Experiments in Operations Research are Hardly Reproducible: A Bike-Sharing Case-Study.

Thomas Barzola, Van-Dat Cung, Nicolas Gast, Vincent Jost

Experiments in Operations Research are Hardly Reproducible: A Bike-Sharing Case-Study.

Thomas Barzola[1], Van-Dat Cung[1], Nicolas Gast[2], Vincent Jost[1]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP*, G-SCOP, 38000 Grenoble, France
{Thomas.Barzola, Van-Dat.Cung, Vincent.Jost}@grenoble-inp.fr

[2] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
nicolas.gast@inria.fr

Mots-clés : *Reproducible Research, Operations Research, Bike-Sharing Systems.*

1 Introduction : the Reproducibility Crisis

Repeatability and reproducibility are cornerstones of the scientific process, necessary to avoid-dissemination of flawed results : The result of an experiment can be considered to be part of the scientific knowledge only if others can reproduce (or invalid) this results. In [1], Baker et al. reported a *reproducibility crisis*, by showing that more than 70% of tested researchers were not able to reproduce the work from other researcher while 50% were not able to reproduce their own work. In computational science, where experimental results consist in running computer programs, ensuring reproducibility seems easier as it suffices to provide all code and data to rerun the experiments. Yet, Collberg and Proebsting showed in [2] that, only few papers contain code that can actually be run. We argue that Operations Research faces the same problem : a vast majority of papers report non reproducible experiments (probably because of one of the excuses listed in [2]).

Many researchers do believe that providing a detailed description of the experiments and the algorithms used is sufficient to guarantee reproducibility. In this paper, we argue that this is largely false. To assert this, we tried to reproduce the experiments of [3]. The authors of this paper were aware of the need for their work to be reproducible : they made their data available in one of the authors' website, and they provide a detailed description of their methodology (only the code is missing). Nevertheless, despite all the care they took, we were not able to reproduce their work and our numerical findings are significantly different from theirs. Without their code, we cannot be sure if there is a bug (in their implementation or in ours) or a difference in the interpretation of the model. This raises a number of ethical questions for the community : what is the validity of science if numerical results cannot be trusted ? Instead of developing new methodology, should we not spend more time reimplementing existing methods, making them available to all ? This may lead to a less productive, yet more trustworthy and reliable, science.

2 Original Work : Management of Bike-Sharing Systems

Bike-sharing systems are now deployed widely around the world. These systems are large (Shanghai system has more than 500,000 bikes) and have been the subject of many papers in the scientific literature. One of the most important questions for Operations Research in such system is management strategies : The heterogeneity of demand patterns leads to areas with high bike demands without available vehicles. This raises the question on how to reallocate

bikes from one area to another, on which several hundreds of papers have been published [5]. Yet, comparing all the proposed management strategies is a difficult task : each paper studies its own metric, on its own data, and provides figures that show that the proposed strategies work well on this particular scenario. In general no code nor data are made available.

The paper [3] is about computing the number of bikes that should be placed at the beginning of the day in each station, in order to minimize the excess-time, which is the difference between the actual journey time and the best journey time possible. As a baseline, the authors choose an initial repartition of bikes by using a method from [4]. The contribution of [3] is to propose a guided local search to improve the performance. The parameters to generate many random scenarios are provided. To generate scenarios, a random number generator is used. Since we do not have have access to the same generator and that no seed is provided in the publication, we expect to have results that should not be equals but comparable up to confidence intervals.

3 Numerical Results

We present in Table 1 the comparison of the results between the excess-time reported in [3] and the excess-time computed by our implementation. In the first two rows, we report the excess-time before the local search is conducted. Each column corresponds to a given instance. We observe that the numbers we obtained are comparable with the numbers reported in [3] but different (16% on average). After exchanging a few emails with the authors of the original work, neither them or us were able to spot a difference between our implementation and their. Yet, one could argue that the difference remains small. The last two rows compare the results after the guided local search and contain more surprising results : Our reproduction outperforms the original work. Still, we have a good confidence in our reproduction because on other indicators our reproduction gives results similar to the original work. Nonetheless, The lack of code and reproducible examples made impossible the explanation of this difference.

TAB. 1 – Excess-time for the baseline (first two rows) and after optimization (last two rows) with 95% confidence intervals when available.

	Hubway.Aug	Capital.Apr	Capital.June	Divvy.Aug	Divvy.Oct
Before optim. [3]	140.43	274.45	318.03	36.08	75.5
Our Implem.	122.45 ±1.0	229.1 ±1.8	265.1 ±2.1	28.37 ±0.6	60.62 ±0.8
After optim. [3]	49.48	46.08	46.69	14.43	17.35
Our implem.	30.42 ±0.5	31.42 ±0.5	32.51 ±0.5	10.05 ±0.25	11.36 ±0.26

Références

[1] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604) :452, 2016.

[2] Christian Collberg and Todd A Proebsting. Repeatability in computer systems research. *Communications of the ACM*, 59(3) :62–69, 2016.

[3] Sharon Datner, Tal Raviv, Michal Tzur, and Daniel Chemla. Setting inventory levels in a bike sharing network. *Transportation Science*, 53(1) :62–76, 2019.

[4] Mor Kaspi, Tal Raviv, and Michal Tzur. Parking reservation policies in one-way vehicle sharing systems. *Transportation Research Part B : Methodological*, 62 :35–50, 2014.

[5] C. S. Shui and W. Y. Szeto. A review of bicycle-sharing service planning problems. *Transportation Research Part C : Emerging Technologies*, 117 :102648, August 2020.