Towards Causal Inference for SpatioTemporal Data

Conflict and Forest Loss in Colombia

Presented by Sumantrak Mukherjee

Context for this work

The Colombian Conflict

- Armed conflict for 50 years
- More than 20,000 fatalities
- Peace agreement in 2016
- Increased forest loss attributed to armed conflict
- Pressure on forest reduced due to armed conflict preventing logging
- A data driven approach ->

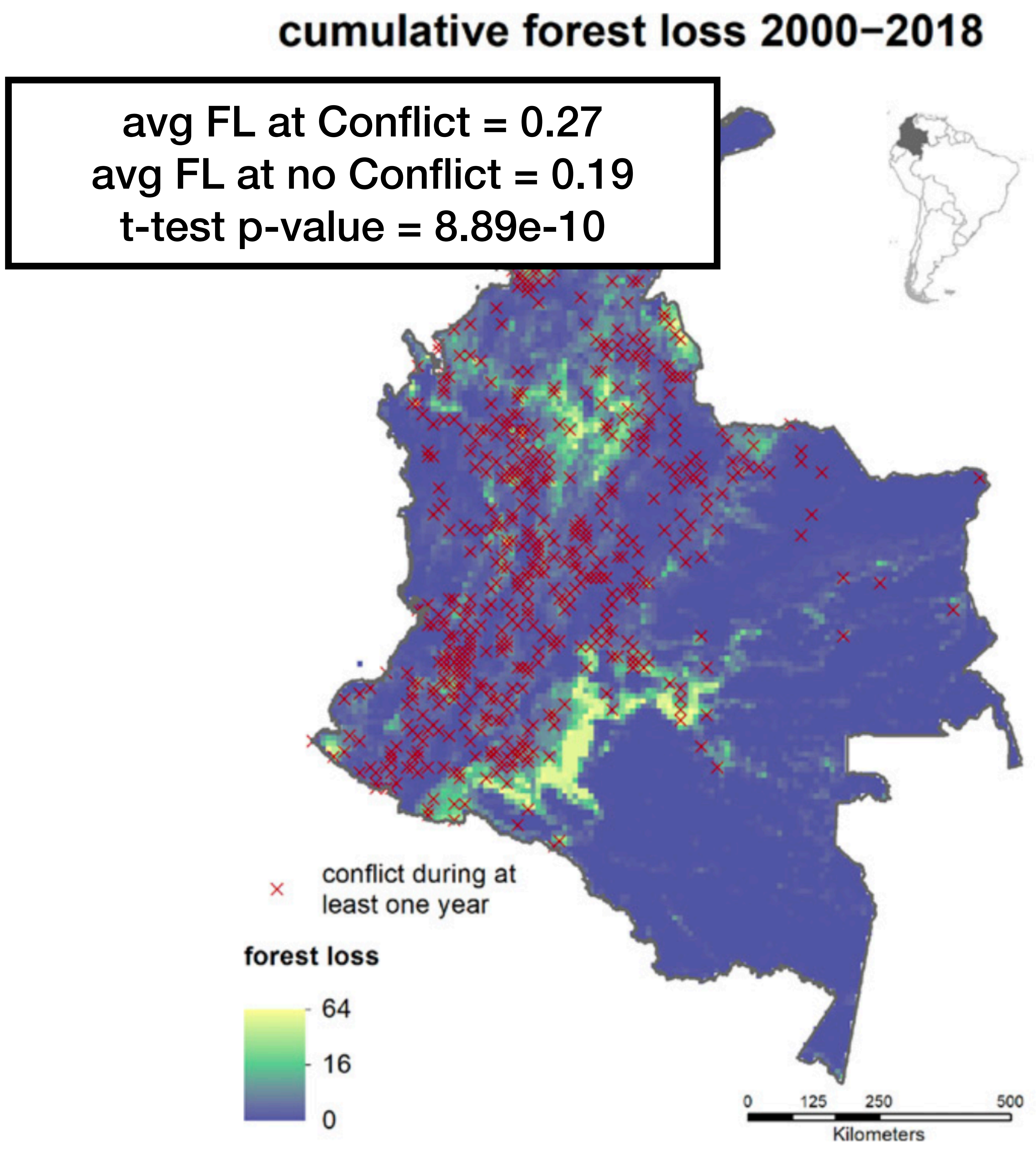
How does conflict affect forest loss?

Data Sources

- Forest loss dataset (2000-2018) (Source: Landsat Satellite Imagery, Resolution 30m x 30m), Complete canopy removal)
- Conflict event dataset (Source: Georeference Events Dataset (GED) from Uppsala Conflict Data Program (UCPD), Armed force by an organised actor resulting in at least one death)
- Accessibility proxy in terms of road distance (Source: https://diva-gis.org)
- Additional data sources for population, industrial presence etc (Source: unavialable)

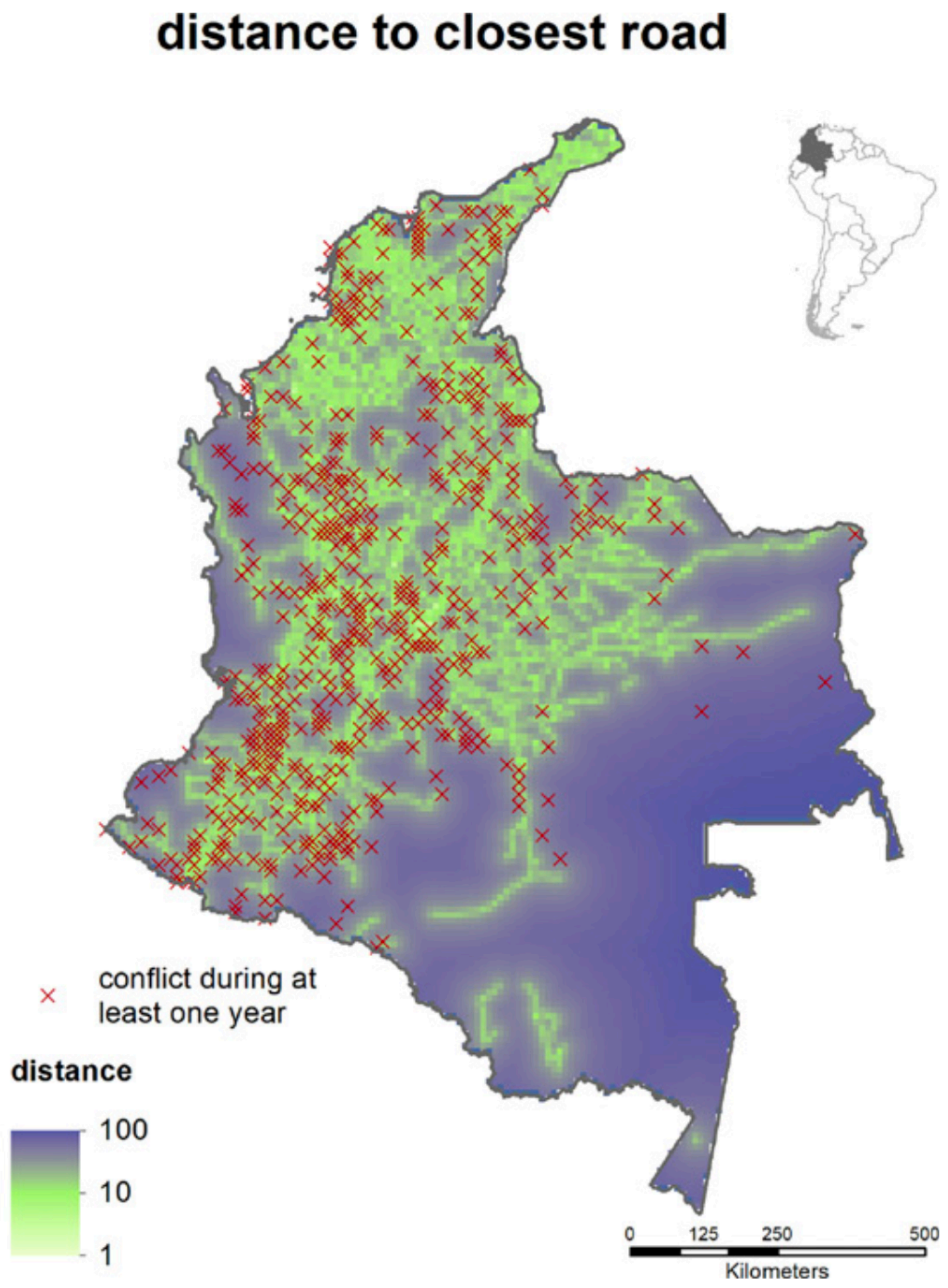
Summary Statistics

- Conflict events aggregated by Counting all events per year in each grids
- Forest loss aggregated by averaging annual loss (and spatial resolution adjusted to 10km x 10km)
- X_s^t : Binary conflict indicator at location s at time t
- Y_s^t : Absolute forest loss in location s from year $t-1$ to t
- Average forest loss in areas of conflict **exceeds** non conflict by almost **50%**



The Fault in our Analysis

- Both conflict and forest loss happen in areas of high accessibility
- W_s^t : Distance from s to the closest road in km
- Other confounders - (population density, market infrastructure etc…)
- Exaggerated significance of t-test due to strong spatial dependencies in X and Y



A method suitable for Spatiotemporal Data

Causal Model for Spatiotemporal Process

Preliminaries

- Every dataset is considered to be a realisation of a spatiotemporal process
- $\mathbf{Z}$ is a p-dimensional spatiotemporal process taking values in $\mathscr{Z}_P$
- Z_s^t marginalising $\mathbf{Z}$ at s and t
- $\mathbf{Z}_s : (Z_s^t)_{t\in\mathbb{N}}$ for time series *and* $\mathbf{Z}^t : (Z_s^t)_{s\in\mathbb{R}^2}$ for spatial process
- *Weakly stationary* if marginal distribution for Z_s^t is same for all s and t
- *Time Invariant* if $\mathbb{P}(\mathbf{Z}^1 = \mathbf{Z}^2 = \ldots) = 1$

Causal Model for Spatiotemporal Process

Causal Graphical Models for Spatiotemporal Processes

- Multivariate Spatiotemporal Processes for joint modelling of phenomenon for eg. Accessibility, population density, market infrastructure, forest loss, conflict
- Modelling causal relations among “disjoint bundles”
- $\mathbf{Z} = (\mathcal{S}, \mathcal{G}, \mathcal{P})$
- $\mathcal{S} = (S_j)_{j=1}^{k}$ for non empty disjoint sets $S_1, \ldots, S_k \subseteq \{1\ldots p\}$and $\cup_{j=1}^{k} S_j = \{1\ldots p\}$
- A Direct Acyclic Graph $\mathcal{G}$ with vertices $S_1, \ldots, S_k$
- Where $\mathcal{P} = \{\mathcal{P}^j\}_{j=1}^{k}$ is $\mathcal{P}^j = \{\mathbb{P}_z^j\}_{z \in Z_{|PA_j|}}$

Causal Model for Spatiotemporal Process

Observations and Interventions

- $\mathbb{P}(F) = \int_{F_1} \dots \int_{F_k} \mathbb{P}^k_{\mathscr{Z}^{PA_k}}(dz^{(S_k)}) \dots \mathbb{P}^1(dz^{(S_1)})$ is the observational distribution
- The conditional distribution of $\mathbf{Z}^{(S_j)}$ given $\mathbf{Z}^{(PA_j)}$ as induced by $\mathbb{P}$ is $\mathscr{P}_j$
- Can therefore be written as $[\mathbf{Z}^{(S_k)} \,|\, \mathbf{Z}^{(PA_k)}] \dots [\mathbf{Z}^{(S_1)}]$
- Intervention is defined as replacing $\mathscr{P}_j$ by $\tilde{\mathscr{P}}_j$
- The new graphical model is therefore $(\mathscr{S}, \mathscr{G}, \tilde{\mathscr{P}})$

Latent Spatial Confounder Model

Definition of an LSCM

- $(\mathbf{X}, \mathbf{Y}, \mathbf{H}) = (X_s^t, Y_s^t, H_s^t)_{(s,t)\in\mathbb{R}^2\times\mathbb{N}}$ where $X_s^t \in \mathbb{R}^d$, $H_s^t \in \mathbb{R}^l$ and real valued Y_s^t
- Causal Structure $[\mathbf{Y} \mid \mathbf{X}, \mathbf{H}][\mathbf{X} \mid \mathbf{H}][\mathbf{H}]$
- Assumptions
    - Latent process $\mathbf{H}$ is weakly stationary and time invariant
    - IID sequence $\epsilon^1, \epsilon^2, \ldots$ of weakly stationary spatial error processes and measurable function $f : \mathbb{R}^{d+l+1} \to \mathbb{R}$ st $Y_s^t = f(X_s^t, H_s^t, \epsilon_s^t)$

Latent Spatial Confounder Model

Average Treatment Effect and Causal Interpretation

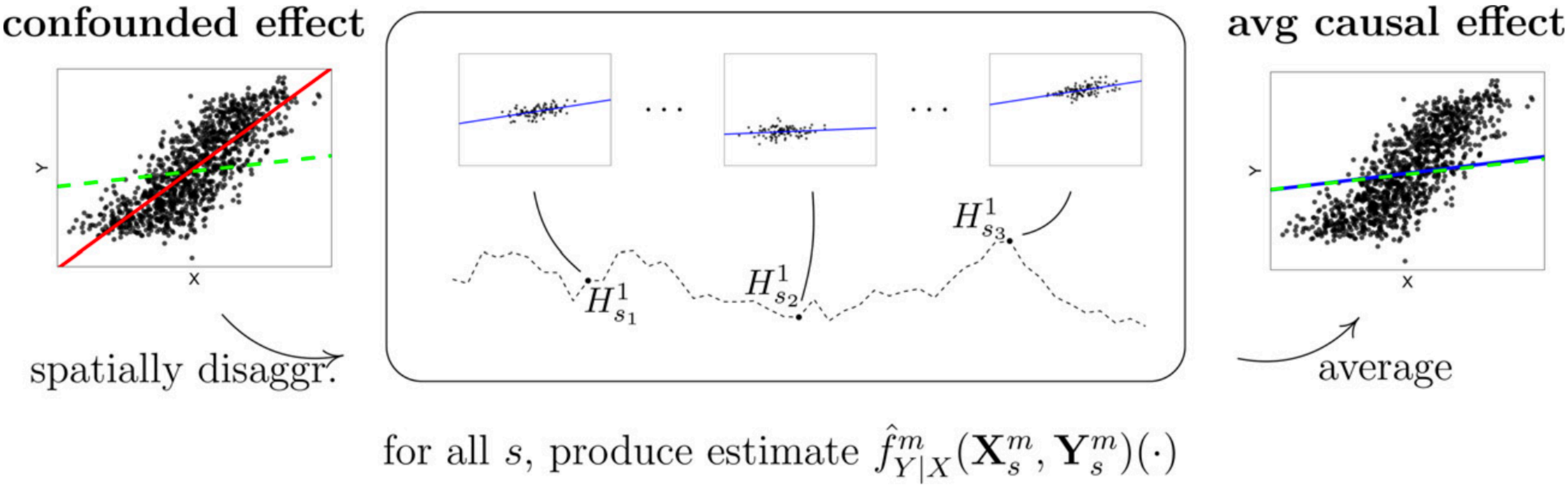
- Average effect $f_{AVE(X \to Y)}(x) := \mathbb{E}[f(x, H_0^1, \epsilon_0^1)]$ expectation over both noise and latent variables
- For fixed x and s, t if an intervention is applied s.t. $X_s^t = x$ holds almost surely in the interventional distribution $\mathbb{P}_x$ then $\mathbb{E}_{\mathbb{P}_X}[Y_s^t] = f_{AVE(X \to Y)}(x)$
- When graph is known we can compute interventional distribution from observational distribution
- $f_{AVE(X \to Y)}(x) := \mathbb{E}[f_{Y|(X,H)}(x, H_0^1)]$ where $f_{Y|(X,H)}$ is the regression function $(x, h) \to \mathbb{E}[Y_s^t | X_s^t = x, H_s^t = h]$

Estimating Average Causal Effect

- We have dataset $(\mathbf{X}_n^m, \mathbf{Y}_n^m) = (X_s^t, Y_s^t)_{(s,t)\in\{s_1,\ldots,s_n\}\times\{1,\ldots,m\}}$
- For every $s \in \{s_1, \ldots, s_n\}$ several time instances with $t \in \{1,\ldots,m\}$with the same conditional $Y_s^t \,|\, (X_s^t, H_s^t)$
- The latent realisation h_s of H_s^1 is not observed but since $\mathbf{H}$ is static for every s we can estimate $f_{Y|(X,H)}(\,\cdot\,, h_s)$
- We need to specify a model class for $f_{Y|(X,H)}(\,\cdot\,, h)$ and a suitable estimator $\hat{f}_{Y|X} = (\hat{f}^m_{Y|X})_{m\in\mathbb{N}}$
- $$\hat{f}^{nm}_{AVE(X\to Y)}(\mathbf{X}_n^m, \mathbf{Y}_n^m)(x) := \frac{1}{n}\sum_{i=1}^{n} \hat{f}^m_{Y|X}(\mathbf{X}_{s_i}^m, \mathbf{Y}_{s_i}^m)(x)$$

Estimating Average Causal Effect from Data

Averaging localised models



Estimating Average Causal Effect from Data

Assumptions

- Law of Large Numbers for LSCM : With increasing number of spatial locations $\hat{f}^{nm}_{AVE(X\to Y)}(\mathbf{X}^m_n, \mathbf{Y}^m_n)(x) := \frac{1}{n}\sum_{i=1}^{n}\hat{f}^m_{Y|X}(\mathbf{X}^m_{s_i}, \mathbf{Y}^m_{s_i})(x)$ approximates $f_{AVE(X\to Y)}(x) := \mathbb{E}[f_{Y|(X,H)}(x, H^1_0)]$ for a stationary Gaussian process $\mathbf{H}^1$ sampled regularly in space

- Consistent estimators of the conditional expectations $\hat{f}^m_{Y|X}(\mathbf{X}^m_s, \mathbf{Y}^m_s)(x) - f_{Y|(X,H)}(x, H^1_s) \to 0$ as $m \to \infty$

- there exists $N \in \mathbb{N}$ s.t. for all $n \geq N$ we can find $M_n \in \mathbb{N}$ s.t. for all $m \geq M_n$ $\mathbb{P}\Big(\mid \hat{f}^{nm}_{AVE(X\to Y)}(\mathbf{X}^m_n, \mathbf{Y}^m_n)(x) - f_{AVE(X\to Y)}(x) \mid > \delta\Big) \leq \alpha$

Asymptotic consistency: An Example LSCM

$\boldsymbol{\zeta}, \boldsymbol{\psi}, \boldsymbol{\xi}^t, \boldsymbol{\varepsilon}^t,\ t \in \mathbb{N}$

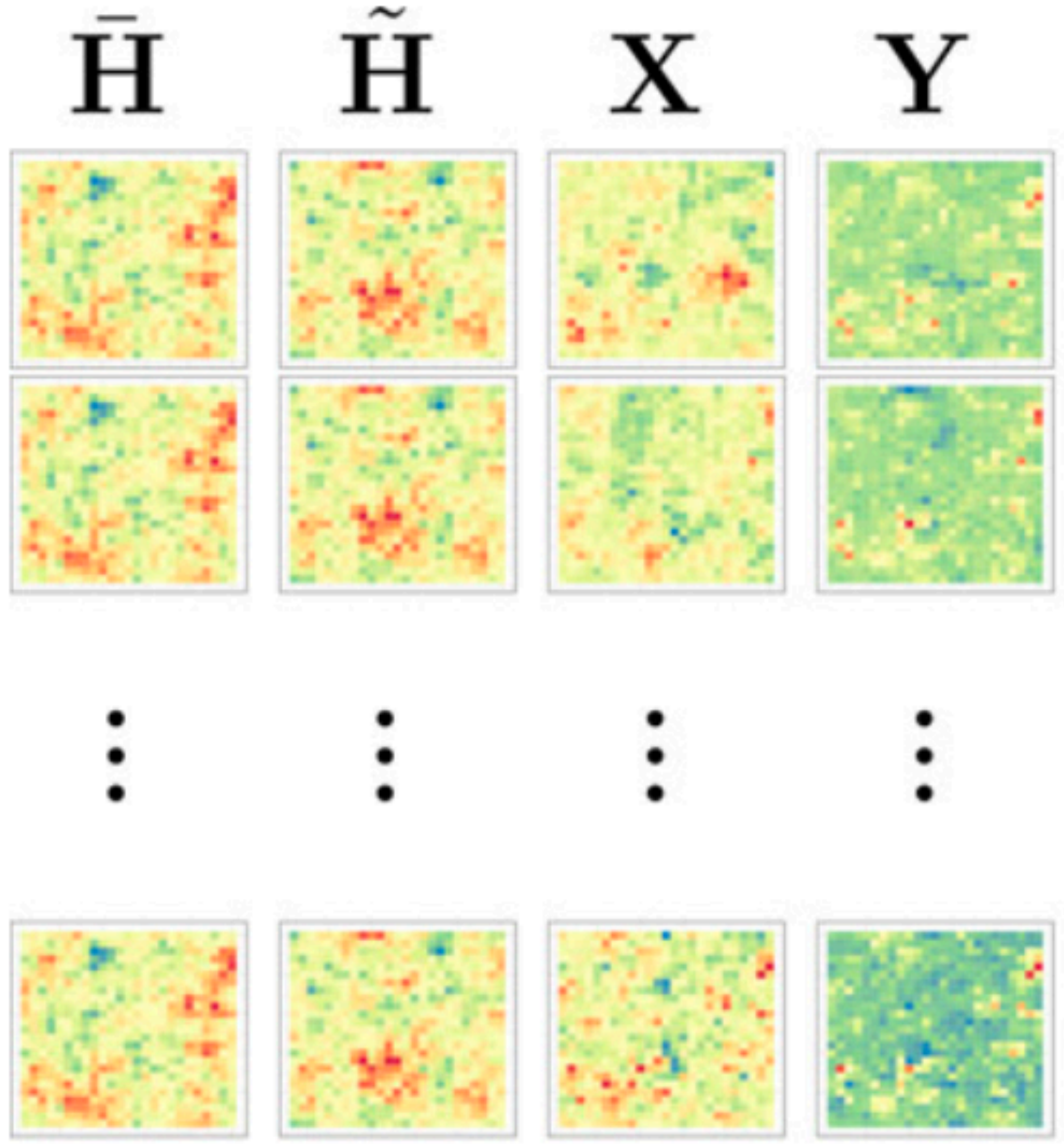
Are independent versions of univariate Gaussian Processes with mean 0 and covariance $u \to \exp(-\frac{1}{2}||u||_2)$

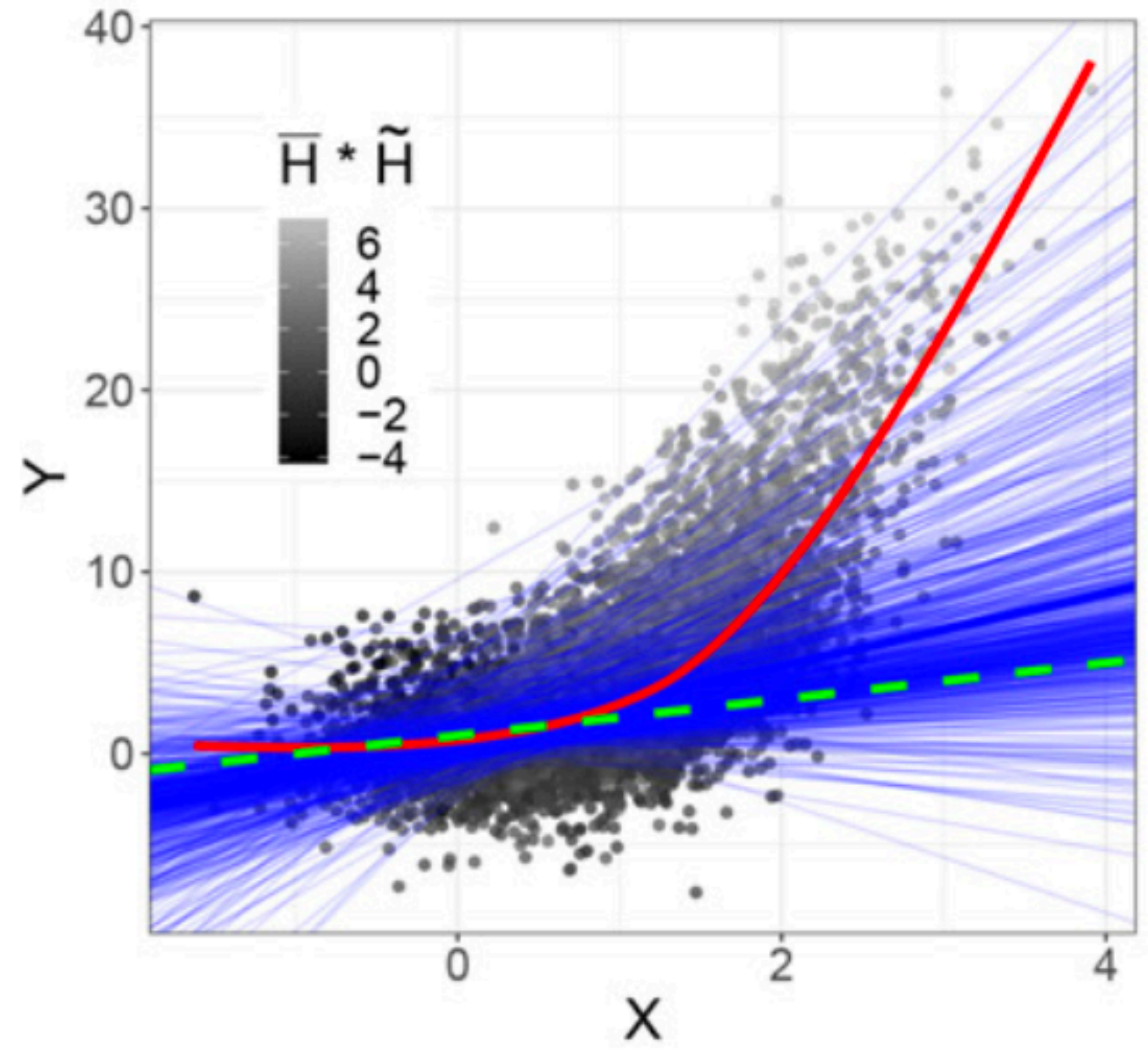
$$H_s^t = (\bar{H}_s^t, \tilde{H}_s^t) = (\zeta_s, 1 + \tfrac{1}{2}\zeta_s + \tfrac{\sqrt{3}}{2}\psi_s),$$

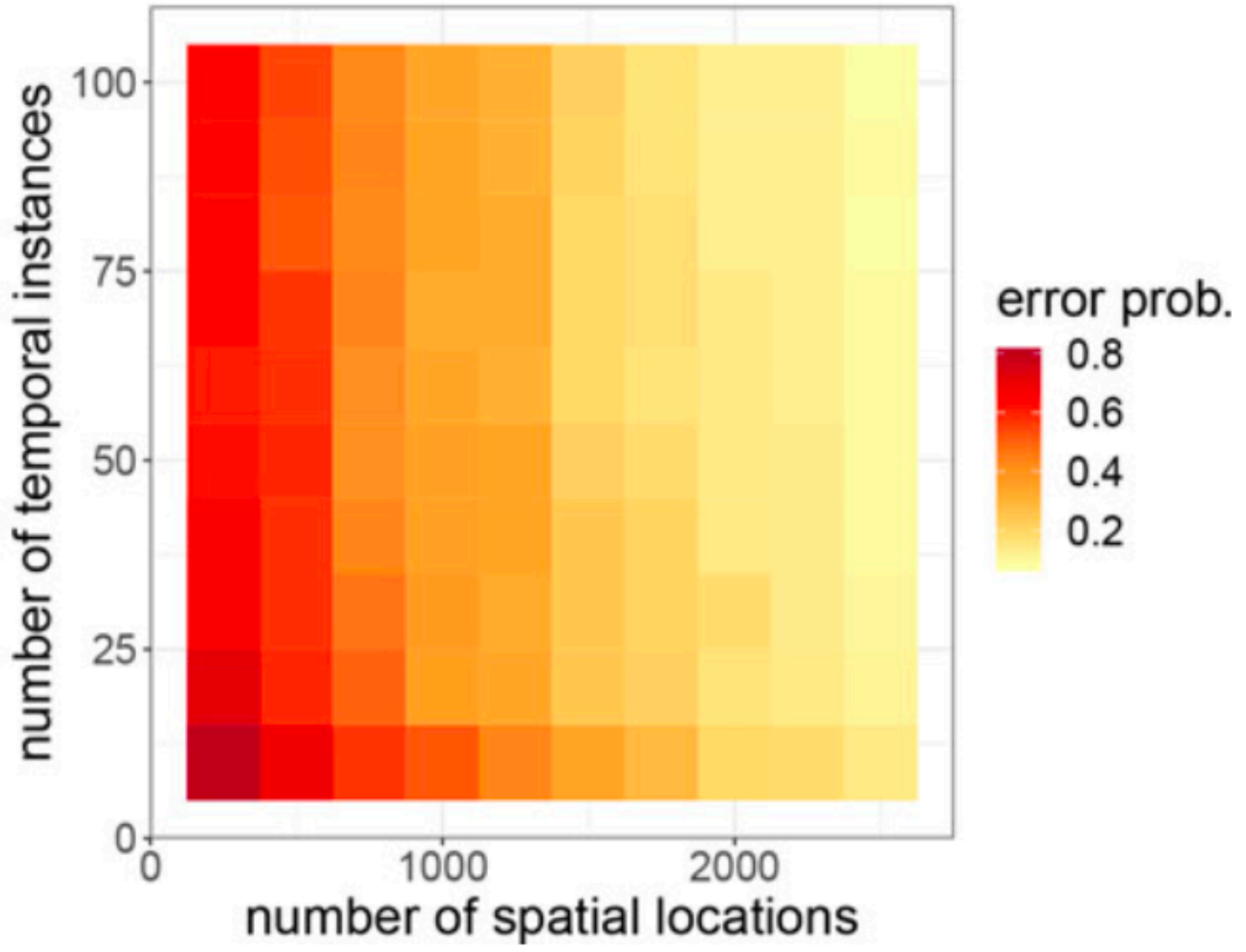
$$X_s^t = \exp(-||s||_2^2/1000) + (0.2 + 0.1 \cdot \sin(2\pi t/100))$$
$$\cdot\ \bar{H}_s^t \cdot \tilde{H}_s^t + 0.5 \cdot \xi_s^t,$$

$$Y_s^t = (1.5 + \bar{H}_s^t \cdot \tilde{H}_s^t) \cdot X_s^t + (\bar{H}_s^t)^2 + |\tilde{H}_s^t| \cdot \varepsilon_s^t.$$

$$Y_s^t = f_1(H_s^t) \cdot X_s^t + f_2(H_s^t, \varepsilon_s^t)$$







Testing for Existence of Causal Effects

Exchangeability and Resampling

- $H_0 : (\mathbf{X}, \mathbf{Y})$ is generated from an LSCM with a function f constant wrt X_s^t
- Formalising "No causal effect of X on Y" within LSCMs
- For a resampling test we use a permutation scheme s.t. for the null hypothesis the distribution of data remains unaffected
- for every $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{(d+1)\times n \times m}$ and permutation σ of elements $\{1, \ldots, m\}$
  $\sigma(\mathbf{x}, \mathbf{y})$ is the permuted array with entries $(\sigma(x, y))_s^t = (x_s^t, y_s^{\sigma(t)})$
- Exchangeability is under H_0 $\sigma(\mathbf{X}_n^m, \mathbf{Y}_n^m)$ has the same distribution as $\mathbf{X}_n^m, \mathbf{Y}_n^m$

How to perform a test

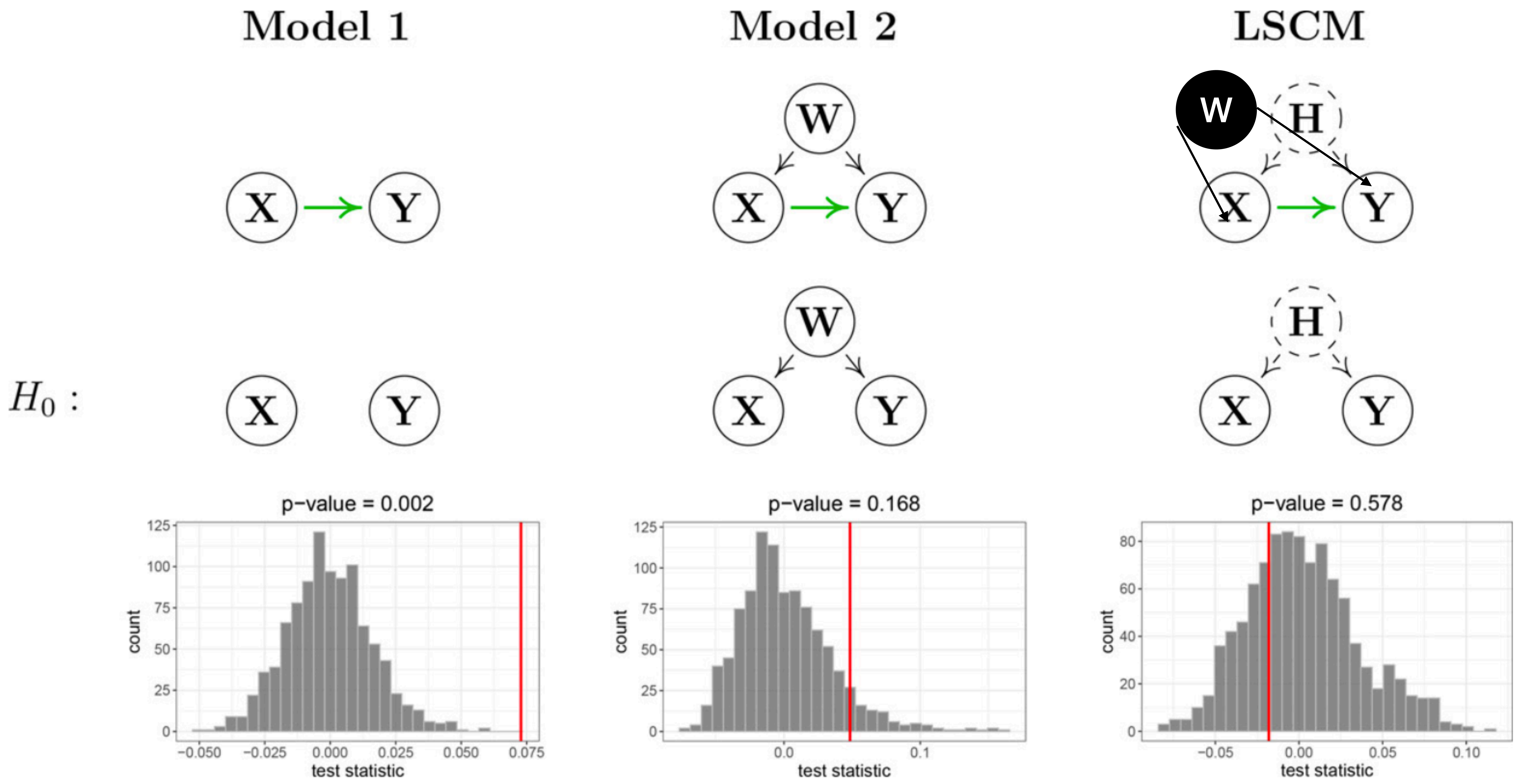
- For $B \in \mathbb{N}$ uniform draws $k_1, \ldots, k_b$ from $\{1,\ldots,M\}$ where $M := m!$
- $p_{\hat{T}}(\mathbf{x}, \mathbf{y}) := \dfrac{1 + |b \in \{1,\ldots,B\} : \hat{T}(\sigma_{k_b}(\mathbf{x}, \mathbf{y})) \geq \hat{T}(\mathbf{x}, \mathbf{y})|}{1 + B}$
- $\phi^{\alpha}_{\hat{T}} = 1 \Leftrightarrow p_{\hat{T}} \leq \alpha$
- $\hat{T}(\mathbf{X}^m_n, \mathbf{Y}^m_n) = \psi(\hat{f}^{nm}_{AVE}(\mathbf{X}^m_n, \mathbf{Y}^m_n))$
- What is a block permutation scheme and why we need it

Applying the model to Data

- $T := f_{AVE(X \to Y)}(1) - f_{AVE(X \to Y)}(0)$ difference in forest loss intervening on conflict
- How to test $H_0 : T = 0$
- Omit all locations which do not have data for both conflict and no conflict
- $\hat{f}^{nm}_{AVE(X \to Y)}(\mathbf{X}^m_n, \mathbf{Y}^m_n)(x) = \frac{1}{|\mathcal{I}^m_n|} \sum_{i \in \mathcal{I}^m_n} \frac{1}{|\{t : X^t_{s_i} = x\}|} \sum_{t : X^t_{s_i} = x} Y^t_{s_i}$

Alternative assumptions on Causal Structure

Quantifying Causal Influence of Conflict on Forest Loss

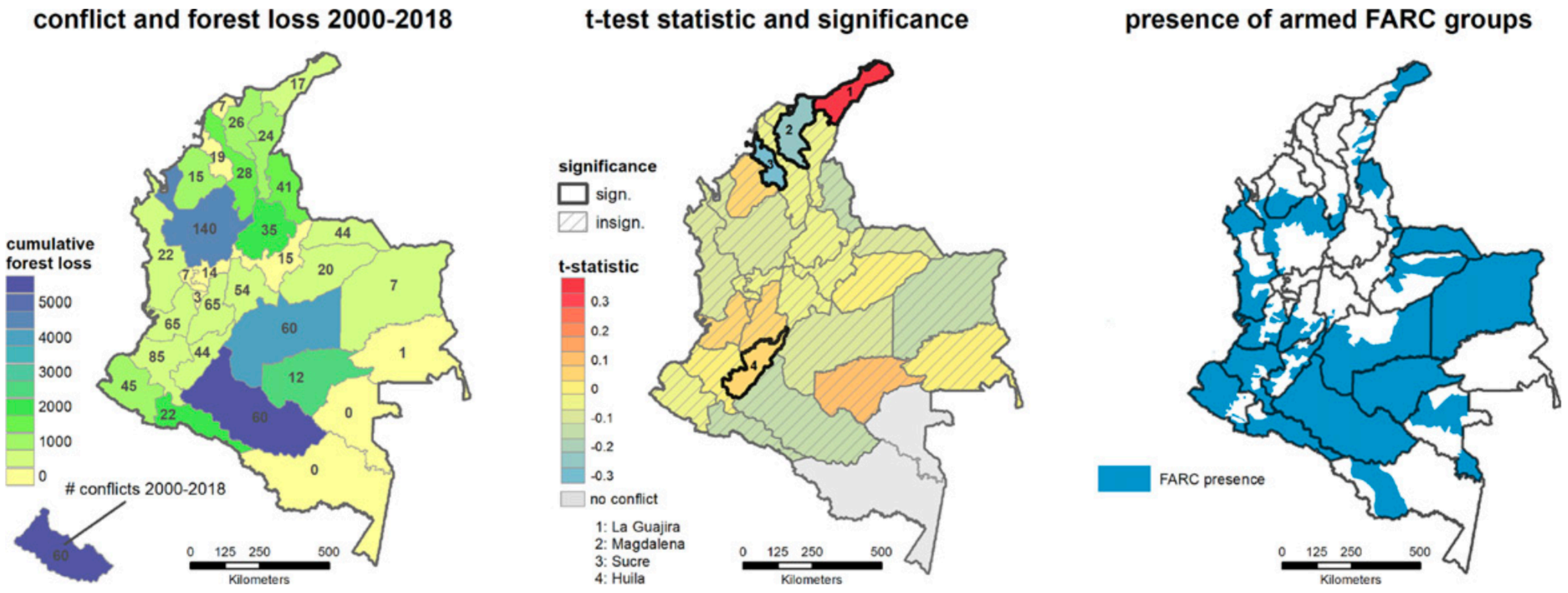


Causal Effects based on Different Models

- Baseline Model: Conflict has a significant positive effect on forest loss ($\hat{T}$ = 0.073, P = 0.002)
- Adjusting for Confounders:
    - Accessibility (**$\hat{T}$ = 0.049, P = 0.168**) and Population Density (**$\hat{T}$ = 0.038, P = 0.214**) reduce effect size and remove significance.
    - Accounting for all time-invariant confounders reverses the sign ($\hat{T}$ = -0.018, P = 0.578) but remains insignificant.
- Spatial and Temporal Adjustments
    - **No evidence** for spatial spill-over effects
    - No significant effect when accounting for time delay ($\hat{T}$ = -0.0293, P = 0.354).
    - Block-permutation tests confirm non-significance.

Regional Analysis of Conflict

Departmental Policies Matter



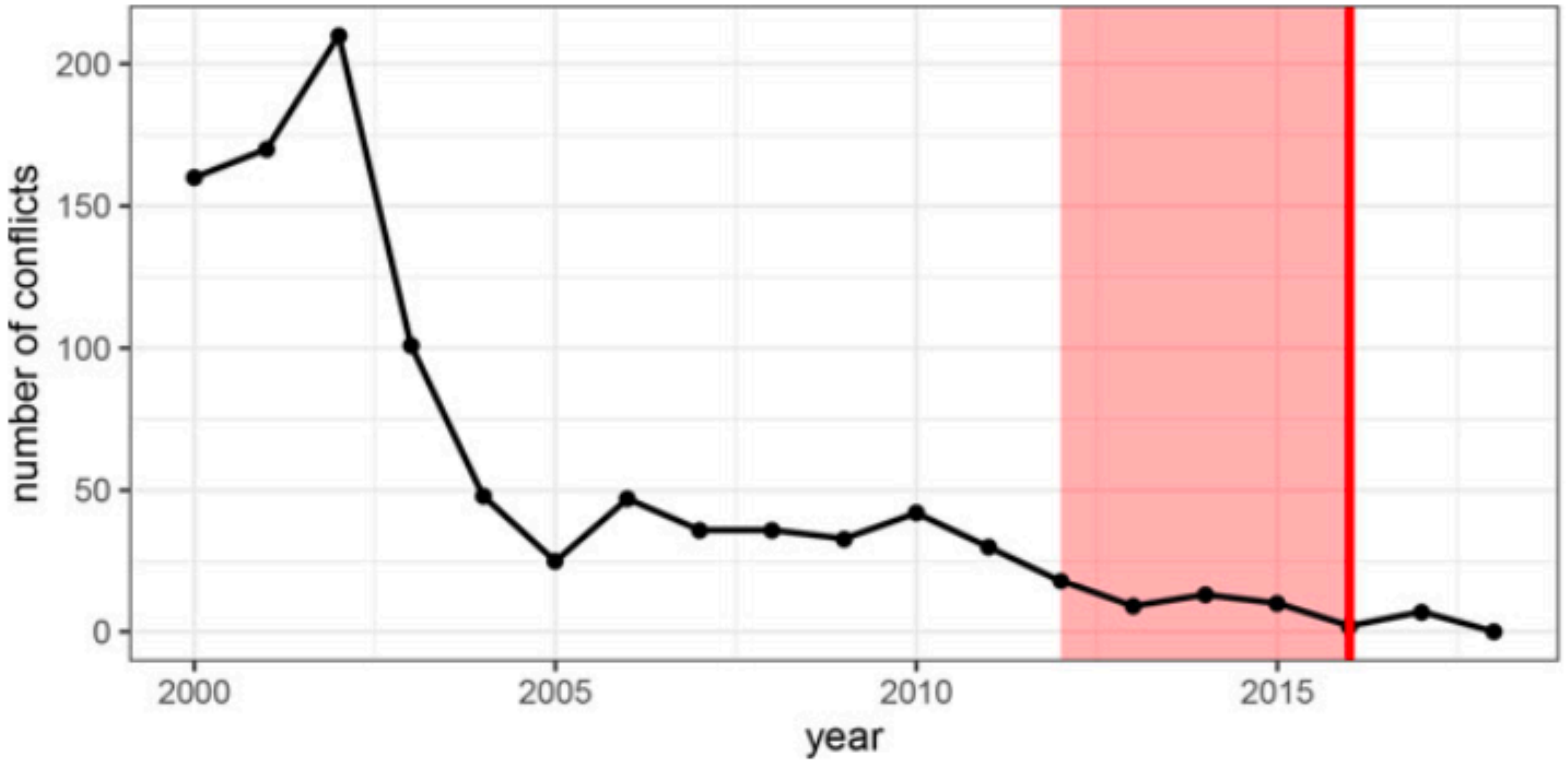
Adjusting for Spatial Heterogeneity

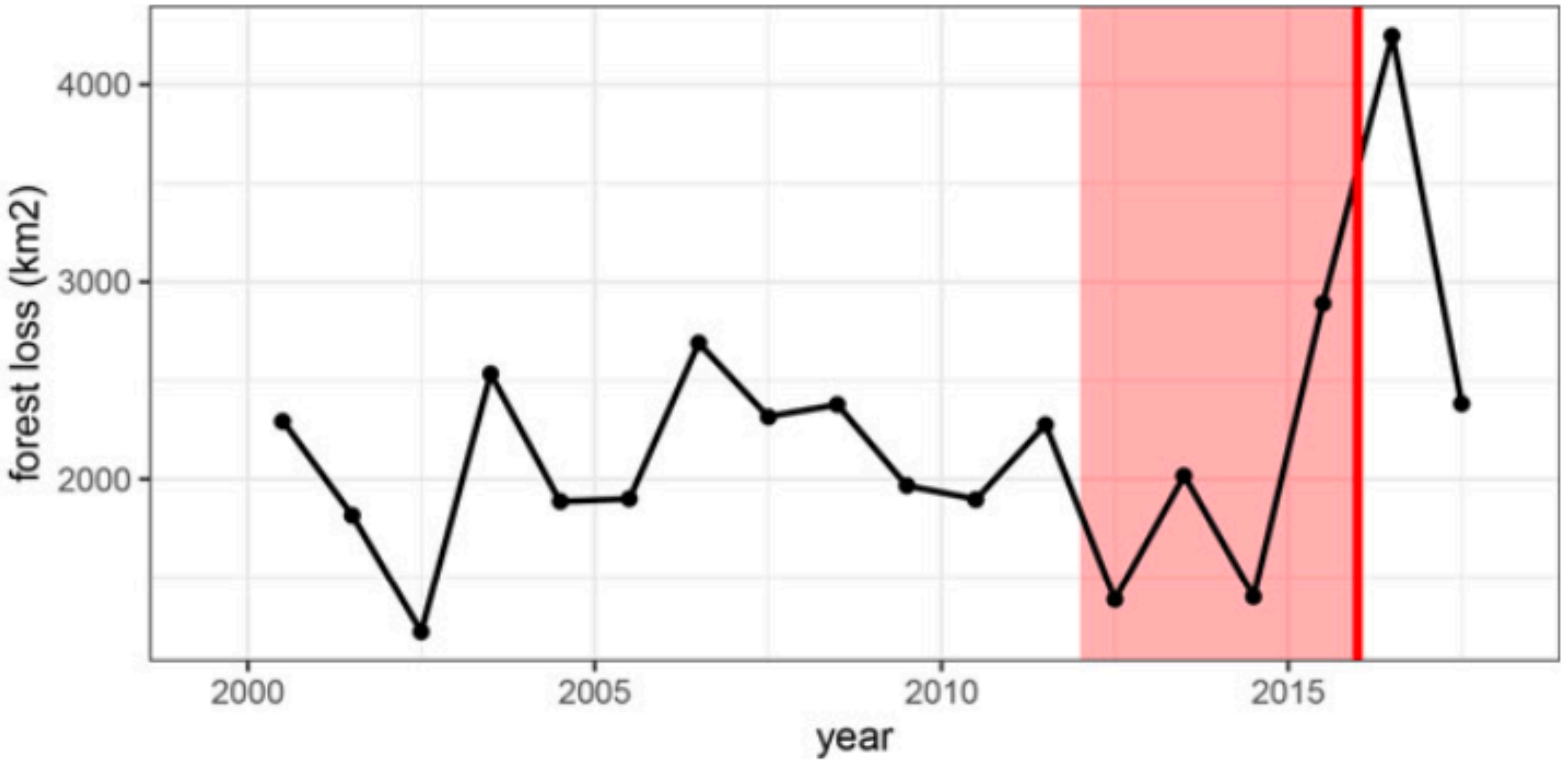
Policy changes across borders result in heterogeneity

- **High heterogeneity** across departments.
- La Guajira ($\hat{T}$ = 0.398, P = 0.047): Strongest positive effect on deforestation.
- Magdalena ($\hat{T}$ = -0.218, P = 0.004): Significant negative effect.
- Huila ($\hat{T}$ = 0.095, P = 0.023): Moderate positive effect.
- FARC-Controlled Areas:
    - 6 out of 8 regions show negative effects of conflict on deforestation.
    - Explanation for positive effect: Forest cover was a **strategic resource** for internal governance and drug production.

Verifying Intervention Effects

Government Interventions leads to resolution of local tensions





Recap

- Causal vs Predictive Analysis
- New causal framework particularly designed for multivariate spate temporal processes
- Estimation and testing of causal effects (non parametric hypothesis test for causal relationships, asymptotic consistency proven through simulations)
- Empirical findings (no country wide effect, regional variability, sociopolitical influence)
- Interpretations of the method (finding align with post conflict deforestation surge, potential bias from time invariant confounders, relative role of conflict)

Extension and Future Directions

Points for discussion

- Combining time variant confounders with unobserved confounders
- Temporally lagged causal effects?
- Space time interchangeability and analysis based on that
- Slowly varying unobserved confounders
- Smoothness Assumptions in space might help with hypothesis testing
- Datasets and other applications where such an analysis might be useful
- Counterfactuals based on the method proposed

Thanks for participating
Volunteers for presenting on 2nd April?