The MDPLIB 2.0 Library of Benchmark Instances for Diversity Problems

Rafael Martí[1], Abraham Duarte[2], Anna Martínez-Gavara[1], and Jesús Sánchez-Oro[2]

[1]Department of Statistics and Operations Research, University of Valencia, Spain
[2]Department of Computer Science, University Rey Juan Carlos, Spain

March 2021

The library MDPLIB 2.0[1] contains 770 instances classified in different subsets according to their source. We consider three sets of instances depending on the type of values in their distance matrices: Euclidean, Real, and Integer. An extensive description of their characteristics follows.

1. **Euclidean instances set.** This data set consists of 160 matrices for which the values were calculated as the Euclidean distances from randomly generated points with coordinates in the 0 to 10 range. It collects two subsets, namely GKD-c, and GKD-d:

   - GKD-c: Duarte and Martí (2007) generated these 20 matrices with 10 coordinates for each point and n = 500 and m = 50.

   - GKD-d: Parreño et al. (2021) generated 70 matrices for which the values were calculated as the Euclidean distances from randomly generated points with two coordinates in the 0 to 100 range. For each value of $n = 25, 50, 100, 250, 500, 1000$, and 2000, they considered 10 instances with $m = \lceil n/10 \rceil$ and 10 instances with $m = 2\lceil n/10 \rceil$, totalizing 140 instances. The main motivation of this new set is to include the original coordinates in the instances files that unfortunately are not publicly available nowadays for the other subsets. In this way, researchers may represent the solutions in line with the work in Parreño et al. (2021).

2. **Real instances set.** This data set consists of 140 matrices with real numbers randomly selected according to a uniform distribution.

   - MDG-a. This data set contains 60 instances. Duarte and Martí (2007) generated 40 matrices with real numbers randomly selected in $[0, 10]$ and called them *Random Type I instances*, 20 of them with $n = 500$ and $m = 50$, and the other 20 with $n = 2000$ and $m = 200$. Parreño et al. (2021) generated 20 additional matrices with $n = 100$ and real numbers randomly selected in $[0, 10]$ that can be solved to optimality.

[1]Please, cite as: Martí, Duarte, and Martínez-Gavara (2021), The MDPLIB 2.0 Library of Benchmark Instances for Diversity Problems, University of Valencia. https://www.uv.es/rmarti/paper/mdp.html.

- MDG-b. This data set contains 60 instances. Originally, Duarte and Martí (2007) created this set with 40 matrices generated with real numbers randomly selected in [0, 1000] and called them *Random Type II instances*. 20 of them have $n = 500$ and $m = 50$, and the other 20 have $n = 2000$ and $m = 200$. Parreño et al. (2021) generated 20 additional matrices with $n = 100$ and real numbers randomly selected in $[0, 1000]$,

- MDG-c. Martí et al. (2013) proposed this data set with very large instances. It consists of 20 matrices with randomly generated numbers according to a uniform distribution in the range [0, 1000], and with $n = 3000$ and $m = 300$, 400, 500 and 600.

3. **Integer instances set.** This data set consists of 170 instances where the distance matrices are integer random numbers generated from an integer uniform distribution.

   - ORLIB: This is a set of 10 instances with $n = 2500$ and $m = 1000$ that were proposed for binary problems (Beasley, 1990). The distances are integers generated at random in $[-100, 100]$ where the diagonal distances are ignored.

   - PI: Palubeckis (2007) generated 10 instances where the distances are integers from a $[0, 100]$ uniform distribution. 5 of them are generated with $n = 3000$ and $m = 0.5n$, and 5 with $n = 5000$ and $m = 0.5n$. The density of the distance matrix is 10%, 30%, 50%, 80% and 100%.

   - SOM-a. These 50 instances were generated by Martí et al. (2010) with a generator developed by Silva et al. (2004) with integer random numbers between 0 and 9 generated from an integer uniform distribution. The instance sizes are such that for $n = 25$, $m = 2$ and 7; for $n = 50$, $m = 5$ and 15; for $n = 100$, $m = 10$ and 30; for $n = 125$, $m = 12$ and 37; and for $n = 150$, $m = 15$ and 45.

   - SOM-b. These 20 instances were generated by Silva et al. (2004) with the same random generator from SOM-a. The instance sizes are such that for $n = 100$, $m = 10, 20, 30$ and 40; for $n = 200$, $m = 20, 40, 60$ and 80; for $n = 300$, $m = 30, 60, 90$ and 120; for $n = 400$, $m = 40, 80, 120$, and 160; and for $n = 500$, $m = 50, 100, 150$ and 200.

   - MGPO: To complement the sets above, Martí et al. (2021) considered 80 large matrices with relatively low m values. Specifically, Martí et al. (2021) generate 40 instances with $n = 1000$ and integer numbers randomly selected in $[1, 100]$, 20 of them with $m = 50$ and 20 with $m = 100$. Similarly, we generate 40 matrices with $n = 2000$ and integer numbers randomly selected in $[1, 100]$, 20 of them with $m = 50$, and 20 with $m = 100$.

A final note on the use of instances is its applicability to the different models. It must be noted that some of them were introduced for the MaxSum model, and could not be adequate for other diversity models. This is especially true in the case of some instances in the SOM set that contain so many 0 values that all feasible solutions have a minimum distance value of 0. The empirical analysis in Martí et al. (2021) shows that 23 instances in the SOM set have an optimal MaxMin value of 0, and therefore if we apply a heuristic and obtain a solution with a value of 0 in the MaxMin objective, this is not a reliable measure of

its assessment. Researchers have to be very careful when using this set to test other models than the classic MaxSum. This 23 instances are:

- `SOM-a_36_n125_m37`, `SOM-a_37_n125_m37`, `SOM-a_38_n125_m37`, `SOM-a_39_n125_m37`, `SOM-a_40_n125_m37`, `SOM-a_46_n150_m45`, `SOM-a_47_n150_m45`, `SOM-a_48_n150_m45`, `SOM-a_49_n150_m45`, `SOM-a_50_n150_m45`.
- `SOM-b_3_n100_m30`, `SOM-b_4_n100_m40`, `SOM-b_7_n200_m60`, `SOM-b_8_n200_m80`, `SOM-b_10_n300_m60`, `SOM-b_11_n300_m90`, `SOM-b_12_n300_m120`, `SOM-b_14_n400_m80`, `SOM-b_15_n400_m120`, `SOM-b_16_n400_m160`, `SOM-b_18_n500_m100`, `SOM-b_19_n500_m150`, `SOM-b_20_n500_m200`.

Table 1 summarizes the library MDPLIB 2.0 for the maximum diversity problems [2]. This table shows the number of instances, type, and the range of n and m in each subset.

Set	**# Instances**	**Type**	**Range of n**	**Range of m**
GKD-c	20	Euclidean	500	50
GKD-d	140		[25, 2000]	[3, 400]
MDG-a	60	Real numbers	[100, 2000]	[50, 200]
MDG-b	60		[100, 2000]	[50, 200]
MDG-c	20		3000	[300, 600]
ORLIB	10	Integer numbers	2500	1000
PI	10		$\{3000, 5000\}$	$\{1500, 2500\}$
SOM-a	50		[25, 150]	[2, 45]
SOM-b	20		[100, 500]	[10, 200]
MGPO	80		[1000, 2000]	[50, 100]
Total	470		[25, 5000]	[2, 2500]

Table 1: MDPLIB 2.0 benchmark library for MDP.

Constrained benchmark instances

The benchmark set of instances in the constrained dispersion problem is derived from the original MDPLIB. Specifically, Peiró et al. (2021) and Martínez-Gavara et al. (2021) select a subset of 50 instances to generate the new benchmark set. Specifically, they are selected from three sets:

- GKD: this set was originally proposed by Glover (1989) for small-size instances, and it was extended for medium- and large-size instances in

[2] without capacity and cost constraints

Duarte and Martí (2007) and Martí et al. (2010), respectively. In particular, 10 instances of size 50, 10 of size 150, and 10 of size 500 are selected.

- MDG: this data set was proposed in Duarte and Martí (2007) and it consists of 100 matrices with real numbers randomly selected from a uniform distribution. In particular, 10 of this set of size 500 are selected.

- SOM: this data set was created by Martí et al. (2010) for the maximum diversity problem, where the objective function is the sum of the distances. The matrices of this set are generated with random numbers of an integer uniform distribution between 0 and 9. In particular, 10 of them of size 50 are selected.

Capacitated Dispersion Problem. In the Capacitated Dispersion Problem (CDP), for each selected original instance, Peiró et al. (2021) randomly generate the capacity value of each node in the range [1, 1000]. Then, they compute the sum of all capacities and set the minimum capacity B as this sum multiplied by a factor φ_b of 0.2 and 0.3 respectively, thus creating two instances for each original one. The benchmark for the Capacitated Diversity Problem thus consists in 100 instances. We named the file that contains these 100 instances, Const-(CDP).

Generalized Dispersion Problem. In the Generalized Dispersion Problem (GDP), Martínez-Gavara et al. (2021) generate the capacity and cost real numbers with a Uniform distribution. Specifically, as in the (CDP) the capacity c_i of a node $i \in V$ is generated by a $U(1, 1000)$, the fix cost a_i is generated from its capacity c_i by a $U(c_i/2, 2\,c_i)$, and finally the variable cost b_i is generated by $U(\min(1, a_i), \max(1, a_i))/100$. The minimum capacity B is computed as the sum of all capacities multiplied by a factor φ_b of 0.2 or 0.3, and the maximum budget is computed as:

$$K = \begin{cases} \varphi_k \sum_{i \in V} a_i & \text{(GDP-f) model,} \\ \varphi_k \sum_{i \in V} (a_i + b_i\ c_i) & \text{(GDP-v) model} \end{cases}$$

where φ_k is a 0.2 or 0.3 factor (see both models in Martínez-Gavara et al. (2021)). Therefore, each original instance in the MDPLIB produces 4 instances, thus obtaining a benchmark set of 200. We named the file that contains these instances 200, Const-(GDP).

We finish the description of the instances for the (CDP) and (GDP), summarizing the library, MDPLIB 2.0 for the constrained problems in Table 2.

References

Beasley, J. E. (1990). OR-Library: Distributing Test Problems by Electronic Mail. *Journal of the Operational Research Society*, 41(11):1069–1072.

Duarte, A. and Martí, R. (2007). Tabu search and GRASP for the maximum diversity problem. *European Journal of Operational Research*, 178(1):71–84.

Set	**#inst.**	**Type**	n	**Cap. factor** (φ_b)	**Cost factor** (φ_k)
100 instances for (CDP) — Const-(CDP)					
GKD-b	40	Euclidean	$50, 150$	$0.2, 0.3$	-
GKD-c	20		500	$0.2, 0.3$	-
MDG-b	20	Real	500	$0.2, 0.3$	-
SOM	20	Integer	50	$0.2, 0.3$	-
200 instances for (GDP) — Const-(GDP)					
GKD-b	80	Euclidean	$50, 150$	$0.2, 0.3$	$0.2, 0.3$
GKD-c	40		500	$0.2, 0.3$	$0.2, 0.3$
MDG-b	40	Real	500	$0.2, 0.3$	$0.2, 0.3$
SOM	40	Integer	50	$0.2, 0.3$	$0.2, 0.3$

Table 2: MDPLIB 2.0 benchmark library for (CDP) and (GDP).

Glover, F. (1989). Tabu Search—Part I. *ORSA Journal on Computing*, 1(3):190–206.

Martí, R., Gallego, M., and Duarte, A. (2010). A branch and bound algorithm for the maximum diversity problem. *European Journal of Operational Research*, 200(1):36–44.

Martí, R., Gallego, M., Duarte, A., and Pardo, E. G. (2013). Heuristics and metaheuristics for the maximum diversity problem. *Journal of Heuristics*, 19(4):591–615.

Martí, R., Martínez-Gavara, A., Pérez, S., and Sánchez-Oro, J. (2021). Discrete diversity and dispersion maximization. A review and an empirical analysis from an OR perspective. *European Journal of Operational Research (Submitted)*.

Martínez-Gavara, A., Corberán, T., and Martí, R. (2021). GRASP and Tabu Search for the Generalized Dispersion Problem. *Expert Systems with Applications*, 173:114703.

Palubeckis, G. (2007). Iterated tabu search for the maximum diversity problem. *Applied Mathematics and Computation*, 189(1):371–383.

Parreño, F., Álvarez-Valdés, R., and Martí, R. (2021). Measuring diversity. a review and an empirical analysis. *European Journal of Operational Research*, 289(2):515–532.

Peiró, J., Jiménez, I., Laguardia, J., and Martí, R. (2021). Heuristics for the capacitated dispersion problem. *International transactions in operational research*, 28(1):119–141.

Silva, G. C., Ochi, L. S., and Martins, S. L. (2004). Experimental comparison of greedy randomized adaptive search procedures for the maximum diversity problem. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3059, pages 498–512.