Received: 3 July 2020 | Accepted: 28 June 2021

DOI: 10.1002/env.2696

RESEARCH ARTICLE

WILEY

Spatial cluster detection with threshold quantile regression

Junho Lee[1] | Ying Sun[1] | Huixia Judy Wang[2]

[1]Statistics Program, CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

[2]Department of Statistics, George Washington University, Washington, DC, USA

Correspondence
Junho Lee, Statistics Program, CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia.
Email: junho.lee@kaust.edu.sa

Funding information
King Abdullah University of Science and Technology, Grant/Award Number: OSR2019-CRG7-3800; National Science Foundation, Grant/Award Numbers: DMS-1712760, IR/D program

Abstract

Spatial cluster detection, which is the identification of spatial units adjacent in space associated with distinctive patterns of data of interest relative to background variation, is useful for discerning spatial heterogeneity in regression coefficients. Some real studies with regression-based models on air quality data show that there exists not only spatial heterogeneity but also heteroscedasticity between air pollution and its predictors. Since the low air quality is a well-known risk factor for mortality, various cardiopulmonary diseases, and preterm birth, the analysis at the tail would be of more interest than the center of air pollution distribution. In this article, we develop a spatial cluster detection approach using a threshold quantile regression model to capture the spatial heterogeneity and heteroscedasticity. We introduce two threshold variables in the quantile regression model to define a spatial cluster. The proposed test statistic for identifying the spatial cluster is the supremum of the Wald process over the space of threshold parameters. We establish the limiting distribution of the test statistic under the null hypothesis that the quantile regression coefficient is the same over the entire spatial domain at the given quantile level. The performance of our proposed method is assessed by simulation studies. The proposed method is also applied to analyze the particulate matter ($PM_{2.5}$) concentration and aerosol optical depth (AOD) data in the Northeastern United States in order to study geographical heterogeneity in the association between AOD and $PM_{2.5}$ at different quantile levels.

KEYWORDS

quantile regression, spatial cluster detection, spatial threshold effect, threshold model, threshold quantile regression

1 | INTRODUCTION

Fine particulate ($PM_{2.5}$) is a well-known harmful air pollutant as a risk factor for mortality (Samoli et al., 2008), various cardiopulmonary diseases (Dominici et al., 2006; Pope & Dockery, 2006), and preterm birth (Chang et al., 2012). The satellite-derived aerosol optical depth (AOD) is a proxy measurement of particle air pollution data since it measures light extinction due to particles (e.g., dust, smoke, pollution) in the atmospheric column. Previous studies showed that $PM_{2.5}$ concentrations have positive associations with AOD (Chu et al., 2016; Grantham et al., 2018; Ma et al., 2016; Yu et al., 2017). For spatial data, it is often assumed that regression coefficients are homogeneous across the entire spatial domain of interest. However, real applications often show spatial heterogeneity in regression coefficients. That is,

https://doi.org/10.1002/env.2696

regression coefficients may be different in specific subregions from the rest of the area. For example, the $PM_{2.5}$–AOD data studied in Section 5 demonstrates spatial heterogeneity in the relationship between $PM_{2.5}$ and AOD (Figure 3).

Spatial heterogeneity in regression coefficients has been addressed by clustered varying coefficient regression for the spatial data. Lawson et al. (2014) proposed the grouped spatial varying coefficient regression when the total number of groups is given. Recently, Lee, Gangnon, and Zhu (2017) and Lee et al. (2020) proposed spatial cluster detection approaches of regression coefficients. Spatial cluster detection is a statistical methodology to identify observations adjacent in space that are associated with distinctive patterns of data of interest relative to background variation (Gangnon, 2010, 2012; Gangnon & Clayton, 2000; Kulldorff, 1997; Kulldorff & Nagarwalla, 1995). However, the aforementioned spatially clustered varying coefficient regression approaches are developed for mean regression. Mean regression model assumes the constant relationship between a response and covariates across the population. These constant regression coefficients are based on the homoscedastic assumption that random errors are drawn from identical distributions. Thus, mean regression is fragile when the homoscedastic assumption is violated, which is often seen in medicine and survival analysis, financial and economic statistics, and environmental modeling (Yu et al., 2003). Grange et al. (2016) studied black carbon (BC) contributions to $PM_{2.5}$ in London, United Kingdom, and showed that these two variables did not follow a mean rate of change; the contribution of BC was getting bigger as $PM_{2.5}$ was moving to the upper tail of its distribution. Recently, Yoshida (2021) showed heteroscedasticity of air quality data in Beijing, China, and considered the model at the tail of $PM_{2.5}$ instead of at central. Since a number of studies have shown that high levels of $PM_{2.5}$ are fatal (Chang et al., 2012; Dominici et al., 2006; Pope & Dockery, 2006; Samoli et al., 2008), in $PM_{2.5}$ studies, upper quantiles would be of more interest than the median or mean. Furthermore, the $PM_{2.5}$–AOD data studied in Section 5 demonstrates not only spatial heterogeneity but also heteroscedasticity (Figure 3).

Quantile regression (Koenker & Bassett, 1978) provides a natural and automatic way to capture the unknown data heteroscedasticity since it enables us to model the impact of predictors at different quantiles of the response distribution; see Koenker (2005) for a more detailed review of quantile regression. There have been a number of studies on threshold quantile regression (Cai, 2010; Cai & Stander, 2008; Caner, 2002; Galvao et al., 2011, 2014; He & Zhu, 2003; Horowitz & Spokoiny, 2002; Lee et al., 2011; Otsu, 2008; Zheng, 1998). More recently, Zhang et al. (2014) and Tang et al. (2015) developed procedures for testing change points due to a covariate threshold, and Kuan et al. (2017) and Su and Xu (2019) studied confidence intervals for the estimated threshold parameter in regression quantiles. Threshold quantile regression models consider piecewise effects in subregions divided by one threshold variable with jumps occurring at the unknown change points. However, although there are some previous studies on quantile regression with spatial data (Hallin et al., 2009; McMillen, 2013), there appears to be very limited work for spatial cluster analysis.

In this article, we propose a novel approach that enables us to not only address the spatial heterogeneity in regression coefficients but also accommodate data heteroscedasticity. We define a set of potential spatial clusters by considering geographical coordinates of the observations as threshold variables and introducing two threshold parameters. And then, we first test if there exists a spatial cluster against the null hypothesis that the quantile regression coefficient is the same over the entire spatial domain at the given quantile level. This test for the existence of a spatial cluster requires the limiting distribution of the supremum test statistic under the null hypothesis to control the Type I error. If the test rejects the null hypothesis, then we choose the cluster that gives the largest test statistic among all candidate clusters, as the spatial cluster estimator. The main challenge in developing our method comes from the fact that the limiting null distribution is not pivotal. In similar situations in the mean regression approach, a parametric bootstrap was adopted to obtain the *p*-value (Lee, Gangnon, & Zhu, 2017; Lee et al., 2020, 2021). However, in the quantile regression setup, a *p*-value via a Monte Carlo method is computationally costly. Thus, we resolve this challenge by proposing a simulation-based algorithm for calculating the critical values. We believe that our proposed method is the first of its kind to address the spatial heterogeneity issue in the quantile regression coefficients. We assess the performance of our proposed method via simulation studies and the analysis of the air quality data in the Northeastern United States. These studies suggest that the proposed method provides better performance than the mean regression approach (Lee, Gangnon, & Zhu, 2017) by producing robust results to the heavy-tailed distribution and capturing the heteroscedasticity.

The remainder of this article is organized as follows. In Section 2, we introduce the proposed spatial threshold quantile regression framework. In Section 3, we define the test statistic, present its asymptotic null distribution, and introduce one simulation-based algorithm for approximating the asymptotic critical values for the test statistic. We also introduce a sequential scheme for the identification and estimation of multiple spatial clusters. In Section 4, we conduct simulation studies to evaluate and compare the proposed method with existing approaches. Section 5 presents a real data application by studying the impacts of AOD on $PM_{2.5}$ from the Northeastern United States. Lastly, Section 6 contains some discussion and conclusions. Proofs are provided in the Supplementary Material.

2 | STATISTICAL MODEL

In this section, we construct a statistical model to capture the heteroscedasticity and spatial heterogeneity in regression coefficients. First, we introduce the conditional quantile function on the spatial data to address the heteroscedasticity. And then, we extend the conditional quantile function with threshold variables to define the spatial cluster. We assign the separate regression coefficient to the cluster to take the spatial heterogeneity into account.

Let $y_{\boldsymbol{s}} \in \mathbb{R}$ and $\boldsymbol{x}_{\boldsymbol{s}} \in \mathbb{R}^p$ denote the dependent variable and the covariate vector at $\boldsymbol{s}$, respectively, where $\boldsymbol{s} = (s_1, s_2)^\top$ is a geographical location on the unit square $[0,1]^2 \in \mathbb{R}^2$. In this article, we assume that $\{(y_{\boldsymbol{s}}, \boldsymbol{w}_{\boldsymbol{s}}^\top)^\top \mid \boldsymbol{s} \in [0,1]^2\}$ is an independent process although the results can be generalized to weak stationary processes, where $\boldsymbol{w}_{\boldsymbol{s}} = (\boldsymbol{s}^\top, \boldsymbol{x}_{\boldsymbol{s}}^\top)^\top$.

Let $Q_{y_{\boldsymbol{s}}}(\tau|\boldsymbol{w}_{\boldsymbol{s}})$ denote the conditional τth quantile of $y_{\boldsymbol{s}}$ given $\boldsymbol{w}_{\boldsymbol{s}}$, where $\tau \in (0,1)$. Then, we assume that when there is no spatial cluster, the effect of $\boldsymbol{x}_{\boldsymbol{s}}$ on the τth quantile of $y_{\boldsymbol{s}}$ is linear and the same across space. If there is a spatial cluster, then we assume that $\boldsymbol{x}_{\boldsymbol{s}}$ shows distinctive association to the τth quantile of $y_{\boldsymbol{s}}$ within the cluster relative to the rest of the region. Thus, for a given $\tau \in \mathcal{T} = [\tau_L, \tau_U] \subset (0,1)$, we could consider the following hypotheses:

$$\begin{aligned} &H_0 : Q_{y_{\boldsymbol{s}}}(\tau|\boldsymbol{w}_{\boldsymbol{s}}) = \boldsymbol{x}_{\boldsymbol{s}}^\top \boldsymbol{\theta}_1(\tau), \\ \text{versus } &H_1 : Q_{y_{\boldsymbol{s}}}(\tau|\boldsymbol{w}_{\boldsymbol{s}}) = \mathcal{I}(s_1 \notin [a_1^*, b_1^*] \text{ or } s_2 \notin [a_2^*, b_2^*]) \cdot \boldsymbol{x}_{\boldsymbol{s}}^\top \boldsymbol{\theta}_1(\tau) \\ &\qquad + \mathcal{I}(s_1 \in [a_1^*, b_1^*],\ s_2 \in [a_2^*, b_2^*]) \cdot \boldsymbol{x}_{\boldsymbol{s}}^\top \boldsymbol{\theta}_2(\tau), \text{ for some } a_1^*, a_2^*, b_1^*, b_2^*, \end{aligned}$$

where $\mathcal{I}(\cdot)$ is the indicator function, a_1^*, a_2^*, b_1^*, and $b_2^* \in [0,1]$ are the threshold parameters such that $a_1^* < b_1^*$, $a_2^* < b_2^*$, and $\boldsymbol{\theta}_1(\tau) \neq \boldsymbol{\theta}_2(\tau)$ for $\boldsymbol{s} \in [a_1^*, b_1^*] \times [a_2^*, b_2^*]$. That is, for a given τ, the quantile regression model has the uniform coefficient $\boldsymbol{\theta}_1(\tau) \in \mathbb{R}^p$ over all $\boldsymbol{s} \in [0,1]^2$ under H_0, while H_1 assumes an additional quantile regression coefficient $\boldsymbol{\theta}_2(\tau) \in \mathbb{R}^p$. A rectangular spatial cluster is defined to be $[a_1^*, b_1^*] \times [a_2^*, b_2^*]$ by the threshold parameters, and each coordinate of $\boldsymbol{s}$, s_1 and s_2, plays a role of the threshold variable in a threshold regression model.

For convenience, we reparameterize as $\boldsymbol{\beta}_{(1)}(\tau) = \boldsymbol{\theta}_1(\tau)$ and $\boldsymbol{\beta}_{(2)}(\tau) = \boldsymbol{\theta}_2(\tau) - \boldsymbol{\theta}_1(\tau)$, and re-express the hypotheses:

$$\begin{aligned} &H_0 : Q_{y_{\boldsymbol{s}}}(\tau|\boldsymbol{w}_{\boldsymbol{s}}) = \boldsymbol{z}_{\boldsymbol{s}}(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)^\top \boldsymbol{\beta}(\tau) \text{ with } \boldsymbol{\beta}_{(2)}(\tau) = \boldsymbol{0}, \text{ for all } (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \in \boldsymbol{\Gamma}^2, \\ \text{versus } &H_1 : Q_{y_{\boldsymbol{s}}}(\tau|\boldsymbol{w}_{\boldsymbol{s}}) = \boldsymbol{z}_{\boldsymbol{s}}(\boldsymbol{\gamma}_1^*, \boldsymbol{\gamma}_2^*)^\top \boldsymbol{\beta}(\tau) \text{ with } \boldsymbol{\beta}_{(2)}(\tau) \neq \boldsymbol{0}, \text{ for some } (\boldsymbol{\gamma}_1^*, \boldsymbol{\gamma}_2^*) \in \boldsymbol{\Gamma}^2, \end{aligned} \tag{1}$$

where $\boldsymbol{z}_{\boldsymbol{s}}(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) = (\boldsymbol{x}_{\boldsymbol{s}}^\top,\ \mathcal{I}(\boldsymbol{s} \in \boldsymbol{\gamma}_1 \times \boldsymbol{\gamma}_2) \cdot \boldsymbol{x}_{\boldsymbol{s}}^\top)^\top$, $\boldsymbol{\gamma}_1 = (a_1, b_1)^\top \in \boldsymbol{\Gamma}$, $\boldsymbol{\gamma}_2 = (a_2, b_2)^\top \in \boldsymbol{\Gamma}$, $\boldsymbol{\Gamma} = \{(\gamma_L, \gamma_U)^\top | \ 0 \leq \gamma_L < \gamma_U \leq 1\}$, $\boldsymbol{\gamma}_1 \times \boldsymbol{\gamma}_2 = \{\boldsymbol{s}_i \mid s_{i1} \in [a_1, b_1],\ s_{i2} \in [a_2, b_2],\ i = 1, \ldots, n\}$, $\boldsymbol{\beta}(\tau) = (\boldsymbol{\beta}_{(1)}(\tau)^\top, \boldsymbol{\beta}_{(2)}(\tau)^\top)^\top$, $\boldsymbol{\beta}_{(1)}(\tau) = \boldsymbol{\theta}_1(\tau)$, and $\boldsymbol{\beta}_{(2)}(\tau) = \boldsymbol{\theta}_2(\tau) - \boldsymbol{\theta}_1(\tau)$. That is, a rectangular spatial cluster is defined to be $\boldsymbol{\gamma}_1 \times \boldsymbol{\gamma}_2$ by two threshold parameter vectors $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. Then, when the threshold parameters $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ are known, with the given data $\{(y_{\boldsymbol{s}_i}, \boldsymbol{w}_{\boldsymbol{s}_i}^\top)^\top\}_{i=1}^n$ and a given $\tau \in \mathcal{T}$, we can estimate the quantile regression coefficient $\boldsymbol{\beta}(\tau)$ by the following estimator:

$$\hat{\boldsymbol{\beta}}(\tau, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) = \arg\min_{\boldsymbol{b} \in \mathbb{R}^{2p}} n^{-1} \sum_{i=1}^n \rho_\tau(y_{\boldsymbol{s}_i} - \boldsymbol{z}_{\boldsymbol{s}_i}(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)^\top \boldsymbol{b}), \tag{2}$$

where $\rho_\tau(u) = u \cdot \{\tau - \mathcal{I}(u \leq 0)\}$ is the check function (Koenker & Bassett, 1978).

3 | TEST STATISTIC AND ESTIMATION

3.1 | Test statistic

As shown in Lemma C.1 in the Supplementary Material, with known $(\boldsymbol{\gamma}_1^*, \boldsymbol{\gamma}_2^*)$, $\hat{\boldsymbol{\beta}}_{(2)}(\tau, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \xrightarrow{p} \boldsymbol{0}$ for each $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \in \boldsymbol{\Gamma} \times \boldsymbol{\Gamma}$ when H_0 is true, while $\hat{\boldsymbol{\beta}}_{(2)}(\tau, \boldsymbol{\gamma}_1^*, \boldsymbol{\gamma}_2^*) \xrightarrow{p} \boldsymbol{\beta}_{(2)}(\tau) \neq \boldsymbol{0}$ when H_1 is true, where "$\xrightarrow{p}$" denotes convergence in probability. Therefore, it is reasonable to reject H_0 if $\hat{\boldsymbol{\beta}}_{(2)}(\tau, \boldsymbol{\gamma}_1^*, \boldsymbol{\gamma}_2^*)$ is far from $\boldsymbol{0}$ enough. However, since the true value of the threshold parameter $(\boldsymbol{\gamma}_1^*, \boldsymbol{\gamma}_2^*)$ is unknown, it is not adequate to choose $\hat{\boldsymbol{\beta}}_{(2)}(\tau, \boldsymbol{\gamma}_1^*, \boldsymbol{\gamma}_2^*)$ as the test statstic for the existence of a spatial cluster. Instead, we can consider to reject H_0 if the magnitude of $\hat{\boldsymbol{\beta}}_{(2)}(\tau, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ is large enough for some $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \in \boldsymbol{\Gamma}^2$ since $\hat{\boldsymbol{\beta}}_{(2)}(\tau, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \approx \boldsymbol{0}$ for any $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \in \boldsymbol{\Gamma}^2$ when H_0 is true. Thus, we choose the supremum of the Wald statistic as the test statistic

$$SW_n(\tau) = \sup_{(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \in \boldsymbol{\Gamma}^2} n \hat{\boldsymbol{\beta}}_{(2)}(\tau, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)^\top \{\boldsymbol{V}_{22}(\tau, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)\}^{-1} \hat{\boldsymbol{\beta}}_{(2)}(\tau, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2), \tag{3}$$

where $\boldsymbol{V}_{22}(\tau, \gamma_1, \gamma_2)$ is the asymptotic covariance matrix of $\sqrt{n}\hat{\boldsymbol{\beta}}_{(2)}(\tau, \gamma_1, \gamma_2)$ under H_0 and can be replaced by a suitable consistent estimate in practice. Further, the distribution of the test statistic $SW_n(\tau)$ is required under H_0 to control the Type I error. Thus, we establish its limiting distribution under suitable regularity conditions and H_0.

Let $\boldsymbol{\beta}^*_{(1)}(\tau) \in \mathbb{R}^p$ denote the unique solution to $\mathrm{E}[(\tau - \mathcal{I}\{y_{\boldsymbol{s}} \le \boldsymbol{x}_{\boldsymbol{s}}^\top \boldsymbol{\beta}^*_{(1)}(\tau)\}) \cdot \boldsymbol{x}_{\boldsymbol{s}}] = \boldsymbol{0}$, and let $\boldsymbol{\beta}^*(\tau) = (\boldsymbol{\beta}^*_{(1)}(\tau)^\top, \boldsymbol{0}^\top)^\top \in \mathbb{R}^{2p}$. Let $\ell^\infty(\mathcal{T} \times \boldsymbol{\Gamma}^2)$ denote the space of all bounded functions on $\mathcal{T} \times \boldsymbol{\Gamma}^2$, and let $(\ell^\infty(\mathcal{T} \times \boldsymbol{\Gamma}^2))^{2p}$ denote the $(2p)$-product space of $\ell^\infty(\mathcal{T} \times \boldsymbol{\Gamma}^2)$. And then, we make the following the regularity conditions C1–C5.

C1: $\{(y_{\boldsymbol{s}}, \boldsymbol{w}_{\boldsymbol{s}}^\top)^\top, \boldsymbol{s} \in [0, 1]^2\}$ is an independent process.
C2: $\mathrm{E}[\|\boldsymbol{x}_{\boldsymbol{s}}\|^q] < \infty$ for some $q > 2$.
C3: Let $F(\cdot|\boldsymbol{w})$ denote the conditional distribution function of $y_{\boldsymbol{s}}$ given $\boldsymbol{w}_{\boldsymbol{s}} = \boldsymbol{w}$. Assume that $F(\cdot|\boldsymbol{w})$ has a Lebesgue density $f(\cdot|\boldsymbol{w})$ such that

(i) $|f(y|\boldsymbol{w})| \le C_f$ on the support of $(y_{\boldsymbol{s}}, \boldsymbol{w}_{\boldsymbol{s}}^\top)^\top$ for some $C_f > 0$,
(ii) $|f(y_1|\boldsymbol{w}) - f(y_2|\boldsymbol{w})| \to 0$ as $|y_1 - y_2| \to 0$ for each fixed $\boldsymbol{w}$.

C4: The threshold variable $\boldsymbol{s}$ has a continuous distribution.
C5: $\boldsymbol{\Omega}_0(\gamma_1, \gamma_2)$ is positive definite for each $(\gamma_1, \gamma_2) \in \boldsymbol{\Gamma}^2$, and $\boldsymbol{\Omega}_1(\tau, \gamma_1, \gamma_2)$ is positive definite for each $(\tau, \gamma_1, \gamma_2) \in \mathcal{T} \times \boldsymbol{\Gamma}^2$, where

(i) $\boldsymbol{\Omega}_0(\gamma_1, \gamma_2) = \mathrm{E}\left[\boldsymbol{z}_{\boldsymbol{s}}(\gamma_1, \gamma_2)\boldsymbol{z}_{\boldsymbol{s}}(\gamma_1, \gamma_2)^\top\right]$ for $\gamma_1, \gamma_2 \in \boldsymbol{\Gamma}$,
(ii) $\boldsymbol{\Omega}_1(\tau, \gamma_1, \gamma_2) = \mathrm{E}\left[f(\boldsymbol{x}_{\boldsymbol{s}}^\top \boldsymbol{\beta}^*_{(1)}(\tau)|\boldsymbol{w}_{\boldsymbol{s}})\boldsymbol{z}_{\boldsymbol{s}}(\gamma_1, \gamma_2)\boldsymbol{z}_{\boldsymbol{s}}(\gamma_1, \gamma_2)^\top\right]$.

C1 guarantees the independent observations, but it can be generalized to weak stationary processes. C2 is a moment condition. C3, for each $\boldsymbol{w}$, guarantees the (i) uniformly bounded and (ii) continuous density which is standard in the quantile regression literatures (Angrist et al., 2006; Galvao et al., 2011, 2014; Su & Xu, 2019). C4 is a standard condition in the threshold quantile regression literatures (Galvao et al., 2011, 2014; Hansen, 1996, 2000; Su & Xu, 2019), and it is satisfied by spatial data (e.g., bivariate uniform distribution on $[0, 1]^2$). C5 guarantees that the matrices $\boldsymbol{\Omega}_0(\gamma_1, \gamma_2)$ and $\boldsymbol{\Omega}_1(\tau, \gamma_1, \gamma_2)$ do not degenerate for each $(\gamma_1, \gamma_2) \in \boldsymbol{\Gamma}^2$ and $(\tau, \gamma_1, \gamma_2) \in \mathcal{T} \times \boldsymbol{\Gamma}^2$, respectively.

Theorem 1. *For a given $\tau \in \mathcal{T}$, and under the regularity conditions C1–C5 and H_0, we have*

$$\sqrt{n}\{\hat{\boldsymbol{\beta}}(\tau, \gamma_1, \gamma_2) - \boldsymbol{\beta}^*(\tau)\} \;\Rightarrow\; \boldsymbol{\Omega}_1(\tau, \gamma_1, \gamma_2)^{-1}\boldsymbol{W}(\tau, \gamma_1, \gamma_2) \text{ in } (\ell^\infty(\mathcal{T} \times \boldsymbol{\Gamma}^2))^{2p},$$

where "$\Rightarrow$" denotes weak convergence, and $\boldsymbol{W}(\tau, \gamma_1, \gamma_2)$ is a zero-mean, continuous Gaussian process on $\mathcal{T} \times \boldsymbol{\Gamma}^2$ with covariance kernel

$$\mathrm{E}[\boldsymbol{W}(\tau, \gamma_1, \gamma_2)\boldsymbol{W}(\tau, \gamma_1, \gamma_2)^\top] = \tau(1 - \tau) \cdot \boldsymbol{\Omega}_0(\gamma_1, \gamma_2).$$

Theorem 1 presents the asymptotic null distribution of the quantile regression estimator, $\hat{\boldsymbol{\beta}}(\tau, \gamma_1, \gamma_2)$ in (2), when the regularity conditions C1–C5 hold. Thus, from Theorem 1, the asymptotic null distribution of $SW_n(\tau)$ in (3) can be derived as in the following corollary.

Corollary 1. *For a given $\tau \in \mathcal{T}$, and under the regularity conditions C1–C5 and H_0,*

$$SW_n(\tau) \;\Rightarrow\; \sup_{(\gamma_1, \gamma_2) \in \boldsymbol{\Gamma}^2} \boldsymbol{S}(\tau, \gamma_1, \gamma_2)^\top\{\boldsymbol{V}_{22}(\tau, \gamma_1, \gamma_2)\}^{-1}\boldsymbol{S}(\tau, \gamma_1, \gamma_2), \tag{4}$$

where $\boldsymbol{S}(\tau, \gamma_1, \gamma_2) = \boldsymbol{R}\boldsymbol{\Omega}_1(\tau, \gamma_1, \gamma_2)^{-1}\boldsymbol{W}(\tau, \gamma_1, \gamma_2)$, $\boldsymbol{R} = [\ \boldsymbol{O},\ \boldsymbol{I}_p\ ]_{p \times 2p}$, and $\boldsymbol{V}_{22}(\tau, \gamma_1, \gamma_2) = E[\boldsymbol{S}(\tau, \gamma_1, \gamma_2)\boldsymbol{S}(\tau, \gamma_1, \gamma_2)^\top] = \tau(1 - \tau)\boldsymbol{R}\boldsymbol{\Omega}_1(\tau, \gamma_1, \gamma_2)^{-1}\boldsymbol{\Omega}_0(\gamma_1, \gamma_2)\boldsymbol{\Omega}_1(\tau, \gamma_1, \gamma_2)^{-1}\boldsymbol{R}^\top$.

3.2 | Implementation

For the implementation of the limiting distribution of $SW_n(\tau)$ given by (4), we consider the estimates of $\boldsymbol{\Omega}_0$, $\boldsymbol{\Omega}_1$, and $\boldsymbol{V}_{22}$ as

$$\hat{\boldsymbol{\Omega}}_0(\gamma_1, \gamma_2) = n^{-1}\sum_{i=1}^{n} \boldsymbol{z}_{\boldsymbol{s}_i}(\gamma_1, \gamma_2)\boldsymbol{z}_{\boldsymbol{s}_i}(\gamma_1, \gamma_2)^\top, \tag{5}$$

$$\hat{\boldsymbol{\Omega}}_1(\tau,\gamma_1,\gamma_2) = n^{-1}\sum_{i=1}^{n} \hat{f}_{\boldsymbol{s}_i}\boldsymbol{z}_{\boldsymbol{s}_i}(\gamma_1,\gamma_2)\boldsymbol{z}_{\boldsymbol{s}_i}(\gamma_1,\gamma_2)^\top, \tag{6}$$

$$\hat{\boldsymbol{V}}_{22}(\tau,\gamma_1,\gamma_2) = \tau(1-\tau)\boldsymbol{R}\hat{\boldsymbol{\Omega}}_1(\tau,\gamma_1,\gamma_2)^{-1}\hat{\boldsymbol{\Omega}}_0(\gamma_1,\gamma_2)\hat{\boldsymbol{\Omega}}_1(\tau,\gamma_1,\gamma_2)^{-1}\boldsymbol{R}^\top, \tag{7}$$

where $\hat{f}_{\boldsymbol{s}_i}$ is estimated via the Hendricks–Koenker sandwich (Hendricks & Koenker, 1992; Koenker, 2005):

$$\hat{f}_{\boldsymbol{s}_i}\left(\boldsymbol{x}_{\boldsymbol{s}_i}^\top\hat{\boldsymbol{\beta}}_{(1)}^*(\tau)\right) = \max\left\{0, \frac{2h_n}{\boldsymbol{x}_{\boldsymbol{s}_i}^\top\left(\hat{\boldsymbol{\beta}}_{(1)}^*(\tau+h_n)-\hat{\boldsymbol{\beta}}_{(1)}^*(\tau-h_n)\right)}\right\}$$

with $h_n = O(n^{-1/3})$. In the implementation, the bandwidth h_n is chosen by the `bandwidth.rq` function in the `quantreg` package for `R` (R Core Team, 2017).

However, the main challenge remains in estimating the critical value directly from the asymptotic result (4). The limiting null distribution is not pivotal, and thus in similar situations in the mean regression approach, a parametric bootstrap was adopted to obtain the p-value (Lee, Gangnon, & Zhu, 2017; Lee et al., 2020). However, a Monte Carlo method is computationally costly in the quantile regression setup. Thus, we propose a simulation-based algorithm to calculate the critical values for a spatial cluster's existence. The simulation-based method was also used in Hansen (1996), Galvao et al. (2014), and Su and Xu (2019) in different quantile regression settings. The quantile regression estimator (2) requires the computational complexity of $O(n^{1.25}p^3\log n)$ (Portnoy & Koenker, 1997). Thus, the Monte Carlo method requires the computational complexity of $O(BCGn^{1.25}p^3\log n)$ to find multiple clusters, where B is the number of simulations to obtain the p-value or a critical value, C is the number of true clusters, and G is the number of potential clusters over $\boldsymbol{\Gamma}^2$. In contrast, the simulation-based approach only requires $O((BG+C)n^{1.25}p^3\log n)$ since we only need single B simulations for the approximate critical value. An approximate critical value for the test statistic $SW_n(\tau)$ can be computed as in the following steps:

(i) Generate $\{u_{\boldsymbol{s}_i}^b\}_{i=1}^n$ independently from the uniform distribution on $[0,1]$ for each $b=1,\ldots,B$, where B is a large positive integer.

(ii) Set $\boldsymbol{W}_n^b(\tau,\gamma_1,\gamma_2) = n^{-1/2}\sum_{i=1}^n\{\tau-\mathcal{I}(u_{\boldsymbol{s}_i}^b\le\tau)\}\boldsymbol{z}_{\boldsymbol{s}_i}(\gamma_1,\gamma_2)$ for each $b=1,\ldots,B$.

(iii) For each $b=1,\ldots,B$, compute $\widehat{SW}_n^b(\tau) = \max_{(\gamma_1,\gamma_2)\in\boldsymbol{\Gamma}^2}\ \widehat{W}_n^b(\tau,\gamma_1,\gamma_2)$, where

$$\begin{aligned}&\widehat{W}_n^b(\tau,\gamma_1,\gamma_2)\\&\quad=\boldsymbol{W}_n^b(\tau,\gamma_1,\gamma_2)^\top\hat{\boldsymbol{\Omega}}_1(\tau,\gamma_1,\gamma_2)^{-1}\boldsymbol{R}^\top\{\hat{\boldsymbol{V}}_{22}(\tau,\gamma_1,\gamma_2)\}^{-1}\boldsymbol{R}\hat{\boldsymbol{\Omega}}_1(\tau,\gamma_1,\gamma_2)^{-1}\boldsymbol{W}_n^b(\tau,\gamma_1,\gamma_2).\end{aligned}$$

(iv) For the significance level $\alpha\in(0,1)$, take the empirical $(1-\alpha)$-quantile of the simulated sample $\{\widehat{SW}_n^1(\tau),\ldots,\widehat{SW}_n^B(\tau)\}$ as the approximate critical value $\hat{c}_{1-\alpha}^B$.

In practice, we can take the maximum of $\widehat{W}_n^b(\tau,\gamma_1,\gamma_2)$ in step (iii) over the discretized $\boldsymbol{\Gamma}^2$. That is, we first discretize the unit interval $[0,1]$ into $\{\gamma_0,\gamma_1,\ldots,\gamma_l\}$, where $\gamma_0=0$, $\gamma_l=1$, and $\gamma_k<\gamma_{k'}$ for $k<k'$. Then, we can get the discretized $\boldsymbol{\Gamma}^2$, where $\boldsymbol{\Gamma}=\left\{(\gamma_L,\gamma_U)^\top \mid \gamma_L,\gamma_U\in\{\gamma_0,\gamma_1,\ldots,\gamma_l\},\ \gamma_L<\gamma_U\right\}$. Thus, the number of potential clusters over the discretized $\boldsymbol{\Gamma}^2$ will be $G=|\boldsymbol{\Gamma}^2|=4^{-1}l^2(l+1)^2$, where $|\cdot|$ denotes the cardinality of a set. And then, we reject H_0 if $SW_n(\tau)>\hat{c}_{1-\alpha}^B$. If the test rejects H_0, then it suggests the presence of threshold effects, and thus we can estimate the spatial cluster together with the threshold parameter. It is reasonable to choose the cluster that gives the largest test statistic among all candidate clusters as the spatial cluster estimator. Furthermore, its corresponding (γ_1,γ_2) is considered as the threshold parameter estimator $(\hat{\gamma}_1^*,\hat{\gamma}_2^*)$:

$$(\hat{\gamma}_1^*,\hat{\gamma}_2^*) = \arg\sup_{(\gamma_1,\gamma_2)\in\boldsymbol{\Gamma}^2} n\hat{\boldsymbol{\beta}}_{(2)}(\tau,\gamma_1,\gamma_2)^\top\{\boldsymbol{V}_{22}(\tau,\gamma_1,\gamma_2)\}^{-1}\hat{\boldsymbol{\beta}}_{(2)}(\tau,\gamma_1,\gamma_2), \tag{8}$$

where $\hat{\gamma}_1^*\times\hat{\gamma}_2^*$ is the spatial cluster estimator.

This implementation is based on the fact that $\sqrt{n}\{\hat{\boldsymbol{\beta}}(\tau,\gamma_1,\gamma_2)-\boldsymbol{\beta}^*(\tau)\}$ admits the Bahadur representation under the regularity conditions and H_0 (see the proof of Theorem 1 in the Supplementary Material). The first term of its

Bahadur representation is $n^{-1/2}\sum_{i=1}^{n}[\tau - \mathcal{I}\{y_{s_i} \leq \boldsymbol{x}_{s_i}^{\top}\boldsymbol{\beta}_{(1)}^{*}(\tau)\}]\boldsymbol{z}_{s_i}(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$. Since $\mathcal{I}\{y_{s_i} \leq \boldsymbol{x}_{s_i}^{\top}\boldsymbol{\beta}_{(1)}^{*}(\tau)\}, i = 1, \ldots, n$, are independent Bernoulli trials with the success probability τ, we could replace them with $\left\{\mathcal{I}(u_{s_i} \leq \tau)\right\}_{i=1}^{n}$, where $u_{s_i}, i = 1, \ldots, n$, are generated from *iid* $\mathcal{U}(0, 1)$.

Let $\boldsymbol{\phi}_\tau = \left(\boldsymbol{\beta}(\tau)^{\top}, \boldsymbol{\gamma}_1^{\top}, \boldsymbol{\gamma}_2^{\top}\right)^{\top} \in \mathbb{R}^{2p} \times \boldsymbol{\Gamma}^2$ and $\boldsymbol{\beta}_*(\tau) \in \mathbb{R}^{2p}$ denote the unique solution to $\mathrm{E}[(\tau - \mathcal{I}\{y_{\boldsymbol{s}} \leq \boldsymbol{z}_{\boldsymbol{s}}(\boldsymbol{\gamma}_1^*, \boldsymbol{\gamma}_2^*)^{\top}\boldsymbol{\beta}_*(\tau)\}) \cdot \boldsymbol{x}_{\boldsymbol{s}}] = \boldsymbol{0}$. And then, we make the following assumption C6 to establish the consistency of $\hat{\boldsymbol{\phi}}_\tau = \left(\hat{\boldsymbol{\beta}}(\tau, \hat{\boldsymbol{\gamma}}_1^{*\top}, \hat{\boldsymbol{\gamma}}_2^{*\top})^{\top}, \hat{\boldsymbol{\gamma}}_1^{*\top}, \hat{\boldsymbol{\gamma}}_2^{*\top}\right)^{\top}$ under H_1.

C6: Let $\Delta(\boldsymbol{z}_{\boldsymbol{s}}, \boldsymbol{\phi}_\tau) = \boldsymbol{z}_{\boldsymbol{s}}(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)^{\top}\boldsymbol{\beta}(\tau) - \boldsymbol{z}_{\boldsymbol{s}}(\boldsymbol{\gamma}_1^*, \boldsymbol{\gamma}_2^*)^{\top}\boldsymbol{\beta}_*(\tau)$. Then, there exists $c^* > 0$ such that $P(|\Delta(\boldsymbol{z}_{\boldsymbol{s}}, \boldsymbol{\phi}_\tau)| > c^*) > 0$ for all $\boldsymbol{\phi}_\tau \in \mathbb{R}^{2p} \times \boldsymbol{\Gamma}^2$ such that $\boldsymbol{\phi}_\tau \neq \boldsymbol{\phi}_\tau^*$, where $\boldsymbol{\phi}_\tau^* = \left(\boldsymbol{\beta}_*(\tau)^{\top}, \boldsymbol{\gamma}_1^{*\top}, \boldsymbol{\gamma}_2^{*\top}\right)^{\top}$.

Theorem 2. *For a given $\tau \in \mathcal{T}$, and under the regularity conditions C1–C6, we have $\hat{\boldsymbol{\phi}}_\tau = \boldsymbol{\phi}_\tau^* + o_p(1)$.*

3.3 | Identification and estimation of multiple clusters

We have introduced the procedure for detecting and estimating the existence of a spatial cluster. However, all of these are based on the single cluster assumption when H_1 is true, while in practice, more than one cluster may exist in the study area. Thus, we propose a sequential procedure to identify multiple clusters, where the non-cluster and the single cluster are also covered in the procedure as special cases. The detailed procedure for a given quantile τ is as follows.

(i) Fit the model in (1) under H_0, $Q_{y_s}(\tau|\boldsymbol{w}_{\boldsymbol{s}}) = \boldsymbol{x}^{\top}\boldsymbol{\beta}_{(1)}(\tau)$, and update the response $\tilde{y}_{s_i} = y_{s_i} - \boldsymbol{x}_{s_i}^{\top}\hat{\boldsymbol{\beta}}_{(1)}(\tau)$.

(ii) Obtain the test statistic $SW_n(\tau)$ in (3) based on the data $\{(\tilde{y}_{s_i}, \boldsymbol{w}_{s_i}^{\top})^{\top}\}_{i=1}^{n}$. If $SW_n(\tau)$ is not significant, do not reject H_0 and stop. If rejecting H_0, move to the next step.

(iii) Identify the spatial cluster and obtain the threshold parameter estimator $(\hat{\boldsymbol{\gamma}}_1^*, \hat{\boldsymbol{\gamma}}_2^*)$ as in (8), and calculate the residual $\tilde{\varepsilon}_{s_i} = \tilde{y}_{s_i} - \boldsymbol{z}_{s_i}(\hat{\boldsymbol{\gamma}}_1^*, \hat{\boldsymbol{\gamma}}_2^*)^{\top}\hat{\boldsymbol{\beta}}(\tau, \hat{\boldsymbol{\gamma}}_1^*, \hat{\boldsymbol{\gamma}}_2^*)$.

(iv) Replace the response with the residual $\tilde{y}_{s_i} = \tilde{\varepsilon}_{s_i}$ to remove the effect of $(\hat{\boldsymbol{\gamma}}_1^*, \hat{\boldsymbol{\gamma}}_2^*)$ from the data. Update $\boldsymbol{\Gamma}^2$ with $\boldsymbol{\Gamma}^2 \setminus \{(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) \mid (\boldsymbol{\gamma}_1 \times \boldsymbol{\gamma}_2) \cap (\hat{\boldsymbol{\gamma}}_1^* \times \hat{\boldsymbol{\gamma}}_2^*) \neq \emptyset\}$ to remove all the cluster candidates, which overlap the previously identified cluster $\hat{\boldsymbol{\gamma}}_1^* \times \hat{\boldsymbol{\gamma}}_2^*$, where $\emptyset$ is the empty set. And then, go to step (ii) to detect and identify a new cluster.

Let $(\hat{\boldsymbol{\gamma}}_1^{*k}, \hat{\boldsymbol{\gamma}}_2^{*k})$ and $\hat{\boldsymbol{\beta}}(\tau, \hat{\boldsymbol{\gamma}}_1^{*k}, \hat{\boldsymbol{\gamma}}_2^{*k})$ be the kth obtained threshold parameter estimator and the corresponding coefficient estimates, respectively. If we identify a total of M clusters and there is an index set $\mathcal{K} \subset \{1, 2, \ldots, M\}$, where $(\hat{\boldsymbol{\gamma}}_1^{*k} \times \hat{\boldsymbol{\gamma}}_2^{*k})$s are adjacent each other and $\hat{\boldsymbol{\beta}}(\tau, \hat{\boldsymbol{\gamma}}_1^{*k}, \hat{\boldsymbol{\gamma}}_2^{*k}) = \hat{\boldsymbol{\beta}}(\tau, \hat{\boldsymbol{\gamma}}_1^{*k'}, \hat{\boldsymbol{\gamma}}_2^{*k'})$ for $k, k' \in \mathcal{K}$, then $\cup_{k\in\mathcal{K}}(\hat{\boldsymbol{\gamma}}_1^{*k} \times \hat{\boldsymbol{\gamma}}_2^{*k})$ can be seen an approximation of an irregular shaped cluster.

4 | SIMULATION STUDIES

In this section, we conduct simulation studies to evaluate our proposed method. Our simulation studies mainly consist of four parts: model set up for the data simulation, false positive and power analysis for the hypothesis testing of the threshold effect, and cluster identification. For the comparison, we apply the mean regression approach (Lee, Gangnon, & Zhu, 2017) as well.

4.1 | Simulation design

We generate data based on the following model:

$$y_{s_i} = x_{s_i} + \delta_1 \cdot \mathcal{I}\left(\boldsymbol{s}_i \in \boldsymbol{\gamma}_1^* \times \boldsymbol{\gamma}_2^*\right) \cdot x_{s_i} + \left\{1 + \delta_2 \cdot \mathcal{I}\left(\boldsymbol{s}_i \in \boldsymbol{\gamma}_3^* \times \boldsymbol{\gamma}_4^*\right) \cdot x_{s_i}\right\} \cdot \varepsilon_{s_i}, \tag{9}$$

where x_{s_i}s are generated from *iid* $\mathcal{U}(0, 1)$, and ε_{s_i}s are *iid* random errors with zero mean and the cumulative distribution function (CDF) $F_\varepsilon(\cdot)$. We predefine two spatial clusters to be $\boldsymbol{\gamma}_1^* \times \boldsymbol{\gamma}_2^*$ and $\boldsymbol{\gamma}_3^* \times \boldsymbol{\gamma}_4^*$ with the corresponding threshold effects δ_1 and δ_2, respectively. Thus, under this model (9), we have

$$\begin{aligned} Q_{y_{s_i}}(\tau|\boldsymbol{w}_{s_i}) = {} & x_{s_i} + \delta_1 \cdot \mathcal{I}\left(\boldsymbol{s}_i \in \boldsymbol{\gamma}_1^* \times \boldsymbol{\gamma}_2^*\right) \cdot x_{s_i} \\ & + \left\{1 + \delta_2 \cdot \mathcal{I}\left(\boldsymbol{s}_i \in \boldsymbol{\gamma}_3^* \times \boldsymbol{\gamma}_4^*\right) \cdot x_{s_i}\right\} \cdot F_\varepsilon^{-1}(\tau), \end{aligned} \tag{10}$$

$$E(y_{s_i}|\boldsymbol{w}_{s_i}) = x_{s_i} + \delta_1 \cdot \mathcal{I}\left(\boldsymbol{s}_i \in \boldsymbol{\gamma}_1^* \times \boldsymbol{\gamma}_2^*\right) \cdot x_{s_i}. \tag{11}$$

TABLE 1 False positive rates in Case 0 at the nominal significance level $\alpha = 0.05$

			Quantile τ			
Errors	**Sample size**	**Grid size**	**0.5**	**0.7**	**0.9**	**Mean**
$\mathcal{N}(0,1)$	$n = 30^2$	10×10	0.051	0.041	0.071	0.047
		20×20	0.060	0.057	0.063	0.053
	$n = 50^2$	10×10	0.044	0.053	0.055	0.052
		20×20	0.049	0.054	0.055	0.063
	$n = 70^2$	10×10	0.048	0.039	0.039	0.043
		20×20	0.056	0.044	0.044	0.043
$\sigma \cdot t(2)$	$n = 30^2$	10×10	0.098	0.108	0.277	0.178
		20×20	0.089	0.140	0.341	0.149
	$n = 50^2$	10×10	0.054	0.057	0.167	0.148
		20×20	0.059	0.066	0.187	0.131
	$n = 70^2$	10×10	0.047	0.046	0.115	0.142
		20×20	0.066	0.048	0.126	0.147
	$n = 100^2$	10×10	0.054	0.052	0.064	0.149
		20×20	0.057	0.055	0.087	0.146
	$n = 150^2$	10×10	0.041	0.046	0.050	0.121
		20×20	0.048	0.046	0.055	0.119

We consider four cases. In Case 0, the threshold effects are set to be $\delta_1 = \delta_2 = 0$ for the false positive analysis. Cases 1 and 2 are for the power evaluation. In Case 1, the threshold effects are set to be $\delta_1 = \delta \neq 0$ and $\delta_2 = 0$ for the homoscedastic errors. That is, the threshold effect is uniformly over the all quantile level $\tau \in \mathcal{T} \subset [0, 1]$. However, in Case 2, $\delta_1 = 0$ and $\delta_2 = \delta \neq 0$ for the heteroscedastic errors. Thus, the threshold effect does not exist at the median ($\tau = 0.5$) while it gets larger as the quantile level is further away from the median. Threshold vectors for the spatial clusters are set to be $\boldsymbol{\gamma}_1^* = \boldsymbol{\gamma}_2^* = \boldsymbol{\gamma}_3^* = \boldsymbol{\gamma}_4^* = \boldsymbol{\gamma} = (0.3, 0.7)^\top$. That is, a cluster is predefined to be $\boldsymbol{\gamma} \times \boldsymbol{\gamma} = [0.3, 0.7]^2$ both in the homoscedastic errors (Case 1) and heteroscedastic errors (Case 2). Lastly, in Case 3, we set $\delta_1 = \delta_2 = \delta \neq 0$ and $\boldsymbol{\gamma}_1^* \times \boldsymbol{\gamma}_2^* \neq \boldsymbol{\gamma}_3^* \times \boldsymbol{\gamma}_4^*$ for the dual-cluster identification. Further, for the random errors, we consider the normal errors from $\mathcal{N}(0, 1)$ or the heavy-tailed errors from $\sigma \times t(2)$, where $t(2)$ is the Student's t-distribution with 2 degrees of freedom. We set $\sigma = 1/1.21054$ so that the normal errors and the heavy-tailed errors would have the same median absolute deviation (MAD) as 0.6745. We consider sample sizes $n = 30^2, 50^2$, or 70^2, corresponding to a 30×30, a 50×50, or a 70×70 square grid in the unit square $[0, 1] \times [0, 1]$, respectively.

From now, we simplify the notations $y_{\boldsymbol{s}_i}, x_{\boldsymbol{s}_i}, \boldsymbol{w}_{\boldsymbol{s}_i}$, and $\varepsilon_{\boldsymbol{s}_i}$ by $y_i, x_i, \boldsymbol{w}_i$, and ε_i, respectively, at the location $\boldsymbol{s}_i$, $i = 1, \ldots, n$. Each simulation case is conducted with 1000 repetitions. Furthermore, we use the empirical critical value $\hat{c}_{1-\alpha}^B$ defined in Section 3.1 with $B = 5000$ at the nominal significance level $\alpha = 0.05$.

4.2 | False positive

To evaluate the false positive rate of the threshold effect test, we generate data from the model (9) when $\delta_1 = \delta_2 = 0$ with the normal errors and the heavy-tailed errors, respectively. And then, we perform the hypothesis testing for the threshold effect in the quantile regression at quantile levels $\tau \in \{0.5, 0.7, 0.9\}$ as well as in the mean regression. The grid size for the discretized $\boldsymbol{\Gamma}^2$ is set to be 10×10 or 20×20. That is, $\boldsymbol{\Gamma} = \left\{(\gamma_L, \gamma_U)^\top | \ 0 \leq \gamma_L < \gamma_U \leq 1\right\} \subset \{0.0, 0.1, \ldots, 1.0\}^2$ or $\{0.00, 0.05, \ldots, 1.00\}^2$. The false positive rate is defined as the proportion of the simulations in which the test statistic $SW_n(\tau)$ is greater than the critical value $\hat{c}_{1-\alpha}^B$.

We consider more sample sizes, $n = 100^2$ and 150^2, for the study with heavy-tailed errors. Table 1 summarizes the empirical false positive rates at the significance level $\alpha = 0.05$. The standard error of the estimated false positive rate is about $\sqrt{(0.05)(0.95)/1000} \approx 0.007$ with 1000 simulations. Thus, with normal errors, $\mathcal{N}(0, 1)$, the false positives are within or lie slightly more than one standard error away from the nominal level 0.05 both in the quantile regression and the

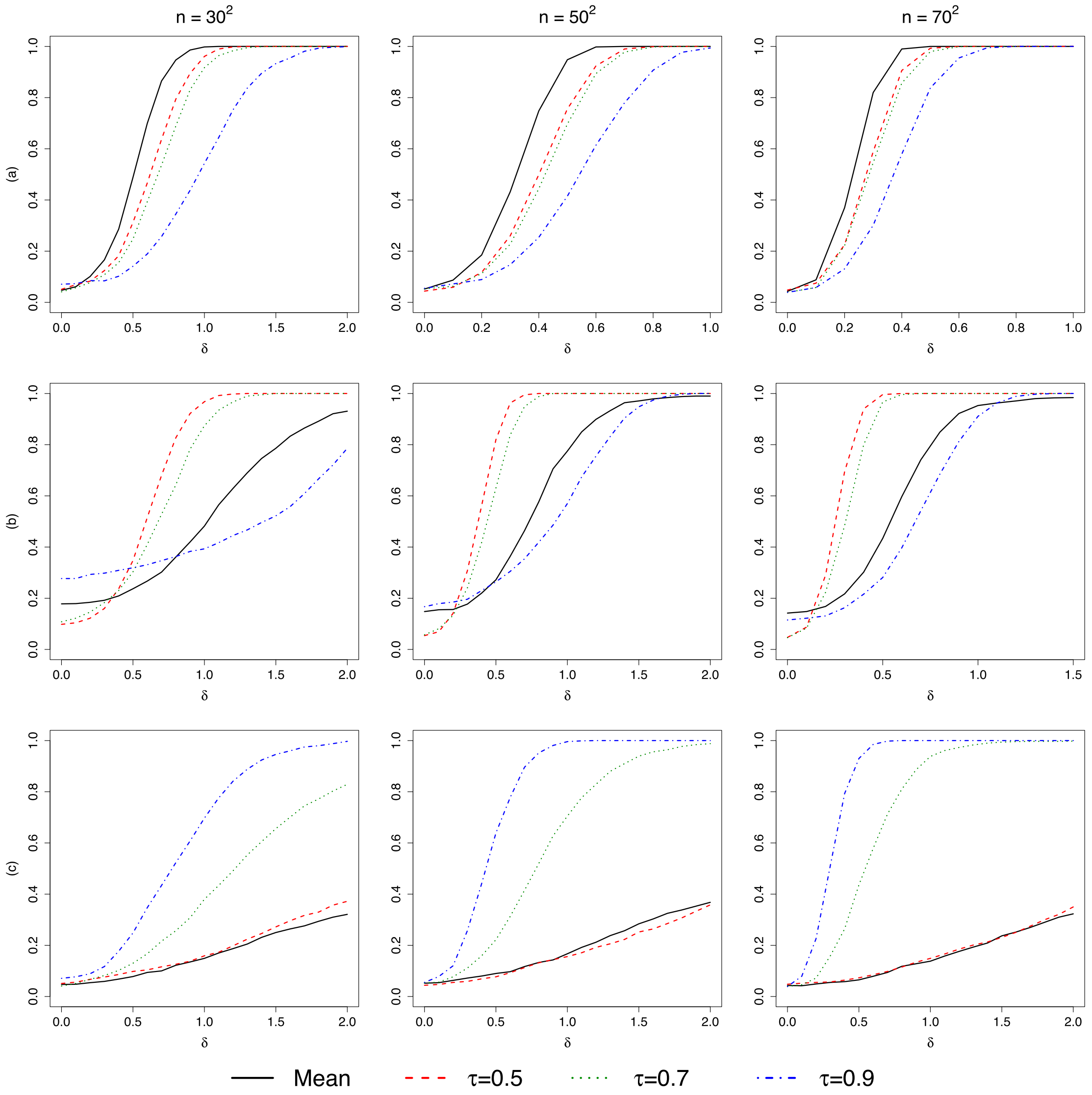


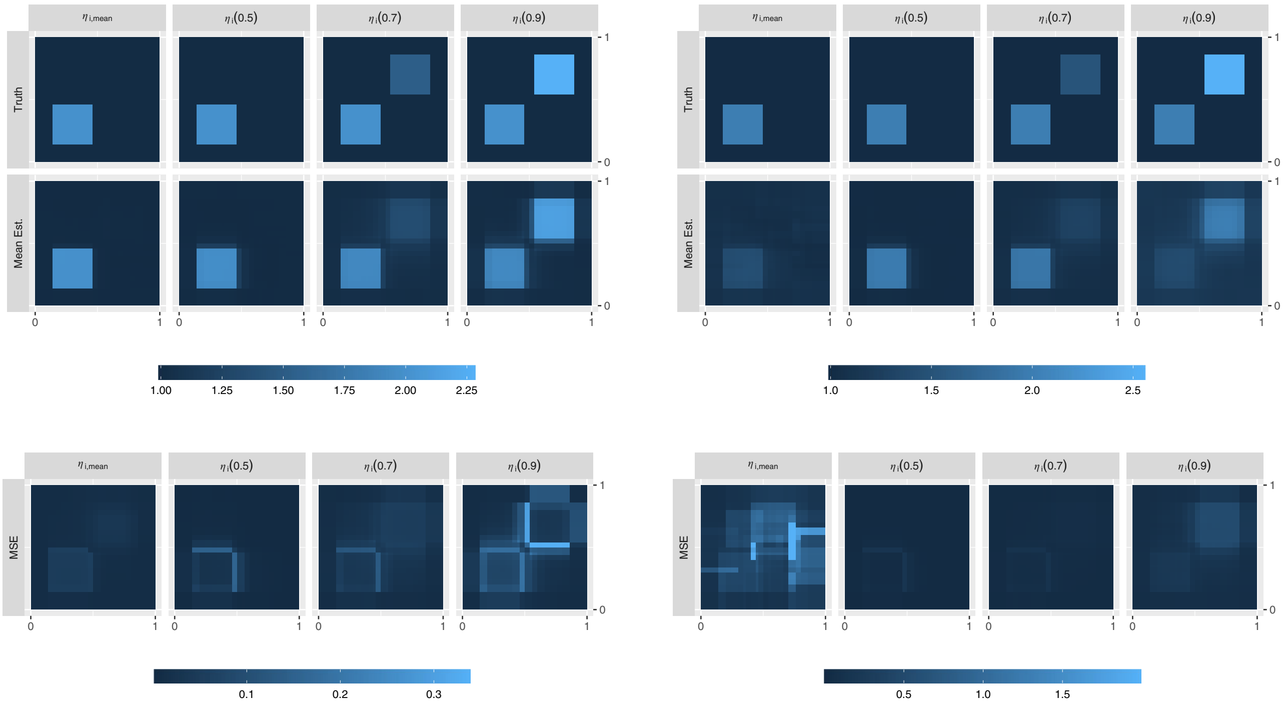
FIGURE 1 Power curves of the mean regression and the quantile regression at $\tau = 0.5, 0.7$, and 0.9. The nominal significance level and the sample size are set to be $\alpha = 0.05$ and $n = 30^2, 50^2$, or 70^2, respectively. (a) Homoscedastic normal errors. (b) Homoscedastic heavy-tailed errors, $\sigma \cdot t(2)$. (c) Heteroscedastic normal errors

mean regression. However, with heavy-tailed errors, $\sigma \cdot t(2)$, the mean regression always produces inflated false positives comparing to the quantile regression at $\tau = 0.5$ and 0.7. The mean regression even shows the higher false positive rates than the higher quantile ($\tau = 0.9$) for relatively large sample sizes ($n = 100^2$ or 150^2). In the meanwhile, the quantile method produces the false positive rates close to the nominal α at $\tau = 0.9$ when the sample sizes are large enough as $n = 100^2$ or 150^2.

4.3 | Power evaluation

To evaluate the power of the threshold effect test, we consider three settings for generating data: homoscedastic normal errors, homoscedastic heavy-tailed errors, and heteroscedastic normal errors. We simulate the data from each setting, and perform the hypothesis testing at quantile levels $\tau \in \{0.5, 0.7, 0.9\}$ as well as at the mean. The grid size for the discretized $\boldsymbol{\Gamma}^2$ is set to be 10×10. That is, $\boldsymbol{\Gamma} = \left\{(\gamma_L, \gamma_U)^\top \mid 0 \le \gamma_L < \gamma_U \le 1\right\} \subset \{0.0, 0.1, \ldots, 1.0\}^2$. We define the power as the proportion of the simulations in which the test statistic $SW_n(\tau)$ is greater than $\hat{c}^B_{1-\alpha}$.

Figure 1 illustrates the power curves for each setting and the sample size. In Figure 1(a,b), where the data are generated with homoscedastic errors, median has the largest power and power at $\tau = 0.9$ is the smallest among the three



(a) Heteroscedastic normal error

(b) Heteroscedastic $t(2)$ error

FIGURE 2 Maps of mean regression slopes $\eta_{i,\text{mean}}$ and quantile regression slopes $\eta_i(\tau)$, where $\tau = 0.5, 0.7$, or 0.9. (a) Normal error. (b) Heavy-tailed error. The first row shows the truth, the second shows the estimated coefficients, and the third row presents the mean squared errors of the estimators

quantile levels. Considered enough sample size ($n \geq 50^2$), both the quantile regression and mean regression approaches provide S-shaped power curves. However, Figure 1(b) shows that mean regression cannot produce enough power with heavy-tailed errors comparing to the results at $\tau = 0.5$ and 0.7.

In Figure 1(c), where the data are generated with the heteroscedastic normal errors, power is largest at $\tau = 0.9$ and smallest at median. This is what we expect since the threshold effect is getting stronger as τ is getting away from the median.

4.4 | Cluster identification

In this study, we evaluate how well our method identifies the true thresholds or clusters which are predefined in the simulation. We generate data from Case 3 ($\delta_1 = \delta_2 = \delta \neq 0$ and $\boldsymbol{\gamma}_1^* \times \boldsymbol{\gamma}_2^* \neq \boldsymbol{\gamma}_3^* \times \boldsymbol{\gamma}_4^*$). The size of the threshold effect is set to be $\delta = 1$ based on the power analysis results, and two clusters are predefined to be $\boldsymbol{\gamma}_1^* \times \boldsymbol{\gamma}_2^* = [0.15, 0.45]^2$ and $\boldsymbol{\gamma}_3^* \times \boldsymbol{\gamma}_4^* = [0.55, 0.85]^2$, respectively. We also consider both heteroscedastic normal and heavy-tailed errors.

For $i = 1, \dots, n$, Equations (10)–(11) can be re-expressed as:

$$Q_{y_i}(\tau|\boldsymbol{w}_i) = \zeta_i(\tau) + \eta_i(\tau) \cdot x_i + F_\varepsilon^{-1}(\tau),$$
$$E(y_i|\boldsymbol{w}_i) = \zeta_{i,\text{mean}} + \eta_{i,\text{mean}} \cdot x_i,$$

where $\zeta_i(\tau) = \zeta_{i,\text{mean}} = 0$, $\eta_i(\tau) = 1 + \delta_1 \cdot \mathcal{I}(\boldsymbol{s}_i \in \boldsymbol{\gamma}_1^* \times \boldsymbol{\gamma}_2^*) + \delta_2 \cdot \mathcal{I}(\boldsymbol{s}_i \in \boldsymbol{\gamma}_3^* \times \boldsymbol{\gamma}_4^*) \cdot F_\varepsilon^{-1}(\tau)$, and $\eta_{i,\text{mean}} = 1 + \delta_1 \cdot \mathcal{I}(\boldsymbol{s}_i \in \boldsymbol{\gamma}_1^* \times \boldsymbol{\gamma}_2^*)$. Thus, at $\tau = 0.5$ and mean, the truth is one cluster. For each simulated dataset, we estimate threshold parameters $(\hat{\boldsymbol{\gamma}}_1^*, \hat{\boldsymbol{\gamma}}_2^*, \hat{\boldsymbol{\gamma}}_3^*$, and $\hat{\boldsymbol{\gamma}}_4^*)$ at quantile levels $\tau \in \{0.5, 0.7, 0.9\}$ as well as at the mean, and estimate the corresponding regression coefficients at each location: $\hat{\zeta}_i(\tau)$s, $\hat{\zeta}_{i,\text{mean}}$, $\hat{\eta}_i(\tau)$s, and $\hat{\eta}_{i,\text{mean}}$ for $i = 1, \dots, n$. And then, we map these regression coefficient estimates.

Figure 2 illustrates the maps of slope estimates $\hat{\eta}_{i,\text{mean}}$ and $\hat{\eta}_i(\tau)$s, where $n = 50^2$ and the grid size for the discretized $\boldsymbol{\Gamma}^2$ is set to be 20×20. That is, $\boldsymbol{\Gamma} = \left\{(\gamma_L, \gamma_U)^\top | \ 0 \leq \gamma_L < \gamma_U \leq 1\right\} \subset \{0.00, 0.05, \dots, 1.00\}^2$. The first two rows are the maps of the true value of the slopes and of the mean of slope estimates across 1000 simulations, respectively. The last row is the maps of the corresponding mean-squared error (MSE). As shown in the figure, true clusters are on the lower left ($\gamma_1^* \times \gamma_2^* = [0.15, 0.45]^2$) and on the upper right ($\gamma_3^* \times \gamma_4^* = [0.55, 0.85]^2$), respectively. While the lower left cluster affects uniformly over the quantile level with the same effect size δ_1, the upper-right cluster has steeper slopes at higher quantiles with the quantile-specific effects as $\delta_2 \cdot F_\varepsilon^{-1}(\tau)$. The results for the other sample sizes, $n = 30^2$ or 70^2, are omitted because findings are similar to those shown in Figure 2.

In Figure 2(a) with heteroscedastic normal errors, we see that only the lower left cluster is identified at $\tau = 0.5$ and mean, while the quantile method could correctly identify both clusters well at $\tau = 0.7$ and 0.9. That is, the quantile method can help identify spatial clusters at tails of the distribution, which are caused by heteroscedasticity, while the mean method may overlook those. In Figure 2(b) with heteroscedastic heavy-tailed errors, we see similar results as in Figure 2(a). However, mean regression cannot identify the lower left cluster clearly with the higher MSE comparing to the median. Thus, there are less chances to detect the cluster at the mean.

5 | DATA APPLICATION

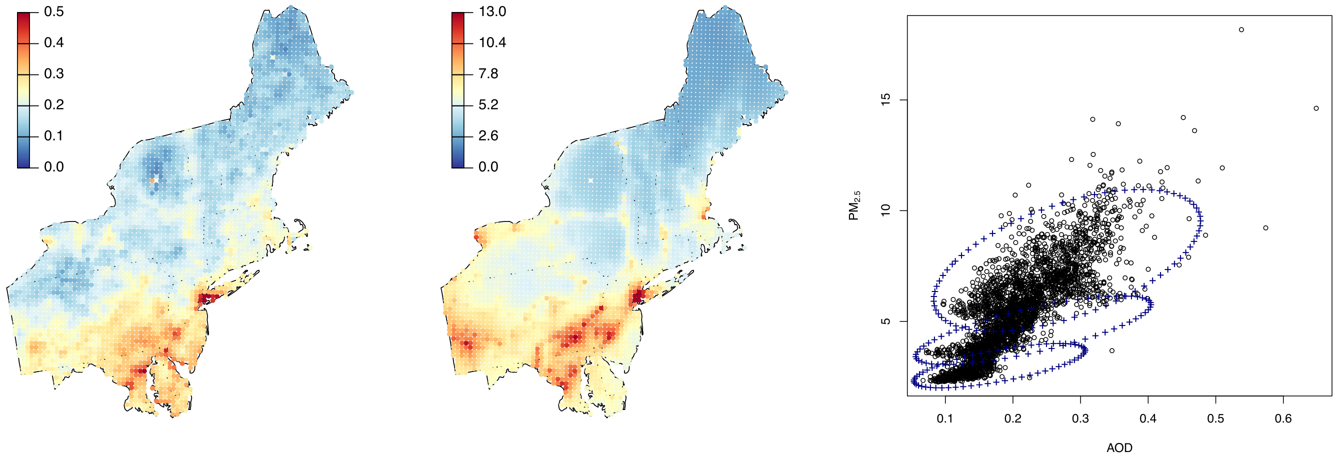
We apply the proposed quantile regression approach to study the impact of AOD on $\mathrm{PM}_{2.5}$ for the summer (June–August) 2012. AOD is a proxy measurement of particle air pollution data since it measures light extinction due to particles in the atmospheric column. We consider the regression model with $\mathrm{PM}_{2.5}$ as the response variable and AOD as the covariate since AOD was shown in previous studies to have positive impacts on $\mathrm{PM}_{2.5}$ (Chu et al., 2016; Grantham et al., 2018; Ma et al., 2016; Yu et al., 2017). The study domain covers the Northeastern United States (Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and District of Columbia), which is defined by the National Climatic Data Center (Karl & Koss, 1984).

We obtain the daily $\mathrm{PM}_{2.5}$ values and satellite-measured AOD data from the Environmental Protection Agency (EPA, https://www.epa.gov/cmaq) and the Moderate Resolution Imaging Spectroradiometer (MODIS, https://modis.gsfc.nasa.gov/data), respectively. Since a $12 \times 12\ \mathrm{km}^2$ grid is the common resolution available when we consider the data from these different sources (EPA and MODIS), we organize the data on a $12 \times 12\ \mathrm{km}^2$ grid up to match the EPA grid cell with a total of 3186 observations. Furthermore, we prepare the data for summer 2012 by averaging each daily variables ($\mathrm{PM}_{2.5}$ and AOD) over June–August, 2012.

Figure 3(a,b) shows the maps of $\mathrm{PM}_{2.5}$ and AOD data averaged over the summer season in 2012. The scatterplot is provided in Figure 3(c), and it shows the positive association between AOD and $\mathrm{PM}_{2.5}$. However, it looks there are at least three chunks of observations, which we circle by navy plus (+) signs on the scatterplot. They look like having different features to one another with respect to the intercepts and slopes. Furthermore, heteroscedasticity is also shown with the funnel-shaped variation within each group. Thus, we suspect the stronger contribution of AOD at the upper tail of the $\mathrm{PM}_{2.5}$ distribution. Our goal is to identify spatial clusters geographically where the spatial observations show similar associations between the response and the covariate. Thus, if those observations are close to each other geographically within some subregions, it is important to find such subregions and model a clustered varying coefficient regression with them. Furthermore, if the association between two variables does not follow a mean rate of change, it is also crucial to take the heteroscedasticity into account. Thus, these $\mathrm{PM}_{2.5}$ and AOD data let us have the scientific motivation to develop statistical models with spatially and quantile level-wise varying AOD effects when estimating $\mathrm{PM}_{2.5}$ by identifying spatial clusters and considering a flexible regression model to capture the heteroscedasticity.

Thus, we apply the proposed quantile regression approach to the $\mathrm{PM}_{2.5}$ and AOD data. The mean regression approach (Lee, Gangnon, & Zhu, 2017) is applied to the same data as well for comparisons. In the simulation studies, we compare two methods when the sample size is at least $n = 50^2$ with the searching grid size 20×20. We use the ratio between these two sizes as a reference to define the searching grid in the real data analysis. Thus, we consider $30 \times 30\ \mathrm{km}^2$ grid resolution ($12\ \mathrm{km} \times \frac{50}{20}$) for the searching grid in the $\mathrm{PM}_{2.5}$ and AOD dataset. The covariate (AOD) is centered to have a zero mean in the application.

Figure 4 illustrates maps of regression coefficient estimates at the mean and quantile levels $\tau = 0.5, 0.7$, and 0.9. The scatterplots between $\mathrm{PM}_{2.5}$ and AOD, with the fitted regression lines for each cluster, are provided below the slope estimate maps. Further, the colors in each scatterplot and the corresponding slope estimate map match each other. In each map and scatterplot, observations are colored based on the coefficient estimates: red colors for high values and blue colors for low values. Mean and median detect four clusters, while three clusters are found at $\tau = 0.7$ and 0.9. At $\tau = 0.5$, the central cluster, beige color in the intercept map, is not found in the slope map. It means that this cluster has the effect in



(a) AOD (b) $PM_{2.5}$ (c) $PM_{2.5}$ v.s. AOD

FIGURE 3 Average of daily aerosol optical depth (AOD) and $PM_{2.5}$ concentration during the summer 2012

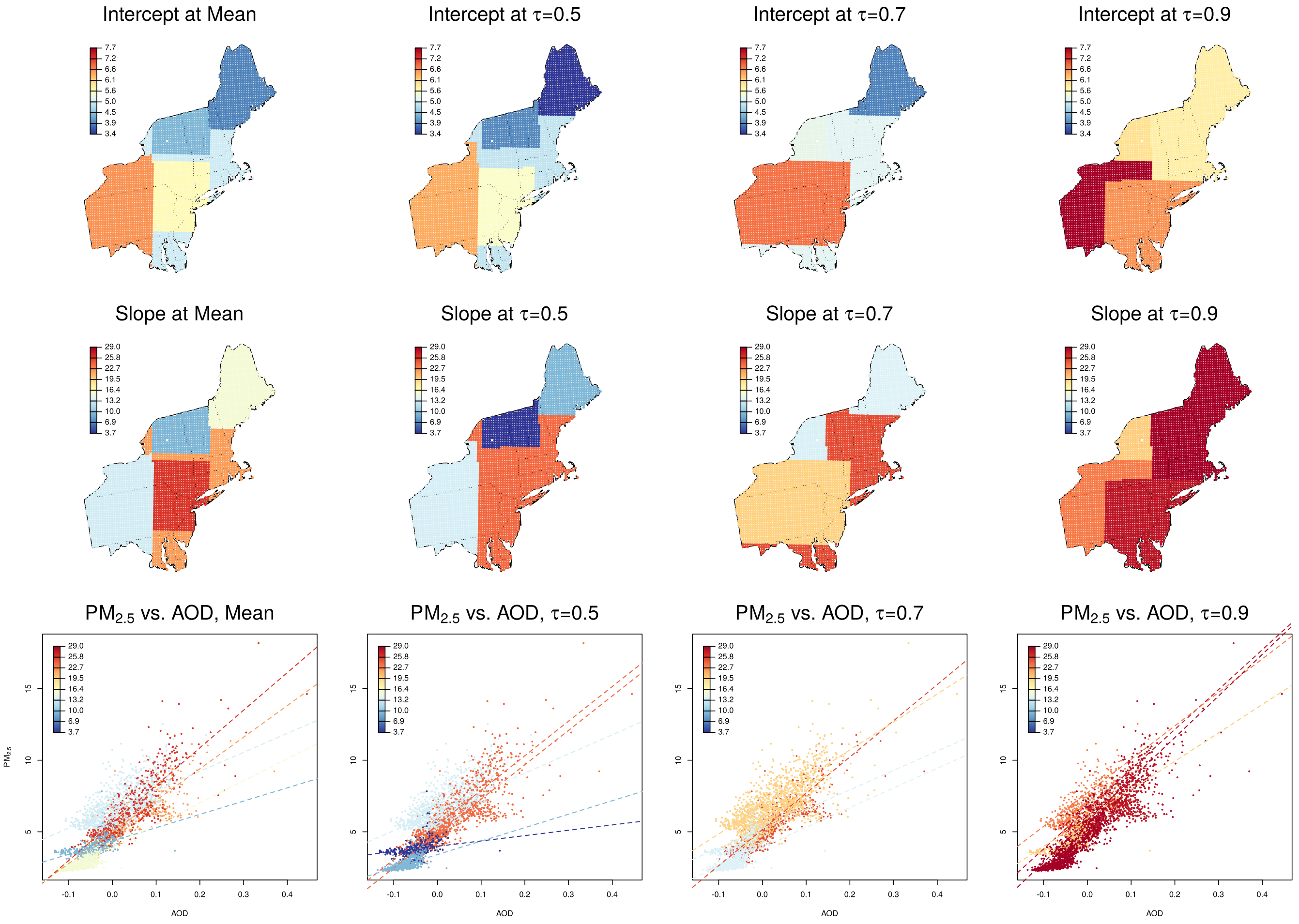


FIGURE 4 Row 1: Maps of the intercept estimates. Row 2: Maps of the slope estimates. Row 3: Scatterplots between $PM_{2.5}$ and aerosol optical depth (AOD) with the fitted regression lines. The colors in each scatterplot match the colors in the corresponding slope estimates map. Observations are colored based on the coefficient estimates: Red colors for high values and blue colors for low values

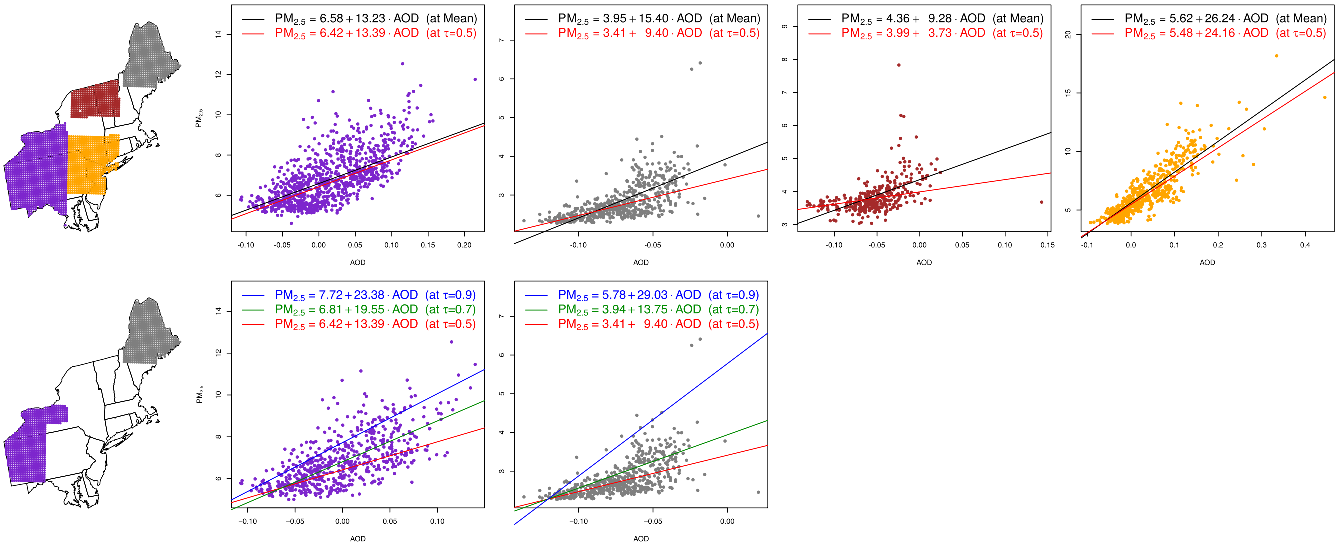


FIGURE 5 Row 1: Common subregions covered by both mean and median. Row 2: Common subregions covered by all quantile levels $\tau = 0.5, 0.7$, and 0.9. Each subregion is indicated in different colors: Purple for the west, gray for the north-east, brown for the upper-central, and orange for the central. Scatterplots are provided next to the maps in the same color of each subregion with fitted regression lines. Fitted line and equation are in black at the mean, in red at $\tau = 0.5$, in green at $\tau = 0.7$, and in blue at $\tau = 0.9$, respectively

the intercept only. On the other hand, at $\tau = 0.7$, the upper-central cluster has the effect in the slope only. The identified clusters are not identical across quantile levels, but they agree with what we observed in Figure 3. We see that there are several areas commonly covered by spatial clusters at the mean and different quantiles. Thus, for further comparisons, we investigate these common areas by looking at the corresponding distributions and fitted regression lines. We call these common areas as subregions to distinguish from the clusters which are separately detected at each quantile.

Figure 5 shows the maps of the subregions commonly covered by the clusters at the mean and median in Row 1, and at different quantiles in Row 2, respectively. That is, Figure 5 illustrates the locations $\{\boldsymbol{s} \mid \boldsymbol{s} \in \cap_\tau \cup_{j=i}^{J_\tau} (\hat{\boldsymbol{\gamma}}_1^{*j} \times \hat{\boldsymbol{\gamma}}_2^{*j})\}$, where J_τ is the number of detected spatial clusters at τ and $\hat{\boldsymbol{\gamma}}_1^{*j} \times \hat{\boldsymbol{\gamma}}_2^{*j}$ is the jth spatial cluster estimator. Each subregion is indicated in different colors: purple for the west, gray for the north-east, brown for the upper-central, and orange for the central. Scatterplots are provided next to the maps in the same color of each subregion with fitted regression lines. A total of four subregions are shared by the clusters at the mean and median. In the purple and orange subregions, the fitted equations are very close each other between the mean and median. However, the mean provides the steeper slope estimates in the gray and brown subregions. That is, $\text{PM}_{2.5}$ is skewed to the right given AOD value in the north-east area and the upper-central area. In the meanwhile, clusters of the quantile regression method at $\tau \in \{0.5, 0.7, 0.9\}$ share two subregions in the west and the north-east. If the homoscedastic assumption holds, observations should show the pipeline-shaped variation, and fitted regression models should provide the uniform slope for all quantiles. However, we see that observations within each scatterplot have the funnel-shaped variation, and that fitted regression models have the steeper slopes for the higher quantiles. That is, the heteroscedasticity exists in the west area and the north-east area. In these two subregions, there is the stronger contribution of AOD at the upper tail of the $\text{PM}_{2.5}$ distribution, as indicated in Figure 3.

Our analysis shows geographical heterogeneity in the AOD–$\text{PM}_{2.5}$ relationship for each given quantile level. The geographical heterogeneity for different quantiles can further facilitate developing statistical models with spatially and quantile level-wise varying AOD effects when estimating $\text{PM}_{2.5}$. Identified clusters at median are qualitatively the same to those from the mean regression. However, median shows robust regression estimates when the distribution of $\text{PM}_{2.5}$ is skewed for given AOD value.

6 | CONCLUSION AND DISCUSSION

In this article, we have proposed a new methodology to identify spatial clusters of regression coefficients on the quantile of the response. The novelty of this article is that we have addressed both issues of the heteroscedasticity and geographical heterogeneity in spatial regression models. Our proposed model addresses spatial heterogeneity through spatial cluster detection and accommodates heteroscedasticity via the quantile regression approaches. In the presence of heteroscedasticity, the quantile regression coefficients $\boldsymbol{\beta}(\tau)$ and the spatial cluster will vary across τ as shown in the upper-right cluster of Figure 2 and Row 2 of Figure 5. Both simulation studies and data application demonstrate that the proposed quantile approach provides better performance than the mean approach (Lee, Gangnon, & Zhu, 2017), especially for distributions with heavy-tails or heteroscedasticity. By conducting analysis at multiple quantiles, the quantile regression framework provides a natural and automatic way to capture the heteroscedasticity. Thus, our proposed method could be an answer in practice with spatial data that do not show the homogeneous features across the study area and that the tail of the distribution is of more interest than mean.

The formal testing for heteroscedasticity is a separate topic from spatial cluster detection. To test for heteroscedasticity, one sufficient way is to test whether $\boldsymbol{\beta}(\tau)$ is constant over τ or not. To construct this testing procedure, the key assumption in the spatial cluster detection (coefficients are constant within the cluster for any given τ) should hold. If not (e.g., continuous coefficients in $\boldsymbol{s}$), it will be more complicated due to the confounding effect of spatial homogeneity, heterogeneity, and heteroscedasticity. Thus, we do not consider a separate formal testing procedure for checking the heteroscedasticity in this article and leave the formal study of heteroscedasticity testing for future research.

Spatial cluster detection approaches, including our proposed method, aim to identify specific clusters with features different from the background. Our model's key advantage over the varying coefficient model with smooth $\{\boldsymbol{\beta}_{\boldsymbol{s}} |\ \boldsymbol{s} \in [0,1]^2\}$ is the explicit identification of specific, contiguous, and compact geographic regions (clusters) associated with different sets of regression coefficients. The associated regression equation applies to a well-identified subset of space in our model. Further, the key assumption of spatial cluster detection approaches is that the number of clusters components is relatively small compared with the number of observations. However, when this assumption is strongly violated, a spatially varying coefficient model, which allows for continuous variation in the regression coefficients (e.g., Gaussian processes, generalized additive models, geographically weighted regression), would be more appropriate. Neither approach is universally optimal or parsimonious. However, in the case of finding spatial hot spots, which aims to identify distinct features comparing to the background, our proposed method seems more suitable than a varying coefficient model with smooth $\{\boldsymbol{\beta}_{\boldsymbol{s}} |\ \boldsymbol{s} \in [0,1]^2\}$.

The proposed method is for a given quantile level and assumes rectangular spatial clusters. And, here are several ideas that can extend this method. The method can be further extended to handle multiple quantile levels simultaneously (e.g., Galvao et al., 2014; Su & Xu, 2019). Further, although we consider the rectangular window for the spatial cluster because we have the regular grid data, it can be modified with other shapes, such as circles, ellipses, squares, and even arbitrary shapes (Assunção et al., 2006; Kulldorff, 1997; Kulldorff et al., 2006; Lee, Gangnon, & Zhu, 2017; Lee et al., 2020, 2021; Tango & Takahashi, 2005). However, in practice, especially with the irregular grid data (e.g., county-level data), considering simple windows (circular or rectangular) is common in cluster detection or scan statistic approaches. Recently, Lee et al. (2021) showed that true clusters in arbitrary shapes are identified effectively, albeit not parsimoniously, by using circular windows. Thus, we believe that the rectangular window will act as well in practice with the irregular grid data.

We can also develop statistical inference for threshold parameters or spatial clusters. Statistical inference for threshold parameters is a challenging problem since the limiting distribution of the estimator is nonstandard due to the nonsmoothness of the indicator function in thresholding. Seo and Linton (2007) proposed a smoothed estimator for mean threshold regression by smoothing the indicator function to achieve asymptotic normality. Lee, Gangnon, Zhu, and Liang (2017) developed statistical inference on the estimated cluster instead of on threshold parameters by defining a confidence set for the true cluster in the one-dimensional space. It is possible to adopt these ideas mentioned above in our setup, and we leave this topic for future research.

ACKNOWLEDGMENTS

This publication is based upon work supported by King Abdullah University of Science and Technology (KAUST), Office of Sponsored Research (OSR) under Award No. OSR-2019-CRG7-3800; by the IR/D program and grant DMS-1712760 from the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors thank Dr. Howard H. Chang for providing the satellite data.

REFERENCES

Angrist, J., Chernozhukov, V., & Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica*, *74*(2), 539–563.

Assunção, R., Costa, M., Tavares, A., & Ferreira, S. (2006). Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine*, *25*(5), 723–742.

Cai, Y. (2010). Forecasting for quantile self-exciting threshold autoregressive time series models. *Biometrika*, *97*(1), 199–208.

Cai, Y., & Stander, J. (2008). Quantile self-exciting threshold autoregressive time series models. *Journal of Time Series Analysis*, *29*(1), 186–202.

Caner, M. (2002). A note on least absolute deviation estimation of a threshold model. *Econometric Theory*, *18*(3), 800–814.

Chang, H. H., Reich, B. J., & Miranda, M. L. (2012). Time-to-event analysis of fine particle air pollution and preterm birth: Results from North Carolina, 2001–2005. *American Journal of Epidemiology*, *175*(2), 91–98.

Chu, Y., Liu, Y., Li, X., Liu, Z., Lu, H., Lu, Y., Mao, Z., Chen, X., Li, N., Ren, M., Liu, F., Tian, L., Zhu, Z., & Xiang, H. (2016). A review on predicting ground $\mathrm{PM}_{2.5}$ concentration using satellite aerosol optical depth. *Atmosphere*, *7*(10), 129.

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., & Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA*, *295*(10), 1127–1134.

Galvao, A. F., Kato, K., Montes-Rojas, G., & Olmo, J. (2014). Testing linearity against threshold effects: Uniform inference in quantile regression. *Annals of the Institute of Statistical Mathematics*, *66*(2), 413–439.

Galvao, A. F., Montes-Rojas, G., & Olmo, J. (2011). Threshold quantile autoregressive models. *Journal of Time Series Analysis*, *32*(3), 253–267.

Gangnon, R. E. (2010). A model for space-time cluster detection using spatial clusters with flexible temporal risk patterns. *Statistics in Medicine*, *29*(22), 2325–2337.

Gangnon, R. E. (2012). Local multiplicity adjustment for the spatial scan statistic using the Gumbel distribution. *Biometrics*, *68*(1), 174–182.

Gangnon, R. E., & Clayton, M. K. (2000). Bayesian detection and modeling of spatial disease clustering. *Biometrics*, *56*(3), 922–935.

Grange, S. K., Lewis, A. C., & Carslaw, D. C. (2016). Source apportionment advances using polar plots of bivariate correlation and regression statistics. *Atmospheric Environment*, *145*, 128–134.

Grantham, N. S., Reich, B. J., Liu, Y., & Chang, H. H. (2018). Spatial regression with an informatively missing covariate: Application to mapping fine particulate matter. *Environmetrics*, *29*(4), e2499.

Hallin, M., Lu, Z., & Yu, K. (2009). Local linear spatial quantile regression. *Bernoulli*, *15*(3), 659–686.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, *64*(2), 413–430.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica*, *68*(3), 575–603.

He, X., & Zhu, L. X. (2003). A lack-of-fit test for quantile regression. *Journal of the American Statistical Association*, *98*(464), 1013–1022.

Hendricks, W., & Koenker, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, *87*(417), 58–68.

Horowitz, J. L., & Spokoiny, V. G. (2002). An adaptive, rate-optimal test of linearity for median regression models. *Journal of the American Statistical Association*, *97*(459), 822–835.

Karl, T., & Koss, W. J. (1984). *Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983*. National Climatic Data Center.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50.

Kuan, C. M., Michalopoulos, C., & Xiao, Z. (2017). Quantile regression on quantile ranges – A threshold approach. *Journal of Time Series Analysis*, *38*(1), 99–119.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics, Part A*, *26*, 1481–1496.

Kulldorff, M., Huang, L., Pickle, L., & Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine*, *25*(22), 3929–3943.

Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine*, *14*(8), 799–810.

Lawson, A. B., Choi, J., & Zhang, J. (2014). Prior choice in discrete latent modeling of spatially referenced cancer survival. *Statistical Methods in Medical Research*, *23*(2), 183–200.

Lee, J., Gangnon, R. E., & Zhu, J. (2017). Cluster detection of spatial regression coefficients. *Statistics in Medicine*, *36*(7), 1118–1133.

Lee, J., Gangnon, R. E., Zhu, J., & Liang, J. (2017). Uncertainty of a detected spatial cluster in 1D: Quantification and visualization. *Stat*, *6*(1), 345–359.

Lee, J., Kamenetsky, M. E., Gangnon, R. E., & Zhu, J. (2021). Clustered spatio-temporal varying coefficient regression model. *Statistics in Medicine*, *40*(2), 465–480.

Lee, J., Sun, Y., & Chang, H. H. (2020). Spatial cluster detection of regression coefficients in a mixed-effects model. *Environmetrics*, *31*(2), e2578.

Lee, S., Seo, M. H., & Shin, Y. (2011). Testing for threshold effects in regression models. *Journal of the American Statistical Association*, *106*(493), 220–231.

Ma, Z., Liu, Y., Zhao, Q., Liu, M., Zhou, Y., & Bi, J. (2016). Satellite-derived high resolution PM2.5concentrations in Yangtze River Delta Region of China using improved linear mixed effects model. *Atmospheric Environment*, *133*(2016), 156–164.

McMillen, D. P. (2013). *Quantile regression for spatial data*. Springer.

Otsu, T. (2008). Conditional empirical likelihood estimation and inference for quantile regression models. *Journal of Econometrics*, *142*(1), 508–538.

Pope, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air & Waste Management Association*, *56*(6), 709–742.

Portnoy, S., & Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statistical Science*, *12*, 279–300.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Samoli, E., Peng, R., Ramsay, T., Pipikou, M., Touloumi, G., Dominici, F., Burnett, R., Cohen, A., Krewski, D., Samet, J., & Katsouyanni, K. (2008). Acute effects of ambient particulate matter on mortality in Europe and North America: Results from the APHENA study. *Environmental Health Perspectives*, *116*(11), 1480–1486.

Seo, M. H., & Linton, O. (2007). A smoothed least squares estimator for threshold regression models. *Journal of Econometrics*, *141*, 704–735.

Su, L., & Xu, P. (2019). Common threshold in quantile regressions with an application to pricing for reputation. *Econometric Reviews*, *38*(4), 417–450.

Tang, Y., Song, X., & Zhu, Z. (2015). Threshold effect test in censored quantile regression. *Statistics & Probability Letters*, *105*, 149–156.

Tango, T., & Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, *4*, 11.

Yoshida, T. (2021). Extreme value inference for quantile regression with varying coefficients. *Communications in Statistics - Theory and Methods*, *50*(3), 685–710.

Yu, K., Lu, Z., & Stander, J. (2003). Quantile regression: Applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*(3), 331–350.

Yu, W., Liu, Y., Ma, Z., & Bi, J. (2017). Improving satellite-based PM2.5 estimates in China using Gaussian processes modeling in a Bayesian hierarchical setting. *Sci Rep*, *7*(1), 7048.

Zhang, L., Wang, H. J., & Zhu, Z. (2014). Testing for change points due to a covariate threshold in quantile regression. *Statistica Sinica*, *24*(4), 1859–1877.

Zheng, J. X. (1998). A consistent nonparametric test of parametric regression models under conditional quantile restrictions. *Econometric Theory*, *14*(1), 123–138.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Lee, J., Sun, Y., & Judy Wang, H. (2021). Spatial cluster detection with threshold quantile regression. *Environmetrics*, e2696. https://doi.org/10.1002/env.2696