Heterogeneous Distributed Lag Models to Estimate Personalized Effects of Maternal Exposures to Air Pollution

Ander Wilson, Colorado State University

Slides and papers available at `anderwilson.github.io`

Thanks to the Team

Harvard University

- **Daniel S. Mork** (formerly CSU)
- Marc Weisskopf
- Brent A. Coull

Columbia University

- Marianthi-Anna Kioumourtzoglou

NIEHS Grants: ES029943, ES028811

USEPA grant: RD-839278

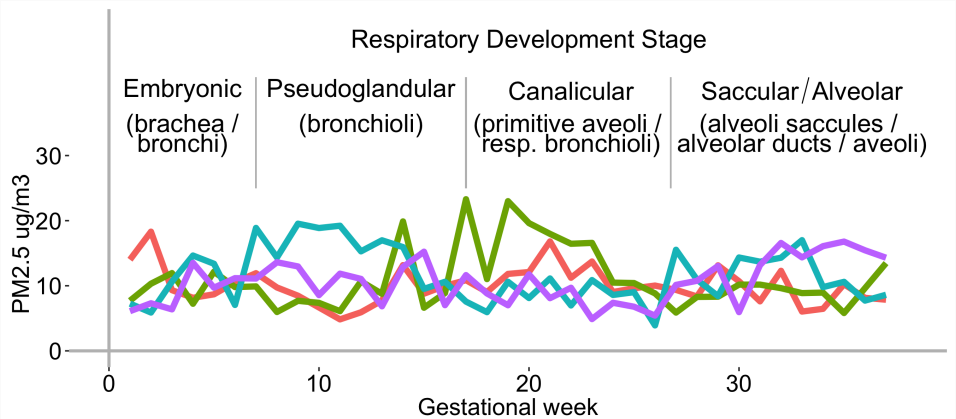
Contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

Air Pollution is Bad

Critical Windows of Susceptibility

Definition

A period in time during which an exposure can alter phenotype.

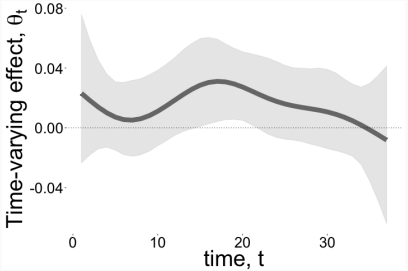


Distributed Lag Model (DLM)

$$y_i = \sum_{t=1}^{T} x_{it}\theta_t + \boldsymbol{z}_i'\boldsymbol{\gamma} + \varepsilon_i$$

- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)'$ *constrained* to vary smoothly in time (e.g. spline, Gaussian process, ...)
  - adds stability to the model
  - conforms with biological hypothesis that exposure at proximal time points are likely to have similar effects

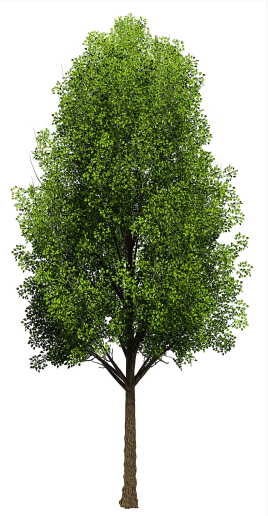
DLM analysis of $PM_{2.5}$ and asthma among boys in the ACCESS cohort.



[1]Figure source: Wilson et al. (2017) *Biostatistics*.

Heterogeneity and Modification with Critical Windows

- Increased focus on vulnerable populations and precision environmental health
- Standard approach is to conduct a stratified analysis
- Bayesian distributed lag interaction models allow for modification by a single categorical factor (Wilson et al, 2017, *Biostatistics*)
- Lack of methods for multiple modifiers
- Heterogeneity by multiple modifiers poses dimensionality and multiple comparison problems
- Our work: How to estimate heterogeneous DLMs using Bayesian additive regression trees (BART)



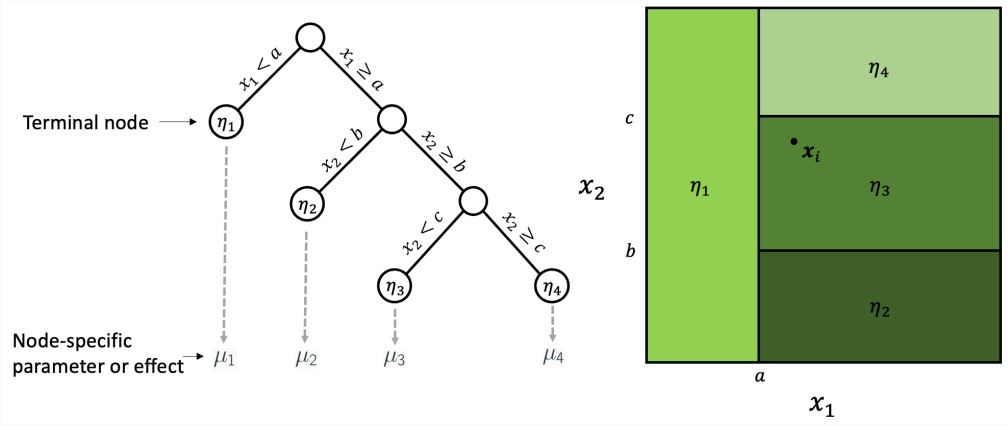
Bayesian Additive Regression Trees (BART)

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i$$

- Proposed by by Chipman, George, McCulloch (1998, *JASA* & 2010, *AOAS*)
- Estimate a general mean function
- State of the art predictive performance
- Allows for coherent Bayesian inference

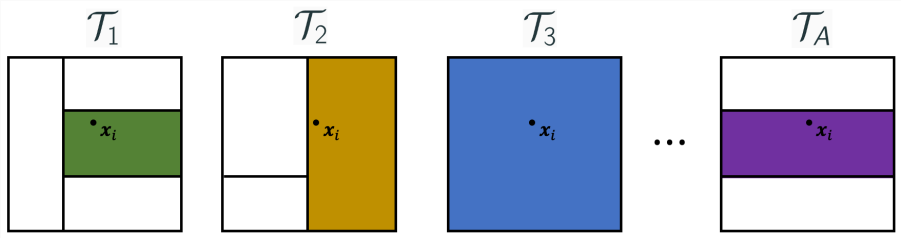
Bayesian Additive Regression Trees (BART)

$$g(\boldsymbol{x}_i, \mathcal{T}) = \mu_b \qquad \text{if } \boldsymbol{x}_i \in \eta_b$$

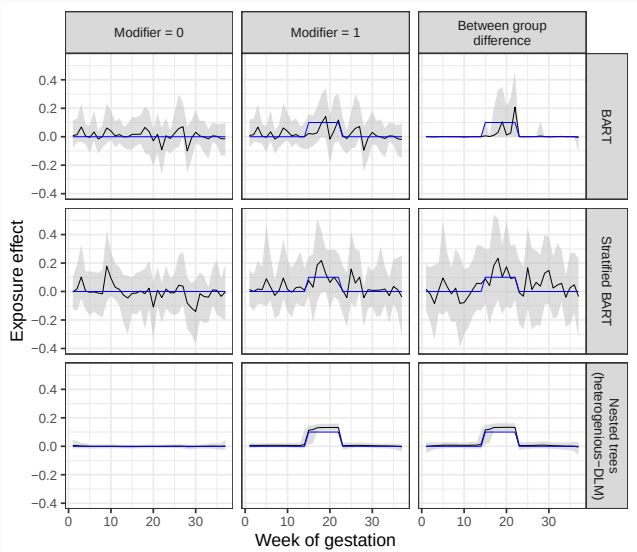


BART

$$f(\boldsymbol{x}_i) = \sum_{a=1}^{A} g(\boldsymbol{x}_i, \mathcal{T}_a)$$



BART Alone Does Not Work

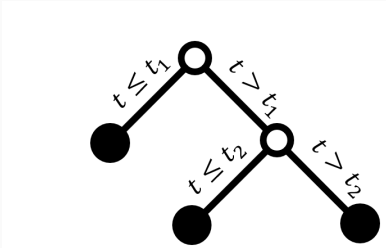


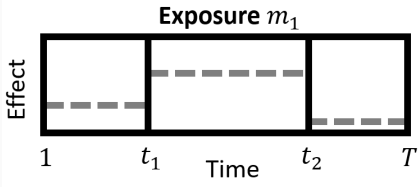
- Does not account for autocorrelation between repeated measures of exposure
- Applies same amount and type of shrinkage and selection to exposures and candidate modifiers

Treed Distributed Lag Model (TDLM) To Add Structure

$$y_i = \sum_{t=1}^{T} x_{it}\theta_t + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_i$$

- Apply BART to time ($t = 1, \ldots, T$) to define structure in the lag function $\theta_1, \ldots, \theta_T$
- Constant effect of exposure in each terminal node or time segment

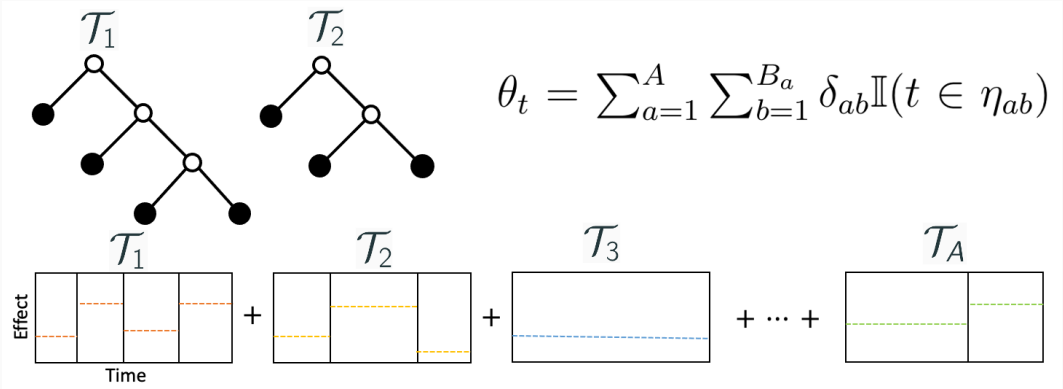




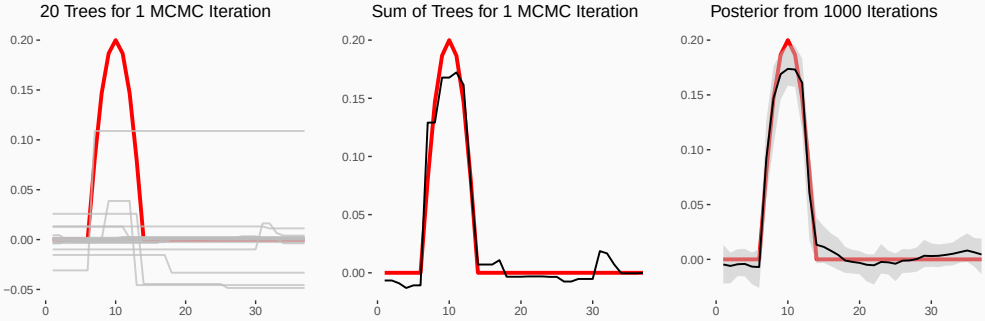
[1] Source: Mork & Wilson (2023) *Biometrics*

TDLM: Ensemble of Trees

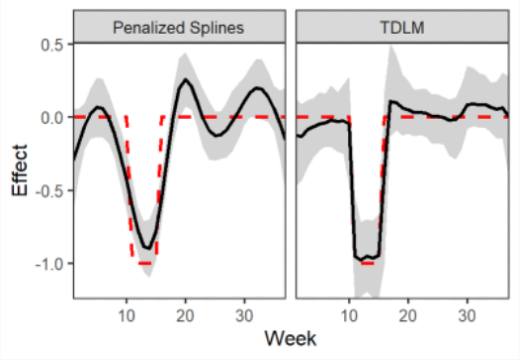
- Use ensemble of A trees
- Adds robustness and can approximate smooth distributed lag functions
- η_{ab} and δ_{ab} is the terminal node and effect for node b on tree a

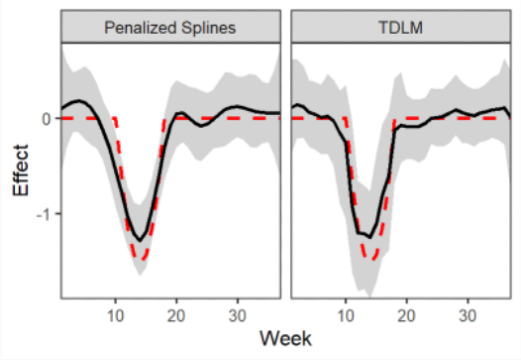


TDLM: Illustrative Example



TDLM: Illustrative Example





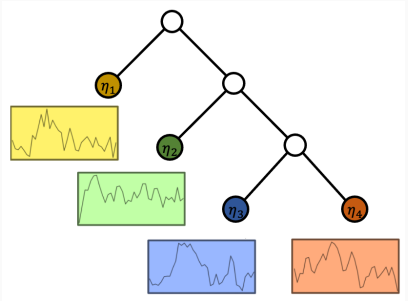
The advantages of trees and TDLM

- More flexible
- Less tuning
- Lower false discovery rate
- More robust in time series studies when adjusting for long-term trends (Leung 2022 et al. *Am. J. Epi.*)
- Extends to mixture exposures
- Extends to heterogeneity with multiple modifiers

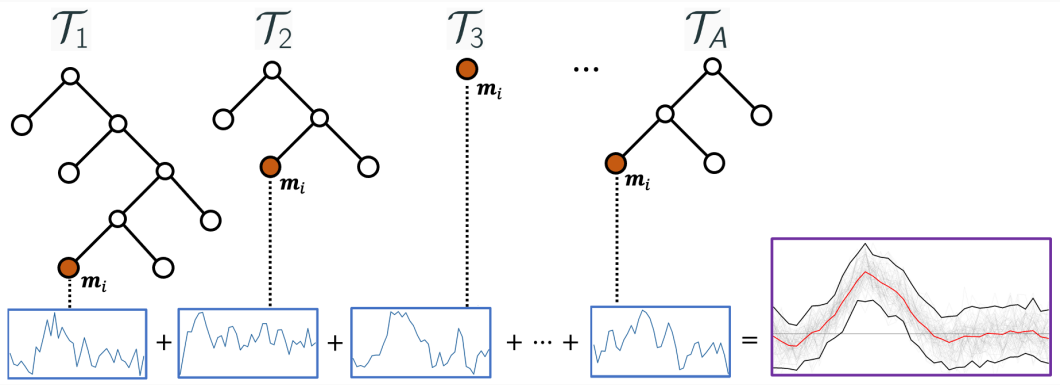
Heterogeneity DLM (HDLM)

$$y_i = \sum_{t=1}^{T} x_{it}\theta_t(\boldsymbol{m}_i) + \boldsymbol{z}_i'\boldsymbol{\gamma} + \varepsilon_i$$

- DLM for a single pollutant with personalized effects based on a vector of modifying factors $\boldsymbol{m}$
- Key idea: use BART to partition modifier space and have a unique distributed lag function for each terminal node
- Allows for multiple modifiers that are continuous, categorical and/or ordinal



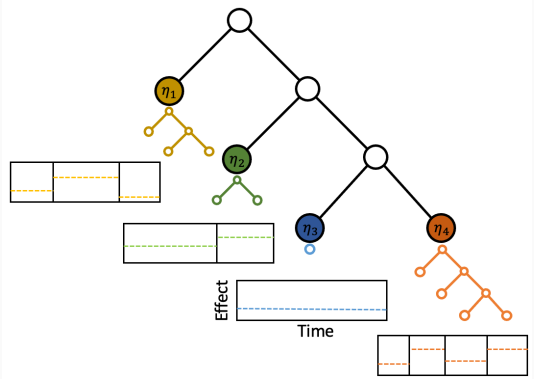
Heterogeneity DLM (HDLM)



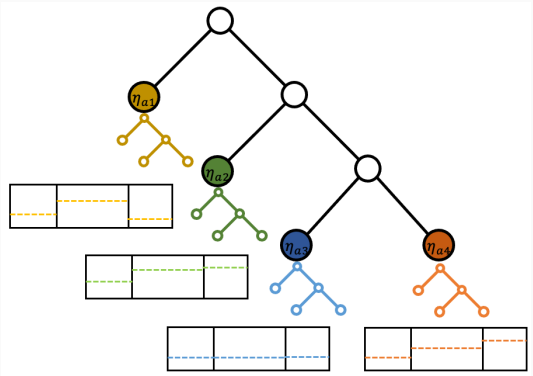
Nested and Shared Tree HDLM

- We can fit the distributed lag function with splines, Gaussian processes, or more trees

Nested Tree HDLM



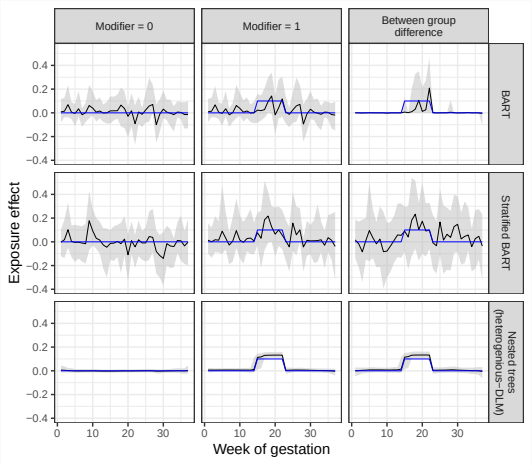
Shared Tree HDLM



Nested and Shared Tree HDLM

- Dirichlet prior to modifier inclusion (Linero 2018, *JASA*)
- Horseshoe-type priors on terminal node parameters
- Estimated with MCMC following original BART algorithm with a few key changes including additional grow step complexity for the modifier trees in the nested tree model

Nested and Shared Tree HDLM

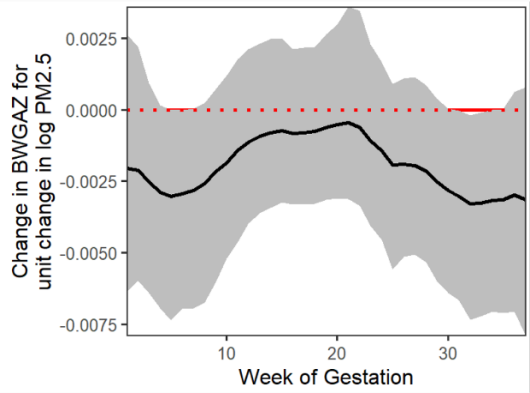


- HDLMs have nominal coverage and low false window detection rates
- Includes true modifiers with high probability
- Includes null modifiers with lower probability (0.6-0.7)
- Comparable to DLM when there is no heterogeneity

Birth Weight Analysis

- 310,236 full term (37 weeks) births from Colorado Front Range with estimated conception dates between 2007 – 2015
- Outcome: birth weight z-score (BWGAZ), adjusted for sex, gestational age
- $PM_{2.5}$ exposure measured weekly during gestation
- Controlled for: mother's age, height, weight, body mass index, income, education, marital status, prenatal care, smoking habits, race, Hispanic, child's sex, year/month of conception, elevation, county, trimester average temperature

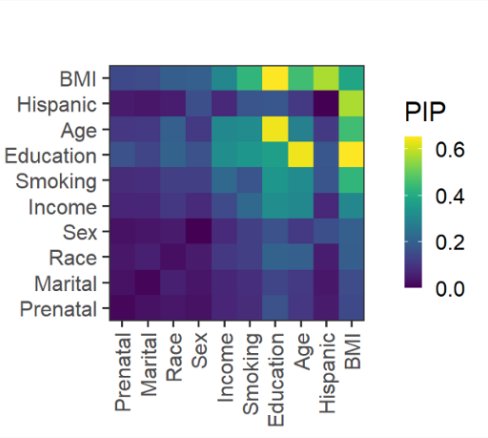
Analysis with DLM (no heterogeneity)



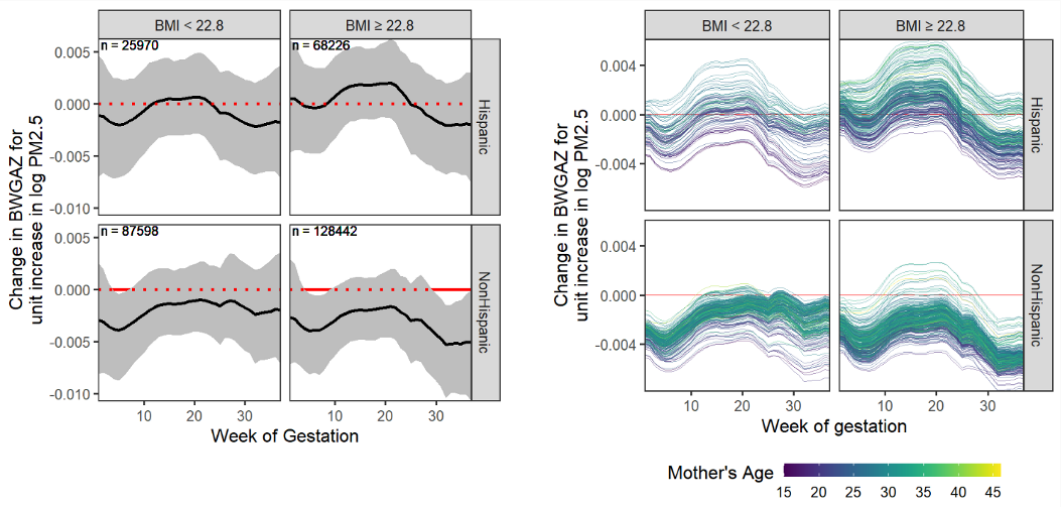
Modifier Selection

Covariate	Type	Modifier	PIP
Age at Conception	Continuous	✓	0.93
Height	Continuous		
Prior Weight	Continuous		
Body Mass Index	Continuous	✓	0.95
Income	Ordinal	✓	0.74
Education	Ordinal	✓	0.90
Marital Status	Categorical	✓	0.50
Prenatal Care	Categorical	✓	0.48
Smoking Habits	Ordinal	✓	0.78
Race	Categorical	✓	0.61
Hispanic	Binary	✓	0.95
Sex of Child	Binary	✓	0.64
County of Residence	Categorical		
Month of Conception	Categorical		
Year of Conception	Categorical		
Avg. Temp per Trimester	Continuous		

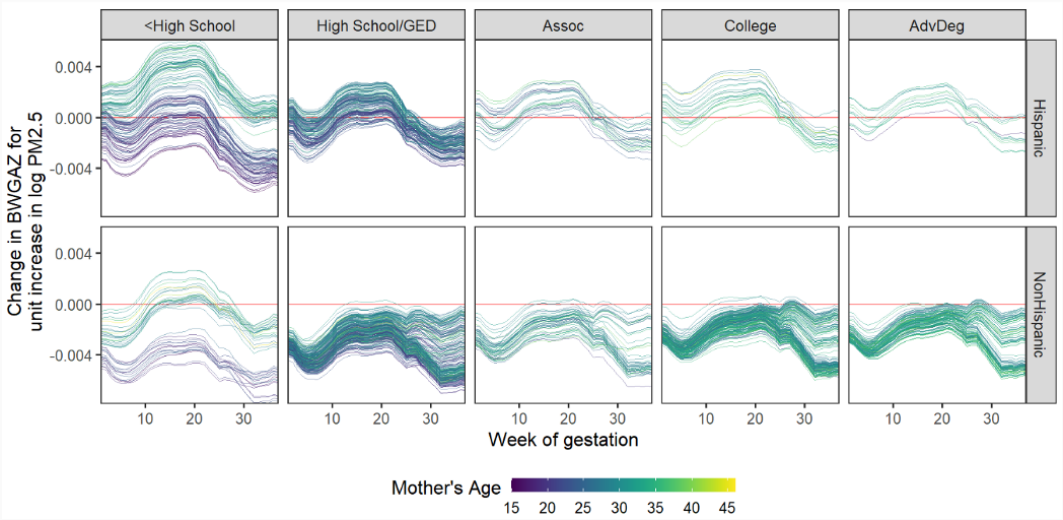
PIP = Posterior Inclusion Probability



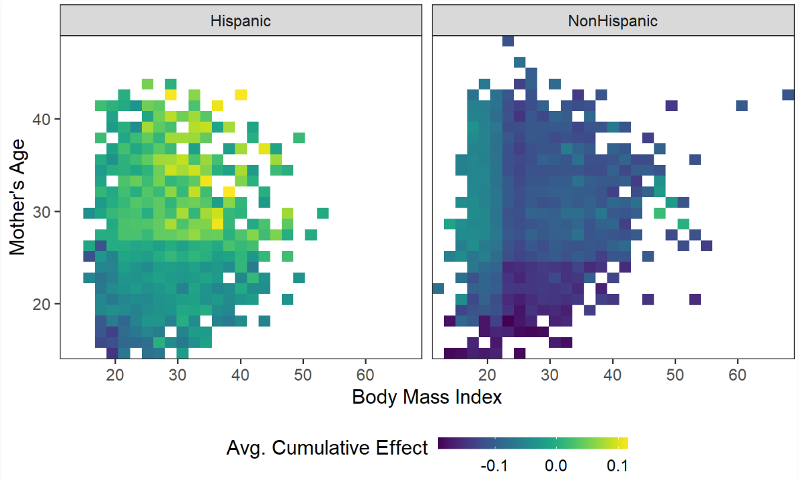
Modification by Maternal BMI and Hispanic Status



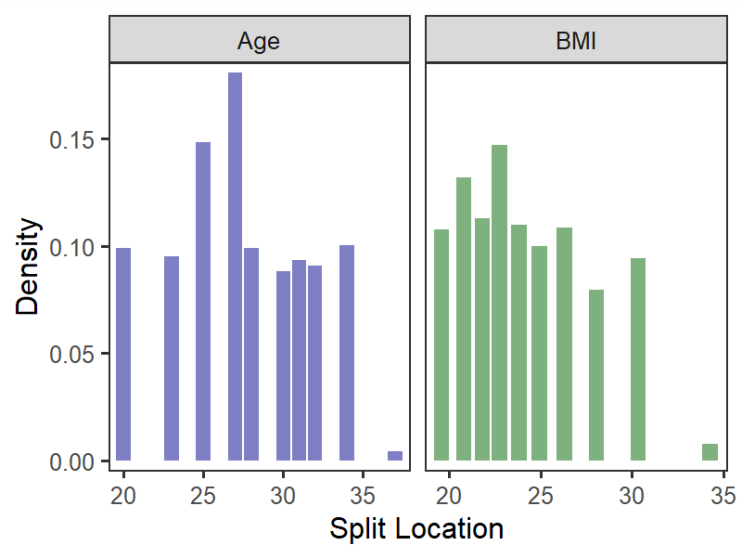
Modification by Maternal Education and Hispanic Status



Cumulative Effect by M. Age, M. BMI and Hispanic Status



Posterior Analysis of Split Points



Summary

- We can add structure to BART to get interpretable estimates of DLMs
- Allows for identifying critical windows
- Allows for heterogeneity
- Overall good finite sample properties
- Available for linear and logistic regression
- R code available: `github.com/danielmork/dlmtree`

Thank You

anderwilson.github.io
ander.wilson@colostate.edu
@ander_wilson

Mork, D., Wilson, A. (2023). Estimating perinatal critical windows of susceptibility to environmental mixtures via structured Bayesian regression tree pairs. *Biometrics*. https://arxiv.org/abs/2102.09071

Mork, D., Kioumourtzoglou, M.-A., Weisskopf, M., Coull, B. A., Wilson, A. (In press). Heterogeneous Distributed Lag Models to Estimate Personalized Effects of Maternal Exposures to Air Pollution. *Journal of the American Statistical Association*. http://arxiv.org/abs/2109.13763.