategy to maximize individual utility.

The core of this solution is a **reinforcement learning approach** modeled as a **Markov Decision Process (MDP)**. The coordinator agent learns over repeated allocation rounds which sequences result in higher cumulative benefits for the system. Since task benefits vary stochastically over time, the learning mechanism must adapt to uncertainty and evolving conditions.

To manage this, the paper models the learning challenge as a **MAB** problem, where the trade-off between *exploration* (trying new priority orders) and *exploitation* (using known good sequences) plays a crucial role. However, unlike standard MABs where actions are independent, DC-MATA introduces dependencies between actions because the sequence in which agents act influences which tasks remain available to others. This makes the problem sequential.

The paper proposes and evaluates four **MAB-inspired learning strategies** for the coordinator. These strategies rely on a **priority-sensitive reward function** that places greater weight on earlier agents in the selection order, because they have a greater impact on final outcomes. A reward table Q_t is maintained and updated each round based on the benefit observed. As shown in Equation 3, learning occurs only when positive regret is achieved, and the magnitude of the update is proportional to the improvement. After each update, the reward table is normalized.

$$Q_t \leftarrow Q_t + (R_w \cdot \max(\rho_t, 0)) \quad Q_t \leftarrow \frac{Q_t}{\sum_{i=1}^N Q_t} \quad (3.2)$$

Here, ρ_t represents the *regret*, calculated as the difference between the current round's total benefit and the maximum benefit observed so far:

$$\rho_t \leftarrow \sum_{i=1}^n b_{ij} x_{ij} - \max_{i=0}^t R \quad (3.3)$$

The vector R_w is a set of weights that emphasizes earlier agent positions in the sequence more heavily. It is defined as:

$$R_w = \frac{1}{t_s} \quad (3.4)$$

where t_s denotes the current step within the round (i.e., the position in the sequence). This weight decreases with the position in the sequence, giving greater learning influence to earlier agents.

To validate the solution, a series of simulation experiments were conducted using benefit matrices with different structures, variances ($\sigma^2 = 0, 2, 4, 8$), and task-agent combinations.

Key findings from the experiments indicate that the ϵ -decay strategy exhibited the most consistent performance across complex scenarios, reliably converging toward near-optimal solutions even under high variance. In contrast, PSAS and ASA strategies generally achieved faster learning in simpler environments but were more sensitive to the structure of the benefit matrix and the presence of noise.

DC-MATA effectively balances individual incentives with global performance. By capturing both individual incentives and global optimisation goals, DC-MATA opens new possibilities for smart task allocation in real-world domains such as agriculture, logistics, and collaborative robotics.

3.7.2 A Contextual Multi-armed Bandit Approach to Personalised Trip Itinerary Planning

An important and directly related contribution to the field of adaptive mobility decision-making is the work **A Contextual Multi-Armed Bandit Approach to Personalised Trip Itinerary Planning** [26]. This study addresses the problem of generating customized travel itineraries by leveraging a Contextual Multi-Armed Bandit (CMAB) framework. The CMAB approach enables the system to consider a variety of contextual features, such as user preferences, trip duration, budget constraints, preferred categories of points of interest (POIs) and accommodation choices, to incrementally learn the most satisfactory travel routes for each individual user. Unlike classical route planning algorithms, which typically provide static or one-size-fits-all solutions, the CMAB framework facilitates dynamic learning over time, progressively refining recommendations based on accumulated user interaction data and contextual observations.

The decision-making process in this work closely parallels the learning paradigm adopted in this dissertation. In both cases, personalization is central to the system's objective. In the itinerary planning scenario, the system selects a sequence of travel elements (transportation, lodging, attractions) that best align with each traveller's inferred preferences.

One notable strength of the article lies in its validation methodology, where the authors demonstrate that CMAB significantly outperforms traditional planning baselines in terms of user satisfaction and computational efficiency. These findings lend empirical support to the use of contextual bandit frameworks in real-world mobility applications, reinforcing the rationale behind employing MAB algorithms in this dissertation. While their work is focused on leisure travel and point-of-interest planning, the underlying learning framework and adaptation strategy are highly transferable to the urban mobility domain, especially when considering user-centric, sustainable transport solutions.

In conclusion, the CMAB approach to personalised trip planning contributes a strong methodological foundation and empirical justification for adaptive, preference-aware mobility coordination systems. It highlights the potential of learning-based solutions in managing the trade-off between individual satisfaction and system performance, a core challenge that is also addressed in this dissertation through the lens of emission-aware, reinforcement learning-driven multimodal transport assignment.

3.7.3 Quality-Aware User Scheduling for Federated Mobile Crowdsensing

Another relevant contribution to sequential decision-making in resource-constrained, personalised environments is presented in the paper **Quality-Aware User Scheduling for Federated Mobile Crowdsensing** [49]. This work tackles the challenge of scheduling mobile users (MUs) in a federated learning (FL) setting where each user has varying quality levels in terms of local data and model updates. In the context of mobile crowdsensing systems, where privacy constraints prevent centralized data collection, FL enables users to train local models on-device and share

only model parameters. However, the heterogeneity of user data quality and the dynamic nature of wireless channels introduce significant uncertainty and variability into the learning process.

To address these issues, the authors model the MU scheduling problem as a Combinatorial Multi-Armed Bandit problem. In this formulation, each user represents an arm, and at each iteration, a subset of users is selected for participation in the federated learning round. The goal is to maximize the quality of local model updates while minimizing communication latency and ensuring robust system performance over time. The authors introduce a novel scheduling algorithm that combines the Upper Confidence Bound (UCB) strategy with a reputation mechanism, which captures the historical reliability and performance of each mobile user. Specifically, the algorithm maintains a balance between exploration (trying underutilized users) and exploitation (selecting high-reputation users), ultimately promoting both accuracy and communication efficiency in the learning process.

Like the federated crowdsensing scenario, the transport coordination problem tackled in this dissertation also involves selecting individuals (travellers) from a population and assigning them actions (transport paths) that are constrained by shared resources (e.g., metro or bus capacities). Moreover, both systems must adapt over time based on limited feedback. In the federated learning context, this feedback comes from the performance of the aggregated model after receiving updates from selected users. In the transport coordination scenario, the system learns from observed outcomes such as emissions and traveller satisfaction following each assignment decision.

Another key parallel lies in the use of personalised, data-driven prioritization to guide sequential allocation. In the user scheduling framework, **individuals are ranked and selected based on their anticipated contribution** to the federated model. In both cases, this results in a dynamic, adaptive system that learns which individuals contribute most effectively to a global objective. For federated learning, this objective is model accuracy and convergence, while in the case of this dissertation, it is minimizing total system-wide emissions through personalised traveller assignments.

In summary, the work on quality-aware user scheduling for federated mobile crowdsensing provides valuable methodological insight and reinforces the applicability of bandit-based coordination strategies in dynamic, multi-agent systems where personalization, feedback uncertainty, and limited resources must be simultaneously managed.

3.7.4 Bandit Learning Based Stable Matching for Decentralized Task Offloading in Dynamic Fog Computing Networks

The article **Bandit Learning Based Stable Matching for Decentralized Task Offloading in Dynamic Fog Computing Networks** [40] presents an interesting contribution to decentralized decision-making under uncertainty. This work addresses the problem of assigning computing tasks from multiple mobile task nodes (TNs) to available helper nodes (HNs) with computing resources in a dynamic and uncertain environment.

The central idea of the paper is to jointly leverage MAB learning and stable matching theory to create a decentralized task offloading system in fog computing. In this setting, each task node

lacks full knowledge of the quality of service (QoS) characteristics of the available helper nodes and must learn over time which nodes yield the best performance for offloading computational tasks. In contrast, the helper nodes act in their own interest and may reject certain task requests to optimise their own resource allocation.

To solve this, the authors model the problem as a matching market with incomplete information, applying Thompson Sampling, a Bayesian bandit strategy, to enable each task node to learn the expected utility of matching with specific helper nodes. At each iteration, task nodes select helper nodes based on sampled expected rewards from posterior distributions, and a decentralized negotiation process resolves conflicts when multiple TNs prefer the same HN. The resulting matching aims to be both learning-aware (improving over time through interaction) and stable (no task-helper pair would rather be matched to each other than their current assignment).

Moreover, the authors demonstrate that their hybrid algorithm outperforms static heuristic baselines, especially in highly dynamic environments with time-varying task arrivals and node conditions. It validates the feasibility and effectiveness of using bandit algorithms to guide multi-agent decision-making, balancing individual agent preferences with collective efficiency.

3.7.5 Neural Constrained Combinatorial Bandits

A contribution to the field of combinatorial contextual bandit models is the work by Wang et al. titled **Neural Constrained Combinatorial Bandits** [41]. This study proposes a novel approach that combines neural function approximation with primal-dual optimisation techniques to address the limitations of existing contextual bandit models in handling complex reward structures and operational constraints. This study proposes a novel approach that combines neural function approximation with primal-dual optimisation techniques to address the limitations of existing contextual bandit models in handling complex reward structures and operational constraints.

Traditional bandit models, such as linear or kernel contextual bandits, assume simple structures for the reward function. However, in many real-world applications, the reward and cost functions are highly nonlinear and dependent on complex contextual inputs. For example, in federated learning, the selection of clients for model training can result in rewards and costs (e.g., energy consumption) that depend on numerous latent system factors. The authors argue that such scenarios are poorly addressed by existing bandit models and propose a framework that utilizes overparameterized multi-layer perceptrons to model both rewards and costs from contextual data.

The core of their method is the Neural-PD (Primal-Dual) algorithm, which performs online optimisation under anytime cumulative constraints. Unlike traditional cumulative constraints that are evaluated only at the end of the decision horizon, anytime constraints must be satisfied at every round, making them significantly more stringent and relevant to systems requiring consistent performance or fairness guarantees. A key feature of the model is its combinatorial selection setting: in each round, a subset of actions (arms) is chosen rather than a single one. The authors design a decision rule that maximizes a weighted score incorporating both the estimated rewards and the estimated penalties from potential constraint violations. This balance is controlled by tunable parameters that can emphasize exploitation, constraint satisfaction, or a mix of both, depending on

system priorities. Notably, their approach does not require prior knowledge of the structure of the reward or cost functions, allowing for flexibility and adaptability in deployment across different applications.

This work significantly extends the applicability of contextual bandit models by addressing non-linear reward estimation, combinatorial action spaces, and stringent operational constraints within a unified learning and optimisation framework. The use of neural networks enables rich modelling of complex interactions in the data, while the primal-dual structure ensures robust handling of constraints over time. These contributions make Neural-PD a promising foundation for constrained learning in systems such as resource-limited IoT networks, personalised recommender systems and task allocation under fairness or efficiency criteria.

3.7.6 Online Worker Selection Towards High Quality Map Collection for Autonomous Driving

Cao et al. in **Online Worker Selection Towards High Quality Map Collection for Autonomous Driving** [13] propose a learning-based approach for dynamically selecting vehicle agents to participate in crowdsourced HD map collection, a task critical for autonomous driving. The work addresses key challenges in dynamic task allocation: the unpredictability of agent behaviour (in this case, vehicles with uncertain routes), the need for high-quality data coverage, and operation under limited budgets. These challenges are representative of many real-world coordination problems involving mobile agents and resource-constrained environments.

Rather than relying on trajectory prediction models, which are often inaccurate in complex urban areas, the authors formulate the worker selection problem as a MAB scenario. This enables the platform to learn agent utility over time while balancing exploration and exploitation. Their solution, named Online Worker Selection (OWS), incorporates two main techniques: performance transfer, which estimates the potential of new agents based on past data, and a UCB-based selection rule that prioritizes agents with high observed utility or high uncertainty.

To measure performance, the system introduces a marginal utility metric that accounts for the spatial importance of different areas and penalizes redundant coverage. The goal is to maximize cumulative utility (i.e., data quality) while staying within a fixed recruitment budget. By focusing on utility rather than trajectory prediction, the method is able to make timely and robust decisions without requiring full prior knowledge of agent behaviour.

3.7.7 Risk-Averse Combinatorial Semi-Bandits

The paper **Risk-Averse Combinatorial Semi-Bandits** [8] introduces a theoretically grounded and practically significant advancement in the field of bandit-based decision-making under uncertainty. Traditional stochastic bandit formulations aim to optimise expected rewards, which assumes a risk-neutral perspective. However, many real-world systems, particularly those involving public resource allocation, critical infrastructure, or environmental constraints, must account for risk sensitivity, especially in the face of high-variance outcomes or limited tolerance for poor decisions.

The authors address this by introducing a risk-averse extension to the combinatorial semi-bandit setting.

The standard combinatorial semi-bandit model allows agents to select subsets of actions (arms) in each round, receiving individual feedback for each chosen component. This is suitable for structured decision spaces, such as selecting a set of transport routes, task assignments, or user-targeted content bundles.

3.7.8 EWS: Towards Cost-Effective Job Scheduling via Combinatorial Multi-Armed Bandit Learning

A significant contribution to the field of online resource allocation and learning-based scheduling is presented in the work by Xiao et al., titled **EWS: Towards Cost-Effective Job Scheduling via Combinatorial Multi-Armed Bandit Learning** [47]. This paper addresses the problem of efficiently scheduling AI-intensive jobs on heterogeneous GPU clusters, where cold-start delays, uncertain execution times, and budget constraints collectively make job scheduling highly challenging. Unlike prior scheduling strategies that often assume deterministic or predictable job durations, the authors propose a framework that explicitly incorporates uncertainty and limited information through an online learning approach based on Combinatorial Multi-Armed Bandits (CMAB).

The proposed algorithm, Exponential Weighting Scheduling (EWS), is designed to make job-instance assignment decisions in real time, using feedback from previously observed outcomes to improve future performance. At its core, EWS maintains a probability distribution over possible job-instance mappings, which it updates dynamically using the EXP3 algorithm, a standard method in adversarial and stochastic bandit settings. This probabilistic approach allows the scheduler to balance exploration of new allocation strategies with exploitation of high-performing ones, without requiring prior knowledge of job durations or instance-specific performance metrics.

The model formulation captures essential constraints of real-world systems. Jobs arrive dynamically with specific deadlines and resource requirements, and instances may be cold (requiring startup latency) or warm. Each job must be assigned to a feasible instance such that its memory requirements are met and its deadline is respected if possible. The overall objective is to minimize the total operating cost for service providers, defined as the monetary cost of running instances, while still meeting user-defined quality-of-service targets.

One of the central contributions of the work is the transformation of the job scheduling problem into a CMAB setting, where each action corresponds to a vector representing a feasible set of job-instance assignments. The reward function is inversely proportional to the cost incurred by a scheduling decision, and rewards are only observed after execution, capturing the semi-bandit feedback structure typical of combinatorial bandits. The EWS algorithm updates its decision distribution using exponential weighting, assigning higher probability to actions with lower observed cost, and adjusting for variability in GPU resource availability.

Extensive experimental evaluations are conducted using real-world traces and simulated workloads, including diverse AI applications such as image generation and multimodal dialogue systems. The results show that EWS consistently reduces total costs and cold-start delays compared to heuristic baselines.

3.8 Query 2 Main Results

This section describes a selection of the most relevant works retrieved through the second query (Q2) of the systematic literature review.

3.8.1 Multi-Agent Reinforcement Learning for Electric Vehicles Joint Routing and Scheduling

Wang et al. [43] present a comprehensive multi-agent reinforcement learning (MARL) framework designed to jointly optimise routing and charging decisions for electric vehicles (EVs) within an integrated transport and energy system. Their primary objective is to reduce carbon emissions by enabling intelligent coordination between the mobility behaviour of electric vehicles and the intensity of carbon in the grid. The problem is modeled as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP), which allows EV agents to make autonomous decisions based on local observations, while a central aggregator provides routing guidance. The learning architecture combines Deep Q-Networks for the aggregator with parameter-sharing Deep Deterministic Policy Gradient for individual agents, enabling decentralized control under uncertainty. The framework is evaluated through realistic simulations that incorporate dynamic electricity prices, carbon intensity signals, and traffic congestion data. The results demonstrate significant reductions in CO₂ emissions compared to baseline strategies, highlighting the potential of MARL for carbon-aware decision-making in mobility systems.

Despite its strengths, this work is primarily constrained to electric vehicle networks and assumes the availability of smart grid infrastructure and vehicle-to-grid (V2G) integration. It does not account for the diversity of traveller behaviours, preferences, or transport modes beyond the EV domain. As a result, the approach is limited in its applicability to broader urban mobility systems, where a mix of public and private modes, infrastructure disparities, and behavioural heterogeneity play central roles in shaping emissions.

3.8.2 Dual-Objective Reinforcement Learning-Based Adaptive Traffic Signal Control for Decarbonization and Efficiency Optimisation

The **Dual-Objective Reinforcement Learning-Based Adaptive Traffic Signal Control for Decarbonization and Efficiency Optimisation** [48] article propose a reinforcement learning-based traffic signal control framework that jointly targets two objectives: minimizing traffic congestion and reducing CO₂ emissions. Their approach leverages a Deep Q-Network (DQN) to optimise traffic light phases dynamically based on real-time traffic conditions, queue lengths, and emission

estimates. The reward function incorporates both environmental and efficiency metrics, enabling the agent to learn policies that balance vehicle throughput with sustainability goals.

The system is evaluated using microscopic traffic simulations in VISSIM, where it is shown to outperform conventional fixed-time and actuated signal control strategies. Notably, the results demonstrate measurable reductions in both average delay and cumulative emissions, illustrating the potential of deep reinforcement learning (DRL) to guide more sustainable urban traffic management policies. The dual-objective formulation is especially relevant from a policy design standpoint, as it aligns technical optimisation with broader environmental objectives.

Despite its contributions, the scope of the work remains limited to intersection-level traffic control. It does not consider traveller behaviour, modal choice, or city-wide coordination mechanisms, which are important for understanding and influencing broader emission patterns in urban mobility systems. Moreover, the solution operates within a fixed spatial structure and does not adaptively assign agents to modes or paths, leaving untapped the potential for traveller-level interventions.

3.8.3 Graph Reinforcement Learning for Carbon-Aware Electric Vehicles in Power-Transport Networks

Zhang et al. [33] present a graph reinforcement learning framework for optimising the routing and charging behaviour of electric vehicles (EVs) with a focus on minimizing carbon emissions in interconnected power-transport networks. The proposed method models the spatial and temporal dependencies between EVs, road traffic, and the power grid using a Graph Convolutional Network (GCN), which is integrated into an actor-critic reinforcement learning architecture. By incorporating real-time carbon intensity data from the power grid and traffic states from the road network, the model enables EV agents to learn environmentally optimal paths that reduce both travel-related and electricity-related emissions.

The study highlights the importance of joint system optimisation, particularly in the context of increasing EV adoption and the corresponding strain on energy infrastructure. Empirical results demonstrate that the graph-based reinforcement learning approach outperforms traditional shortest-path and cost-based baselines in terms of both CO₂ reduction and route efficiency. However, the approach is narrowly tailored to electric vehicles and assumes access to fine-grained, real-time data across energy and transport sectors. It does not consider broader mobility behaviour.

3.8.4 Multi-Agent Reinforcement Learning for Electric Vehicle Decarbonized Routing and Scheduling

A relevant contribution to the intersection of urban mobility and emission mitigation is presented in the work **Multi-Agent Reinforcement Learning for Electric Vehicle Decarbonized Routing and Scheduling** [44]. This study proposes a hierarchical MARL framework that jointly optimises eco-routing and charging schedules for electric vehicles (EVs), with the explicit aim of reducing carbon emissions across both transport and power networks.

Unlike traditional models that treat EV routing and energy scheduling as separate problems, this work introduces a unified solution by modelling the joint problem as a partially observable Markov game. The authors employ a hybrid policy based on Proximal Policy optimisation (PPO), which simultaneously handles discrete routing decisions and continuous charging schedules, allowing for scalable decision-making under dynamic and uncertain conditions.

Two main case studies are used to demonstrate the system's effectiveness: a synthetic network and a realistic central London scenario. Results show that the proposed model can significantly reduce total CO₂ emissions by integrating route choice with vehicle-to-grid (V2G) services, allowing EVs to adjust their energy consumption based on grid carbon intensity in real time.

A notable strength of this approach is its ability to operate without requiring prior knowledge of traffic or power systems, which increases its applicability in real-world deployments. Furthermore, the framework supports decentralized agent behaviour, enabling each EV to learn and adapt independently based on localized observations.

However, the method remains focused on electric vehicles and infrastructure that supports bidirectional energy flow, such as smart grids and V2G integration. This narrows its applicability to urban systems with high EV penetration and advanced power networks. The model does not account for broader traveller heterogeneity or multimodal transport systems that include public transit, active transport modes, or varying user preferences. Additionally, the model's reliance on fine-grained energy data and infrastructure may pose challenges in cities with less mature digital and energy infrastructures.

Still, the work marks an important step toward coordinated, low-carbon mobility and energy management, offering a compelling proof-of-concept for how reinforcement learning can be used to align transportation and energy systems with decarbonization goals.

3.9 Multimodal Mobility Systems and Infrastructure Planning

Urban transport systems are increasingly shaped by the interplay between infrastructure design, user behaviour, and environmental sustainability goals. This section highlights relevant works that explore the planning and coordination of multimodal mobility networks, focusing on how infrastructure, user preferences, and policy mechanisms can be integrated to improve transport efficiency and reduce environmental impacts. These studies provide critical background for understanding the structural, behavioural, and algorithmic challenges that this dissertation seeks to address, especially regarding traveller assignment, system coordination, and emissions-aware decision-making.

Multimodal Urban Mobility and Multilayer Transport Networks [5] presents a comprehensive survey of multimodal transport systems through the framework of multilayer network science. The study examines how urban mobility unfolds across various transport modes (walking, cycling, public transit, and automobiles) and how these are modeled as interconnected layers, each with distinct structural and dynamical properties. This multilayer perspective supports a detailed analysis of how different modes interact to influence accessibility, congestion, and overall system efficiency.

The authors introduce metrics such as betweenness centrality, interdependence, overlap census, and spatial outreach to evaluate the integration and resilience of transport networks. These tools provide insight into the relative importance of each transport layer and help quantify the effects of structural constraints, intermodal connectivity, and network design on mobility performance. Particular attention is given to how limitations in capacity and asymmetric availability of modes shape system behaviour.

In addition to infrastructural analysis, the study reviews empirical findings on traveller behaviour, noting that decision-making is often suboptimal due to bounded rationality, incomplete information, and heterogeneous preferences. This highlights the need for adaptive, preference-aware coordination mechanisms capable of responding to dynamic urban conditions. The paper also discusses the role of policy interventions and simulation tools in supporting sustainable mobility strategies, underscoring the importance of integrating behavioural and structural considerations in transport modelling.

Optimisation of the location and capacity of shared multimodal mobility hubs to maximize travel utility in urban areas [46] presents a bilevel optimisation model for determining the optimal location and capacity of shared multimodal mobility hubs in urban areas. The model integrates infrastructure planning with individual travel behaviour, aiming to maximize overall travel utility while accounting for network constraints, demand distribution, and service characteristics. At the upper level, the problem addresses the strategic design of hubs, including their placement and capacities; at the lower level, it simulates user equilibrium by modelling traveller mode and hub choices based on generalized cost functions.

The study emphasizes the importance of aligning infrastructure investments with user preferences and travel patterns to enhance network efficiency and accessibility. By incorporating factors such as walking distance, transfer times, and hub congestion into utility-based mode choice models, the framework captures the trade-offs travellers face in multimodal systems. The application of the model to the city of Amsterdam demonstrates its ability to identify configurations that significantly improve network performance and modal integration.

This work contributes to the broader discussion on demand-responsive and sustainable urban mobility by showing how integrated planning can optimise the effectiveness of shared mobility services. The modelling approach also illustrates how behavioural considerations can be embedded within optimisation routines to support more adaptive and user-centric transport strategies.

Planning shared mobility hubs in European cities: A methodological framework using MCDA and GIS applied to Barcelona [6] presents a methodological framework for planning shared mobility hubs in urban contexts, integrating spatial analysis with decision-support techniques. The framework combines Multi-Criteria Decision Analysis (MCDA), the Analytic Hierarchy Process (AHP), and Geographic Information Systems (GIS) to assess and visualize the suitability of potential locations for shared mobility infrastructure. Their decision-support tool, M2MHub, is designed as a QGIS plug-in and enables the generation of desirability maps based on a wide range of urban and transport indicators. The framework incorporates six main location criteria, mobility, compactness, diversity of uses, profile of potential users, spatial configuration,

and environment, each supported by a weighted set of sub-indicators. These indicators were derived through literature review and expert consultation with stakeholders across European cities. By accounting for both transport-specific and urban design factors, the methodology supports a holistic approach to locating intermodal hubs that can serve as both functional transfer points and socially integrated urban spaces. The application of the framework in Barcelona demonstrates its flexibility and utility in identifying optimal locations for mobility hubs, particularly within dense, multimodal urban settings. Results show that criteria such as network accessibility, walkability, air quality, and land-use diversity significantly influence the spatial desirability of hub locations. Importantly, the tool supports scenario testing, allowing planners to adjust indicator weights and simulate different policy priorities or future development plans.

The development of cooperative multimodal transport systems has gained traction as cities seek to improve efficiency and sustainability in mobility services. The recent study **Revenue sharing and resource allocation for cooperative multimodal transport systems** [16] proposes a two-stage hierarchical game-theoretical model for revenue sharing and resource allocation among transport service providers (TSPs). In their framework, a cooperative game defines a revenue-sharing rule to ensure coalition stability, while a non-cooperative game models the decentralized decisions of TSPs regarding resource allocation. This approach reflects the reality that while co-operation can enhance service integration and efficiency, TSPs still act in self-interest and require incentive-compatible mechanisms to participate. The study highlights the significance of designing revenue-sharing schemes that ensure Pareto efficiency and coalition stability under demand uncertainty, offering a computationally tractable alternative to classical methods like the Shapley value. These insights underscore the complex interplay between cooperation and competition in multimodal systems and support the integration of strategic agent behaviour into urban mobility planning frameworks.

The integration of electric vehicles (EVs) into urban transport systems introduces new challenges for routing and scheduling, particularly due to battery limitations and the coupling between mobility and power infrastructures. The work **Energy-efficient multimodal mobility networks in transportation digital twins: Strategies and optimisation** [27] presents a MARL framework that jointly optimises the routing and charging scheduling of EVs. The proposed model accounts for traffic conditions, energy consumption, and charging infrastructure constraints, enabling agents to learn coordinated strategies that minimize both travel delay and energy cost. By modelling EVs as decentralized learners operating within a shared environment, the approach captures the interdependence between route selection and charging behaviour. The results demonstrate that MARL can significantly enhance system performance compared to baseline heuristics, especially in scenarios with high demand variability and limited charging resources. This work emphasizes the necessity of coordination among autonomous agents in multi-modal and energy-aware transport networks.

Modelling the multimodal mobility behaviour of urban commuters has increasingly relied on advanced learning techniques that can accommodate both individual preferences and group-level trends. In **Deep Reinforcement Learning of Simulated Students Multimodal Mobility Be-**

haviour: Application to the City of Toulouse [23], Kermali et al. (2025) propose a Deep Reinforcement Learning (DRL) framework to predict multimodal mobility choices of students based on mesoscopic factors such as vehicle ownership and origin-destination regions. Their approach integrates a multilayer graph representation of the transportation network with a Proximal Policy optimisation (PPO) agent, enabling the learning of personalised yet generalizable commuting patterns. By employing graph attention networks and LSTM-based encodings, the model captures both structural relationships in the network and sequential decision behaviour. Experiments using simulated data from Toulouse demonstrated the model's robustness in both unimodal and multimodal scenarios, offering a scalable solution for anticipating mobility flows and informing sustainable urban transport planning.

Chapter 4

Methodology

This chapter presents the methodological foundations of the proposed learning-based coordination framework, which aims to support low-emission traveller assignment in multimodal urban transport systems. The approach models the coordination problem as a variant of the Multi-Armed Bandit framework. The methodology operates within the context of complex transport networks with restricted capacity, where multiple travellers, with diverse preferences, must be dynamically assigned to emission aware travel paths. By integrating urban mobility planning with regret-based learning and traveller's preference-aware utility scoring, the framework captures the trade-offs between individual preferences, system capacity, and environmental impact. The following sections formalize the problem, describe the learning architecture, and detail the simulation dynamics used to evaluate the strategy.

4.1 Markov Decision Process

In classical Markov Decision Processes (MDPs), the agent interacts with an environment over multiple time steps, observing a sequence of state transitions. However, as explained by Albrecht *et al.* [4], the multi-armed bandit problem is a special case of the MDP, characterized by a single state ($|S| = 1$) and no state transitions. After each round, the environment resets and a new round begins. There is no carryover of state information between rounds.

Thus, the environment satisfies the conditions of a stateless bandit problem:

- S contains only one state, which is the same at the start of every round,
- A is the set of actions corresponding to priority orderings of travellers (i.e., choosing the next traveller to assign),
- R is a stochastic reward function based on total emissions saved,
- There is no transition function P , as the process completes in one "shot" per round and the environment resets.

The coordinator agent's objective is to learn a policy π that selects effective priority orders in each round to maximize the total reward across a finite horizon H . This policy is guided by an estimated action-value function $Q_t(a)$, which is updated over time using observed rewards and an exploration-exploitation sampling strategy, as described in the next section.

4.2 Proposed Approach

This work builds upon an agent-based coordination framework to support sustainable transport mode assignment in urban mobility systems. Inspired by the **Dynamic and Cooperative Multi-Agent Task Allocation (DC-MATA)** framework [14], which uses dynamic agent prioritization to improve task allocation under uncertainty, a similar abstraction is adopted to address the challenge of minimizing transport-related emissions in a multimodal setting, where multiple types of transportation modes are available and can be combined.

A **dynamic and distributed transport assignment problem** is considered in a simulated city composed of predefined zones and synthetic populations. The city is modeled as a directed multigraph $G = (V, E)$, where each node $v \in V$ represents a fictional city zone, and each edge $e \in E$ corresponds to a transport connection between zones. Each edge is associated with a specific transport mode and its corresponding capacity. The set of transport modes $J = \{\text{metro, bus, car, foot}\}$ differ in capacity and emissions. Multiple transport modes between the same zone pair are represented as parallel edges in the multigraph, though not all modes are available between every pair.

Each traveller $i \in I$ is defined by:

- a distinct **origin-destination (OD)** pair of zones,
- a personalised **preference vector** over the four transport modes, with values ranging from 1 to 100.

The OD pairs and preferences are synthetically generated to reflect diverse behavioural tendencies (see Section 5.1 in Chapter 5).

The coordination task is modeled as a **Multi-Armed Bandit** problem, where each traveller corresponds to an arm. The learning objective is to discover an optimal **priority order** $O = \{o_1, o_2, \dots, o_N\}$, where o_k denotes the index (i) of the traveller assigned at position k in the sequence, and $N = |I|$ is the total number of travellers. The values in O form a permutation of traveller indices, meaning each traveller is assigned exactly once per round.

The coordinator agent seeks to learn such an ordering to optimise overall system performance. Specifically, in each round t , the objective is to maximize the total reward obtained from all traveller assignments:

$$\max R^t = \sum_{i=1}^N R_i^t \quad (4.1)$$

where R_t^i is the reward received by traveller i in round t . The total reward R^t captures system-level performance, particularly in terms of emissions reduction, and is influenced by the interdependencies introduced by sequential traveller decisions. Further details on how individual rewards are computed can be found in [Subsection 4.2.2](#).

A **coordinator agent** manages the assignment process over a finite time horizon H . In each round $t \in H$, the coordinator selects a priority order O determining the sequence in which travellers select a preferred path based on their preferences. When selected by the coordinator, each traveller evaluates all feasible multimodal paths from their origin to destination and acts greedily, always choosing the path with the highest **utility score** u , which reflects how well the path aligns with their individual transport preferences. The score function is defined and discussed in detail in [Subsection 4.2.1](#). As a result, early assignments influence the availability and quality of options for later travellers, introducing interdependencies among decisions.

To track which travellers are still to be assigned, a traveller mask is maintained. This is a binary vector where each value indicates whether the corresponding traveller is unassigned (1) or already assigned (0). The mask is updated after each step of the assignment process and reset at the end of each round.

Unlike traditional MABs with independent arms, this setting involves interdependencies: early assignments affect later ones by changing capacity and influencing emissions. To guide learning, the coordinator maintains a **reward table** $Q_t \in \mathbb{R}^N$, which estimates the value of prioritizing each traveller. After each round, the table is updated using a regret-weighted rule:

$$Q_t \leftarrow Q_t + (R_w \cdot \max(\rho_t, 0)) \quad (4.2)$$

$$Q_t \leftarrow \frac{Q_t}{\sum Q_t} \quad (4.3)$$

where ρ_t is the regret in round t , defined as:

$$\rho_t = R^t - \max_{k=1}^{t-1} R^k \quad (4.4)$$

In this expression R^t is the total system-wide reward obtained in round t , representing the aggregated reward from all travellers and $\max_{k=1}^{t-1} R^k$ is the highest total reward achieved in any round before t , that is, the best-known performance so far. The regret value ρ_t reflects how much worse (or better) the current round's performance is compared to the historical best. A positive ρ_t indicates improvement over past performance, while a negative value denotes underperformance.

The vector $R_w \in \mathbb{R}^N$ applies position-based weighting: $(R_w)_i = \frac{1}{s_i}$, where s_i is the position of traveller i in the priority sequence. This gives more learning weight to early decisions, which have greater downstream influence on system efficiency.

To simulate real-world variability in transport, Gaussian noise is applied to emissions at the start of each round (see [Subsection 4.2.3](#) for more details).

To address the exploration-exploitation trade-off in identifying effective priority orders, the coordinator employs an ε -greedy strategy with exponential decay (see [Subsection 4.2.4](#) for more details).

4.2.1 Traveller Path Utility Score

Travellers evaluate all the **paths** connecting their origin and destination. A path consists of a sequence of zones, and for each segment (i.e., each zone-to-zone transition), multiple transport modes may be available. Only *simple paths*, paths that do not revisit any zone, are considered and therefore they do not contain cycles. For every such path, the traveller considers all possible combinations of available transport modes across its segments. This results in a set of multimodal path options, where each option uses a specific transport mode for each segment. Only combinations that respect current transport capacities are included. Each traveller follows a **greedy strategy**: when it is their turn to act, they select the available path that maximizes a **utility score** u , defined in [Equation 4.5](#), based on their individual transport preferences. Once a path is selected, the transport capacity for each used segment is reduced by one unit. Every traveller is guaranteed to be assigned a path, even if it is suboptimal relative to their preference.

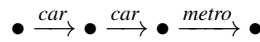
The utility score is computed as:

$$u = \frac{p}{(c+1) \cdot \ell^2} \quad (4.5)$$

where:

- u : utility score of the path for the traveller,
- p : sum of the traveller's preference values for the transport modes used in the path,
- c : number of times the transport mode changes between consecutive segments along the path,
- ℓ : total number of segments (zone-to-zone transitions) in the path.

For example:



In this case:

- p : depends on the traveller's preference vector. If the traveller has a preference of 70 for *car* and 50 for *metro*, then $p = 70 + 70 + 50 = 190$;
- $c = 1$: the mode changes once (from *car* to *metro*);
- $\ell = 3$: there are three zone-to-zone transitions.

This formulation is designed to balance preference satisfaction with path simplicity. It reflects three key priorities:

1. **Individual preference alignment:** A higher total preference p increases the score, favoring paths composed of modes the traveller prefers.
2. **Mode consistency:** The score penalizes frequent switching between transport modes through the $c + 1$ term, encouraging coherence in travel mode selection.
3. **Path compactness:** The squared length ℓ^2 strongly discourages longer, more circuitous routes.

As a result, travellers are guided toward selecting paths that best suit their personal preferences while implicitly favoring efficiency and structural simplicity.

4.2.2 Reward Function

The reward function is defined in terms of the **total emissions saved** in each round.

The reward R_i^t for traveller i in round t is computed based on the total emissions of the selected path P_i^t , normalized relative to a worst-case emission baseline:

$$R_i^t = 1 - \frac{\text{Em}(P_i^t)}{\text{Em}_{\text{baseline}}(P_i^t)} \quad (4.6)$$

where:

- $\text{Em}(P_i^t)$: sum of the emission values of the transport modes used in each segment of the path P_i^t ,
- $\text{Em}_{\text{baseline}}(P_i^t) = |P_i^t| \cdot e_{\text{car}}$: the emissions assuming all segments of the path used the highest-emission mode (car),
- $R_i^t \in [0, 1]$: with 1 representing an ideal zero-emission path.

Emission values for each transport mode are based on *well-to-wheel* estimates reported by the European Environment Agency [2], and are summarized in Table 4.1. Walking is treated as emission-free, as it produces no direct or energy-related emissions.

Table 4.1: Emissions by transport mode.

Transport Mode	Emission (gCO ₂ e per passenger-kilometre)
Car	160
Bus	80
Metro	33
Walking	0

This reward formulation was designed to reflect the **carbon efficiency** of the selected path. It is inversely proportional to the emissions produced: lower emissions result in higher rewards. Conceptually, the reward represents the system's ability to avoid emissions and thereby reduce the

social cost of carbon, which refers to the combined environmental, health, and economic impacts associated with CO₂ emissions. In real world terms, the social cost of carbon is influenced by factors such as distance travelled, travel time, and traffic congestion. Although these factors are not explicitly modeled here, they can be abstractly captured through a constant coefficient β , which represents the **marginal cost** per unit of emission saved. Under this interpretation, the reward approximates:

$$R_i^t \propto \beta \cdot (\text{Em}_{\text{baseline}}(P_i^t) - \text{Em}(P_i^t)) \quad (4.7)$$

where β is an unknown or variable factor reflecting externalities like distance or congestion, and the term $\text{Em}_{\text{baseline}} - \text{Em}$ quantifies the **emission savings** achieved by the selected path. Thus, the reward quantifies the system's effectiveness in reducing unnecessary emissions and mitigating their associated social and environmental impacts.

4.2.3 Environmental Dynamics

To capture day-to-day variability in transport efficiency, Gaussian noise is injected into the emissions matrix at the start of every round. This environmental randomization is applied to reflect the uncertainty of the environment itself and of human behaviour and its interactions with external conditions. In real-world urban settings, emissions for the same origin-destination path can vary due to a wide range of unpredictable factors, such as traffic signals, traffic congestion, weather conditions, interactions with other drivers or pedestrians, and delays at stops. These are externalities that are independent of the traveller's intention or control but nonetheless influence the overall emission outcome. This stochastic component models the fact that human decision-making is naturally subject to uncertainty. Even if a traveller follows the same origin-destination path each day, the resulting emissions are not guaranteed to be identical.

Let $Em \in \mathbb{R}^{n \times 4}$ be the base (deterministic) emissions matrix and σ^2 a variance hyper-parameter. A perturbed matrix Em' is produced as

$$\text{Em}'_{i,j} = \begin{cases} \text{Em}_{i,j} + \delta_{\text{OD}(i)}^{(\text{metro})} & j = 0 \quad (\text{metro}) \\ \text{Em}_{i,j} + \delta_{\text{OD}(i)}^{(\text{bus})} & j = 1 \quad (\text{bus}) \\ \text{Em}_{i,j} + \mathcal{N}(0, \sigma^2) & j = 2 \quad (\text{car}) \\ \text{Em}_{i,j} & j = 3 \quad (\text{foot}) \end{cases} \quad (4.8)$$

Here, $\text{OD}(i)$ denotes the origin–destination pair of traveller i . For each OD pair, the scalars

$$\delta_{\text{OD}}^{(\text{metro})} \sim \mathcal{N}(0, \sigma^2), \quad \delta_{\text{OD}}^{(\text{bus})} \sim \mathcal{N}(0, \sigma^2) \quad (4.9)$$

are sampled *once per OD pair* at the start of each round. That is, all travellers sharing the same OD pair will see the same metro and bus perturbations. Thus:

- **Metro and bus:** The same offset is applied to all travellers with the same OD pair, reflecting shared operational characteristics for each route segment.
- **Car:** A different noise sample is drawn for each traveller i , modelling variation in individual vehicles and routes.
- **Foot:** Walking remains deterministic with zero emissions.

Negative values generated by the noise are clipped to zero. Introducing stochastic emissions in this structured way forces the learning algorithm to generalize across fluctuating conditions while respecting the physical differences between collective (metro, bus) and individual (car, foot) transport modes.

4.2.4 Epsilon-Decay Sampling

The ϵ -greedy algorithm is employed to balance exploration and exploitation. At each step:

- with probability ϵ , the coordinator selects a random unassigned traveller (**exploration**);
- with probability $1 - \epsilon$, the traveller with the highest value in Q_t is selected (**exploitation**).

To improve long-term performance, an **exponential decay** is applied to ϵ over time, following:

$$\epsilon_t = \epsilon_0 \cdot \exp\left(\frac{\ln(\epsilon_{\min}/\epsilon_0)}{T_{\text{target}}} \cdot t\right), \quad (4.10)$$

where ϵ_0 is the initial exploration rate, $\epsilon_{\min}$ is the target minimum value, and T_{target} specifies the number of training rounds over which ϵ is expected to decay toward $\epsilon_{\min}$. After T_{target} , ϵ may continue decreasing slightly, following the same exponential trend.

This adaptive schedule ensures sufficient exploration in the early stages of training, while gradually shifting toward exploitation as learning progresses. The use of a *decaying exploration strategy* is inspired by empirical results from the DC-MATA framework [14], which highlighted ϵ -decay as a consistently robust approach in varied and noisy environments. It enabled reliable convergence to near-optimal solutions, even under high variance, by effectively balancing the need for early exploration with increasing confidence in the learned decision priorities.

Chapter 5

Empirical Demonstration

This chapter presents a series of simulation-based experiments designed to evaluate the effectiveness, scalability, and adaptability of the proposed methodology in supporting emission-aware traveller assignment. By simulating different city layouts, transport infrastructures, and population distributions, the experiments aim to reflect the varied aspects of urban mobility systems and test how the learning agent performs under practical constraints. The scenarios also allow us to explore the influence of behavioural incentives and algorithmic design choices on both system efficiency and sustainability outcomes. Together, these demonstrations provide evidence of how adaptive coordination mechanisms can contribute to the goal of reducing transport emissions in complex urban environments.

5.1 Experimental Setup

The experiments were carried out using multiple synthetic city configurations. Each experiment runs for **1000 rounds**, during which the learning agent updates traveller priority weights based on regret and observed outcomes. These scenarios simulate diverse city configurations, traveller populations, and transport conditions, allowing us to observe the system’s performance under varying structural and behavioural constraints.

As stated before, each traveller $i \in I$ is defined by:

- a distinct **origin-destination (OD)** pair of zones,
- a personalised **preference vector** over the four transport modes.

The preferences are encoded as integer values ranging from 1 to 100, where higher values represent stronger preferences. Specifically, for each traveller and each transport mode, a value is sampled independently within a predefined range. These intervals reflect assumptions about perceived utility and general behavioural trends. Public transport modes (metro and bus) are assigned moderate ranges (30–70), representing varying but typically non-extreme attitudes toward shared transport. Car is given a wider and higher range (40–100) to capture stronger preferences often associated with private transport in urban settings. Foot travel receives a broader but generally lower

range (1–50), recognizing that while some individuals strongly favor walking, many consider it a less preferred option. Origin-destination (OD) pairs are created from all valid combinations of distinct zones. To reach the total number of travellers, these pairs are replicated in proportion and then randomly ordered, ensuring a balanced and diverse distribution of trips across the network.

In every scenario, approximately half of travellers are designated as car owners, reflecting the European average of around 500 cars per 1,000 inhabitants [18].

Each scenario varies in the following parameters:

- **Population size:** 360 to 1540 individuals,
- **Number of zones:** 4-zones and 6-zones city representations,
- **Connection Type:** including fully connected layouts, topologies with missing diagonal connections, and configurations with limited access to public transport.

To contextualize the spatial constraints influencing learning, Figure 5.1 presents the network connection types used in the experiments, including full connectivity, no-diagonal constraints, and PT-limited access. For example, in a 4-zone network with no diagonal connections, a traveller cannot go directly from zone 1 to zone 4; instead, such trips require passing through intermediate zones (e.g. zones 2 or 3), increasing coordination complexity. In the visualizations, blue links represent car and walking paths, pink denotes metro, and red corresponds to bus.

Table 5.2 summarizes the capacity limits associated with each transport mode in the simulation. Public transport modes have fixed capacity limits, but car and walking are assumed to have unlimited availability, so anyone can drive their own car if they have it.

Table 5.1: Transport Mode Capacity Limits.

Transport Mode	Capacity
Metro	50 travellers per link
Bus	20 travellers per link
Car	Unlimited
Walking	Unlimited

Table 5.2: Transport Mode Capacity Limits.

Transport Mode	Preference Range
Metro	30-70
Bus	30-70
Car	40-100
Walking	1-50

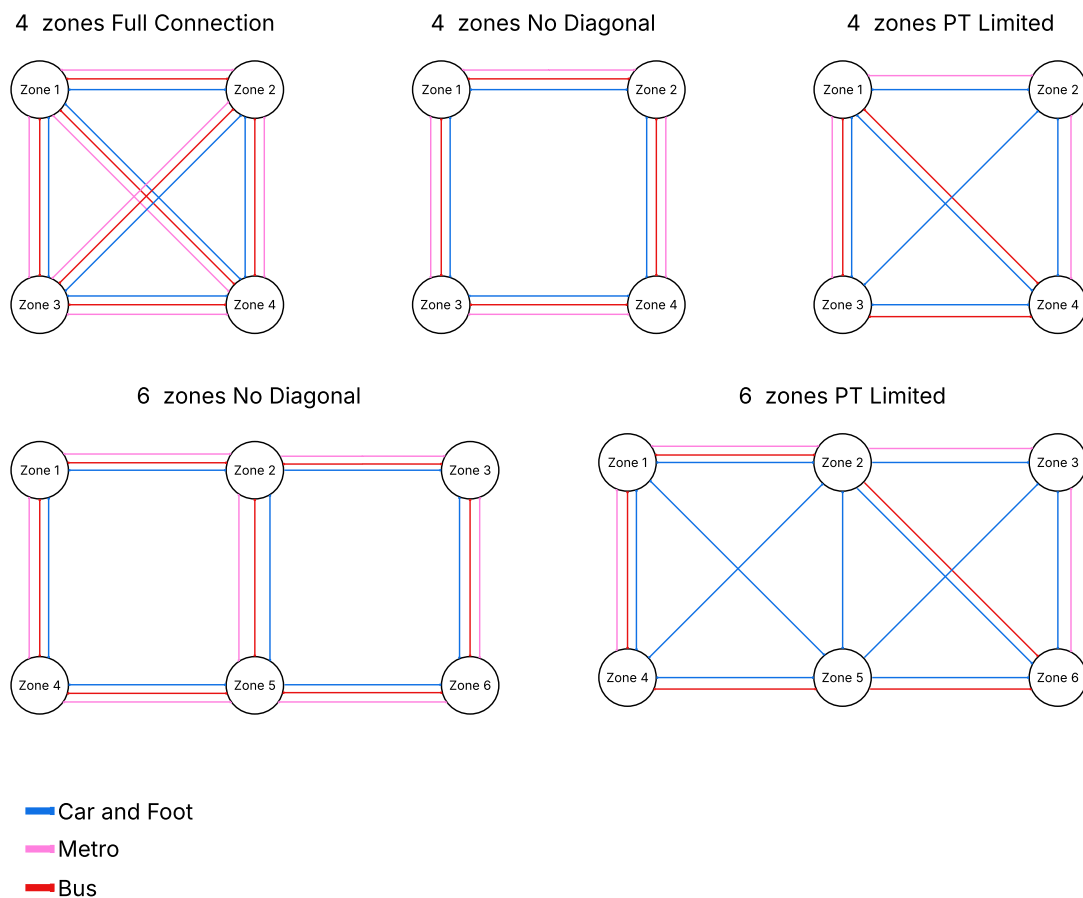


Figure 5.1: Illustration of the Connection Types

5.2 Evaluation Objectives

These metrics are recorded and saved for each scenario, enabling comparative analysis and visualizations in the subsequent results sections.

The experimental setting has three distinct objectives, each testing a different aspect of the suggested solution. These studies evaluate:

1. Evaluating the Learning-Based Coordination Strategy for Emission Reduction
2. Influence of Behavioural Incentives on Transport Mode Choices
3. Ablation Study on the Exploration Parameter ϵ and Decay Strategies

5.3 Experiment 1: Evaluating the Learning-Based Coordination Strategy for Emission Reduction

The primary focus of this work is to evaluate whether the MAB-based coordinator can effectively reduce total transport emissions by learning optimal priority orders for assigning travellers. In addition, the experiments investigate how the coordination strategy adapts to different network structures and population sizes. This allows us to assess the scalability of the framework, its robustness to increasing complexity, and its sensitivity to variations in transport connectivity and capacity constraints.

All tests in this evaluation were conducted using an exploration rate of $\epsilon = 0.9$ with an exponential decay strategy.

For each scenario, Table 5.3 depicts a board summarizing the performance of each learning-based coordination strategy in terms of reward progression and emission reduction. It reports the initial and final moving average rewards, emission levels at the start and end of the simulation, and their respective differences. Additionally, it includes the emission from a baseline scenario where all individuals exclusively use cars (All by Car) to go from their origin to their destination, providing a reference value representing the maximum potential emissions for each scenario.

Table 5.3: Performance summary for all experimental scenarios.

Population Size	Number Zones	Connection Type	First MA Reward	Last MA Reward	Reward Δ	First Emission	Last Emission	Emission Δ	Emission All by Car
1540	6	PTlimited	727.44	746.9	2.67%	174800.03	169958.63	-2.77%	320134.66
1540	6	noDiagonalConnect	734.14	760.4	3.58%	217391.93	208046.88	-4.3%	410678.38
1080	6	PTlimited	485.07	497.94	2.65%	128168.39	124775.34	-2.65%	224618.06
1080	6	noDiagonalConnect	504.18	529.52	5.03%	156033.7	145938.53	-6.47%	287968.27
1080	4	PTlimited	500.53	516.84	3.26%	95138.81	92329.68	-2.95%	172785.34
1080	4	fullConn	493.66	497.81	0.84%	94394.32	93200.31	-1.26%	172785.34
1080	4	noDiagonalConnect	522.44	545.24	4.36%	119402.91	113568.48	-4.89%	230361.5
840	4	PTlimited	385.1	398.06	3.37%	75615.91	72306.71	-4.38%	134394.85
840	4	fullConn	371.49	376.53	1.36%	75380.89	74270.85	-1.47%	134394.85
840	4	noDiagonalConnect	393.22	410.9	4.5%	96761.32	92569.06	-4.33%	179201.4
560	4	PTlimited	241.92	243.85	0.8%	51897.04	51825.54	-0.14%	89600.16
560	4	fullConn	235.94	235.86	-0.03%	51802.83	51857.12	0.1%	89600.16
560	4	noDiagonalConnect	247.84	253.27	2.19%	66166.08	65035.49	-1.71%	119678.39
360	4	PTlimited	151.44	151.53	0.06%	34061.2	34078.96	0.05%	76805.58
360	4	fullConn	151.97	151.94	-0.02%	33250.34	33308.51	0.17%	57607.4
360	4	noDiagonalConnect	156.05	159.72	2.35%	43578.9	42476.4	-2.53%	76805.58

Scenarios with larger populations naturally yield higher absolute values for both reward and emissions, as these metrics are calculated by summing the outcomes of each individual traveller's choice. More people making transport decisions leads to a greater cumulative impact on both system performance and environmental output.

At the start of the first round, the coordinator has no prior knowledge about the travellers or their expected rewards. The internal reward table, used to estimate the value of each traveller as an arm in the multi-armed bandit formulation, is initialized to zero for all travellers. As a result, during the initial round, exploitation behaviour (i.e., selecting the most promising traveller based on estimated rewards) is uninformed. When all travellers have equal estimated rewards (i.e., all zeros), the exploitation strategy simply defaults to selecting travellers in the order they are stored in memory. Exploration acts as normal, randomly selecting an available traveller.

Figure 5.2 presents a line plot for each population size, illustrating the evolution of the moving average of the total reward per round across different scenarios. The total reward represents the sum of the individual rewards associated with each traveller's path choice in a given round. The shaded region around each line depicts the variability (standard deviation) of the moving average, calculated using a rolling window of 30 rounds. This visual band tends to be wider at the beginning and gradually narrows over time. This occurs because the exploration rate ϵ decays as learning progresses, reducing the randomness in decision-making. As a result, the coordinator's behaviour stabilizes, and the total reward becomes more consistent across rounds, leading to smaller variance and a tighter shaded region.

The evolution of total emissions, calculated as the sum of the emissions associated with the chosen path of each traveller per round, and the evolution of transport usage, which refers to the number of travellers selecting each transport mode (metro, bus, car, or walking) at each round, are presented in the Appendix A.

Population of 1540

In the **six-zone network without diagonal connections**, the moving average reward improved from 734.14 to 760.40, a gain of 26.26 units (3.58%). Emissions decreased from 217,391.93 to 208,046.88, a 4.30% reduction (9,345 gCO₂e). Car usage dropped by 61 users, foot travel increased by 49, metro increased by 4, and bus remained stable.

In the **PT-limited six-zone scenario**, the reward increased by 19.46 units (2.67%), and emissions dropped by 4,841 gCO₂e (2.77%). Car usage decreased by 30 users, foot increased by 33, metro declined by 5, and bus usage remained unchanged.

Population of 1080

In the **fully connected four-zone** configuration with generated preferences, the system exhibited a very slight learning effect. The moving average reward rose from 493.66 to 497.81, a 0.84% increase (4.15 units). Emissions decreased from 94,394.32 to 93,200.31, yielding a 1.26% reduction. This modest progress was accompanied by a decline of 9 users in car usage, a gain of 5 walkers, and a small rise in metro usage.

In the **no-diagonal four-zone network**, the reward increased significantly, from 522.44 to 545.24—a 4.36% improvement. Emissions dropped by 5,834 gCO₂e (4.89%), with a corresponding 37-user drop in car use and an equivalent rise in foot travel.

In the **PT-limited four-zone scenario**, the final reward was 16.31 units higher than the initial, while emissions fell by 2.95% (2,809 gCO₂e). Car usage declined by 18 users, and walking increased by 26.

The **six-zone network without diagonal links** saw rewards rise from 504.18 to 529.52 (5.02%), and emissions dropped by 10,095 gCO₂e (6.47%). Metro usage decreased by 3 users, bus usage increased by 1 user, car usage dropped by 63 users, and foot traffic grew by 63 users.

In the **PT-limited six-zone scenario**, the reward increased by 12.86 units (2.65%), and emissions were reduced by 3,393 gCO₂e (2.65%). Metro increased by 2 users, bus decreased by 3, car decreased by 20, and foot increased by 17 users.

Population of 840

In the **fully connected four-zone network**, the average reward increased from 371.49 to 376.53, a 5.04-unit gain (1.36%). Emissions decreased by 1,110 gCO₂e (1.47%). Car usage declined by 9 users, foot increased by 4, and metro increased by 6.

In the **no-diagonal four-zone configuration**, the reward increased by 17.68 units (4.50%) and emissions dropped by 4,192 gCO₂e (4.33%). Car usage decreased by 27 users, foot traffic increased by 30, and metro increased by 3.

In the **PT-limited four-zone scenario**, reward rose by 12.97 units (3.37%), and emissions decreased by 3,309 gCO₂e (4.38%). Car usage dropped by 20 users, foot increased by 16, metro declined by 6, and bus remained unchanged.

Population of 560

In the **fully connected four-zone scenario**, both reward and emissions remained effectively **unchanged**. Reward showed a trivial variation (-0.08 units), and emissions increased slightly by 54 gCO₂e (0.10%), due to variance applied per round. Transport mode usage did not change.

In the **no-diagonal four-zone configuration**, rewards increased by 5.43 units (2.19%), and emissions dropped by 1,130 gCO₂e (1.71%). Car usage decreased by 6, foot increased by 3, metro increased by 5, and bus dropped by 4.

In the **PT-limited scenario**, reward increased by 1.93 units (0.80%), and emissions decreased slightly by 71.50 gCO₂e (0.14%). Metro increased by 11, while bus, car, and foot usage slightly decreased.

Population of 360

In the **fully connected four-zone network**, there was no observable learning or emission change. Reward remained nearly flat and usage across all modes remained fixed.

In the **no-diagonal configuration**, reward increased by 3.67 units (2.35%), and emissions decreased by 1,102 gCO₂e (2.53%). Car usage decreased by 4, foot increased by 5, metro increased by 9, and bus dropped by 10.

In the **PT-limited scenario**, reward and emissions stayed nearly constant. No shift was observed in mode usage.

In the first and third cases, the number of travellers is not sufficient to create consistent conflicts or competition over transport modes. As a result, the coordination system faces fewer impactful decisions, reducing the opportunities to explore and refine effective traveller orderings. Although small differences in reward and emission values are reported in the table, like -0.02% and 0.06%, these variations are not indicative of actual learning. Instead, they stem from random fluctuations caused by the variability applied to transport mode emissions in each round. This is evident

when observing the corresponding emission and reward plots, where minor changes occur without sustained trends, confirming the plateaued behaviour typical in low-demand environments.

Observations and Conclusions

Across all population groups, a consistent pattern emerges: scenarios with greater structural constraints, such as limited connections or public transport, result in more pronounced learning and stronger behavioural shifts.

Results consistently show that emission reductions and reward improvements are more significant in scenarios with greater spatial constraints, structural complexity, or larger populations. Spatially constrained scenarios, such as those without diagonal connections (**noDiagonalConnect**) between zones, limit the number of possible routes, forcing travellers to depend on fewer direct links. Structurally complex setups, like six-zone networks or configurations with limited access to public transport (**PTlimited**), add further restrictions by increasing the scale or limiting available transport options.

For a fixed population of 1080 individuals in a four-zone setup, reward improvement in the fully connected scenario is just **0.84%**, whereas in the **noDiagonalConnect** and **PTlimited** scenarios it jumps to **4.36%** and **3.37%**, respectively. Emissions follow the same trend: the fully connected case achieves a reduction of only **1.26%**, compared to **4.89%** and **2.95%** in the more constrained configurations.

Larger populations also contribute to improved learning. Even when the number of zones and the network structure remain the same, a greater number of travellers competing for limited transport capacity increases system pressure and intensifies the need for coordination. For example, in the **4zones-noDiagonalConnect** configuration, the reward improvement for 360 individuals is just **2.35%**, with an emission reduction of **2.53%**. With 840 individuals, the reward gain grows to **4.50%**, and emissions are reduced by **4.33%**. At 1080 individuals, the improvements become even more pronounced: reward increases by **4.36%** and emissions drop by **4.90%**. A similar pattern emerges in the **PTlimited** scenario: for 360 people, both reward and emission changes are negligible, whereas for 1080 people, the reward increases by **3.37%** and emissions are reduced by **2.95%**. These results highlight how larger populations intensify constraints and make the order of traveller assignments more influential—providing the learning agent with richer opportunities to discover effective coordination strategies.

A similar effect is observed when comparing the number of zones while keeping the population and connection type constant. In the **noDiagonalConnect** configuration with 1080 individuals, increasing from four zones to six zones results in a slightly greater reward improvement—from **4.36%** to **5.02%**—and a much stronger emission reduction—from **4.89%** to **6.47%**.

These characteristics encourage better learning because they make it harder for travellers to follow their default preferences. In less constrained environments, such as fully connected four-zone networks, most individuals can reach their destinations using their top-ranked transport modes with little interference, resulting in low coordination pressure. However, as constraints increase, the preferred options are often no longer directly available or efficient, especially under

capacity limitations. In these conditions, the coordinator is exposed to a broader decision space and more opportunities to learn how different priority orders affect the overall outcome. Over time, the system adapts by shifting users away from high-emission modes, such as private cars, and promoting more sustainable choices like walking or public transport, maximizing collective efficiency and minimizing emissions.

In highly connected environments with fewer constraints (**fullConnection**), traveller behaviour tends to stabilize quickly due to the limited need for coordination and fewer decision-making opportunities. This results in stable patterns of reward, emissions, and transport usage over time. Similarly, in some scenarios with smaller populations, the system may not exhibit strong adaptation, as the lower number of travellers reduces the complexity of interactions and limits the potential for learning. Although slight improvements in emissions can still be observed in these stable cases, the magnitude of change is modest. It is also important to note that minor fluctuations in emission values across rounds are expected due to the variance applied to the emissions during simulation and do not necessarily indicate meaningful behavioural change.

Overall, the results validate the role of environmental pressure in driving intelligent, sustainable adaptation in multi-agent systems.

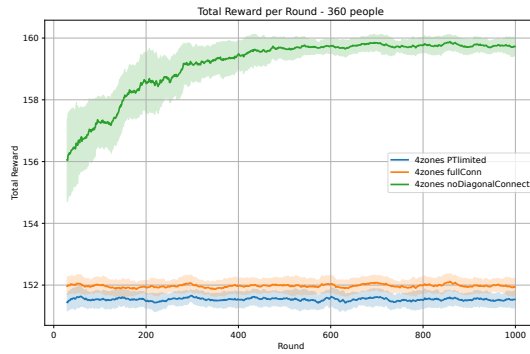
5.4 Experiment 2: Influence of Behavioural Incentives on Transport Mode Choices

This set of experiments explores how external incentives can influence transport behaviour and reduce emissions. Each incentive scenario reuses the same city configuration and traveller population structure as the baseline but adjusts the preference vectors accordingly. The learning process and assignment mechanisms remain unchanged. The complete set of graphics for all scenarios can be found in **Appendix B**. To simulate the effect of real-world policy changes, traveller preferences were modified in two distinct ways:

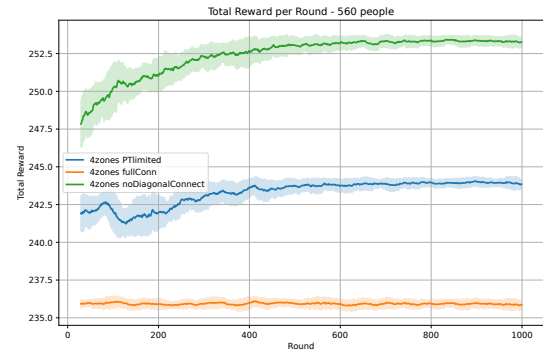
- **Free Public Transport (FreePT):** To represent a intervention where the city provides a universal free pass for all residents, each traveller's preference values for metro and bus were doubled. This simulates increased attractiveness and perceived utility of public transport options.
- **Increased Fuel Costs (expensiveFuel):** To reflect scenarios where fuel becomes more expensive (e.g., due to taxation or market fluctuations), preference values for car usage were reduced by half, making private transport less appealing.

These incentives demonstrate distinct impacts depending on both the connectivity of the transport network and the size of the population. The detailed results for reward progression, emissions, and transport mode usage under different incentive scenarios are presented in **Appendix A**.

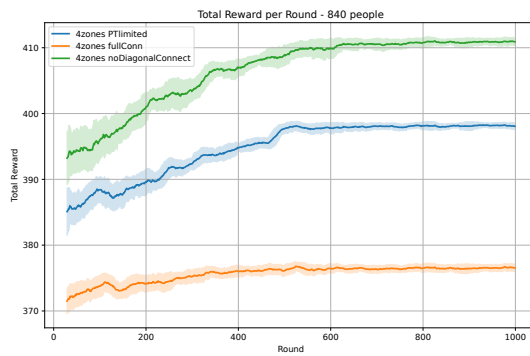
In networks with limited connectivity, such as **noDiagonalConnect** and **PTlimited** layouts, both incentive strategies successfully reduced car usage. However, the effects differed in nature.



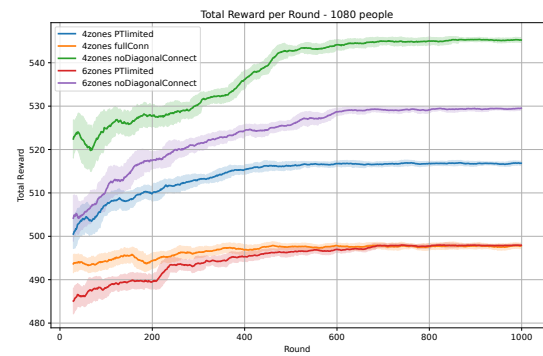
(a) 360 people



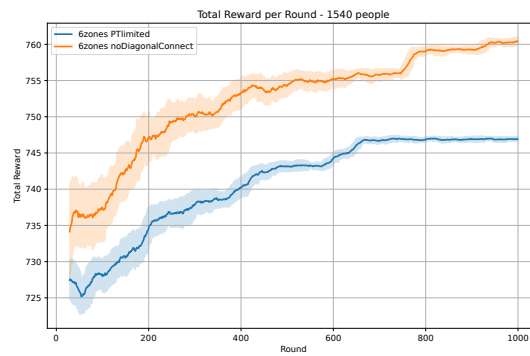
(b) 560 people



(c) 840 people



(d) 1080 people



(e) 1540 people

Figure 5.2: Moving average of rewards in different population sizes.

FreePT primarily encouraged increased use of public transport modes (metro and bus) wherever they were accessible. In contrast, when public transport options were more restricted, the **expensiveFuel** scenario prompted a more noticeable shift toward walking.

In fully connected networks, the **FreePT** incentive consistently outperformed **expensiveFuel** in terms of both emissions reduction and cumulative reward. The broad availability of metro and bus options in these settings allowed travellers to take full advantage of the increased attractiveness of public transport. On the other hand, **FreePT** exhibited diminishing returns in PT-limited environments, where travellers could not leverage the incentive effectively due to limited availability and coverage of public transport options.

Population size also played a critical role in the system’s responsiveness to incentives. In larger populations (e.g., 1080 or 1540 individuals), the results show statistically significant behavioural shifts, including greater reductions in emissions and car dependency. These populations offer more behavioural diversity and learning opportunities, enabling the coordinated assignment strategy to more effectively detect and exploit patterns aligned with sustainable behaviour. Conversely, in smaller populations (e.g., 360 individuals), the influence of incentives was limited. The reduced number of travellers leads to less variability in decision-making and weakens the learning signal necessary for adaptation.

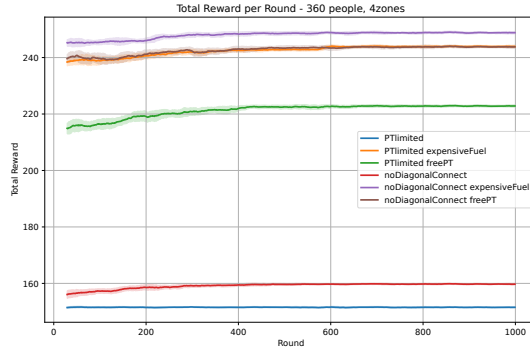
In summary, **FreePT** proves most effective in well-connected networks, where travellers can easily switch to public transport. In contrast, **expensiveFuel** tend to be more robust in constrained environments, where walking remains a universally accessible low-emission alternative.

5.5 Experiment 3: Ablation Study on the Exploration Parameter ϵ and Decay Strategies

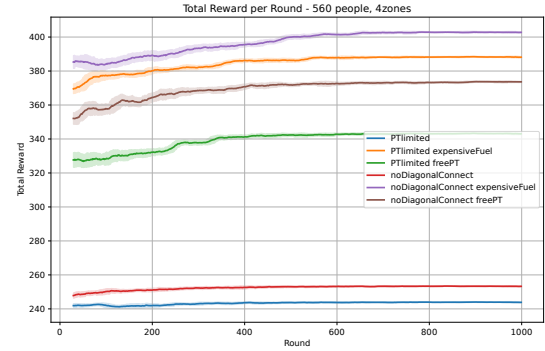
The goal of this ablation study is to evaluate how different exploration strategies influence the performance of the learning system. Specifically, we analyze the role of the ϵ parameter in the ϵ -greedy sampling mechanism, and how its decay over time affects outcomes. Three decay strategies were implemented and compared: linear decay, exponential decay, and simulated annealing. The complete set of graphics for all scenarios can be found in **Appendix C**.

Linear Decay: $\epsilon_t = \max(0, \epsilon_0 - k \cdot t)$, where k is chosen such that ϵ reaches zero after approximately 80% of the total rounds. This results in a gradual and uniform decrease in exploration probability.

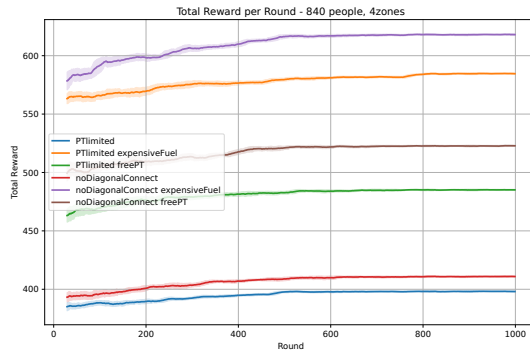
Exponential Decay: $\epsilon_t = \epsilon_0 \cdot \alpha^t$, with $\alpha < 1$. The decay factor is calculated so that ϵ reaches a small threshold (e.g., 0.001) after 80% of the total rounds. This strategy allows for a slower decay at first, with exploration decreasing more sharply in later rounds.



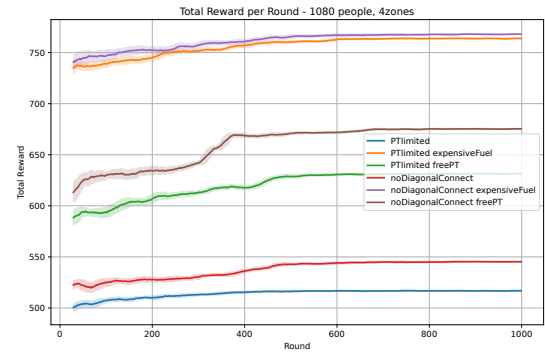
(a) 360 people, 4 zones



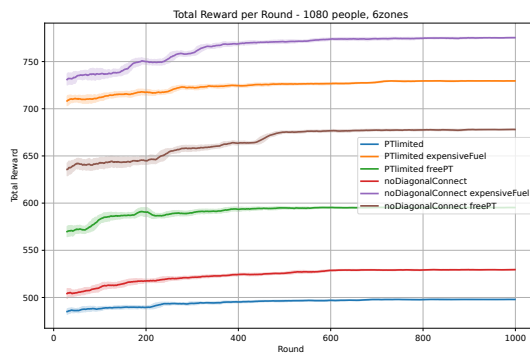
(b) 560 people, 4 zones



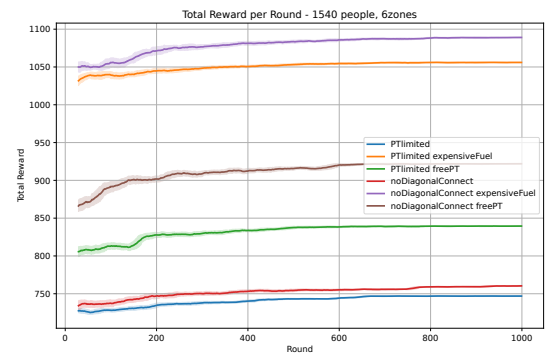
(c) 840 people, 4 zones



(d) 1080 people, 4 zones



(e) 1080 people, 6 zones



(f) 1540 people, 6 zones

Figure 5.3: Moving average of reward in different population sizes under three incentive conditions.

Simulated Annealing: This strategy uses a temperature-based acceptance probability, initialized with a temperature derived from ε using a fixed mapping. The temperature decays multiplicatively via a cooling factor (typically 0.996). The probability of selecting a suboptimal action is governed by the Boltzmann probability based on the regret at each round.

To configure Simulated Annealing (SA) to behave similarly to the epsilon-greedy strategy at the start of training, we set the initial probability of accepting a worse solution in SA equal to the initial epsilon value. In SA, the probability of accepting a worse solution is given by:

$$P_{\text{accept}} = \exp\left(\frac{\Delta R}{T}\right) \quad (5.1)$$

Where ΔR is the change in reward and T is the current temperature and P_{accept} is the probability of accepting a worse solution.

We assume a typical reward decrease of $\Delta R = -0.5$ and solve for the temperature T that yields a target probability ε of accepting such worse solutions:

$$\varepsilon = \exp\left(\frac{-0.5}{T}\right) \Rightarrow T = \frac{-0.5}{\ln(\varepsilon)} \quad (5.2)$$

Using this formula, we compute the following initial temperatures corresponding to different values of ε :

Table 5.4: Initial temperatures values derived from epsilon.

Epsilon (ε)	Initial Temperature (T)
0.3	0.415
0.5	0.721
0.7	1.402
0.9	4.744

Each strategy is tested over identical simulation conditions (same initial preferences, city layout, and transport network), and their convergence behaviour, stability, and final emission outcomes are compared.

5.5.1 By Population

Population of 1080

In the configuration with 1080 people and four zones, strategies using low epsilon ($\varepsilon = 0.3$) with exponential decay performed poorly. This is because exploration ends too early in a large action space, preventing the algorithm from discovering better strategies. In the variant with six zones, the challenges intensified for strategies with high epsilon ($\varepsilon = 0.9$) and linear decay. These configurations maintained high exploration for too long and failed to converge effectively within the given number of rounds. Overall, almost all linear decay strategies in this scenario would likely need more iterations to achieve competitive performance.

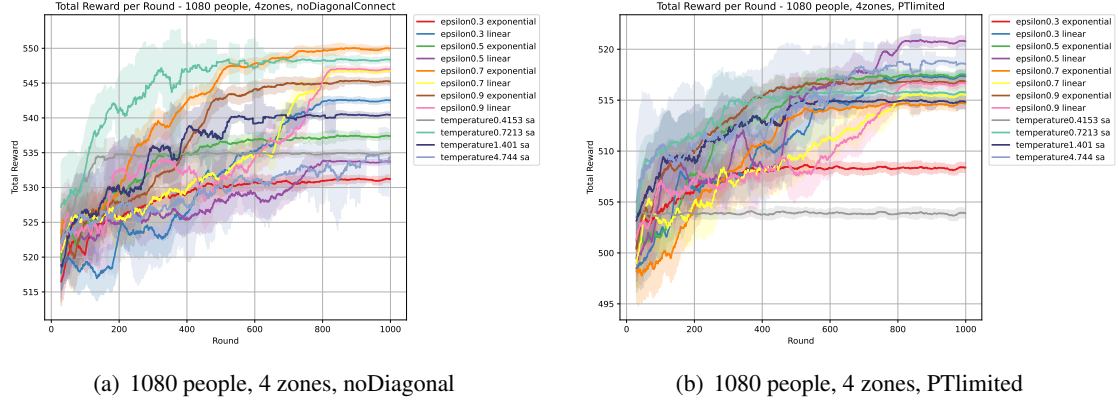


Figure 5.4: Moving average of rewards for 1080 people in four-zone under different ϵ -decay strategies.

Population of 840

With an increased population of 840, the problem space becomes more complex and the impact of exploration strategies becomes more evident. Simulated annealing with low temperatures (0.4153 and 0.7213) performed significantly worse, suggesting that these configurations converge too early without sufficient exploration. This issue did not manifest in the 360-person case, highlighting how performance is sensitive to problem complexity.

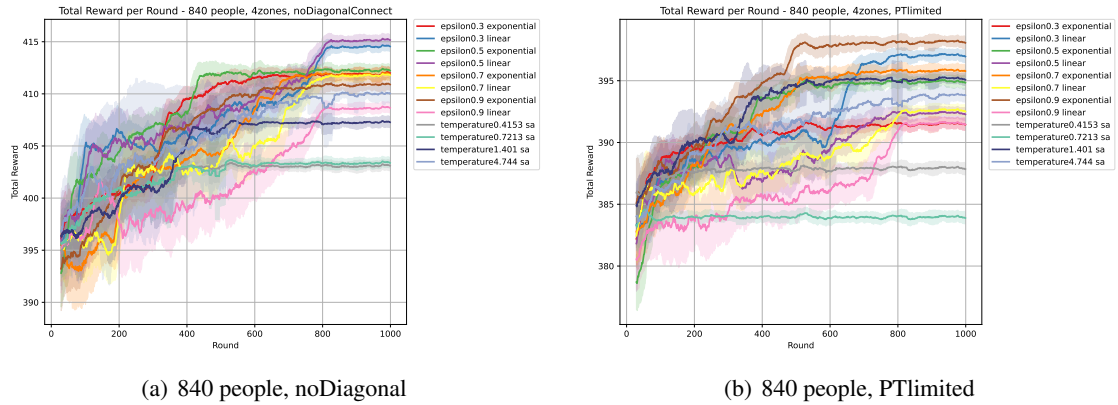


Figure 5.5: Moving average of rewards for 840 people in four-zone under different ϵ -decay strategies.

Population of 360

In the 360-person scenarios, we observed a relatively small gap between initial and final rewards. This is due to the environment's low complexity, with fewer zones and fewer possible configurations, which simplifies the agent's learning process. Most strategies converged quickly

regardless of exploration parameters. The exponential decay strategy consistently stood out, frequently achieving the highest final rewards across scenarios. While linear decay showed competitive performance for $\epsilon=0.7$, its other configurations generally underperformed. Interestingly, simulated annealing with low temperatures (e.g., 0.4153 or 0.7213) was not particularly detrimental in this population, unlike in more complex settings.

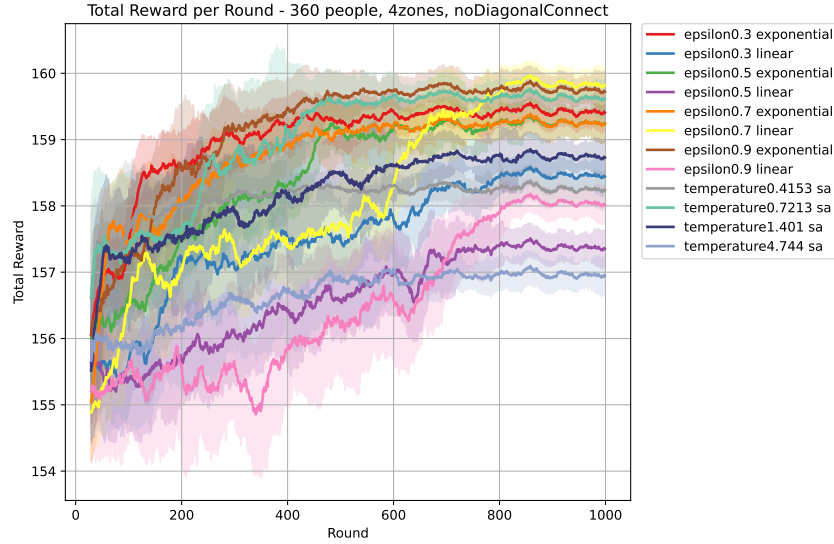


Figure 5.6: Moving average of rewards for 360 people under different ϵ -decay strategies.

5.5.2 By Decay Strategy

Linear Decay

The linear decay strategy exhibited mixed performance across the different experimental conditions. Its strength was most evident when paired with moderate to low epsilon values (0.3–0.5). In these cases, it enabled stable convergence, particularly in structured environments like PTlimited, and in mid-complexity populations (e.g., 560 and 840 people). Notably, the combination of $\epsilon = 0.3$ with linear decay ranked among the top-performing strategies overall. However, linear decay showed clear limitations in high-complexity environments, especially with high epsilon (e.g., $\epsilon = 0.9$) or in scenarios with many zones. It maintained exploration for too long, creating instability and preventing effective convergence. These shortcomings were particularly evident in the 1080 population with 6 zones, where most linear configurations underperformed. In summary, linear decay is effective for simple or moderately complex problems, but may require longer training horizons or careful tuning in more dynamic environments.

Exponential Decay

It consistently performed well across a diverse set of environments, when paired with high epsilon value (e.g., $\epsilon = 0.9$). This pairing allowed exploration to persist during the early rounds and taper off more gently, a behaviour that was especially beneficial in complex scenarios with many

zones, large populations, or limited connectivity (e.g., noDiagonalConnect). However, exponential decay was not universally optimal. When combined with low epsilon (e.g., $\epsilon = 0.3$ or 0.5), it led to poor performance, as the decay compounded the already limited exploration and caused early convergence to suboptimal solutions. These combinations appeared among the worst overall. Therefore, while exponential decay is highly effective with strong initial exploration, it must be paired carefully to avoid premature exploitation in sparse-search problems.

Simulated Annealing

Simulated annealing (SA) displayed more variable results, and in general, it was less competitive than epsilon-decay strategies. Its performance was highly sensitive to the temperature parameter: low temperatures (e.g., 0.4153 and 0.7213) consistently led to early convergence and poor rewards, especially in more complex settings like the 840 and 1080 populations. On the other hand, moderate to high temperatures (e.g., 1.401 and 4.744) were more effective, occasionally achieving top-3 performance in structured layouts like PTlimited.

5.5.3 Epsilon-Based Combinations

The top three combinations were: (i) $\epsilon = 0.9$ with exponential decay, (ii) $\epsilon = 0.3$ with linear decay, and (iii) $\epsilon = 0.5$ with linear decay. The first combination maintained high exploration for longer, which was especially beneficial in complex scenarios with larger populations and restricted connectivity. The second and third combinations, while more conservative, performed particularly well in structured environments, offering stable convergence and solid final rewards. These results highlight that both sustained exploration and early stabilization can be effective, depending on the environment's complexity.

The three worst-performing combinations using epsilon (excluding simulated annealing) were: (i) $\epsilon = 0.3$ with exponential decay, (ii) $\epsilon = 0.9$, linear decay, and (iii) $\epsilon = 0.5$, exponential decay. In the first and third, the exploration ended too early, leading to poor performance in complex environments. For the second, exploration persisted too long, destabilizing the learning process.

These results emphasize that both the intensity of exploration and the decay schedule must be well-matched to the problem's complexity. Intermediate configurations often performed inconsistently unless precisely tuned.

5.6 Summary

The experiments were designed to explore how the learning agent behaves under varying environmental complexities, diverse network topologies, incentive conditions, and different exploration configurations. By systematically adjusting parameters such as population size, transport connectivity, and algorithmic exploration mechanisms, the study provides an evaluation of the robustness, adaptability, and effectiveness of the proposed coordination strategy across a diverse and representative set of urban mobility scenarios.

The first experiment focused on evaluating the emission reduction potential of the proposed learning-based coordination strategy. The results show that this methodology consistently improves overall emission outcomes, particularly in structurally constrained scenarios (e.g., areas with limited public transport coverage or networks lacking diagonal zone connections). Larger populations amplified the learning signal, enabling more effective prioritisation, better coordination, and greater system-wide emission savings. These findings also demonstrated the system's scalability and robustness, as improvements became more pronounced in denser, more populated, and more structurally complex environments.

The second experiment assessed the impact of behaviour-oriented incentive strategies, such as free public transport and increased fuel costs, on transport mode choice and carbon emissions. The findings indicate that such external incentives can influence individual and collective transport behaviour, especially when they are well-aligned with the structural characteristics of the transport network. Free public transport emerged as the most effective strategy in highly connected environments where public options are widely accessible, while increased fuel costs had a stronger behavioural impact in constrained settings, where limited options forced reconsideration of car usage. Population size also played a notable role: larger populations responded more strongly and consistently to incentive shifts, whereas smaller populations showed limited behavioural adaptation due to weaker competition for resources and reduced learning signals over time.

Finally, the third experiment explored how the configuration of the ϵ parameter and its decay strategy directly affects the learning performance and system stability. Among the strategies tested, high initial exploration with exponential decay ($\epsilon = 0.9$) demonstrated strong and reliable performance across structurally complex environments. Linear decay proved effective when paired with lower ϵ values in more stable or simpler network settings, allowing for early convergence without sacrificing too much exploration. Simulated annealing achieved good results only when the initial temperature was carefully set high enough to allow sufficient early-stage exploration, as configurations with lower starting temperatures caused premature convergence to suboptimal solutions. These results collectively highlight the critical importance of aligning exploration strategies with the complexity of the environment, ensuring a balance between learning flexibility and convergence efficiency throughout the simulation horizon.

Chapter 6

Conclusions

This chapter summarises the key contributions and findings of the dissertation, reflects on how the proposed approach addresses the defined research questions, and outlines promising directions for future research. The work presented here contributes to the development of emissions-aware mobility coordination through reinforcement learning, with a focus on practical implementation in multimodal urban transport systems.

6.1 Main Contributions

This dissertation proposed a reinforcement learning–based coordination framework to support carbon-efficient decision-making in urban multimodal transport systems. Motivated by the growing environmental challenges posed by urban mobility, this work addresses the need for adaptive and preference-aware systems capable of reducing greenhouse gas emissions.

By modelling the city as a network of zones and travellers with personalised preferences, the coordination challenge was framed as a Multi-Armed Bandit problem. A central learning agent determines the traveller priority order for path assignment, learning to favor sequences that yield lower cumulative emissions.

Extensive experiments demonstrated the system’s capacity to reduce emissions and shift mobility behaviours toward more sustainable options, especially under structural limitations where learning plays a critical role. The integration of policy incentives, such as free public transport and increased fuel costs, further illustrated the model’s adaptability to external interventions. Additionally, a sensitivity analysis on exploration strategies confirmed that exponential decay of ϵ offers the most stable and effective balance between exploration and exploitation.

In summary, this work advances the use of reinforcement learning for sustainable urban mobility by proposing a practical, scalable, and environmentally aware coordination strategy. The proposed model can help city planners and policymakers in simulating, evaluating, and promoting low-emission transport behaviours while respecting individual preferences. Future directions include extending the framework to decentralized settings, incorporating real-time data, and applying it to real-world mobility datasets for validation.

6.2 Addressing the Research Questions

The outcomes of this study provide clear evidence in support of the research questions outlined in Chapter 3. Regarding **RQ1**, this work demonstrates that Multi-Armed Bandit (MAB) algorithms can effectively support personalised and dynamic coordination of travel assignments in urban settings. By leveraging adaptive prioritisation based on individual preferences and contextual emissions, the MAB-based framework outperforms static and non-learning baselines in carbon efficiency across different scenarios and mobility constraints. As for **RQ2**, the results show that Reinforcement Learning, when instantiated through simplified bandit formulations, can contribute meaningfully to pollution mitigation. Specifically, the use of MAB for real-time transport mode selection enables carbon-aware policy enforcement without full knowledge of the environment, promoting reductions in greenhouse gas emissions while still respecting traveller preferences. The simulation-based evaluation validates the practical applicability of learning-based strategies in achieving scalable and sustainable urban mobility coordination.

6.3 Future Work

The present work offers a foundational framework for coordination in sustainable transport mode assignment using multi-agent learning. However, several directions remain open for further exploration and development. These can be grouped into two main perspectives: direct improvements to the current framework, and broader research directions that can be developed based on it.

From the first perspective, several lines of improvement are suggested:

First, while this study uses synthetically generated OD matrices and preference distributions, future extensions could integrate mobility datasets from real urban regions. Historical trip data from urban regions could be used to infer realistic OD matrices and traveller profiles, increasing the ecological validity of the experimental scenarios.

Second, all travellers in the current framework follow a homogeneous a selfish strategy, picking the path with highest score for them. An extension would be to introduce heterogeneous traveller behaviours, allowing some travellers to adopt alternative selection strategies. This would allow the simulation to better reflect diverse decision-making behaviours present in real populations.

Third, the current problem formulation treats mode and route assignment as a single joint decision. A possible extension would be to decouple these two steps, allowing the system to first determine the transport mode and then compute the route separately. This would support more refined behaviours, such as preferences for specific transport modes along certain segments of a trip. At present, path options are constructed arbitrarily by combining modes and routes.

Fourth, other reinforcement learning architectures could be explored, including extensions to a Multi-Agent Reinforcement Learning (MARL) setting. In such a framework, agents could represent transport services or drivers themselves, learning independent or cooperative strategies within the shared urban environment. This research line opens up a wide road of opportunities

to study more complex scenarios in urban mobility, accounting for the interactions of different players.

Fifth, future work could employ more accurate emission models and incorporate a greater variety of transport modes and vehicle types, such as electric and hybrid cars. This would allow the framework to more precisely evaluate the impact of different policy scenarios, energy transitions, and vehicle technologies on urban emissions.

Beyond improving the current framework, this work provides a foundation for developing new directions of research:

Firstly, integrate the coordination model into a carbon market framework. In such a system, emissions would not only contribute to reward signals but could also be translated into tradable carbon credits. Agents would then learn and coordinate under economic constraints, simulating the dynamics of market-based climate regulation.

Secondly, broaden the sustainability objectives of the model. In addition to CO₂ emissions, future research could consider energy efficiency, traffic congestion, or battery consumption in electric mobility systems. These added layers would support multi-objective optimisation, enhancing the system's applicability to smart city planning.

Finally, another area for exploration is the inclusion of autonomous-vehicle ecosystems. In this scenario, vehicles could act as decision-making agents that optimise their own routes and modes in real-time, shifting the coordination burden from centralised assignment to distributed agent behaviour.

Pursuing these lines of investigation will not only broaden the technical scope of the framework but also improve its alignment with real-world goals of sustainable, intelligent, and human-centred mobility systems.

References

- [1] Alekh Agarwal et al. *Making Contextual Decisions with Low Technical Debt*. 2017. DOI: [10.48550/arXiv.1606.03966](https://doi.org/10.48550/arXiv.1606.03966). arXiv: [1606.03966](https://arxiv.org/abs/1606.03966) [cs.LG].
- [2] European Environment Agency. *Decarbonising Road Transport: The Role of Vehicles, Fuels and Transport Demand*. Publications Office of the EU, 2022. ISBN: 978-92-9480-473-0. DOI: [doi/10.2800/68902](https://doi.org/10.2800/68902).
- [3] Shipra Agrawal and Navin Goyal. “Analysis of Thompson Sampling for the Multi-armed Bandit Problem”. In: *Proceedings of the 25th Annual Conference on Learning Theory*. Vol. 23. JMLR Workshop and Conference Proceedings, June 2012, pp. 39.1–39.26.
- [4] Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. Cambridge, Massachusetts: The MIT Press, 2024. ISBN: 978-0-262-04937-5.
- [5] Laura Alessandretti et al. “Multimodal Urban Mobility and Multilayer Transport Networks”. In: *Environment and Planning B: Urban Analytics and City Science* 50.8 (Oct. 2023), pp. 2038–2070. ISSN: 2399-8083. DOI: [10.1177/23998083221108190](https://doi.org/10.1177/23998083221108190). (Visited on 07/01/2025).
- [6] Inés Aquilué Junyent et al. “Planning Shared Mobility Hubs in European Cities: A Methodological Framework Using MCDA and GIS Applied to Barcelona”. In: *Sustainable Cities and Society* 106 (July 2024), p. 105377. ISSN: 2210-6707. DOI: [10.1016/j.scs.2024.105377](https://doi.org/10.1016/j.scs.2024.105377). (Visited on 07/01/2025).
- [7] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. “Finite-time Analysis of the Multiarmed Bandit Problem”. In: *Machine Learning* 47.2–3 (2002). ISSN: 0885-6125. DOI: [10.1023/A:1013689704352](https://doi.org/10.1023/A:1013689704352).
- [8] Ranga Shaarad Ayyagari and Ambedkar Dukkupati. “Risk-Averse Combinatorial Semi-Bandits”. In: *2023 IEEE International Symposium on Information Theory (ISIT)*. 2023, pp. 1472–1477. DOI: [10.1109/ISIT54713.2023.10206452](https://doi.org/10.1109/ISIT54713.2023.10206452).
- [9] Avishan Bagherinezhad, Mohammad Mehdi Hosseini, and Masood Parvania. “Real-Time Coordinated Operation of Power and Autonomous Electric Ride-Hailing Systems”. In: *IEEE Transactions on Smart Grid* 14.3 (2023), pp. 2214–2225. DOI: [10.1109/TSG.2023.3247780](https://doi.org/10.1109/TSG.2023.3247780).
- [10] Ali Baheri. “Multilevel Constrained Bandits: A Hierarchical Upper Confidence Bound Approach with Safety Guarantees”. In: *Mathematics* 13.1 (2025). Cited by: 1; All Open Access, Gold Open Access, Green Open Access. DOI: [10.3390/math13010149](https://doi.org/10.3390/math13010149).
- [11] Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. “Survey on Applications of Multi-Armed and Contextual Bandits”. In: *2020 IEEE Congress on Evolutionary Computation (CEC)*. 2020, pp. 1–8. DOI: [10.1109/CEC48606.2020.9185782](https://doi.org/10.1109/CEC48606.2020.9185782).

- [12] Apostolos N. Burnetas, Odysseas Kanavetas, and Michael N. Katehakis. “Optimal data driven resource allocation under multi-armed bandit observations”. In: *Annals of Operations Research* (2025). Cited by: 0. DOI: [10.1007/s10479-025-06554-3](https://doi.org/10.1007/s10479-025-06554-3).
- [13] Xiaofeng Cao et al. “Online Worker Selection Towards High Quality Map Collection for Autonomous Driving”. In: *2019 IEEE Global Communications Conference (GLOBECOM)*. 2019, pp. 1–6. DOI: [10.1109/GLOBECOM38437.2019.9013516](https://doi.org/10.1109/GLOBECOM38437.2019.9013516).
- [14] António Ribeiro da Costa et al. “Dynamic and Cooperative Multi-Agent Task Allocation: Enhancing Nash Equilibrium Through Learning”. In: *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2024, pp. 1860–1866. DOI: [10.1109/SMC54092.2024.10831689](https://doi.org/10.1109/SMC54092.2024.10831689).
- [15] Shuhong Dai et al. “MARP: A Cooperative Multiagent DRL System for Connected Autonomous Vehicle Platooning”. In: *IEEE Internet of Things Journal* 11.20 (2024), pp. 32454–32463. DOI: [10.1109/JIOT.2024.3432119](https://doi.org/10.1109/JIOT.2024.3432119).
- [16] Xiaoshu Ding and Sisi Jian. “Revenue Sharing and Resource Allocation for Cooperative Multimodal Transport Systems”. In: *Transportation Research Part C: Emerging Technologies* 164 (July 2024), p. 104666. ISSN: 0968-090X. DOI: [10.1016/j.trc.2024.104666](https://doi.org/10.1016/j.trc.2024.104666). (Visited on 07/01/2025).
- [17] European Environment Agency. “Greenhouse Gas Emissions from Transport in Europe”. In: *European Environment Agency* (Oct. 2024). URL: <https://www.eea.europa.eu/en/analysis/indicators/greenhouse-gas-emissions-from-transport> (visited on 06/29/2025).
- [18] Eurostat. *Passenger cars by type of motor energy and age of vehicle*. 2025. DOI: [10.2908/road_eqs_carhab](https://doi.org/10.2908/road_eqs_carhab). URL: https://ec.europa.eu/eurostat/databrowser/view/road_eqs_carhab/default/table (visited on 06/29/2025).
- [19] Jiaying Guo, Long Cheng, and Shen Wang. “CoTV: Cooperative Control for Traffic Light Signals and Connected Autonomous Vehicles Using Deep Reinforcement Learning”. In: *2024 IEEE Intelligent Vehicles Symposium (IV)*. 2024, pp. 3155–3155. DOI: [10.1109/IV55156.2024.10588855](https://doi.org/10.1109/IV55156.2024.10588855).
- [20] Nicolas Gutowski et al. “Gorthaur-EXP3: Bandit-based selection from a portfolio of recommendation algorithms balancing the accuracy-diversity dilemma”. In: *Information Sciences* 546 (2021), pp. 378–396. ISSN: 0020-0255. DOI: <https://doi.org/10.1016/j.ins.2020.08.106>.
- [21] *IEEE*. URL: <https://ieeexplore.ieee.org/>.
- [22] International Energy Agency. “CO2 Emissions in 2023”. In: *International Energy Agency* (Mar. 2024). URL: <https://www.iea.org/reports/co2-emissions-in-2023> (visited on 06/29/2025).
- [23] Abdelfatah Kermali et al. “Deep Reinforcement Learning of Simulated Students Multimodal Mobility Behavior: Application to the City of Toulouse”. In: *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*. SAC ’25. New York, NY, USA: ACM, May 2025, pp. 882–888. ISBN: 979-8-4007-0629-5. DOI: [10.1145/3672608.3707830](https://doi.org/10.1145/3672608.3707830). (Visited on 07/01/2025).
- [24] Romain Laroche and Raphael Feraud. “Reinforcement Learning Algorithm Selection”. In: *ICL 2018: 6th International Conference on Learning Representations*. Feb. 2018.

- [25] Alexandre Letard et al. “Bandit algorithms: A comprehensive review and their dynamic selection from a portfolio for multicriteria top-k recommendation”. In: *Expert Systems with Applications* 246 (2024), p. 123151. ISSN: 0957-4174. DOI: [10.1016/j.eswa.2024.123151](https://doi.org/10.1016/j.eswa.2024.123151).
- [26] Haowei Li et al. “A Contextual Multi-armed Bandit Approach to Personalized Trip Itinerary Planning”. In: *2024 IEEE International Conference on Smart Mobility (SM)*. 2024, pp. 55–60. DOI: [10.1109/SM63044.2024.10733530](https://doi.org/10.1109/SM63044.2024.10733530).
- [27] Wan Li et al. “Energy-Efficient Multimodal Mobility Networks in Transportation Digital Twins: Strategies and Optimization”. In: *Energy* 318 (Mar. 2025), p. 134587. ISSN: 0360-5442. DOI: [10.1016/j.energy.2025.134587](https://doi.org/10.1016/j.energy.2025.134587). (Visited on 07/01/2025).
- [28] Yuxi Li. *Reinforcement Learning in Practice: Opportunities and Challenges*. 2022. DOI: [10.48550/arXiv.2202.11296](https://doi.org/10.48550/arXiv.2202.11296). arXiv: 2202.11296 [cs.LG].
- [29] Haoyang Liu, Keqin Liu, and Qing Zhao. “Learning in a Changing World: Restless Multi-armed Bandit With Unknown Dynamics”. In: *IEEE Transactions on Information Theory* 59.3 (2013), pp. 1902–1916. DOI: [10.1109/TIT.2012.2230215](https://doi.org/10.1109/TIT.2012.2230215).
- [30] Miguel Morales. *Grokking Deep Reinforcement Learning*. Manning, 2020. ISBN: 9781617295454.
- [31] Andi Nika, Sepehr Elahi, and Cem Tekin. “Contextual Combinatorial Volatile Multi-armed Bandit with Adaptive Discretization”. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Vol. 108. 2020, pp. 1486–1496.
- [32] Doruk Öner et al. *Combinatorial Multi-Objective Multi-Armed Bandit Problem*. 2018. DOI: [10.48550/arXiv.1803.04039](https://doi.org/10.48550/arXiv.1803.04039). arXiv: 1803.04039 [cs.LG].
- [33] Dawei Qiu et al. “Graph Reinforcement Learning for Carbon-Aware Electric Vehicles in Power-Transport Networks”. In: *IEEE Transactions on Smart Grid* 15.4 (2024), pp. 3919–3935. DOI: [10.1109/TSG.2024.3359289](https://doi.org/10.1109/TSG.2024.3359289).
- [34] Rayyan – Intelligent Systematic Review. URL: <https://www.rayyan.ai/>.
- [35] Carolin Schwalm, Hans Schily, and Alexander Charlish. “Using multi-armed bandits for search application in a space surveillance radar network”. In: *Proc. 2nd NEO and Debris Detection Conference*. Vol. 2. Darmstadt, Germany: ESA Space Debris Office, 2023.
- [36] Scopus. URL: <https://www.scopus.com/>.
- [37] Aleksandrs Slivkins. *Introduction to Multi-Armed Bandits*. 2024. DOI: [10.48550/arXiv.1904.07272](https://doi.org/10.48550/arXiv.1904.07272). arXiv: 1904.07272 [cs.LG].
- [38] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts: The MIT Press, 2018. ISBN: 978-0-262-03924-6.
- [39] Chengyu Tang and Sanjeev Baskiyar. “Adaptive and Responsive Traffic Signal Control Using Reinforcement Learning and Fog Computing”. In: *2024 IEEE Cloud Summit*. 2024, pp. 36–41. DOI: [10.1109/Cloud-Summit61220.2024.00013](https://doi.org/10.1109/Cloud-Summit61220.2024.00013).
- [40] Hoa Tran-Dang and Dong-Seong Kim. “Bandit learning based stable matching for decentralized task offloading in dynamic fog computing networks”. In: *Journal of Communications and Networks* 26.3 (2024), pp. 356–365. DOI: [10.23919/JCN.2024.000017](https://doi.org/10.23919/JCN.2024.000017).
- [41] Shangshang Wang et al. “Neural Constrained Combinatorial Bandits”. In: *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*. 2023, pp. 1–10. DOI: [10.1109/INFOCOM53939.2023.10228958](https://doi.org/10.1109/INFOCOM53939.2023.10228958).

- [42] Xin Wang et al. “Enabling Efficient Vehicle–Road Cooperation Through AIoT: A Deep Learning Approach to Computational Offloading”. In: *IEEE Internet of Things Journal* 11.22 (2024), pp. 36127–36139. DOI: [10.1109/JIOT.2024.3445642](https://doi.org/10.1109/JIOT.2024.3445642).
- [43] Yi Wang, Dawei Qiu, and Goran Strbac. “Multi-agent reinforcement learning for electric vehicles joint routing and scheduling strategies”. In: *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. 2022, pp. 3044–3049. DOI: [10.1109/ITSC55140.2022.9921744](https://doi.org/10.1109/ITSC55140.2022.9921744).
- [44] Yi Wang et al. “Multi-agent reinforcement learning for electric vehicle decarbonized routing and scheduling”. In: *Energy* 284 (2023), p. 129335. ISSN: 0360-5442. DOI: <https://doi.org/10.1016/j.energy.2023.129335>.
- [45] *Web of Science*. URL: <https://www.webofscience.com/wos/woscc/basic-search>.
- [46] Stavros Xanthopoulos et al. “Optimization of the Location and Capacity of Shared Multimodal Mobility Hubs to Maximize Travel Utility in Urban Areas”. In: *Transportation Research Part A: Policy and Practice* 179 (Jan. 2024), p. 103934. ISSN: 0965-8564. DOI: [10.1016/j.tra.2023.103934](https://doi.org/10.1016/j.tra.2023.103934). (Visited on 07/01/2025).
- [47] Linchang Xiao et al. “EWS: Towards Cost-Effective Job Scheduling via Combinatorial Multi-Armed Bandit Learning”. In: *International Conference on Networking, Architecture and Storage, NAS 2024* (2024). Cited by: 0. DOI: [10.1109/NAS63802.2024.10781367](https://doi.org/10.1109/NAS63802.2024.10781367).
- [48] Gongquan Zhang et al. “Dual-Objective Reinforcement Learning-Based Adaptive Traffic Signal Control for Decarbonization and Efficiency Optimization”. In: *Mathematics* 12.13 (2024). DOI: [10.3390/math12132056](https://doi.org/10.3390/math12132056).
- [49] Hui Zhang et al. “Quality-Aware User Scheduling for Federated Mobile Crowdsensing”. In: *2022 14th International Conference on Wireless Communications and Signal Processing (WCSP)*. 2022, pp. 187–191. DOI: [10.1109/WCSP55476.2022.10039253](https://doi.org/10.1109/WCSP55476.2022.10039253).
- [50] Weilong Zhu et al. “An Intelligent Route Guidance Strategy based on Congestion Type for ITS”. In: *ICC 2022 - IEEE International Conference on Communications*. 2022, pp. 5047–5052. DOI: [10.1109/ICC45855.2022.9838612](https://doi.org/10.1109/ICC45855.2022.9838612).

Appendix A

Experiment 1: Extended Results

This appendix includes the complete set of plots and visual analyses for **Experiment 1: Evaluating the Learning-Based Coordination Strategy for Emission Reduction**.

A.0.1 Population Size of 1540

Table A.1: Difference in transport mode usage for scenarios with 1540 people.

People	Zones	Connection Type	Transport Mode	First Round	Last Round	Difference
1540	6	PTlimited	metro	392	387	-5
			bus	180	180	0
			car	922	892	-30
			foot	634	667	33
1540	6	noDiagonalConnect	metro	696	700	4
			bus	280	280	0
			car	1076	1015	-61
			foot	541	590	49

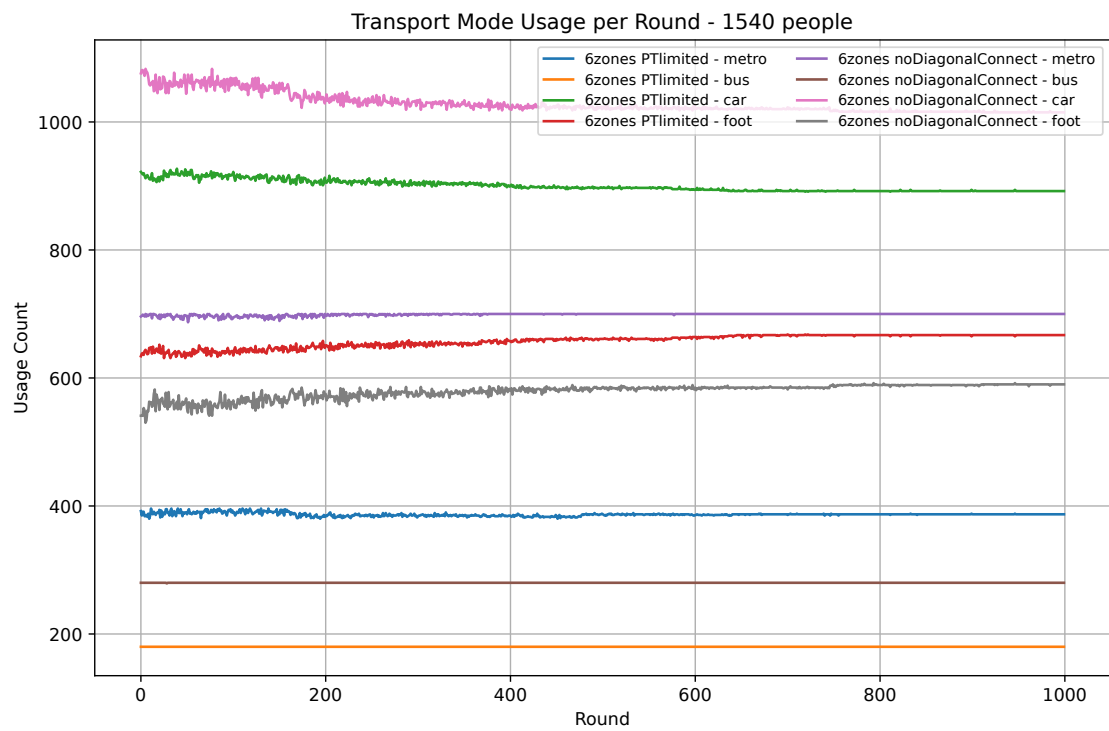


Figure A.1: Transport mode usage evolution in six-zone for scenarios with 1540 individuals.

A.0.2 Population Size of 1080

Table A.2: Difference in transport mode usage for scenarios with 1080 people.

People	Zones	Connection Type	Transport Mode	First Round	Last Round	Difference
1080	4	PTlimited	metro	300	300	0
			bus	120	120	0
			car	473	455	-18
			foot	253	279	26
1080	4	fullConn	metro	419	423	4
			bus	240	240	0
			car	384	375	-9
			foot	37	42	5
1080	4	noDiagonalConnect	metro	400	400	0
			bus	160	160	0
			car	584	547	-37
			foot	304	341	37
1080	6	PTlimited	metro	341	343	2
			bus	180	177	-3
			car	641	621	-20
			foot	356	373	17
1080	6	noDiagonalConnect	metro	653	650	-3
			bus	279	280	1
			car	702	639	-63
			foot	194	257	63

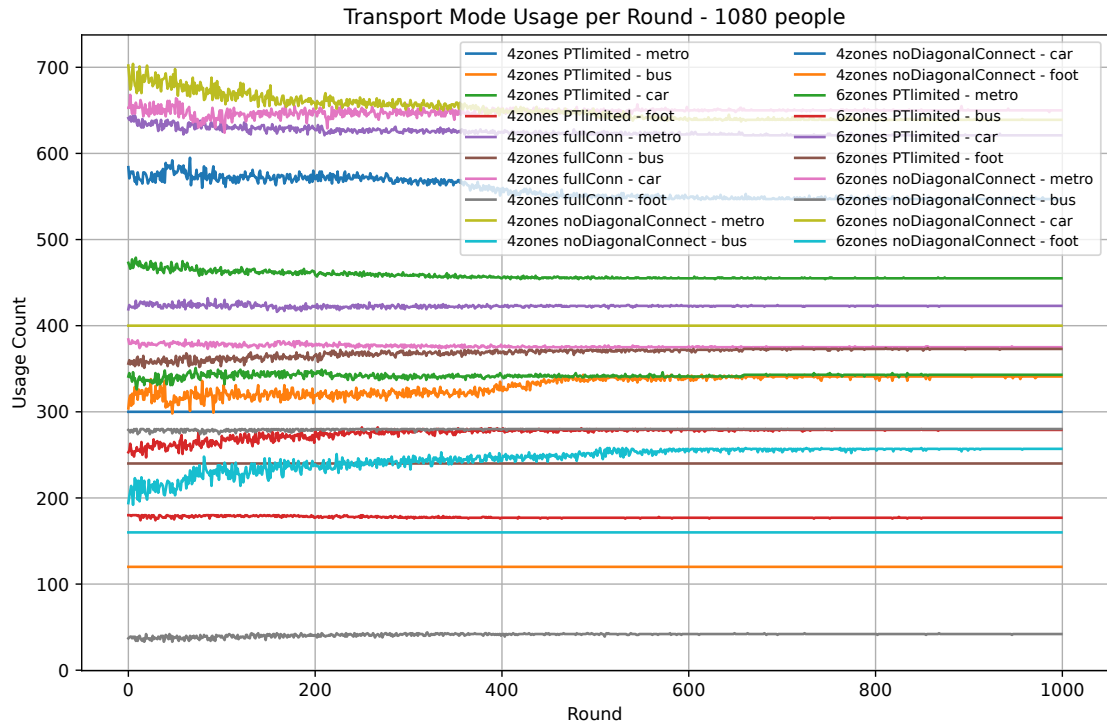


Figure A.2: Transport mode usage evolution for scenarios with 1080 individuals.

A.0.3 Population Size of 840

Table A.3: Difference in transport mode usage for scenarios with 840 people.

People	Zones	Connection Type	Transport Mode	First Round	Last Round	Difference
840	4	PTlimited	metro	271	265	-6
			bus	120	120	0
			car	357	337	-20
			foot	151	167	16
840	4	fullConn	metro	287	293	6
			bus	235	235	0
			car	295	286	-9
			foot	23	27	4
840	4	noDiagonalConnect	metro	397	400	3
			bus	160	160	0
			car	443	416	-27
			foot	128	158	30

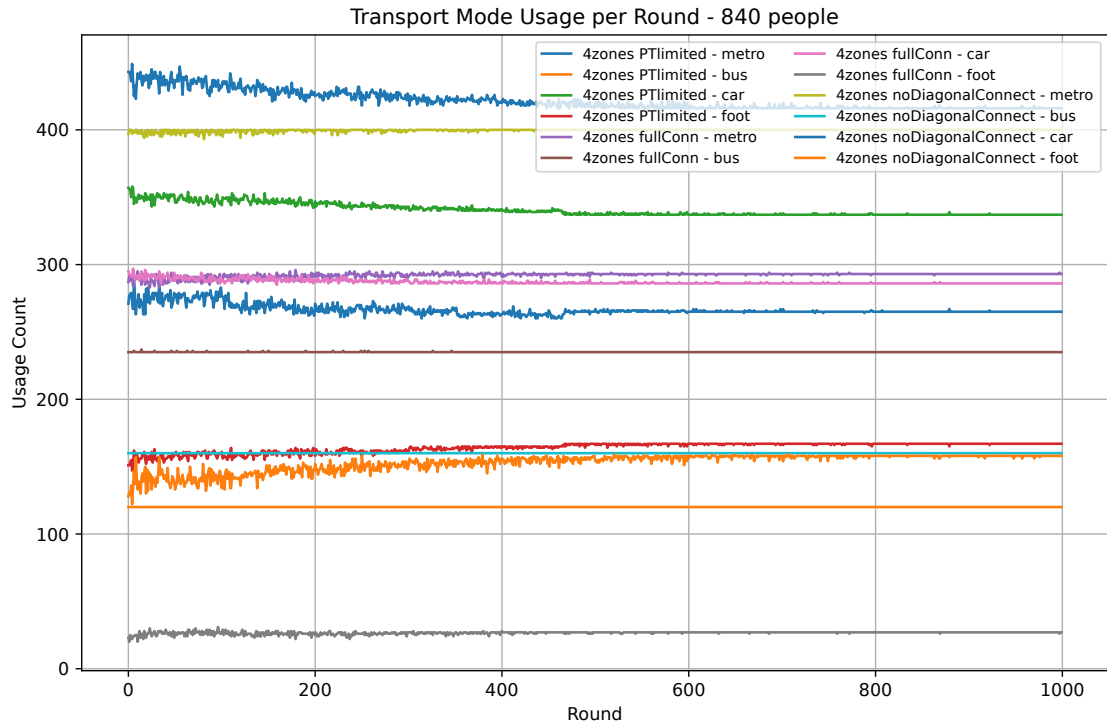


Figure A.3: Transport mode usage evolution for scenarios with 840 individuals.

A.0.4 Population Size of 560

Table A.4: Difference in transport mode usage for scenarios with 560 people.

People	Zones	Connection Type	Transport Mode	First Round	Last Round	Difference
560	4	PTlimited	metro	182	193	11
			bus	104	103	-1
			car	235	233	-2
			foot	73	71	-2
560	4	fullConn	metro	170	170	0
			bus	170	170	0
			car	204	204	0
			foot	16	16	0
560	4	noDiagonalConnect	metro	287	292	5
			bus	157	153	-4
			car	276	270	-6
			foot	30	33	3

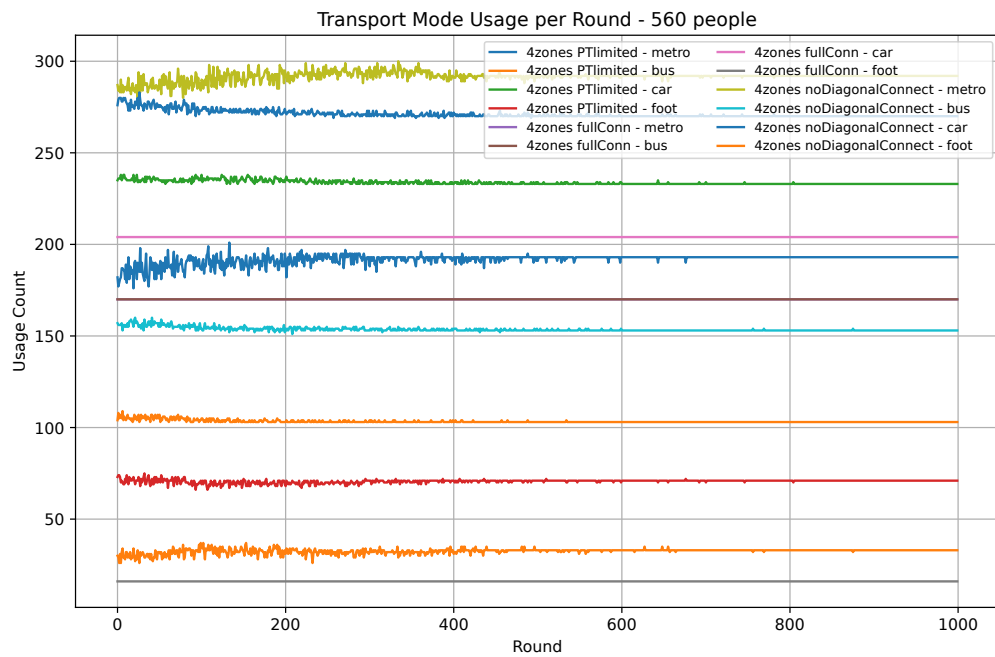


Figure A.4: Transport mode usage evolution for scenarios with 560 individuals.

A.0.5 Population Size of 360

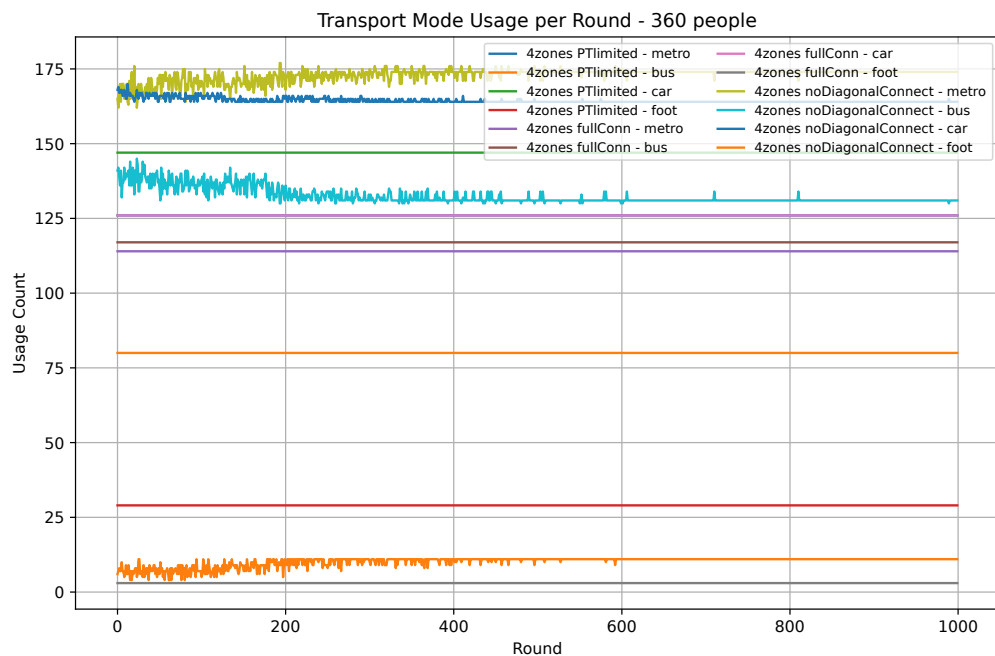


Figure A.5: Transport mode usage evolution for scenarios with 360 individuals.

Table A.5: Difference in transport mode usage for scenarios with 360 people.

People	Zones	Connection Type	Transport Mode	First Round	Last Round	Difference
360	4	PTlimited	metro	126	126	0
			bus	80	80	0
			car	147	147	0
			foot	29	29	0
360	4	fullConn	metro	114	114	0
			bus	117	117	0
			car	126	126	0
			foot	3	3	0
360	4	noDiagonalConnect	metro	165	174	9
			bus	141	131	-10
			car	168	164	-4
			foot	6	11	5

Appendix B

Experiment 2: Extended Results

This appendix includes the complete set of plots and visual analyses for **Experiment 2: Influence of Behavioural Incentives on Transport Mode Choices**.

B.0.1 Population Size of 1540

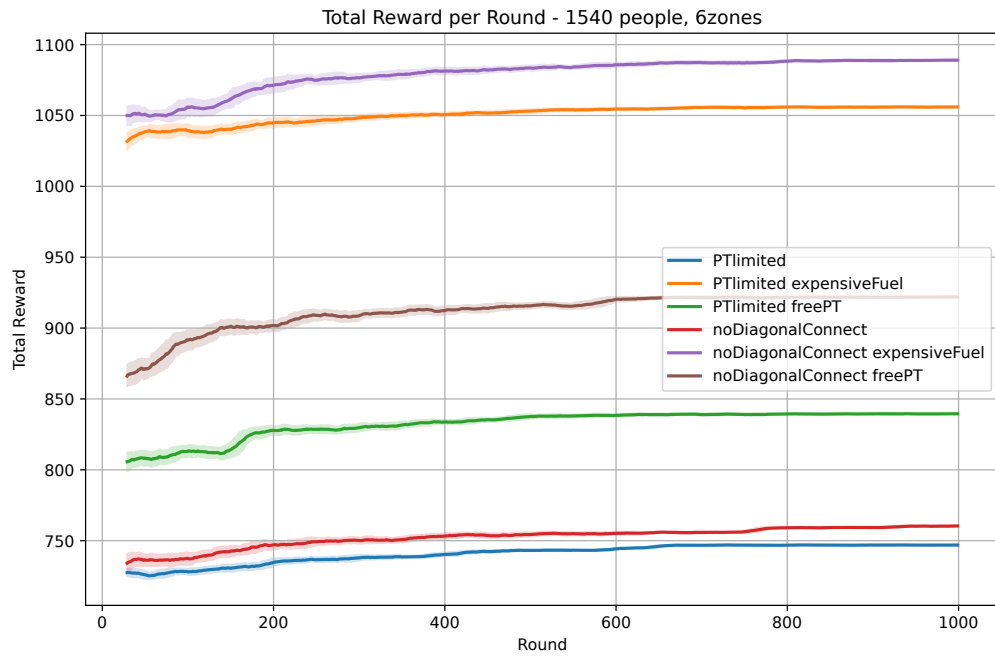


Figure B.1: Moving average of reward for scenarios with 1540 individuals in six-zone under different incentive strategies.

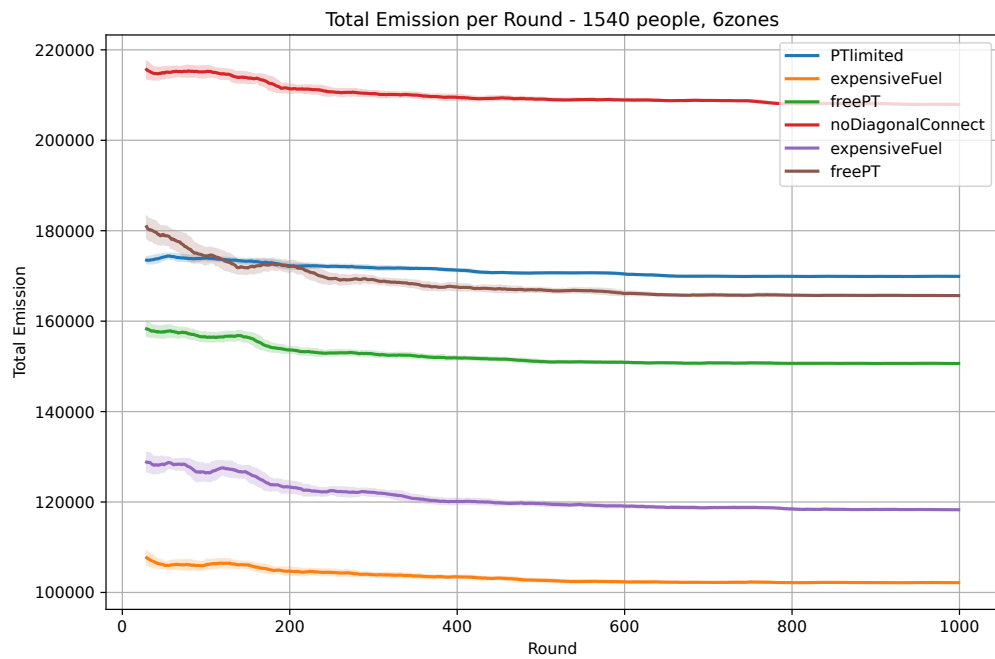


Figure B.2: Emission levels for scenarios with 1540 individuals in six-zone under different incentive strategies.

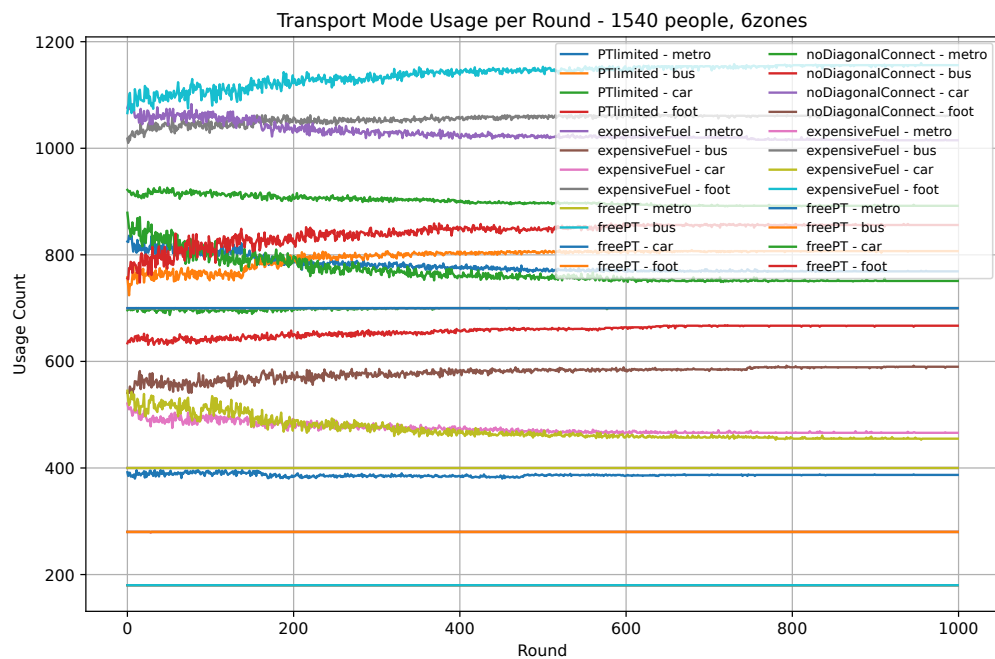


Figure B.3: Transport mode usage evolution for 1540 individuals in six-zone scenarios under different incentive strategies.

B.0.2 Population Size of 1080

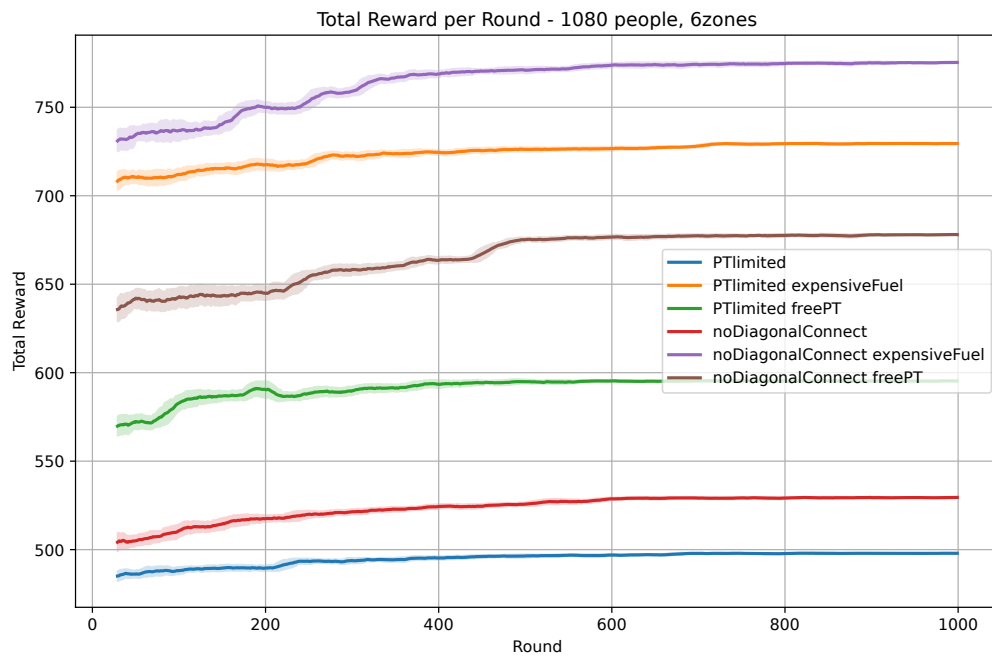


Figure B.4: Moving average of reward for scenarios with 1080 individuals in six-zone under different incentive strategies.

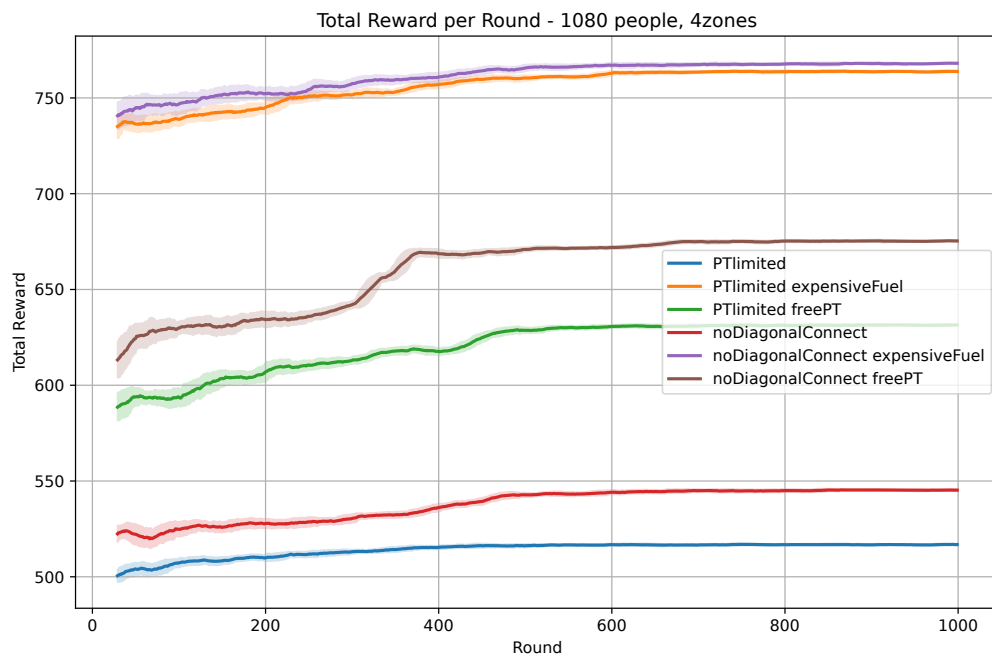


Figure B.5: Moving average of reward for scenarios with 1080 individuals in four-zone under different incentive strategies.

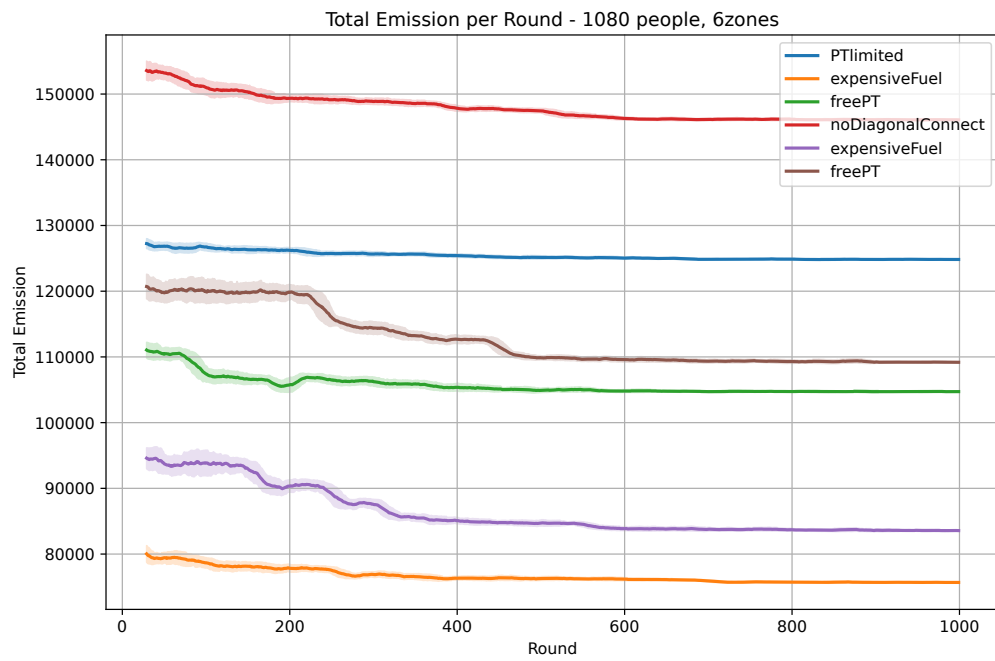


Figure B.6: Emission levels for scenarios with 1080 individuals in six-zone under different incentive strategies.

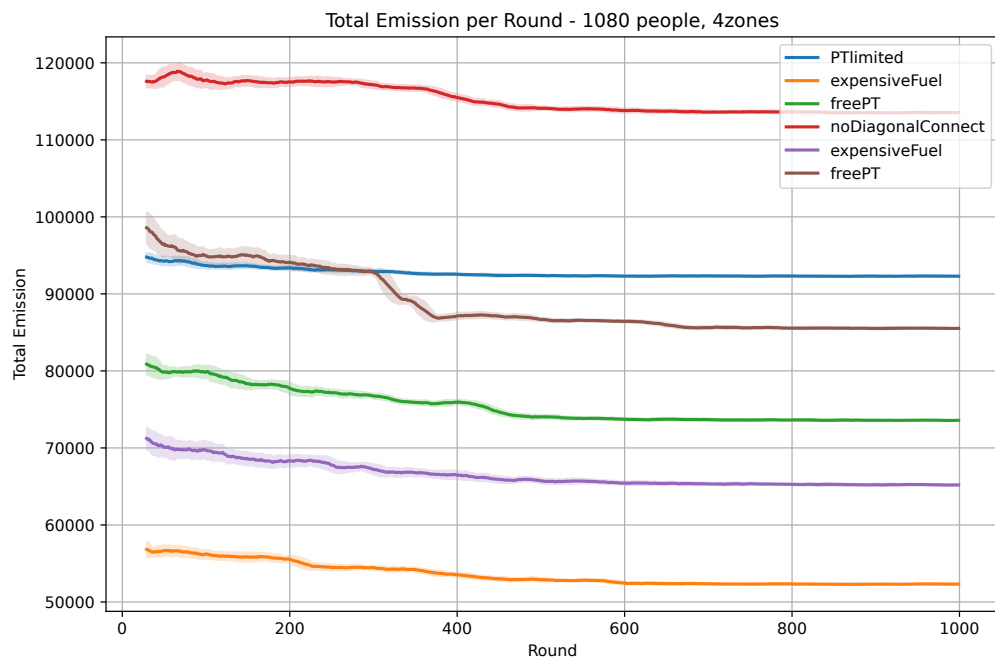


Figure B.7: Emission levels for scenarios with 1080 individuals in four-zone under different incentive strategies.

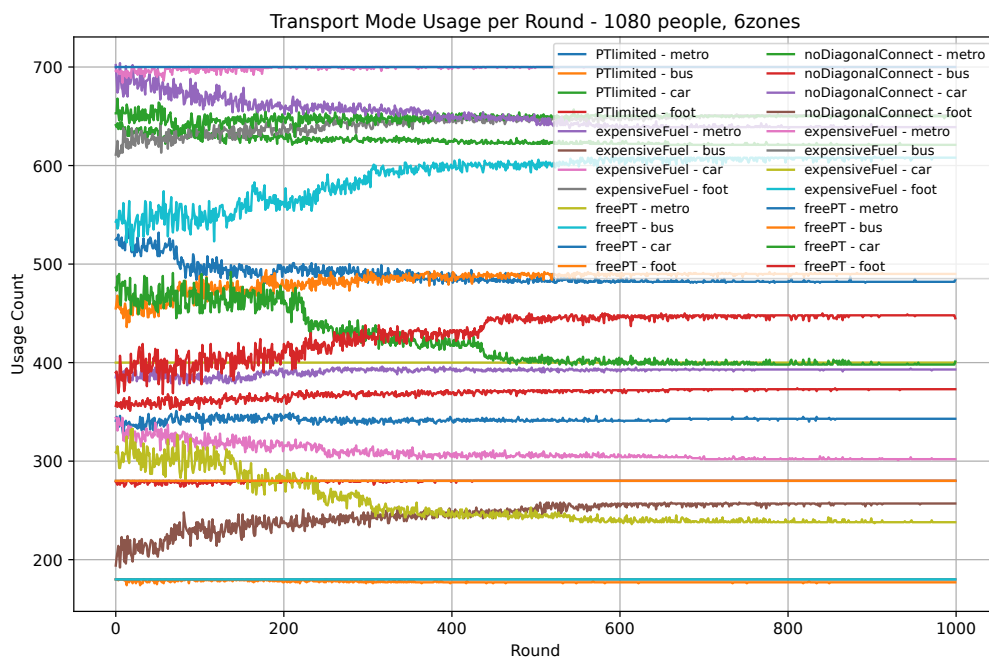


Figure B.8: Transport mode usage evolution for 1080 individuals in six-zone scenarios under different incentive strategies.

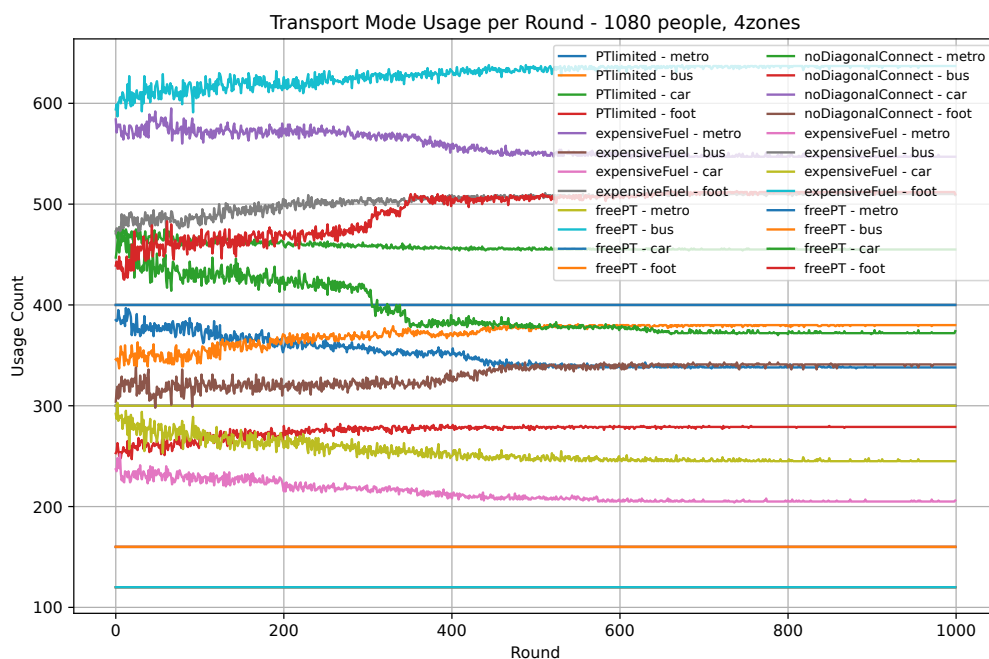


Figure B.9: Transport mode usage evolution over time for 1080 individuals in four-zone scenarios under different incentive strategies.

B.0.3 Population Size of 840

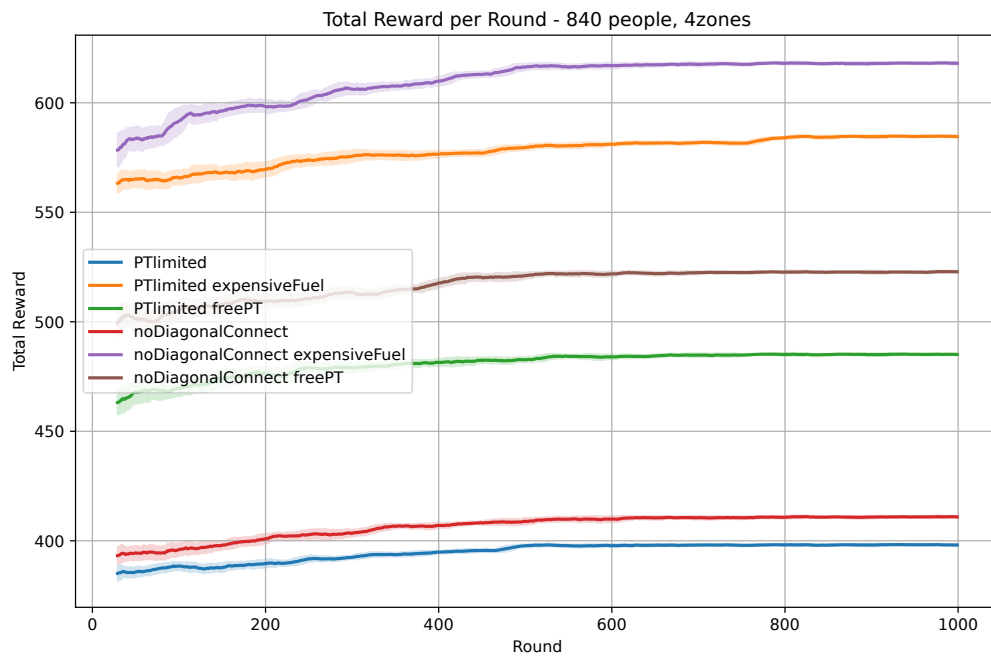


Figure B.10: Moving average of reward for scenarios with 840 individuals in four-zone under different incentive strategies.

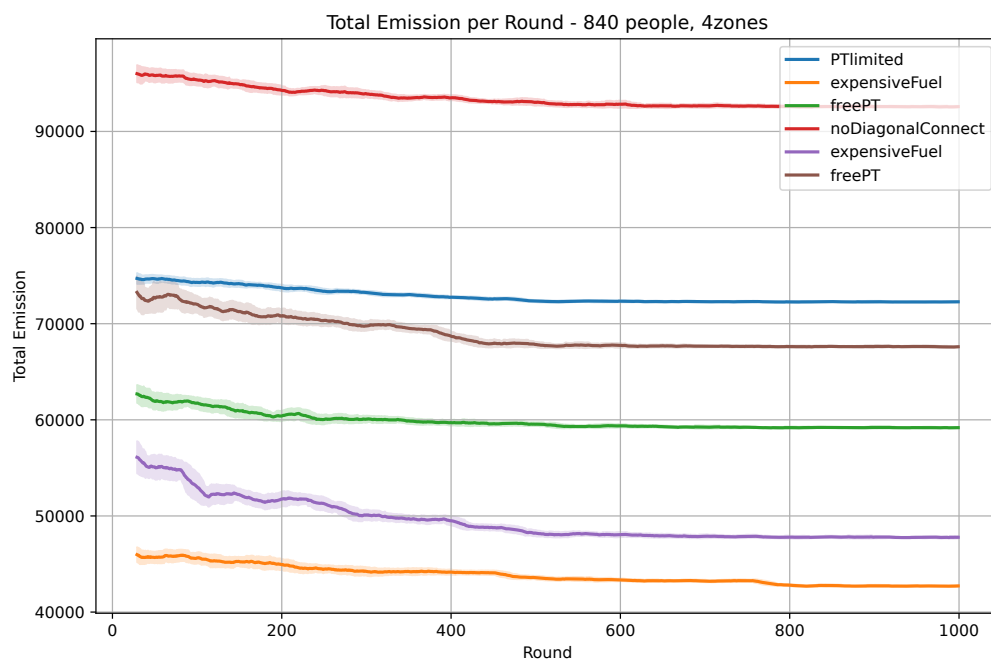


Figure B.11: Emission levels for scenarios with 840 individuals in four-zone under different incentive strategies.

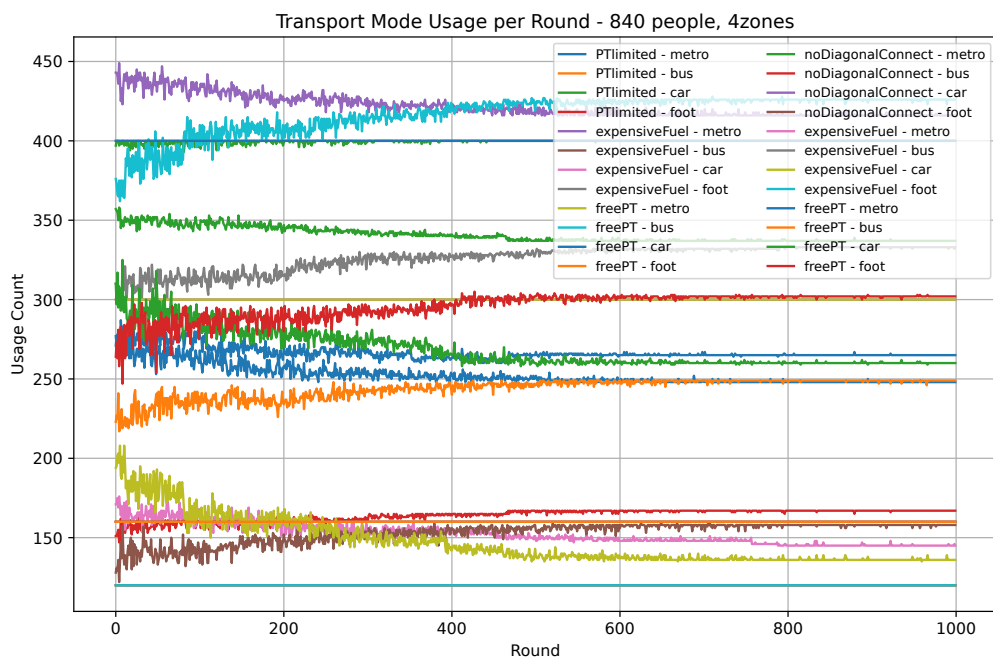


Figure B.12: Transport mode usage evolution over time for 840 individuals in four-zone scenarios under different incentive strategies.

B.0.4 Population Size of 560

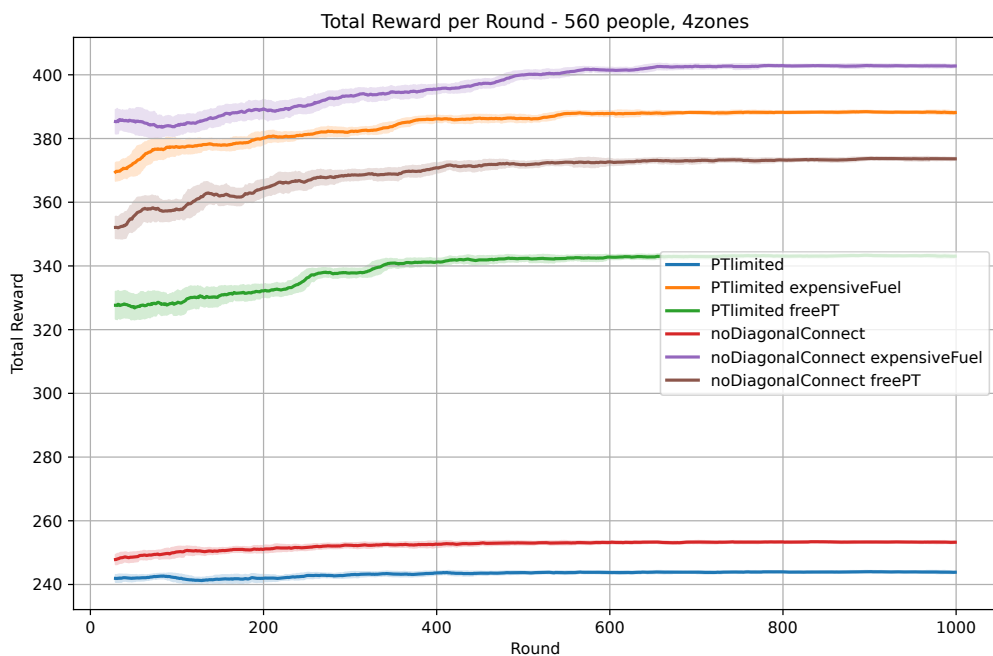


Figure B.13: Moving average of reward over time for scenarios with 560 individuals in four-zone under different incentive strategies.

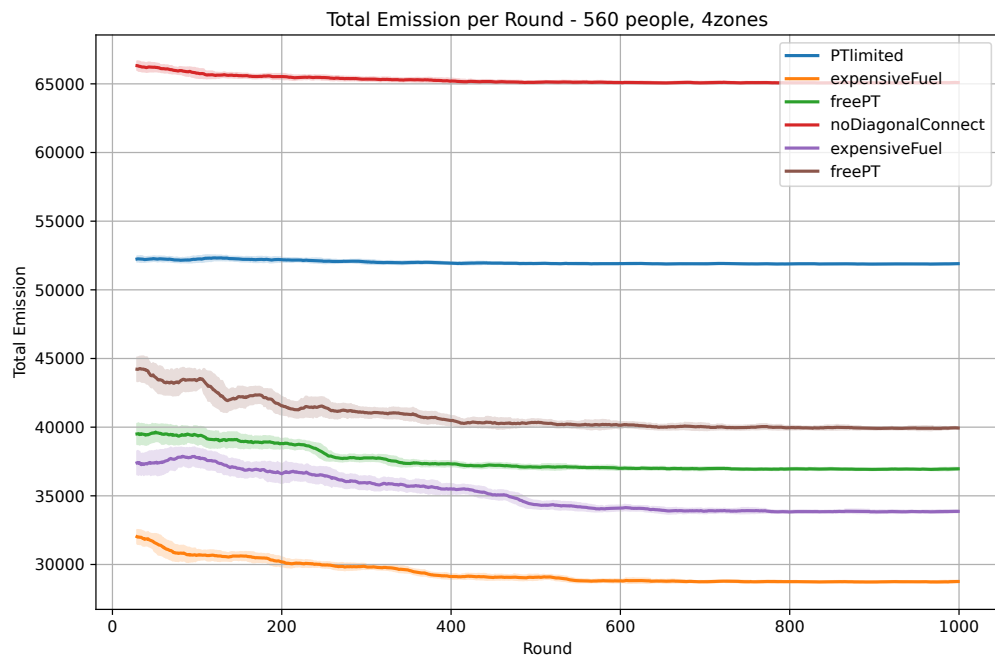


Figure B.14: Emission levels for scenarios with 560 individuals in four-zone under different incentive strategies.

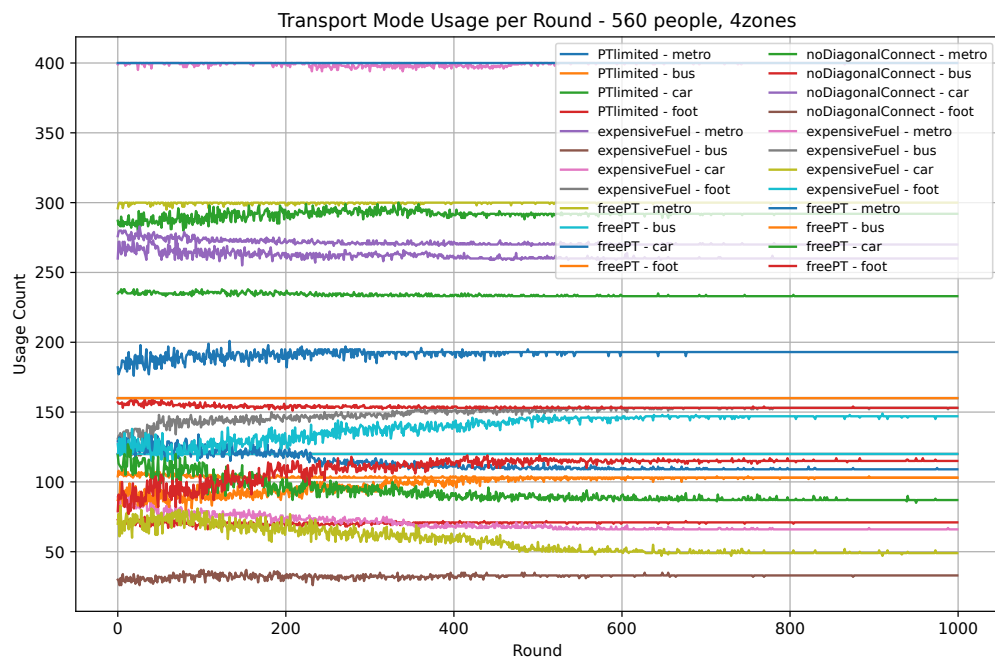


Figure B.15: Transport mode usage evolution over time for 560 individuals in four-zone scenarios under different incentive strategies.

B.0.5 Population Size of 360

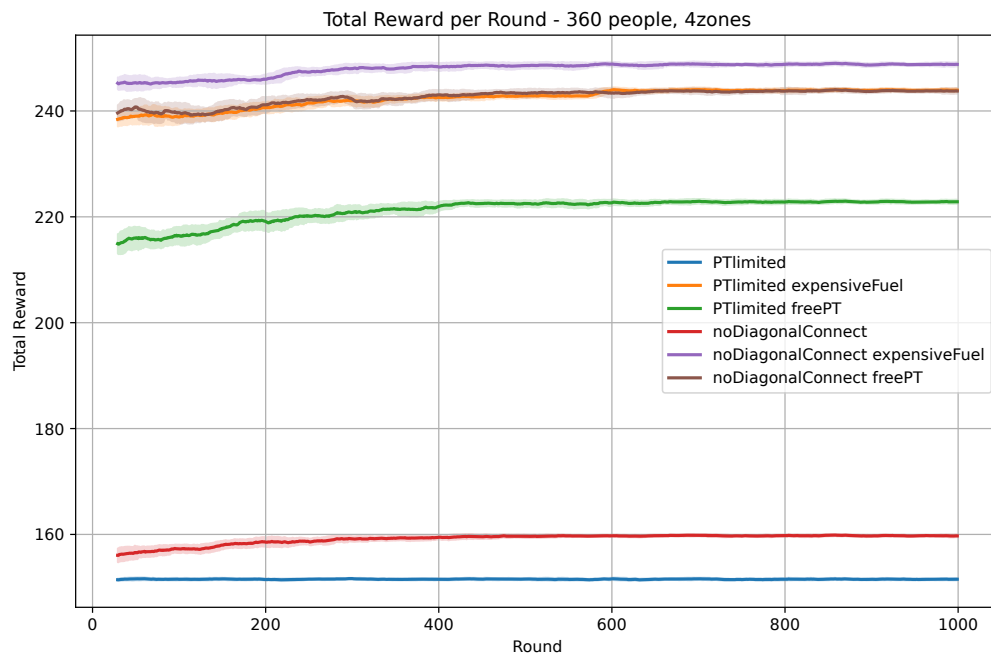


Figure B.16: Moving average of reward for scenarios with 360 individuals in four-zone under different incentive strategies.

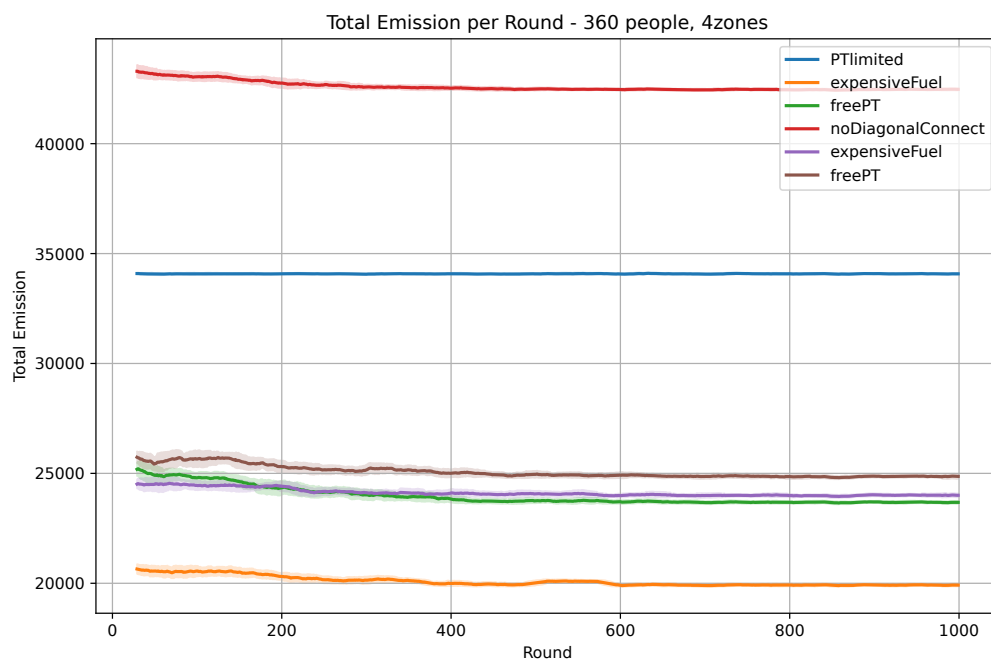


Figure B.17: Emission levels for scenarios with 360 individuals in four-zone under different incentive strategies.

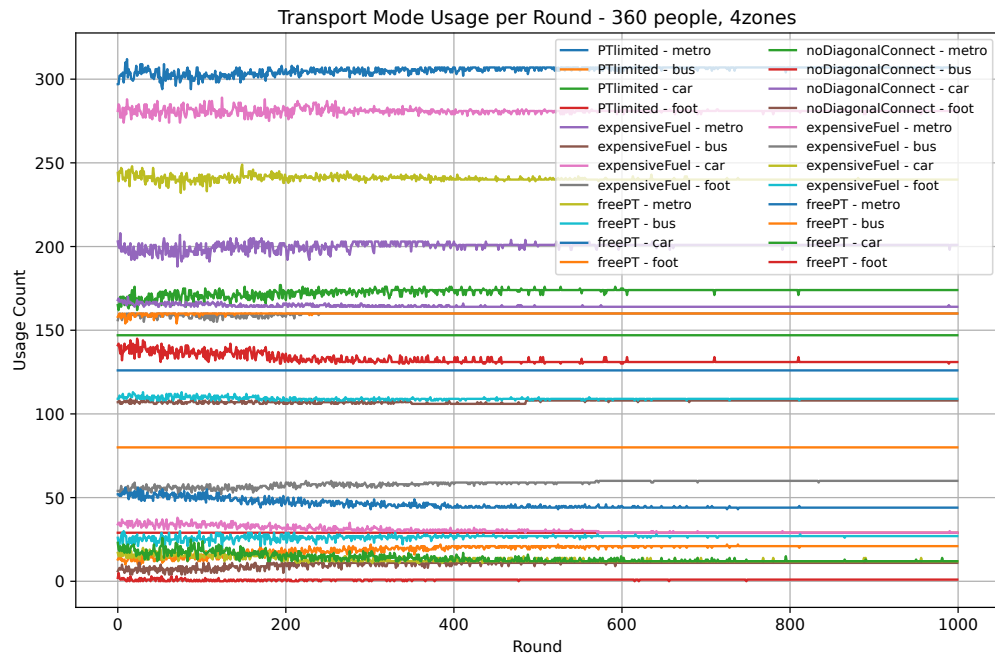


Figure B.18: Transport mode usage evolution for 360 people in four-zone under different incentive strategies.

Appendix C

Experiment 3: Extended Results

This appendix includes the complete set of plots and visual analyses for **Experiment 3: Ablation Study on the Exploration Parameter ϵ and Decay Strategies**.

C.0.1 Population Size of 1080

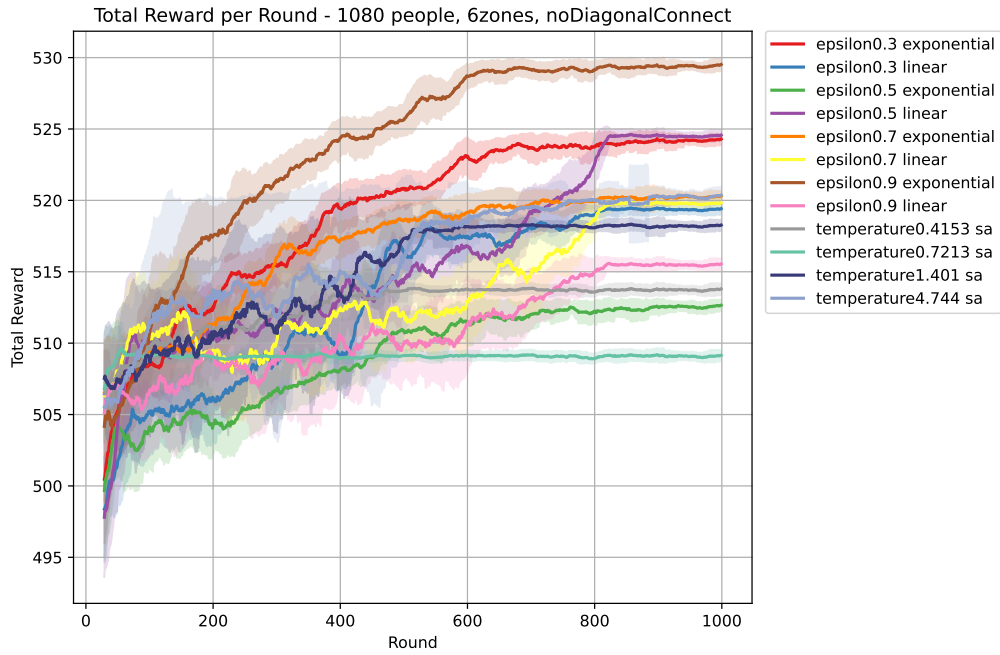


Figure C.1: Moving average of rewards for 1080 individuals in six-zone (no diagonals) under different ϵ -decay strategies.

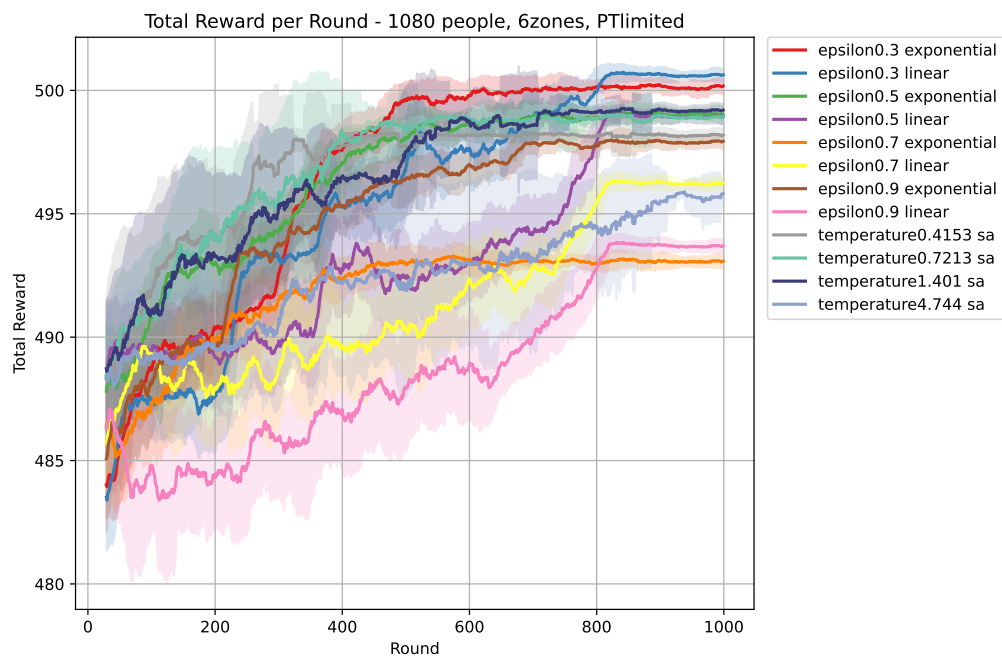


Figure C.2: Moving average of rewards for 1080 individuals in six-zone (PT Limited) under different ϵ -decay strategies.

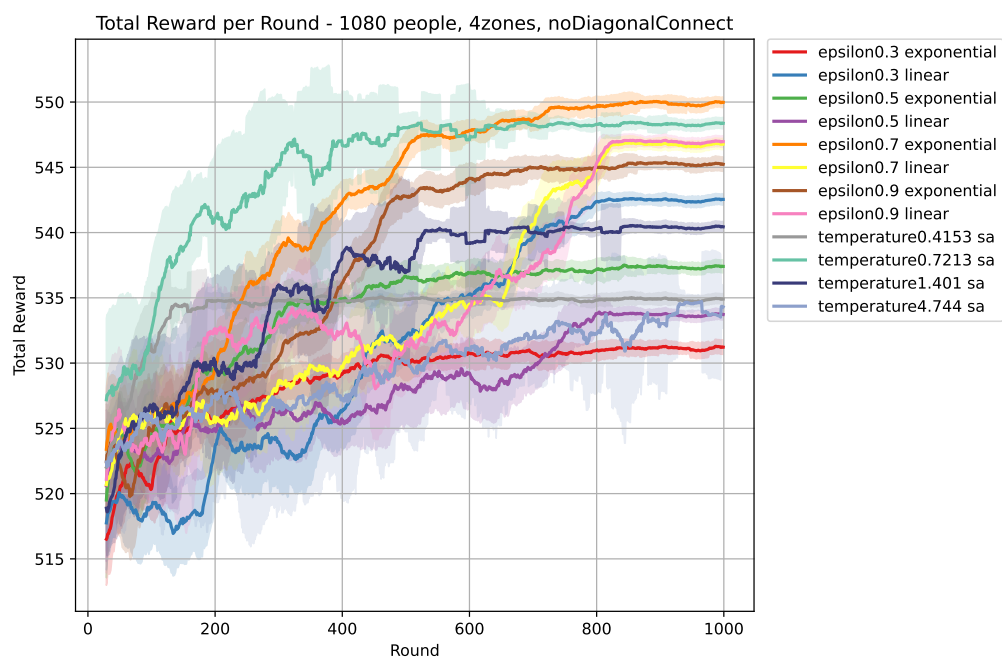


Figure C.3: Moving average of rewards for 1080 individuals in four-zone (no diagonals) under different ϵ -decay strategies.

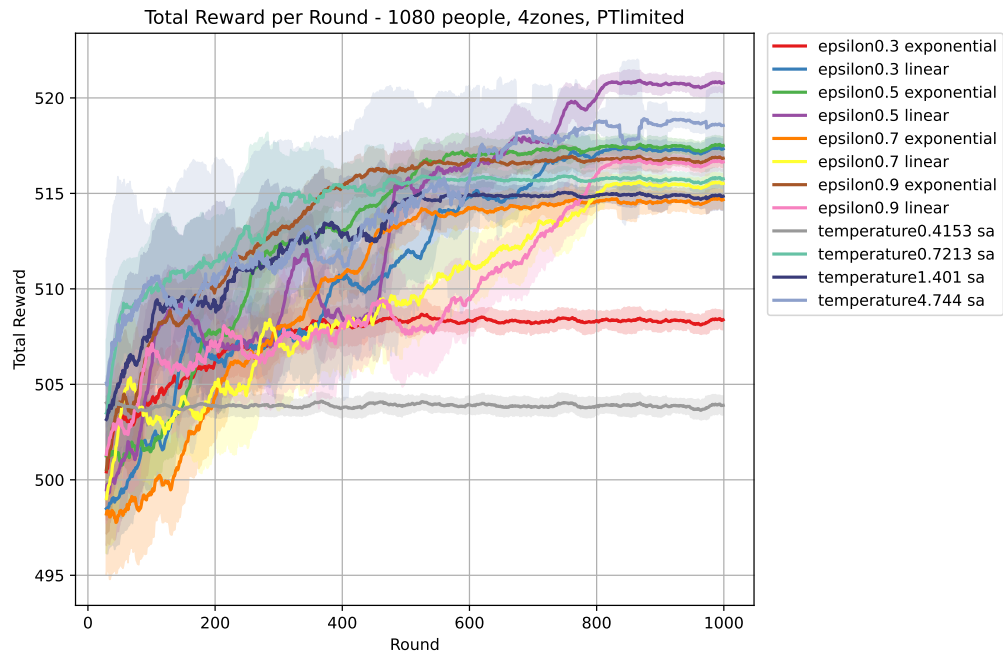


Figure C.4: Moving average of rewards for 1080 individuals in four-zone (PT Limited) under different ϵ -decay strategies.

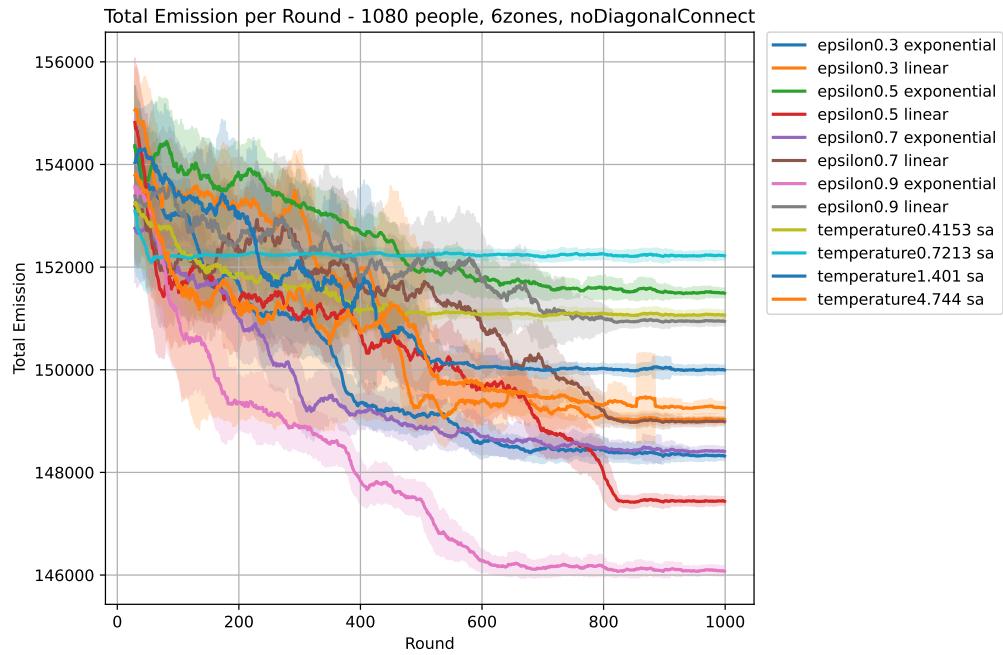


Figure C.5: Emission levels for scenarios with 1080 individuals in six-zone (no diagonals) under different ϵ -decay strategies.

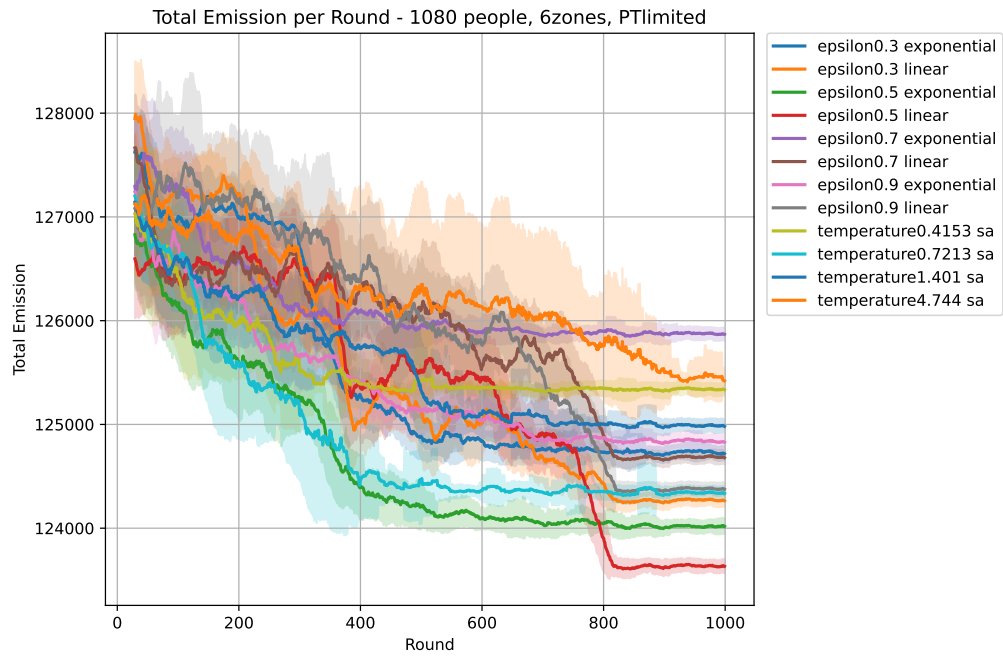


Figure C.6: Emission levels for scenarios with 1080 individuals in six-zone (PT Limited) under different ϵ -decay strategies.

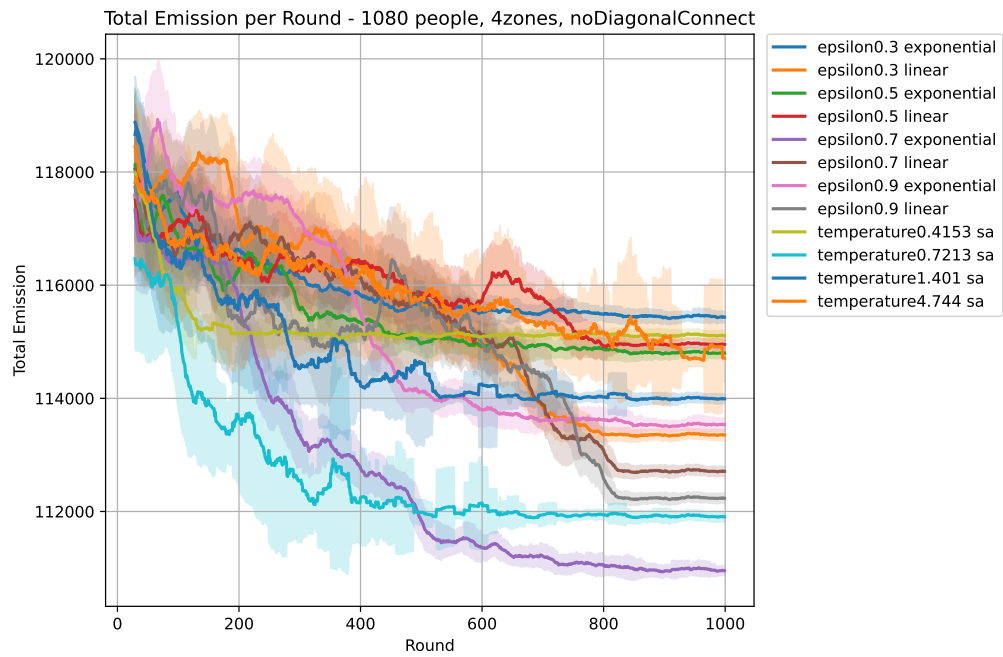


Figure C.7: Emission levels for scenarios with 1080 individuals in four-zone (no diagonals).

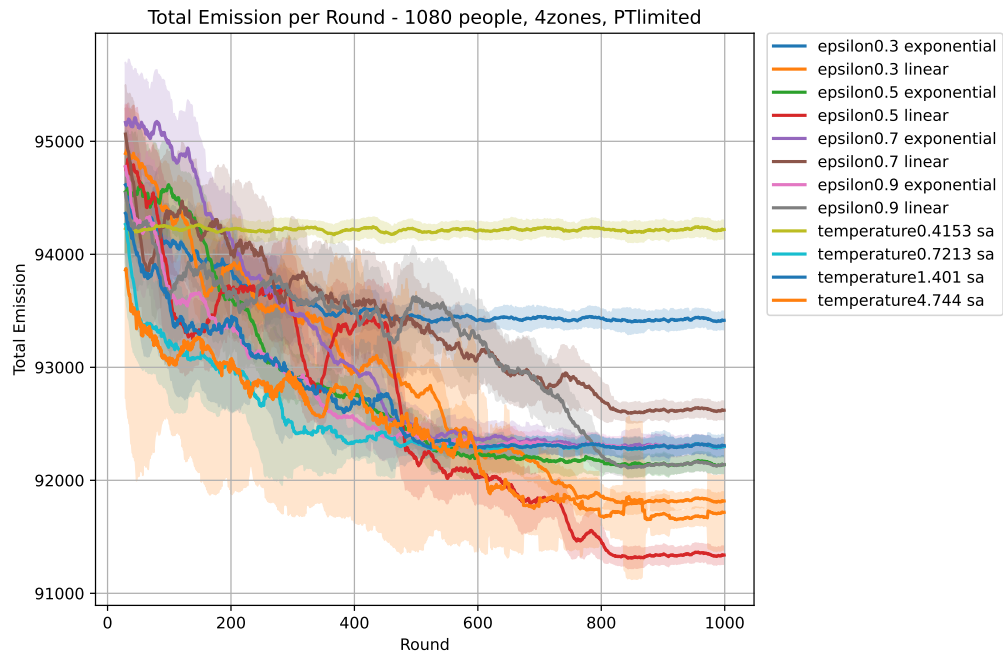


Figure C.8: Emission levels for scenarios with 1080 individuals in four-zone (PT Limited) under different ϵ -decay strategies.

C.0.2 Population Size of 840

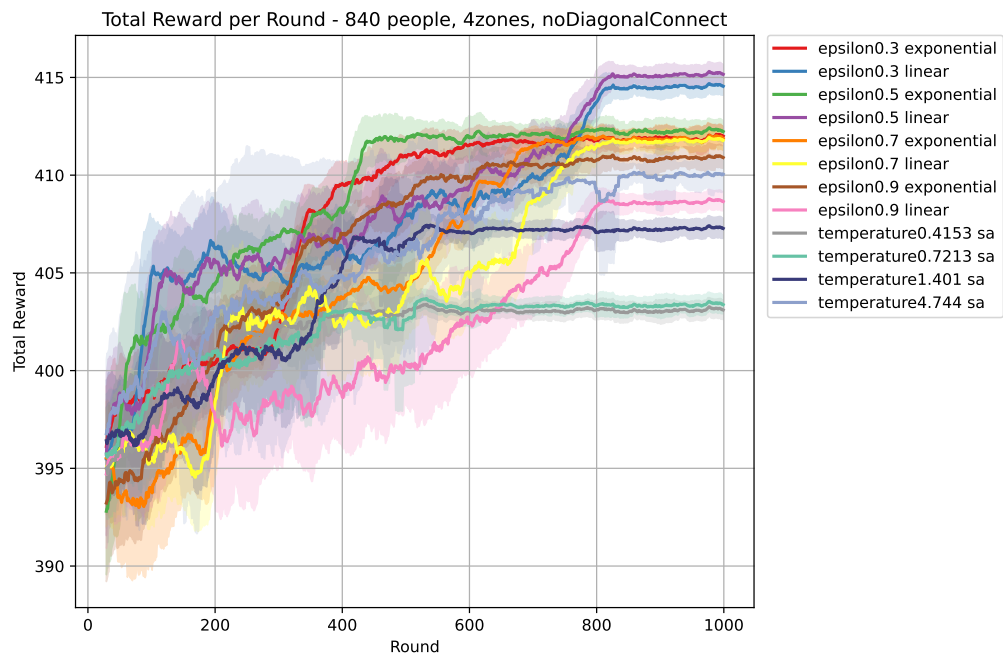


Figure C.9: Moving average of rewards for 840 people in four-zone (no diagonals) under different ϵ -decay strategies.

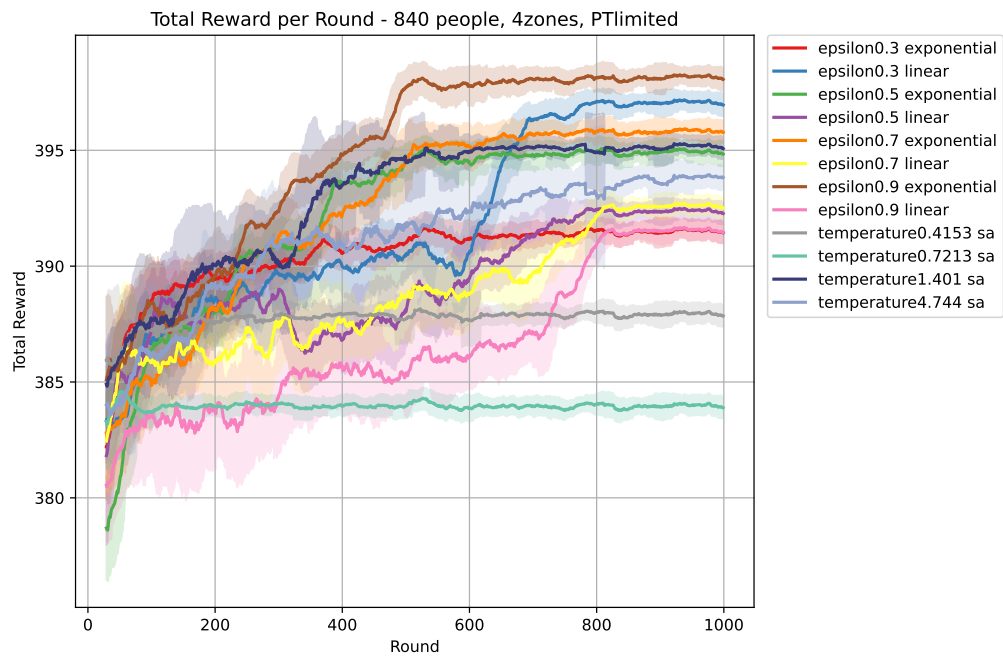


Figure C.10: Moving average of rewards for 840 people in four-zone (PT Limited) under different ϵ -decay strategies.

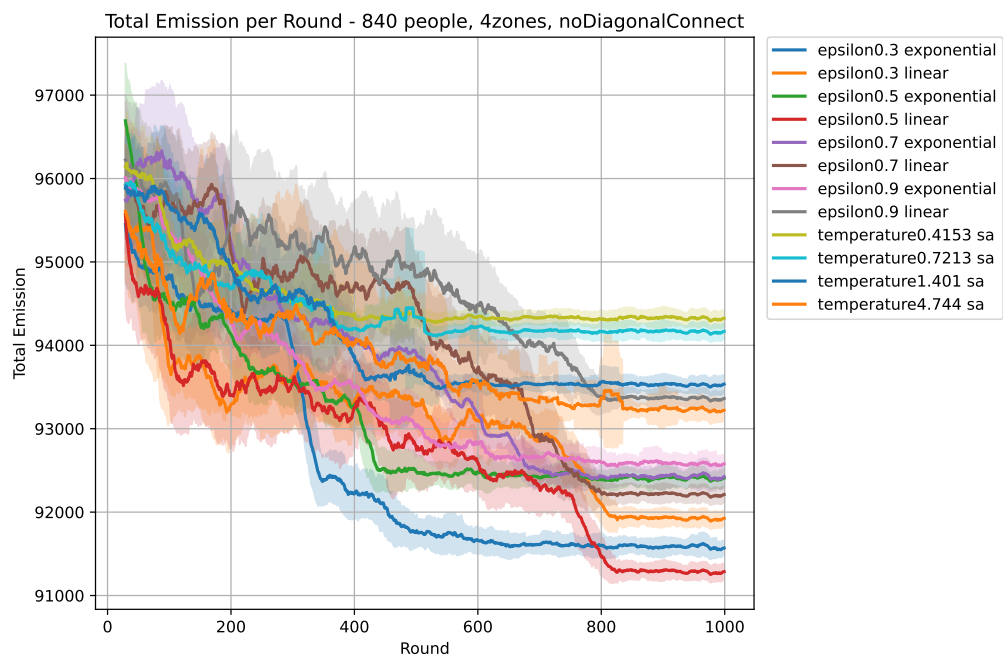


Figure C.11: Emission levels for scenarios with 840 individuals in four-zone (no diagonals) under different ϵ -decay strategies.

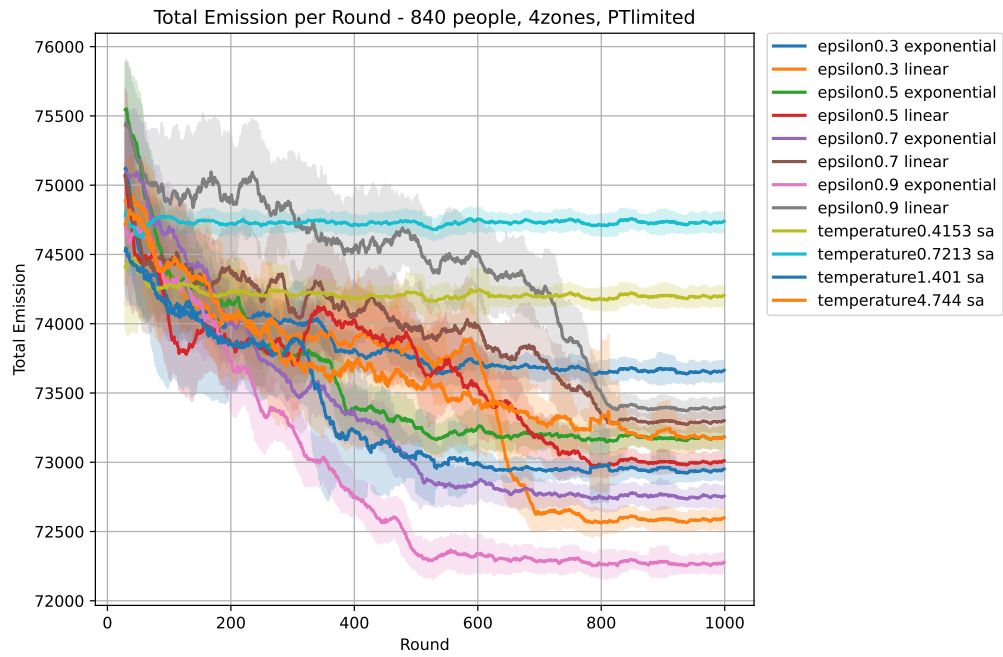


Figure C.12: Emission levels for scenarios with 840 individuals in four-zone (PT Limited) under different ϵ -decay strategies.

C.0.3 Population Size of 560

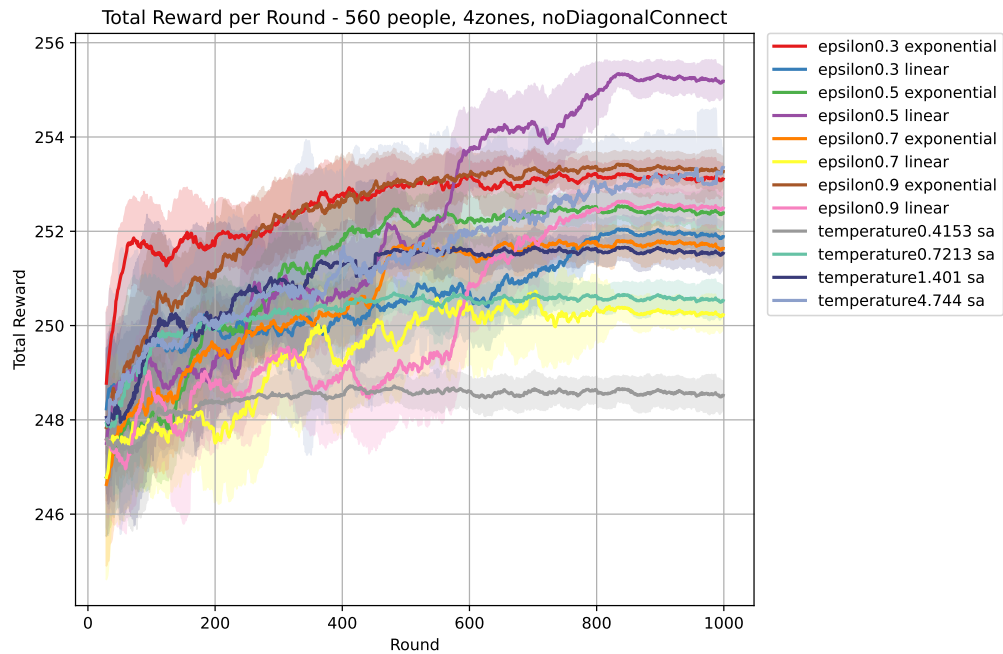


Figure C.13: Moving average of rewards for 560 people in four-zone (no diagonals) under different ϵ -decay strategies.

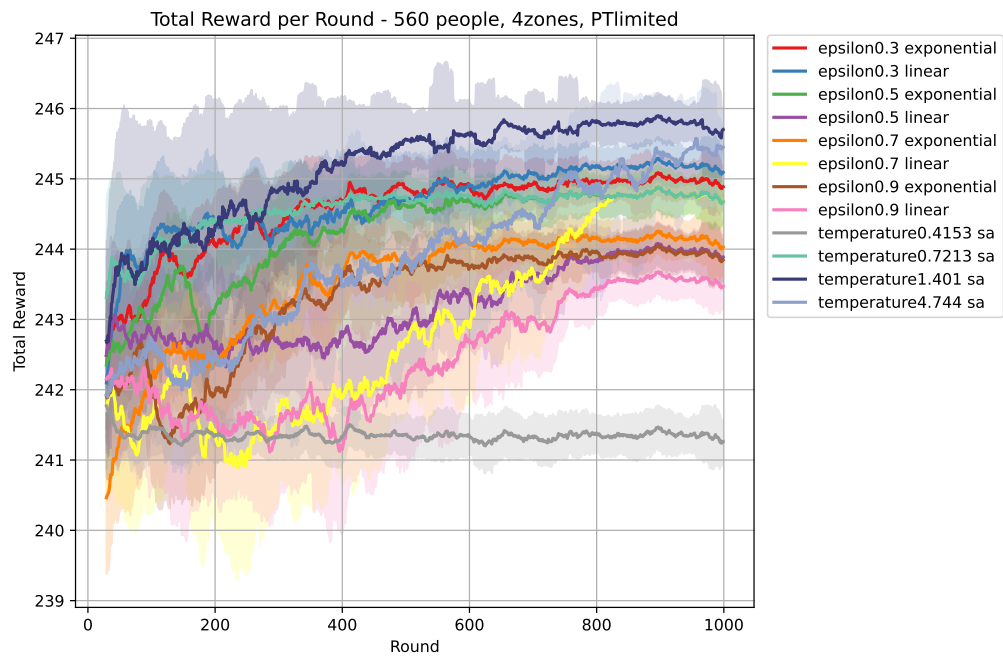


Figure C.14: Moving average of rewards for 560 people in four-zone (PT Limited) under different ϵ -decay strategies.

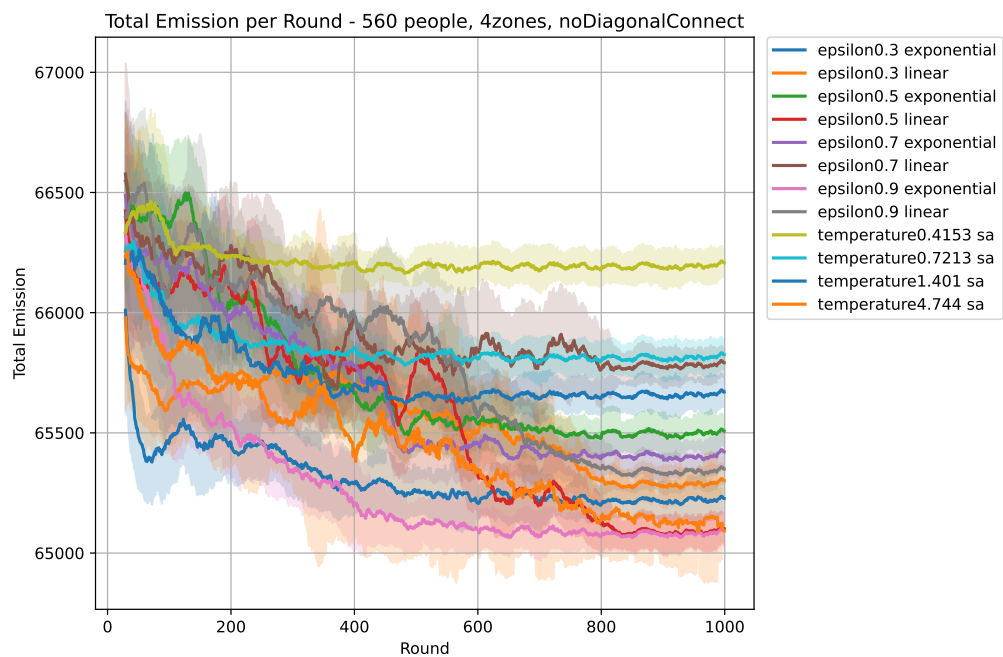


Figure C.15: Emission levels for scenarios with 560 individuals in four-zone (no diagonals) under different ϵ -decay strategies.

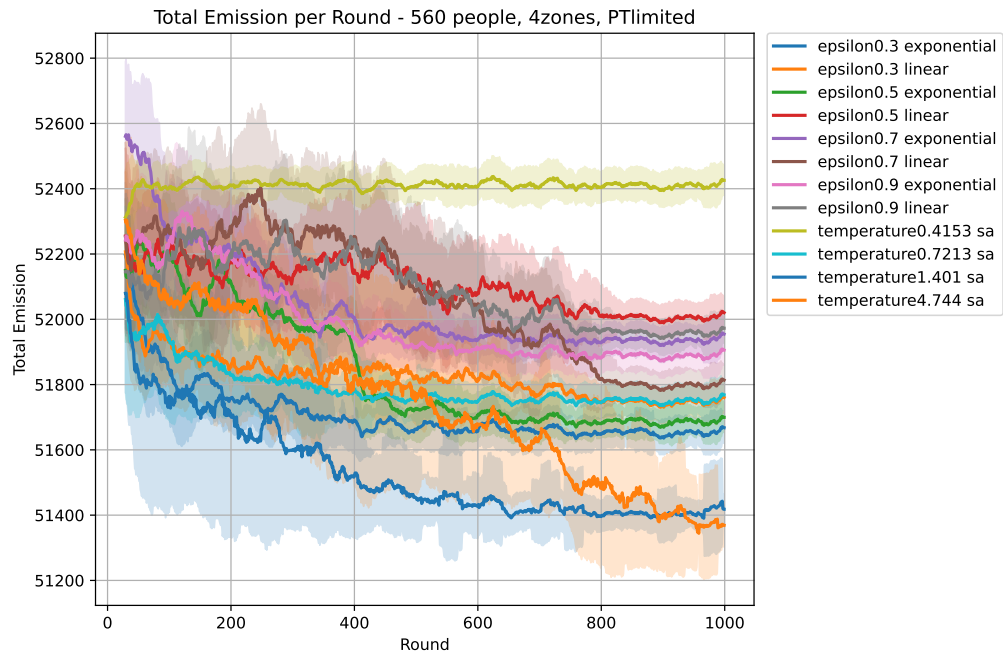


Figure C.16: Emission levels for scenarios with 560 individuals in four-zone (PT Limited) under different ϵ -decay strategies.

C.0.4 Population Size of 360

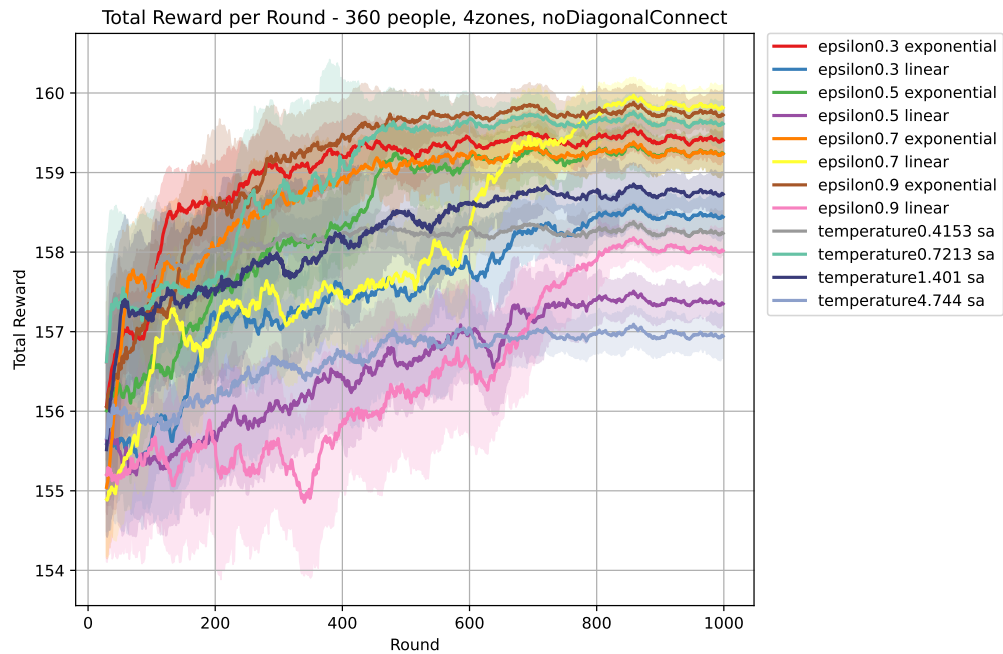


Figure C.17: Moving average of rewards for 360 people in four-zone (no diagonals) under different ϵ -decay strategies.

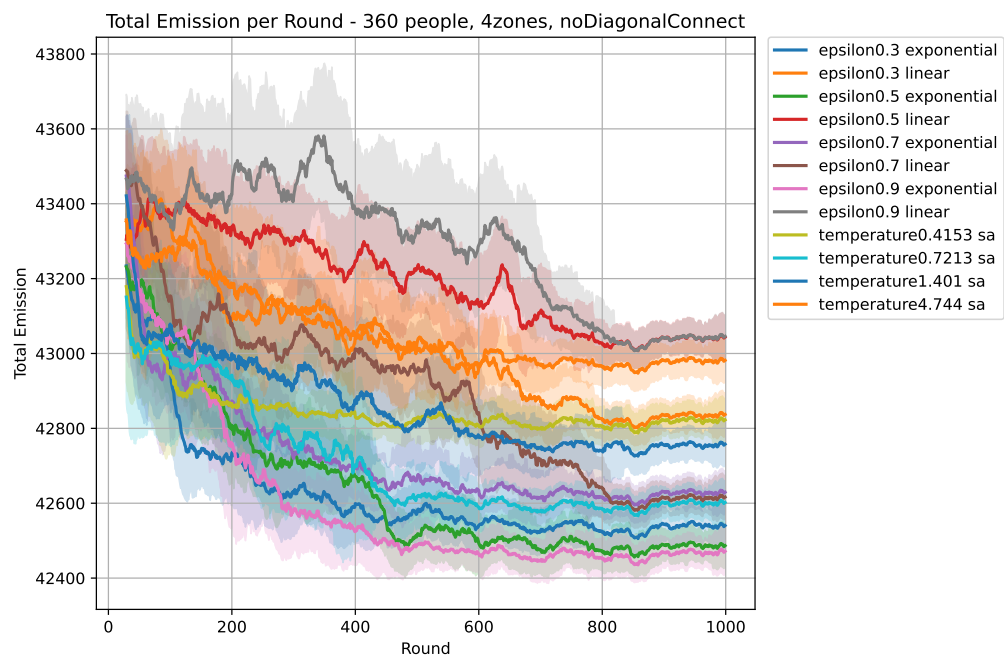


Figure C.18: Emission levels for scenarios with 360 individuals in four-zone (no diagonals).