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

Ander Wilson

Colorado State University

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

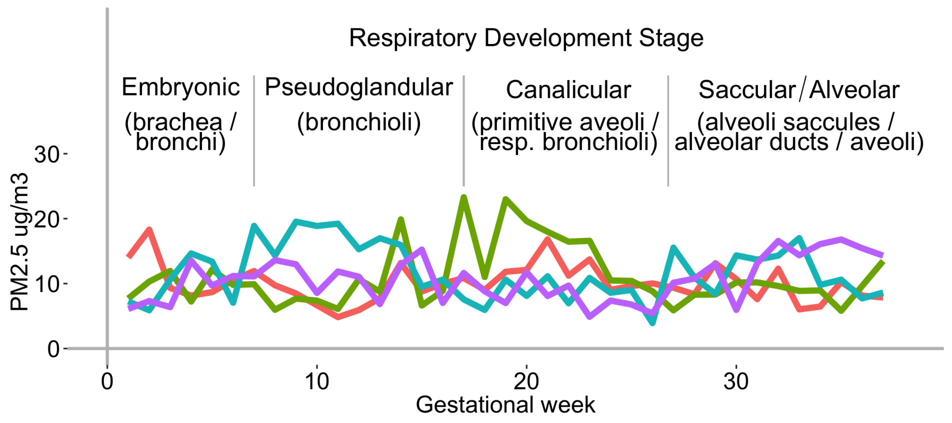
Contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

Air Pollution is Bad

Critical Windows of Susceptibility

Definition

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

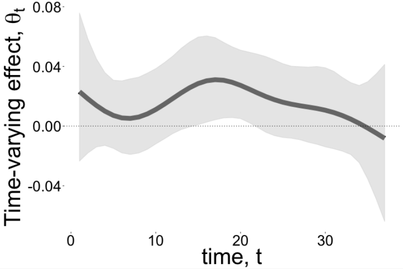


Distributed Lag Model (DLM)

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

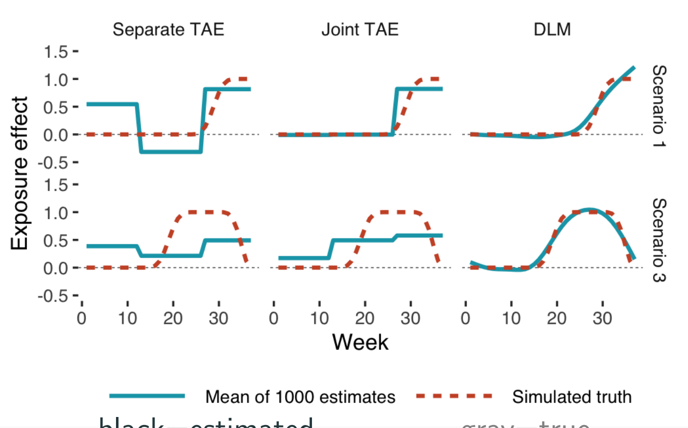
- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)'$ *constrained* to vary smoothly in time (e.g. spline, Gaussian process, ...)
  - adds stability to the model
  - conforms with biological hypothesis that exposure at proximal time points are likely to have similar effects

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



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

DLM Compared to Trimester Average Exposures



black=estimated gray=true

Advantages of DLM

- Simultaneous adjustment for exposures at all time points
- Data driven identification of windows

[1] Wilson et al. (2017b) *American Journal of Epidemiology*

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

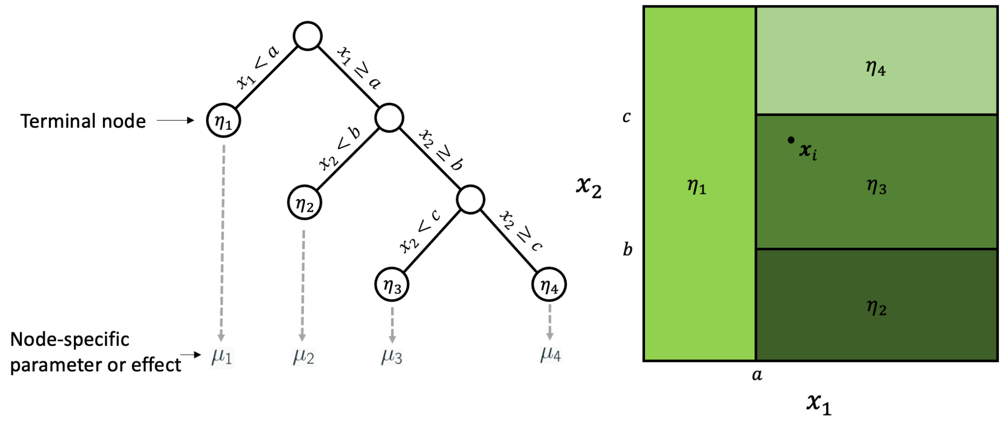
Bayesian Additive Regression Trees (BART)

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

- 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

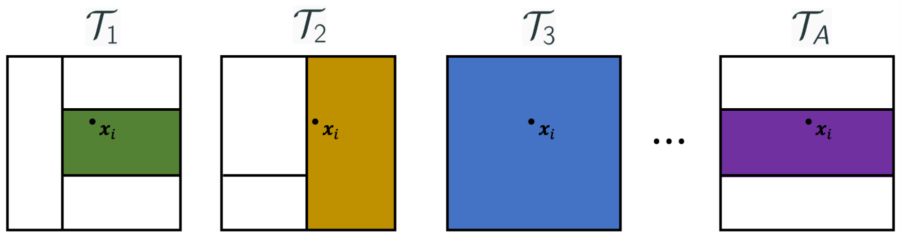
Bayesian Additive Regression Trees (BART)

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



BART

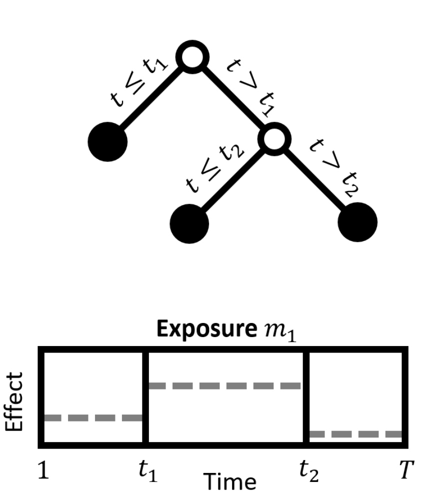
$$f(x_i) = \sum_{a=1}^{A} g(x_i, \mathcal{T}_a)$$



Treed Distributed Lag Model (TDLM)

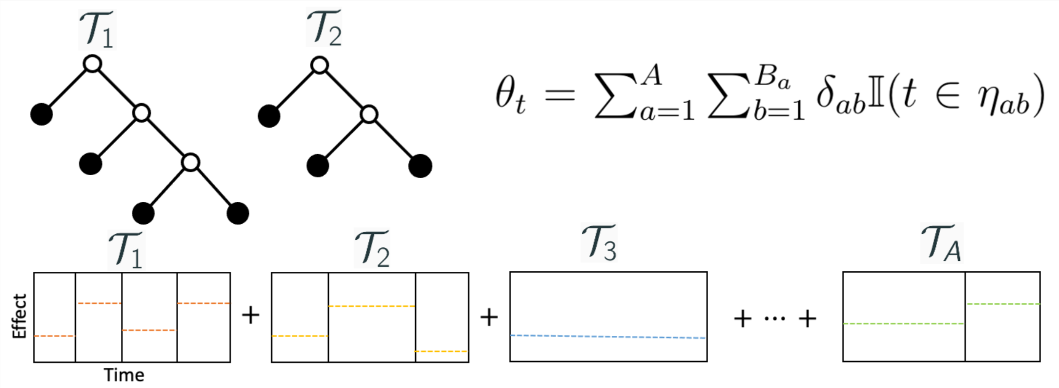
$$y_i = \sum_{t=1}^{T} x_{it}\theta_t + z_i'\gamma + \varepsilon_i$$

- 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



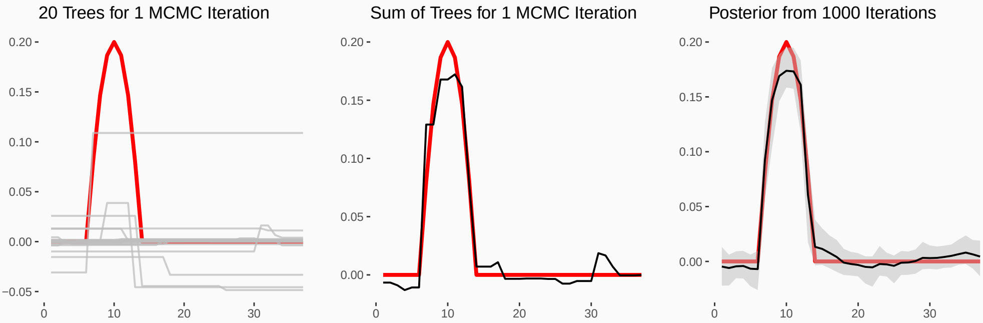
TDLM: Ensemble of Trees

- 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*

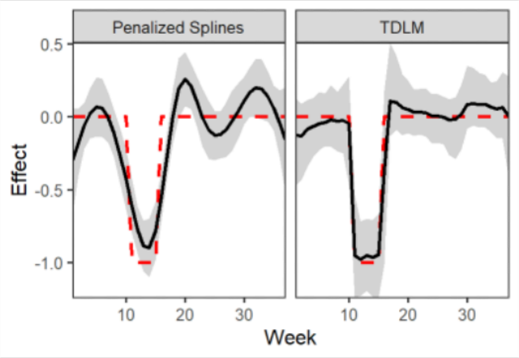


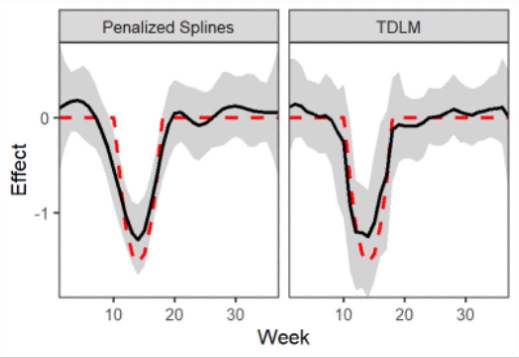
$$\theta_t = \sum_{a=1}^{A} \sum_{b=1}^{B_a} \delta_{ab} \mathbb{I}(t \in \eta_{ab})$$

TDLM: Illustrative Example



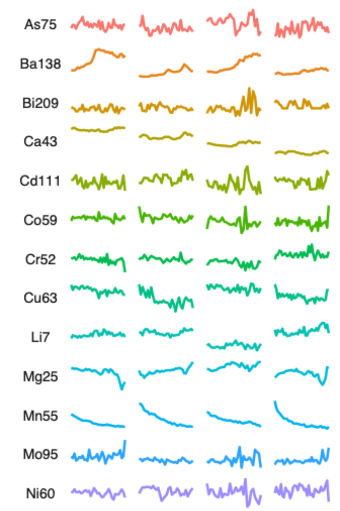
TDLM: Illustrative Example





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

Five approaches

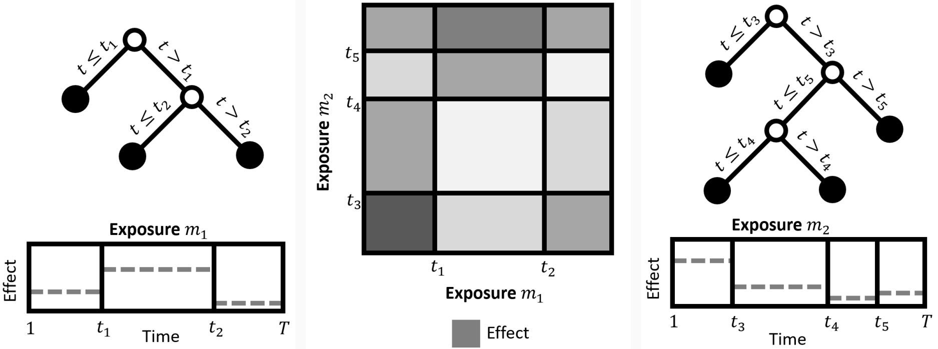
- Bayesian kernel machine regression DLM (Wilson et al., 2021, *AOAS*)
- Treed distributed lag mixture models (Mork and Wilson, 2021, *Biometrics*)
- Spline based component selection (Antonelli, Wilson and Coull, 2021, preprint)
- Critical window variable selection for mixtures (Warren et al., 2021,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

Distributed Lag Mixture Model (DLMM)

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

- 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)

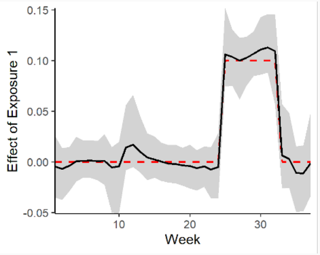
Tree Pairs & Exposure Selection

- 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)



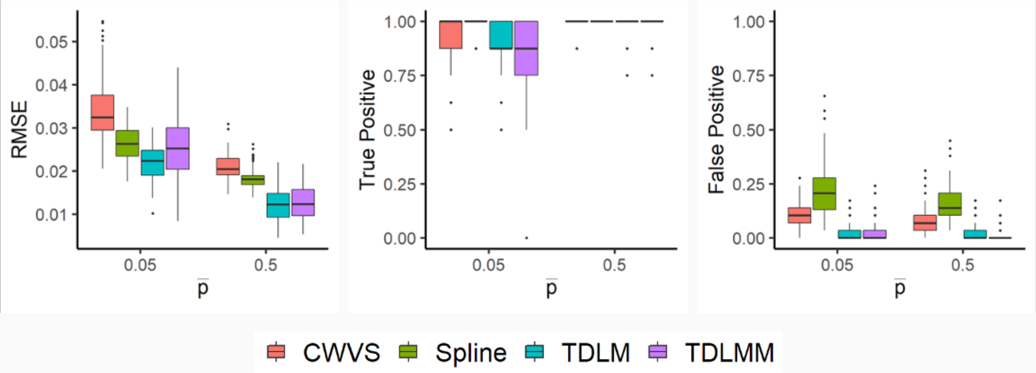
- Scenario 1: Binary outcome, single exposure
- $n = 5000$, two different average probabilities of success (0.05, 0.5)
- Randomly placed, eight-week critical window
- Real Colorado exposure data for $PM_{2.5}$
- Compare:
    - TDLM with a single exposure
    - Penalized cubic regression splines[1]
    - 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)

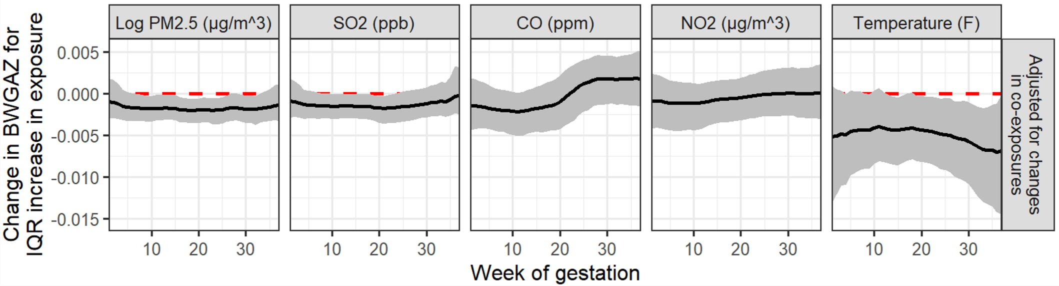
- Second simulation from a mixture with time-sensitive interactions
- Gaussian model
- Overall good performance
  - acceptable RMSE
  - proper 95% interval coverage
  - high precision identifying windows
  - high rate of selecting correct exposures and lower rate of selecting incorrect exposures

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

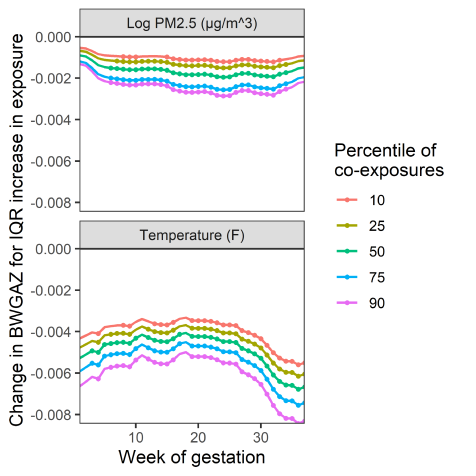
- Many "main effects"
- Here: IQR change of one exposure and the expected corresponding change in the co-exposures

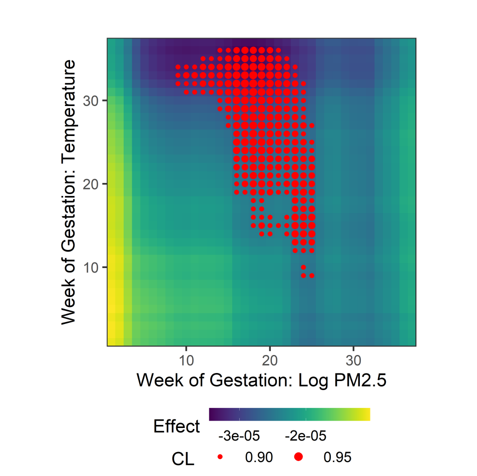


$$E\left[Y\middle|\widetilde{x}_t = E\left\{\boldsymbol{x}_t\middle|x_{mt} = x_{m(0.75)}\right\}, \widetilde{x}_{[t]} = \overline{\boldsymbol{x}}, \boldsymbol{z} = \boldsymbol{z}_0\right]$$

$$- E\left[Y\middle|\widetilde{x}_t = E\left\{\boldsymbol{x}_t\middle|x_{mt} = x_{m(0.25)}\right\}, \widetilde{x}_{[t]} = \overline{\boldsymbol{x}}, \boldsymbol{z} = \boldsymbol{z}_0\right]$$

Temperature-$PM_{2.5}$ Interaction





Heterogeneous Critical Windows

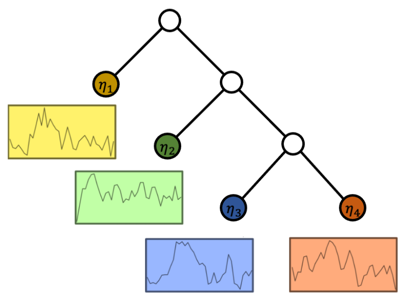
Heterogeneity and Modification with Critical Windows

- 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

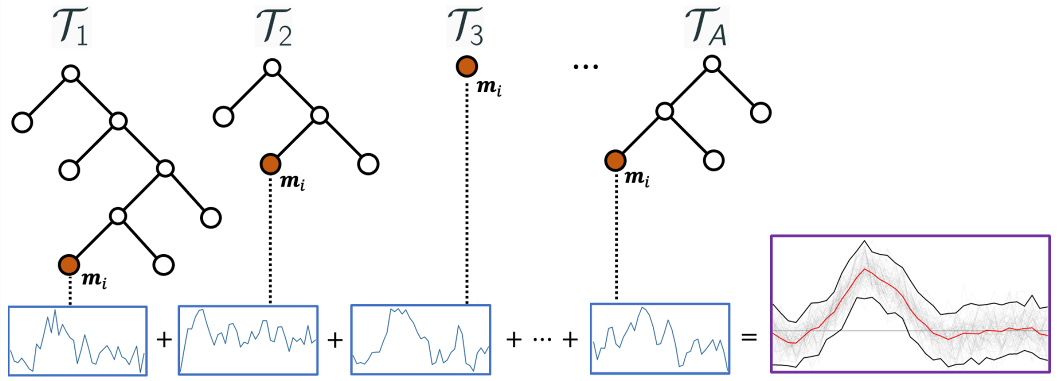
Heterogeneity DLM (HDLM)

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

- 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



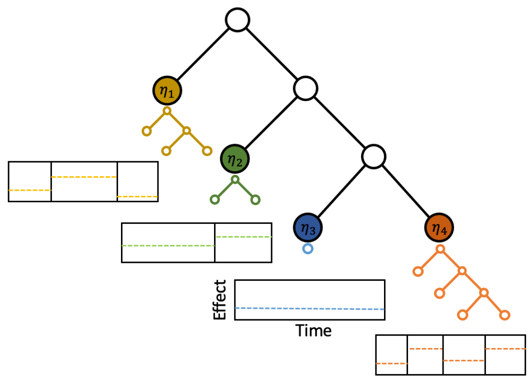
Heterogeneity DLM (HDLM)



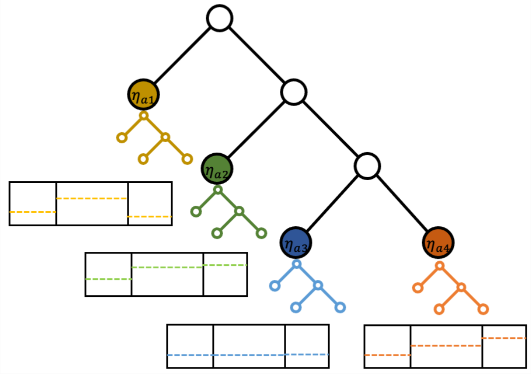
Nested and Shared Tree HDLM

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

Nested Tree HDLM



Shared Tree HDLM



Simulation

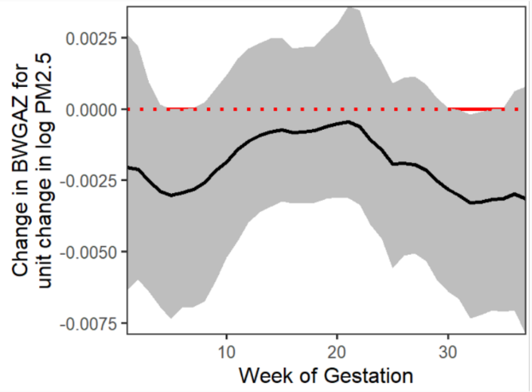
- HDLMs have nominal coverage and low false window detection rates
- Includes true modifiers with high probability
- Includes null modifiers with lower probability (0.6-0.7)
- Treed-DLM approaches better than GP-DLM when subgroups effects vary in smoothness
- Comparable to DLM when there is no heterogeneity

Birth Weight Analysis

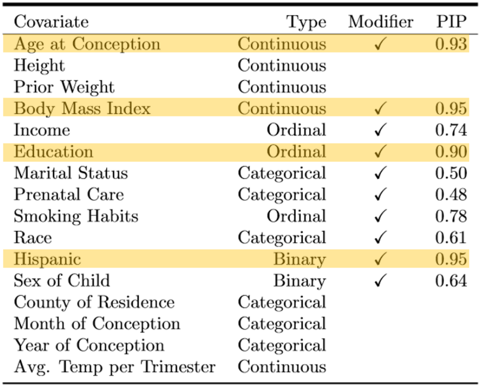
ort Collins
Denver
Unite
Grand Junction
COLORADO
Colorado Springs
Pueblo

- 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)

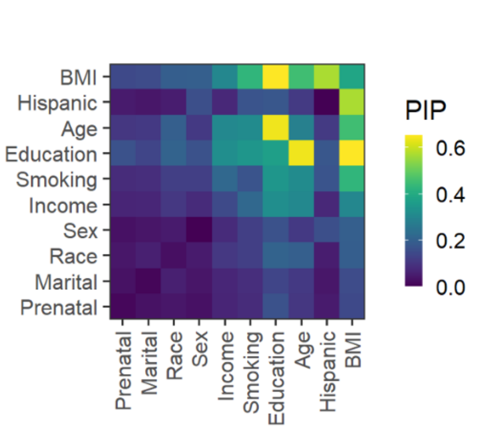


Modifier Selection

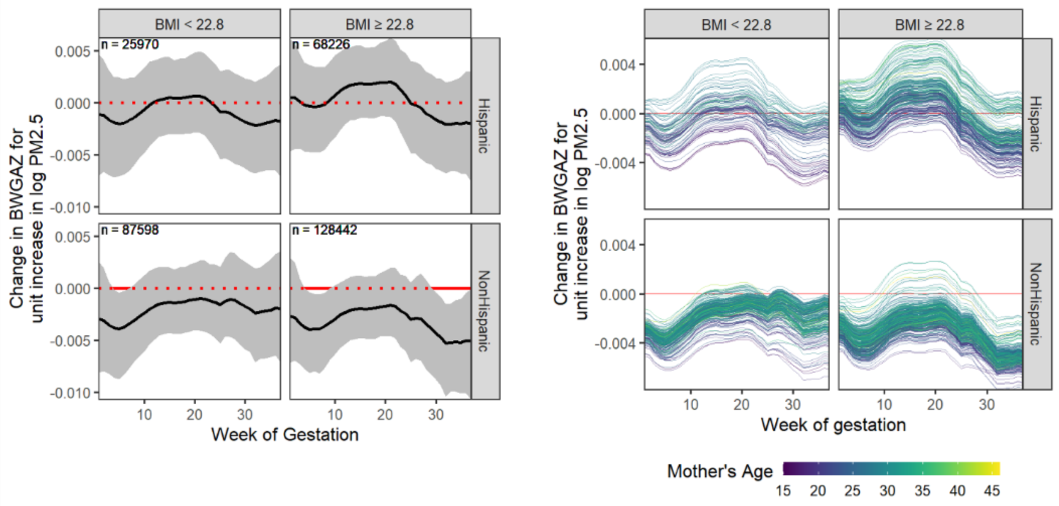


Covariate	Type	Modifier	PIP
Age at Conception	Continuous	✓	0.93
Height	Continuous		
Prior Weight	Continuous		
Body Mass Index	Continuous	✓	0.95
Income	Ordinal	✓	0.74
Education	Ordinal	✓	0.90
Marital Status	Categorical	✓	0.50
Prenatal Care	Categorical	✓	0.48
Smoking Habits	Ordinal	✓	0.78
Race	Categorical	✓	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		

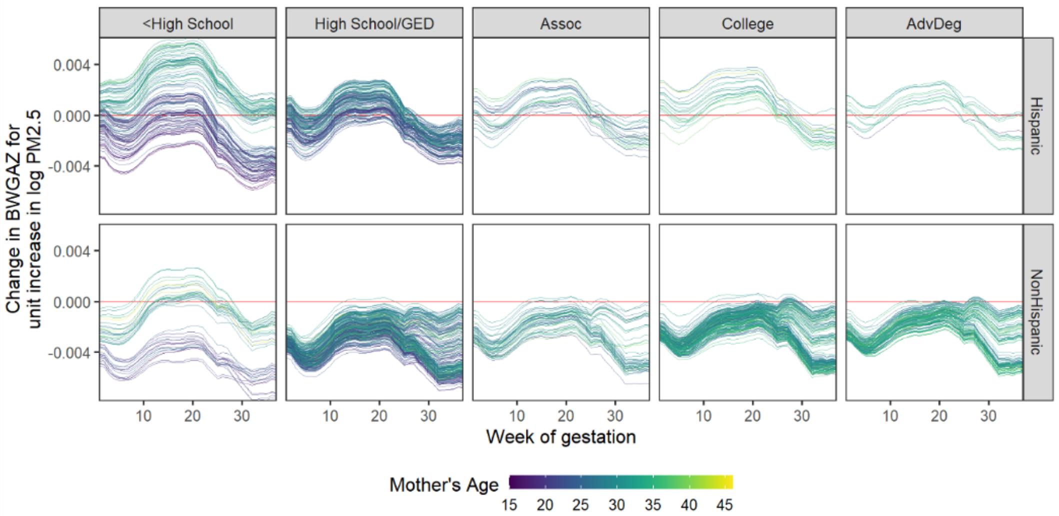
PIP = Posterior Inclusion Probability



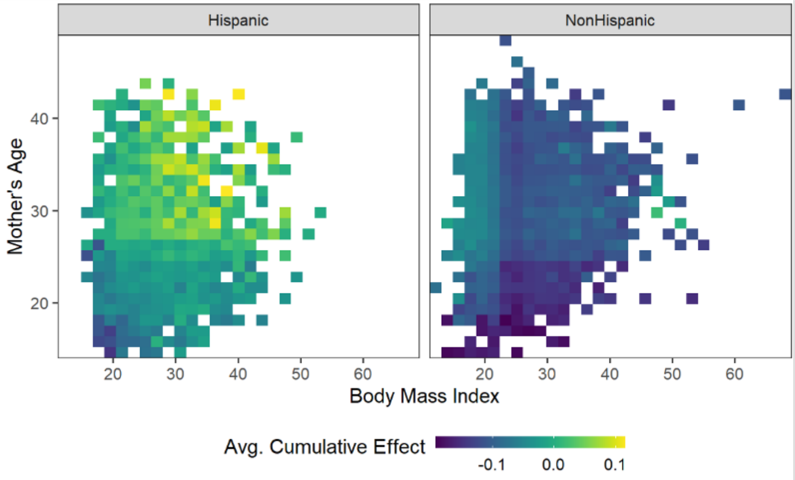
Modification by Maternal BMI and Hispanic Status



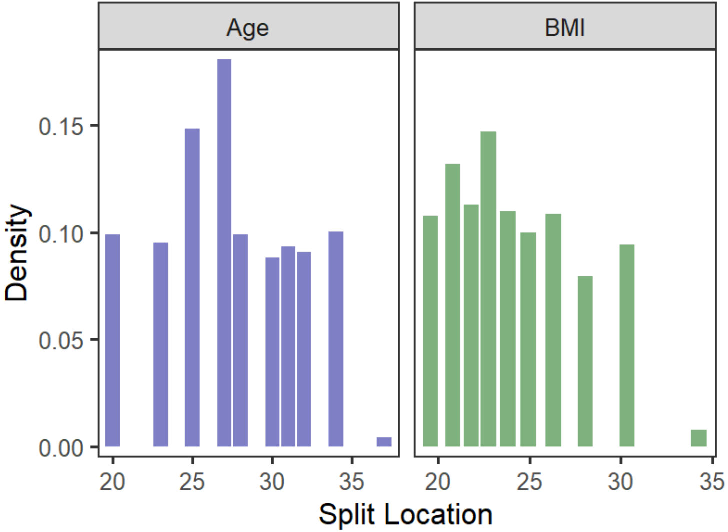
Modification by Maternal Education and Hispanic Status



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. (2021+). 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).

Lee, A, HHL Hsu, YHM Chiu, S Bose, MJ Rosa, I Kloog, A Wilson, J Schwartz, S Cohen, BA Coull, RO Wright, and RJ Wright (2018). "Prenatal fine particulate exposure and early childhood asthma: Effect of maternal stress and fetal sex". *Journal of Allergy and Clinical Immunology* 141 (5).

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 (2021a). "Estimating perinatal critical windows of susceptibility to environmental mixtures via structured Bayesian regression tree pairs". *Biometrics*.

Mork, D and A Wilson (2021b). "Treed distributed lag nonlinear models". *Biostatistics*.

Warren, JL, HH Chang, LK Warren, MJ Strickland, LA Darrow, and JA Mulholland (2021). *Critical Window Variable Selection for Mixtures: Estimating the Impact of Multiple Air Pollutants on Stillbirth*. 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).

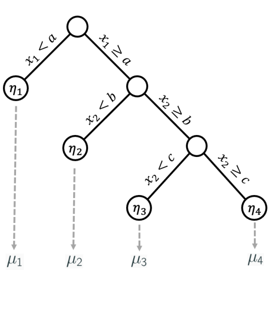
Wilson, A, YHM Chiu, HHL Hsu, RO Wright, RJ Wright, and BA Coull (2017a). "Bayesian distributed lag interaction models to identify perinatal windows of vulnerability in children's health". *Biostatistics* 18 (3).

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).

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.

BART

BART Priors

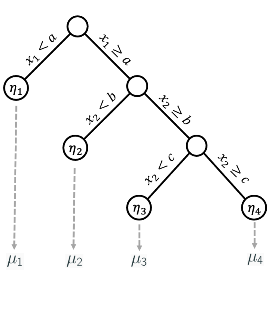


- Implicit prior based on tree generating process
- Three parts:
  - Prior that a node at tree depth d splits
  - Prior on variable that is split at a node (e.g. uniform from all variables)

$$\alpha(1+d)^{-\beta} \qquad \alpha \in (0,1),\ \beta \in [0,\infty)$$

  - Prior on a rule that splits that variable (e.g. uniform breaks in range or uniform of subgroups of categorical variables)
- Independent Gaussian priors on μs

BART Computation



- μs can be integrated out to avoid changing parameter space problem
- Bayesian backfitting updates one tree at a time with Metropolis–Hastings
- Four possible tree-update steps
  - Grow
  - Prune
  - Change splitting rule
  - Swap parent and child node order
- Update other parameters with Gibbs

TDLMM

TDLM Priors

$$\begin{aligned}
\delta_{ab} | \tau_a^2, \nu^2, \sigma^2 &\sim \mathcal{N}(0, \tau_a^2 \nu^2 \sigma^2) \\
\nu &\sim \mathcal{C}^+(0,1) \\
\tau_a &\sim \mathcal{C}^+(0,1)
\end{aligned}$$

$$\begin{aligned}
\sigma &\sim \mathcal{C}^+(0,1) \\
\boldsymbol{\gamma} &\sim \mathcal{MVN}(0, \sigma^2 c \boldsymbol{I})
\end{aligned}$$

$$\alpha = 0.95, \; \beta = 2$$

TDLMM Priors

$$\begin{aligned}
\delta_{ajb} | \mu^2_{S_{aj}}, \nu^2, \sigma^2 &\sim \mathcal{N}(0, \mu^2_{S_{aj}} \nu^2 \sigma^2) \qquad \text{(main effects)} \\
\mu_{S_{aj}} &\sim \mathcal{C}^+(0, 1) \\
\zeta_{ab_1 b_2} | \mu^2_{S_{a1} S_{a2}}, \nu^2, \sigma^2 &\sim \mathcal{N}(0, \mu^2_{S_{a1} S_{a2}} \nu^2 \sigma^2) \qquad \text{(interactions terms)} \\
\mu_{S_{a1} S_{a2}} &\sim \mathcal{C}^+(0, 1) \\
\nu &\sim \mathcal{C}^+(0, 1)
\end{aligned}$$

$$\begin{aligned}
\sigma &\sim \mathcal{C}^+(0, 1) \\
\boldsymbol{\gamma} &\sim \mathcal{MVN}(0, \sigma^2 c \boldsymbol{I})
\end{aligned}$$

$$\alpha = 0.95, \ \beta = 2$$

TDLMM Computation

Key modifications to the BART MCMC algorithm:

- Integrate out fixed effect when estimating trees and distributed lag effects
- New proposal step: switch exposure, accepted with Metropolis-Hastings algorithm Simultaneous integration over all distributed lag effects during tree update
- Multivariate draw of tree terminal node and interaction parameters
- Logistic regression method for regression trees using Polya Gamma latent variable (Polson, Scott, Windle, 2013, *JASA*)
- Methods for zero inflated count data coming soon.
- Posterior analysis of tree structures, exposure, and estimates gives distributed lag effects and uncertainty

HDLM

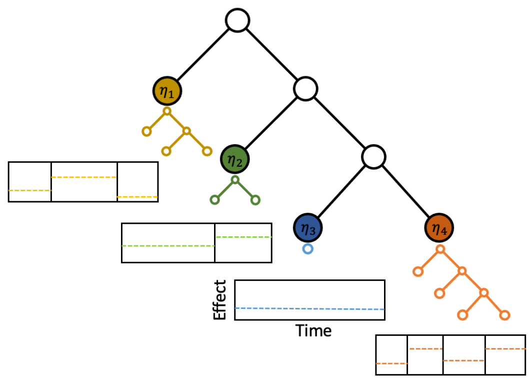
HDLM Priors

$$
\begin{aligned}
\delta_{abc}|\tau_a, \nu, \sigma &\sim \mathcal{N}(0, \tau_a^2 \nu^2 \sigma^2) \\
\nu &\sim \mathcal{C}^+(0, 1) \\
\tau_a &\sim \mathcal{C}^+(0, 1)
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\gamma} &\sim \mathcal{MVN}(0, d\sigma^2 I_p) \\
\sigma &\sim \mathcal{C}^+(0, 1)
\end{aligned}
$$

HDLM Computation

Nested Tree HDLM



- Iterate updating modifier tree and distributed lag trees
- For nested tree, new proposal step to grow or prune a modifier tree that relies on proposing a new distributed lag tree for each new terminal node
- Reduced computation for shared tree
- DLM trees updated using standard approaches

HDLM Computation for Nested Tree Model

1. Integrate over all fixed effects
2. For each tree in ensemble, conditional on other trees
    2.1 Propose new modifier tree using modifierselection probabilities
        - If grow/prune, draw new treed DLM
        - Simultaneously integrate over all distributed lag effects, accept proposal with Metropolis-Hastings (MH)

    2.2 For each nested tree, propose new treed DLM
        - Integrate over distributed lag effects
        - Accept proposal with MH

    2.3 Draw distributed lag effects for each nested tree from full conditional
3. Draw fixed effects conditional on exposure effects, variance parameters, andmodifier selection probabilities
4. Iterate 2-3 until posterior distribution has been adequately sampled

HLDM Simulation 1, early late effect

	Effect ($z_{i1} > 0$)				No Effect ($z_{i1} \leq 0$)			
Model	RMSE*	Coverage	TP	FP	RMSE*	Coverage	FP	MSPE
Nested Tree HDLM	6.72	0.91	0.88	0.04	2.67	1.00	0.00	0.977
Shared Tree HDLM	7.39	0.90	0.86	0.05	3.18	0.99	0.01	0.978
Gaussian Process HDLM	7.14	0.95	0.90	0.02	3.94	1.00	0.00	0.978
Nested Tree: Truth	5.29	0.96	0.94	0.02	1.96	1.00	0.00	0.974
Gaussian Process: Truth	6.43	0.96	0.97	0.02	3.63	1.00	0.00	0.987
Treed DLM	11.45	0.64	0.84	0.18	5.52	0.68	0.32	1.000
Gaussian Process DLM	11.49	0.70	0.61	0.11	5.55	0.78	0.22	0.999

HLDM Simulation 1, Early Late Effect

	Effect ($z_{i1} > 0$)				No Effect ($z_{i1} \leq 0$)			
Model	RMSE*	Coverage	TP	FP	RMSE*	Coverage	FP	MSPE
Nested Tree HDLM	6.72	0.91	0.88	0.04	2.67	1.00	0.00	0.977
Shared Tree HDLM	7.39	0.90	0.86	0.05	3.18	0.99	0.01	0.978
Gaussian Process HDLM	7.14	0.95	0.90	0.02	3.94	1.00	0.00	0.978
Nested Tree: Truth	5.29	0.96	0.94	0.02	1.96	1.00	0.00	0.974
Gaussian Process: Truth	6.43	0.96	0.97	0.02	3.63	1.00	0.00	0.987
Treed DLM	11.45	0.64	0.84	0.18	5.52	0.68	0.32	1.000
Gaussian Process DLM	11.49	0.70	0.61	0.11	5.55	0.78	0.22	0.999

HLDM Simulation 1, Scaled Effect

	Effect ($z_{i1} > 0$)				No Effect ($z_{i1} \leq 0$)			
Model	RMSE*	Coverage	TP	FP	RMSE*	Coverage	FP	MSPE
$\sigma^2 = 25$								
Nested Tree HDLM	6.62	0.92	0.66	0.01	2.59	1.00	0.00	0.969
Shared Tree HDLM	6.53	0.93	0.76	0.01	2.85	0.99	0.01	0.969
Gaussian Process HDLM	7.53	0.94	0.63	0.01	3.98	1.00	0.00	0.971
Treed DLM	10.92	0.82	0.92	0.02	5.72	0.78	0.22	1.000
Gaussian Process DLM	11.12	0.84	0.86	0.01	5.71	0.80	0.20	1.000

HLDM Simulation 3, No Heterogeneity

Model	RMSE×100	Coverage	TP	FP	MSPE
Nested Tree HDLM	3.42	0.97	0.76	0.00	1.001
Shared Tree HDLM	3.35	0.97	0.76	0.00	1.001
Gaussian Process HDLM	3.84	1.00	0.76	0.00	1.001
Treed DLM	3.01	0.98	0.82	0.01	1.000
Gaussian Process DLM	3.64	0.99	0.89	0.01	1.003