Atmos. Chem. Phys. Discuss.,
doi:10.5194/acp-2017-26-EC1, 2017

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Status Update: Is smoke on your mind? Using social media to determine smoke exposure" *by* Bonne Ford et al.

D. Topping (Editor)

david.topping@manchester.ac.uk

Received and published: 24 April 2017

In this paper the authors explore the power of social media data to improve data coverage on smoke exposure. As the need for increased data density in atmospheric exposure generally progresses, it is highly likely that more studies will rely on the 'citizen sensors' approach. I find the study a refreshing addition to the often stagnant observation based literature. It adds to the already wealthy cross-disciplinary arm of ACP and I enjoyed reading it.

I would like to see this published in ACP after some comments are addressed below.

Could you remove multiple contributions from a particular individual from Facebook? How do you remove a particular biased commentary from a subset of users? I was interested in the lack of threshold PM2.5 concentration for people to start posting. Could

this be a factor? If a small number of users are relying on available monitoring data, then reporting this, they might be driving a wider response. This isn't necessarily a negative feature, of course, but has parallels in social media coverage of viral outbreaks.

What percentage of facebook users are you actually obtaining? For example, twitter restricts access to a small percentage unless a fee is paid. Could you add this information to the manuscript?

I often wonder how much an individual response is due to reporting on a news item/political debate rather than commentary on conditions experienced at any point in time. As with some practices in sentiment analysis, it might be useful to analyse bigrams/trigrams for a given post. Is that data available?

I appreciate the difficulty in providing social media data, having been personally rejected from other journals on this commonly known technicality. Would it be possible to provide a little more detail on the process of Facebook data retrieval for those who might want to replicate a similar study at least?

Regarding the regression model, was there a particular reason to opt for linear combinations of predictor variables? I wonder if the accuracy of your technique might be increased by even a simple decision tree, or ensemble method, an additional variables. Using k-folds cross validation and variable selection this might generate a more widely applicable method.

A minor comment on the line: 'social media datasets could currently improve estimates without the costly investment of computer modeling.' I would add this really depends on the application. If you were to fit a multivariate regression model to actual post content, with access to many hundreds of thousands of posts, the time to train a model varies with amount of data used. I leave it to the authors to decide on whether to retain this.

Interactive comment on Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2017-26, 2017.