Spatiotemporal quantile regression for detecting distributional changes in environmental processes

Brian Reich, NC State

JSM 2012

San Diego, CA
Climate change versus global warming

- Global warming refers to an increase in mean temperature.

- Climate change is more general, and refers to changes in the distribution of climate variables.

- This includes changes in the mean, increased variability, and severity of extreme events.

- We propose using spatiotemporal QR as a flexible and interpretable method for simultaneously detecting changes in several features of the distribution of climate variables.
Let $Y_t$ be the value in year $t$.

Here the $\tau^{th}$ quantile is modeled as $\beta_0(\tau) + \beta_1(\tau)t$. 
Quantile regression

- Mean regression: $E(Y) = X^T \beta$.

- Quantile regression: $Q(\tau) = X^T \beta(\tau)$, where $Q(\tau)$ is the $\tau^{th}$ quantile of $Y$ so that $P[Y < Q(\tau)] = \tau \in [0, 1]$.

- $\beta(\tau)$ gives the covariate effects on the $\tau^{th}$ quantile; by varying $\tau$ we study different aspects of the response distribution:
  - **Center**: $\tau = 0.5$ gives median regression.
  - **Extreme events**: With $\tau = 0.99$ we study the magnitude of extreme values.
  - **Variability**: $Q(0.75) - Q(0.25) = X^T [\beta(0.75) - \beta(0.25)]$ is the IQR.
Let $Y_t(s)$ be the response for location $s$ in year $t$.

- We model the $\tau^{th}$ quantile of $Y_t(s)$ as

$$Q_t(\tau|s) = \beta_0(\tau|s) + \beta_1(\tau|s)t.$$ 

- $\beta_1(\tau|s)$ is the increase per year at site $s$ in the $\tau^{th}$ quantile.

- All sites and quantiles are modeled simultaneously to borrow strength where appropriate.

- We perform inference on the evolution of different aspects of the distribution.
Spatiotemporal quantile regression

Our model has three essential components:

1. Select a prior for $\beta_0(\tau|s)$ and $\beta_1(\tau|s)$ so that the quantile function is valid (monotonically increasing) at $s$ for all $t$.

2. Smooth $\beta_0(\tau|s)$ and $\beta_1(\tau|s)$ over $s$ to borrow strength across nearby locations to estimate the quantile function.

3. Account for dependence between nearby observations in the same year using a copula.
Basis expansion for $\beta_0$ and $\beta_1$

Let $\beta_j(\tau|s) = \theta_{j0}(s) + \sum_{l=1}^{L} B_l(\tau) \theta_{jl}(s)$.

- $\theta_{j0}(s)$ is a location parameter (the median for this model).
- $B_l(\tau)$ are known basis functions.
- $\theta_{jl}(s)$ determine the shape of the quantile function.
- The $\theta_{jl}(s)$ vary spatially to allow for a different distribution at each site.
- This gives $2L$ parameters at each site.
We use piecewise Gaussian basis functions

- Let $0 = \kappa_1 < \ldots < \kappa_{L+1} = 1$ be a grid of knots.

- For $l$ with $\kappa_l < 0.5$, 
  
  $$
  B_l(\tau) = \begin{cases} 
  \Phi^{-1}(\kappa_l) - \Phi^{-1}(\kappa_{l+1}) & \tau < \kappa_l \\
  \Phi^{-1}(\tau) - \Phi^{-1}(\kappa_{l+1}) & \kappa_l \leq \tau < \kappa_{l+1} \\
  0 & \kappa_{l+1} \leq \tau,
  \end{cases}
  $$

- For $l$ such that $\kappa_l \geq 0.5$, 
  
  $$
  B_l(\tau) = \begin{cases} 
  0 & \tau < \kappa_l \\
  \Phi^{-1}(\tau) - \Phi^{-1}(\kappa_l) & \kappa_l \leq \tau < \kappa_{l+1} \\
  \Phi^{-1}(\kappa_{l+1}) - \Phi^{-1}(\kappa_l) & \kappa_{l+1} \leq \tau.
  \end{cases}
  $$
We use piecewise Gaussian basis functions

- Left: basis functions $B_l(\tau)$ with $L = 6$.
- Middle: quantile function $\beta_0(\tau) = \sum_{l=1}^{L} B_l(\tau) b_l$.
- Right: The corresponding density with $t = 0$.
Properties of these basis functions

- If $\theta_{jl}(s)$ are the same for all basis functions $l = 1, ..., L$, then $Y_t(s) \sim N[\theta_{j0}(s), \theta_{j1}(s)^2].$

- In this sense, the model is centered on the heteroskedastic Gaussian spatial model.

- If we restrict attention to quantile levels $\kappa_2, ..., \kappa_L$ and time points $t = 0, 1$, then this model spans the entire set of valid quantile functions defined on this support.

- Therefore, for large $L$ it should be possible to approximate a wide class of quantile functions.
The likelihood function

- The density corresponding to the quantile function is

\[
\sum_{l=1}^{L} \mathbb{I}[Q(\kappa_l) \leq Y_t(s) < Q(\kappa_{l+1})] \mathcal{N} \left[ Y_t(s) \mid a_l(s, t), b_l(s, t)^2 \right].
\]

- \(a\) and \(b\) are functions of \(\theta\).

- We refer to this as a multiply-split normal density.

- Other densities could be “split” to center the quantile function on a different prior.
Non-crossing quantiles

- In order for the quantile function to be valid, it must be increasing in $\tau$ for all $s$ and $t$.

- For this choice of basis function, this happens if and only if

  1. We restrict $t$ to a finite interval, e.g., $t \in [0, 1]$

  2. $\theta_{0l}(s) > 0$ for all $l = 1, \ldots, L$

  3. $\theta_{0l}(s) + \theta_{1l}(s) > 0$ for all $l = 1, \ldots, L$

- To ensure the constraints are satisfied, let $\theta_{jl}^*$ be unconstrained parameters, and

$$
\theta_{jl} = \theta_{jl}^* [\theta_{0l}(s) > 0 \text{ and } \theta_{0l}(s) + \theta_{1l}(s) > 0] + \epsilon.
$$
The parameters that control the spatial variability of the quantile function, $\theta_{jl}^*$, are given Gaussian process priors.

They are independent over $j$ and $l$.

$\mathbb{E}[\theta_{jl}^*(s)] = \bar{\theta}_{jl}$.

$\text{Cov}[\theta_{jl}^*(s), \theta_{jl}^*(s')] = \sigma_{jl}^2 \exp \left( \frac{||s - s'||}{\rho_{jl}} \right)$. 
Residual correlation

- Allowing the quantile function to vary spatially accounts for some spatial dependence.

- However, there remains residual spatial dependence.

- We account for this with a Gaussian copula.

- Let $\epsilon_t(s)$ be a latent Gaussian process with mean zero, variance one, and correlation over $s$.

- We take $Y_t(s) = Q_t[U_t(s)|s]$ where $U_t(s) = \Phi[\epsilon_t(s)]$.

- The likelihood for year $t$ is

$$f[Y_t(s_1), \ldots, Y_t(s_n)] = \phi_n[\epsilon_t(s_1), \ldots, \epsilon_t(s_n)] \prod_{i=1}^{n} \frac{f[Y_t(s_i)]}{\phi_1[\epsilon_t(s_i)]}$$

where $\phi_n$ is the $n$-dimensional MNV density.
Analysis of the eastern US temp data

- \( Y_t(s) \) is temperature for month \( t \) at location \( s \).

- Data are used from 1931-2009 at 191 sites in SE US.

- We analyze data from June, July, and August, with a separate quantile function by month.

- A simple space-time copula is used for the residual dependence.

- We analyze monthly minimum, maximum and mean temperature separately.

- We first select the number of knots \( (L) \) using several model fit criteria.
Issues with rounding

- The data are rounded to the nearest degree (F).

- To account for rounding we model $Y_t(s) = \mu_t(s) + \epsilon_t(s)$, where $\epsilon_t(s) \overset{iid}{\sim} N(0, \sigma^2)$.

- $\mu_t(s)$ is the true value, and modeled using quantile regression.

- It turns out that in this case if we include the latent $\theta_t(s)$ in the MCMC, they have conjugate full conditionals.
Model fit criteria for the number of basis functions, $L$

$L = 1$ is the usual Gaussian model with spatially varying mean and standard deviation.

<table>
<thead>
<tr>
<th>Data</th>
<th>$L$</th>
<th>DIC</th>
<th>EPD</th>
<th>CRPS</th>
<th>MSE</th>
<th>MAD</th>
<th>COV90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1</td>
<td>52,134</td>
<td>0.270</td>
<td>0.620</td>
<td>0.347</td>
<td>0.432</td>
<td>0.912</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>51,727</td>
<td>0.259</td>
<td>0.608</td>
<td>0.343</td>
<td>0.431</td>
<td>0.916</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>51,956</td>
<td>0.260</td>
<td>0.609</td>
<td>0.347</td>
<td>0.433</td>
<td>0.910</td>
</tr>
<tr>
<td>Min</td>
<td>1</td>
<td>155,779</td>
<td>5.162</td>
<td>2.706</td>
<td>3.883</td>
<td>1.463</td>
<td>0.912</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>153,692</td>
<td>4.742</td>
<td>2.595</td>
<td>3.986</td>
<td>1.474</td>
<td>0.916</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>153,522</td>
<td>4.633</td>
<td>2.566</td>
<td>4.038</td>
<td>1.502</td>
<td>0.907</td>
</tr>
<tr>
<td>Max</td>
<td>1</td>
<td>132,352</td>
<td>2.687</td>
<td>1.955</td>
<td>2.341</td>
<td>1.154</td>
<td>0.907</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>132,717</td>
<td>2.704</td>
<td>1.960</td>
<td>2.479</td>
<td>1.177</td>
<td>0.909</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>132,903</td>
<td>2.725</td>
<td>1.967</td>
<td>2.193</td>
<td>1.132</td>
<td>0.914</td>
</tr>
</tbody>
</table>

Brian Reich  
North Carolina State University
Posterior mean of $\beta_1(\tau|s)$ by month for three sites

Brian Reich  
North Carolina State University
$\beta_1(\tau | s)$ for $\tau = 0.1$ (left) and $\tau = 0.9$

- Shaded symbols represent sites with 95% posterior intervals that exclude zero.
- The largest trend is in the lower quantile in the mountains.
Future work:

- Currently not equipped to handle large $n$.
- More flexible models may be needed if there are several covariates that need to be included.

References:


Thanks!