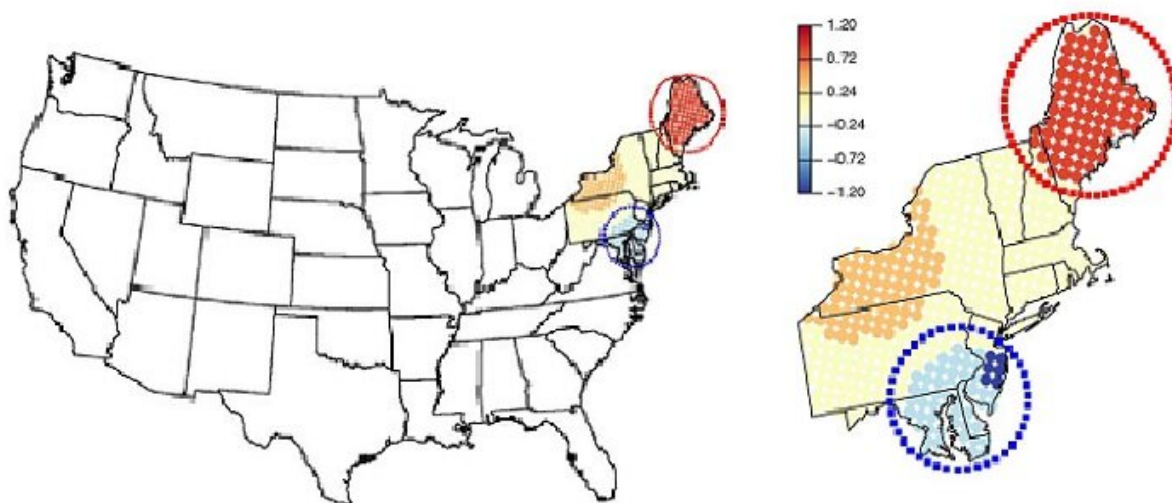# More accurate detection of hotspot clusters provides new insights into the behavior of air pollution

September 16 2019

The mixed effect model allows more accurate identification of hotspots in which atmospheric variables relate differently compared with other areas. Credit: John Wiley & Sons Ltd.

A more reliable method for identifying regions with different relationships between air pollution and weather conditions improves the detection of pollution hotspots.

The relationship between weather conditions and air pollution is complex and can vary wildly from location to location. This makes it difficult to pinpoint the sources of pollution and predict its behavior in the atmosphere. While data scientists and statisticians have made significant progress in wrestling with this problem, the enormous volumes of environmental data and multitude of variables, such as wind speed, temperature and pollution component, require compromises to make the problem manageable.

For example, most existing approaches to detecting "hotspots" in the correlation between variables in spatial data involve constructing a grid in which the relationship between variables in a cell is treated independently of all others. Although this is not entirely realistic—there is often dependence between spatial areas particularly in weather and air pollution data—it is extraordinarily difficult to find spatial hotspots and determine the spatial dependence structure at the same time.

Ying Sun and Junho Lee from KAUST's Environmental Statistics Laboratory have made a leap forward in addressing this problem with the development of a "mixed effect model" for hotspot detection.
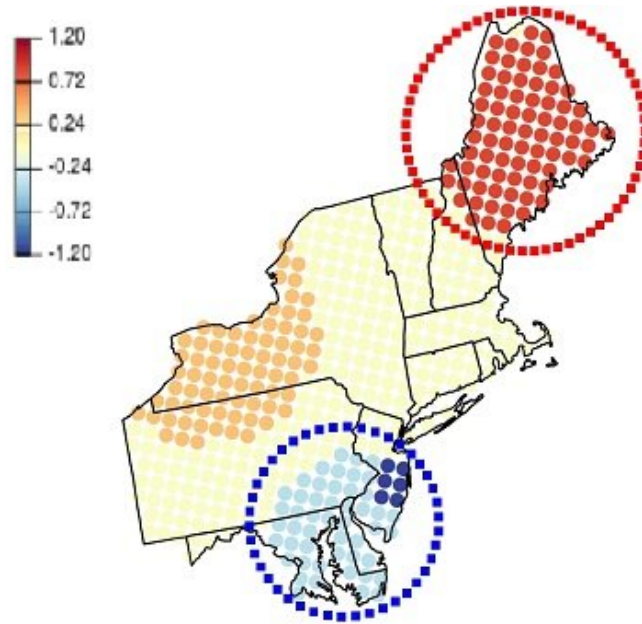
This map shows how the mixed effect model breaks the northeastern U.S. into blocks, allowing them to identify "hotspots." Credit: John Wiley & Sons Ltd

"We address the problem by using a simple spatial block structure to approximate the spatial dependency," says Lee. "This allows us to find spatial hotspots showing distinct patterns while reducing the rate of false positives due to spatial dependence."

The approach, developed in collaboration with Howard Chang from Emory University in the United States, involves breaking the region into blocks and sequentially applying random effects to the blocks to tease out strong correlations from background variability or "noise." This has the added benefit of being able to identify any number of hotspot clusters in the data, including clusters that may overlap.

"The main challenge was how to decide an appropriate block size for the random effects," says Lee. "We settled on matching the block size to the

range of spatial dependence in the data."

The team applied their method to analyze air pollution data over the northeastern United States. They found that in summer, the concentrations of micrometer-scale particulate matter in the air (PM2.5) increased with temperature and decreased with relative humidity across most of the region.

"However, with our approach, we could find distinct areas with the opposite trend, such as in the Chesapeake Bay area, where there is a negative association between PM2.5 and temperature, and around Maine where there is a positive correlation between PM2.5 and relative humidity," says Lee.

Provided by King Abdullah University of Science and Technology