Reinforcement Learning for Vehicle Routing Problems with Time Windows

A Project

Presented to the

Faculty of

California State Polytechnic University, Pomona

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science in Business

Analytics

In
The College of Business Administration

By

Faria Haque

Govardhini Bandla

Kanwarpreet Singh

Norberto Limon

Minh Vong

SIGNATURE PAGE

PROJECT/THESIS: Reinforcement Learning for Vehicle Routing Problems with Time Windows

AUTHOR: Faria Haque
Govardhini Bandla
Kanwarpreet Singh
Norberto Limon
Minh Vong

DATE SUBMITTED: Summer 2023
College of Business Administration

Dr. Honggang Wang
Project Committee Chair
Professor of Technology and
Operations Management ______________________________

Jae Min Jung, Ph.D.
Project Committee Member
Professor of Marketing ______________________________

ABSTRACT

Purpose: The objective of the project is to determine an optimal set of routes for a fleet of vehicles to service customers with specific time constraints while minimizing overall operational costs. This project also investigates the suitability and effectiveness of different Reinforcement Learning (RL) approaches, such as Deep Q Networks (DQNs), and Policy Gradient methods Advantage-Actor-Critic (A2C) and Proximal Policy Optimization (PPO), in tackling the Vehicle Routing Problem with Time Windows (VRPTW).

Methodologies: Reinforcement Learning has emerged as a powerful technique for solving complex decision-making problems, offering promising solutions to the VRPTW. This project explores the application of RL algorithms to address the VRPTW, aiming to design efficient and adaptive routing strategies.

Findings: Reinforcement Learning approaches are suitable and effective for VRP problems, and based on numerical results gathered, RL solves the challenging VRPTW problem better or comparable than the existing popular methods in terms of solution quality and efficiency.

Originality and Value: This study innovatively applies RL techniques to solve the VRPTW, contributing to improved routing strategies for time-constrained logistics.

Practical implications: The research provides practical insights to enhance routing optimization, resource utilization and operational efficiency across real-world scenarios. The outcomes of this research have the potential to significantly advance the state-of-the-art in vehicle routing optimization, providing valuable insights into the feasibility and practicality of employing RL algorithms to address complex, time-constrained routing problems.

Keywords: Reinforcement Learning, Vehicle Routing Problem, Logistics, Route-Optimization

SUMMARY OF INDIVIDUAL CONTRIBUTION

Faria Haque

In the course of this project, I have been extensively involved in conducting various literature reviews concerning the Vehicle Routing Problem, as well as exploring solutions that leverage Reinforcement Learning techniques. Additionally, I have been responsible for identifying key analytical objectives, evaluating a range of methodologies, performing comprehensive data analysis, creating informative visualizations, and delving into the Hybrid Genetic Search (HGS) Algorithm for baseline analysis. I've collaborated with my team members to compose detailed reports and craft engaging presentations.

Govardhini Bandla

As part of this project, I contributed by reviewing multiple articles on Vehicle routing problems using the time window to understand different algorithms and data types to implement the right method to perform analysis based on our data type. I closely worked with raw data, performed data cleaning processes, data wrangling, analysis, and visualization using Tableau, PyTorch, and Python Libraries, and identified meaningful insights from it.

Kanwarpreet Singh

I have contributed more towards project management and coming up with the initial plan for our goals. I did a literature review for this project and understood the significance of the problem. I worked to find the analytical objectives of this project which were Minimizing the total distance traveled by the vehicles, Minimizing the total travel time for the vehicles, and Delivering goods in their preferred time windows. Also, I contributed to the work on the PPO model for reinforcement learning in addition to Multi-Agent RL. In the last part of the project, I worked on the presentation slides to make them organized and follow the flow of our presentation.

Norberto Limon

For this project, my contribution consisted primarily of conducting research on the viability of some of the Reinforcement Learning Algorithms we tested on our problem set. This consisted of Literature Reviews for Reinforcement Learning, with particular attention to methods utilizing PPO, A2C, and A3C for VRP. I also helped to write the code that was used to evaluate model performance and create our custom OpenAI Gym environments.

Minh Vong

With the project, I have been actively engaged in reviewing multiple literatures as reinforcement learning is a rather challenging topic. I have dedicated my time to the following tasks: Google OR tools implementation in solving VRP problems. DQN implementation in solving VRP problem. Moreover, I was also responsible for the model comparison report as well as recording them in the written report in our Data Analysis section and Conclusions and Recommendations. In addition, I also collaborated closely with my team members in presentation preparation and time management.

Table of Contents

Introduction

A. Background:

The classical version of the Vehicle Routing Problem is used to find a set of optimal routes for a fleet of homogenous vehicles that need to serve a set of customers with routes beginning and ending at a specific depot. Each vehicle route must start and end in the depot. It was first introduced by Dantzig and Ramser in 1959 in Dantzig and Ramser (1959), who proposed a simple matching-based heuristic to route gasoline delivery trucks. Since then, a large part of the OR (Operational Research) community has been devoted to this problem which naturally arises in a large variety of practical applications. This trend has been reinforced by the explosion of consumer direct delivery at the beginning of the century. As an example of this explosion, UPS Ground delivered around 11.5 million packages in 1931, against around 5.5 billion packages in 2019.

B. Managerial Problem Statement:

The VRP is a common business problem in today's complex logistics environment. Nowadays, considering the increased environmental issues, the Vehicle Routing Problem has definitely gained importance. The basic purpose of our research is to analyze the Vehicle routing problem with time windows and find the optimum solution with Reinforcement Learning. The general goal for our solutions to such problems is to find routes that can serve the maximum number of customers with the lowest possible cost. The challenge is to design optimal routes from a base starting point to a set of customer endpoints while accounting for business-specific constraints such as routing options, vehicle capacity, limited resources, pickup/delivery, time windows, etc. We will explore different methodologies within reinforcement learning and compare them to find the best possible solution considering the current state of the environment.

C. Marketing Research Problem Definition

We have found that in the past, there were many solutions provided for VRPTW. Considering those solutions, there is still a need to identify the most promising approaches, optimize the provided solutions and come up with new results. Also, we identified several different variants of VRP like Time Window, Capacitated Vehicle Routing Problem, Vehicle Routing Problem with Pickup and Delivery, VRP with Backhauls, etc for which these solutions would provide benefit.

D. Research Objective and Research Questions

With this research paper, the goal is to shed more light on this challenging topic in regard to its practical implications.

- Conduct exploratory analysis, visualize the data, and extract statistical insights.
- Investigate the feasibility of applying reinforcement learning to vehicle routing problems in the sense of vehicle navigation.
- Build reinforcement learning models in combination with deep learning models
- Build reinforcement learning models in combination with heuristics methods.
- Identifying the models that yield the best results on run time and optimization level.

E. Significance of the Research

Vehicle Routing Problem (VRP) has become an essential issue worldwide under the development of e-commerce, transportation cost has become one of the most pressing factors in business operations. Overall, domestic shipping rates for good transportation by road and rail are up about 23% from 2020. All of which leads to the rising importance and need for routing optimization in the supply chain process. This research will provide insight into how to best optimize existing resources under time constraints using Reinforcement Learning and related methods. VRP focuses on determining optimal routes for a fleet of vehicles given operational constraints such as time window, route length, etc. This is important and is one of our analytical objectives when it comes to solving the issue of rising costs in the supply chain process. Not only for business purposes, VRP also has very practical implications for the commercial use of vehicle navigation systems. For example, Google map is a well-known navigation app here in the States. However, it's a different story in Southeast Asian countries because commuting is a whole lot different with more scooter bikes and fewer cars, which gives locals the flexibility to maneuver through smaller routes of the city that cars wouldn't be able to go through, which leads to a more complex and intricate road map. Throughout our project, our goal is to gain a better look at the solution to the VRP by researching the use of reinforcement learning to achieve the best results on run time and optimization levels.

BACKGROUND AND ANALYTICS OBJECTIVES

Industry Background

The Vehicle Routing Problem is a well-known optimization problem in the field of operations research and logistics. It involves finding the most efficient way to deliver goods or services to customers using a fleet of vehicles. The VRP has many real-world applications in various industries, including transportation, delivery services, etc. Logistics companies are using VRP to optimize the routing of their delivery vehicles to reduce transportation costs, increase delivery efficiency, and improve customer satisfaction. In the transportation industry, VRP can be used to optimize the routing and scheduling of public transit systems, school buses, and taxi services. For example, a school bus company can use VRP to determine the most efficient routes and schedules for picking up and dropping off students. Similar to the Waste Management industry and Field service management industry, VRP can be used to optimize the routes for their respective services in terms of scheduling the service to determine the most efficient routes to travel. Overall, the Vehicle Routing Problem is an important problem that has wide applications across various industries

Mission statement and objectives of target industries

Although a specific mission statement for the Vehicle Routing Problem might not exist, the primary objective of VRP is to enhance transportation and logistics efficiency. By identifying optimal solutions for routing and scheduling a fleet of vehicles, the VRP aims to ensure prompt delivery of goods or services to customers while reducing transportation costs and enhancing overall operational efficiency. This mission underscores the potential advantages that VRP offers across different industries.

Environmental Analysis

Economic factors

The surge in e-commerce users has led to a notable rise in road freight transportation volume. According to Statista's June 2022 report, the United States saw a total of 268 million e-commerce buyers in 2022, with projections indicating an increase to nearly 285 million by 2025 (*Statista, June 2022*). As the number of online orders and home deliveries continues to grow, customers now expect faster delivery times. To address this demand and promote economic growth, solving the vehicle routing problem has become crucial. By optimizing routes, companies can effectively reduce logistic and transportation costs, thereby minimizing travel time and lowering fuel expenses.

Legal/Regulatory factors

I. Artificial Intelligence and Data Privacy Laws

The Artificial Intelligence and Logistics Industries have until recently seen a fairly consistent regulatory environment. With regard to AI, concerns have been voiced at an increasing rate around the globe over data privacy and data sovereignty issues. As such it has become increasingly common for various countries and zones of influence to enact their own data privacy and data localization regulations and frameworks. In some countries like the US and India, privacy is recognized as a fundamental right but federal data privacy regulation has been lagging behind in an effort to strike a balance between the competing interest of fostering innovation and economic growth. In 2018 and 2019 two drafts of the Data Protection Bill were created and ultimately rejected but in 2021 a new draft was introduced in light of a recent Supreme Court Judgement against Pegasus, a spyware tool that was being used by foreign governments to spy on citizens in India and globally. The new bill is believed to have much stronger support and aims to clarify the sensitive circumstances under which data privacy is to be subject to stronger legal protections and also introduces a provision extending data privacy rights of its citizens to apply to any foreign companies processing sensitive data of Indian citizens.

Where data collection is permissible, the bill purports a policy of data minimization and data localization with the implementation of consent mechanisms and data rights management to be made available to all users in India's jurisdiction. Details are currently being worked out concerning the definition of terms such as sensitive data. This is in line with the global trend and is being balanced against the need for cybersecurity and national security which would provide the government with some capabilities to store and protect user data that is given consensually.

This is primarily being addressed by the recently codified Information Technology Guidelines and Digital Media Ethics Code Rules, 2021. (Wadhwa, Bains 2022)

In other parts of the world such as in the EU, Data privacy has been largely addressed with the introduction of GDPR in 2012 and its passing in 2018. In addition to provisions on data portability and the right to be forgotten, it creates a centralized “one-stop-shop” for all of its member states. It deals with the localization issue via the standard of consent which is not deemed to apply to the public sector, thus creating a localized silo for public data as a byproduct instead of explicit provision. It also enacted an independent European Data Protection Board but allows for “legitimate interest” to be used as a reason for the public sector to process user data without the need for consent. (Sponselee, Vreeman, 2018) In a more extreme case of this, China has also recently passed Data Privacy legislation in the form of the Personal Information Protection Law (PIPL) which targets the private sector, restricts communications storage abroad but actually *requires* open data access from Chinese companies by the People’s Liberation Army, including those with data centers outside of its territorial jurisdiction. (Ke, Liu, Luo, Yu, 2021)

In the US, Data privacy has long been a pressing issue but legislation has been slow and patchwork, with different states passing their own data privacy rules such as California’s CCPA. Companies are generally not obligated by the government to provide open access to data but disclosures must be made whenever there is a major data breach to bolster national cybersecurity efforts. In this way, although there is still no federal privacy bill in the US, there is a bill undergoing active review and the legislation around data privacy otherwise largely reflects the data privacy posture of India. The bill currently under consideration is the American Data Privacy and Protection Act, which despite ongoing political polarization is expected to have a real chance at achieving bipartisan support. The bill attempts to curb the issue of perceived algorithmic bias from the acquisition of sensitive data by large companies. Otherwise, the legislative environment is becoming similar in many ways to India’s data privacy posture, providing breathing room for local companies while providing capabilities for centralized national security efforts and generally respecting user privacy via consent but enacting more targeted regulations with regard to jurisdictions believed to be hostile towards its national security posture. (Datagrail, 2022)

The AI field is one that is subject to international regulation and the particular regulations differ across jurisdictions. While some countries like Mexico have yet to even begin properly deliberating issues around data privacy in a significant manner through official channels, there are data protection measures in place that have a minor impact on the AI industry within its borders and the debate is beginning to form within academic circles. (Recio, 2017) However, in general and especially as countries around the world mature with regards to their management of data, the trend is generally moving towards prioritization of data localization measures and data privacy restrictions that vary slightly with regards to private and public sectors pursuant to the local priorities and concerns within that jurisdiction. The impact this may have on the AI industry is likely to be that companies will struggle briefly to adapt to the new normal as some jurisdictions are cut off or become more difficult to operate in but in the long term will operate with much more clarity and certainty, especially incentivizing new startups, possibly leading to a delayed surge in the AI space across the board, including in the use of reinforcement learning to solve the VRP and other challenges.

II. Logistics

While the industries impacted by the VRP are wide-ranging from transportation and transit to fleets of electric vehicles, one of the most heavily impacted sectors would be the logistics industry. This too would vary as different regulations affect different types of logistics in different ways. Overseas logistics for instance would be subject to maritime laws, airborne logistics would be subject to their own standards and land-based vehicles would also be subject to different regulations depending on jurisdiction. However, the most closely aligned industry to the problem which is the subject of our research is vehicle logistics, particularly with regard to trucking and smaller delivery vehicles.

In light of the broad range of jurisdictions, we will instead focus our attention on some of the recent impacts on the industry and how they play out in two divergent approaches that are likely to be representative of a wider decision faced by logistics companies around the world. The decision in question is whether or not to use additional labor from independent contractors to help fulfill an ever-increasing demand for delivered goods and how both decisions are affected by the broader notion of increasing demand for delivered goods. The two examples we will be looking at are California and South Africa in the post-Covid industry. In addition to increased pressures from mandatory stay-at-home orders and holiday spending patterns, there are an ever-increasing number of companies emerging to crowdsource deliveries. The likes of Uber and Amazon have contributed greatly to this trend but similar platforms like the company Grab in Thailand contribute in a similar manner that can impact their counterpart industries in other parts of the globe in the same manner.

First with regard to the question of hiring independent contractors, the trucking industry is one that has historically relied heavily on the use of independent contractors for dealing with elasticity in the demand for delivered goods. As demand increases across the board, the use of independent drivers also increases. However, with this comes the question of local labor laws pertaining to a mandatory minimum number of rest periods. In California, at the onset of the

Covid-19 pandemic, the attorney general attempted to impose stronger worker protections by reclassifying large segments of the independent driver pool to be deemed employees instead of independent contractors. This would put a higher burden on logistics companies as they would be subject to more stringent pay and mandatory rest period requirements. This was seen by some as an attempt to provide additional worker protection during a strenuous period but would also have the effect of shifting that burden to logistics companies which may or may not be able to afford the changes. (Smith, 2020) Although in this case, a federal judge struck down the injunction for this reclassification, the matter is being considered for appeal, and at any rate, if assumed to be successful it would represent the move in the direction towards increased labor protections at the expense of the logistics industry more broadly. This is one possible decision that could be made by many other jurisdictions and so it is safe to assume that at least some will move in that direction, leading to a net added constraint on the logistics industry that yearns for the need for increased route optimization in order to survive the ensuing circumstances of regions affected by decisions of this fashion.

On the other side of this debate is an example from an article discussing a similar debate in South Africa. Here the author advocates for the emergence of a Temporary Employment Services partner (TES) to make up for the increased demand. This is in line with the reasoning behind the independent contractor thesis discussed above but potentially allowing for workers to receive the benefits of being full-time employees of these agencies and working for different companies as third parties to provide the flexibility that is offered by the independent contractor model. (Govender, 2022) One example that might help to illustrate the author’s point is the emergence of Uber as a means of satisfying the increased demand pressure. Although there might be an added expense with outsourcing labor temporarily to a platform like this (assuming the actual TES had their workers on board as employees), by being able to meet the demand, the increased cost in the short term should theoretically be outweighed in the medium and long term by the increase in business and continued growth resulting from maintaining their brands and meeting customer service expectations by successfully satisfying orders as they are placed in a timely fashion. In any case, the line of reasoning is similar. More demand requires more workers and there may be ways to meet this demand with increased labor without the concern of exploitative labor environments.

One thing that these approaches do not seem to account for however is the fact that delivery-based goods and services are increasing across the board even without accounting for the recent effects of Covid-19. This means that even with staffing concerns being addressed by the TES model there will still be external pressures to optimize to keep up with the demand. While this may or may not be temporarily stifled by economic conditions and arguably diminishing birth rates, the broader trend is likely to persist over time. In either case, this could create ripe conditions where solutions to the VRP are increasingly sought out, leading to further applicability and relevance for AI-based optimization models.

Technological and Environmental Factors

In terms of technology, the vehicle routing problem has a strong effect on the environment in which the problem arises. Vehicle routing is a prevalent issue in urban areas especially due to the influence of traffic flows and the density of street networks, which is a constant state of changing and renovating. Municipal authorities have attempted to restrict transport with access limitations, hours rules, or even European Emission Standard regulations for vehicles to tackle traffic problems. In addition, there are technological factors that affect VRP:

1. GPS technology: the application can be used to track vehicles, monitor their movement, and calculate real-time traffic conditions, which helps to optimize and reduce travel time and fuel costs.
2. Vehicle telematics: A closely related to GPS technology that can provide real-time information about vehicle location, speed, and fuel consumption. This information can be used to optimize routes, and plan for fuel stops.
3. Internet of Things (IoT): sensors can be used to monitor traffic flow, weather conditions, and other factors that can impact the VRP, whose application can also be used to optimize routes and reduce travel time

Furthermore, VRP has a tremendous impact on the environment with its purposes. The ultimate goal of VRP is to optimize the total distance traveled in order to best reduce time and money, which positively relates to reducing the amount of emission.

Social Factors

Social factors play a crucial role in the Vehicle Routing Problem, and it is essential to consider them while devising optimal routes for a fleet of vehicles. The VRP is influenced by various social factors, including:

1. Urbanization: With cities experiencing rapid growth and increasing population density, there is a foreseeable rise in demand for delivery services. However, navigating through congested urban areas may pose challenges, leading to longer travel times and higher transportation costs.

2. Labor Costs: The expenses related to labor, including wages and benefits, can significantly impact transportation costs. Companies may adjust their route designs to minimize labor expenses and optimize efficiency.

3. Traffic: Traffic congestion is a major social factor that directly affects vehicle travel times and delivery schedules. Dealing with heavy traffic can result in increased transportation costs, longer delivery times, and dissatisfied customers.

4. Environmental Concerns: The emphasis on environmental sustainability has led to a greater focus on reducing carbon emissions and the ecological impact of transportation. Companies are increasingly seeking optimized routes that minimize fuel consumption and contribute to a greener supply chain.

Taking these social factors into account during the VRP planning process is vital to ensure efficient and cost-effective delivery operations while addressing environmental and societal concerns.

Cultural factors

The vehicle routing problem with time windows is an optimization problem, in which a set of vehicles must visit a set of customers within a given time window while minimizing costs such as distance, travel time, and number of vehicles used. This problem can be solved by learning optimal policies through trial and error interaction with the environment.

Some cultural factors that impact the development of reinforcement learning in this context include:

1. Local Customers and Practices: local customers and practices impact the design of the reinforcement learning algorithm.
2. Organizational culture: the culture of the organization using the reinforcement learning algorithm will impact the adoption and implementation of the algorithm.
3. Education and training: it will impact the adoption and implementation of the reinforcement learning algorithm.
4. Language and communication: the language used to describe the problem and the reinforcement learning algorithm will impact the ease of adoption and implementation.
5. Social norms: social norms will impact the willingness of users to adapt to new technologies.

Industry Analysis

Industry size and outlook

The Fourth Industrial Revolution serves companies in developing a modern supply chain (MSC) system when they are faced with a dynamic process. Vehicle routing problem (VRP) is a good framework that focuses on mobility and real-time integration for the industry.

The United States logistics industry is a highly integrated supply chain network. Supply chain networks link consumers and producers through different transportation modes. In 2020, the North American logistics market was sized at over two trillion U.S. dollars and ranked second after Asia-Pacific. The Asia-Pacific region's logistics market reached more than 3.9 trillion U.S. dollars that year (*Statista,*2022).

Leading logistics companies like UPS Supply Chain Solutions, FedEx Freight, Amazon, DHL Supply Chain, J.B.Hunt Transport Services., USPS, C.H. Robinson, and XPO Logistics Inc. are a few of the potential beneficiaries that can use vehicle routing problems for mobility and real-time integration.

Four major competitors and their strategies

- Target market

  1. Transportation and logistics:
  2. Municipalities and government:
  3. Manufacturing and distribution:
  4. Retailers and e-commerce:

- Positioning

Positioning of vehicle routing problems is a solution to help companies optimize the transportation and logistics process. Mainly VRP is positioned as a solution for the following:

By optimizing vehicle routes and schedules, organizations can reduce transportation costs, including fuel, labor. VRP solutions help organizations improve efficiency of transportation and logistics by reducing time to complete services.

- Product/service

The product and services of Vehicle Routing Problem (VRP) solutions typically include software solutions that enable businesses to optimize their routing and scheduling of delivery vehicles.

These solutions can be offered as off-the-shelf products or customized to meet the specific needs of a business

- Price

The price of implementing a VRP solution in transportation and logistics can vary depending on several factors like the complexity of the problem, size of fleet, number of deports and customers.

- Distribution

Vehicle Routing Problem (VRP) solutions are used in transportation and logistics to optimize the routing and scheduling of delivery vehicles. For instance: Retail, E-commerce, manufacturing, healthcare, etc.

- Summary table of comparison

Target market	Transportation and logistics	Municipalities and government	Manufacturing and distribution	Retailers and e-commerce
Positioning	Improve efficiency	Reduce environmental impact	Flexibility and scalability	Cost savings
Product/service	Furnitures, shifting, import, export	Courier, mails, collections	Retail, industrial	Data analysis and reporting

Distribution	export	Healthcare, waste management	manufacturing	Delivery to customers
Promotion	Marketing, ads	Education and training	Trade shows and events	Email marketing

Customer Analysis

The Vehicle Routing Problem is a broad topic that can be applied to all industry segments because logistics of goods is essentially the soul to a living economy. Along with that, VRP is an ongoing problem for a large customer base. To further research the customer analysis, it is best to have a look into the logistic aspect based on product segments.

Segments and their brief justification

Consumer goods: Consumer goods refer to a class of stocks and companies that cater to items purchased by individuals and households, rather than manufacturers and bulk buyers. Within this sector, companies produce and sell products intended for direct use and enjoyment by the consumers. It encompasses various types of goods, including convenience goods, shopping goods, unsought goods, and specialty goods, all of which fall under the same umbrella of consumer goods.

E-commerce : On the other hand, e-commerce businesses deal with different categories of goods. Firstly, there are physical goods, such as books, gadgets, furniture, and appliances, which are tangible products shipped to customers. Secondly, digital goods form another category, including intangible items like software, e-books, music, text, images, and videos, which can be downloaded or accessed online. Finally, e-commerce platforms also offer services such as ticket sales and insurance, further diversifying the range of products available to consumers.

Research Framework/Model

Background

Reinforcement learning is a type of machine learning wherein an agent interacts with its environment and learns by trial and error to achieve the optimal policy over time. This is done in a similar way to how positive and negative incentives stimulate learning in humans and other living organisms in relation to their environments. It uses these incentives to train an agent to take some desired action in an environment. Reinforcement learning is a broad field of study that encompasses several techniques, a few of which we will be discussing below as prime candidates for solving the Vehicle Routing Problem. They are as follows:

A. Reinforcement Learning with Tree-based Regression

B. Reinforcement Learning with Variable Neighborhood Search Method

C. Multi-Agent Reinforcement Learning

D. Deep Reinforcement Learning

E. Deep Reinforcement Learning with Local Search

F. Integrated Deep Reinforcement Learning with Heuristics Approach

Additional Context: When referencing reinforcement learning techniques, it is helpful to understand a few contextual concepts.

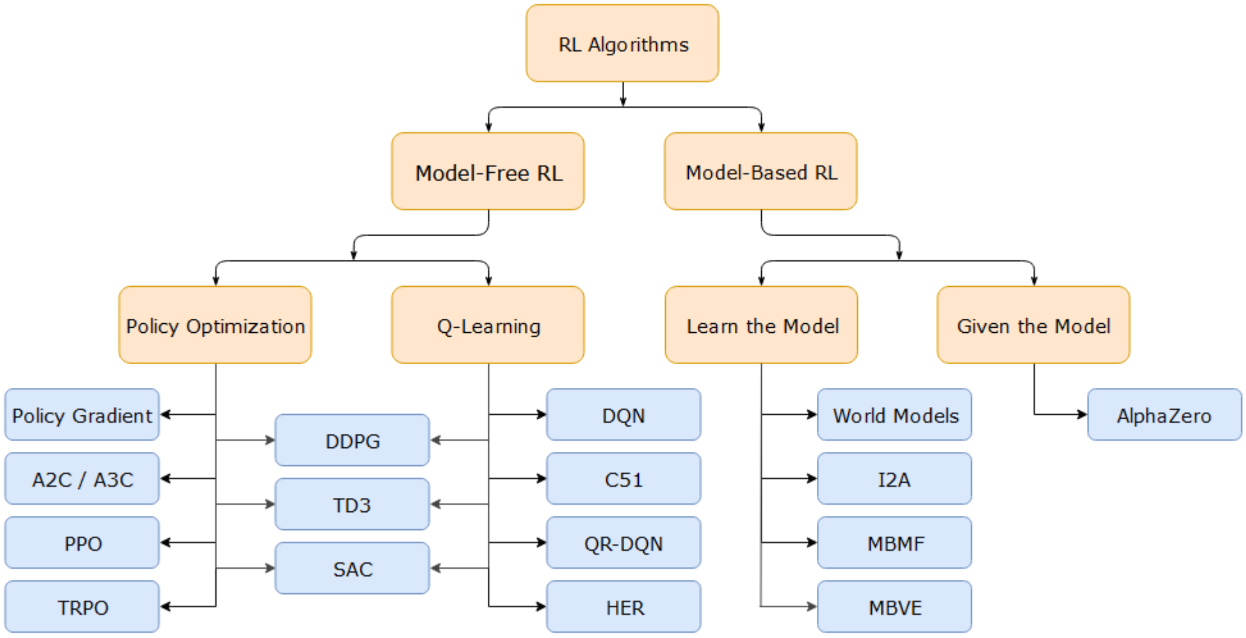
1. The Markov Decision Processes (MDP)

2. The Bellman Equation

3. Q-Networks

4. Policy Gradients

Taxonomy of Frameworks/Models: All reinforcement learning frameworks can be classified as either model-free or model-based, referring to whether or not they have access to a model of their environment, which consists of a function that predicts state transitions and rewards. Here is an illustration from OpenAI's documentation on reinforcement learning that includes a non-exhaustive taxonomy of reinforcement learning algorithms classified under either model-free or model-based to demonstrate this notion.



Source: https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html

Both approaches have their pros and cons but in short, a model-based approach would provide the agent with some guidelines known as policies that allow it to plan accordingly. On the other hand, a model-free approach would not have this set of policies which can be both a benefit and a detriment in that it takes much longer to learn but it learns without inheriting the biases in the policies that might not be relevant or appropriate for the environment in which the model is deployed. Most RL techniques fall in one classification or the other but deep reinforcement learning can be either or depending on whether or not the forward dynamics of the environment are estimated before deployment. We will elaborate on this further as we cover deep reinforcement learning.

A. Reinforcement Learning with Tree-based Regression

Although there are several ways to use RL for solving the VRP, one approach that seeks to optimize vehicle routes while simultaneously considering changing requirements in a dynamic environment is the use of traditional Reinforcement Learning in conjunction with a tree-based algorithm. This approach was introduced in 2021 by researchers Thananut Phiboonbanakit, Teerayut Horanont, Van-Nam Huynh, and Thepchai Supnithi. (Phiboonbanakit et al., 2021)

In many use cases for which the VRP is relevant, there are several factors that can change the conditions under which the RL agent is expected to operate. For instance, a traffic collision, weather conditions, or changing customer requirements might have a sudden effect on route planning that might require the need for route changes or deployment of additional vehicles to re-calibrate to changes in the environment and meet the demands of the users.

In order to deal with the changes, this approach proposes that the state and actions of the RL agent be fed to a tree-based regression model that is used to assess whether the current route is feasible and the response is returned to the RL agent in order to adjust actions for optimizing the vehicle routing task. This adds the ability to respond to changes in a real-world environment that is limited by pure traditional Reinforcement Learning. With Model-free RL, the agent can become stuck in an infinite search when it lacks feedback from the external environment. On the other hand with Model-based RL the model is able to update itself but can be rigid in its approach with changes that do not keep up with the changing ecosystem. This method thereby attempts to take principles from both in a "hybrid" model approach that has the flexibility of model-free and the reactivity of model-based RL.

B. Reinforcement Learning with Variable Neighborhood Search Method

The Variable Neighborhood Search (VNS) method, initially introduced by Mladenović and Hansen in 1997, is a popular metaheuristic approach that has undergone continuous development to address a range of optimization problems (Hansen et al., 2019a; Mladenović et al., 2021) and also machine learning tasks (Mladenović et al., 2021). The VNS algorithm utilizes a local search approach that involves systematic modifications to neighborhood search structures. These modifications occur during both the descent phase, where the algorithm seeks to find a local optimum, and the perturbation phase, which enables the algorithm to escape from the corresponding valley (Hansen et al., 2017). VNS was selected due to its scalability and versatility in addressing a wide range of optimization problems.

The VNS approach involves a systematic alteration of neighborhood structures during the search process to facilitate comprehensive exploration of the search space. VNS is a stochastic metaheuristic technique that includes either a basic local search component or a more potent version known as Variable Neighborhood Descent (VND). The VND methodology employs a deterministic alternation of neighborhoods and relies on three strategies to explore them: (i) randomly, (ii) deterministically, or (iii) a mixture of both in a deterministic and random order (Duarte et al., 2018). The utilization of VND as the local search procedure within VNS gives rise to a more comprehensive technique called General Variable Neighborhood Search (GVNS). The implementation of this generalized approach has resulted in numerous successful applications documented in the literature. In addition to GVNS, the study employed reinforcement learning techniques by introducing Bandit VNS. The computational experiments conducted on both methodologies yielded improved results and facilitated a comparative analysis between them.

To conclude, the framework presented in this study integrates reinforcement learning and metaheuristics (two well-studied topics) to address the capacitated vehicle routing problem and yields promising methods along with several avenues for future research.

C. Multi-Agent Reinforcement Learning

Multi-Agent Reinforcement Learning (MARL) is a subfield of machine learning that deals with developing algorithms and techniques for coordinating the behavior of multiple agents in an environment. The agents typically interact with each other and with the environment, and their goal is to learn how to optimize some objective function while taking into account the actions of other agents.

Multi-Agent vehicle routing problem with time windows is an indispensable constituent in urban logistics distribution systems. Over the past decade, numerous methods for MARL have been proposed, but most are based on heuristic rules that require a large amount of computation time. One approach to solving the VRP using MARL is to use a decentralized approach, where each agent makes its own decisions based on local observations and communication with other agents. In this case, each agent is trained using reinforcement learning, where the reward signal is a function of the overall performance of the fleet. The agents can communicate with each other to exchange information about their local observations and actions, and they can use this information to coordinate their behavior and improve their performance.

The benefits and challenges of multi-agent reinforcement learning are described, a central challenge in the field is the formal statement of a multi-agent learning goal. Several multi-agent reinforcement learning algorithms are applied to an illustrative example involving the coordinated transportation of an object by two cooperative robots. In an outlook for the multi-agent reinforcement learning field, a set of important open issues are identified, and promising research directions to address these issues are outlined.

D. Deep Reinforcement Learning

Deep Reinforcement Learning (DRL) is a specialized area within machine learning that combines the power of deep learning and reinforcement learning techniques. Its primary goal is to empower agents to learn and make decisions in complex and dynamic environments. In recent years, deep learning, particularly deep neural networks, has garnered significant attention in the field of reinforcement learning. This integration has been successfully applied in various domains such as games, robotics, and natural language processing. Deep learning leverages neural networks with multiple layers to comprehend intricate representations of data, making it especially adept at learning from unstructured information like images, speech, and text.

DRL has proven effective in achieving optimal solutions for diverse tasks and objectives, including mastering games, controlling robots, and managing traffic flow. However, a notable limitation is that training DRL agents can be computationally expensive and demands substantial amounts of data. To address this concern, current research in DRL is focused on enhancing the efficiency and scalability of DRL algorithms.

E. Deep Reinforcement Learning with Local Search

The Vehicle Routing Problem mainly gives an approximate solution by using heuristic methods. Construction algorithms and local search algorithms are two main categories of heuristic algorithms that are commonly used in VRP. Construction algorithms often sacrifice the solution quality for its high efficacy whereas local search uses different search operators to get better solutions.

In deep reinforcement learning, local search is a well-known technique to solve combinatorial optimization problems. The main concept of local search is to iteratively move from a solution to a neighboring solution by applying local perturbations until it gives an optimum solution.

The local search algorithm contains a series of perturbations called moves. Each move makes small changes in the present solution, to get an alternative solution that is closer to the previous solution. From these previously selected alternative solutions, the next solution is selected which is called the neighborhood. The neighborhood is generated by operators. Each operator defines a type of perturbation to apply. Real-world examples where local search has been applied are the vertex cover problem, the traveling salesman problem, the boolean satisfiability problem, the nurse scheduling problem, the k-medoid clustering problem, and the Hopfield neural network problem.

The local search approach greatly increases performance and gives optimal solutions within a short period of time. For big instances where the search space is too large, the local search algorithm works well within a reasonable time. It also consumes low memory, as the neighborhood considered at each step is comparatively small compared to the whole search space.

F. Integrated Deep Reinforcement Learning with Heuristics Approach

Another way of solving the VRP with Deep Learning involves the fusion of deep learning with a classical heuristics approach. Although it would be theoretically possible to find solutions to problems like the VRP using heuristics methods alone, these methods take an extraordinary amount of time that makes applications to the VRP in real-time infeasible. Therefore this approach attempts to address this challenge with a two-layer method that uses integrated recurrent neural networks with an attention mechanism as well as heuristics methods.

For this context it may be helpful to equate heuristics methods with what would otherwise be commonly referred to as brute force approaches. In other words, these are approaches that are extensive and may not be optimal but are capable of eventually finding the solution to a given problem. This provides the timeliness of deep learning with the robustness of heuristics methods such as local search algorithms, genetic algorithms, colony algorithms, and Lin Kernighan Heuristics (LKH3) which is an improved version of a local search algorithm called 2-opt that has previously been applied to the Traveling Salesman Problem, a predecessor to the VRP.

The proposed method works as follows:

1) An initial solution method is calculated with deep reinforcement learning

2) The initial solution is fed to a secondary results solution method which performs optimization using LKH3. By acting on the DRL results instead of the entire data set, the Heuristics model is able to compute much faster.
3) Both the initial DRL solution method and the secondary heuristics optimization method are fed to an evaluation model
4) A final solution is determined by comparing the two models as selecting the more optimal solution.

The results of this approach were tested using three types of decoders (Greedy DRL, Beam DRL, and SampleDRL). It was discovered that this approach can be very effective in small-sized problems but the contribution of the hybrid approach compared to the strictly DRL approach becomes more pronounced with larger samples of data. (Aktas et al., 2022)

Analytics Methods (Ch03)

A. Data and Sampling

Population: For the Vehicle Routing Problem (VRP), there is not one specific population data. However, this problem can be applied to various real time scenarios such as logistics and transportation, where the population data may be relevant. In such cases, the population data would refer to the set of customers that need to be served by the fleet of vehicles. The population data typically includes information such as location of each customer, their demand (i.e., the amount of goods to be delivered to them), and their service time (i.e., the time required to make deliveries).

Some of the methods to gather data for the VRP:

1) Randomly generate customer locations: A random number generator to generate the customer locations within a defined geographical area. Subsequently, the distance between each customer and the depot can be calculated to determine the distance matrix.
2) Real-world data: real-world data such as customer addresses or GPS coordinates to generate the customer locations. The dataset used in our project comprises 250 static VRPTW instances. These instances use data from a US-based grocery delivery service containing between 200 and 900 customers.
3) Use synthetic data: data can also be generated from a simulation or a statistical model. Synthetic data is useful when you need to test an algorithm's performance on different problem sizes and complexities.

Sampling frame: Dynamic instances were created by sampling 100 requests

Sampling method: Created by the environment from the static set of customers during a number of epochs that is between 5 and 9, depending on the instance.

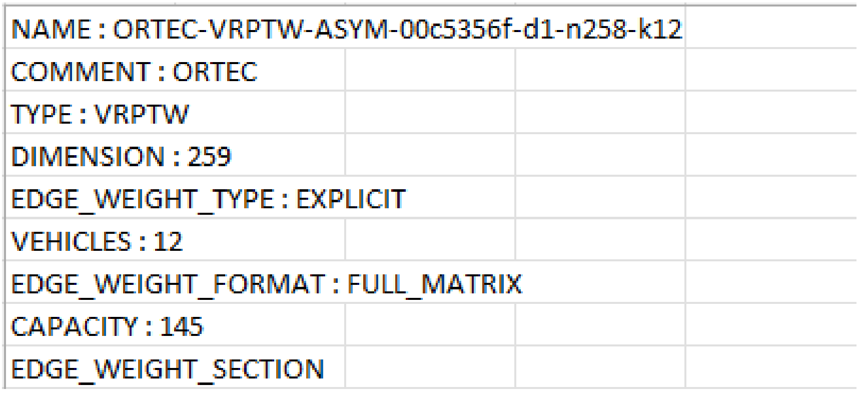
B. Data Wrangling Plan

Data Collection: A public dataset with 250+ instances is provided by ORTEC. This dataset is real-time data from a US-based grocery delivery service that is anonymous. These instances contain between 200 - 900 customers.

Data Wrangling:

Import: pandas

Tidy: pandas



NAME : ORTEC-VRPTW-ASYM-00c5356f-d1-n258-k12
COMMENT : ORTEC
TYPE : VRPTW
DIMENSION : 259
EDGE_WEIGHT_TYPE : EXPLICIT
VEHICLES : 12
EDGE_WEIGHT_FORMAT : FULL_MATRIX
CAPACITY : 145
EDGE_WEIGHT_SECTION

Understand Data: The image above briefly introduces the elements of our project. We have a distance matrix of 259 with 12 vehicles and each vehicle has a capacity of 145 item units. Our instances have a similar format containing:

1. Edge_Wright_Section: contains the full duration matrix.
2. Node_Coord_Section: This matrix provides
3. Demand_Section: This provides us with the demand of each customer per location.
4. Depot_Section: All instances are provided in the VRPLib format, following the convention used by LKH3. Following this format, the depot is referred to as node 1, and the remaining locations are 2 to n, but we generally refer to the depot as node 0 and locations as 1 to n-1.
5. Service_Time_Section: This variable shows us the time it takes to complete a service at each location.
6. Time_Window_Section: This showcases the time constraint per location.

1) Programming: Our program consists of the following functions, Methods, and Objects:

    1. Agent class/object

        a. choose_action method

    b. learn method

2. Environment class/object

    a. reset method

    b. step method

C. Analytics Methods to Employ

The goal of this project is to explore reinforcement learning methods for solving the Vehicle Routing Problem. To achieve this goal we will complete the following analytic methods:

1: Conduct Exploratory Analysis, Visualize the Data and Extract Statistical Insights

Exploratory analysis plays a crucial role in understanding the data before delving into more advanced techniques, particularly in the context of the Vehicle Routing Problem. By thoroughly examining the dataset containing information about customers, their locations, demands, and available vehicles with their capacities, we aim to uncover data characteristics and patterns.

This initial analysis involves exploring variables and their distributions, detecting any missing values or outliers, and gaining a comprehensive understanding of the data structure. The insights obtained during this exploration help inform decisions regarding data preprocessing and feature engineering steps.

Visualizing the data is a fundamental step to gain valuable insights and identify patterns or trends that may exist. Various visual representations like scatter plots, histograms, or heatmaps are created to understand the spatial distribution of customers, their demands, and vehicle capacities. Additionally, we utilize plots to analyze temporal aspects, such as customer demand fluctuations over time or vehicle availability. These visualizations offer valuable knowledge about the problem's characteristics and potential areas for improvement.

Statistical analysis complements the exploration by revealing meaningful patterns and relationships within the data. Descriptive statistics are computed to understand central tendencies, variabilities, and distributions of relevant variables related to the VRP. Measures like mean, median, standard deviation, or skewness provide insights into customer demands, distances between locations, and vehicle capacities. Furthermore, hypothesis testing, and correlation analysis may be conducted to examine relationships between variables. Such statistical insights guide decision-making in the implementation of reinforcement learning methods for tackling the VRP.

2: Investigate the feasibility of applying RL to VRPTW for vehicle navigation

The use of Reinforcement Learning has shown promise in addressing the Vehicle Routing Problem, where the goal is to optimize routes for a fleet of vehicles serving customers while considering various constraints. RL offers flexibility, efficiency, and adaptability to dynamic environments, making it a suitable candidate for VRP-related navigation challenges.

To apply RL to VRP, a framework is proposed, defining the state representation, action space, and reward function for RL algorithms. Among the popular libraries in Python for RL development are TensorFlow, PyTorch, Keras, and Stable Baselines. For R programming, the 'rlang' package provides the necessary functionalities.

Several techniques, including Q-Learning, Hybrid-Genetic Search, Advantage-Actor Critic, and Proximal Policy Optimization, were tested to explore their viability in solving VRP.

While some techniques produced sensible results, others posed challenges, indicating that RL can be used to solve VRP but requires careful consideration of each model's characteristics for effective data ingestion.

However, certain challenges and limitations with RL for VRP navigation were identified. These include scalability issues with larger problem sizes, the exploration-exploitation trade-off, and the computational resources needed for complex RL models. Despite these challenges, RL remains a promising approach for VRP navigation, and further research and experimentation are required to address these limitations and fully unlock its potential in solving VRP challenges.

3: Build RL models in combination with Q-learning models

The model implements a Q-learning algorithm for solving the vehicle routing problem. From the previous recap, VRP is a combinatorial optimization problem in which a fleet of vehicles with limited capacity must serve a set of customers with known demands while minimizing the total distance traveled.

The code defines a VRP environment class that simulates the VRP environment and implements the step() function, which takes action (i.e., choosing a customer to visit) for each vehicle and returns the next state, the reward, and whether the episode is done,

The train-vrp_agent() function uses the Q-learning algorithm to learn the optimal policy for the VRP. It initializes a Q-table with a shape that can hold all possible states and actions and trains the agent for a specified number of episodes. The code uses an epsilon-greedy exploration strategy to balance exploration and exploitation.

The test-vrp-agent() function tests the learned policy on a new set of VRP instances. It chooses an action for each vehicle based on the Q-value and updates the state and reward accordingly. The function returns the total distance traveled by all vehicles.

The Q-learning update rule is as followed:

$$Q(s, a) \leftarrow Q(s, a) + \alpha(r + \gamma \ \mathrm{max\_a'} \ Q(s', a') - Q(s, a))$$

- $Q(s, a)$ represents the Q-value of state-action pair (s, a)
- α is the learning rate
- r is the reward obtained by taking action a in state s
- γ is the discount factor
- $\mathrm{max\_a'} \ Q(s', a')$ is the maximum Q-value over all possible actions a' in the next state s'

Epsilon-greedy exploration:

```
if random.uniform(0, 1) < $\varepsilon$:
    a ← random action
else:
    a ← argmax_a' Q(s, a')
```

- ε is the probability of taking a random action (exploration)
- $\mathrm{argmax\_a'} \ Q(s, a')$ is the action that maximizes the Q-value in state s

4: Explore and Test RL models in combination with heuristics and multi-agent models

Service time is a principal component in determining customer satisfaction. In order to optimize on time, we first needed to explore several RL algorithms to see which would be suitable for solving the VRP. In order to compare a representative variety of the types of

Reinforcement Learning Algorithms that are available for solving the VRPTW, we came up with three categorical approaches to VRP, each with a collection of specific algorithms that can be deployed and fine tuned to fit our use case. These three categories ate:

1) Reinforcement Learning with Deep Learning:
2) Reinforcement Learning with Heuristics
3) Multi-Agent Reinforcement Learning

To represent these approaches, we decided to look at some simple ways we could test the algorithms against one another in a uniform manner. This is where we discovered stablebaselines3 and OpenAI's Gymnasium library. Stablebaselines3 offered us the ability to quickly deploy various different RL models. Used in conjunction with Gymnasium, a library that helps to standardize testing environments, we were able to easily compare the performance of several algorithms that represented our 3 categorical approaches. The algorithms we selected for analysis were as follows:

1) DQN - Reinforcement Learning with Deep Learning
2) SAC / A2C - Reinforcement Learning with Heuristics
3) PPO - Multi-Agent Reinforcement Learning

With this approach we were able to quickly compare metrics across several models although not with the same level of specificity as our OR-Tools or the HGS (Hybrid Genetic Search Algorithm) approaches used for our baseline. This expanding body of knowledge will help facilitate the subsequent evaluation. Preliminary results seem to indicate that DQN is not well equipped to handle route optimization in our test environment without further modification but SAC, A2C and PPO were able to produce meaningful results out of the box. Based in part on the metrics gathered we were able to produce some visualizations through the use of tensorboard that illustrated the performance differences in the algorithms on our test environment. Further visuals were also generated to extract insights from the raw dataset. Our next step is to deploy our algorithms in a custom environment tailored to our dataset and obtain performance metrics as we did in the Gymnasium test environment..

5: Identify the models that yield the best results on run time and optimization level

Due to the dynamic nature and high complexity of VRPTW, identifying RL models that yield efficient results in terms of both run time and optimization level can be challenging. Nevertheless, a few of the RL models that demonstrated promise for handling this type of problem are as follows:

Deep Q-Network (DQN): DQN is a popular RL algorithm that combines deep learning with Q-learning. It has been applied to VRPTW by representing the problem as a sequential decision-making process. DQN-based models can achieve good results in terms of solution quality, but their run time may vary depending on the complexity of the VRPTW instance. However, with this project, DQN did not yield a desirable solution.

Proximal Policy Optimization (PPO): PPO is a policy optimization algorithm that has been used in VRPTW to learn policies that map the state of the problem to actions (e.g., selecting the next customer to visit). PPO offers a balance between exploration and exploitation, which can lead to efficient solutions. However, the run time of PPO can be influenced by factors such as the size of the problem and the complexity of the policy network.

Deep Deterministic Policy Gradient (DDPG): DDPG is an RL algorithm that combines deep learning with deterministic policy gradients. It has been applied to VRPTW to learn policies for vehicle routing decisions. DDPG can handle continuous action spaces, which is useful in VRPTW where decisions involve selecting coordinates for the next customer. The run time of DDPG may depend on the complexity of the problem and the training process.

Multi-Agent Reinforcement Learning (MARL): VRPTW can be formulated as a multi-agent problem, where each vehicle acts as an agent making routing decisions. MARL algorithms, such as Independent Q-Learning or Deep Multi-Agent Reinforcement Learning, have been employed to optimize vehicle routing decisions. The run time and optimization level of MARL models can vary based on the number of vehicles, problem complexity, and coordination among agents.

It's important to note that RL models often require significant computational resources and extensive training to achieve good results. The run time and optimization level can be influenced by various factors, including the problem size, the RL model architecture, hyperparameter settings, and the quality of the reward function. Experimentation and tuning are essential to find the best-performing RL model for a specific VRPTW problem instance.

Building RL models

The data we are working with for this project is a sample dataset consisting of route instances with the following noteworthy features:

a. Type

b. Dimension

c. Edge Weight Type

d. Vehicles

e. Edge Weight Format

f. Capacity

g. Edge Weight Section

h. Node Coordinate Section

i. Demand Section

j. Depot Section

k. Service Time Section

l. Time Window Section

Each instance here represents a single route mapped to a set of location coordinates with weights representing cost in terms of time & distance. In addition to optimizing on this cost, we need these optimizations to take place within the bounds of time window constraints according to each customer.

The data in this case was provided as an anonymous sample of supermarket delivery data. To import the data we will be using the Python library Pandas. This library will help us to import the data into a dataframe, which will make it easier to work with. Once the data is in a workable format we can easily check for null values and outliers such as routes that do not conform to the allocated time constraints. Once the data is cleaned and verified, this would be a good time to split the data into our testing and training datasets.

a) Build Agent

In this context, the agent is synonymous with the solver for our program. It will take steps to explore the environment and learn about it in order to inform itself about the most optimal route to achieving its destination.

b) Build Environment

In this step we will write a class object that defines the environment using the OpenAI gym interface. OpenAI gym is a library that provides environments for reinforcement learning and other machine learning algorithms to train against. These gym environments are equipped with methods that are used to initialize our environment and return observations to our agent. The two critical functions for our environment are as follows:

1. Reset - initializes the environment and returns an observation, also provides extra information that may be utilized by the solver.
2. Step - accepts an action and returns an observation, a reward, and a flag indicating whether the environment is complete and may also provide additional information that may be utilized by the solver.

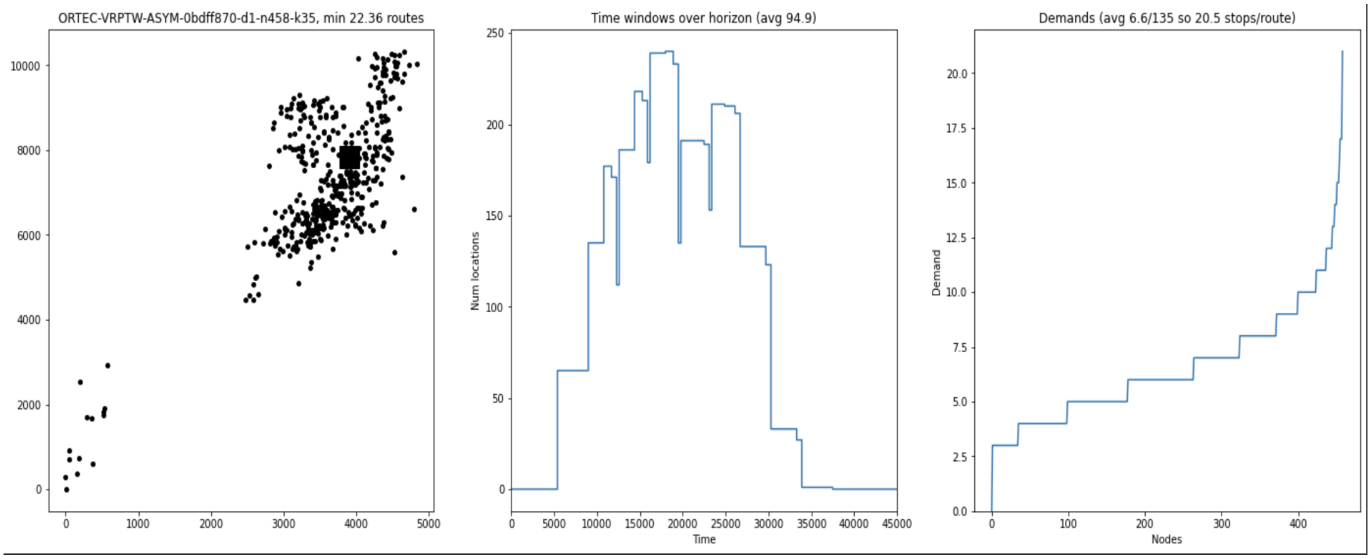
c) Build Reinforcement Learning Model

For purposes of reinforcement learning we will write a function to initialize an environment, deploying the agent and then tasking the agent with making selections about which node to travel to next. These decision points which should happen at every node are referred to as episodes or epochs. These epochs then represent stages that are the building blocks in the agent's travel toward solving the VRP. At each epoch, the agent must decide on an active state at which point it connects with an adjacent node and receives some consequence in the form of a cost.

The goal of our solver is to learn from multiple route instances about which instances result in the most optimal ways to solve their routes. By having the environment function return additional information to the agent, the agent can use that information in its decision-making when determining what factors might indicate a good choice for optimization when the model is generated from the training instance data.

Data Analysis and Results

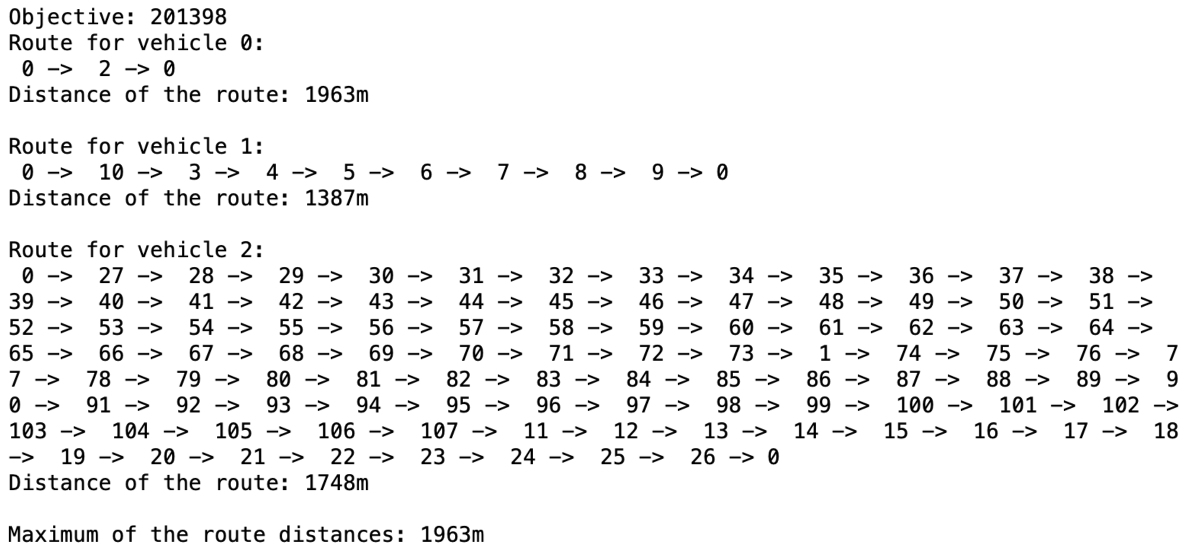
Descriptive analytics and results:



The descriptive report showcases the properties of the data instance for this project including the coordination point records, the demands of the customers, the capacity of the vehicles, time window, service time for each delivery as well as the duration matrix. The first graph, it visualizes the nodes in a clear individual route to which they belong. The second graph plots the number of locations/customers that can be delivered at any time during the day with an average time of 94.9 minutes. The last chart illustrates the demand property of the data for the nodes recorded with an average of 6.6/1.35 so about 20.5 stops per route.

Google OR tools Result:

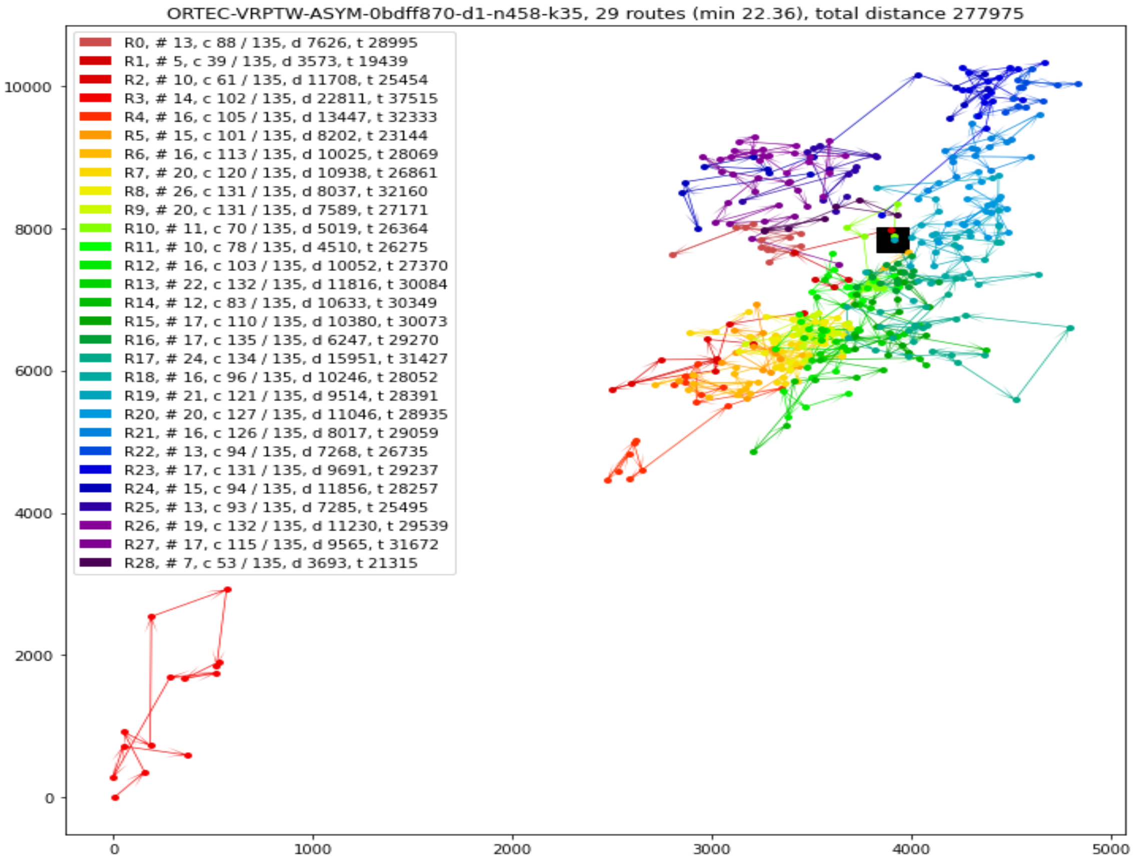
It is worth noting that the Google OR tools package does not fall under the scope of Reinforcement Learning. However, its results give a general evaluation of how different algorithms can be used to address the issue. Looking at the result, Google OR tools did yield a better objective on distance 201,398 compared to that of HGS. However, the algorithm allocation services to the vehicle is not ideal and it did not take into account the capacity of each vehicle.



```
Objective: 201398
Route for vehicle 0:
 0 ->  2 -> 0
Distance of the route: 1963m

Route for vehicle 1:
 0 ->  10 ->  3 ->  4 ->  5 ->  6 ->  7 ->  8 ->  9 -> 0
Distance of the route: 1387m

Route for vehicle 2:
 0 ->  27 ->  28 ->  29 ->  30 ->  31 ->  32 ->  33 ->  34 ->  35 ->  36 ->  37 ->  38 ->
39 ->  40 ->  41 ->  42 ->  43 ->  44 ->  45 ->  46 ->  47 ->  48 ->  49 ->  50 ->  51 ->
52 ->  53 ->  54 ->  55 ->  56 ->  57 ->  58 ->  59 ->  60 ->  61 ->  62 ->  63 ->  64 ->
65 ->  66 ->  67 ->  68 ->  69 ->  70 ->  71 ->  72 ->  73 ->  1 ->  74 ->  75 ->  76 ->  7
7 ->  78 ->  79 ->  80 ->  81 ->  82 ->  83 ->  84 ->  85 ->  86 ->  87 ->  88 ->  89 ->  9
0 ->  91 ->  92 ->  93 ->  94 ->  95 ->  96 ->  97 ->  98 ->  99 ->  100 ->  101 ->  102 ->
103 ->  104 ->  105 ->  106 ->  107 ->  11 ->  12 ->  13 ->  14 ->  15 ->  16 ->  17 ->  18
->  19 ->  20 ->  21 ->  22 ->  23 ->  24 ->  25 ->  26 -> 0
Distance of the route: 1748m

Maximum of the route distances: 1963m
```

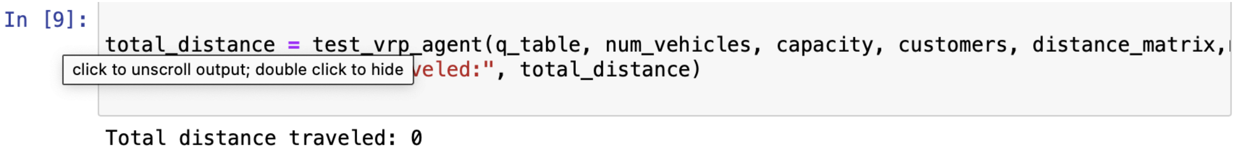
HGS Result:

To best satisfy the total 200-900 demand distributed over the nodes within the desired time window, HGS yielded the optimal results containing 29 routes with the estimated total distance traveled at 277,975. The graph below showcases the visualization of the solution within the assigned territory. The dark square in the center is the depot, which is the starting point where vehicles load inventory and start their delivery services. HGS yields a more expensive solution. However, the algorithm is more popular for solving Vehicle Routing Problems.



DQN result:

With DQN, the goal was to initialize a VRP environment class that simulates the step() function. Then trained VRP agents use Q-learning algorithms to learn the optimal solution. Within the process, epsilon value was adjusted strategically to balance between exploration and exploitation for best result. To explain this further, when the model exploits the reward knowledge, it will only maintain the performance at the highest value of reward per step without exploring other options. However, with exploration, the model keeps on testing the water for higher reward values. However, when testing it on a new instance, DQN did not yield a desirable solution.

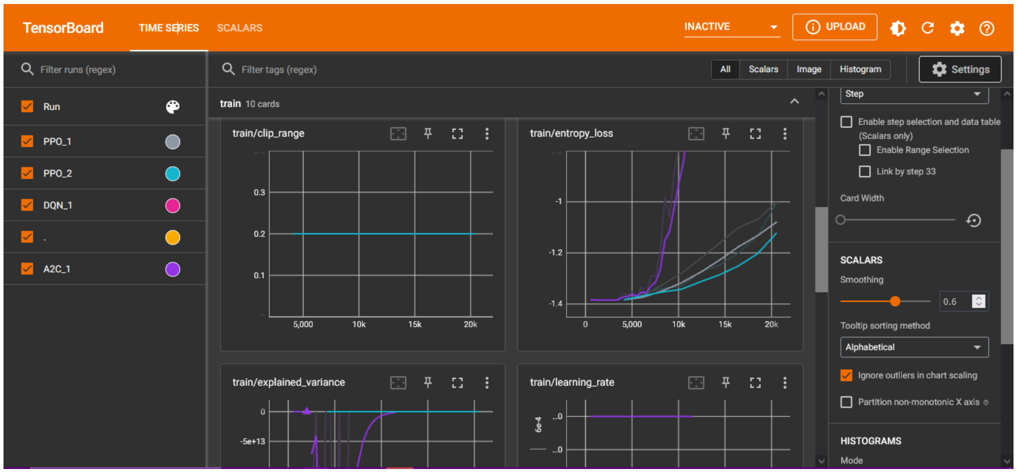


```
In [9]:
total_distance = test_vrp_agent(q_table, num_vehicles, capacity, customers, distance_matrix,
click to unscroll output; double click to hide veled:", total_distance)

Total distance traveled: 0
```

Policy Gradient RL Model Performance Results:

For our final analysis we used tensorboard and the python libraries stablebaselines3 and gymnasium to compare model performance in terms of both training and evaluations of performance, accuracy and efficiency. The following charts provide a comparison of our entropy loss by model type. This can be used to measure accuracy where high entropy loss is indicative of low accuracy and vice versa. Based on these charts, PPO seems to be the most accurate model. However for route optimization, it is more accurate to start with the definition of entropy which is a measure of "how random the decisions of the model are. [Entropy] should slowly decrease during a successful training process. If it decreases too quickly the beta hyperparameter should be increased." (AurelianTactics, 2018)

If we consider the entropy loss, which is the elimination of entropy, the inverse should be true which shows itself to be the case below. The second point about sharp decreases in entropy is even reflected inversely as entropy loss rapidly increases in comparison with a more gradual PPO plot, indicative of the aforementioned beta hyperparameter issue. In simple words, this means that A2C could be prone to overcorrecting or converging too soon.



☑ Run
☑ PPO_1
☑ PPO_2
☑ DQN_1
☑ .
☑ A2C_1

Step
□ Enable step selection and data table (Scalars only)
□ Enable Range Selection
□ Link by step 33

Card Width

SCALARS

Smoothing 0.6

Tooltip sorting method: Alphabetical

☑ Ignore outliers in chart scaling

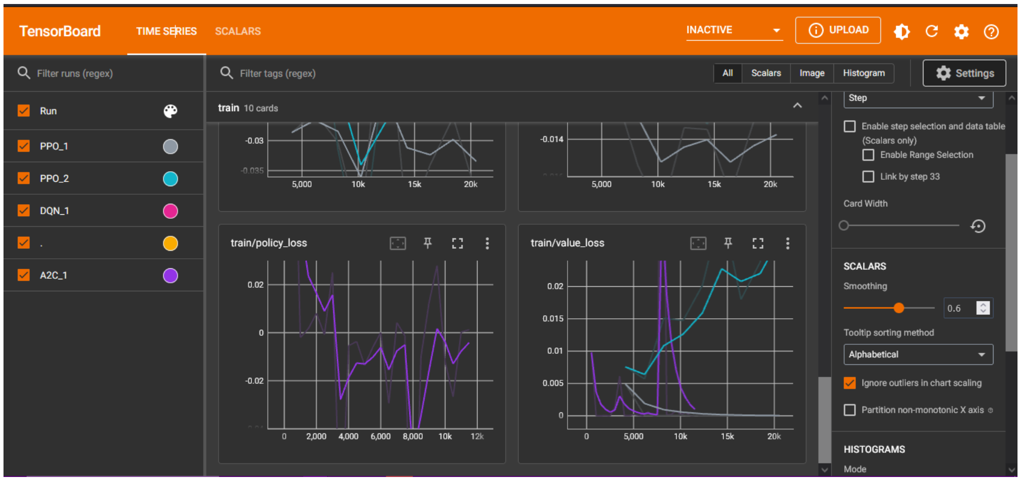
□ Partition non-monotonic X axis

HISTOGRAMS

Mode

We can obtain additional insights from other visuals such as our value loss chart, which showcases the value function loss for on-policy algorithms. In this case that applies to DQN, A2C and PPO. The results suggest that for PPO, value loss rapidly increases with training. The value loss, according to a recent article, “correlates with how well the model is able to predict the value of each state. This should increase while the agent is learning and then decrease once the reward stabilizes. These values will increase as the reward increases, and then should decrease once reward becomes stable.” (AurelianTactics, 2018)

Our model reflects this pattern which suggests it is working aas expected.



☑ Run
☑ PPO_1
☑ PPO_2
☑ DQN_1
☑ .
☑ A2C_1

Step

☐ Enable step selection and data table (Scalars only)

☐ Enable Range Selection

☐ Link by step 33

Card Width

SCALARS

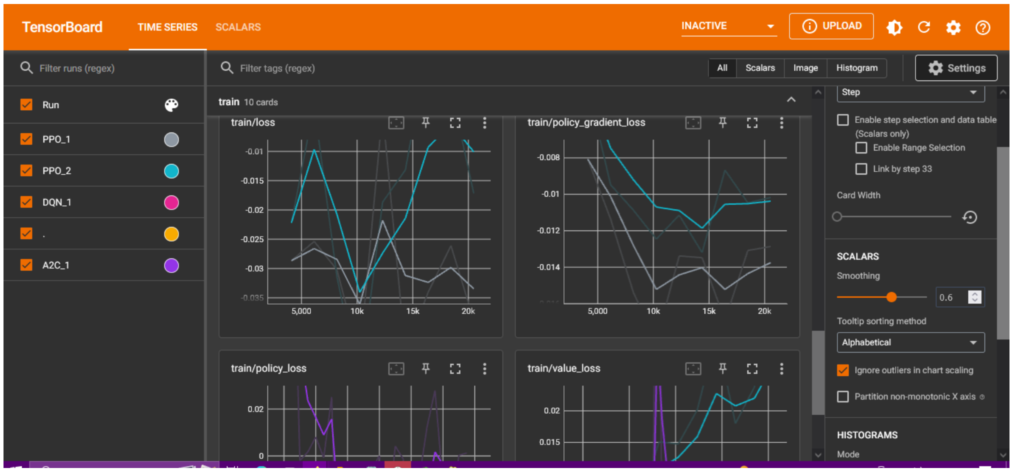
Smoothing 0.6

Tooltip sorting method: Alphabetical

☑ Ignore outliers in chart scaling

☐ Partition non-monotonic X axis

Finally our policy gradient loss chart is a chart that is particularly applicable to the policy gradient family of RL algorithms, in our case, PPO. This indicates that the intensity of the learning starts high and rapidly stabilizes. This is premised on the idea that the policy attempts to guide learning by grading the agent’s actions based on assumed probability estimates of success, taking greater action where the assumed probability estimate of success is high. Other algorithms for whom this applies includes A3C, an improved version of A2C. (Ecoffet, 2018) In its current form, A2C can take similar performance measurements with the policy loss function illustrated on the left-hand side of the chart above.



☑ Run
☑ PPO_1 ●
☑ PPO_2 ●
☑ DQN_1 ●
☑ . ●
☑ A2C_1 ●

Step

□ Enable step selection and data table (Scalars only)
□ Enable Range Selection
□ Link by step 33

Smoothing 0.6

Tooltip sorting method: Alphabetical

☑ Ignore outliers in chart scaling

□ Partition non-monotonic X axis

Conclusion and Recommendations

Summary of findings:

In summary, solving Vehicle Routing Problem using Reinforcement Learning is an interesting topic with a high level of practicality in today's ever changing delivery and transportation industry. With the application of RL, the team have had a chance to dive deep not only in theoretical aspects of RL machine learning as well as its sub categories (DQN, PPO, A2C) but also the real life business knowledge from researching.

With the project, the team decided to solve the topic of Vehicle Routing with various techniques of RL and discovered that some models yield better results than the other. The team also incorporated Google OR Tools package in comparison to RL.

To best satisfy the total 200-900 demand distributed over the nodes within the desired time window, HGS yielded the optimal results containing 29 routes with the estimated total distance traveled at 277,975 while Google OR tools did yield a better objective on distance 201,398 compared to that of HGS. However, the algorithm allocation services to the vehicle is not ideal and it did not take into account the capacity of each vehicle.

Recommendations for Analytic Objective #1 (Minimizing the total distance traveled by the vehicles)

***Use OR-Tools to develop a working prototype without RL*-** From the previous interpretation of the method, Google OR does not fall under the scope of reinforcement learning. However, it is worth taking into consideration when it comes to the solution of the Vehicle Routing Problem. With the result of applying the method to the assigned data set. Google OR tools did yield a better objective on distance 201,398 compared to that of HGS. However, the algorithm allocation services to the vehicle is not ideal and it did not take into account the capacity of each vehicle. Therefore, one recommendation is to include the variable of vehicles' capacity. The pictures below explain the concept and result in the process.

Train an agent using a PPO RL model for general-purpose routing- Of the methods tested in our study, for minimizing total distance traveled we recommend deploying a PPO agent for general route optimization. The efficiency by which it is able to travel toward a goal while avoiding obstacles (as demonstrated in our Frozen Lake experiments) makes PPO a good choice for optimizing distance traveled.

Train an agent using an A2C RL model for more efficient local alternatives- A2C may be more efficient at finding optimal solutions by distance in frequently visited areas of known dimensions. Based on the entropy loss graph, we can conclude that the A2C agent is rapidly converging on a solution, but possibly failing to generalize as well as the PPO model. This makes A2C or PPO a viable choice for efficiency gains on datasets with less road variation and PPO the preferred method for more generalized and robust solutions.

Use HGS to benchmark the routes generated by our RL models- HGS serves as a valuable tool for evaluating the performance of RL models in route generation by comparing their results with human-generated routes. This technique helps to understand the capabilities and limitations of RL models and drives advancements in optimizing route planning algorithms.

Visualize your route sequence using Network Graphs- Visualizing the VRPTW route sequence using network graphs provides a representation of an optimized solution, allowing for easy interpretation and analysis. By minimizing the total distance traveled by vehicles we were able to achieve cost savings, improved efficiency and better customer service. One recommendation is to consider specific requirements and constraints of VRPTW problem instances when implementing the visuals and optimization techniques.

Recommendations for Analytic Objective #2 (Minimizing the total travel time for the vehicles)

Minimize the service time between nodes with an objective function that minimizes the SERVICE_TIME_SECTION column in our dataset- For optimizing total travel time, we can still minimize the cost of going from one node to the next. This cost would need to be measured in units of time instead of distance. We could use the SERVICE_TIME_SECTION column in our data to minimize our objective function and obtain results optimized for time instead of distance.

Recommendations for Analytic Objective #3 (Delivering goods in their preferred time windows)

Integrate time window data as dynamic variables- During our input embedding, we can add the remaining time and remaining packages to be delivered as dynamic variables that are updated with every node traversal. Traversal to the depot node indicates the end of a route and should only be done once a day to preserve route data integrity and optimize drive time.

Optimize for num vehicles, then distance traveled- In order to address the question of time windows, we suggest deriving a remaining time column and optimizing our model in two steps: first, minimize the number of vehicles, then minimize distance traveled (2020, Poullet).

Works Cited:

Smith, J. (2021, December 20). Shipping and Logistics Costs Are Expected to Keep Rising in 2022. Retrieved December 16, 2022, from

https://www.wsj.com/articles/shipping-and-logistics-costs-are-expected-to-keep-rising-in-2022-11639918804

Asghari, M., & Mirzapour Al-e-hashem, S. M. J. (2021). Green vehicle routing problem: A state-of-the-art review. *International Journal of Production Economics*, *231*, 107899. https://doi.org/10.1016/j.ijpe.2020.107899

Abdirad, Krishnan, Gupta; "A Two-Stage Metaheuristic Algorithm for the Dynamic Vehicle Routing Problem in Industry 4.0 approach",form

https://arxiv.org/ftp/arxiv/papers/2008/2008.04355.pdf

Rishi Wadhwa GB. (2022, March 17) The Evolution of India's data privacy regime in 2021. The evolution of India's data privacy regime in 2021.

https://iapp.org/news/a/the-evolution-of-indias-data-privacy-regime-in-2021/.

van Duin, Annika Sponselee (2021, April 1) GDPR in the Public Sector: Cyber Security: Privacy. Deloitte Netherlands.

https://www2.deloitte.com/nl/nl/pages/risk/articles/cyber-security-privacy-gdpr-in-the-public-sector.html.

Recio M. (2020, May 6) GDPR MATCHUP: Mexico's Federal Data Protection Law held by private parties and its regulations. GDPR matchup: Mexico's Federal Data Protection Law Held by Private Parties and its Regulations.

https://iapp.org/news/a/gdpr-matchup-mexicos-federal-data-protection-law-held-by-private-parties-and-its-regulations/.

Liu V, Ke X, Luo Y, Yu Z. (2021, August 24) Analyzing China's PIPL and how it compares to the EU's GDPR. https://iapp.org/news/a/analyzing-chinas-pipl-and-how-it-compares-to-the-eus-gdpr/

DataGrail. (2022, December 10) The ADPPA explained: Everything you need to know. DataGrail. https://www.datagrail.io/blog/data-privacy/the-adppa-explained-everything-you-need-to-know/

Smith, Jennifer. (2020, Jan 16) Wall Street Journal (Online); New York, N.Y. [New York, N.Y]. https://www.proquest.com/abicomplete/docview/2339343866/90BF9D2F8F3943D6PQ/18?accountid=10357

Govender, Tania. (2022, Oct 31) Bizcommunity.com; Cape Town Cape Town: SyndiGate Media Inc. https://www.proquest.com/abicomplete/docview/2730387443/FB2DBC6738D74C87PQ/20?accountid=10357

Zhao Jiuxia ,Mao Minjia, Zhao Xi (2021,November) ; A Hybrid of Deep Reinforcement Learning and Local Search for the Vehicle Routing Problems.

Mladenović, Hansen (1997) Variable neighborhood search. Computers & Operations Research, 24(11), 1097-1100. https://doi.org/10.1016/S0305-0548(97)00031-2

Ecoffet, A. L. (2018, October 5). An Intuitive Explaination of Policy Gradient. Retrieved July 15, 2023, from https://towardsdatascience.com/an-intuitive-explanation-of-policy-gradient-part-1-reinforce-aa4392cbfd3c.

aureliantactics. (2018, December 13). [web log]. Retrieved July 15, 2023, from https://medium.com/aureliantactics/understanding-ppo-plots-in-tensorboard-cbc3199b9ba2

Poullet, J. (2020). *Leveraging machine learning to solve the vehicle routing problem with time windows* (thesis). Massachusetts Institute of Technology, Cambridge.https://hdl.handle.net/1721.1/127285

Coding Reference link:

https://colab.research.google.com/drive/1S4r_6lMfVwtuL73CJm09TNfRCfGuLZcY?usp=sharing#scrollTo=i067jqXGUFqt