JASA ACS Reproducibility Initiative - Author Contributions Checklist Form

The purpose of the Author Contributions Checklist (ACC) Form is to document the artifacts associated with a manuscript (e.g., code and data supporting the computational findings), and describe how to reproduce the findings. The final version of this document will be included as online supplemental material with the published paper and referenced in the abstract.

Data

Abstract (Mandatory)

City-level daily $PM_{2.5}$ concentrations of 338 Chinese cities from Jan 1st, 2015 to Dec 31st, 2016; Station-level monthly $PM_{2.5}$ concentrations from 73 stations in the BTH region from January 2013 to December 2016; Topographic information (longitude and latitude) of 338 Chinese cities and 73 BTH stations.

Availability (Mandatory)

Available at
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LMPZFZ

Description (Mandatory if data available)

- The authors have legitimate access to the data.
- The data are approved for noncommercial use.
- The data are deposited in JASA Dataverse.
- The $PM_{2.5}$ concentration and longitude/latitude data are provided by China National Environmental Monitoring Center (CNEMC), while the China's elevation data are publicly accessible in the platform of China's Geospatial Data Cloud (http://www.gscloud.cn/).
- The $PM_{2.5}$ concentration and longitude/latitude data are CSV files while China's elevation data are TIF files.
- Metadata: "city", "station", "lon", "lat", "year", "month", "days", etc.
- Version information (N/A)

Optional Information (complete as necessary)

N/A

Code

Abstract (Mandatory)

R codes for both simulation and real data analysis

Description (Mandatory)

- R codes
- Licensing information: MIT License

- We agree that the code will be deposited with data in JASA Dataverse if the manuscript is accepted.
- Version information (N/A)
- Supporting software requirements:
  R version: 3.6.1 (We recommend R version 3.4 or higher);
  R packages and versions (in parentheses):
  "fda" (2.4.8), "funcy" (1.0.1), "mclust" (5.4.5), "fields" (9.8.3), "MASS" (7.3.51.4), "zoo" (1.8.6), "ggmap" (3.0.0), "rgdal" (1.4.8), "PottsUtils" (0.3.3), "maptools" (0.9.5), "plyr" (1.8.4), "scales" (1.0.0), "doParallel" (1.0.15), and "foreach" (1.4.7).

Optional Information (complete as necessary)

Instructions for Use

Reproducibility (Mandatory)

- What is to be reproduced: All tables and figures in the manuscript.
- How to reproduce analyses
  - The "homo" folder provides code for the homoscedastic case in simulation, with "Simulation_homo.R" as the main function and "SimulateMRF.R" to simulate Markov random fields.
  - The "hetero" folder provides code for the heteroscedastic case in simulation, with "Simulation_hetero.R" as the main function and "SimulateHeteroData.R" to simulate heteroscedastic data.
  - "Regionalization_China.R" is the main function for real data analysis of China, with "GetNeighbors.R" performing neighborhood selection by Markov random field combining geographic information.
  - "Regionalization_BTH.R" is the main function for real data analysis of the BTH region.
- Expected run-time of the workflow (and information about particularly slow steps in workflow, if any). If possible, give the approximate time to run on a standard desktop machine.
  - The expected run-time for real data analysis of China is about 2 hours; the expected run-time for real data analysis of the BTH region is about 15 minutes on a standard desktop machine.
  - The expected run-time for one trial of simulation (homoscedastic case) is about 40 minutes.
  - The expected run-time for one trial of simulation (heteroscedastic case) using the method "HSFMM-MRF" or "HSFMM" is about 30 hours. The Gibbs sampling procedure in the Monte Carlo EM algorithm is the most time-consuming step. To expedite the simulation procedure, 50 trials of simulations was parallelized using 50 cores simultaneously on two servers. This process took approximately 5 days.