Bias amplification and variance inflation in distributed lag models using low spatial resolution data

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Michael Leung[1], Sebastian T Rowland[2], Brent A Coull[1], Anna M Modest[3,4], Michele R Hacker[3,4], Joel Schwartz[1], Marianthi-Anna Kioumourtzoglou[2], Marc G Weisskopf[1], Ander Wilson[5]

[1]Harvard T.H. Chan School of Public Health, Boston, MA; [2]Columbia University Mailman School of Public Health, New York, NY; [3]Beth Israel Deaconess Medical Center, Boston, MA; [4]Harvard Medical School, Boston, MA; [5]Colorado State University, Fort Collins, CO

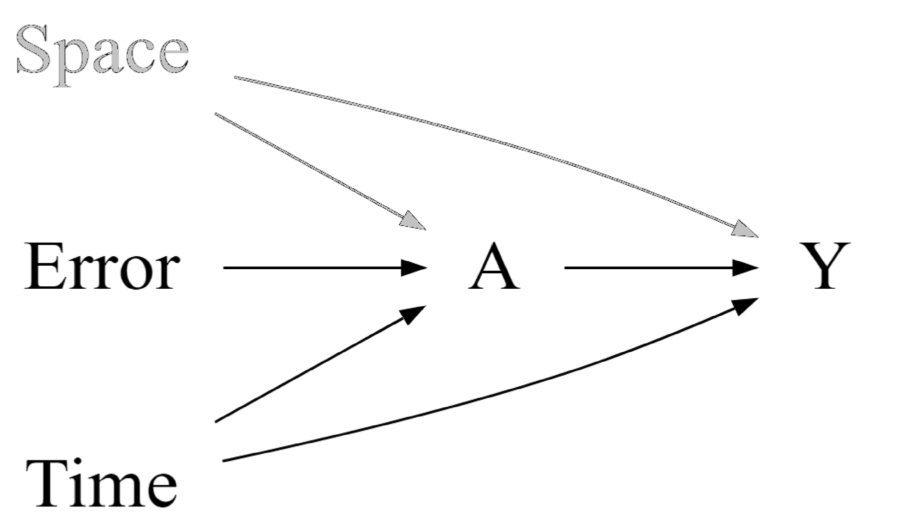

Fig 1. DAG for bias amplification. Removing the spatial component can amplify bias by time trends.

Background

- Distributed lag models (DLMs) are often used to estimate lagged associations & identify critical exposure windows
- DLMs are commonly fit to low-spatial-resolution exposure data (e.g., single monitor or weather etc.)
- We focus on two perils that can occur when fitting DLMs to low-spatial-resolution exposure data:
  1. Bias amplification ("Z-bias") in the presence of residual confounding by time trends — by removing the spatial component, time trends explain a larger fraction of exposure variation; thus, any residual confounding by time trends manifests as a larger bias (Fig 1)
  2. Variance inflation from concurvity between a distributed lag function and a secular function of time
- We demonstrate these issues using the example of NO_2 and birth weight through a real-data analysis and simulation study

Methods

- **Data Analysis**: We first estimate the NO_2-birth weight lag-response function (per 10 ppb) using DLMs in a Massachusetts-based cohort
  - Birth weight (grams) from 41,653 deliveries at Beth Israel Deaconess Medical Center, 2000-2015
  - NO_2 data were derived from a well-validated ensemble model that estimates daily NO_2 concentration for each 1-km grid in the US (R^2=0.79)
  - All models were adjusted for sociodemographic characteristics, time trends, and temperature
- **Simulation Study**: We simulated data based on the lag-response and seasonality/long-term trends from the real data (Fig 2). We varied the following parameters and simulated each scenario 500 times:
  1. Spatial resolution – High (1-km grid), Low (county-level), No resolution
  2. DLM estimation approach – Natural spline-based, Tree-based
  3. Time trend adjustment methods – No Adjustment, Long-term Spline, Year + Spline, Year + Harmonics, Year + Season

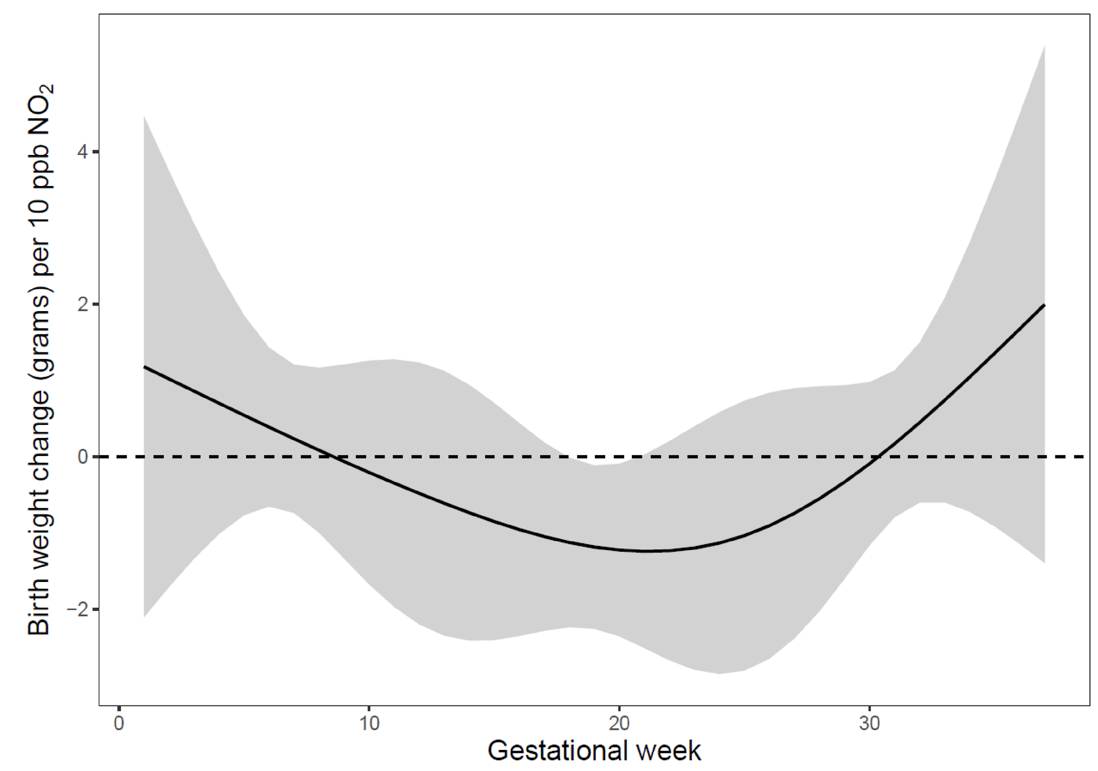

Fig 2. Lag-response for Massachusetts data. Negative association for NO_2 in weeks 15-30. Black line is lag-response, grey band is 95% confidence interval, dashed line is null

Results

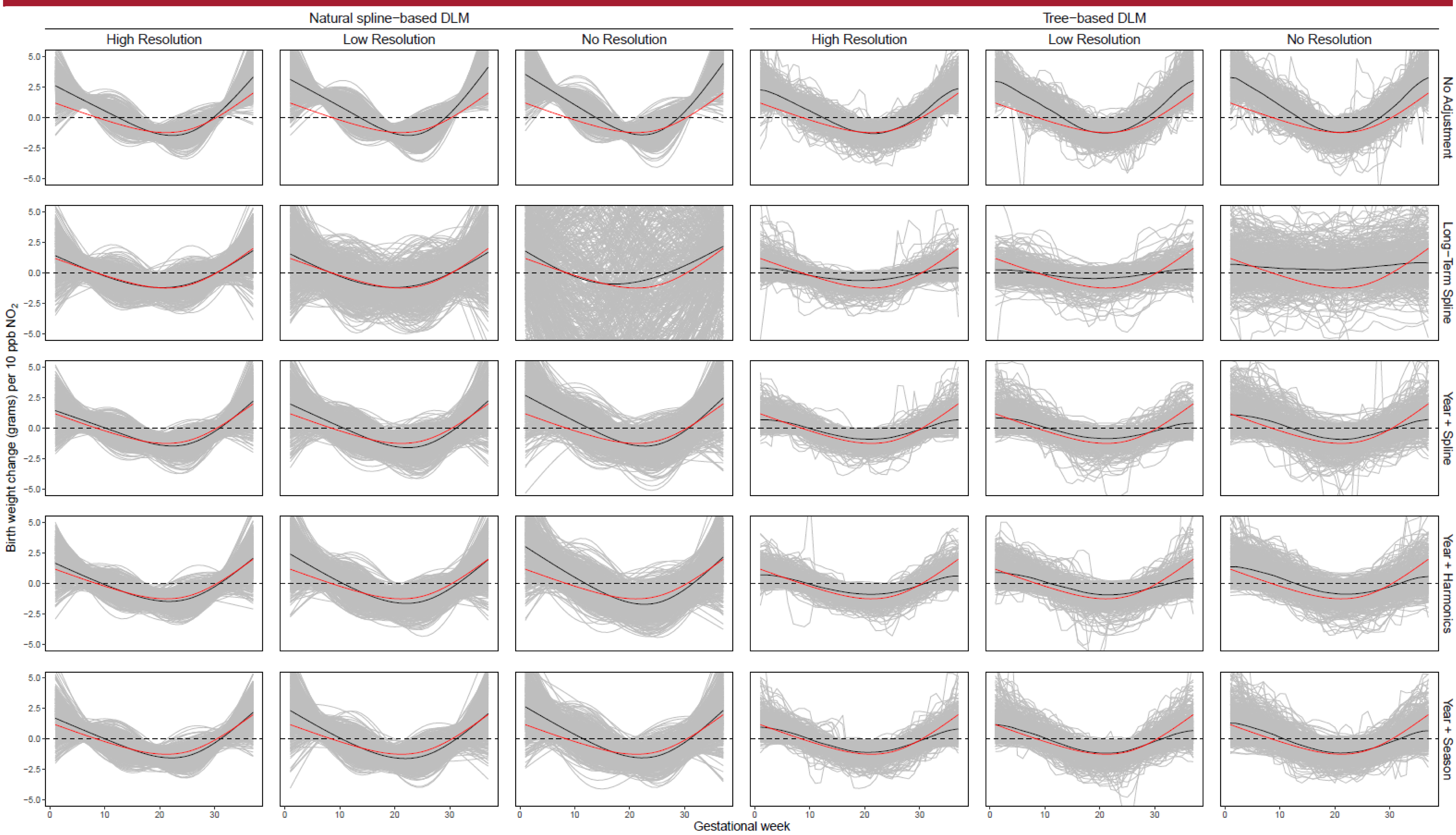

Fig 3. Simulation lag-responses for weekly NO_2 and birth weight estimated from natural spline-based and tree-based DLMs. Red line is true lag-response from Massachusetts data, grey line is lag-response from each iteration, black line is average lag-response, dashed line is null

Key findings:

- In a Massachusetts-based cohort, we found that associations were negative for NO_2 exposure experienced in weeks 15-30
- When using low- or no-spatial-resolution exposures, bias due to time trends was amplified
- Variance inflation was higher in low- or no-spatial-resolution DLMs when using a long-term spline to adjust for seasonality and long-term trends (concurvity)
- High-spatial-resolution data produced low bias and nominal coverage of the distributed lag estimator

Conclusion

- DLM analyses should jointly consider the spatial resolution of exposure data and the parameterization of the time trend adjustment and lag constraints
- If high-spatial-resolution exposure data are unavailable, then we recommend year + harmonics or year + season as they performed relatively well

Acknowledgements

Funding: NIEHS P30 ES009098, P30 ES000002, R01 ES030616, R01 ES029943, and R01 AG065276 and USEPA RD-835872-01

Contact: Michael Leung (mleung@hsph.harvard.edu); @epimikey