SUBOPTIMAL FINGERPRINTING?

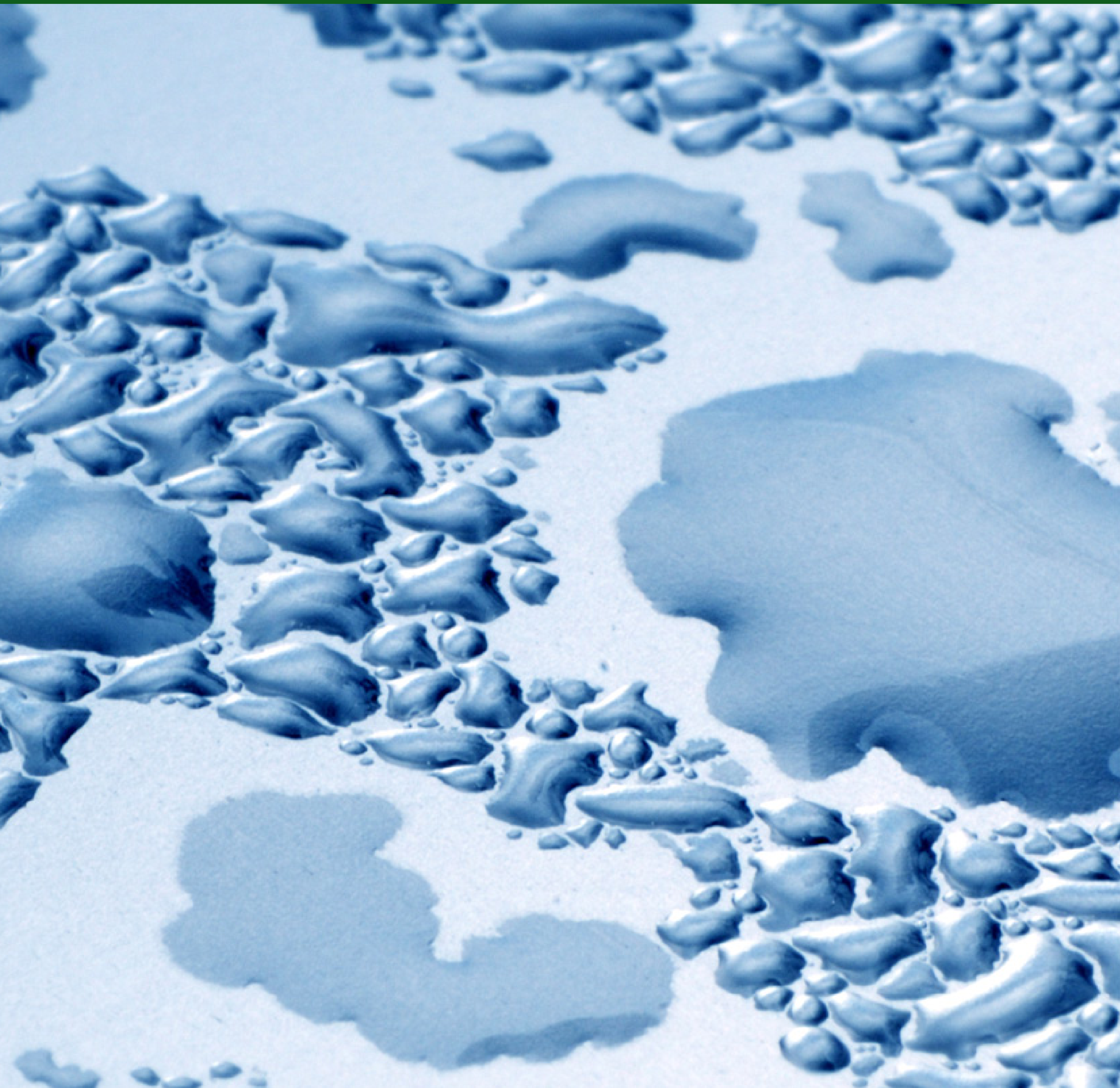
Ross McKitrick

The Global Warming Policy Foundation
GWPF Note 26

Suboptimal Fingerprinting?

Ross McKitrick

Note 26, The Global Warming Policy Foundation



Contents

About the author

Ross McKitrick is Professor of Economics at the University of Guelph, Canada. He specialises in environmental economics and policy.

Introduction

In 1999, Myles Allen and Simon Tett published an article in the journal *Climate Dynamics* (henceforth denoted 'AT99'), which formalised a procedure – optimal fingerprinting – for attributing observed climate changes to underlying causes, with a specific focus on forcing due to greenhouse gases.[1] They also proposed a method called the Residual Consistency (RC) test, for ascertaining if the statistical model was valid.

Optimal fingerprinting, which is sometimes called optimal detection, was instantly embraced and promoted by the IPCC in its 2001 Third Assessment Report (TAR),[2] and has been referenced in every IPCC Assessment Report since. TAR Appendix 12.1 was headlined 'Optimal detection is regression', and began:

> The detection technique that has been used in most 'optimal detection' studies performed to date has several equivalent representations...It has recently been recognised that it can be cast as a multiple regression problem with respect to generalised least squares (Allen and Tett, 1999; see also Hasselmann, 1993, 1997)

In 2014 a group of authors led by Jara Imbers, which included Myles Allen as coauthor, pointed to the impact the statistical method had had over the intervening years:

> The Intergovernmental Panel on Climate Change's (IPCC) 'very likely' statement that anthropogenic emissions are affecting climate is based on a statistical detection and attribution methodology that strongly depends on the characterization of internal climate variability…as simulated by [climate models].[3]

The IPCC's promotion of and reliance on optimal fingerprinting continues today.[4] It has been used in dozens and possibly hundreds of studies over the years. Wherever you begin in the literature in the field, all paths lead back to Allen and Tett (often via the follow-up paper Allen and Stott 2003[5]). Furthermore, the literature has relied almost exclusively on the RC test for checking the validity of results. So, the errors and deficiencies in the paper matter acutely, even two decades later.

I have published an article in *Climate Dynamics* showing that the optimal fingerprinting method, as set out in AT99 and the follow-up paper, is theoretically flawed and gives meaningless results.[6] It does not prove that all the results from using this method are wrong, but it does show that the basis on which they were believed to be correct is non-existent.

On logic and the implications of my findings

A careful statement of the implications of my finding must note an elementary principle of logic. We can say without fear of contradiction:

> Suppose A implies B. Then if A is true, B is true.

As an example, all dogs have fur. A beagle is a dog; therefore a beagle has fur. However, we cannot say this:

> Suppose A implies B. A is not true, therefore B is not true.

Example: all dogs have fur; a cat is not a dog, therefore a cat does not have fur. But of course we can say:

> Suppose A implies B; A is not true therefore we do not know if B is true.

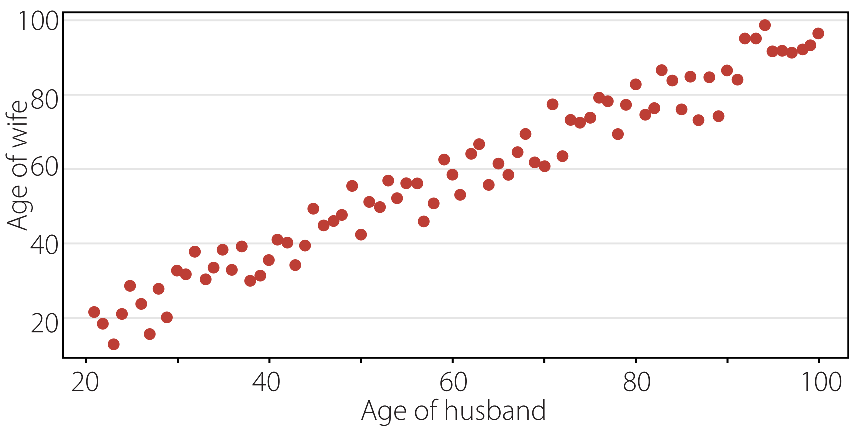
Example: all dogs have fur. A dolphin is not a dog, therefore we do not know if a dolphin has fur.

When looking at the implications of my findings, 'A' is the mathematical argument that Allen and Tett invoked to prove 'B' – the claim that their model yields unbiased and accurate results. In my critique, I showed that 'A', their mathematical argument, is erroneous. So we have no basis to say anything about 'B', and certainly not that their model yields unbiased and accurate results. The critique also applies to the RC test: it yields meaningless answers. In my article, I list the conditions needing to be proven to validate their claims about their method. I don't think it can be done, for reasons stated in the paper, but I leave open the possibility. Absent such proof, applications of their method over the past 20 years leave us uninformed about the influence of GHGs on the climate. Here I will try to explain the main elements of the statistical argument.

Regression

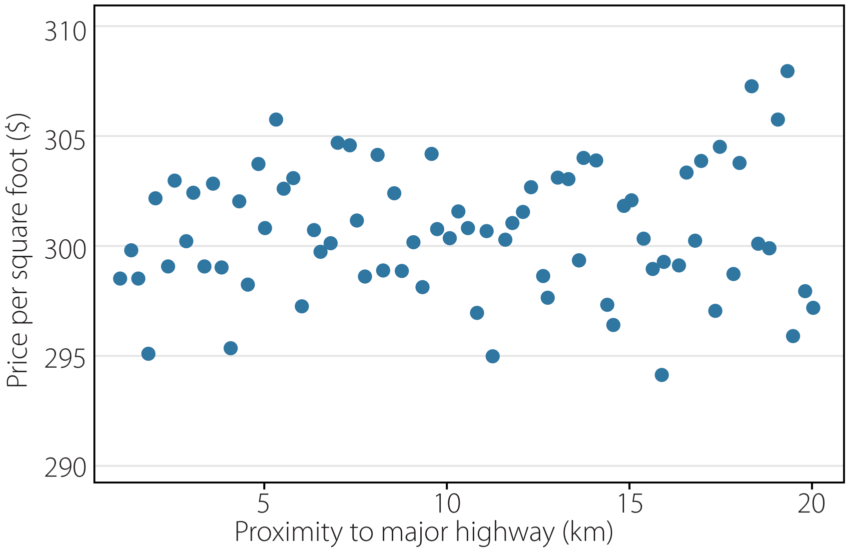
Most people with some level of scientific education are familiar with the idea of drawing a line of best fit through a scatter of data. This is called linear regression. Consider Figure 1, which shows, for a sample of married couples, the wife's age plotted against the husband's age.

Figure 1: Married couples: age of wife versus age of husband



Clearly the two are correlated: older men have older wives and vice versa. You can easily picture drawing a straight line of best fit through the points. It is customary to refer to the horizontal axis as the *x*-axis and the vertical axis as the *y*-axis. The line can be defined using two numbers: the slope and the intercept on the y-axis. If the slope is positive, higher values along the *x*-axis are associated with higher values along the *y*-axis too. This is clearly the case in the above example; any reasonable line through the sample would slope upwards. But in other cases, it is not so obvious. For example, Figure 2 shows the value of retail property in relation to its proximity to a major highway:

Figure 2: Retail property values versus proximity to highway.



Here, a line of best fit might be nearly horizontal, but might also slope up. For the purpose of picturing why statistical theory becomes important for interpreting regression analysis it is better to have in mind Figure 2 rather than Figure 1. We rarely have data where the relationship is as obvious as it is in the husband–wife example. We are more often trying to get subtle patterns out of much noisier data.

It can be particularly difficult to tell if slope lines are positive if we are working in multiple dimensions. In Figure 2 there are many other variables besides proximity to a highway that might account for variations in retail property values. If, say, there are three possible drivers of retail property values, we need to estimate the slope parameter for each, as well as the intercept.

Note that regression models can establish correlation, but correlation is not causation. Older men do not cause their wives to be older; it is just that people who marry tend to be of the same age group. If we found properties far from highways to be more valuable, it might mean distance to a highway affects property values, or it might mean that highways tend to be built on land that was less valuable for unrelated reasons. Regression models can help support interpretations of causality if there are other grounds for making such a connection, but it must be done very cautiously and only after rigorously testing whether the model has omitted important explanatory variables.

Sampling and variance

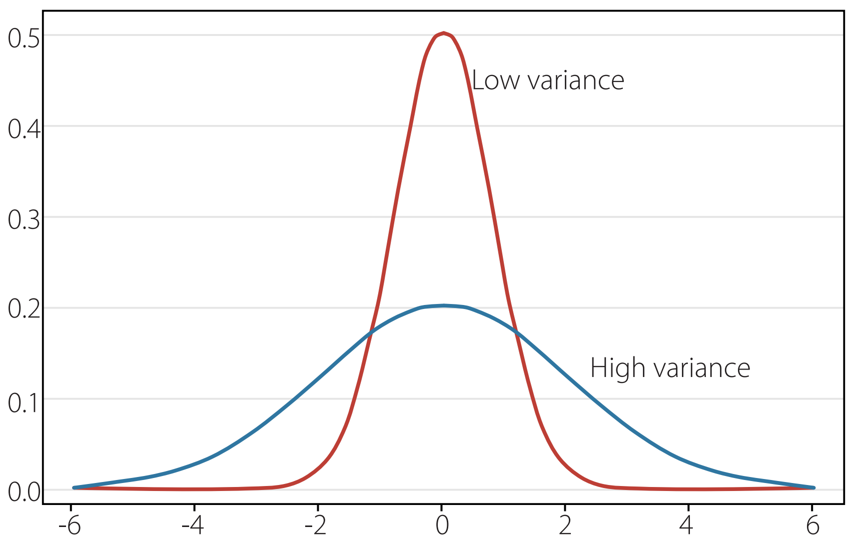
Regardless of how we estimate the slope parameter (or parameters), we will need some way of testing if it is definitely positive or not. That requires a bit more theory. Figure 1 was a plot of a *sample* of data. It is clearly not the entire collection of husbands and wives in the world. A sample is a subset of a *population*. When we do statistical analysis, we have to take account of the fact that we are working with a sample rather than the entire population. (In principle, the larger the sample, the more representative it is of the entire population.)

The line of best fit through the sample can only ever yield an estimate of the true value of the slope, and because it is an estimate, we can only really talk about a range of possible values. Regression therefore yields a distribution of possible estimates, some more likely than others. If you fit a line through data using a simple program such as Excel, it might only report the central estimate, but what the underlying theory yields is a distribution of possible values.

Most people are familiar with the idea of a 'bell curve', which summarises data. An example would be the distribution of grades in a class, where many values are clustered around the mean, but progressively fewer as you go further away from it. The wideness of a distribution is summarised by a number called the *variance*. If the variance is low, the distribution is narrow, and if it is high, the distribution is wide (Figure 3).

So, as well as the estimate of the slope, regression analysis yields an estimate of the variance. A closely related concept is the standard deviation of the slope – again, a measure of how tightly clustered the points in the sample are around the straight line we have fitted through it.[7] Statistical theory tells us that as long as the regression model satisfies a certain set of conditions, there is a 95% probability that the true value of the slope (the one you'd get if you were able to sample the whole population) is within approximately plus or minus two standard deviations of the estimate of the slope. This is called the 95% confidence interval.

Figure 3: Variance



So, we can use regression methods to fit a line through a sample of data – say hurricane frequency and temperature – and if the slope estimate is more than two standard deviations above zero, we can say we are 'confident' that an increase in temperature leads to an increase in hurricanes. If it isn't, we say that the relationship is positive but *statistically insignificant*.

Bias, efficiency and consistency

The slope estimate is obtained using a formula that takes in the sample data and generates a number. There are many formulas that can be used. The most popular one is called Ordinary Least Squares (OLS).[8] OLS also yields an estimate of the variances of each coefficient.

It is possible to distil the distribution of slope estimates down to a single value by means of a probability-weighted average. In statistics, this is known as the expected value. Statistical theory can be used to show that as long as the regression model satisfies a certain set of conditions, the expected value is the same as the value for the population as a whole. In this case we say the estimator is *unbiased*. If the set of conditions referred to above is met, the variance estimate is also unbiased.

Since there are many possible estimation formulas besides OLS, we need to think about why we would prefer OLS to the others. One reason is that, among all the options that yield unbiased estimates, OLS yields the smallest variance.[9] So it makes the best use of the available data and gives us the smallest 95% confidence interval. We call this *efficiency*.

Some formulas (or 'estimators', in the jargon) give us estimated slope coefficients or variances that are biased when the sample size is small but, as the sample size gets larger, the bias disappears and the variance goes to zero, so the distribution collapses onto the true value. This is called *consistency*. An *inconsistent* estimator has the undesirable property that as we get more and more data we have no assurance that our coefficient estimates get closer to the truth. With inconsistent estimators, the variance may shrink as the sample size increases, but the bias never reaches zero, which means the estimate does not converge on the true value.

When are estimates reliable?

I have several times referred to 'a certain set of conditions' that a regression model needs to satisfy in order for OLS to yield unbiased, efficient and consistent estimates. These conditions are listed in any introductory econometrics textbook, and they are called the Gauss-Markov (GM) conditions. Much of the field of econometrics (which is the branch of statistics that tries to use regression analysis to build economic models) is focused on testing for failures of the GM conditions and, when they are found, proposing remedies.

Some failures of the GM conditions imply only that the variance estimates are biased; the slope estimates remain unbiased. In other words, we get a decent estimate of the slope parameter but our judgment of whether it is significant or not will be unreliable. Other failures of the GM conditions imply that the estimates of both the slope and the variance are biased. In this case the analysis may be spurious and totally meaningless.

As an example of a bad research design, suppose we have data from hundreds of US cities, over many years, showing both the annual number of crimes in the city and the number of police officers on the streets. We can fit a line through the data to test if crime goes down when more police are deployed. However, there are several problems that would likely lead to failure of several of the GM conditions. First, the sample consists of small and large cities together, and we can expect very different crime statistics in larger and smaller cities. If we don't take account of this, we will get biased estimates of the variances of the slope coefficients. Second, there will be lag effects: a change in police officer numbers might lead to a change in crime only after a certain amount of time has passed. This too can bias the slope and variance estimates. Third, while crime may depend on policing, policing levels may also depend on the amount of crime, so both variables are determined by each other: one is not clearly determined outside the model. This can severely bias the coefficients and lead to spurious conclusions (such as that more policing leads to higher crime levels). Finally, both crime and policing depend on factors not included in the model, and unless those outside factors are uncorrelated with the level of policing, the slope and variance estimates will be biased.

It is therefore critical to test for failures of the GM conditions. There is a huge literature in econometrics on this topic, which is called *specification testing*. Students who learn regression analysis learn specification testing all the way along. If a regression model is used for economics research, the results would never be taken at face value without at least some elementary specification tests being reported.

For some violations of the GM conditions, the remedy consists of a transformation of the data before applying OLS. One example would be converting all the data into per-capita terms. When we apply a data transformation to remedy a violation of one or more GM conditions, we then say we are using Generalized Least Squares (GLS). Having applied a GLS transformation doesn't mean we can assume the GM conditions automatically hold; they still have to be tested. In some cases a GLS transformation is still not enough and other modifications to the model are needed to achieve unbiased and consistent estimates.

The method of Allen and Tett

Various authors prior to AT99 had proposed comparing observed climate measures – changes in temperature or hurricane frequency or the occurrence of heatwaves, for example – to climate simulations with and without greenhouse gases. If including greenhouse gases gave a significantly better match to the observations, then scientists could point the finger at human emissions as the cause. The method is referred to as 'fingerprinting' or 'optimal fingerprinting'.[10]

Those authors had also argued that the analysis would need to be aided by adjusting the data for local climatic variability, putting more weight on areas where the climate is inherently more stable and less weight on areas where it is 'noisier'. The weightings required are calculated from something called the 'climate noise covariance matrix', which measures the variability of the climate in each location and, for each pair of locations, how their climate conditions correlate with each other. The mathematics involved is beyond the scope of this paper, but for the purposes of understanding the key issues, it is simply necessary to understand that one of the steps required is to calculate the inverse of the matrix.

But this proved difficult in practice. Rather than using observed data to compute the matrix, climatologists have long preferred to use climate models. While there were reasons for this choice, it created many problems (which I discuss in my paper). One of these was that climate models don't have enough resolution to identify all elements of the matrix independently. This meant that the matrix had no inverse.[11] So the scientists were forced to use an approximation called a 'pseudo-inverse' to compute the needed weights. This created further problems.

The error

Allen and Tett's argument was something like this. They noted that applying a weighting scheme makes the fingerprinting model similar to a GLS regression. And since a properly-specified GLS model satisfies the GM conditions, their method (they said) yields unbiased and efficient results. That slightly oversimplifies their argument, but not by much. And the main error is obvious. You can't know if a model satisfies the GM conditions unless you test for specific violations. AT99 stated the GM conditions incorrectly, leaving an important one out altogether, and failed to propose any tests for violations.

In fact, they derailed the whole idea of specification testing by arguing that they only needed to test that the climate model noise covariance estimates were 'reliable' (their term—which they did not define), and they proposed something called the Residual Consistency (RC) test for that purpose. They didn't offer any proof that the RC test does what they claimed it does.[12] In fact, they didn't even provide a mathematical statement of *what* it tests; they only said that if the formula they proposed pops out a small number, the fingerprinting regression is valid. In my paper, I explained that there can easily be cases where the RC test would yield a small number, even in models that are known to be misspecified and unreliable.

So, in summary, Allen and Tett's method failed to ensure the GM conditions were met, and so failed to assess whether their estimates were reliable. In fact, as I argued in my paper, the Allen and Tett method, as set out in their paper, *automatically* fails at least one GM condition, and probably more. So the results must be assumed to be unreliable.

In the years since its publication, however, no-one noticed the errors in the AT99 discussion of the GM conditions, no-one minded the absence of a derivation of the RC test, and none of the subsequent applications of the AT99 method were subject to conventional specification testing. That means we have no basis for accepting any claims that rely on the optimal fingerprinting method.

An aside: the slight modification

Allen and Tett's optimal fingerprinting approach, with only one slight modification, has been the one used by the climate science profession for 20 years.

The slight modification came in 2003, when Myles Allen and a different coauthor, Peter Stott, proposed shifting from GLS to another estimator called Total Least Squares (TLS).[13] It still involves weighting for climate variability, but the slope coefficients are estimated using a different formula. Their rationale for TLS was that the climate model-generated variables in the fingerprinting regression are themselves quite 'noisy', and this can cause GLS to yield coefficient estimates that are biased downwards. This is true, but econometricians deal with this problem using a technique called Instrumental Variables (IV). We don't use TLS (in fact almost no-one outside of climatology uses it) because, among other things, if the regression model is misspecified, TLS over-corrects and imparts an upward bias to the results. It is also extremely inefficient compared to OLS. IV models can be shown to be consistent and unbiased. TLS models can't, unless the researcher makes some restrictive assumptions about the variances in the dataset that themselves can't be tested; in other words, unless the modeler 'assumes the problem away.' I will discuss these issues in detail in a forthcoming paper.

Implications and next steps

Optimal fingerprinting fails the GM conditions. Allen and Tett erroneously claimed the opposite, and later authors quoted and relied on this claim. The method (including the TLS variant) yields results that might by chance be right, but in general will be biased and inconsistent and therefore cannot be assumed to be reliable. Nothing in the method itself (including use of the RC test) allows scientists to claim more than that.

In addition to examining the biases introduced by using TLS in fingerprinting regressions, I am working on a paper exploring the effects of applying basic specification testing to fingerprinting regressions and remedying the resulting failures.

Replying to responses

Optimal fingerprinting has been heavily used in the climate literature for establishing attribution; studies applying it have been cited thousands of times, and it has been prominently featured by the IPCC since it first appeared. There has been nearly exclusive reliance on the RC test to defend fingerprinting analysis results. A number of commentators on my paper have tried to shrug off my criticism as unimportant or irrelevant. But if none of the issues raised in my paper 'matter', then we might as well say nothing in the climatology literature matters.

More specifically, in considering any response to my paper, it will be important to note whether it actually disagrees with or disproves my arguments, or simply tries to change the subject. I anticipate that a lot of respondents will implicitly concede that my paper is correct, but argue it doesn't matter because so much time has gone by. However, as a matter of the scientific record it is important to understand and acknowledge if Allen and Tett made errors in their mathematical presentation, and whether the subsequent literature corrected them or simply carried them forward. As far as I have seen, they were carried forward, in the sense that people still to this day rely on the RC test and they still use AT99-type regression models without testing for specification errors associated with the GM conditions.

Also, and more generally, if major errors in the methodology went unnoticed for so long, it calls into question how much confidence we can have in other statistical methodologies that have been developed by climate scientists in subsequent years. Having worked on paleoclimate reconstruction methods, trend estimation and comparison methods, and now on optimal fingerprinting, I conclude that climate journals, unlike statistics or econometrics journals, seem to rely on referees who don't know how to ask the right questions when confronted with a novel statistical method. My paper is, in effect, the referee report that Allen and Tett would have received had they submitted their paper to a statistics or econometrics journal.

One line of response to my paper has been that Allen and Tett's methodology has been superseded by other methods,[14] which get the same results (sometimes). But my critique still applies. Specifically, there remains the problem of showing that the resulting estimator is consistent.[15] The regularization-based fingerprinting literature has never revisited the claims around whether the GM conditions are satisfied.[16] Regularisation is a computational improvement, possibly, but not a theoretical one.

There are some other recent attribution methods, including time-series methods, that do not make any use of climate models.[17] My critique does not specifically apply to these. There may be other issues, but I haven't looked at them in detail. The ones I have seen have largely been confined to analysing the time series of global average surface temperatures, and have considered only a very limited number of explanatory variables.

Notes

1. MR Allen and SFB Tett, Checking for model consistency in optimal fingerprinting. *Climate Dynamics* 1999; 15:419–434.

2. See TAR Chapter 12, Box 12.1, Section 12.4.3 and Appendix 12.1.

3. Imbers, J, A Lopez, C Huntingford and M Allen, Sensitivity of climate change detection and attribution to the characterization of internal climate variability. *Journal of Climate,* 2014; 27.

4. See AR6 Section 3.2.1 https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Chapter_03.pdf.

5. https://link.springer.com/content/pdf/10.1007/s00382-003-0313-9.pdf.

6. McKitrick, R. Checking for model consistency in optimal fingerprinting: a comment. *Climate Dynamics* (2021).

7. The standard error, or standard deviation, is the square root of the variance.

8. It is derived by supposing that the straight line allows us to predict the value of *Y* that corresponds with each value of *X*, but there will be an error in each such prediction, and we should choose the slope and intercept estimates that minimise the sum of the squared errors.

9. This assumes we are only considering linear estimators, which is a detail we can ignore for the present purpose.

10. A rather witty response to this research on social media was to suggest that the second 'r' in 'optimal fingerprinting' should have been an 'o'.

11. In mathematical terms we say the matrix is 'rank deficient', an implication of which is that the inverse does not exist.

12. There is a question in statistical theory of whether the residuals of a regression provide consistent estimates of the unknown error terms, but the RC test of Allen and Tett has nothing to do with this.

13. Allen, MR and PA Stott. Estimating signal amplitudes in optimal finger-printing, Part I: Theory. *Climate Dynamics* 2003; 21:477–491. DOI 10.1007/s00382-003-0313-9.

14. Specifically, regularisation methods, associated with authors Ribes, Terray, Hannart and so forth. These are covered in my paper in a couple of places. Regularisation is an alternative way of estimating the inverse of the non-invertible climate noise matrix. It yields a full-rank approximation so there is no longer a dependence on the rank truncation parameter.

15. See condition [N3] in my paper.

16. For instance it has never discussed the conditional independence assumption, which is a key GM condition.

17. Such as cointegrating vector autoregressions or CVAR.

Review process

GWPF publishes papers in a number of different formats, with a different review process pertaining to each.

- Our flagship long-form GWPF Reports, are all reviewed by our Academic Advisory Panel.
- GWPF Briefings and Notes are shorter documents and are reviewed internally and/or externally as required.

In addition, for most publications, we invite external reviews from a parties who we would expect be critical. If these critics have substantive comments, we offer to publish these alongside the main paper. In this way, we hope to encourage open debate on the important areas in which we work.

The review process for GWPF papers is therefore somewhat more in depth than a typical review for an academic journal.

- More potential reviewers are involved
- The number of substantive comments typically exceeds journal peer review
- The identity of the author is known to the potential reviewers.

As an organisation that is subject to sometimes very hostile criticism, our review process has to be very careful. All parties involved therefore treat the reviews with the utmost seriousness.

Final responsibility for publication rests with the Chairman of the Trustees, Terence Mordaunt, and the GWPF Director, Dr Benny Peiser. But In every case, the views expressed are those of the author. GWPF has never had any corporate position.

About the Global Warming Policy Foundation

The Global Warming Policy Foundation is an all-party and non-party think tank and a registered educational charity which, while openminded on the contested science of global warming, is deeply concerned about the costs and other implications of many of the policies currently being advocated.

Our main focus is to analyse global warming policies and their economic and other implications. Our aim is to provide the most robust and reliable economic analysis and advice. Above all we seek to inform the media, politicians and the public, in a newsworthy way, on the subject in general and on the misinformation to which they are all too frequently being subjected at the present time.

The key to the success of the GWPF is the trust and credibility that we have earned in the eyes of a growing number of policy makers, journalists and the interested public. The GWPF is funded overwhelmingly by voluntary donations from a number of private individuals and charitable trusts. In order to make clear its complete independence, it does not accept gifts from either energy companies or anyone with a significant interest in an energy company.

Views expressed in the publications of the Global Warming Policy Foundation are those of the authors, not those of the GWPF, its trustees, its Academic Advisory Council members or its directors.

GWPF NOTES