Final project

CIVIL-557 Decision Aid Methodologies in Transportation

Nour Dougui, Tim Hillel, Selin Ataç, Stefano Bortolomiol, Marija Kukic, Nicola Ortelli, Janody Pougala, Negar Rezvany

25 May 2021

The *Decision Aid Methodologies in Transportation* final project consists of two separate assignments, each covering one of the two fields presented in the course: operations research and machine learning. The two assignments are given in §1-2 respectively. Details of the submissions, deadlines, and final exam are then given in §3.

1 Mode choice prediction for non-car owning households in the USA *(40 points)*

For this project you will be analysing and predicting the mode choice behaviour for individuals from non-car owning households across the USA. You will be using data from the National Household Travel Survey (NHTS), carried out between March 2016 and May 2017 by the Federal Highway Administration (FHWA) (U.S. Department of Transportation, Federal Highway Administration 2017).

You should approach this project as though working as a machine-learning consultant for a client, and will be expected to explain the modelling decisions made.

1.1 Problem description

The client for the project is the US Federal Highway Administration (FHWA). The FHWA wish to better understand potential impacts that decreasing car ownership would have on the modal split in the USA. Specifically, they wish to model the mode choice behaviour of non-car owning households. Following on from your work analysing mode choice behaviour in Lausanne (for *Transports publics de la région lausannoise*), and London (for *Transport for London*), you have been asked to develop an individual mode choice model the target population. The client wishes to explore machine learning approaches for this task.

Your overall goal for this project is to build the highest performing mode choice prediction model possible for non-car owning households in the USA, using the supplied dataset. As the model will be used for simulation, the client has requested for the model to output choice probabilities for each mode, out of the eight modes of *walking, cycling, rail public transport, bus public transport, private vehicle driver, private vehicle passenger, taxi,* and *other modes.* At this stage in the development, the client is not concerned about computational complexity or operating times and instead has asked you to focus on achieving the highest possible predictive power.

The client has provided anonymised trip data from the NHTS with which to build the model. Specifically, they have provided the trip records for all houses in the full NHTS which do not own a car. The client has provided the data in its raw format (including continuous, binary, and categorical features as well as missing values), and has asked you to carry out and investigate data pre-processing. Note some categorical features are numeric, whilst others are text. A data dictionary has been provided which details all of the features in the dataset, and specifies the possible values for each feature.

The client is aware of a recent machine learning study which makes use of the full NHTS dataset (Wang et al. 2021), alongside two other datasets. They have therefore asked you to review and critique the methodology used in this study, in addition to the requested experimental work.

1.2 Tasks

The project is broken into the following tasks:

1. Review and critique (within the context of the material introduced in the course) the experimental methodology used in the comparative study of (Wang et al. 2021). Provide recommended solutions for any issues you identify. Your answer should consider:
   - model hyperparameter selection,
   - validation schemes used,
   - sampling for model validation,
   - model performance metrics, and
   - data imbalance.
2. Develop the highest performing mode-choice classifier you can using the supplied data (Kaggle competition). The model will be evaluated on the *negative cross-entropy loss.* Note, you are *not limited* to methods taught in the course, and are actively encouraged to try new algorithms and approaches. You will need to decide and implement:
   - data pre-processing,
   - the model evaluation scheme,
   - hyperparameter selection, and
   - model comparison and selection.

3. Prepare a concise modelling report which summarises your experimental methodology. The report should explain in clear language the choices you have made (data pre-processing, models tested, model validation scheme, hyperparameter search, etc). Provide brief motivations for each decision. Additionally, you should report to the client what **extra data or variables** could be collected or added to the dataset that you feel could improve the quality of the mode-choice classifier.

1.3 Data

The provided dataset is adapted from the US National Household Travel Survey (U.S. Department of Transportation, Federal Highway Administration 2017), conducted between March 2016 and May 2017. The dataset comprises of one-day travel diaries, provided by all members of the household. The data is therefore hierarchical.

The data has been processed to include only trips with known mode-choice from households which do not own a car or private vehicle. It contains 23 180 trips from 4719 households.

The data contains both categorical and numerical features, and will need to be preprocessed. A dataset dictionary (`nhts_dictionary.xslx`) is provided, that details the different features in the data, and the different values they can take. The choice of mode for each trip is made between eight main modes: walking (*WALK*), cycling (*CYCLE*), rail public transport (*RAIL*), bus public transport (*BUS*), private vehicle driver (*DRIVE*), private vehicle passenger (*PASSENGER*), private hire vehicle (*TAXI*), and other modes (*OTHER*).

The data is provided as two sets. The train/validate data (`nhts_train_validate.csv`) contains 70% of trips, and includes the choice label. This data is available to use as you see fit for model development, hyperparameter selection, model selection etc. The test dataset (`nhts_test.csv`) contains the remaining 30% of trips for model evaluation, and **does not** include the choice label. Once you have finalised your model specification, you will need to generate the choice probabilities for the trips in the test set, and submit them to the Kaggle competition page. An example submission file (`random_example.csv`) is provided to illustrate the format required for the Kaggle submissions. Note: you will need to select up to two models for final submission. The datasets are additionally available on the Kaggle competition page.

In addition to the data dictionary, data files, and example submission file, we have also provided a user guide (`NHTS2017_UsersGuide.pdf`) for the data, which explains how the full NHTS dataset was collected and processed.

1.4 Kaggle competition

The evaluation of your models will be performed publicly using a Kaggle competition. You will need to create a Kaggle account to register. You can either create a single account for your team, or it is possible for each team member to create an account, and then register as being part of the same team for the competition.

Once you have created an account, you can register for the competition at the following URL: `https://www.kaggle.com/t/5723a0b5f7774882a2aa7727b81ee69d`. Please see the Kaggle competition page for further details. Note this is a private (InClass) competition, created for the course, and will not award Kaggle ranking points or count towards tiers. Please do not share this URL with those outside the course.

Submissions are made using a CSV file of the model output on the test dataset (see Kaggle page). There is a public leaderboard, visible to all members of the course, which is kept up-to-date as you submit your model predictions.

You are able to submit two new files for evaluation on the public leaderboard per day. At the end of the project (i.e. before midnight on the 25th June 2021) you will need to select up to two files to submit for the final leaderboard. The public leaderboard performance is calculated on different rows of the test dataset to the final leaderboard. As such, you are encouraged not to try to optimise your models to score highly on the public leaderboard, but instead validate internally on the train/validate data, and select the best performing models from here.

Note, you will also need to submit separately by email *all* of your code (in the form of a Jupyter notebook). This code should be clearly commented, and include **all modelling stages** (i.e. data pre-processing, model initialisation, hyerparameter search, training the final model, and generating the results files).

1.5 Deliverables

The submission for the project is in three parts.

1. Send by email before **23:59 on the 25th June 2021**:
   - PDF of containing review of paper by Wang et al. (2021) and modelling report.
   - Jupyter notebook containing all code (i.e. data pre-processing, model initialisation, hyerparameter search, training the final model, and generating the results files). The code should be able to be run directly on another machine, and produce all results submitted to the Kaggle competition. The code should also contain comments or markdown cells explaining the operation of each code block.
2. Submit your results on Kaggle by **23:59 on the 25th June 2021**. Sign-up at the following url `https://www.kaggle.com/t/5723a0b5f7774882a2aa7727b81ee69d`.
3. Send by email before **23:59 on the 1st July 2021** your finalised presentation slides covering the paper review and modelling process for the presentation on 2nd July 2021.

References

U.S. Department of Transportation, Federal Highway Administration (2017). *2017 National Household Travel Survey.* http://nhts.ornl.gov.

Wang, Shenhao et al. (2021). "Comparing Hundreds of Machine Learning Classifiers and Discrete Choice Models in Predicting Travel Behavior: An Empirical Benchmark". In: *arXiv:2102.01130 [cs, econ]*. arXiv: `2102.01130 [cs, econ]`.

2 The multi-compartment vehicle routing problem (MCVRP) *(40 points)*

A multi-compartment vehicle routing problem (MCVRP) is a variant of the classic capacitated vehicle routing problem (CVRP). Recent truck models are equipped with a special device which allows for introducing bulkheads in predefined positions of the loading space such that it can be split into different compartments. These truck models are called *multi-compartment vehicles* (MCVs). They can be used in grocery distribution (e.g., supply of stores with different temperature zones), waste collection (e.g., glass waste) and fuel distribution (e.g., supply of petrol stations).

2.1 Problem description

Our problem is motivated by an application in grocery distribution where a high product variety with particular temperature requirements needs to be managed and an efficient supply chain is essential. Products with similar characteristics and temperature requirements are usually denoted as product segments or simply *segments* in grocery distribution. In the past, only one product segment could be transported within the same vehicle as the temperature could only be set up at one level at a time. However, MCVs are able to split the loading area flexibly into compartments with different temperatures whilst there is no loss in capacity (see Figure 1).

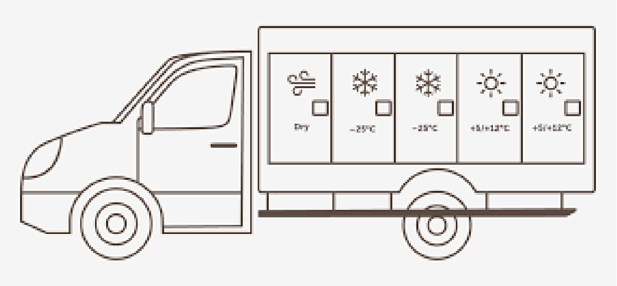

Figure 1: A multi-compartment vehicle with multi-temperature body.

Grocery retailers need to simultaneously manage four to five different temperatures (e.g., frozen, chilled and ambient) across their logistics subsystems. The majority of products are distributed via a Distribution Center (DC). The distinct products can be allocated to the same DC, but separate warehouse zones at different temperatures are required to prevent spoilage. The same reasoning applies to the transportation process, during which the preferred temperature for each product needs to be maintained to guarantee high product quality and to adhere to legal regulations.

Using MCVs poses some new challenges for the planning as the joint distribution of segments influences not only the transportation process but also the upstream and downstream supply chain operations. Indeed, different gates have to be approached by an MCV to collect different segments from distinct DC temperature zones. This leads to an increase in *loading costs* that depend on the number of segments assigned to a vehicle. On the other hand, separate deliveries with a single-compartment vehicle for each segment may be avoided, reducing the number of visits to a store and the total travel time.

In our problem, we consider that splitting the capacity is done according to the following properties:

- The size of each compartment **is not fixed in advance** but can be determined individually for each vehicle.
- The size of the compartments can only be varied **discretely**: the walls separating the compartments from each other can only be introduced in specific predefined positions.
- The number of compartments, into which the capacity of a vehicle is divided, **can be equal** to the number of product segments.

Consequently, in the MCVRP applied to grocery distribution, we not only have to determine the vehicle tours, but we have also to decide **for each vehicle**:

- into how many compartments the capacity should be divided,
- the size of each compartment,
- which product segment should be assigned to each compartment.

2.2 Problem formulation

The MCVRP applied to grocery distribution is defined on a complete, undirected and weighted graph $G = (N, E)$, where $N = \{0, 1, ..., n\}$ is the set of nodes and $E = \{(i, j) : i, j \in N\}$ is the set of edges. Node $i = 0$ represents the DC location. The other nodes represent the stores. Each store $i \in I = N \backslash \{0\}$ is associated with a service time ser_i which correspond to the time spent to unload the products. Each edge $(i, j) \in E$ is associated with a travelling cost c_{ij} and a travelling time t_{ij}.

Due to legal restrictions, retailers usually define a maximum tour duration. Therefore, we assume that each vehicle that depart from the DC must return before time T (the maximum tour duration).

The DC is responsible for the distribution of products from $|S|$ segments. At each store $i \in I$ exists a demand d_{is} of each product segment $s \in S$. The segments must be transported from the DC to the stores without being mixed. Therefore, each segment is transported into a dedicated compartment. Each vehicle can visit a store just **one time**. On the other hand, a store may be visited by **several** vehicles in order to deliver different segments. However, if being delivered, each segment demand must be loaded in total. In other words, a split delivery of a single segment demand **is not allowed**.

For the purpose of transportation, a set V of homogeneous vehicles is available at the DC, each having a total capacity Q. For each vehicle $v \in V$, the total capacity Q can be divided into a limited number k of compartments, $k \leq |S|$, which allows to load different segments while keeping them separated during transportation.

The size of the compartments can be varied discretely in equal step sizes: each compartment size and also the total vehicle capacity Q, is an **integer multiple** of a basic compartment unit size q_{unit}. Let the set of these multiples be denoted by $M = \{1, 2, ..., m_{max}\}$ where $m_{max} = \frac{Q}{q_{unit}}$. Then $q_m = m \times q_{unit}$, $m \in M$, denotes a compartment size.

Example:

Let's consider that the vehicle capacity Q is equal to 100 units and the basic compartment unit size q_{unit} is equal to 10 units. Hence, only compartment sizes of 10, 20, 30,...,100 units can be selected. Accordingly, m_{max} is equal to $\frac{Q}{q_{unit}} = \frac{100}{10} = 10$. Besides, for $m = 4$ for example q_4 is equal to 40 units (see Figure 2).

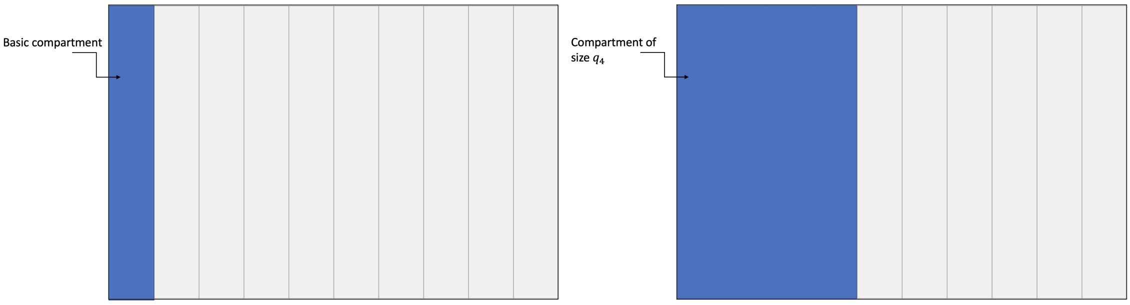

Figure 2: Discrete compartment size.

Let's consider that the retailer needs to manage 3 segments: frozen, chilled and ambient. A possible allocation of the segments in the vehicle is given by Figure 3. The frozen segment is assigned to a compartment of size 30 units. The chilled segment is assigned to a compartment of size 50 units. The ambient segment is assigned to a compartment of size 20 units.

Due to the characteristics of MCVRP applied to grocery distribution that have been explained, loading and unloading costs are also decision relevant costs in addition to the travelling costs. In line with this, ℓ_k represents the loading cost of a vehicle dependent on the number k of product segments transported, and u indicates the unloading cost of a store. The cost of unloading is the same for each store.

The objective of the problem is to minimize the total routing costs, considering travelling, loading and unloading costs, while satisfying the demand of the stores and the maximum tour duration constraint. As a consequence, the problem involves the following partial decisions that are made simultaneously:

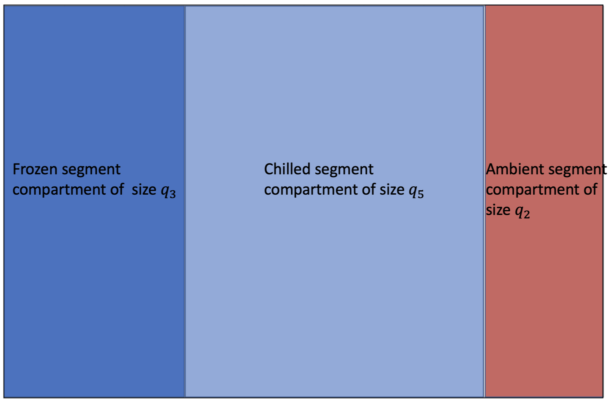
Frozen segment compartment of size q_3
Chilled segment compartment of size q_5
Ambient segment compartment of size q_2

Figure 3: Segments assignment.

- Sequence of store visits, as in every CVRP (this decision specifies the order in which each vehicle should perform the delivery of the orders assigned).
- Assignment of orders to tours/vehicles (this decision determines which segments are delivered by the vehicle).
- Size of each compartment (this decision fixes for each vehicle how its total capacity is split into compartments).

The formal definitions of the sets and the parameters for the MCVRP is as follows.

Sets

$N = \{0, 1, ..., n\}$	Set of stores and distribution center.
$I = \{1, ..., n\}$	Set of stores.
V	Set of vehicles.
S	Set of product segments.
$M = \{1, ..., m_{max}\}$	Set of basic compartements.

Parameters

d_{is}	Demand for each store i of each product segment s.
c_{ij}	Travelling cost from node i to node j.
ℓ_k	Loading cost of a vehicle transporting k segments (loading space split in k compartments).
u	Unloading cost per store.
T	Maximum tour duration.
t_{ij}	Travelling time from node i to node j.
ser_i	Service time for each store i.
Q	Vehicles capacity.
q_{unit}	Basic compartment size.

2.3 Instance description

In this assignment, you are asked to solve the MCVRP on an instance inspired by the grocery distribution problem. We assume that the grocery retailer operates a fleet of 4 homogeneous MCVs, which are parked overnight in the distribution center. The loading capacity of each MCV is 400 kg and q_{unit} the basic compartment size is 20 kg.

The retailer manages 4 product segments: ambient, chilled, deep chilled and frozen. He must serve the demand for each segment of 10 stores. The corresponding demand matrix is given in file *MCVRP_General.txt*.

In order to adhere to legal regulations, each MCV tour must not exceed a maximum duration of 8 hours. The travelling times from the distribution center to each store and from each store to another are given in file *MCVRP_Time.txt*. Besides, the service time at each store is given in the same file.

The travelling costs from the distribution center to each store and from each store to another as well as the loading cost of a vehicle dependent on the number k of product segments transported, are given in file *MCVRP_Cost.txt*. The unloading cost per store is 40 units.

2.4 Exercise questions

1. Write the mathematical model for the MCVRP using the sets and parameter introduced in §2.2 and identifying the appropriate decision variables.

   Hint:

   - you may use the following principal decision variables:
     - $x_{ijv} = 1$: if vehicle $v \in V$ travels from node i to node j with $(i, j) \in E$, 0 otherwise.
     - $del_{isv} = 1$ if segment $s \in S$ is delivered by vehicle $v \in V$ to store $i \in I$, 0 otherwise.
     - $y_{svm} = 1$ if size q_m is selected for segment $s \in S$ in vehicle $v \in V$, 0 otherwise.

   Other auxiliary variables are needed.

   - In order to insure sub-tour elimination, adapt the corresponding constraint seen in the course for the CVRP problem. Recall that in the CVRP problem each store can be visited by **just one** vehicle.

2. Implement the model in OPL and use it to solve the problem for the instance described in §2.3. *Note: The input data of the instance are already prepared in the file* `data_2021.dat`.

   Illustrate the obtained optimal solution through graphs and/or tables.

3. Without solving the problem, analyse how the optimal solution is impacted when we modify the parameters of the problem:

   - **case 1**: more MCVs are available at the Distribution Center.
   - **case 2**: the unloading cost u per store increases drastically.
   - **case 3**: the loading cost ℓ is a constant and does not depend on the number k of segments transported by a given vehicle.

4. You now want the solve the MCVRP applied to grocery distribution on a much larger network. A meta-heuristic such as simulated annealing or variable neighbourhood search is therefore needed to solve this large instance. Thus, you need to design neighbourhood structures.

   To illustrate the neighbourhood structures you are going to design, consider the following initial feasible solution of the instance of the problem described in §2.3:

   The routes of the 4 vehicles are shown in Figure 4: vehicle 1 (red) follows the route $0 \to 5 \to 8 \to 10 \to 3 \to 4 \to 2 \to 0$; vehicle 2 (blue) follows the route $0 \to 6 \to 2 \to 4 \to 1 \to 0$; vehicle 3 (green) follows the route $0 \to 6 \to 0$; vehicle 4 (gray) follows the route $0 \to 7 \to 10 \to 9 \to 3 \to 0$. The product segments delivered by the 4 vehicles to the stores are summarized in Table 1. We give an example to illustrate how to read the table: if we consider store (column) 4, product segments 1 and 3 are delivered by vehicle 1, while product segments 2 and 4 are delivered by vehicle 2.

		Stores									
		1	2	3	4	5	6	7	8	9	10
Segments	1	2	2	1	2	1	2	4	1	4	4
	2	2	2	1	1	1	2	4	1	4	4
	3	2	2	4	2	1	2	4	1	4	1
	4	2	2	1	1	1	3	4	1	4	4

Table 1: Feasible solution: product assignment

   - This solution isn't the optimal solution of the problem. Suggest at least three neighbourhoods that you would design and illustrate the result of applying those neighbourhoods on the initial feasible solution. Explain why you think these neighbourhoods are appropriate to solve the problem.

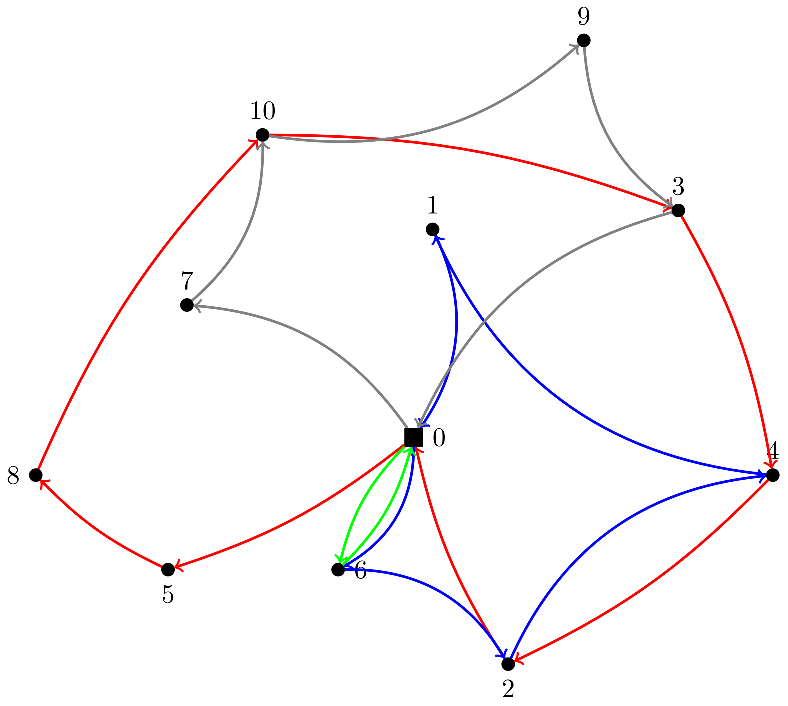

Figure 4: Feasible solution: routes

The MCVRP problem is a special case of the CVRP problem. Make sure to show the differences between the neighbourhood seen in the course for the CVRP problem and the specific neighbourhoods of the MCVRP problem.

The neighbourhoods can be presented in words, graphics, or with pseudo-code. You should not implement the meta-heuristic nor the neighbourhoods.

- When generating a neighbour using the neighbourhood suggested in the previous question, you may end up with an infeasible solution. Suggest repair procedure(s) in order to transform the obtained neighbour in a feasible solution.

  The repair procedure(s) can be presented in words, graphics, or with pseudo-code. You should not implement it.

2.5 Deliverables

The submission for the project is in two parts.

1. Send by email before **23:59 on the 25th June 2021**:
   - OPL code of the developed models.
   - A report containing the model descriptions, the results analysis, the answer to the

questions, at least three neighbourhoods that you designed to solve this problem and the corresponding repair procedure(s).

2. Send by email before **23:59 on the 1st July 2021** your finalised presentation slides containing the problem and model descriptions, results analysis and the neighbourhoods for the presentation on 2nd July 2021.

3 Submission and presentations

Teams should be three people where possible, otherwise two. Please send all emails to `tim.hillel@epfl.ch` and `nourelhouda.dougui@epfl.ch`, **copying in all members of your team**.

3.1 Deadlines

- Please send an email with detail of your team by **23:59 on the 30th May 2021**.
- Please send in a single email by **23:59 on the 25th June 2021** the deliverables for the two projects:
  - For the mode choice prediction project:
    1. the PDF containing the paper review and modelling report.
    2. the Jupyter notebook containing all code.
  - For the MCVRP project:
    1. OPL code of the developed models.
    2. the project report.
- Please submit on Kaggle by **23:59 on the 25th June 2021** your final model submissions for the mode choice prediction task.
- Please send in a single email by **23:59 on the 1st July 2021** your presentation slides for the two project, to be presented the next day.

3.2 Assessment

The final presentations will take place **online** on the **2nd July 2021**. Slots will be 40 minutes per team - with a 20 minute presentation covering the two projects (i.e. 10 minutes per problem), and 20 minutes for questions. All team members are expected to present and respond to questions approximately equally.

The schedule for the final exam schedule will be sent to you on or before the **25th June 2021**.