Economics
The Open-Access, Open-Assessment E-Journal

Vol. 12, 2018-34 | June 06, 2018 | http://dx.doi.org/10.5018/economics-ejournal.ja.2018-34

Replicating "Predicting the present with Google trends" by Hyunyoung Choi and Hal Varian (The Economic Record, 2012)

Tom Coupé

Abstract
In this paper, the author describes different ways in which one can replicate a paper and illustrate them by applying them to the study by Choi and Varian (Predicting the Present with Google Trends, *The Economic Record* 2012).

(Published in Special Issue The practice of replication)

JEL A1 C1 C8 C53
Keywords Replication

Authors
Tom Coupé, University of Canterbury, Christchurch, New Zealand, tom.coupe@canterbury.ac.nz

The author thanks the editor and 3 referees for very helpful comments on earlier drafts of this paper.

Citation Tom Coupé (2018). Replicating "Predicting the present with Google trends" by Hyunyoung Choi and Hal Varian (The Economic Record, 2012). *Economics: The Open-Access, Open-Assessment E-Journal*, 12 (2018-34): 1–8. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-34

Received August 30, 2017 **Published as Economics Discussion Paper** September 28, 2017
Accepted May 29, 2018 **Published** June 6, 2018

Introduction – the three types of replication

In this paper, I describe different ways in which one can replicate a paper and illustrate them by applying them to the study by Choi and Varian (2012).

The most basic type of replication (from here, a type I replication) checks whether one can replicate the results of a paper using the data and code provided by the authors. One would expect that, with the help of the authors' data and code, one can exactly replicate the numerical values reported in a paper. If this is the case, one could state that the paper is fully replicable. If not all the values can be replicated, one could compute the share of values that can be replicated exactly and state that X% of the paper can be replicated. This indicator reflects well how careful the original authors have been.

However, this might not be an ideal indicator of whether one can use this paper to build future research on. First, not all numerical values are equally important. A non-replicable R^2 is typically less important than a non-replicable coefficient estimate, and a non-replicable coefficient in an auxiliary regression is less important than a non-replicable coefficient in the main regression. Second, small deviations between the published value and the replication value often will not materially affect the conclusions of the paper. Hence, one might want to create a weighted index in which the weights reflect both the importance of the numerical values and the relative difference between the published values and the replication values.

Admittedly, defining these weights would involve a certain degree of subjectivity, but one could for example give a zero weight to all the numerical values except for those that are used in the conclusions. This would allow one to distinguish between the replicability of the paper as a whole and the replicability of the conclusions. If the conclusions are 100% replicable, one might not worry too much that the paper as a whole is not 100% replicable. If the conclusion is not 100% replicable, one could compute the average of the absolute value of the percentage difference between the published values that are used in the conclusion and their replicated values, computing a 'replicability gap' which indicates the extent of the gap between the published conclusion and the conclusion based on the replication. For example, the published values used in the conclusion on average could be 10% different from the replication values of the conclusion. While not perfect, these two numbers combined, the share of the replicable values used in the conclusion and the gap between the published conclusion and its replication, will, in most cases, give a reasonable idea of the extent to which one can trust the conclusion and build on this paper for future research.

Rather than using the data and code provided by the authors, one could also try to collect the data and write the code based on the description in the paper that one is attempting to replicate (from here, a Type II replication). Ideally, such a replication would be able to recreate the same variables, analyse the same specification and lead to exactly the same numerical estimates. In this Type II replication, one can distinguish between the replication of the values in the database and the replication of the paper itself. If one cannot replicate many of the values in the database, one is likely to have a hard time replicating the numerical values in the paper, but there might be cases where there is still a fairly small 'replicability gap' between the conclusion of the published paper and the replicated conclusion. On the other hand, if one has a replicable database, one might still have a big 'replicability gap' between the published conclusion and its replication.

Note that, given the likely presence of implicit assumptions, small unreported data manipulations, updated data sources or misunderstandings by the person undertaking the replication, the replication statistics computed for Type II replications are likely to be smaller than those for Type I replications.

The two above-mentioned approaches to replication are important, as these kinds of checks can provide an incentive for researchers to put more effort into avoiding mistakes, to check their own results and, ideally, to make publicly available their code, data and the details of how these were created. At the same time, these two approaches to replication are rather narrow: they only check whether one obtains similar results if one follows the same method for the same time period and the same country and the same data source. However, since most people will only read abstracts, introductions and/or conclusions, and since we all have the tendency (and wish) to extrapolate the results of studies of specific situations to general laws of economics, 'broader' replication is also needed.

These more comprehensive checks (Type III replication) constitute replication attempts of the overall conclusion of the paper that one wished to replicate rather than of the exact numerical estimates reported in the paper. That is, can one generalize conclusions across many countries, series, time periods and even regression specifications or techniques? As a consequence, the focus of the replication is no longer on whether the numerical values in the paper can be replicated exactly but rather on how the conclusions of the original paper change when one modifies certain features of the original study. Such replication attempts are important, as they provide an incentive for researchers to avoid cherry picking results and to perform various robustness checks. In addition, such checks should push researchers to be very careful in reaching their conclusion and make it clear to the reader that their results apply to a specific data set and setting rather than being seen as an illustration of an economic law that is true always and everywhere. Such comprehensive replication efforts are similar in nature to meta-analyses or literature surveys.

Illustrating the types of replication using Choi and Varian (2012)

Let me now turn to the paper that I will use to illustrate the above types of replication, the paper by Choi and Varian (2012).

Choi and Varian (2012) include four examples of macroeconomic statistics that can be forecast more accurately if time series, reflecting the search intensity of terms related to the macroeconomic statistics, are included in the regression used to forecast the macroeconomic statistics. They show, for example, that, if one augments an autoregressive model of the sales of motor vehicles and parts in the United States with series that reflect the evolution of the search intensity for 'trucks and SUVs' and 'auto insurance', both the in-sample and the out-of-sample forecast accuracy improves by about 10%.[1] Other examples focus on the forecasting of United

[1] These are categories of searches rather than specific search terms; hence, this is the evolution of the intensity of the search terms that fall into the categories 'trucks and SUVs' and 'auto insurance'. It is not clear why these two categories are used. In the paper, the authors write: 'A little experimentation shows that two of these categories, Trucks & SUVs and Automotive Insurance, significantly improve in-sample fit when added to this regression.'

States' unemployment benefit claims, visitor arrivals to Hong Kong and consumer confidence in Australia.

Choi and Varian's (2012) paper is highly cited, having over 1000 citations in Google Scholar. Other papers that use search intensity to forecast economic series include those by Ettredge et al. (2005), which focuses on unemployment, and Goel et al. (2010), which focuses on the box office revenue of films, the sales of video games and the popularity of songs. These papers have not only led to academic citations but also inspired many organizations outside academia to experiment with search intensity series to improve their forecasts (see below for details).

The paper by Choi and Varian (2012) can be replicated in the various ways described above. Table 1 compares the results reported in the paper with the results of these different replication types, focusing on the example of forecasting the sales of motor vehicles and parts in the United States.

Column (1) of Table 1 presents the results of the regression of US motor vehicles and parts sales on its lags and two indicators of search intensity, as published on page 4 of Choi and Varian's (2012) paper. Column (2) provides the results of a Type I replication, that is, the results that I obtain when using the data and code provided on Varian's website.[2] Using Choi and Varian's (2012) data and code, I obtain exactly the same results as are published. If a further check of the other examples in the paper also resulted in this conclusion, this Type I replication could be deemed to be a perfect replication.

In column 3 of Table 1, I make an attempt to collect the data and write the code based on the description in Choi and Varian's paper, rather than using the code and data that they provide on Varian's website. For papers based on Google's search intensity, this kind of replication is unlikely to lead to exactly the same numerical estimates. As Choi and Varian (2012) state in their paper: 'Note that Google Trends data is computed using a sampling method, and the results therefore vary a few per cent from day to day'. Hence, it is unlikely that, six or so years after Choi and Varian created their search intensity series, I would obtain exactly the same series.[3] Similarly, the US sales series data available from the link provided in the paper are slightly different from the data provided on Varian's website.[4] Hence, the replicability of the data is low

---

[2] http://www.sims.berkeley.edu/~hal/Papers/2011/Data.zip

[3] For the same reason, this replication will not be replicable, as on different days I obtain slightly different numbers when downloading the search intensity series.

[4] The link to the data provided in Footnote 2 of Choi and Varian (2012) refers to the link https://www.census.gov/retail/marts/www/timeseries.html, which gives the 'Adjusted Monthly Sales for Retail and Food Services' and the adjustment coefficients, which are rounded to three digits. Multiplying these two quantities results in what should be the unadjusted series used by Choi and Varian (2012). The unadjusted time series of the estimates can also be obtained from the Census bureau:

https://www.census.gov/econ/currentdata/dbsearch?program=MARTS&startYear=1992&endYear=2017&categories%5B%5D=441&dataType=SM&geoLevel=US&notAdjusted=1&submit=GET+DATA&releaseScheduleId These two 'unadjusted' series are not exactly the same but very similar; the differences could be related to rounding. Both series are slightly different from the data provided on Varian's website, possibly due to revisions of the original series. The search intensity data on Varian's website are also standardized in a different way from those available from Google Trends.

Table 1: the contribution of search intensity series to the prediction of car sales under various scenarios

	Choi–Varian (2012) paper (Jan 2004–June 2011)	Replication of Choi–Varian (2012) using their data and code (Jan 2004– June 2011)	Replication of Choi–Varian (2012) using my data and code (Jan 2004– June 2011)	Replication of Choi–Varian (2012) using my data and code and longer period (Jan 2004–July 2017)	Replication of Choi–Varian (2012) using my data and code for NZ (Jan 2004–June 2011)	Replication of Choi–Varian (2012) using my data and code for NZ, and longer period (Jan 2004–July 2017)
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	–0.45798	–0.45798	–0.6818	0.4856	1.5787*	–0.5113
lag(y, 1)	0.61947***	0.61947***	0.5298***	0.4955***	0.2984***	0.4002***
lag(y, 12)	0.42865***	0.42865***	0.3565***	0.4027***	0.3973***	0.5893***
Trucks & SUVs	1.05721***	1.05721***	0.95***	0.336***	0.0243	0.2159*
Auto Insurance	–0.52966***	–0.52966***	–0.5098***	–0.1899***	0.2428	–0.0647
Adjusted R2 with search series	0.808	0.808	0.783	0.864	0.528	0.772
Adjusted R2 without search series	0.7111	0.7111	0.714	0.847	0.482	0.767

In the table, the estimates based on an OLS regression analysis are given. *** means significant at 1% significance level, * means significant at 10% significance level. The dependent variable of regressions 1 to 4 reflects US seasonally unadjusted monthly sales in the 'motor vehicle & parts' category (see footnote 5). For regressions 5 and 6, the dependent variable is NZ car sales. Besides lagged dependent variables, the explanatory variables include indices reflecting the search intensity for words in the Trucks & SUVs and the Auto Insurance category of Google Trends for respectively the US (1–4) and NZ (5–6).

for this paper. As a consequence, the paper itself is likely to have a small share of replicable coefficients in our Type II replication. Column 3 indeed shows that, when I use the data that I downloaded from Google Trends, I obtain slightly different estimates from those published in the paper.

Choi and Varian (2012) conclude, based on their analysis of four series, as follows: 'We have found that simple seasonal AR models that include relevant Google Trends variables tend to outperform models that exclude these predictors by 5 per cent to 20 per cent.' While the share of numerical results used to come to this conclusion (the reduction in forecast error for the four examples) that is replicable is likely to be small, the fact that for the example I checked that the published coefficients and their replication are similar, both in the statistical and in the economic sense, suggests that the 'replicability gap' for this Type II replication of the Choi and Varian (2012) paper will be small.

The Type III replication similarly focuses on checking whether Choi and Varian's (2012) conclusion survives when applied to other settings than those checked in the paper.

In column 4, I extend the data set from 2011 to 2017 and find that the in-sample predictive impact of adding search intensity terms is weaker. While the coefficients of the search intensity variables are significant, the adjusted R^2 increases by less than 5%. Similarly, using Choi and Varian's model and time period on NZ car sales data shows insignificant coefficients for both search intensity variables, though the increase in the adjusted R^2 is still more than 5% (column 5). Extending the data to 2017 for NZ further shows only a negligible increase in the adjusted R^2 (column 6). Of course, these are just some counterexamples.[5] Varying the countries or time periods further, one might find that differences due to the longer period or the NZ data are the exception rather than the rule. Hence, a replication should check systematically, varying one aspect of the original study, whether adding Google search intensity series adds predictive power.

For example, one could start by collecting data for a given series for many countries and estimate the same model suggested by Choi and Varian (2012) for all the countries, that is, adding search intensity data for 'trucks and SUVs' and for 'auto insurance' to an AR(1,12) model.[6] This would allow one to investigate the extent to which there is heterogeneity in the contribution of these two search intensity series. If such heterogeneity is found, one could try to explain it. For example, in countries where there is a greater number of searches (like the US), the predictive contribution of search intensity series could be greater than in countries with a smaller number of searches (like New Zealand).

Similarly, one could vary other aspects of the Choi–Varian (2012) paper. One could continue to focus just on the US but check how changing the model affects the contribution of the search intensity series. One could, for example, include additional autoregressive terms, use series of other search queries or include other macro-economic series.[7] Alternatively, one could

---

[5] These counterexamples are only based on in-sample forecasting performance. Choi and Varian (2012) also check the out-of-sample forecast performance.

[6] One example of a step in this direction is the paper by Tuhkuri (2016), which studies how Google Search can help to predict unemployment not only at the US level but also at the level of the various different US states.

[7] See for example Li (2016), who, in addition to search intensity series, uses 29 economic variables to forecast the US jobless initial claims and employment.

vary the time period of the analysis, applying the Choi–Varian (2012) model to shorter or longer periods of data or to various non-overlapping sub-periods, or keep the period fixed but change the frequency of the data.[8] Finally, one could try predict series that were not covered by the examples in the Choi–Varian (2012) paper.[9]

It is hard to predict the outcome of this third replication type, though my guess is that Choi and Varian's (2012) results would be true not just for the examples that they tested but for some places and some time periods rather than always and everywhere. Choi and Varian's (2012) conclusion is written such that it allows for some uncertainty: 'relevant Google Trends variables **tend** [bold added] to outperform models that exclude these predictors by 5 per cent to 20 per cent'.[10] Thus, the result of the third-step replication would help to quantify the '**tend** [bold added] to outperform' from Choi and Varian's conclusion. If Type III replication reached the conclusion that adding search intensity series improves forecasts only for the cases that they describe in the paper, one could argue that this Type III replication was not successful and that their results are not externally replicable. However, just as it is unlikely that a single study will consider all possible circumstances, it is unlikely that any replication will be able to consider all possible circumstances. Hence, rather than leading to a binary conclusion (the external replication is successful or not successful), the replication could estimate how often, and in which circumstances, adding search intensity series improves forecasts by 5 to 20 per cent, or more generally present a more complete description of the distribution of the improvements in forecast accuracy.

So far, I have focused on three replication types, of which the first two types focus on the internal validity of the paper while the third type focuses on the external validity of the paper. These three types all focus on 'academic' replication, however. Replication could also leave the ivory tower and check whether a paper is replicable in 'real life'; that is, are the conclusions of a paper, after the paper has been made public, actually used by decision makers? In the case of Choi and Varian (2012), even if academic studies find that Google Search can help with forecasting, the ultimate 'replication' would involve businesses and governments actually using this knowledge in real life and, because of this, gaining a competitive advantage. That would be the 'real' proof that the predictive power of Google search intensity is not just an academic gimmick but provides forecasters with an economic, meaningful advantage. Mui (2014) writes that 'The Bank of Israel and the Bank of England incorporate Google analytics into some of their forecasts', suggesting that adding Google Search intensity might even pass this advanced replication benchmark, at least in some cases.

---

[8] Choi and Varian (2012), for example, point out that search intensity might help to predict turning points better.

[9] For example, while Choi and Varian (2012) use searches related to unemployment to forecast unemployment claims, Baker and Fradkin (2011) relate job search to extensions of unemployment payments.

[10] Additionally, one can vary several dimensions at the same time.

References

Baker, S., and Fradkin, A. (2011). What Drives Job Search? Evidence from Google Search Data. Discussion Paper No. 10-020, Stanford Institute for Economic Policy Research. https://econpapers.repec.org/paper/sipdpaper/10-020.htm

Choi, H. and Varian, H. (2012). Predicting the Present with Google Trends. *The Economic Record*, 88(s1): 2–9. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-4932.2012.00809.x

Ettredge, M., Gerdes, J. and Karuga, G. (2005). Using Web-based Search Data to Predict Macroeconomic Statistics. *Communications of the ACM*, 48(11): 87–92. https://cacm.acm.org/magazines/2005/11/6078-using-web-based-search-data-to-predict-macroeconomic-statistics/abstract

Goel, S., Hofman, J.M., Lahaie, S., Pennock, D.M. and Watts, D.J. (2010). Predicting Consumer Behavior with Web Search, *Proceedings of the National Academy of Sciences*, 7 (41): 17486–17490. http://www.pnas.org/content/107/41/17486

Li, Xinyuan (2016). Nowcasting with Big Data: is Google useful in Presence of Other Information? Policy Research Working Paper 7398, World Bank Group. https://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-7398

Mui, Q. (2014). What Search Engines Tell Governments about the Economic Here and Now. *The Washington Post*, June 14. https://www.theguardian.com/technology/2014/jun/14/big-data-google-economics-twitter

Tuhkuri, J. (2016). Forecasting Unemployment with Google Searches. ETLA Working Papers 35, The Research Institute of the Finnish Economy. https://www.etla.fi/wp-content/uploads/ETLA-Working-Papers-35.pdf