EPFL
ENAC TRANSP-OR
Dr. Virginie Lurkin
Dr. Nikola Obrenovic

Decision-aid Methodologies in Transportation
Spring 2018

The *Decision-aid Methodologies in Transportation* final project consists of two assignments, covering the two fields presented in the course: operations research and data mining. The two assignments are given in Sections 1 and 2, respectively.

1 Vehicle routing problem with time windows *(70 points)*

1.1 Problem description

One of the best-known operations research problem is the travelling salesman problem (TSP). A number of cities have to be visited by a salesman who must return to the same city where he started. The route has to be constructed in order to minimize the distance to be travelled. The vehicle routing problem (VRP) is the m-TSP where a demand is associated with each city, and vehicle has a certain capacity. If we add a time window to each customer we get the vehicle routing problem with time windows (VRPTW) . In addition to the capacity constraint, a vehicle now has to visit a customer within a certain time frame. The vehicle may arrive before the time window opens but the customer cannot be serviced until the time windows open. It is not allowed to arrive after the time window has closed.

An application of the vehicle routing problem with time windows is the problem of cash collection. Frequently, banks send out vehicles to their branches to collect cash left by depositors. Imagine that you work in the logistics department of a small bank which, on a particular day, needs to collect cash from 15 of its branches. Table 1 lists the 15 branches (c1 to c15). For each branch, it gives the amount of cash to be collected, and the time interval $[\lambda; \mu]$ (in terms of time of the day) within which the cash can be collected. Table 2, moreover, provides the complete distance matrix in kilometers between the bank vault (v0) and the 15 bank branches (c1-c15). Your vehicle fleet consists of 3 identical vehicles, which are assumed to travel at a constant speed of 60 km/h.

Your task is to construct the vehicle tours with the objective of minimizing the total length and duration, with equal weights, of the complete tour schedule.

1.2 PART 1 *(50 points)*

- Provide a mathematical formulation for the VRPTW applied to cash collection with the objective of minimizing the total length and duration, with equal weights, of the complete tour schedule.

The following constraints are imposed:

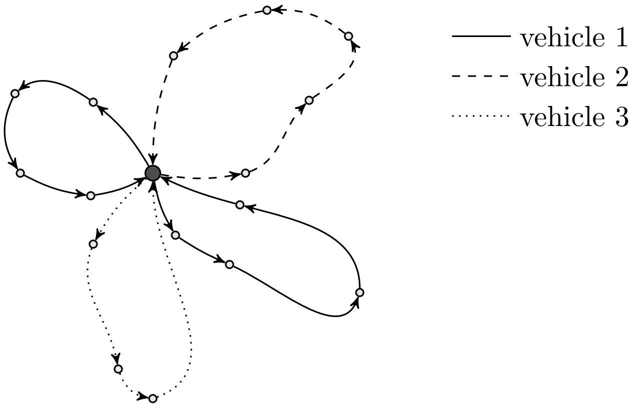
- Each available vehicle, if it is used, leaves the vault and returns to it only once.
- The duration of the collection operation is 30 minutes.
- The amount of cash carried by a vehicle is limited to CHF 600'000.
- Each branch should be visited exactly once by one vehicle.
- The collection at each branch should begin and complete within the specified time interval $[\lambda; \mu]$. If the vehicle arrives at a branch before λ it must wait until λ to start the operation. However, the collection operation cannot continue after μ. To give an example, take branch c2. The operation at this branch can start at 12:00. If the vehicle arrives at 11:00, it should wait for 1 hour before starting collection. Given that the duration of the operation is 30m, if the vehicle arrives between 12:00 and 15:30, it can start collecting immediately, so there will be no waiting. Finally, the branch must be visited by one of the vehicles at 15:30 at the latest so that the service does not continue after 16:00.
- All tours should start and finish at the bank vault, denoted by v0 in Table 2.
- All tours should start and finish in the interval [8:00, 18:00]. In other words, all used vehicles can leave the vault at 8:00 at the earliest and should return to the vault at 18:00 at the latest.

- Test your model with the data provided and analyze the results.

***Hint**: (1) You may want to duplicate the bank vault into "origin vault" and "destination vault" for the sake of modeling. They will have a distance of 0 km between each other, the same distance to, and the same distance from any other point. (2) Your model should determine how many (maybe all, maybe not all) of the available vehicles should be used.*

1.3 PART 2 *(20 points)*

The assumption that each available vehicle, if it is used, leaves the vault and returns to it once may lead to situations where the available fleet is insufficient to service all demands. To overcome this problem, we can introduce a **multi-trip** collection strategy, where in each **trip** the vehicle starts from the vault, collects cash from some branches and returns to the vault to offload the collected cash. A vehicle **tour** may be composed of one or more trips. This is illustrated in the figure below, where vehicle 1 executes a tour composed of 2 trips, and vehicles 2 and 3 execute tours composed of 1 trip each.



- Provide a mathematical formulation for the multi-trip VRPTW applied to cash collection with the objective of minimizing the total length and duration, with equal weights, of the complete tour schedule.

  You need to collect cash from 15 bank branches listed as c1 to c15 in Table 1. For each branch, Table 1 gives the amount of cash to be collected, the duration of the collection operation (in minutes), and the time interval $[\lambda, \mu]$ (in terms of time of the day) within which the cash can be collected. Table 2, moreover, provides the complete distance matrix in kilometers between the bank vault (v0) and the 15 bank branches (c1-c15). Your vehicle fleet consists of 3 identical vehicles, which are assumed to travel at a constant speed of 60 km/h.

  The following constraints are imposed:

  - The duration of the collection operation is 20 minutes.
  - The amount of cash carried by a vehicle is limited to CHF 400'000.
  - Each branch should be visited exactly once by one vehicle.
  - The collection at each branch should begin and complete within the specified time interval $[\lambda, \mu]$. If the vehicle arrives at a branch before λ it must wait until λ to start the operation. However, the collection operation cannot continue after μ.
  - All tours should start and finish at the bank vault, denoted by v0 in Table 2.
  - All tours should start and finish in the interval [8:00, 18:00].
  - Every offloading operation at the vault takes 30 min and should complete in the interval [8:00, 18:00].

  Hint: Use the vault replication hint from the first part and extend the idea to model the potential intermediate visits to the vault.

- The solution time of an (mixed) integer (linear) program can sometimes be improved by so-called valid inequalities. These are constraints that restrict the search away from exploration of undesired regions. However, they should not eliminate any feasible solutions.

  - One set of valid inequalities appropriate for your problem concerns tour durations. The way tour durations are minimized in your objective function can lead to slow convergence. Propose a valid bound from below for the tour durations. In other words, formulate constraints about the minimum tour durations.

  - Another frequently encountered class of valid inequalities, usually referred to as symmetry breaking constraints, may also help in case of identical vehicles. To illustrate the concept, take the following example. If you have 2 tours and 2 identical vehicles, the tours can be assigned to the vehicles in 2 different ways, both producing the same value of the objective function. The number of permutations grows if you have more identical vehicles. The branch-and-bound method is likely to evaluate many such permutations unnecessarily. Your task is to propose constraints that limit the evaluated permutations to a minimum. ***Hint:*** *Tours have some attributes, for example total collected cash value or total distance, among others. You can define constraints that assign the tour with the highest attribute value to the first unused vehicle, the tour with the second highest attribute value to the second unused vehicle, etc.*

- Test your model with and without the valid inequalities you implemented. Do you observe any difference in solution time? Which valid inequality seems to be the most beneficial? Do you get the same value of the objective function with and without the valid inequalities you implemented? Do you obtain the same solution as in PART 1? Discuss.

2 Flight Delay Analysis *(10 points)*

2.1 Problem description

Due to the high demand for commercial air travel, flight delays and cancellations happen frequently. According to the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS), approximately 20% of commercial flights arrived late over the past ten years. Although all carriers have troubles with being on time, some of them tend to be late more often than the others. For example, in May 2017, Virgin America flights were on time in just 58.74% cases, in contrast to 82.75% of the flights operated by Delta Air Lines. The 2017 yearly averages for these two companies were 70.00% and 85.40%, respectively.

Bad weather is quite often a cause of flight delays. Since we cannot forecast weather conditions far ahead with high certainty, we also cannot predict flight delays caused by this. On the other hand, some earlier analysis of flight data, recorded by BTS, suggest that several factors, other than weather conditions, can be good indicators of flight delays. Hence, we would like to identify those factors and develop data mining models for prediction of the flight punctuality, by using the BTS dataset. The ultimate usage of the developed models would be to help travelers decide which ticket to buy. Since the purchases are often made far in advance, we cannot rely on the weather condition forecasts and the flight delay predictions should be made without reference to them.

2.2 Dataset

Since October 1987, the BTS tracks the on-time performance of domestic US flights operated by large air carriers. For each flight, BTS records data such as the carrier, origin, destination, scheduled and actual, departure and arrival times, flight distance, etc. For this project, you will use a subset of this data, covering the year 2008. This data set is provided in the *CSV* format at *Kaggle.com*. The detailed description of the attributes is available in the same website.

2.3 Task

- Download the described data about the flights from *Kaggle.com* using this link: *https://www.kaggle.com/giovamata/airlinedelaycauses*. The data file of interest is *DelayedFlights.csv*.
- Get to know the data and understand the available features.
- Decide how to divide the data on the training and test datasets.
- By using the data mining algorithms, seen in the lectures and exercises, create the following classifiers:
  - A binary classifier that will classify all future flights as expected to arrive late or not.
  - A classifier that will classify the future flights into one of the four categories: 1-on time arrival; 2-slight arrival delay (max. 15 minutes); 3-moderate arrival delay (15-60 minutes); 4-substantial arrival delay (over 1 hour).
- For the first classifier, select and try out two algorithms to create it. On the other hand, for the second classifier, it will be enough to use only one algorithm of your choice.
- Also, for each classifier and each used algorithm, test at least 3 different sets of predictor features. Justify the selection of features.
- Assess all created classifiers with the confusion matrix and measures of precision and recall, and suggest the best solution for each of the required classifiers.

3 Technical Information

- For the final exam, you will work and present as a group.
- Please send your presentation and your code by email to `virginie.lurkin@epfl.ch` on June 22, 2018, i.e. one week before the final exam. It will be considered as finalized and will be loaded for you on the computer in the exam room.
- The presentation must be in English. It must be self-contained, i.e. you should not assume that all examiners know the details of the assigned problems. Also, try to make it as interesting as possible.
- Split the presentation equally among the group members. You will have 20 minutes for presentation, followed by 15-20 minutes of questions from the examiners where each student will be individually assessed on any material covered in the course.
- All the concepts needed to answer the given problem have been covered during lectures and/or tutorial sessions.

Bonne chance.

Data Appendix

Table 1: Service specifications for branches

Branch	Amount of cash	λ (o'clock)	μ (o'clock)
	to collect		
c1	CHF 110'000	12:00	16:00
c2	CHF 150'000	12:00	16:00
c3	CHF 130'000	12:00	16:00
c4	CHF 80'000	14:00	16:00
c5	CHF 90'000	9:00	12:00
c6	CHF 90'000	9:00	12:00
c7	CHF 80'000	9:00	12:00
c8	CHF 80'000	11:00	14:00
c9	CHF 100'000	11:00	14:00
c10	CHF 110'000	11:00	14:00
c11	CHF 100'000	14:00	16:00
c12	CHF 50'000	14:00	16:00
c13	CHF 90'000	15:00	18:00
c14	CHF 100'000	15:00	18:00
c15	CHF 100'000	16:00	18:00

Table 2: Distance matrix (in km)

	v0	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15
v0	0	31	77	67	18	14	51	73	73	27	79	66	20	47	54	11
c1	31	0	109	94	32	18	81	102	105	55	87	89	46	70	76	21
c2	77	109	0	39	86	91	31	21	11	64	99	73	70	56	59	87
c3	67	94	39	0	83	77	23	19	46	69	61	96	71	26	24	73
c4	18	32	86	83	0	22	64	86	80	25	97	57	17	65	72	23
c5	14	18	91	77	22	0	63	85	87	40	79	77	32	54	61	4
c6	51	81	31	23	64	63	0	22	32	47	73	73	50	27	32	59
c7	73	102	21	19	86	85	22	0	30	67	80	86	72	40	41	81
c8	73	105	11	46	80	87	32	30	0	56	104	63	64	59	63	84
c9	27	55	64	69	25	40	47	67	56	0	99	39	9	57	65	38
c10	79	87	99	61	97	79	73	80	104	99	0	137	96	47	41	76
c11	66	89	73	96	57	77	73	86	63	39	137	0	46	92	99	76
c12	20	46	70	71	17	32	50	72	64	9	96	46	0	57	65	31
c13	47	70	56	26	65	54	27	40	59	57	47	92	57	0	8	51
c14	54	76	59	24	72	61	32	41	63	65	41	99	65	8	0	57
c15	11	21	87	73	23	4	59	81	84	38	76	76	31	51	57	0