HEIG VD | IIDE Institut interdisciplinaire du Développement de l'entreprise

EPFL

Resampling estimation of discrete choice models

11th Symposium of the European Association for Research in Transportation
6–8 September 2023 | ETH Zurich, Switzerland

Nicola Ortelli[1,2], Matthieu de Lapparent[1], Michel Bierlaire[2]

[1] IIDE, HEIG-VD
[2] TRANSP-OR, EPFL

Motivation

DCMs in the era of big data

- Specifying DCMs is time-consuming.
- Ever-larger datasets:
  - "Wider" data — more variables;
  - "Taller" data — more observations.
- Two distinct problems: specification and estimation.

Motivation

DCMs in the era of big data

- Specifying DCMs is time-consuming.
- Ever-larger datasets:
  - "Wider" data — more variables;
  - "Taller" data — more observations.
- Two distinct problems: specification and estimation.

Speeding up model estimation

- Obtain estimation results faster.
- Spend more time looking for good specifications.

Intuition

Maximum likelihood estimation

- Let $\mathcal{N} = \{(\boldsymbol{x}_n, i_n) : n = 1, \ldots, N\}$ be a choice dataset.
- Log likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log P\left(i_n \mid \boldsymbol{x}_n; \boldsymbol{\theta}\right).$$

- Computational time is linear in N.

Intuition

Maximum likelihood estimation

- Let $\mathcal{N} = \{(\boldsymbol{x}_n, i_n) : n = 1, \ldots, N\}$ be a choice dataset.
- Log likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log P\left(i_n \mid \boldsymbol{x}_n; \boldsymbol{\theta}\right).$$

- Computational time is linear in N.

Factoring-out redundancy

- Suppose that some observations in $\mathcal{N}$ are identical.
- For $U < N$ unique observations:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{u=1}^{U} N_u \cdot \log P\left(i_u \mid \boldsymbol{x}_u; \boldsymbol{\theta}\right).$$

- Computation takes $\approx \frac{U}{N}$ less time!

Intuition

Maximum likelihood estimation

- Let $\mathcal{N} = \{(\boldsymbol{x}_n, i_n) : n = 1, \ldots, N\}$ be a choice dataset.
- Log likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log P\left(i_n \mid \boldsymbol{x}_n; \boldsymbol{\theta}\right).$$

- Computational time is linear in N.

Factoring-out redundancy

- Suppose that some observations in $\mathcal{N}$ are identical.
- For $U < N$ unique observations:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{u=1}^{U} N_u \cdot \log P\left(i_u \mid \boldsymbol{x}_u; \boldsymbol{\theta}\right).$$

- Computation takes $\approx \frac{U}{N}$ less time!

$\Rightarrow$ **Extend factorization to "nearly identical" observations!**

Resampling estimation of DCMs

Procedure

1. Group similar observations.
2. Sample from each group.
3. Weight based on group sizes.
4. Use weighted maximum likelihood estimation.

Resampling estimation of DCMs

Procedure

1. Group similar observations.
2. Sample from each group.
3. Weight based on group sizes.
4. Use weighted maximum likelihood estimation.

Challenges

- Minimize information loss.
- Clustering must be fast!

Resampling estimation of DCMs

Procedure

1. Group similar observations.
2. Sample from each group.
3. Weight based on group sizes.
4. Use weighted maximum likelihood estimation.

Challenges

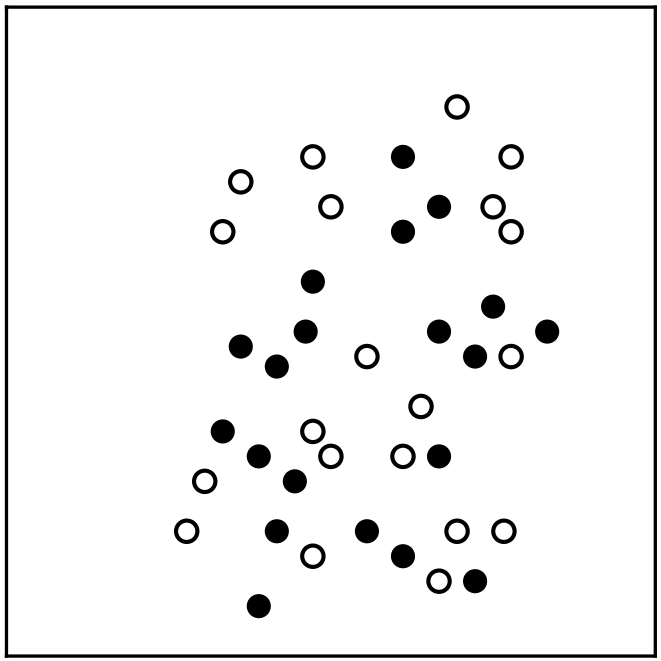
- Minimize information loss.
- Clustering must be fast!

$\Rightarrow$ **Use locality-sensitive hashing (LSH).**

LSH-based dataset reduction (LSH-DR)

Toy data

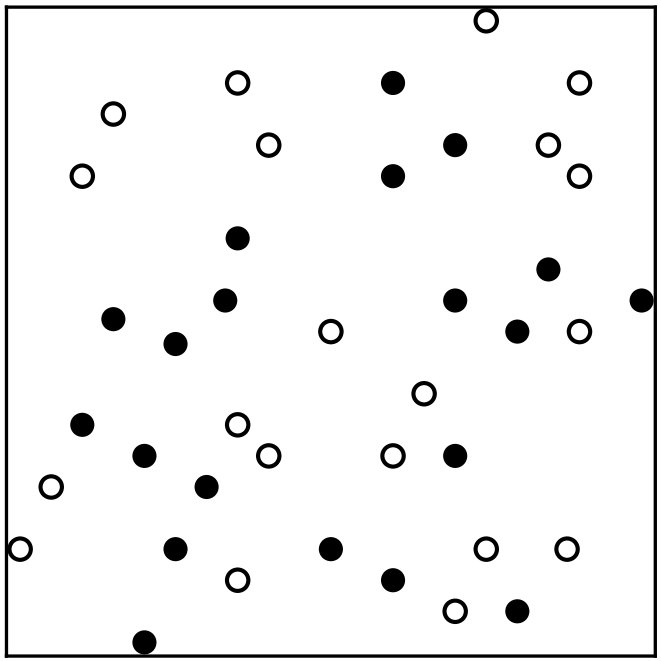
- 2 alternatives.
- 2 explanatory variables.



LSH-based dataset reduction (LSH-DR)

Min-max normalization

$$\frac{\boldsymbol{x} - \min(\boldsymbol{x})}{\max(\boldsymbol{x}) - \min(\boldsymbol{x})}$$

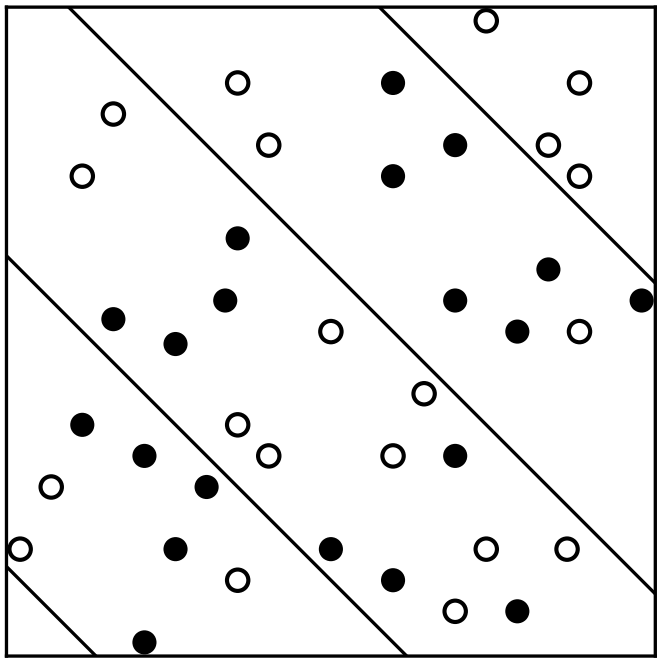


LSH-based dataset reduction (LSH-DR)

A single LSH function

$$h_{\boldsymbol{a},b}(\boldsymbol{x}_n) = \left\lfloor \frac{\boldsymbol{a} \cdot \boldsymbol{x}_n + b}{w} \right\rfloor$$

- $\boldsymbol{a} \overset{iid}{\sim} \mathcal{N}(0,1)$.
- $b \sim \mathcal{U}(0,w)$.
- w is the bucket width.

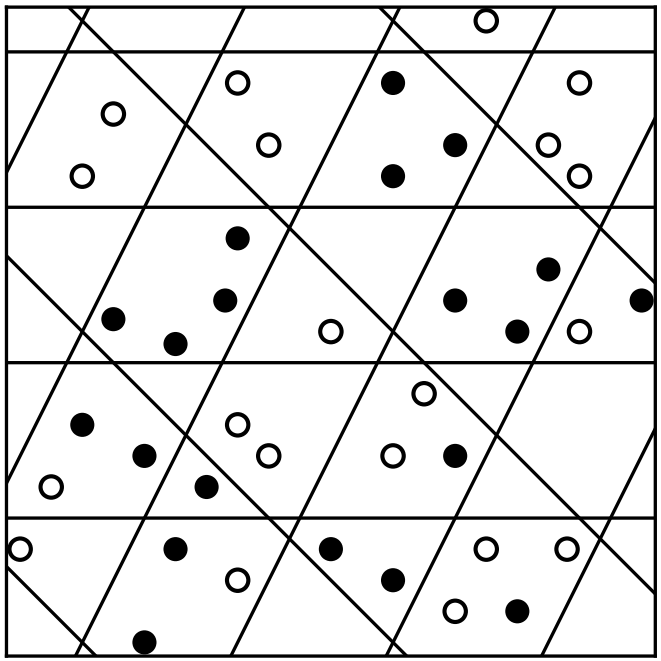


LSH-based dataset reduction (LSH-DR)

Combining LSH functions

$$H_{\boldsymbol{A},B}(\boldsymbol{x}_n) = H_{\boldsymbol{A},B}(\boldsymbol{x}_p) \Leftrightarrow$$
$$h_{\boldsymbol{a}_r,b_r}(\boldsymbol{x}_n) = h_{\boldsymbol{a}_r,b_r}(\boldsymbol{x}_p)\ \forall r.$$

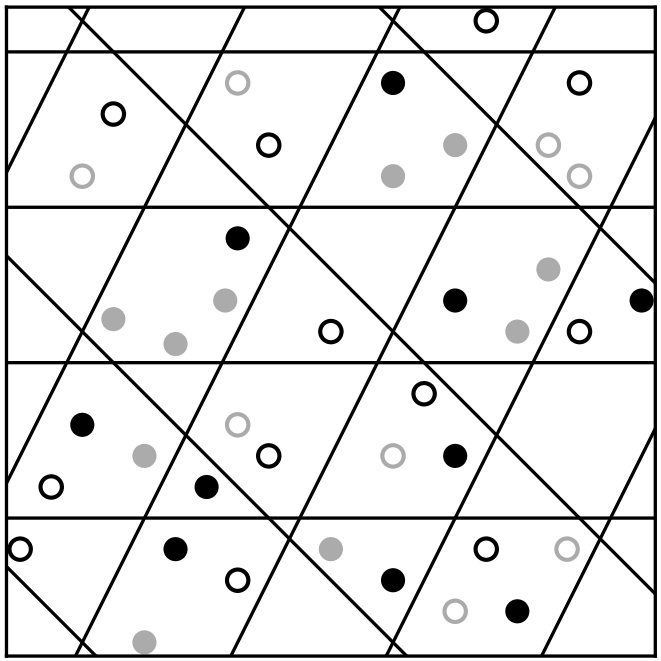
- $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_R)$.
- $B = (b_1, \ldots, b_R)$.



LSH-based dataset reduction (LSH-DR)

Sampling

- In each bucket, sample one observation per alternative.

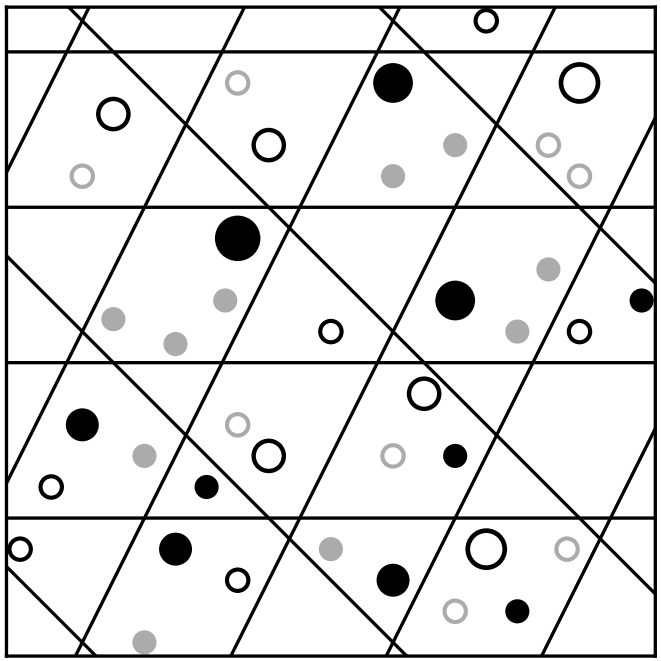


LSH-based dataset reduction (LSH-DR)

Weighting

- Each selected observation $(\boldsymbol{x}_g, i_g)$ is given a weight N_g:

$$N_g = |\{(\boldsymbol{x}_n, i_n) : i_g = i_n, \\ H_{\boldsymbol{A},B}(\boldsymbol{x}_g) = H_{\boldsymbol{A},B}(\boldsymbol{x}_n), \}|$$



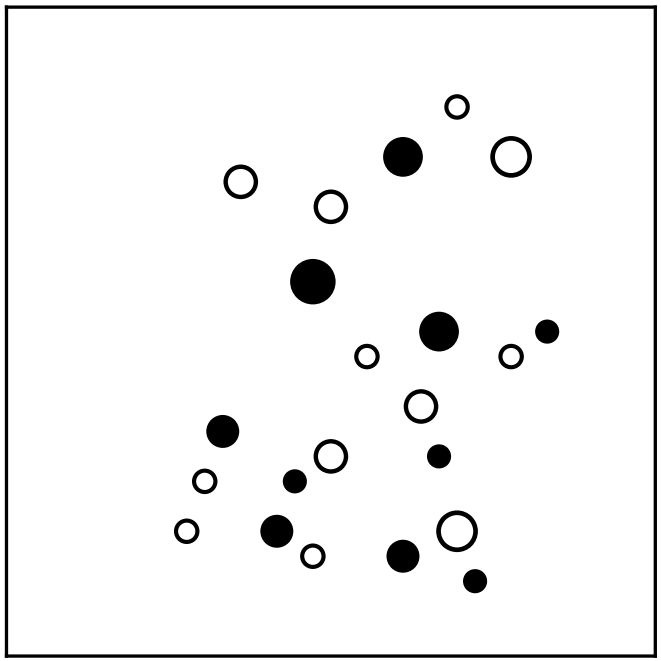
LSH-based dataset reduction (LSH-DR)

Weighted subsample

$$\{(\boldsymbol{x}_g, i_g, N_g) : g = 1, \ldots, G\}$$

Weighted likelihood

$$\mathcal{L}^w(\boldsymbol{\theta}) = \sum_{g=1}^{G} N_g \cdot \log P\left(i_g \mid \boldsymbol{x}_g; \boldsymbol{\theta}\right)$$



Experimental design

LPMC data [Hillel *et al.*, 2018]

- Mode choice, 4 alternatives.
- 81k observations (55k/26k).

Experimental design

LPMC data [Hillel *et al.*, 2018]

- Mode choice, 4 alternatives.
- 81k observations (55k/26k).

Models [Hillel, 2019; Ortelli *et al.*, 2023]

Logit-S

- 10 cont. variables.
- 0 dummies.
- **13 parameters**.

Nested-S

- 10 cont. variables.
- 0 dummies.
- **14 parameters**.

Logit-L

- 11 cont. variables.
- 15 dummies.
- **53 parameters**.

Experimental design

LPMC data [Hillel *et al.*, 2018]

- Mode choice, 4 alternatives.
- 81k observations (55k/26k).

Models [Hillel, 2019; Ortelli *et al.*, 2023]

Logit-S

- 10 cont. variables.
- 0 dummies.
- **13 parameters**.

Nested-S

- 10 cont. variables.
- 0 dummies.
- **14 parameters**.

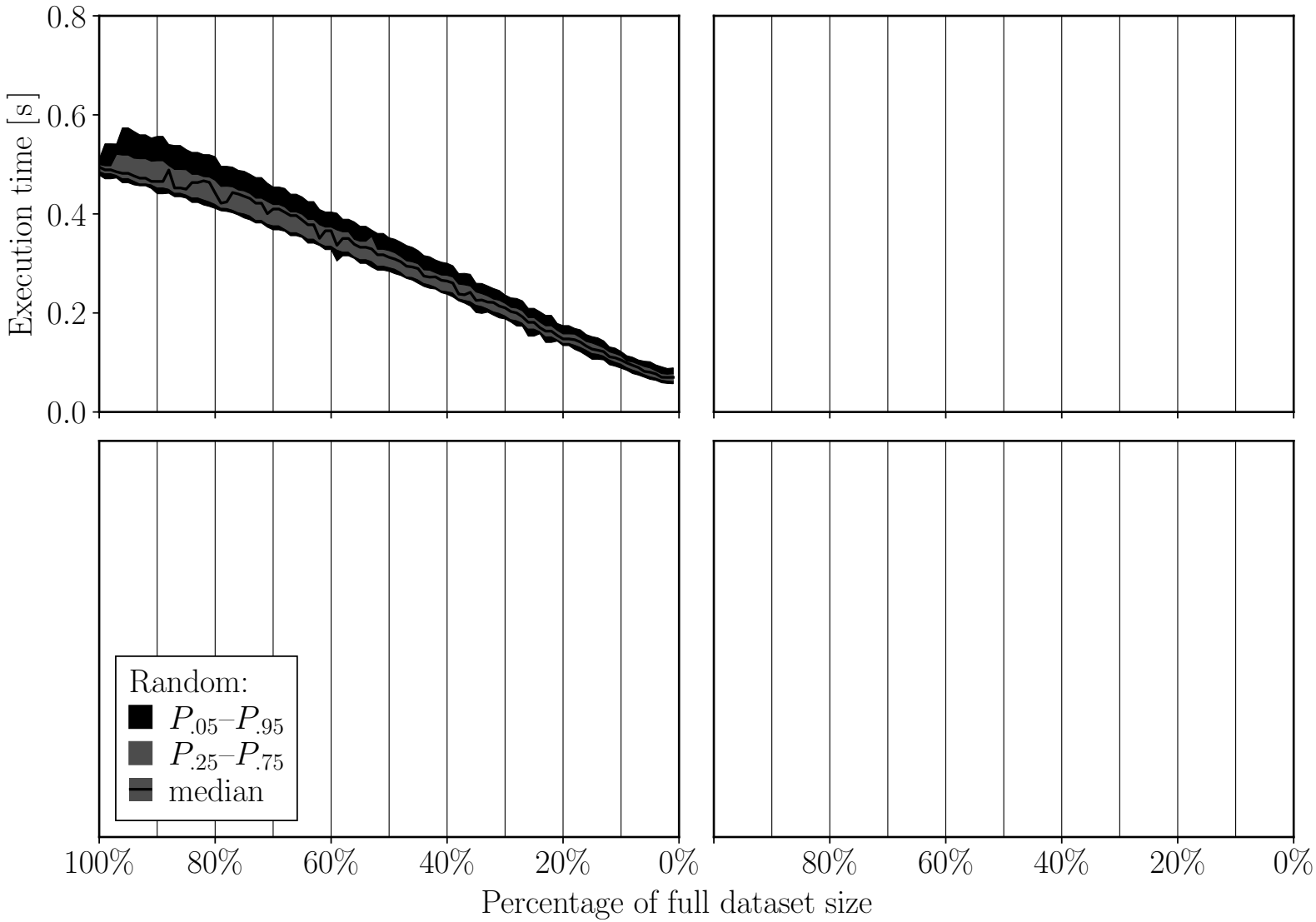
Logit-L

- 11 cont. variables.
- 15 dummies.
- **53 parameters**.

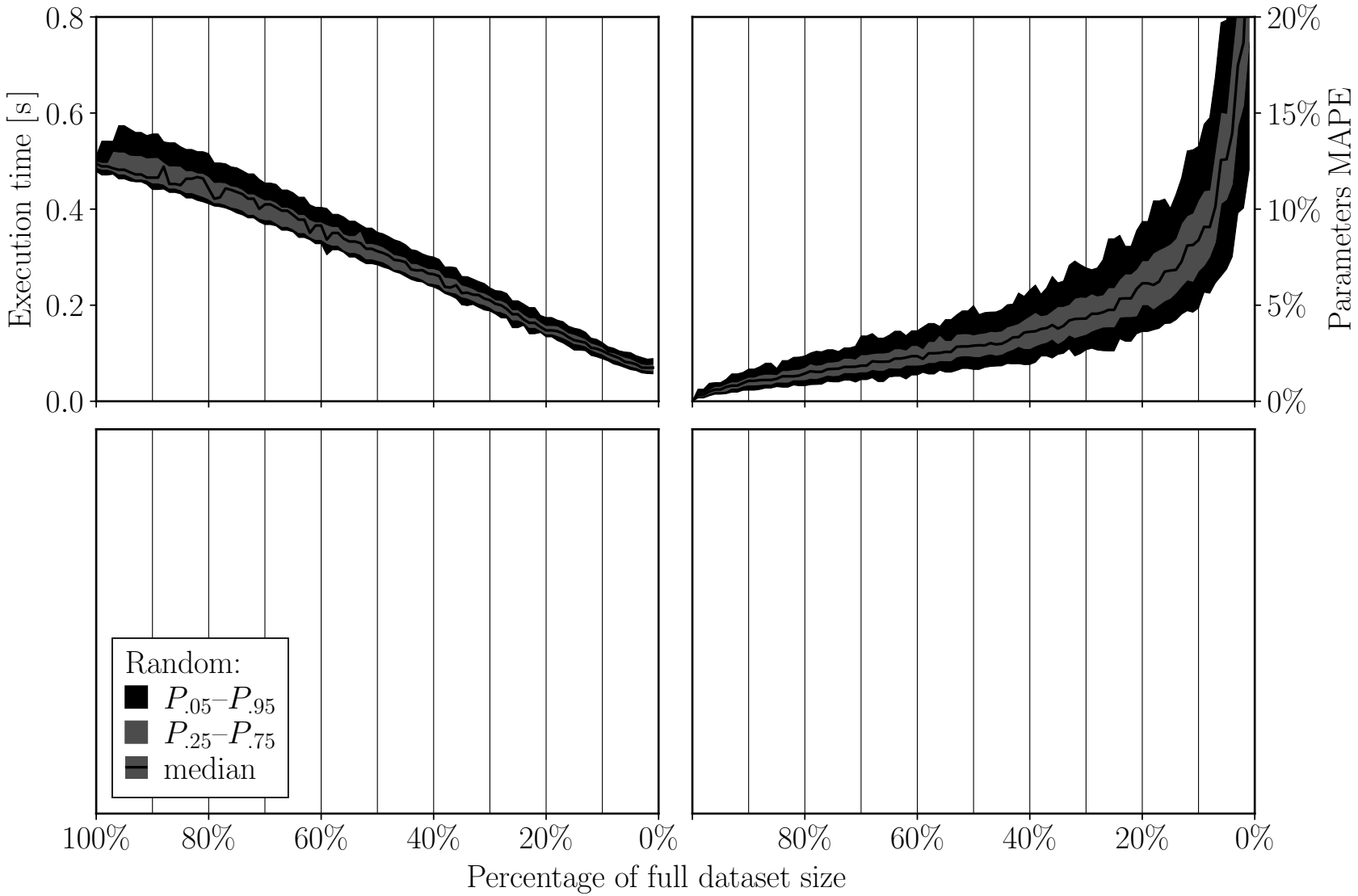
Estimation

- Biogeme. [Bierlaire, 2023]
- 36-core processor @2.4 GHz, 256 GB of RAM.

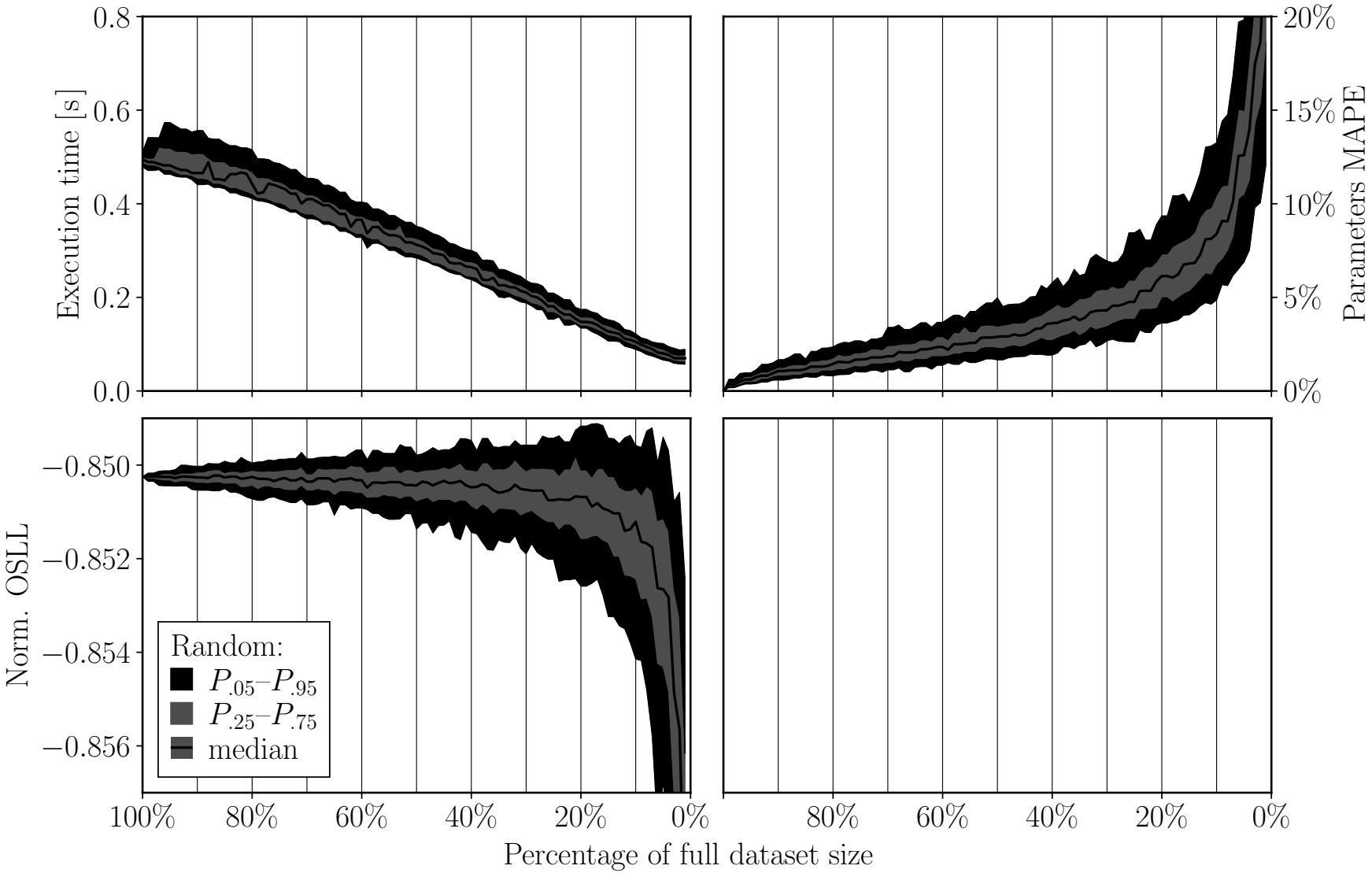
Results — Logit-S



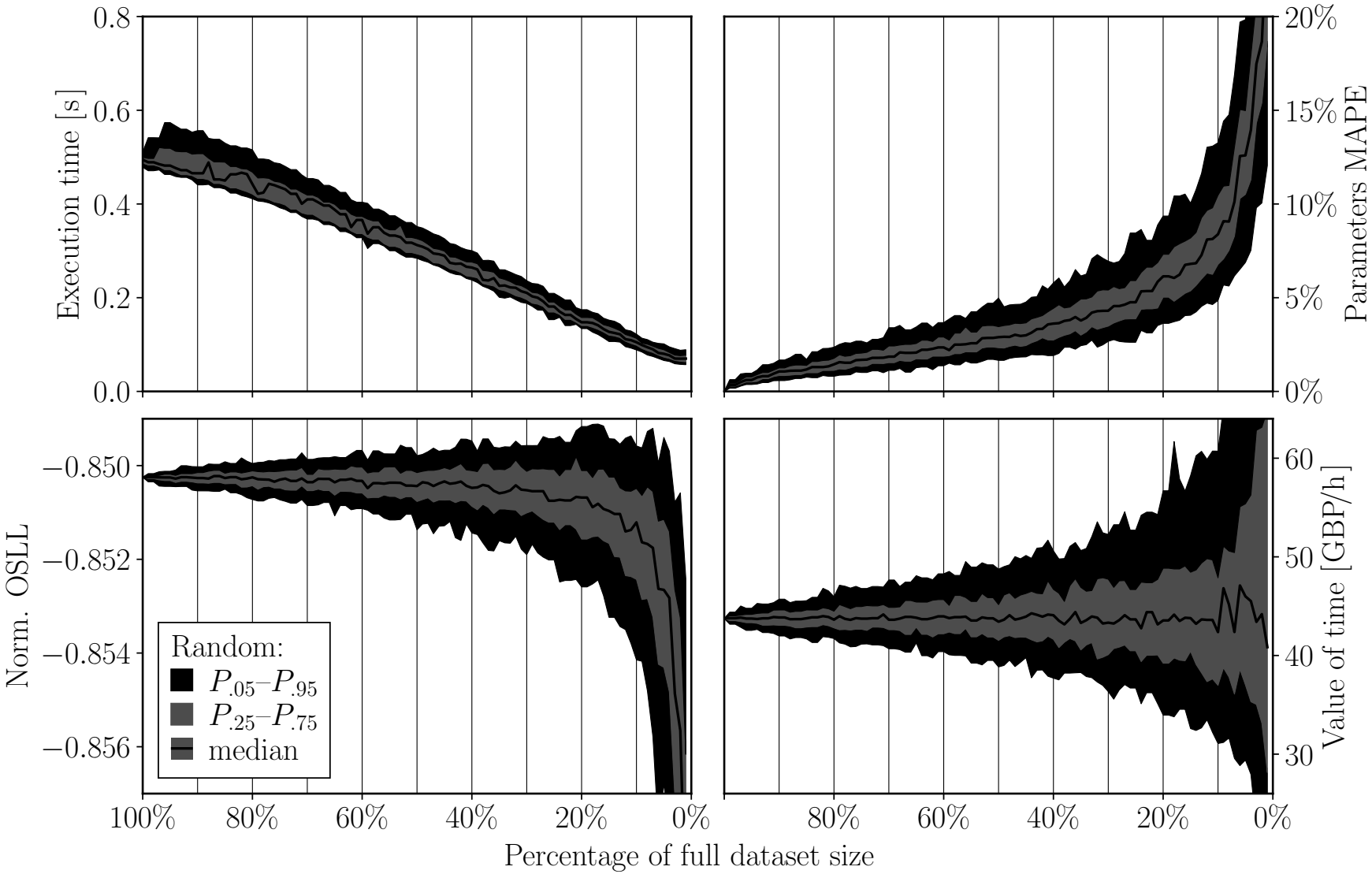
Results — Logit-S



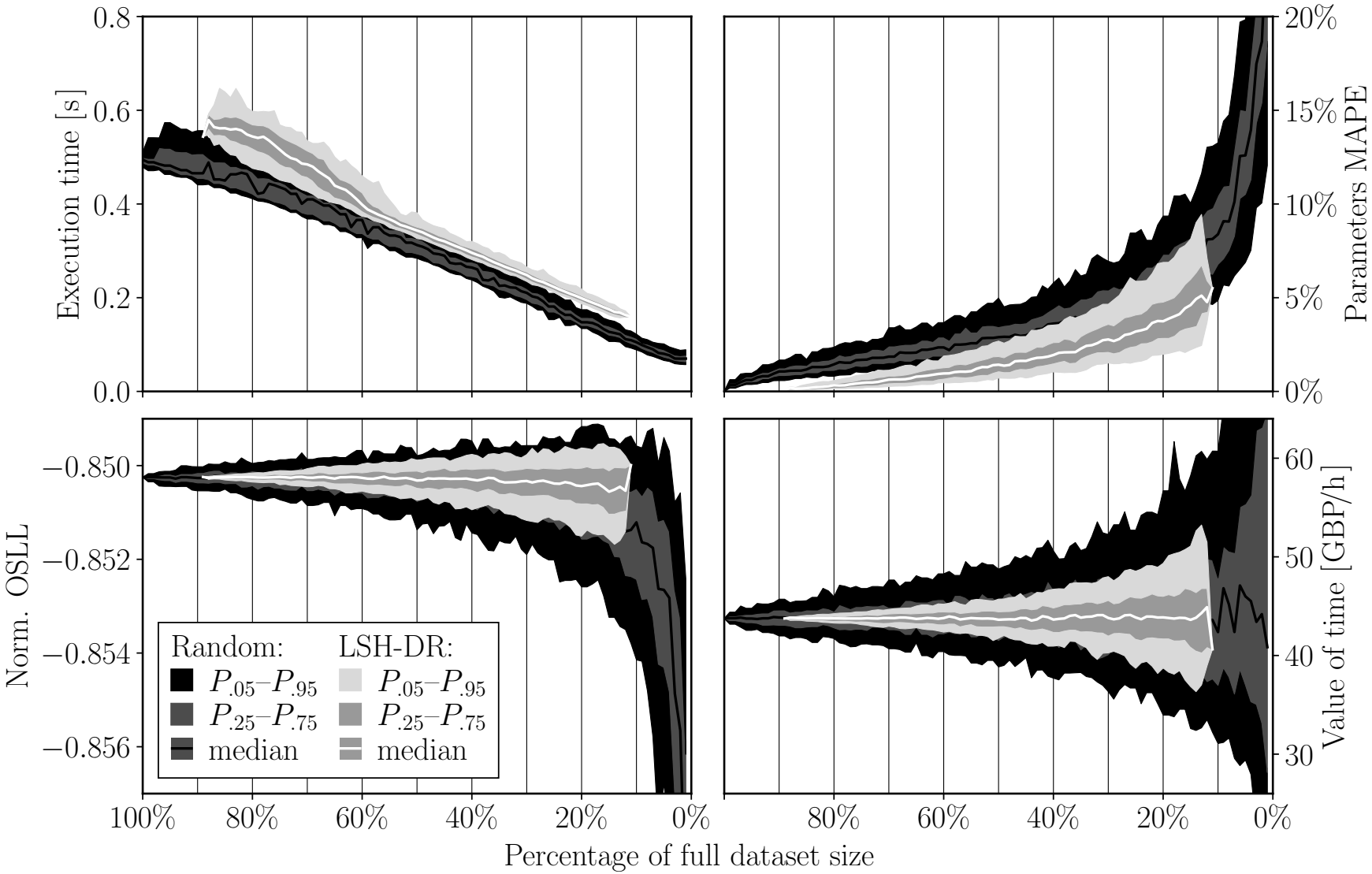
Results — Logit-S



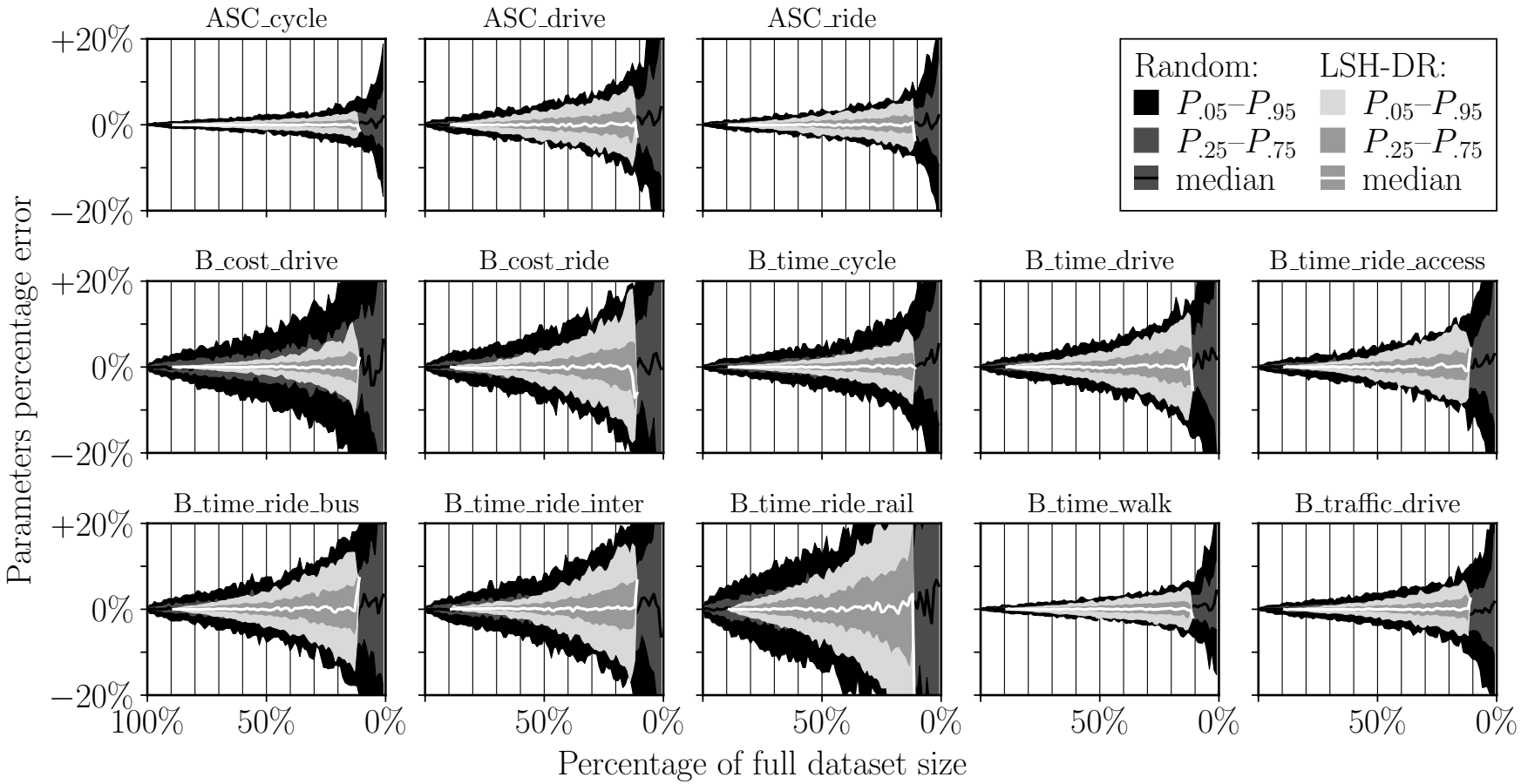
Results — Logit-S



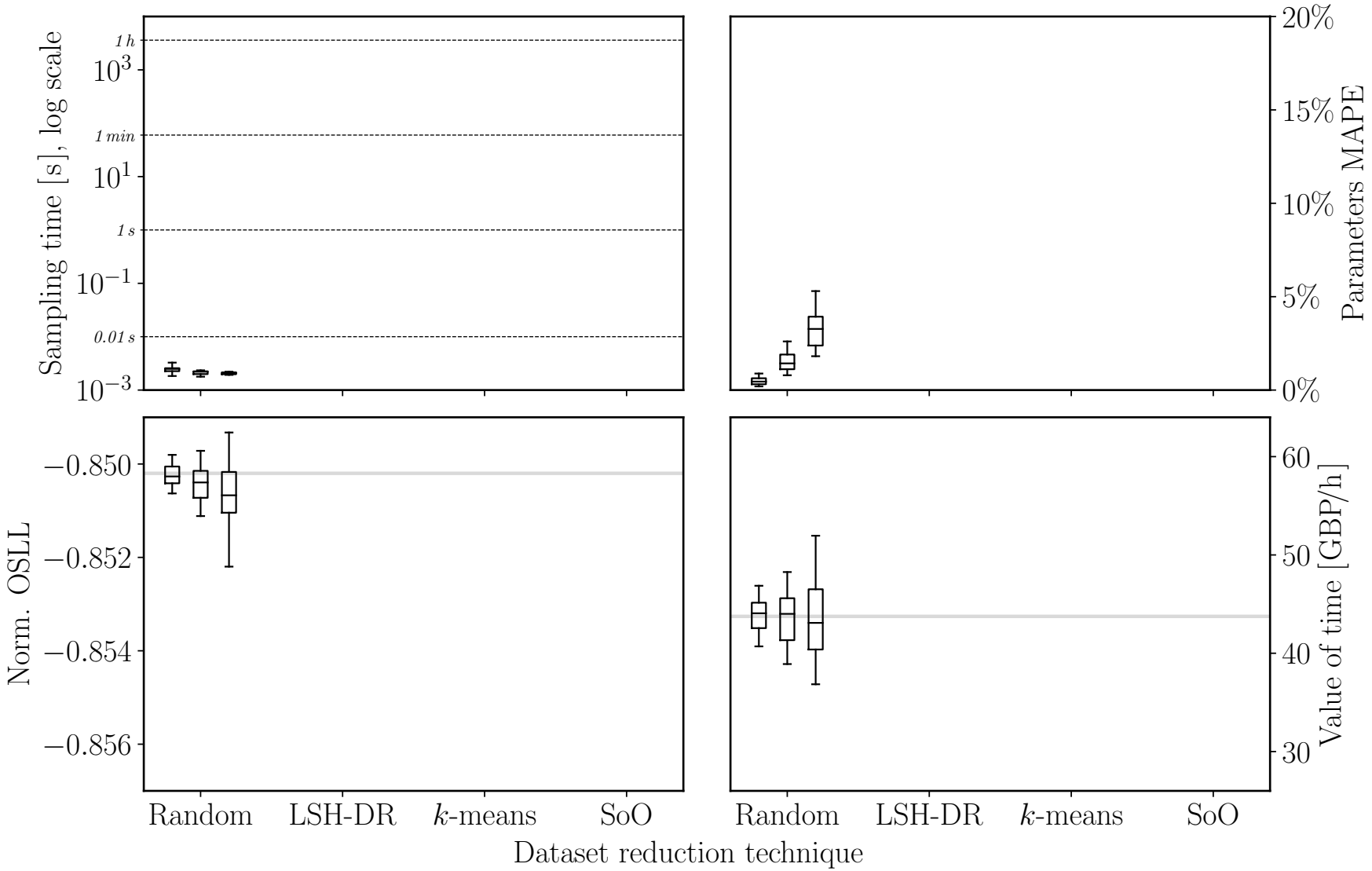
Results — Logit-S



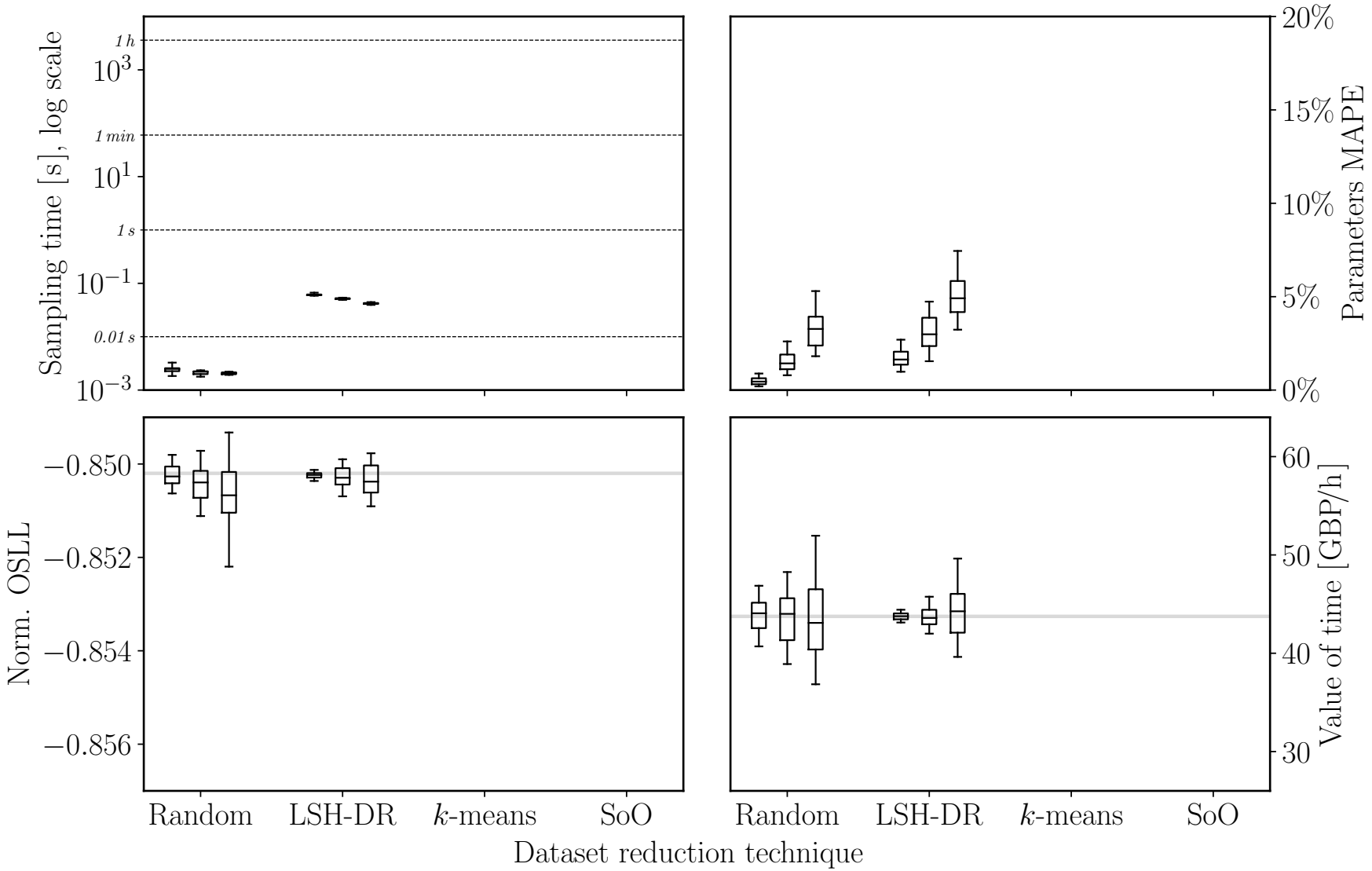
Results — Logit-S: Parameters



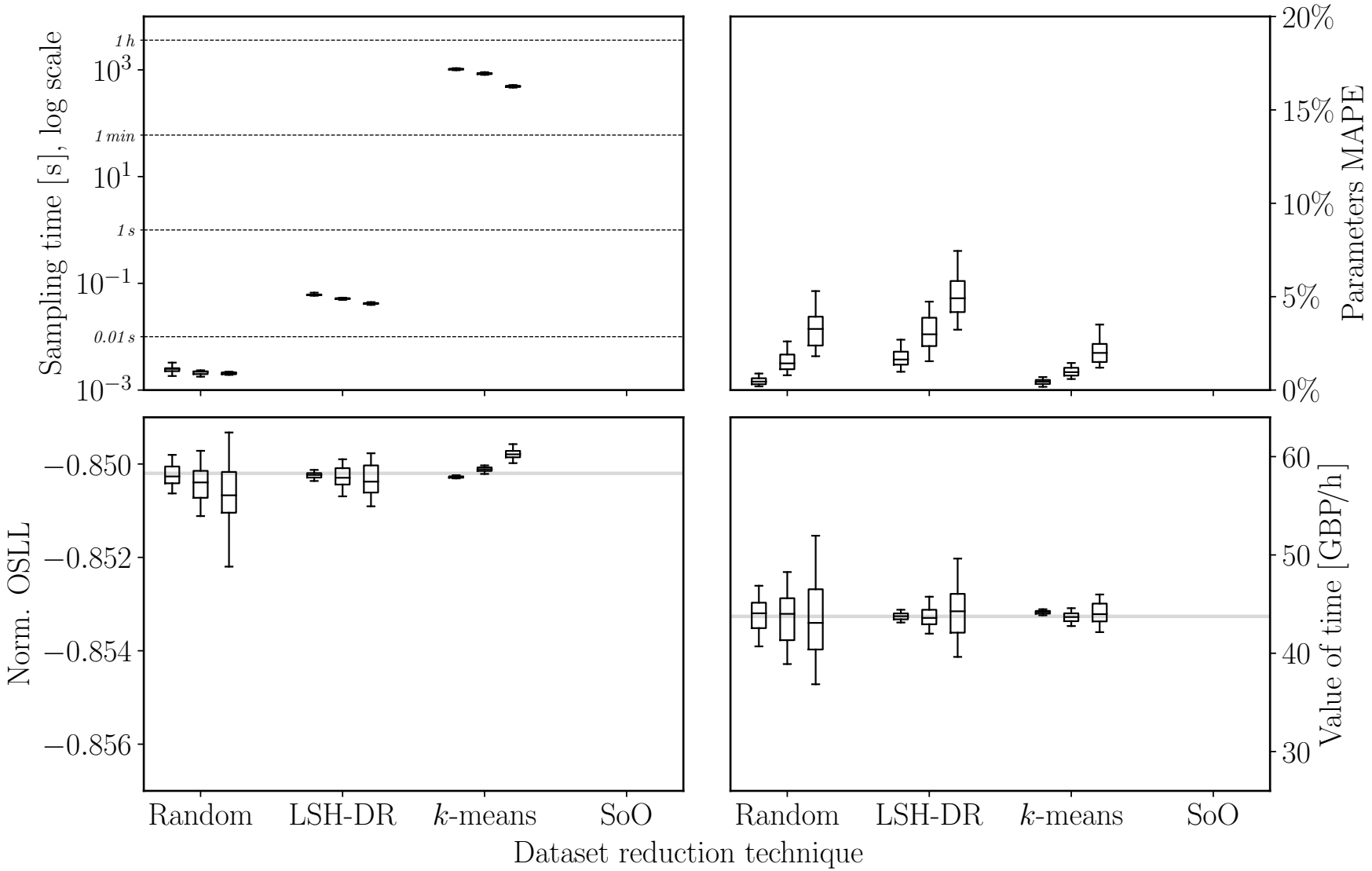
Results — Logit-S: Comparison



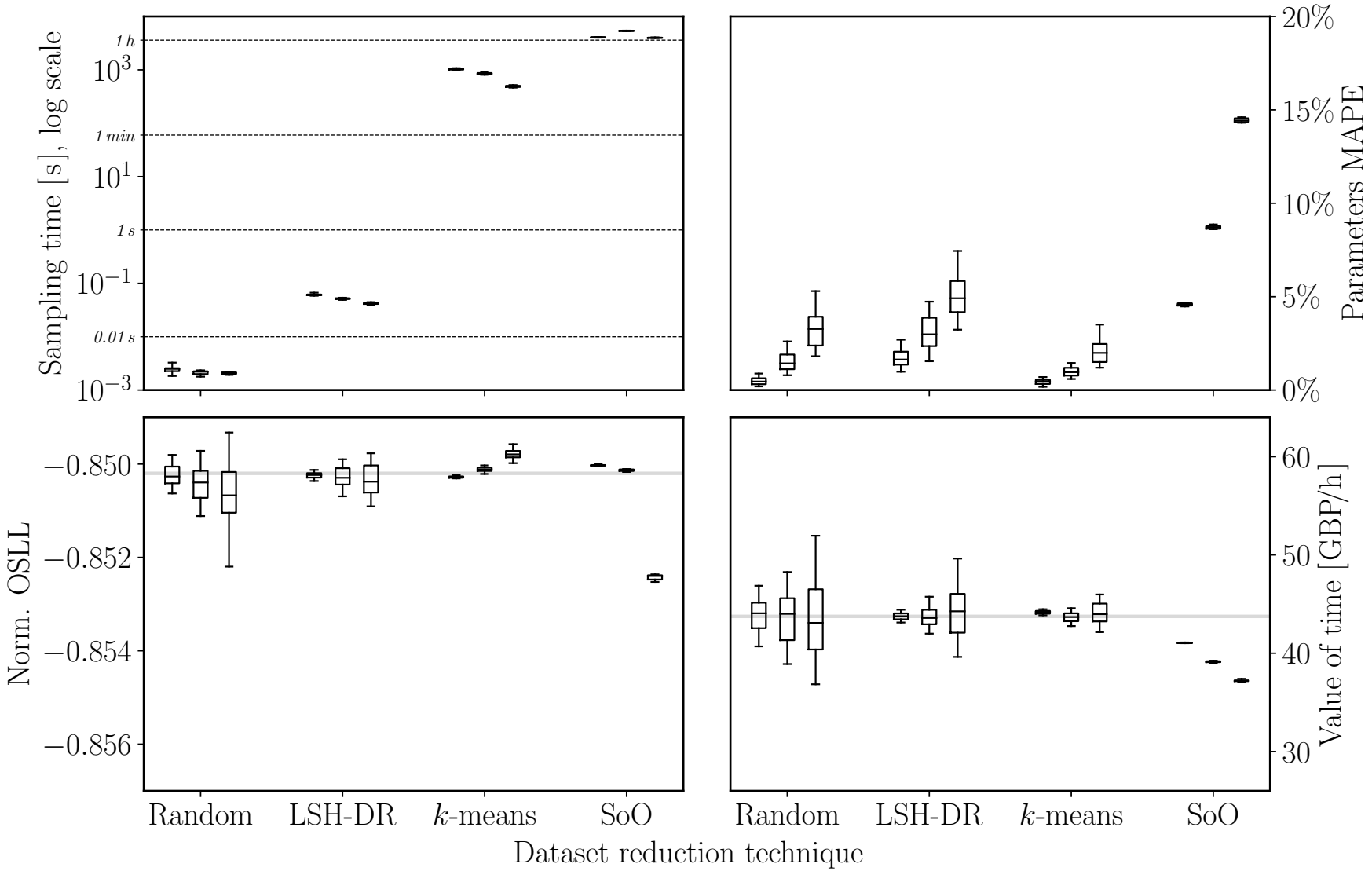
Results — Logit-S: Comparison



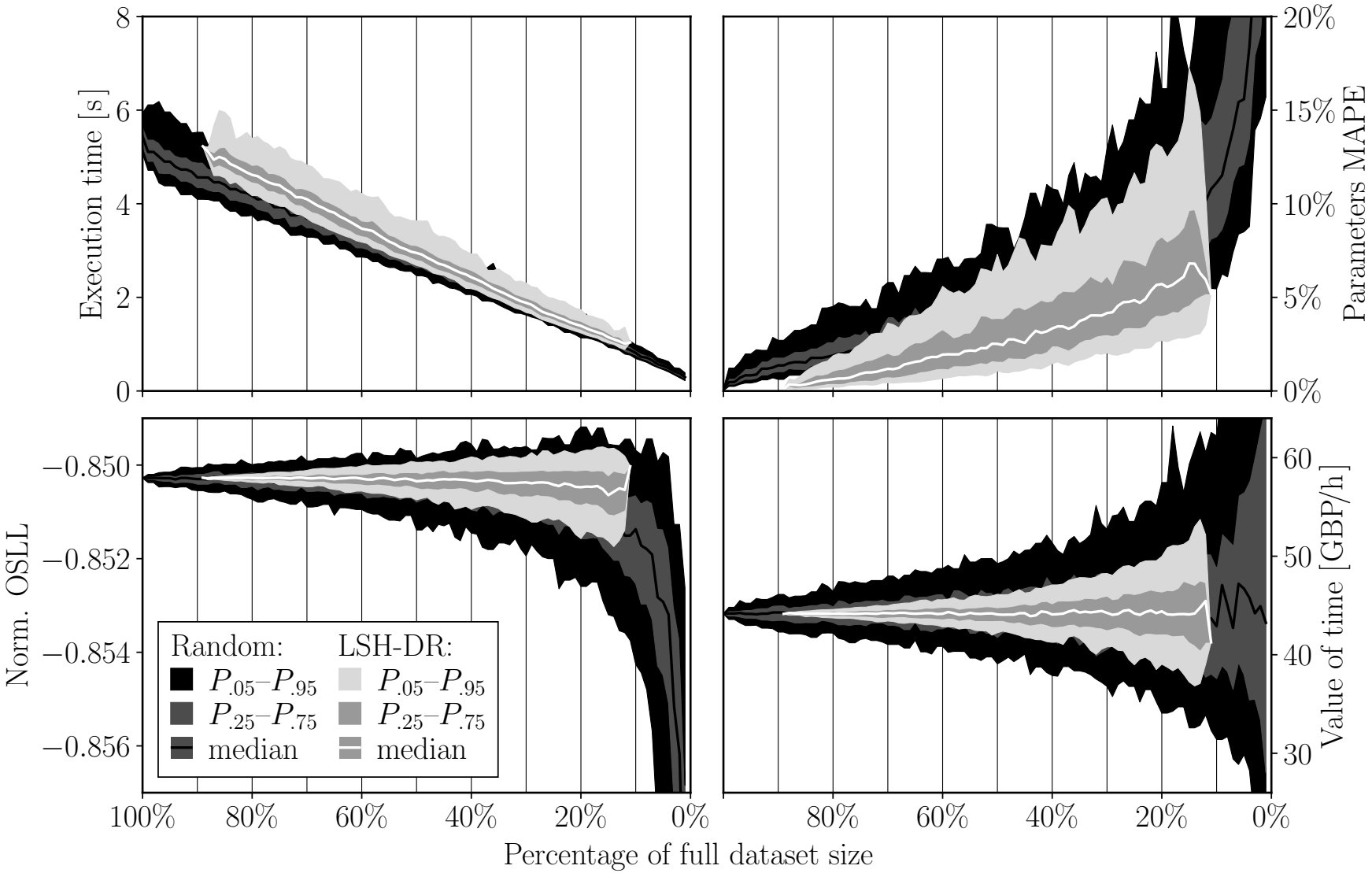
Results — Logit-S: Comparison



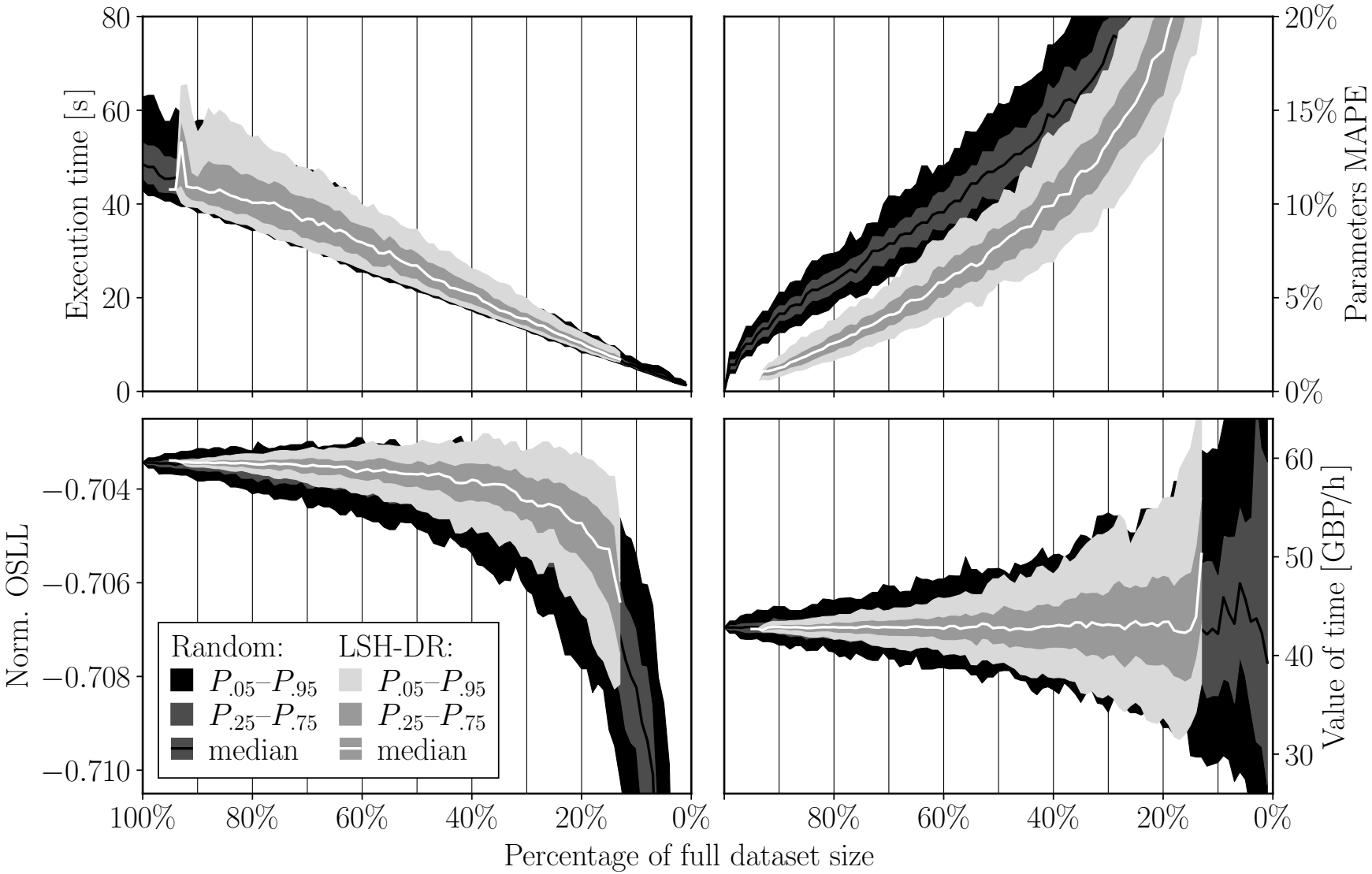
Results — Logit-S: Comparison



Results — Nested-S



Results — Logit-L



Conclusion

Summary

- Resampling technique designed to speed up DCM estimation.
- Time savings come at the cost of deteriorating the estimation results.
- Factor out redundancy, but keep diversity!

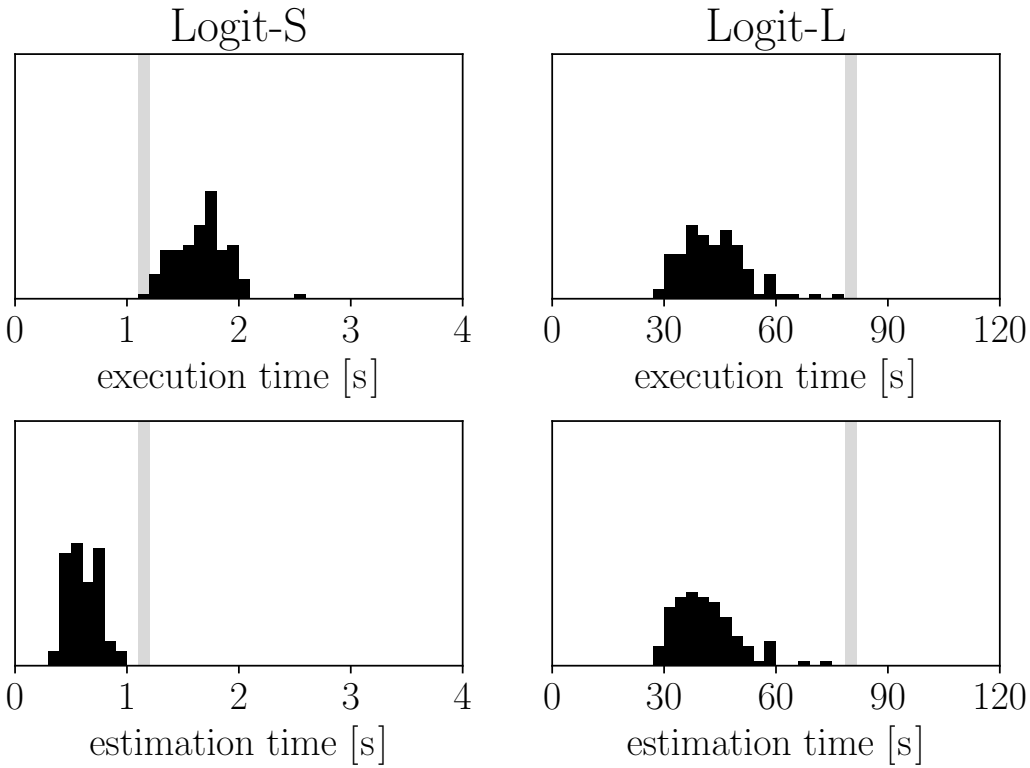
Next steps

- Synthetic prototypical observations.
- Knowledge-based LSH functions.
- Embed resampling within the model estimation process.

Spin-off

Adaptive resampling

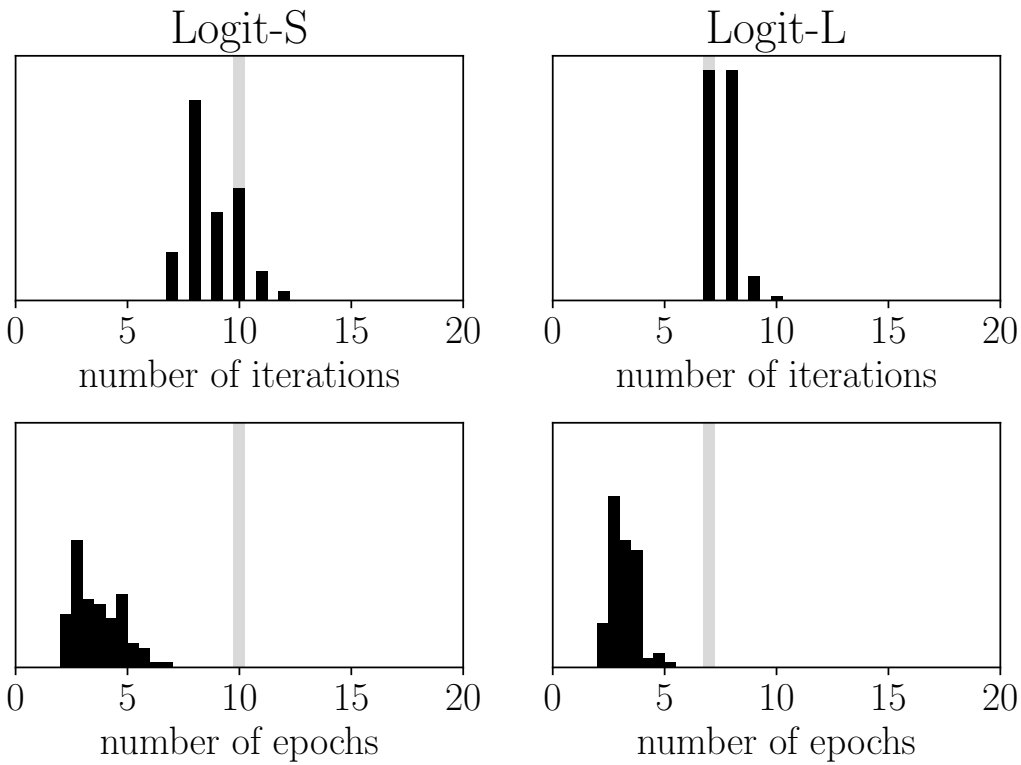
- Generate weighted batches for stochastic optimization. [Lederrey et al., 2021]
- Start small and increase batch size dynamically.



Spin-off

Adaptive resampling

- Generate weighted batches for stochastic optimization. [Lederrey et al., 2021]
- Start small and increase batch size dynamically.



References

LSH-DR

- Ortelli, N., de Lapparent, M. and Bierlaire, M. (2023). Resampling estimation of discrete choice models, Technical Report, TRANSP-OR 230330. Transport and Mobility Laboratory, ENAC, EPFL.

Direct precedents

- van Cranenburgh, S. and Bliemer, M. C. (2019). Information theoretic-based sampling of observations, Journal of choice modelling 31: 181–197.
- Schmid, B., Becker, F., Molloy, J., Axhausen, K. W., Lüdering, J., Hagen, J. and Blome, A. (2022). Modeling train route decisions during track works, Journal of Rail Transport Planning & Management 22: 100320.
- Lederrey, G., Lurkin, V., Hillel, T. and Bierlaire, M. (2021). Estimation of discrete choice models with hybrid stochastic adaptive batch size algorithms, Journal of choice modelling 38.

Dataset & models

- Hillel, T., Elshafie, M. Z. and Jin, Y. (2018). Recreating passenger mode choice-sets for transport simulation: A case study of London, UK, Proceedings of the Institution of Civil Engineers-Smart Infrastructure and Construction 171(1).
- Hillel, T. (2019). Understanding travel mode choice: A new approach for city scale simulation, PhD thesis, University of Cambridge.

HEIG VD | IIDE Institut interdisciplinaire du Développement de l'entreprise

EPFL

Resampling estimation of discrete choice models

11th Symposium of the European Association for Research in Transportation
6–8 September 2023 | ETH Zurich, Switzerland

Nicola Ortelli[1,2], Matthieu de Lapparent[1], Michel Bierlaire[2]

[1] IIDE, HEIG-VD
[2] TRANSP-OR, EPFL

Results — Logit-S: Unweighted