UNDER REVIEW

RESCIENCE C

Reproduction / Ten Years Reproducibility Challenge / Epidemiology

[Rp] Computational replication of estimating trends and seasonality in coronary heart disease

Adrian Barnett[1, ID]

[1]School of Public Health and Social Work, Queensland University of Technology, Brisbane, Australia

Edited by
(Editor)

Reviewed by
(Reviewer 1)
(Reviewer 2)

Received
08 March 2020

Published
–

DOI
–

A replication of Barnett2004.

Taking the ReScience Ten Years Reproducibility Challenge, I tried to reproduce my own code from a paper published in 2004 that estimated the trend and seasonality in coronary heart disease in multiple countries. The original SAS code and text data were relatively easy to find online and on a compact disc, however I had overwritten some of the code for new analyses and I could not find the exact code to replicate the published results. I managed to adumbrate the key figure from the original analysis, although the Bayesian credible intervals were noticeably narrower for the trend, which may be because of a difference in the parameter that controlled the non-linearity of the trend. The estimates of the seasonal patterns were similar, but not identical, and this could be due to my ambiguous description in the original paper of how additional seasonal components were added.

1 Introduction

1.1 Historical context

This paper aims to computationally reproduce my own statistical analysis from 16 years ago as part of the 10-year reproducibility challenge[1]. The analysis examined seasonal patterns in coronary heart disease in multiple countries[2]. In most countries coronary heart disease deaths and emergency admissions to hospital have a strong seasonal pattern with a peak in winter and nadir in summer[3]. There is an interesting variability between-countries in the size and timing of the winter peak. A better understanding of the differences between countries could help our understanding of the underlying causes of the winter peak in disease.

The analysis examined monthly time series of coronary events (fatal and non-fatal) from 35 locations in 21 countries. The time series were 8 to 14 years in length and the earliest year of data was 1980. The analysis aimed to split the time series into a long-term trend, seasonal pattern(s) and remaining noise. There were two approaches, one that used a two-stage approach by first removing the trend, and another that estimated the trend and season together. The trend was estimated using the Kalman filter and the seasonal patterns were estimated using sinusoids. The estimates were made using Markov chain Monte Carlo.

Correspondence should be addressed to Adrian G. Barnett (a.barnett@qut.edu.au)
The authors have declared that no competing interests exists.
Code is available at https://github.com/agbarnett/tenyears.

1.2 Original source code

The original code was written in SAS (version 8.00 for Windows) and chain convergence was checked using the "coda" package (version unknown) in R[4]. The code was based on Matlab code that I wrote during my PhD[5] at The University of Queensland. I converted the code to SAS because I did not have access to Matlab after finishing my PhD and I hoped that more people would be able to re-use my SAS code to apply to their own time series data. The SAS code was first published in November 2004 with an update in August 2006, but I did not record what changes were made. The paper states the code was published in September 2002 but this is likely a typo in the year given that the paper was not submitted until January 2004.

The original code was published in an online appendix to the paper that was separate from the journal in November 2004 at this address: http://www4.ktl.fi/publications/monica/chd_seasonal/appendix.htm. However, the site has since moved to: https://www.thl.fi/publications/monica/chd_seasonal/appendix.htm (accessed 25 January 2020). The web site was created by the WHO MONICA Project specifically for appendices to papers that used the MONICA data. MONICA stands for monitoring trends and determinants in cardiovascular disease. The MONICA Project was a large multi-country study that aimed to examine several aspects of coronary heart disease[6]. It was a well-managed project that had staff who assisted with access to the data and helped add my code to the web.

2 Results

2.1 Retrieval of the software

Most of my SAS code was easy to find, both on the web and on an old compact disc in a desk draw. I made this compact disc of my files when I moved jobs. I did not keep the data because I was more interested in other people using my code for their own data rather than replicating our published results. I believe that another reason for deleting the data was to save space, as the files were relatively large by the standards of the day (around 31,000 kilobytes for data in text format and 70,000 kilobytes for data in SAS format). Luckily the original data files were available in text files in fixed-width format on a compact disc attached to the MONICA monograph published in 2003[6].

My SAS code contained the macros needed to run the two statistical analyses. There was also SAS code to simulate a seasonal time series as an example data set. I found SAS code to read the MONICA data on my compact disc, but not on the web. I could not find the SAS code that applied the two methods to each location.

2.2 Replication execution

I used SAS version 9.4 for Windows (Microsoft Windows 10) to replicate the results; there was no need to use an earlier version of SAS.

My SAS files had relatively good instructions with a detailed header at the top of every file and comments throughout. However, some files had been adapted from the original analysis to secondary analyses, making the code a palimpsest with some commands commented out with notes such as, "changed for weather analysis" and "sensitivity analysis of Ghent". The correct data set was also ambiguous because alternative data sets had also been used (e.g., fatal vs non-fatal events). In hindsight I should have kept the exact data and syntax files needed to re-create the published results.

The macros to run the analyses mainly needed only cosmetic edits, and I also put each macro in its own file rather than one overall file of macros. I also moved part of one macro, that estimated the standard deviation of the noise, into the macro that tested the periodogram for remaining seasonal structure. The largest change was having to re-write the files that applied the macros to the MONICA data. My SAS programming was rusty and I could not automate the process for each centre, so instead I created a new file for each analysis in each of three centres.

I did not repeat my analysis in all 35 locations, but instead did the three locations in the first figure as this felt sufficient to test the code.

The code to estimate the trend without the seasonal pattern used a slightly different parameterisation with "tau" as a ratio rather than an absolute value. "tau" controls the amount of change over time in the trend, with smaller values creating more linear trends. I cannot be sure that this version was the same code as the original.

2.3 Closeness of the replicated results to the original

The original figure is shown in Figure 1 and my replication in Figure 2. The figures show the estimates of the trend and season for three locations. The results are similar except for the confidence intervals for the trend for the two-stage method, which are much narrower for the replication, but are centred on a similar mean. This may be because I could not be certain that I had found the original code to perform the Kalman filter smoothing because of the difference in how "tau" was parameterised. I also needed some trial-and-error to select "tau" which controls the smoothness of the trend and is defined by the user rather than being estimated by an algorithm.

The estimates of the seasonal patterns for the three locations are replicated in Table 1 which also shows the original results. The results are similar for the two-stage method, with most differences being within $\pm$ 0.1. The biggest difference being for the estimate of the noise standard deviation in Belfast (9.0 versus 8.8).

There were differences in the frequencies for the combined method, as two seasonal frequencies were not included in the new results, one in Perth and one in Belfast. Although the amplitudes were relatively small (1.6 or less), hence these are smaller seasonal patterns that are marginally important.

The difference in seasonal frequencies could be because of a key ambiguity in the original paper regarding the decision to add additional seasonal components for the combined approach. After fitting a candidate model, the residuals were tested to look for additional seasonal structure. If structure existed, then the model was modified to add an additional seasonal component as necessary, followed by another test of the residuals. The code produced both the periodogram and the spectrum, which is a smoothed version of the periodogram. The periodogram tended to flag significant seasonal patterns when the spectrum did not, and I used the spectrum to make the decisions because some of the statistically significant patterns found by the periodogram were only just above the 0.05 threshold used to judge statistical significance. This part of the replication was therefore somewhat subjective.

Some of the discrepancy in results in Table 1 could be because the estimates are made using Markov chain Monte Carlo which uses random steps to make estimates. Hence the estimates will depend on the original random number seed. However, I used 5,000 estimates with a burn-in of 500, so these differences should be minor and may explain some of the small differences in the first decimal place for the amplitude and noise estimates. This potential discrepancy would have been avoided if the original code had specified the random number seed.

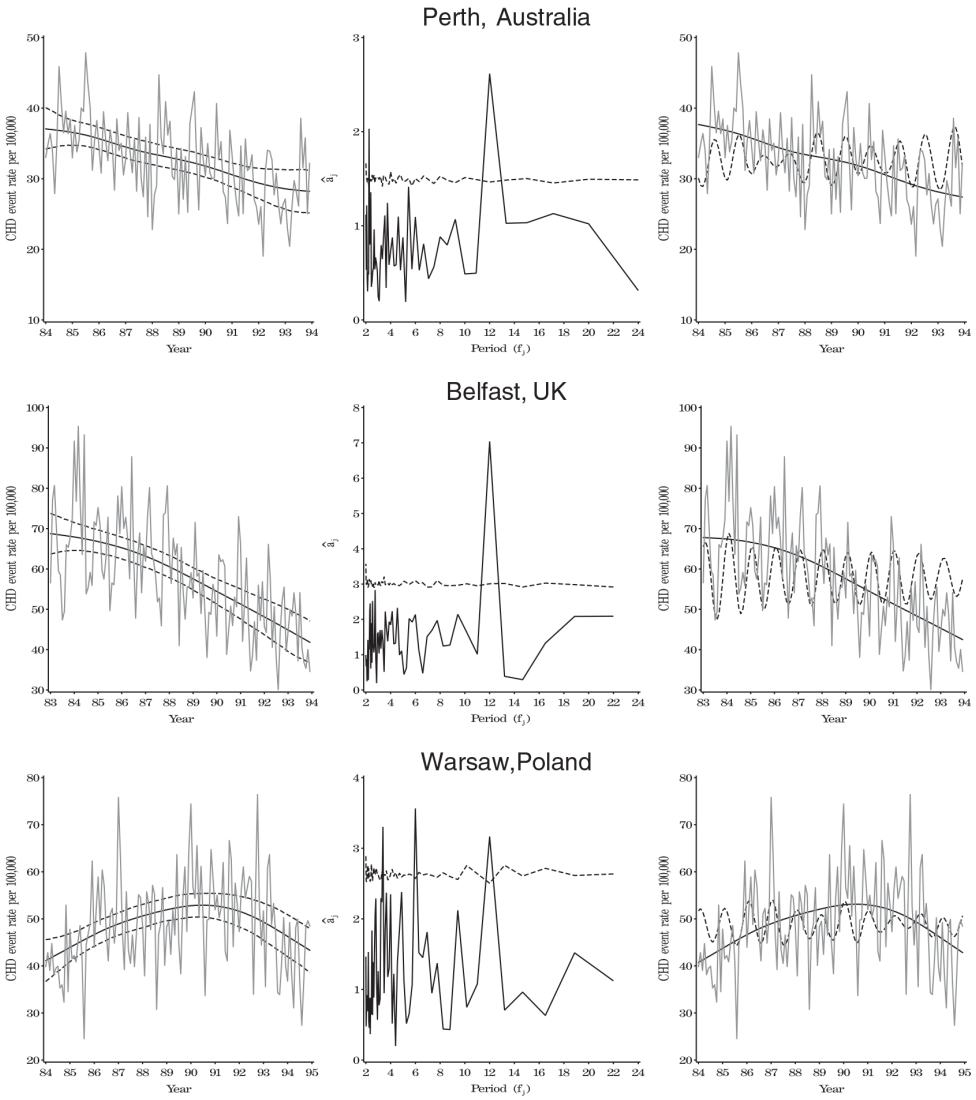

Figure 1. Original results for three locations using the two-stage method (left and centre column) and combined method (right column). The left column shows the observed data, estimated trend and 95% confidence interval. The middle column shows the periodogram of the de-trended data and limit for the test of seasonal structure (dotted line). The right column shows the observed data, estimated trend, and estimated annual seasonal pattern.

Overall the replication took about 21 hours. This was mostly the fiddly task of working out which macros to run and re-creating the figures and tables. Given the amount of code that had to be re-created, I think that only someone familiar with SAS would have been able to replicate the results. My familiarity with the MONICA Project also helped me create the new code, hence I am not certain that an outsider could have replicated the results, perhaps only by trial-and-error.

3 Discussion

It was an interesting exercise to bring this old code back to life. It would have been easier and faster if my original code had been properly arranged and prepared solely to

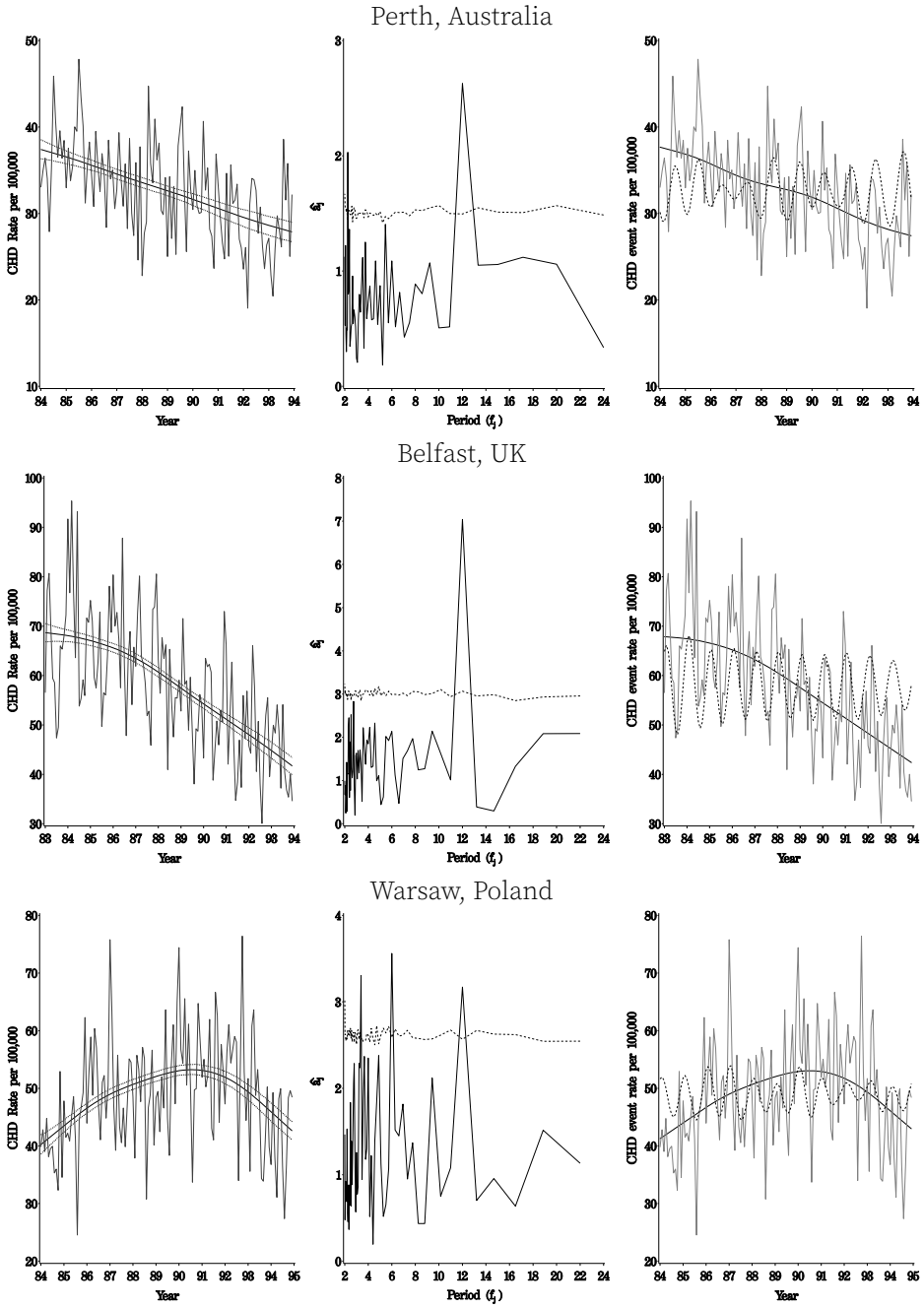

Figure 2. Replicated results for Figure 1

repeat the original analyses. However, I was not aware of guidelines for curating code back in 2004. My original code had many comments which were useful, but there were problems with the overall file structure and how the files related to one another. There were also ambiguities with the dates of the code, and the changes I made after posting the original files. This is now solved by sites that track and date-stamp all changes, such as *GitHub*.

My main reason for sharing the original code was so that others could adapt it to their own data. However, of the 21 papers that cited the original paper which I could access, none of them stated that they re-used my SAS code.

Location	Frequency, months	Amplitude, rate per 100,000	Noise SD
Two-stage method			
Perth, Australia	2.3	2.0 (2.0)	4.3 (4.2)
	12	2.6 (2.6)	
Belfast, UK	12	7.0 (7.0)	8.7 (8.7)
Warsaw, Poland	3.4	3.3 (3.3)	7.5 (7.5)
	6	3.6 (3.6)	
	12	3.2 (3.2)	
Combined method			
Perth, Australia	2.3	1.0 (1.0)	3.9 (3.9)
	2.7	—- (0.8)	
	3.8	1.3 (1.3)	
	5.5	1.5 (1.4)	
	12	2.7 (2.7)	
Belfast, UK	2.8	—- (1.6)	9.0 (8.8)
	12	7.1 (7.0)	
Warsaw, Poland	3.4	3.4 (3.4)	7.9 (7.8)
	6	3.7 (3.7)	
	12	3.3 (3.3)	

Table 1. Estimates of seasonal patterns using the two-stage and combined methods for three locations. The original results are in brackets. Each row in the table corresponds to a seasonal frequency at a location. SD = standard deviation.

Overall the replication was broadly successful in terms of re-creating the key figure and estimates for three locations. There was subjectivity in modelling decisions concerning the smoothness of the trend and number of seasonal components, which made perfect replication difficult.

References

1. K. Hinsen and N. Rougier. "Challenge to test reproducibility of old computer code." In: **Nature** 574.7780 (Oct. 2019), pp. 634–634.
2. A. G. Barnett, A. J. Dobson, and For the WHO MONICA (monitoring trends and determinants in cardiovascular disease) Project. "Estimating trends and seasonality in coronary heart disease." In: **Statistics in Medicine** 23.22 (2004), pp. 3505–3523.
3. S. Stewart, A. K. Keates, A. Redfern, and J. J. V. McMurray. "Seasonal variations in cardiovascular disease." In: **Nature Reviews Cardiology** 14.11 (May 2017), pp. 654–664.
4. M. Plummer, N. Best, K. Cowles, and K. Vines. "CODA: Convergence Diagnosis and Output Analysis for MCMC." In: **R News** 6.1 (2006), pp. 7–11.
5. A. G. Barnett. "On the use of the bispectrum to detect and model non-linearity." In: **Bulletin of the Australian Mathematical Society** 67.3 (June 2003), pp. 527–528.
6. H. Tunstall-Pedoe, W. H. O. M. Project, W. M. Project, and W. H. Organization. **MONICA, Monograph and Multimedia Sourcebook: World's Largest Study of Heart Disease, Stroke, Risk Factors, and Population Trends 1979-2002**. MONICA, Monograph and Multimedia Sourcebook: World's Largest Study of Heart Disease, Stroke, Risk Factors, and Population Trends 1979-2002 p. 1. World Health Organization, 2003.