Identifying Perinatal Critical Windows with Mixtures and Heterogeneity via Structured Regression Trees

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

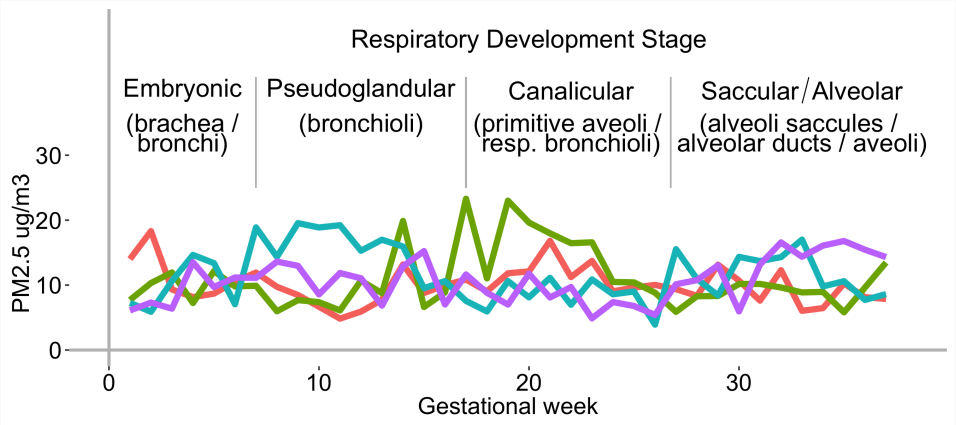
USEPA grant: RD-839278

Air Pollution is Bad

Critical Windows of Susceptibility

Definition

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

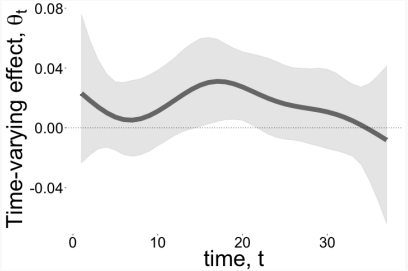


Distributed Lag Model (DLM)

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

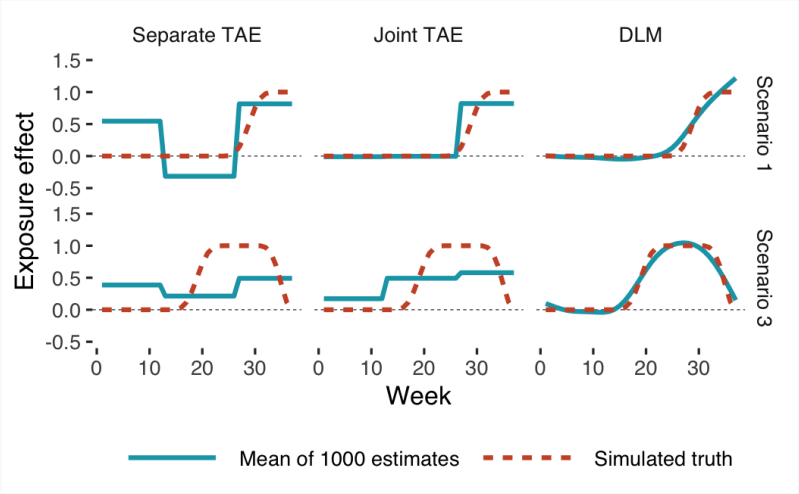
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)'$ *constrained* to vary smoothly in time (e.g. spline, Gaussian process, ...)
  - adds stability to the model
  - conforms with biological hypothesis that exposure at proximal time points are likely to have similar effects

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



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

The Advantage of DLMs



[1]Source: Wilson et al. (2017) *Am. J. Epi.*

Limitations of DLM

- Tendency to over-smooth the distributed lag function
- Lack of DLM methods for mixtures
- Lack of DLM methods for modification or effect heterogeneity
- This talk: How to use Bayesian additive regression trees (BART) to solve all these problems

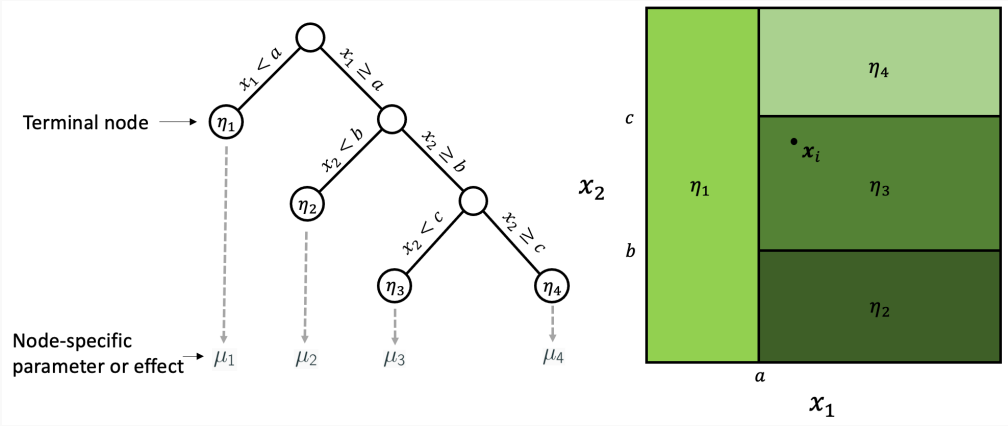
Bayesian Additive Regression Trees (BART)

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

- 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

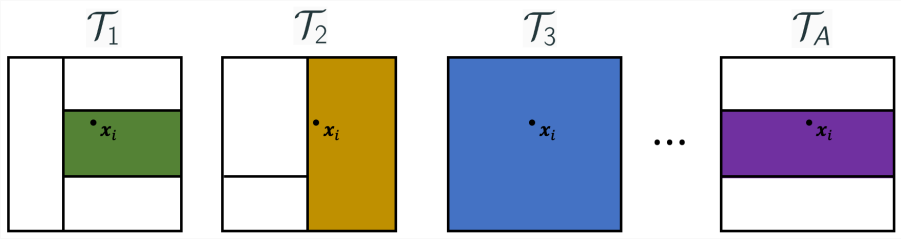
Bayesian Additive Regression Trees (BART)

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



BART

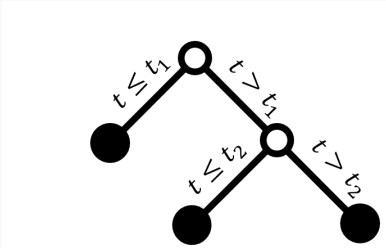
$$f(\boldsymbol{x}_i) = \sum_{a=1}^{A} g(\boldsymbol{x}_i, \mathcal{T}_a)$$

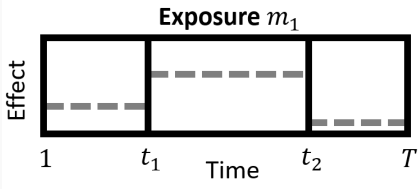


Treed Distributed Lag Model (TDLM)

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

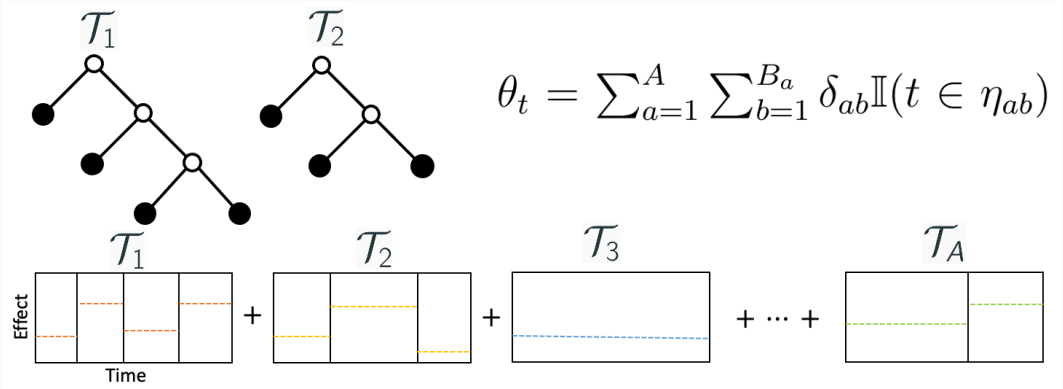
- Apply BART to time ($t = 1, \ldots, T$) to define structure in the lag function $\theta_1, \ldots, \theta_T$
- Constant effect of exposure in each terminal node or time segment



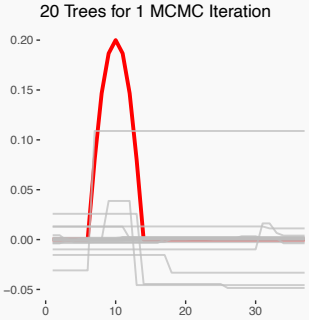


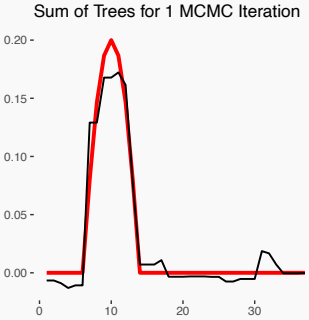
TDLM: Ensemble of Trees

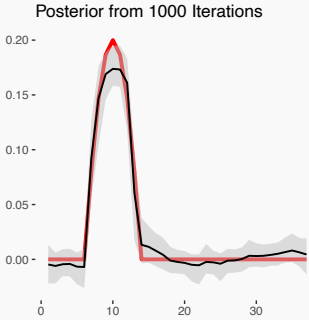
- Use ensemble of A trees
- Adds robustness and can approximate smooth distributed lag functions
- η_{ab} and δ_{ab} is the terminal node and effect for node b on tree a



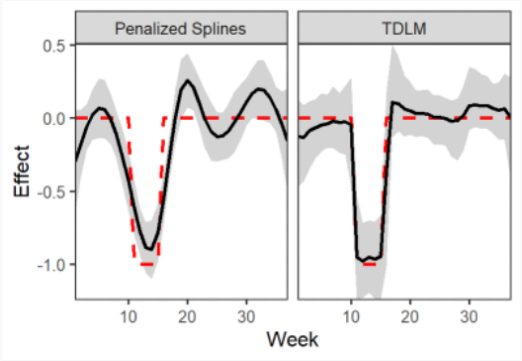
TDLM: Illustrative Example

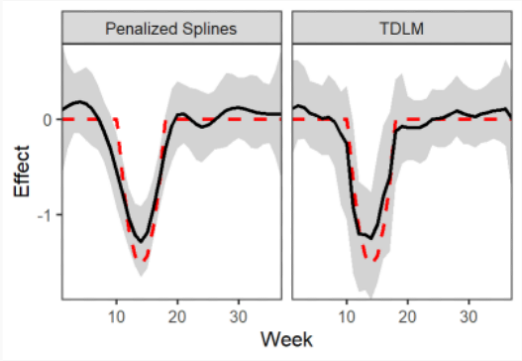






TDLM: Illustrative Example



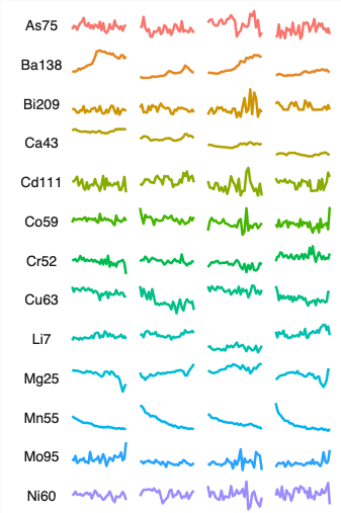


The advantages of trees and TDLM

- 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

Critical Windows with Mixtures

Critical Windows with Mixtures



Challenges of Mixtures Assessed at Longitudinally

- High dimensional exposure space
- High correlation between mixture components
- High autocorrelation within each component
- Nonlinear associations
- Interactions between components including time-sensitive interactions (e.g. priming)

Critical Windows with Mixtures

6 approaches

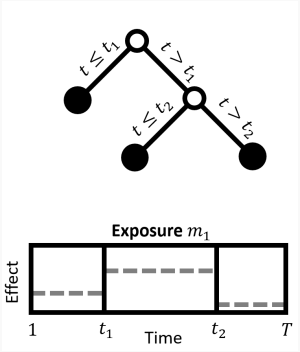
- Bayesian kernel machine regression DLM (Wilson et al., 2021, *AOAS*)
- Treed distributed lag mixture models (Mork and Wilson, 2021, *Biometrics*)
- Spline based component selection (Antonelli, Wilson and Coull, 2021, *Biostatistics*)
- Critical window variable selection for mixtures (Warren et al., 2021, *AOAS*)
- Lagged weighted quantile sums (Bello et al., 2017, *Env. Res.*)
- Partial least squares for quantile regression (Wang et al., 2022 *Biometrics*)

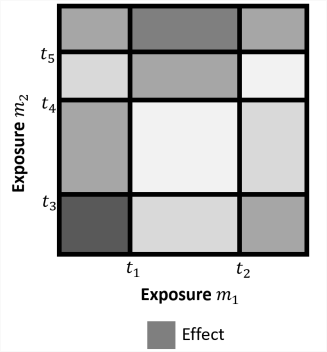
Distributed Lag Mixture Model (DLMM)

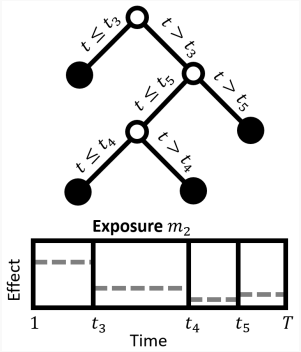
$$y_i = \sum_{m=1}^{M}\sum_{t=1}^{T} x_{imt}\theta_{mt} + \sum_{m_1=1}^{M}\sum_{m_2=m_1}^{M}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T} x_{im_1t_1}x_{im_2t_2}\theta_{m_1m_2t_1t_2} + \mathbf{z}_i'\boldsymbol{\gamma} + \varepsilon_i$$

- θ_{mt} is the main effect of exposure m ($m = 1, \ldots, M$) at time t
- $\theta_{m_1m_2t_1t_2}$ is the interaction among exposures m_1 at time t_1 and m_2 at time t_2
- Includes time-sensitive interactions
- Includes quadratic main effects if we include self interactions
- $MT + \binom{M+1}{2} T^2$ parameters (20,720 in our analysis with $M = 5$ and $T = 37$)

Treed Distributed Lag Mixture Model (TDLMM)







- Structured regression tree pairs add structure to the θ's
- Tree pairs define the main effect and pairwise interaction for two exposures (or a self interaction / quadratic)

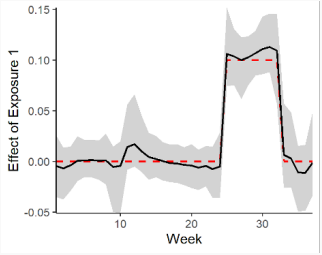
Tree Pairs & Exposure Selection

- Prior on the exposure that each tree is applied to

$$
\begin{aligned}
S_{aj} &= m \qquad \text{if tree } j \text{ in pair } a \text{ is applied to exposure } m \\
S_{aj}|\mathcal{E} &\sim \text{Categorical}(\mathcal{E}) \\
\mathcal{E} &\sim \text{Dirichlet}(\kappa, \ldots, \kappa)
\end{aligned}
$$

- New tree proposal update: switch exposure
- If no tree uses exposure m, that exposure is selected out of the model
- Enforces hierarchical variable selection

TDLM Simulation (single pollutant)

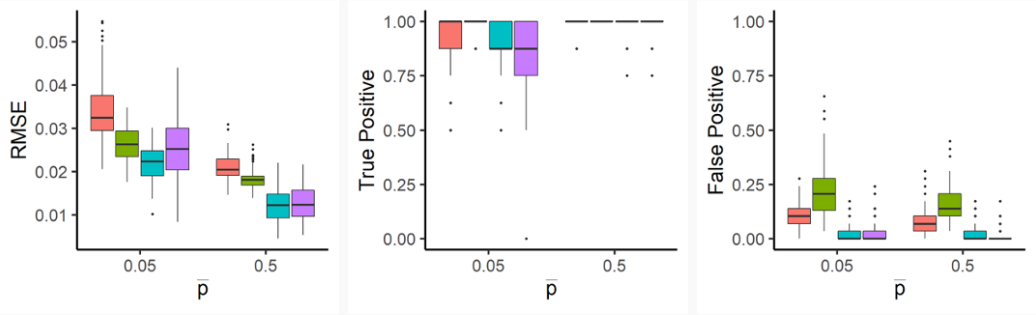


- Scenario 1: Binary outcome, single exposure
- $n = 5000$, two different average probabilities of success (0.05, 0.5)
- Randomly placed, eight-week critical window
- Real Colorado exposure data for $PM_{2.5}$
- Compare:
  - TDLM with a single exposure
  - Penalized cubic regression splines[1]
  - Critical window variable selection (CWVS)[2]
  - TDLMM with four additional exposures in mixture model (NO_2, SO_2, CO, temperature)

[1] Gasparrini et al. (2017) *Biometrics*
[2] Warren et al. (2020) *Biostatistics*

TDLM Simulation (single pollutant)

- Better distributed lag function estimation
- More accurate critical window detection
- Minimal penalty for using TDLMM when only one exposure has a true effect



TDLMM Simulation (mixture with five components)

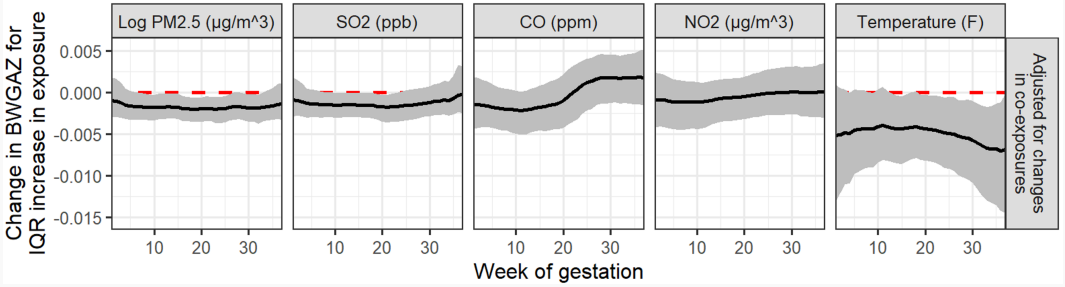
- Second simulation from a mixture with time-sensitive interactions
- Gaussian model
- Overall good performance
  - acceptable RMSE
  - proper 95% interval coverage
  - high precision identifying windows
  - high rate of selecting correct exposures and lower rate of selecting incorrect exposures

Analysis of Colorado Administrative Birth Cohort

- 195,701 full term (37 weeks) births
- Outcome: birth weight z-score (BWGAZ), adjusted for sex, gestational age
- Five exposures assessed weekly during gestation: $PM_{2.5}$, NO_2, SO_2, CO, temperature
- Controlled for: maternal age, weight, income, education, smoking, prenatal care, race, Hispanic, county, elevation, year and month of conception

Main Effects

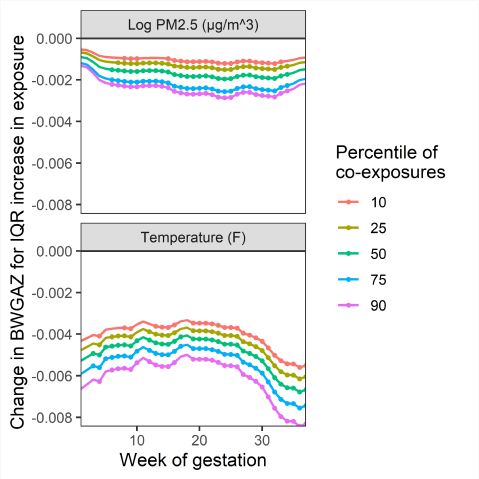
- Many "main effects"
- Here: IQR change of one exposure and the expected corresponding change in the co-exposures

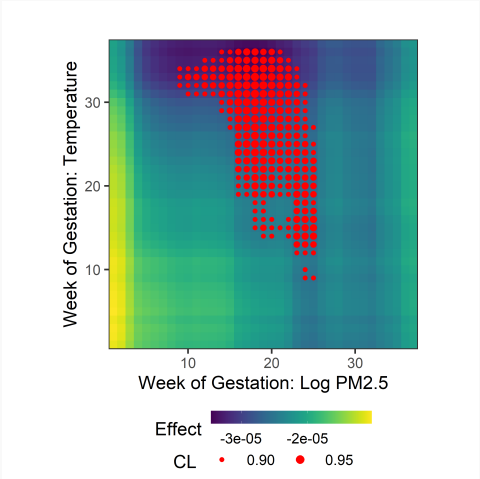


$$\mathrm{E}\left[Y\middle|\widetilde{x}_t = \mathrm{E}\left\{x_t\middle|x_{mt} = x_{m(0.75)}\right\}, \widetilde{x}_{[t]} = \overline{x}, z = z_0\right]$$

$$- \mathrm{E}\left[Y\middle|\widetilde{x}_t = \mathrm{E}\left\{x_t\middle|x_{mt} = x_{m(0.25)}\right\}, \widetilde{x}_{[t]} = \overline{x}, z = z_0\right]$$

Temperature-$PM_{2.5}$ Interaction





Heterogeneous Critical Windows

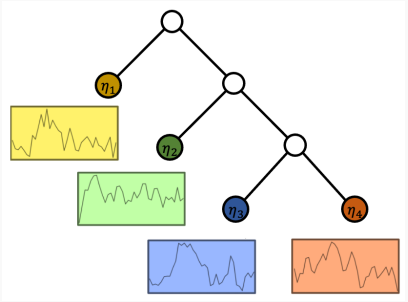
Heterogeneity and Modification with Critical Windows

- 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 continuous modifying factors and multiple modifiers
- Heterogeneity by multiple modifiers poses dimensionality and multiple comparison problems

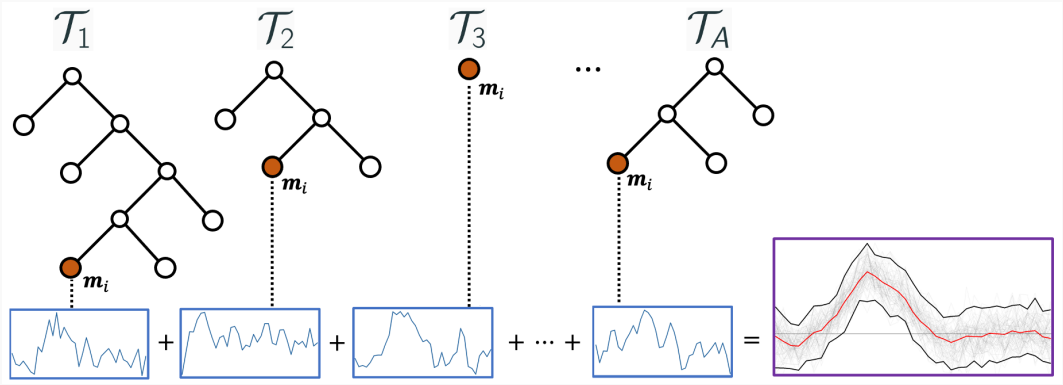
Heterogeneity DLM (HDLM)

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

- 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



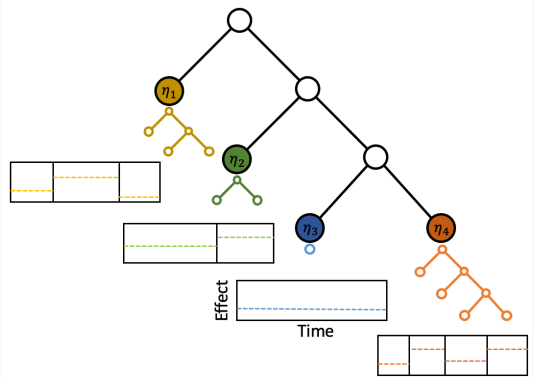
Heterogeneity DLM (HDLM)



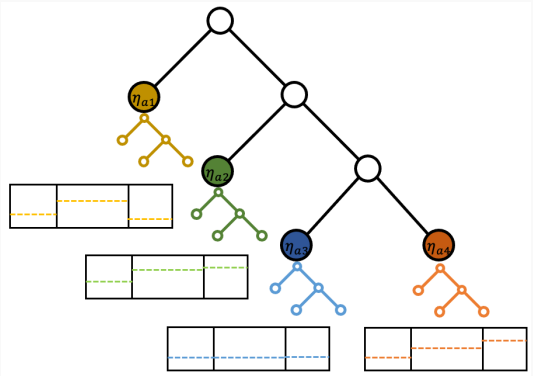
Nested and Shared Tree HDLM

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

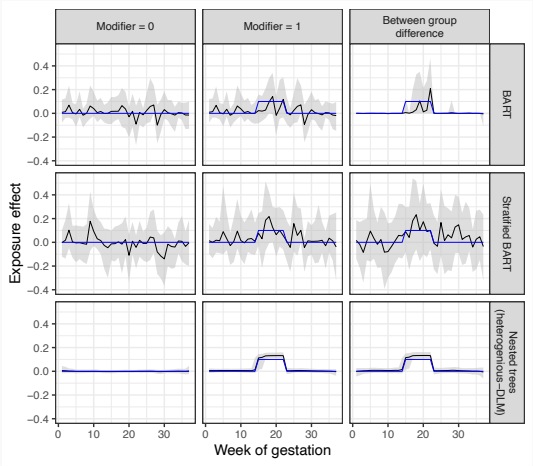
Nested Tree HDLM



Shared Tree HDLM



Nested and Shared Tree HDLM

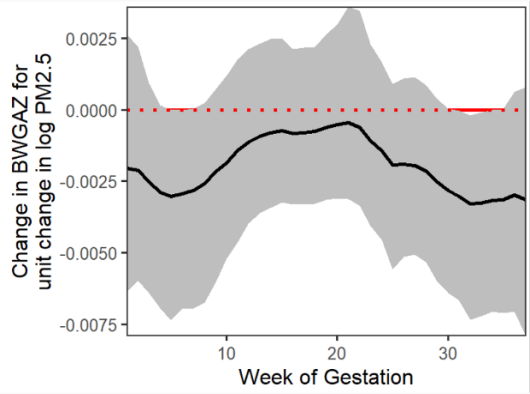


- HDLMs have nominal coverage and low false window detection rates
- Includes true modifiers with high probability
- Includes null modifiers with lower probability (0.6-0.7)
- Treed-DLM approaches better than GP-DLM when subgroups effects vary in smoothness
- Comparable to DLM when there is no heterogeneity

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

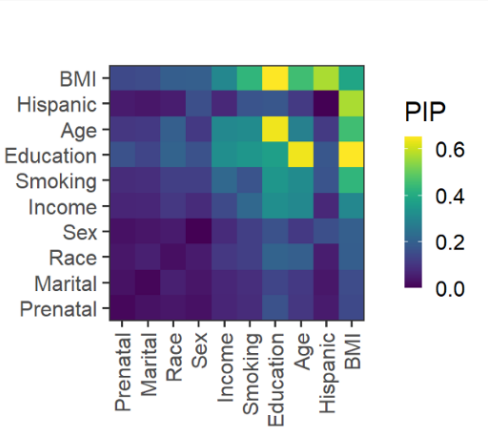
Analysis with DLM (no heterogeneity)



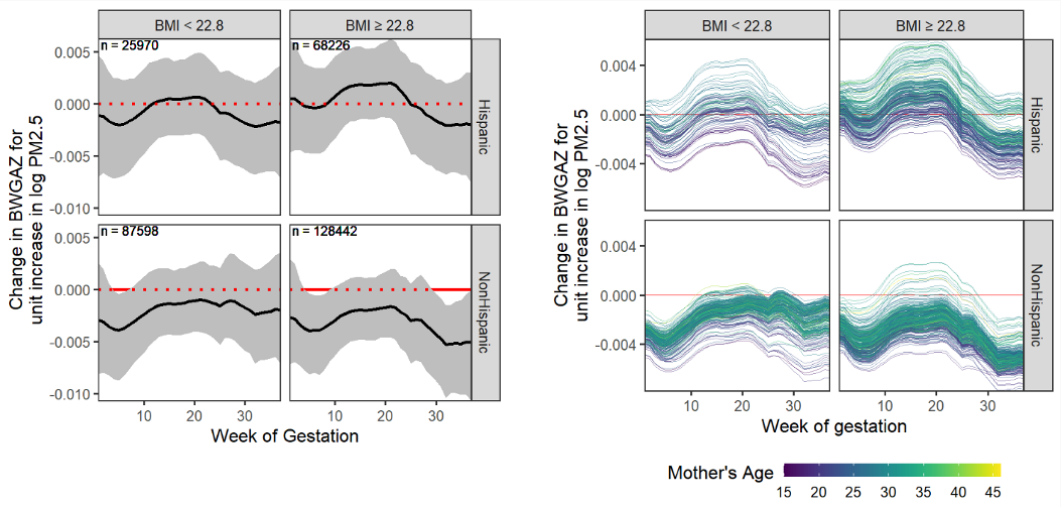
Modifier Selection

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		

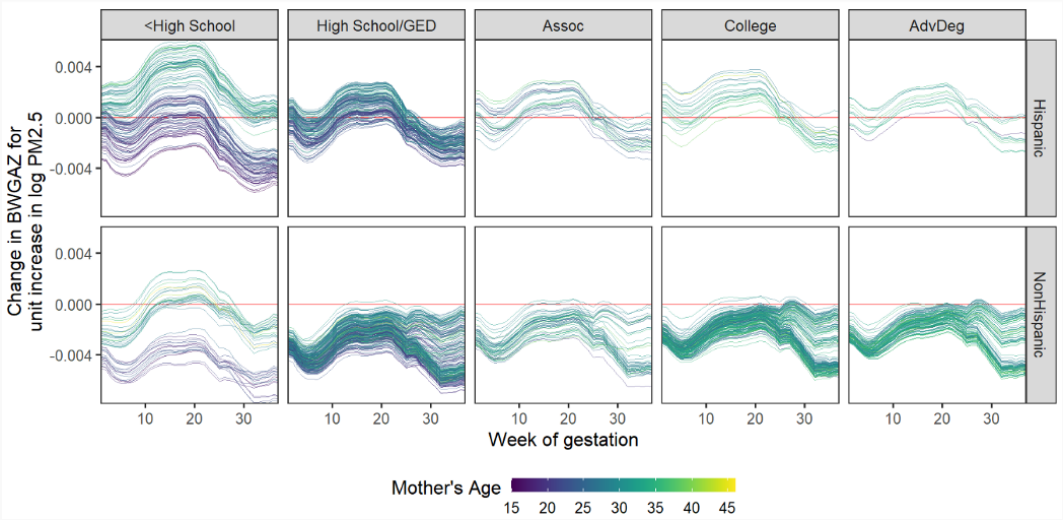
PIP = Posterior Inclusion Probability



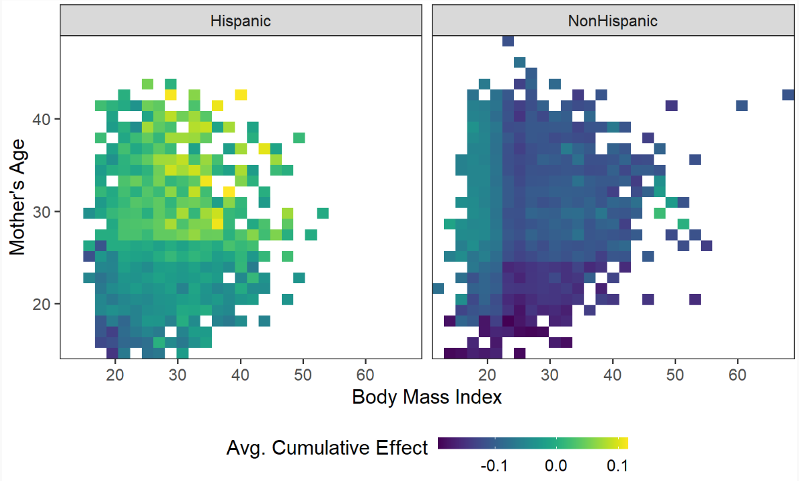
Modification by Maternal BMI and Hispanic Status



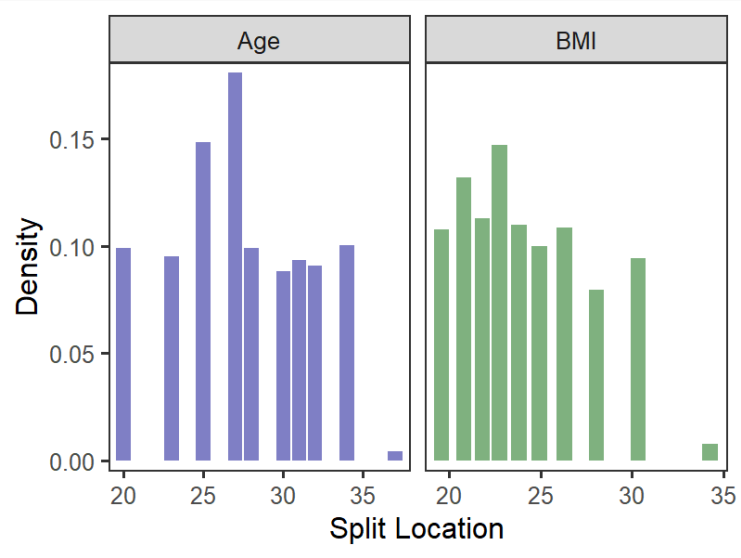
Modification by Maternal Education and Hispanic Status



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 mixtures
- Allows for heterogeneity
- Overall good finite sample properties
- Available for linear and logistic regression (zero inflated count data coming soon)
- Treed distributed lag nonlinear model also available (Mork and Wilson 2021, *Biostatistics*)
- R code available: `github.com/danielmork/dlmtree`

Thank You

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

Mork, D., Wilson, A. (In press). 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. (2022). Heterogeneous Distributed Lag Models to Estimate Personalized Effects of Maternal Exposures to Air Pollution. http://arxiv.org/abs/2109.13763

References

References i

Antonelli, J, A Wilson, and B Coull (2021). *Multiple exposure distributed lag models with variable selection*. arXiv: 2107.14567.

Bello, GA, M Arora, C Austin, MK Horton, RO Wright, and C Gennings (2017). "Extending the Distributed Lag Model framework to handle chemical mixtures". *Environmental Research* 156 (December 2016).

Chipman, HA, EI George, and RE McCulloch (2010). "BART: Bayesian additive regression trees". *The Annals of Applied Statistics* 4 (1).

Chipman, HA, EI George, and RE McCulloch (1998). "Bayesian CART Model Search". *Journal of the American Statistical Association* 93 (443).

Gasparrini, A, F Scheipl, B Armstrong, and MG Kenward (2017). "A penalized framework for distributed lag non-linear models". *Biometrics* 73 (3).

Leung, M, ST Rowland, BA Coull, AM Modest, MR Hacker, J Schwartz, MA Kioumourtzoglou, MG Weisskopf, and A Wilson (2022). "Bias Amplification and Variance Inflation in Distributed Lag Models Using Low-Spatial-Resolution Data". *American Journal of Epidemiology*.

Mork, D, MA Kioumourtzoglou, M Weisskopf, BA Coull, and A Wilson (2021). "Heterogeneous Distributed Lag Models to Estimate Personalized Effects of Maternal Exposures to Air Pollution". arXiv: 2109.13763.

References ii

Mork, D and A Wilson (2021). "Estimating perinatal critical windows of susceptibility to environmental mixtures via structured Bayesian regression tree pairs". *Biometrics*.

Mork, D and A Wilson (2022). "Treed distributed lag nonlinear models". *Biostatistics* 23.3.

Wang, Y, A Ghassabian, B Gu, Y Afanasyeva, Y Li, L Trasande, and M Liu (2022). "Semiparametric distributed lag quantile regression for modeling time-dependent exposure mixtures". *Biometrics*.

Warren, JL, HH Chang, LK Warren, MJ Strickland, LA Darrow, and JA Mulholland (2022). "Critical window variable selection for mixtures: Estimating the impact of multiple air pollutants on stillbirth". *The Annals of Applied Statistics* 16.3. arXiv: 2104.09730.

Warren, JL, W Kong, TJ Luben, and HH Chang (2020). "Critical window variable selection: estimating the impact of air pollution on very preterm birth". *Biostatistics* 21 (4).

Wilson, A, YHM Chiu, HHL Hsu, RO Wright, RJ Wright, and BA Coull (2017a). "Bayesian distributed lag interaction models to identify perinatal windows of vulnerability in children's health". *Biostatistics* 18 (3).

Wilson, A, YHM Chiu, HHL Hsu, RO Wright, RJ Wright, and BA Coull (2017b). "Potential for Bias When Estimating Critical Windows for Air Pollution in Children's Health". *American Journal of Epidemiology* 186.11.

References iii

Wilson, A, HHL Hsu, YHM Chiu, RO Wright, RJ Wright, and BA Coull (In press). "Kernel Machine and Distributed Lag Models for Assessing Windows of Susceptibility to Environmental Mixtures in Children's Health Studies". *Annals of Applied Statistics*. arXiv: 1904.12417.