Replicability and Reproducibility in Evolutionary Optimization

Luís Paquete

paquete@dei.uc.pt
http://www.uc.pt/go/paquete

University of Coimbra, Portugal

Manuel López-Ibáñez

manuel.lopez-ibanez@uma.es
http://lopez-ibanez.eu

University of Málaga, Spain

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3461405

Main reference

Manuel López-Ibáñez, Juergen Branke and Luís Paquete.
Reproducibility in Evolutionary Computation.
Arxiv preprint arXiv:20102.03380 [cs.AI], 2021.
https://arxiv.org/abs/2102.03380

Instructors

Luís Paquete is Associate Professor at the Department of Informatics Engineering, University of Coimbra, Portugal. He received his Ph.D. in Computer Science from the Technical University of Darmstadt, Germany, in 2005 and a M.S. in Systems Engineering and Computer Science from the University of Algarve, Portugal, in 2001. His research interest is mainly focused on exact and heuristic solution methods for multiobjective combinatorial optimization problems. He is in editorial board of Operations Research Perspectives and Area Editor at ACM Transactions on Evolutionary Learning and Optimization.

http://www.uc.pt/go/paquete

Manuel López-Ibáñez is a "Beatriz Galindo" Senior Distinguished Researcher at University of Málaga since 2020 and a Senior Lecturer (Assistant Professor) at the Alliance Manchester Business School, University of Manchester, UK, since 2015. He received the M.S. degree in computer science from the University of Granada, Granada, Spain, in 2004, and the Ph.D. degree from Edinburgh Napier University, U.K., in 2009. Between 2011 and 2015, he was a Postdoctoral Researcher of the Belgian F.R.S.-FNRS at the IRIDIA laboratory in the Université Libre de Bruxelles (ULB), Brussels, Belgium.

http://lopez-ibanez.eu

Part I

Overview

EC as a scientific discipline

> *Computer science is in part a* ***scientific*** *discipline concerned with the empirical study of a class of phenomena, in part a* ***mathematical*** *discipline concerned with the formal properties of certain classes of abstract structures, and in part a* ***technological*** *discipline concerned with the cost-effective design and construction of commercially and socially valuable products*

[Wegner, 1976]

EC as an empirical scientific discipline

Scientific Method

1. Observe a phenomenon
   ☞ EAX crossover shows local optimisation in the TSP
   [Nagata & Kobayashi, 1997]
2. Construct a hypothesis
   ☞ EA + EAX produces better solutions for the TSP
3. Conduct an experiment
4. Draw conclusion about hypothesis:
   either provisionally accepted or it is *falsified*
   (with some statistical confidence)

Falsifiability + Reproducibility

⇒ build research community consensus

⇒ "*Laws of qualitative structure*" [Newell & Simon, 1976]

Why reproducibility?

- Falsifiability and community consensus (Scientific Method)
- Building upon the work of others
  ☞ Typical first step: reproduce previous results
- Quality control and error correction
  - *"Most published research findings are false"* [Ioannidis, 2005]
    ⇒ Reproducibility crisis
  - Yet, very few corrections (Errata) published
    ⇒ Lack of reproducibility studies?

Reproducibility crisis in EC?

- *"Most published research findings are false"* [Ioannidis, 2005]
- Nature survey of 1 500 researchers: [Baker, 2016]
  +70% failed to reproduce another researcher's experiments
  +50% have failed to reproduce their own previous results
- Signs that the situation is CS is no better [Cockburn et al., 2020]
- In EC:
  - Very few published cases [Sörensen et al., 2017]
  - But no reason to think EC is special

What is Reproducibility?

- No consensus in terminology [Claerbout & Karrenbach, 1992] [Plesser, 2018]
- ACM distinguishes between:
  Repeatability, *Reproducibility* and *Replicability*
- López-Ibáñez, Branke, and Paquete [2021] define the terms more precisely and distinguish between:
  Repeatability, *Reproducibility*, *Replicability* and *Generalisability*

Terminology

Artifact [ACM, 2020]

"A digital object that was either created by the authors to be used as part of the study or generated by the experiment itself"

algorithm implementations, benchmark instances, data pre/post-processing scripts, ...

Measurement [López-Ibáñez, Branke, and Paquete, 2021]

"data that results from an experiment"

- measures of quality, computational effort, etc.
- NOT summary statistics

ACM's Terminology [ACM, 2020]

Repeatability (Same team, same experimental setup)

"*The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials.*"

Reproducibility (Different team, same experimental setup)

Replicability (Different team, different experimental setup)

ACM's Terminology [ACM, 2020]

Repeatability (Same team, same experimental setup)

Reproducibility (Different team, same experimental setup)

"*The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials.* [...] *[A]n independent group can obtain the same result using the author's own artifacts.*"

Replicability (Different team, different experimental setup)

ACM's Terminology [ACM, 2020]

Association for Computing Machinery

Repeatability (Same team, same experimental setup)

Reproducibility (Different team, same experimental setup)

Replicability (Different team, different experimental setup)

> *The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials.*[...]
> *[A]n independent group can obtain the same result using artifacts which they develop completely independently.*

Dimensions of reproducibility

- *Artifacts:* Re-use of the original artifacts should allow to repeat the exact same experiments as described in the original publication
- *Random factor:*
  - the experiment only evaluates a random sample
  - the experimental claim applies to range or distribution
  - Random seeds
- *Fixed factor:*
  - the experiment only evaluates specific chosen values
  - experimental claim only supported for those specific values
  - parameter settings, benchmark problems, computational budget

  ... *unless randomized*

Proposed terminology of reproducibility studies [López-Ibáñez, Branke, and Paquete, 2021]

Label	Artifacts	Random factors	Fixed factors	Comment
Repeatability	Original	Original	Original	Exactly repeat the original experiment, generating precisely the same results.
Reproducibility	Original	New	Original	Test whether the original results were dependent on specific values of random factors and, hence, only a statistical anomaly.
Replicability	New	New	Original	Test whether it is possible to independently reach the same conclusion without relying on original artifacts.
Generalisability	Original or New	New	New	Test whether the conclusion extends beyond the experimental setup of the original paper. When new artifacts are used, generalisability should come after a replicability study.

Obstacles to reproducibility

- Cultural obstacles
- Technical obstacles

Cultural Obstacles

- Disincentives to publish artifacts
  - ✘ Additional effort ⇒ Fewer papers
  - ✘ Error detection ⇒ Rejection / retraction
  - ✘ Not *required* by journals *pre*-publication
- Difficulty to publish reproducibility studies
  - ✘ More effort than a new algorithm / survey
  - ✘ Low chances of publication
  - ✘ Biases against both negative results and corrections
- Insufficient description
  - ✘ No or bad artifacts ⇒ reproducibility impossible
  - ✘ Paper often not enough for replication
  - ☞ "Obsolete" and correct code is **better** than no / incorrect code

Technical Obstacles

- ✘ Restrictive licenses, privacy and commercially sensitive data, etc
- ✘ Binary only artifacts (no source code):
  ☞ "Obsolete" source code is **better** than "working" black-box executable
- ✘ Unreproducible or unreplicable computational environment
  any difference may distort CPU-time, RNG, floating-point, etc
- ✘ Prohibitive or unavailable computational resources
  Years of CPU time or specific hardware (GPUs)
- Verification of artifacts
  - ✘ Manual verification ⇒ Tremendous effort
  - ✘ Lack of re-implementations ⇒ Error propagation [Brockhoff, 2015]

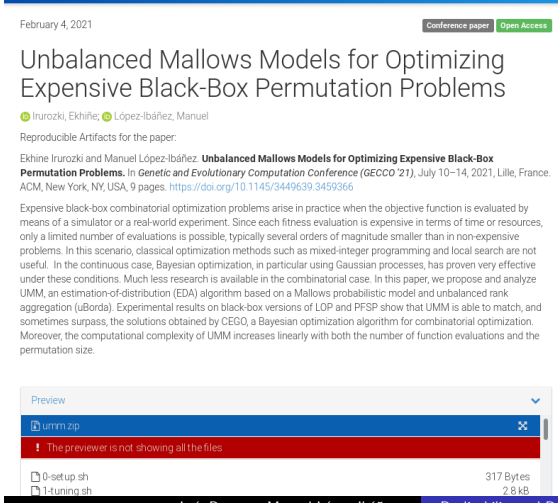
Part II

Guidelines and Tools

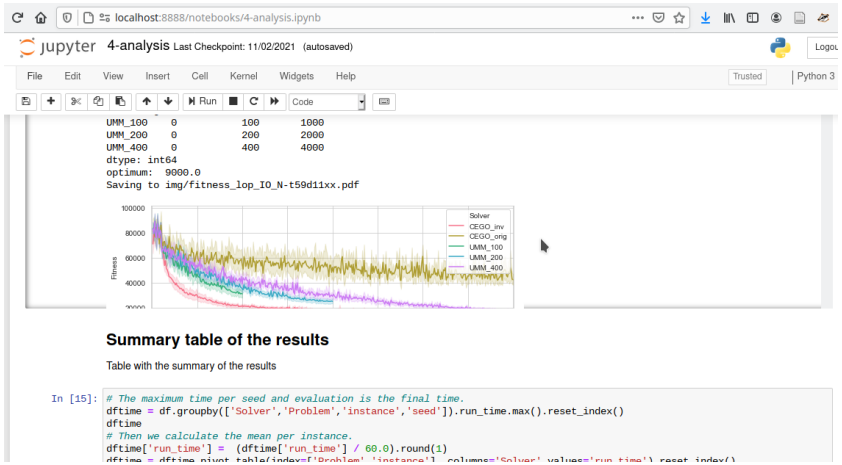
Ensuring reproducibility: Artifacts

- Permanently accessible (ACM badge *artifact available*):
  - ✘ Personal / research group webpages repositories
  - ✔ Git tag / SHA commit: https://github.com/NEO-Research-Group/irace-sumo/tree/62304739940199b3326cf8b34837c540cad6a68d
  - ✔✔ Figshare (https://www.figshare.com), Zenodo (https://www.zenodo.org) Open Science Foundation (https://www.osf.io)
- Complete:
  - ✔ All source code and input data: Pre-processing code, Algorithm code, Analysis code and Presentation code
  - ✔✔ Step-by-step documentation, flexible reproduction scripts and raw intermediate data
  - ✔✔✔ Decision vectors (actual solutions), testsuite, independent solution checkers, ...
- Useful: Open-source license and open-data formats

Zenodo Example https://doi.org/10.5281/zenodo.4500973



Jupyter Python notebook https://doi.org/10.5281/zenodo.4500973



Ensuring reproducibility: Detailed experimental conditions

- ✘ Many results are sensitive to computational platform:

  CPU speed, cache sizes, floating-point arithmetic, library bugs (Linux's `pow()` bug 13932), ...

  - ✔ Virtual machines, containers docker, platforms CODE OCEAN, OSF
  - ✔ Provide *calibration benchmark* running times

- ✘ Hidden / unfair parameter tuning:
  - Report: Effort, Domains, Training problem instances
  - Parameter tuning procedure should be reproducible:
    - ✔ Design of Experiments [Montgomery, 2012]
    - ✔ Automatic configuration tools, e.g., irace [López-Ibáñez et al., 2016]

A checklist for reproducibility (1): Artifacts

- ☐ Step-by-step documentation to reproduce the experiment AND analysis
- ☐ All source code
- ☐ All input data: problem instances, random seeds, ...
- ☐ Analysis and presentation scripts
- ☐ Raw generated data (objective and decision vectors)
- ☐ Testsuite and solution checker
- ☐ Computation time calibration code and running times
- ☐ Open-data formats (✔ CSV ✔ MySQL ✘ Excel, ✘ Oracle, etc.)
- ☐ Open-source license (reading, distributing, running and reusing)
- ☐ Permanent link / DOI to specific version ~~GitHub repo~~
- ☐ Long-term (permanently) accessible repository ~~Personal website~~

A checklist for reproducibility (2) Report / Document

- ☐ Relevant hardware details (CPU details, memory / cache sizes)
  - ☐ Provide a container (e.g., Docker)
  - ☐ Provide link to virtual platform (e.g., Code Ocean)
  - ☐ Provide reviewer access to special hardware (e.g., GPUs)
- ☐ Precise versions of any additional software, packages, simulators, compilers / interpreters, and OS
- ☐ (Hyper-)parameters, including types and domains
- ☐ Parameter tuning process (also reproducible)
- ☐ Separate problem instances for development/tuning and for benchmarking / hypothesis testing
- ☐ Confidence intervals (or p-values), size effects

Encouraging reproducibility efforts

- Journals should *require* that artifacts are provided to reviewers
  - ☞ *Mathematical Programming Computation (Springer)* requires source code
- (Ideally) Journals adopt the *Transparency and Openness Promotion (TOP)* guidelines: [Nosek et al., 2015; Stodden et al., 2016]
  - Reproducibility checks
  - Independent replication
  - . . . among other requirements before publication
- ACM badges provide a way to recognise different degrees of reproducibility

Reproducibility at ACM TELO

1. Authors request a reproducibility level for their work
2. A member of the reproducibility board at ACM TELO checks whether the work is at the requested level and may interact with the authors to improve it in terms of reproducibility
3. Once approved, a *reproducibility badge* is provided

ACM Badges

Artifacts available

Artifacts are permanently available for retrieval

ACM Badges

Artifacts evaluated
Functional

The artifacts are well documented and (functional) requirements are met

Artifacts evaluated
Reusable

Same as above plus careful documentation in order to allow reusability

ACM Badges

Results validated
Reproduced

The main results are obtained using the author-supplied artifacts

Results validated
Replicated

The main results are obtained without the use of author-supplied artifacts (not yet at ACM TELO)

Conclusions

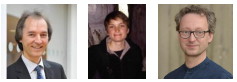
- Reproducibility implies a *cultural shift*
- Authors, journals editors, conference organisers and funding institutes should be active players on promoting this shift
- Extra effort is rewarded with *faster scientific progress* and *higher reputation* for the field as a whole
- But is it such amount of effort? It only implies adopting good practices of verification, documentation and reporting

At GECCO, in the near future: "where is your code?"

Acknowledgments

The tutorial has been benefited from collaborations and discussions with our colleagues:

Juergen Branke, Carola Doerr, Mike Preuss



The research leading to the results presented here has received funding from diverse projects:

- National funds through the FCT - Foundation for Science and Technology, I.P. within the scope of the project CISUC – UID/CEC/00326/2020.
- M. López-Ibáñez is a "Beatriz Galindo" Senior Distinguished Researcher (BEAGAL 18/00053) funded by the Spanish Ministry of Science and Innovation (MICINN).

References I

ACM. Artifact review and badging version 1.1. https://www.acm.org/publications/policies/artifact-review-and-badging-current, Aug. 2020.

M. Baker. Is there a reproducibility crisis? *Nature*, 533:452–454, 2016.

D. Brockhoff. A bug in the multiobjective optimizer IBEA: Salutary lessons for code release and a performance re-assessment. In A. Gaspar-Cunha, C. H. Antunes, and C. A. Coello Coello, editors, *Evolutionary Multi-criterion Optimization, EMO 2015 Part I*, volume 9018 of *Lecture Notes in Computer Science*, pages 187–201. Springer, Heidelberg, Germany, 2015. doi: 10.1007/978-3-319-15934-8_13.

J. Claerbout and M. Karrenbach. Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992*, pages 601–604. Society of Exploration Geophysicists, 1992. doi: 10.1190/1.1822162.

A. Cockburn, P. Dragicevic, L. Besançon, and C. Gutwin. Threats of a replication crisis in empirical computer science. *Communications of the ACM*, 63(8):70–79, July 2020. doi: 10.1145/3360311.

J. P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005. doi: 10.1371/journal.pmed.0020124.

M. López-Ibáñez, J. Dubois-Lacoste, L. Pérez Cáceres, T. Stützle, and M. Birattari. The irace package: Iterated racing for automatic algorithm configuration. *Operations Research Perspectives*, 3:43–58, 2016. doi: 10.1016/j.orp.2016.09.002.

M. López-Ibáñez, J. Branke, and L. Paquete. Reproducibility in evolutionary computation. *Arxiv preprint arXiv:20102.03380 [cs.AI]*, 2021. URL https://arxiv.org/abs/2102.03380.

D. C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, New York, NY, 8th edition, 2012.

Y. Nagata and S. Kobayashi. Edge assembly crossover: A high-power genetic algorithm for the traveling salesman problem. In T. Bäck, editor, *ICGA*, pages 450–457. Morgan Kaufmann Publishers, San Francisco, CA, 1997.

A. Newell and H. A. Simon. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3): 113–126, Mar. 1976. ISSN 0001-0782. doi: 10.1145/360018.360022.

B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. L. Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni. Promoting an open research culture. *Science*, 348(6242):1422–1425, June 2015. doi: 10.1126/science.aab2374.

References II

H. E. Plesser. Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, 11, Jan. 2018. doi: 10.3389/fninf.2017.00076.

K. Sörensen, F. Arnold, and D. Palhazi Cuervo. A critical analysis of the "improved clarke and wright savings algorithm". *International Transactions in Operational Research*, 26(1):54–63, 2017. doi: 10.1111/itor.12443.

V. Stodden, M. McNutt, D. H. Bailey, E. Deelman, Y. Gil, B. Hanson, M. A. Heroux, J. P. A. Ioannidis, and M. Taufer. Enhancing reproducibility for computational methods. *Science*, 354(6317):1240–1241, Dec. 2016. doi: 10.1126/science.aah6168.

P. Wegner. Research paradigms in computer science. In *ICSE'76: Proceedings of the 2nd international conference on Software engineering*, pages 322–330, Oct. 1976.

Replicability and Reproducibility in Evolutionary Optimization

Luís Paquete

paquete@dei.uc.pt
http://www.uc.pt/go/paquete

University of Coimbra, Portugal

Manuel López-Ibáñez

manuel.lopez-ibanez@uma.es
http://lopez-ibanez.eu

University of Málaga, Spain