Code Sharing Is Associated with Research Impact in Image Processing

In computational sciences such as image processing, publishing usually isn't enough to allow other researchers to verify results. Often, supplementary materials such as source code and measurement data are required. Yet most researchers choose not to make their code available because of the extra time required to prepare it. Are such efforts actually worthwhile, though?

1521-9615/12/$31.00 © 2012 IEEE
Copublished by the IEEE CS and the AIP

Patrick Vandewalle

How often have you attempted to implement and reproduce the results of another person's published paper? And when doing so, was this a straightforward process, similar to following a cookbook recipe, or rather a lengthy and painful endeavor? In my personal experience, it's unfortunately too common that such a reimplementation is a complex process, with many pitfalls. Parameters or initialization procedures are omitted, or certain pieces of an algorithm can be understood in multiple ways. Moreover, at the end of the process, I never felt sure that my implementation was the same as the author's—I always worried that I had forgotten something, or that my implementation didn't perform as well

Similarly, when writing an article, I often tend to forget to describe such "details" myself. I'm too excited about my latest theory, analysis, or algorithm, and don't want to let the article's flow be disrupted by practical implementation issues. This is even more the case when hard page limits are imposed. Because of time pressure, we researchers often even forget to note the precise settings by which we obtained a figure's nice results. This makes it (almost) impossible, even for us as authors, to repeat the same experiments with the same results a year after the paper was written.

Yet, you would expect that in our field of computational sciences, it should be easy to share not only the information written down in the paper, but also the whole software environment in which the experiments were performed. A simple way of doing this could be to wrap all the code and data in an archive and make it available online. Smarter and more robust ways of making environments available to other researchers are discussed in other articles in this special issue. This way of working is generally called *reproducible research*.[1,2] When researchers publish in this manner, they share the whole research environment from which they obtained their results. In practice, this typically means the software code and data or measurements, along with sufficient information about the platform (such as version numbers and parameter settings), are posted online.

When discussing research methods and reproducibility with our signal- and image-processing colleagues, there's wide agreement that these basic principles of the scientific method should be followed: results shouldn't be one-of-a-kind—they should be reproducible; and a paper should sufficiently describe the presented research results such that a colleague can fully understand the results and how they were obtained. At the same time, it's rare that papers offer supplementary material (such as code) online.

The most important obstacles for researchers in making code and data available online are the time required to do this and the lack of a direct benefit for the authors and their careers.[3] However, although it's undoubtedly true that the work invested in making code available online isn't as highly regarded as an extra publication, I argue that there could be a clear benefit to authors who do share their code online: a chance of increased impact. To illustrate this, I present two associative analyses and discuss the results. The code and data used in these analyses are available online at http://rr.epfl.ch/37.

Why Share Your Code?

From my experience, I see multiple benefits of making my research reproducible. A first example is the high number of downloads that my colleagues and I receive for reproducible papers and the related code and data. For instance, during the first six months after publishing our red-eye removal paper (for which a Java applet is available),[4] it was the top download at the École Polytechnique Fédérale de Lausanne's (EPFL's) publication database. In another paper on super-resolution imaging[5] (for which my colleagues and I have Matlab source code available—including a GUI to compare methods), the code is still downloaded more than 200 times each month five years after publication. For those papers, we also regularly get feedback by email, which is encouraging to continue this work.

Next, code and data availability also facilitated collaboration, as it's much easier now for a collaborator to pick up our results and apply them in another domain or use a different solver for the same equations. Finally, we've received requests from other colleagues to use our algorithms in commercial applications, as well as other researchers wanting to apply our techniques for image enhancement in domains that we never envisioned.

However, the strongest possible demonstration of reproducible research's increased impact is to show that reproducible papers have more citations than their nonreproducible equivalents. Such an argument requires a large-scale controlled experiment: the relationship between reproducibility and the number of citations should be analyzed with respect to a set of control parameters, such as the journal, the number of authors, the home institution, amount of funding, and the authors' seniority. In this article, I perform two preliminary associative analyses for such a study. These show a correlation between the availability of source code and the number of citations for image-processing papers. The causality of this relation can be demonstrated only in a controlled experiment, which is a subject for future work.

Instead of checking the presented results' full reproducibility, I'll concentrate here on the availability of source code implementing the work presented in the paper. In image processing, where many papers describe new algorithms for image enhancement or analysis, source code (and pseudocode, which is often published in the paper) is extremely important when trying to reproduce results. Because such code is tightly related to the paper, it's also common practice in image processing to cite the related paper when using the source code. It's worth noting that although making source code available generally provides a big step toward results' reproducibility, the two aren't necessarily the same. Some papers might be reproducible without providing code, through the detailed description in the paper, and some code provides an implementation of the presented algorithm for use in other applications, without reproducing all of the results presented in the paper.

So we know, then, that providing the code can be highly beneficial, but how often is it provided? In earlier work on reproducible research in signal processing, my colleagues and I performed a reproducible research review study.[6] In this study, we asked reviewers a number of questions on the reproducibility of articles published in *IEEE Transactions on Image Processing* (*TIP*) in 2004. One of the questions was related to the online availability of code, which about 10 percent of the study's papers offered. Also, a few analyses of the relation between the online availability of datasets and citation counts were made recently: in the field of peace research by Nils Petter Gleditsch and Havard Strand,[7] in cancer research by Heather A. Piwowar and her colleagues,[8] and in astronomy by Edward A. Henneken and Alberto Accomazzi.[9] On the topic of open access and its relation to increased citations, many studies have already been performed (see, for example, the initial study by Steve Lawrence[10] or the online bibliography on this topic available at http://opcit.eprints.org/oacitation-biblio.html).

Table 1. Summary of the first study: *TIP* papers (2004–2006).*

Criteria	All	2004	2005	2006
Number of papers	645	134	182	329
Code available	66 (10%)	12 (9%)	19 (10%)	35 (11%)
Average citations (not RR)	41	62	45	31
Average citations (RR)	198	438	202	114
Median citations (not RR)	25	37	26	21
Median citations (RR)	76	88	111	67
Significance level	6.6 e–11	4.8 e–2	1.4 e–5	9.7 e–7

**TIP = IEEE Transactions on Image Processing.* Papers with code available online are denoted as "RR" (all others are denoted as "not RR").

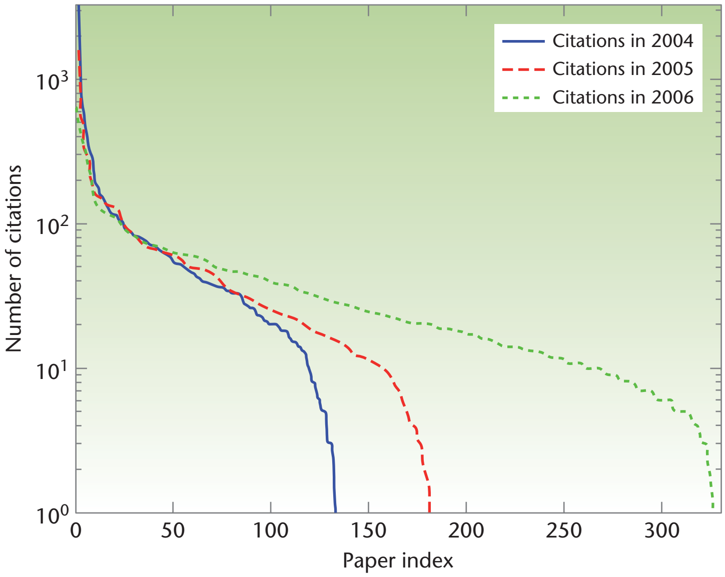

Figure 1. Sorted number of citations for each paper in the first study (with a logarithmic scale on the vertical axis), as measured using Google Scholar. The best-cited article has 3,253 citations, but the distribution of the citations also has a very long tail of poorly cited papers.

First Study: *TIP*, 2004–2006

In the first study, I analyzed all the papers published in *TIP* between 2004 and 2006. In total, I analyzed 645 papers (134 for 2004, 182 for 2005, and 329 for 2006). For each paper, I searched for available source code. I did this by first scanning the article for Web links and checking those. Next, I performed an Internet search (with Google) using the title (in quotation marks) and "source" as search terms. I analyzed the first page of search results, looking further among those links to see if relevant results, such as an author's webpage, showed up. In total, I found code for 66 papers, representing 10 percent of the papers (see Table 1).

For the citation counts, I used Google Scholar. A similar effect is obtained with Web of Science citations, but at typically lower citation rates. Web of Science tends to be more selective in counting citations. I should also remark that I didn't discard self-citations. The best-cited article in the analysis has 3,253 citations, but the distribution of the citations also has a very long tail of poorly cited papers (see Figure 1).

Next, I split the set of papers into two categories: those with and without code available. As Table 1 shows, the average number of citations increases with a factor of 4.8, from 41 citations for the papers without code available to 198 citations for the papers that have code available. However, the average citation counts are strongly influenced by a few highly cited papers. We therefore also computed the median number of citations. The median number of citations for the papers without code online is 25, compared to 76 for the papers with code available online, showing an increase with a factor of 3. The significance of this difference in medians is tested using a Mann-Whitney U-test, and shows indeed that the median of the papers with code available online is significantly higher than for the papers with no code online ($p =$ 6.6 e–11). Table 1 shows separate results per year, including the p-values for separate significance tests on the data per year. Figure 2 shows a box plot per year for the two sets (with a logarithmic scale on the vertical axis). As you can see, for both the combined data and for each individual analyzed year, there's a significant difference ($p < 0.05$) between the median number of citations for papers that have code available and papers that don't have code available (see the bottom row of Table 1 for p significance values).

As a further test to see whether a few highly cited papers with code available were responsible for these results, I also performed significance tests (again using the Mann-Whitney U-test) on the data after removing an increasing number of top-cited papers from the set with code available.

The $p < 0.05$ null hypothesis is rejected until as many as 26 papers are removed from this set, which is almost half of those papers (the total size is 66). This illustrates that the results are not solely determined by a subset of highly cited papers.

Second Study: Most Highly Cited Papers, 2004–2008

The most highly cited papers for a set of high-profile journals in signal and image processing are the subject of the second study. I analyzed papers published in *TIP*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*TPAMI*), and *IEEE Transactions on Signal Processing* (*TSP*). For each journal and each analyzed year (2004–2008), I selected the three most highly cited articles per year, resulting in a total of 45 articles. Note that the number of selected articles per year is chosen rather arbitrarily, as no clear cut-off could be determined from the citation results (see, for example, Figure 1).

For each paper in this study, I searched again for available code. Table 2 summarizes the results. It's remarkable to see that for both *TIP* and *TPAMI*, code is available for 13 out of the 15 analyzed articles (87 percent). This can be compared with the overall code availability of 10 percent for all *TIP* papers in the first study. However, for *TSP*, only two out of the 15 most highly cited papers have code available online (13 percent). A possible explanation for this difference could be that articles in *TSP* generally present theory and algorithms based on a model of the (typically one-dimensional) signals, and don't put a lot of emphasis on results for real data, whereas articles in *TIP* and *TPAMI* do. Another possible explanation is a difference in standards and expected publishing methods between the communities. Because these results are exploratory and obtained from a small set of papers, this difference should be further analyzed in a larger study, including a broader range of journals and more influencing factors.

Contextualizing the Results

The results shown in both studies (except for the *TSP* analysis) indicate that papers with code available online are more highly cited than those without. In the first study, the median number of citations for the papers studied increases with a factor of 3 when code is available online; this difference is significant ($p = 6.6$ e–11). The second study shows that for the best-cited image-processing papers (the top three in *TIP*), 87 percent had code available online (compared to 10 percent as a global average). There's also a large difference among journals, with 87 percent of papers having code available online for *TIP* and *TPAMI*, and only 13 percent for *TSP*. There can be multiple reasons for this difference—such as the difference between more theoretical and more applied papers, and different practices within research communities; on this point, further study is required.

I should mention that for this exploratory work, I simplified reproducibility to online code availability related to a paper. In a computational research domain such as image processing, making code available online provides a big step forward in making articles more reproducible and in making the presented results more repeatable. However, with this simplification I don't claim that papers that don't have code available

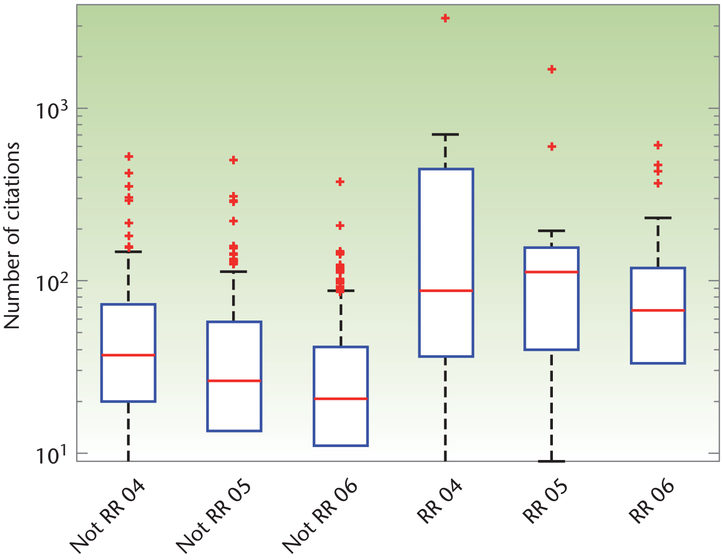

Figure 2. Box plots of the number of citations for papers in the first study with and without code available online. Papers with code available online are denoted as "RR" (all others are denoted as "not RR"). The top and bottom of the boxes indicate the 25th and 75th percentiles, respectively. The line inside the box represents the median value, and the outer bars show the extremes (without the outliers). Note the logarithmic scale on the vertical axis. The bottom bars that aren't shown extend to 0.

Table 2. Summary of the second study: top-cited papers in high-profile journals (2004–2008).*

Criteria	*TIP*	*TPAMI*	*TSP*
Number of papers	15	15	15
Code available	13 (87%)	13 (87%)	2 (13%)

**TIP = IEEE Transactions on Image Processing; TPAMI = IEEE Transactions on Pattern Analysis and Machine Intelligence; TSP = IEEE Transactions on Signal Processing.*

online aren't reproducible. As I mentioned before, it might be that through a paper's detailed description, the results can be reproduced just as well. A theoretical paper usually doesn't require code. Conversely, reproduced results through online code might become meaningless if the code has bugs.

Also, I should clarify that for one paper in the first study, source code was included in the paper itself. While I didn't find this source code separately online, I considered it as part of the set where source code was available. As can be seen in the open-access citation studies (such as the one by Lawrence[10]), papers with an online version freely available have an increased number of citations. I didn't take this into account in my analyses by adding the open-access availability as another variable. Papers that have the code available generally also have an online version of the paper. The increased citation effect is therefore the combined effect of the open-access availability of the paper and the code.

This article's results show a correlation between the online availability of code and the number of citations of the studied papers. This could indicate a causal relation between the two. There are, however, other possible explanations of the results. As is sometimes argued for open-access papers, it's likely that authors spend more effort making code available for their best papers (self-selection bias). In this scenario, the online code isn't the reason for the increased number of citations, but rather the anticipated consequence. Similarly, I noticed that for some papers, the code was made available by researchers other than the authors. This is generally done by colleagues after the work has become popular.

The data described in this article provides a snapshot of code availability and citation counts at a specific moment in time. For some papers, code might have been added only later, while for others, the link to code provided in the paper has become invalid. I've encountered several examples of this. The need for code to repeat a paper's results also depends on the topic and type of paper (for example, theoretical or experimental). To verify whether a causal relation exists between the code's availability and the number of citations, a controlled experiment should be performed at a larger scale, taking all these issues into account as control parameters, together with other parameters influencing the number of citations such as the journal, the number of authors, the home institution, the amount of funding, and the authors' seniority.

Although these results can be a motivation for making code (and data) available online, numerous Internet search queries have also clearly raised some obstacles. First, the lifetime of most webpages is extremely short. Websites are renewed, researchers move from one institute to another, and software changes; in many cases, this causes an end to the code's availability. It therefore seems like a good idea to make the code and data available together with the publication in institutional repositories. Typically, such repositories are set up and maintained with a long-term perspective. An example can be found at http://rr.epfl.ch, together with some setup information. Second, for industrial research (or industry-funded research), it's generally difficult to make the code available online. A similar argument holds for researchers who want to create a startup based on their research results. From my own experience in industrial research, I would strongly recommend establishing a standard of making results reproducible internally.

Acknowledgments

I thank the guest editors for organizing this special issue, and the anonymous reviewers for their feedback. I'm also grateful to Jelte Vink for his help on automating parts of the analysis, and to Martin Vetterli and Victoria Stodden for the inspiring exchange of ideas on reproducible research and statistical analysis, and for their feedback on drafts of this article.

References

1. M. Schwab, M. Karrenbach, and J. Claerbout, "Making Scientific Computations Reproducible," *Computing in Science & Eng.*, vol. 2, no. 6, 2000, pp. 61–67; http://dx.doi.org/10.1109/5992.881708.
2. J.B. Buckheit and D.L. Donoho, *WaveLab and Reproducible Research*, tech. report 474, Dept. of Statistics, Stanford Univ., 1995; http://www-stat.stanford.edu/~donoho/Reports/1995/wavelab.pdf.
3. V. Stodden, "The Scientific Method in Practice: Reproducibility in the Computational Sciences," *MIT Sloan Research Paper Series*, no. 4773–10, 2010; http://ssrn.com/abstract=1550193.
4. F. Volken, J. Terrier, and P. Vandewalle, "Automatic Red-Eye Removal Based on Sclera and Skin Tone

Detection," *Proc. Imaging Science and Technology 3rd European Conf. Color in Graphics, Imaging, and Vision* (CGIV), Soc. for Imaging Science and Technology, 2006, pp. 359–364; http://infoscience.epfl.ch/record/63852/files/VolkenTV06.pdf.

5. P. Vandewalle, S. Süsstrunk, and M. Vetterli, "A Frequency Domain Approach to Registration of Aliased Images with Application to Superresolution," *Eurasip J. Applied Signal Processing*, Hindawi Publishing, vol. 2006, article ID 71459; doi:10.1155/ASP/2006/71459.
6. P. Vandewalle, J. Kovacevic, and M. Vetterli, "Reproducible Research in Signal Processing," *IEEE Signal Processing*, vol. 26, no. 3, 2009, pp. 37–47; doi:10.1109/MSP.2009.932122.
7. N.P. Gleditsch and H. Strand, "Posting Your Data: Will You Be Scooped or Will You Be Famous?" *Int'l Studies Perspectives*, vol. 4, no. 1, 2003, pp. 89–97.
8. H.A. Piwowar, R.S. Day, and D.B. Fridsma, "Sharing Detailed Research Data Is Associated with Increased Citation Rate," *PLoS ONE*, vol. 2, no. 3, 2007, p. e308; www.plosone.org/article/info:doi/10.1371/journal.pone.0000308.
9. E.A. Henneken and A. Accomazzi, "Linking to Data—Effect on Citation Rates in Astronomy," *Proc. Astronomical Data Analysis Software and Systems*, Astronomical Soc. of the Pacific, 2011; http://arxiv.org/pdf/1111.3618v1.pdf.
10. S. Lawrence, "Free Online Availability Substantially Increases a Paper's Impact," *Nature*, vol. 411, no. 6837, 2001, p. 521; www.nature.com/nature/journal/v411/n6837/full/411521a0.html.

***Patrick Vandewalle** is a senior scientist at Philips Research, Eindhoven, The Netherlands. His research interests are in signal and image processing, sampling, digital photography, 3D, and reproducible research. Vandewalle has a PhD from the École Polytechnique Fédérale de Lausanne (EPFL). He's a member of IEEE. Contact him at patrick.vandewalle@a3.epfl.ch or visit the website that he maintains (www.reproducibleresearch.org).*