Reviewing: Trends in fine particulate air pollution and ground-level ozone in the United States, 2003-2012

Advanced Methods III 140.753

Abstract

Note that I think that this analysis was very well done. Kudos! The review I wrote is rather lengthly as sometimes I get into small details. Keep it up!

1 Summary

The authors of the study *Trends in fine particulate air pollution and ground-level ozone in the United States, 2003-2012* obtained data from the Environmental Protection Agency from 2003 to 2012 on both particulate matter (PM 2.5) and ground-level ozone measurements. Their overall goal was to explore the trends at the national (US only) level and between the East and West states. To do so, they summarized the information for each variable by taking the median at the state-month resolution. For exploring the national trends, they further took the mean per month from the previous medians. Similarly, for the East and West states they took the median per month for those states only. Once the data had been reduced, they used a Hodrick-Prescott filter to decompose the time series for PM 2.5 and ozone respectively. Using the results from this model along with a linear regression model, they visualized the national, western and eastern trends. This information, along with other tests like the Wilcoxon's rank sum test, allowed the authors to conclude that PM 2.5 has been decreasing while the ozone levels are increasing. Both slope terms are significantly different from zero. In addition, the PM 2.5 levels are lower for the western states yet the ozone levels are slighly higher for the eastern ones. The overall analysis is exploratory in nature and no inference was carried out.

Another topic that the authors investigated was whether they could detect any change in regulation. However, they could not find evidence of such effect with the ozone data.

The authors of this study recognize that there are other approaches that could be more prudent. In addition, they point out that their analysis overlooks complications such as the increasing number of monitor sites.

In terms of reproducibility, the authors provided instructions on the order to run their scripts. They also provided their code and summarized versions of the data set.

2 Major revisions

I think that the report is very nicely written as the problem's context is described from the impact it has on human lives. In addition, the report is well referenced[1] and explains the major pre-processing steps along with the statistical models used. The plots are nicely documented by including information on the intercept estimates. Furthermore, statistical summaries such as p-values are provided to support their claims.

However, I find intriguing that the authors don't highlight the seasonality trend of the data such as higher PM 2.5 values in the summer and lower ones in the winter. They even try to remove this when doing their statistical analysis. I find this contradictory at least at the intuitive level. Further discussion on why they try to remove the short-term fluctuations and seasonality could be well worth it.

[1] I never expected a 1997 reference from the *Journal of Money, Credit and Banking*. Wow!

From the reproducibility point of view, their code files are almost devoid of comments. I'm not asking for abundant descriptions but more information on what logical steps are being carried out would be much appreciated.

3 Minor revisions

The model fits are said to be excellent ($R^2 = 0.9$) for each i in $1, \ldots, T$. It would be interesting for the authors to describe how this should impact the final Hodrick-Prescott trends shown in figures 1 and 2. I mean, with such a high R^2 I expected them to be closer to the trends seen in the data. Maybe I'm not understanding some details here.

For figures 1 and 2, would it be feasible to display the confidence bands for the estimates? I believe that this information is useful for the reader to determine whether the model is explaining the observed data.

Figures A1 and A3 seem to do a better job at explaining the East vs West differences. I understand that plotting the regression lines on top of that is too much.

The authors claim that taking the median from the median state-month values enables the states to contribute equally. I'm not entirely convinced that this is true. I think that the medians per state-month values do help here, but the final median per month doesn't. I understand that this part was not easy to describe, but maybe using some notation could help to clarify which *median* they are talking about.

Code style does not seem to be consistent, for example on the use of white space. I personally prefer the second form shown below.

```
foo(argument=value)
foo(argument = value)
```

In addition, some variables names are a bit confusing such as *datee* and *datee1* when the first one has data fro the West region and the second one for the East. Furthermore, the authors don't use indentation in loops such as *for* statements. Finally, the code for file for PM 2.5 seems to be almost the same as the one for ozone except for the data used and descriptions in plots. I believe that this strategy can be error prone.

It would have been excellent to highlight the largest drops in figures 1 and 2 that the authors mention when looking for regulation changes. Yet I'm totally aware that doing so is challenging with *lattice*.