Hierarchical Additive Modeling of Nonlinear Association with Spatial Correlations

Qingzhao Yu, Bin Li, Richard Scribner

Louisiana State University, School of Public Health

Supported by: NIAAA-5 R01 AA013810-03 (PI Richard Scribner)

LSU Health Sciences Center
NEW ORLEANS

Outline

- Objective
- Data and Challenges
- Multivariate Additive Regression Trees
- Conditional Autoregressive Model
- Hierarchical Additive Model
- Model Convergence
- Results

LSU Health Sciences Center
NEW ORLEANS

Objective

The aim is to determine whether a change in the alcohol environment results in a change of assaultive violence rates in the neighborhood.

Hypothesis: Neighborhood alcohol outlet density is positively associated with alcohol related crimes, such as the assaultive crime.

Research Environment

- In 1992, a civil unrest occurred over a large area of south central Los Angeles. During the civil unrest, a number of alcohol outlets were damaged, so that the alcohol density decreased in that area.
- We restrict our analysis to the areas affected by civil unrest thereby controlling for a possible global effect of the unrest on outcomes.
- We control for the other factors that are associated with both alcohol density and the related crime rates.

LSU Health Sciences Center
NEW ORLEANS

Data

- There are a total of 290 census tracts.
- All data are collected at the census tract level.
- We collected 10 years' data: 1990 to 1999.

Data

- There are a total of 290 census tracts.
- All data are collected at the census tract level.
- We collected 10 years' data: 1990 to 1999.
- Alcohol Environment: data are obtained from the PRC and California ABC.

Data

- There are a total of 290 census tracts.
- All data are collected at the census tract level.
- We collected 10 years' data: 1990 to 1999.
- Alcohol Environment: data are obtained from the PRC and California ABC.
  - outlet license surrender data following the 1992 Civil Unrest: revealing 279 outlets suspended operation following the 1992 unrest, 252 of those outlets were off-sale type outlets.
  - annual count of offsale and onsale alcohol outlets

LSU Health Sciences Center NEW ORLEANS

Data (Cont'd)

- Sociodemographic Data. These data are collected through US Census and LA county department of health services for the years 1990 to 1999. The variables we used include:
  - % Black, % White, % Hispanic, % Asian, % male, % adults and % households in poverty

LSU Health Sciences Center
NEW ORLEANS

Data (Cont'd)

- Sociodemographic Data. These data are collected through US Census and LA county department of health services for the years 1990 to 1999. The variables we used include:
  - % Black, % White, % Hispanic, % Asian, % male, % adults and % households in poverty
- Outcomes. The crime rates are collected from the LAPD annual police reports for the years 1990 to 1999. The data collected include:
  - Assaultive Violence Rate: homicide, rape, robbery and assault.
  - Complicated interactions might exist among covariates.

LSU Health Sciences Center
NEW ORLEANS

Analysis Challenges

- The association between the assault violence rates and other covariate maybe nonlinear.

LSU Health Sciences Center
NEW ORLEANS

Analysis Challenges

- The association between the assault violence rates and other covariate maybe nonlinear.
- Complicated interactions might exist among covariates.

LSU Health Sciences Center
NEW ORLEANS

Analysis Challenges

- The association between the assault violence rates and other covariate maybe nonlinear.
- Complicated interactions might exist among covariates.
- More than 7% of the observations are missing one or more values for some covariates.

LSU Health Sciences Center
NEW ORLEANS

Analysis Challenges

- The association between the assault violence rates and other covariate maybe nonlinear.
- Complicated interactions might exist among covariates.
- More than 7% of the observations are missing one or more values for some covariates.
- we should take into account the spatial correlations among observations in adjacent tracts.

Multivariate Additive Regression Trees

- MART is a special case of the generic gradient boosting approach.

LSU Health Sciences Center
NEW ORLEANS

Multivariate Additive Regression Trees

- MART is a special case of the generic gradient boosting approach.
- MART approximates the target function $F(\mathbf{x})$ by an additive expansion of trees

$$\hat{f}(x) = \sum_{m=1}^{M} \nu b_H(\mathbf{x}; \gamma_m).$$

Multivariate Additive Regression Trees

- MART is a special case of the generic gradient boosting approach.
- MART approximates the target function $F(\mathbf{x})$ by an additive expansion of trees

$$\hat{f}(x) = \sum_{m=1}^{M} \nu b_H(\mathbf{x}; \gamma_m).$$

- Friedman, 2001, Friedman and Meulman, 2003.

LSU Health Sciences Center NEW ORLEANS

Why MART?

- MART is able to capitalize on the nonlinear relationships between the dependent and independent variables with no need for specifying the basis functions.

Why MART?

- MART is able to capitalize on the nonlinear relationships between the dependent and independent variables with no need for specifying the basis functions.
- MART is able to capture complex and/or high order interaction effects.

Why MART?

- MART is able to capitalize on the nonlinear relationships between the dependent and independent variables with no need for specifying the basis functions.
- MART is able to capture complex and/or high order interaction effects.
- MART can handle mixed-type predictors and missing values in covariates.

Conditional Autoregressive Model

- We use the vector $\{\phi_{T_i,C_i}\}$ to capture spatial autocorrelations among areas C_i at time T_i.
- Assume that an area C_i is correlated with only the areas that are adjacent to it.
- y with a hierarchical structure on its mean function:

$$y_i \sim N(\mu_i, \sigma^2) \text{ and } \mu_i = f(\mathbf{x}_i) + \phi_{T_i,C_i}; \quad (1)$$

- The hierarchical CAR structure for $\{\phi_{T_i,C_i}\}$ has the form

$$\phi_{T_i,C_i=j} | \phi_{T_i,C_i \neq j} \sim N\left(\sum_{k\sim j} \frac{1}{n_j}\phi_{T_i,k}, \frac{1}{n_j \tau_{T_i}}\right).$$

LSU Health Sciences Center
NEW ORLEANS

The Two-Stage Iteration Algorithm

1. Let $\phi^0_{T_i,C_i} = 0$ where $C_i \in \{1, \ldots, C\}$, $T_i \in \{1, \ldots, T\}$; $q = 0$, $\delta = 1000$, $\mu_{1i} = 0$ and $i = 1, \ldots, n$.
2. If $\delta < \Delta$, go to step 3, otherwise q=q+1 and
   (a) Let $z_i = y_i - \phi^{[q-1]}_{T_i,C_i}$. Fit MART $f^{[q]}(\mathbf{x})$ on z and the covariates $\mathbf{x}$.
   (b) Let $e_i = y_i - f^{[q]}(\mathbf{x}_i)$, calculate the Moran's I of e_i within time slots 1 to T. Let S be the collection of time slots in which the spatial correlation test show a p-value smaller than 0.01(this indicates a strong spatial correlation in e_i).

LSU Health Sciences Center
NEW ORLEANS

The Two-Stage Iteration Algorithm (Cont'd)

(c) If S is empty, let $\phi^{[q]}_{T_i,C_i} = 0$ and go to step 3; otherwise for the observations $i \in \{i : T_i \in S\}$, let the $f(\mathbf{x}_i)$ in Equation (1) be $f^{[q]}(\mathbf{x}_i)$ and calculate the generalized MLEs of $\hat{\phi}_{T_i,C_i}$. Let $\phi^{[q]}_{T_i \in S,C_i} = \hat{\phi}_{T_i,C_i}$ and $\phi^{[q]}_{T_i \notin S,C_i} = 0$.

(d) Let $\mu_{0i} = \mu_{1i}$, $\mu_{1i} = f^{[q]}(\mathbf{x}_i) + \phi^{[q]}_{T_i,C_i}$ and let $\delta = \frac{\sum_{i=1}^n (\mu_{1i} - \mu_{0i})^2}{\sum_{i=1}^n \mu_{1i}^2}$, go back to 2.

3. Output the results from step q.

Algorithm Convergence

- Little is known of the convergence properties of the above procedure.

Algorithm Convergence

- Little is known of the convergence properties of the above procedure.
- ICM (Iterated Conditional Modes), Besag (1986).

Algorithm Convergence

- Little is known of the convergence properties of the above procedure.
- ICM (Iterated Conditional Modes), Besag (1986).
- Under certain conditions, iterative conditional maximization converges to local maximizers. Meng and Rubin (1993).

Algorithm Convergence

- Little is known of the convergence properties of the above procedure.
- ICM (Iterated Conditional Modes), Besag (1986).
- Under certain conditions, iterative conditional maximization converges to local maximizers. Meng and Rubin (1993).
- Backfitting process, Hastie and Tibshirani, 2000; Buja et al. 1989.

Algorithm Convergence

- Little is known of the convergence properties of the above procedure.
- ICM (Iterated Conditional Modes), Besag (1986).
- Under certain conditions, iterative conditional maximization converges to local maximizers. Meng and Rubin (1993).
- Backfitting process, Hastie and Tibshirani, 2000; Buja et al. 1989.
- Let $\Delta = 10^{-7}$ in our analysis.

LSU Health Sciences Center
NEW ORLEANS

Relative Variable Importance

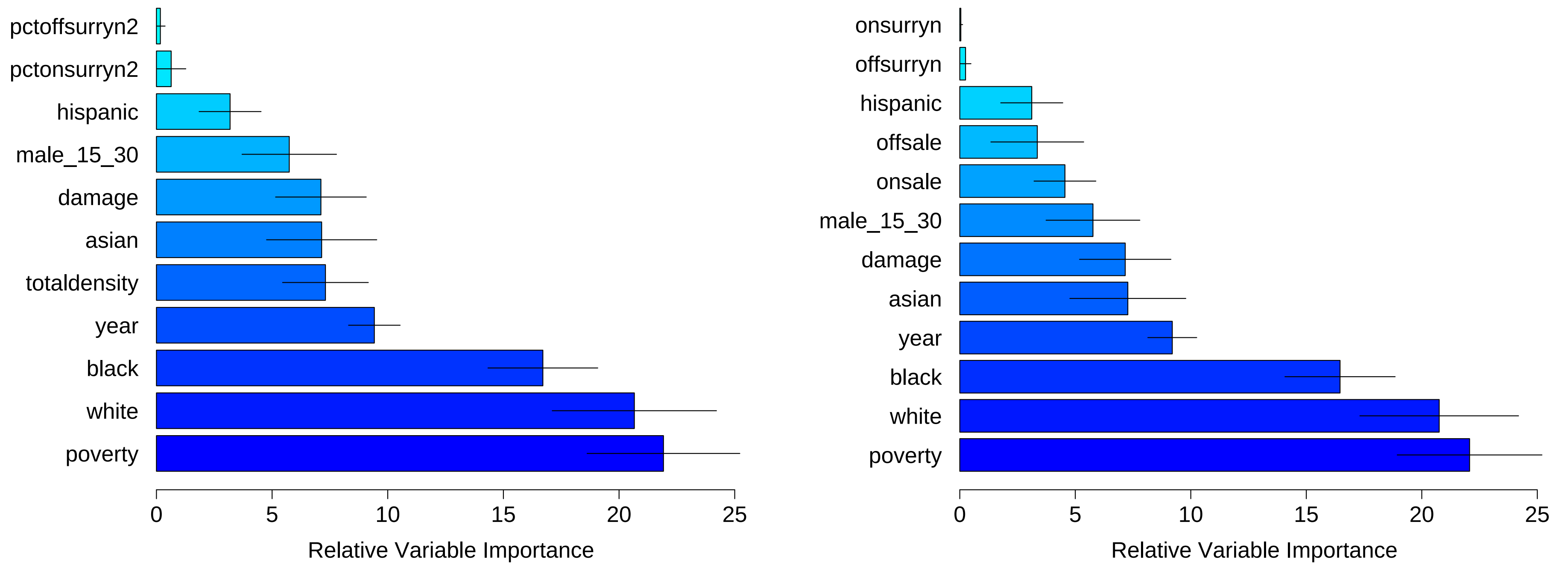
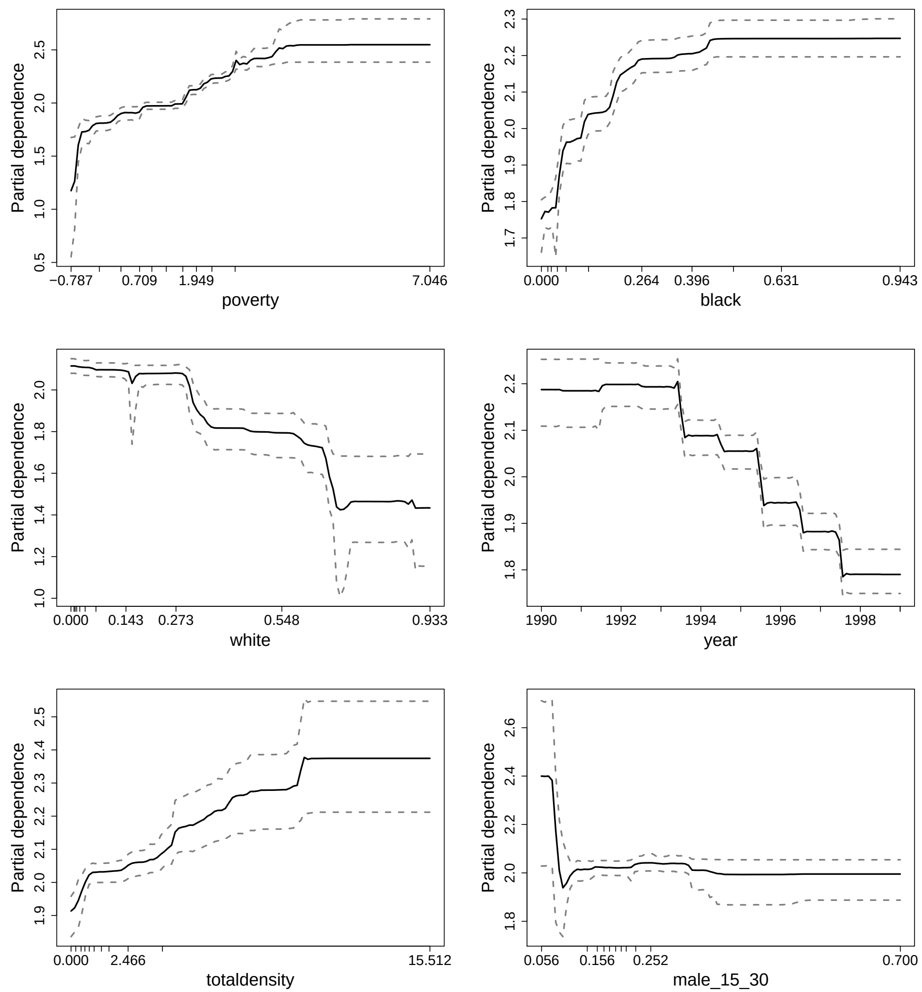


Figure 1: *Relative variable importance in MART-fitted models.*

LSU Health Sciences Center
NEW ORLEANS

Partial Dependence



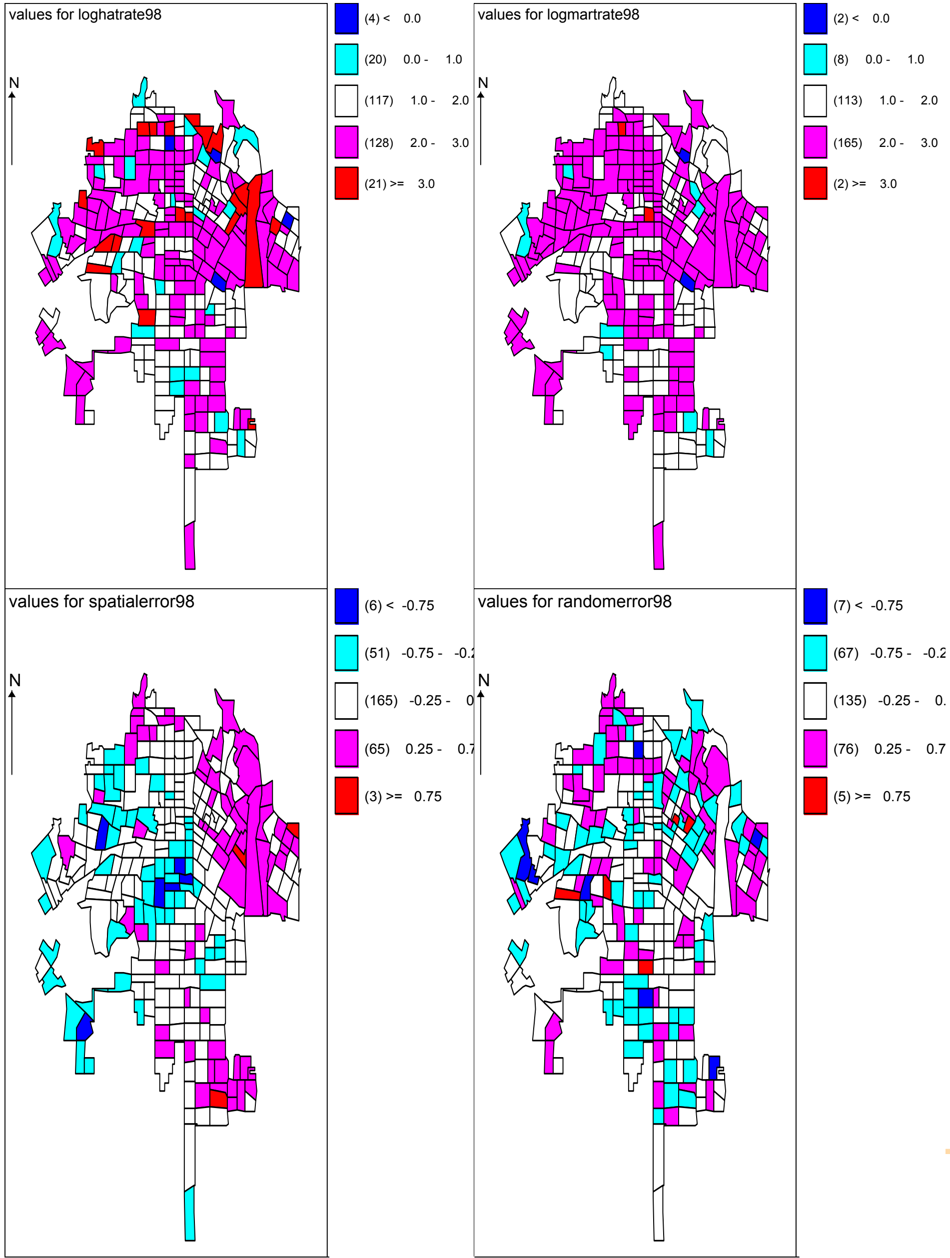
LSU Health Sciences Center
NEW ORLEANS

Figure 2: *Partial dependence plots.*

Spatial Correlation

Year	Origin	P-value	Res1	P-value	Res2	P-value
1999	0.46	< 0.0001	0.13	1.23e-04	-0.09	0.017
1998	0.52	< 0.0001	0.29	0.00e-00	-0.04	0.277
1997	0.50	< 0.0001	0.25	1.46e-12	-0.02	0.688
1996	0.29	< 0.0001	0.05	4.78e-02	0.05	0.048
1995	0.49	< 0.0001	0.23	2.35e-11	-0.01	0.484
1994	0.49	< 0.0001	0.24	4.98e-12	-0.04	0.335
1993	0.52	< 0.0001	0.21	4.83e-09	-0.01	0.825
1992	0.30	< 0.0001	0.14	1.77e-07	-0.16	0.000
1991	0.18	< 0.0001	0.02	4.43e-01	0.02	0.443
1990	0.44	< 0.0001	0.16	1.26e-06	-0.11	0.001

Maps of Log Assault Rates



LSU Health Sciences Center
NEW ORLEANS

Questions