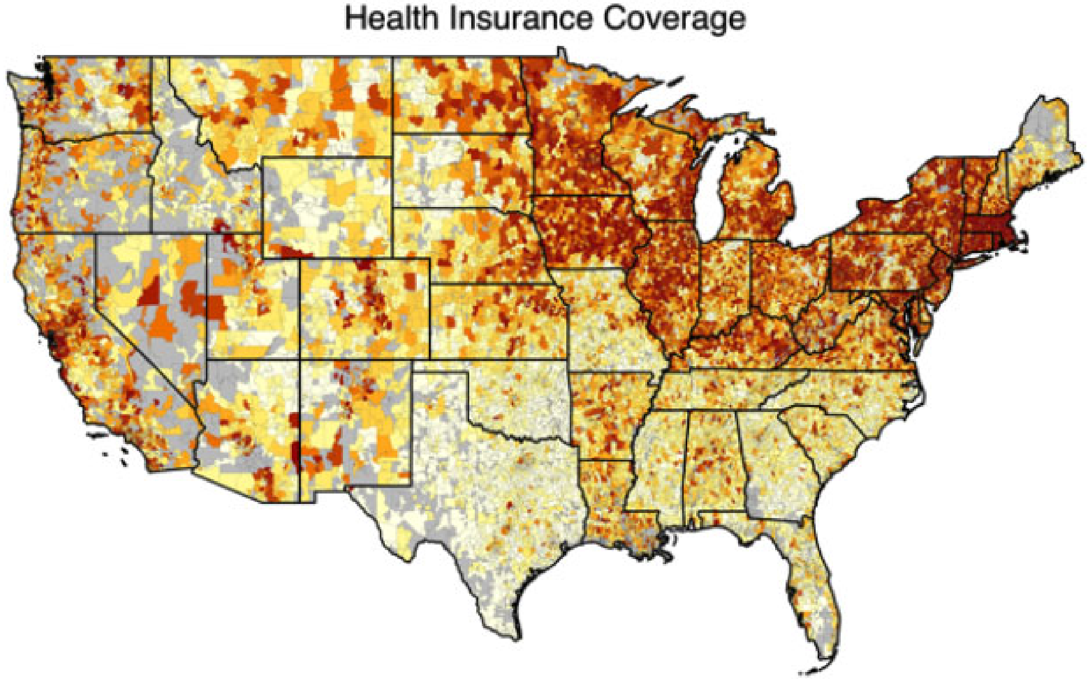
Testing Coefficient Variability in Spatial Regressions

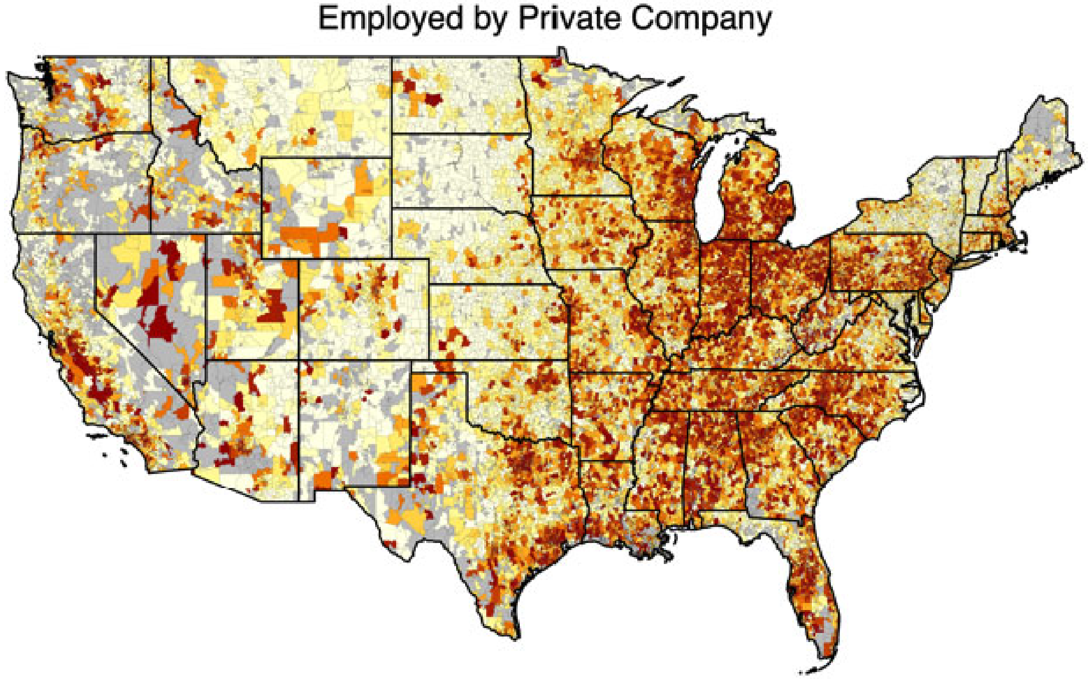
Ulrich K. Müller and Mark W. Watson
Princeton University

Munich Econometrics Workshop 2024

Motivating Example: A Bivariate Spatial Regression

Data in Levels

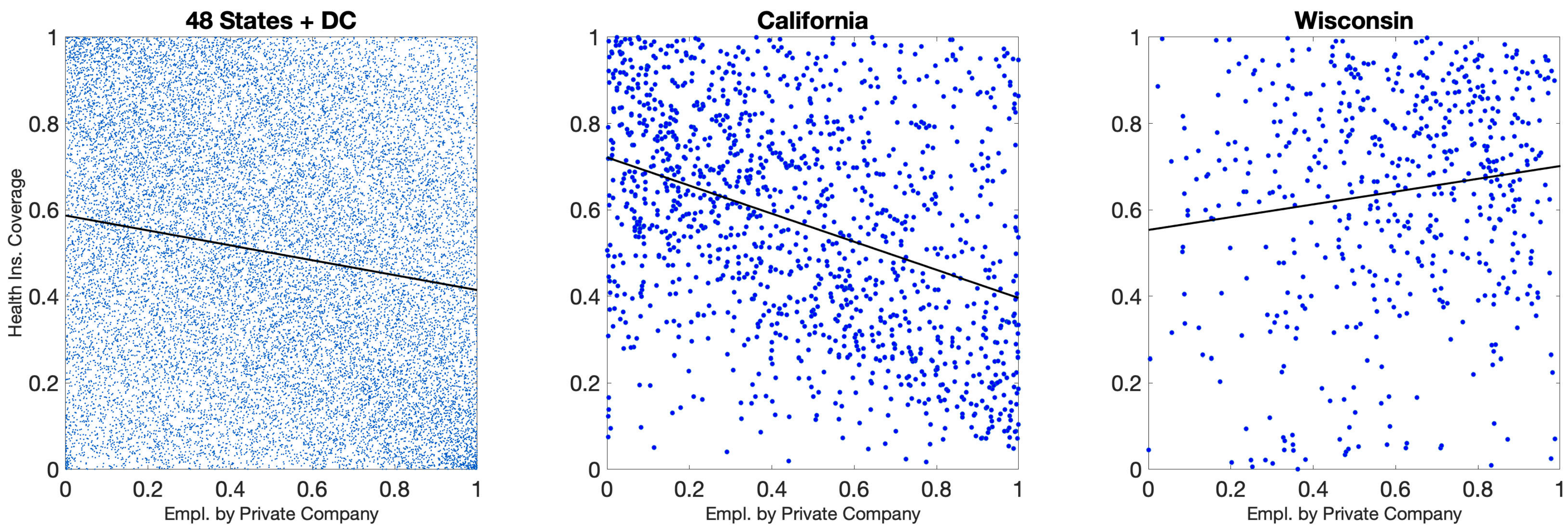




Variables are measured in percentiles across the 21k zip codes

Motivating Example (ctd)

Scatter Plots and OLS estimates: Levels

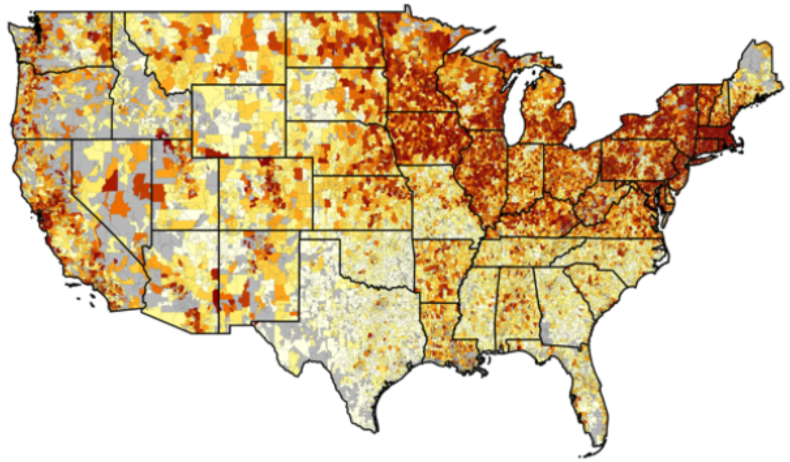


$\hat{\beta}$ = -0.17 $\quad$ $\hat{\beta}$ = -0.33 $\quad$ $\hat{\beta}$ = 0.15

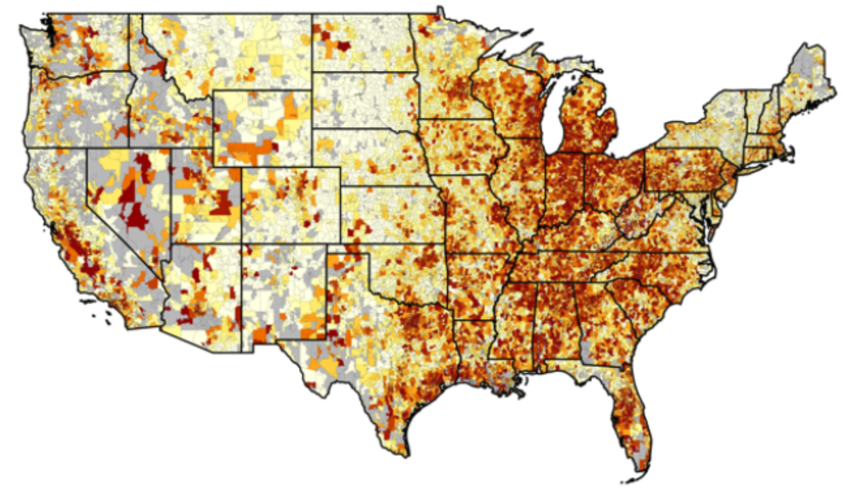
GLS Spatial Difference to Avoid Spurious Regression

Data by zip code: Levels

Health Insurance Coverage

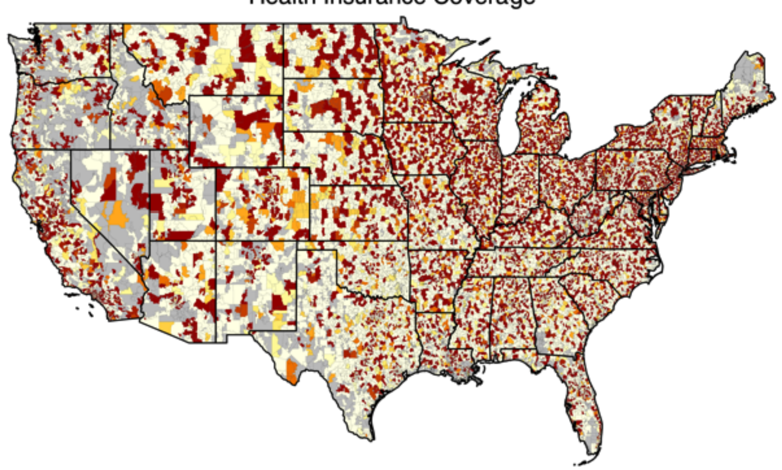


Employed by Private Company

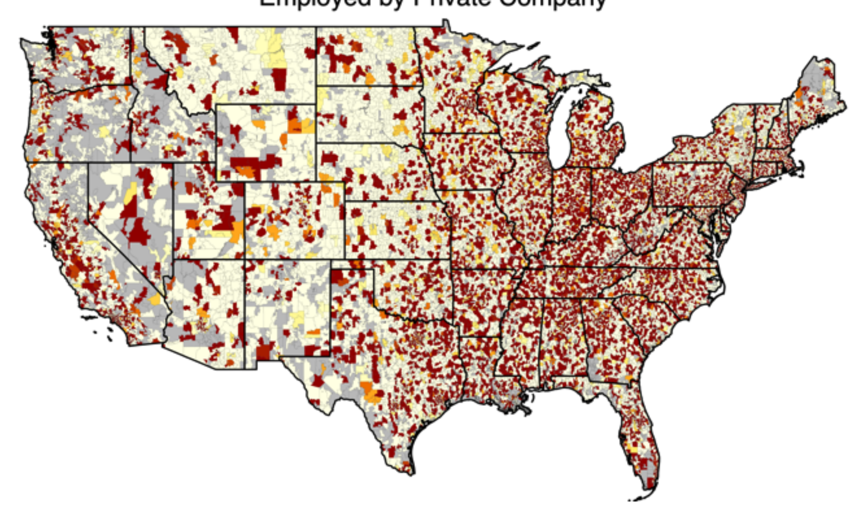


Data by zip code: GLS transformed

Health Insurance Coverage

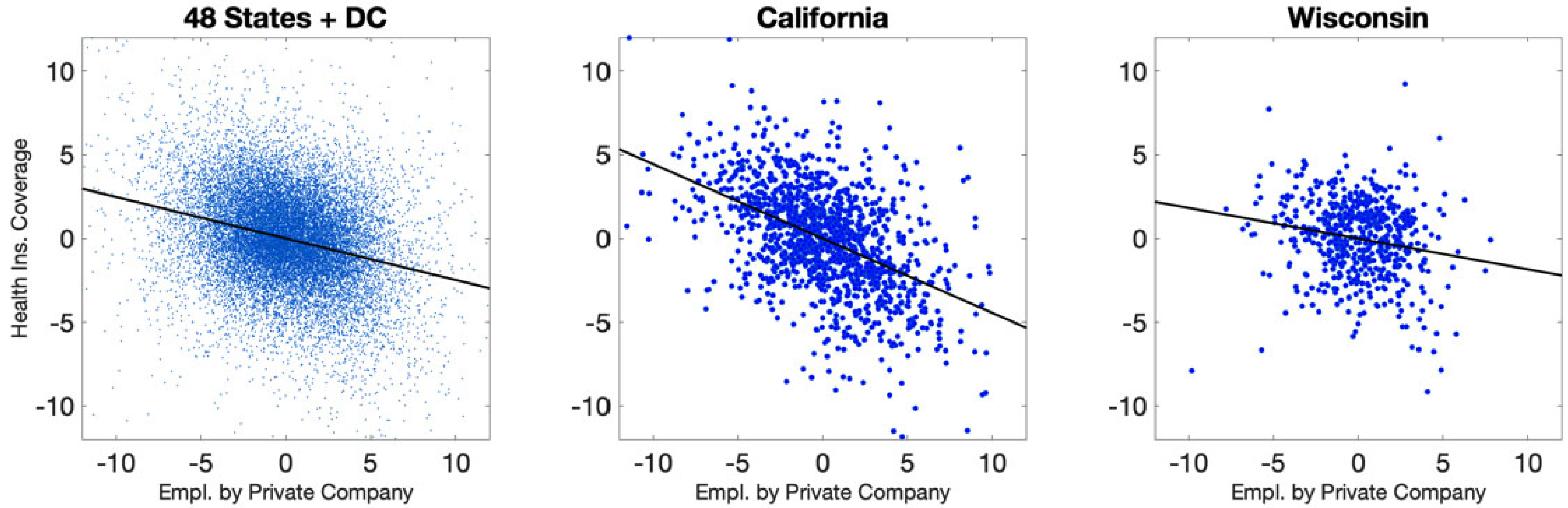


Employed by Private Company



Parameter Stability?

Scatter plots and OLS estimates: GLS transformed data



$\hat{\beta}$ = -0.25 (0.02) $\hat{\beta}$ = -0.44 (0.03) $\hat{\beta}$ = -0.18 (0.07)

Familiar issues: Why these two states? Why states in the first place?

$\Rightarrow$ How to test generic null of parameter constancy in spatial regression?

Literature

Time series regression stability tests

- Discrete breaks: Chow (1960), Quandt (1960), ...
- Validity under general conditions: Andrews (1993), ...
- Martingale Variation: Nyblom (1989), Elliott and Müller (2006), ...
- Time variation in second moments: Hansen (2000), ...

Spatial regression stability tests

- Chow test (with autoregressive spatial errors): Anselin (1990)
- Local spatial regressions: Fotheringham et al. (2002, 2024) (inference assumes i.i.d. errors)

This Paper

- Nyblom (1989)-like test for spatial variation in regression coefficients

  $\Rightarrow$ Locally best test against Lévy-Brownian motion variation in canonical model

- Valid under general conditions

  - Allows for (weakly) spatially correlated, non-Gaussian errors (analogous to Andrews (1993))

  - Accommodates spatially varying second moments (analogous to Hansen (2000))

Outline

1. Canonical Gaussian model and locally best test

2. Validity under general conditions

3. Monte Carlo results

4. Application to 1514 bivariate zip-code level regressions using American Community Survey (ACS) data

Canonical Model

$$\begin{aligned} y_l &= x_l\beta_l + \ldots + \ u_l, \quad l = 1, \ldots, n \\ &= x_l\beta + e_l \quad \text{with} \quad e_l = u_l + x_l(\beta_l - \beta) \end{aligned}$$

- $(y_l, x_l) \in \mathbb{R}^2$ associated with observed location $s_l \in \mathcal{S} \subset \mathbb{R}^d$, $d \geq 1$
- $u_l \sim iid\mathcal{N}(0, 1)$, x_l nonstochastic
- Hypotheses of interest

$$H_0 : \beta_l = \beta \quad \text{vs} \quad H_a : \beta_l \neq \beta_\ell \text{ for some } 1 \leq l, \ell \leq n$$

- Impose invariance $Y \to Y + Xb$

  $\Rightarrow$ Test is a function of OLS residuals $\hat{e}_l$

- Best invariant test against $\{\beta_l\}_{l=1}^n = \{\beta_l^1\}_{l=1}^n$ rejects when $\sum_{l=1}^n \beta_l^1 x_l \hat{e}_l$ is large $\Rightarrow$ no UMP test

Locally Best Test

- Maximize weighted average power

  $\Rightarrow$ Same as maximizing power against β_l stochastic with p.d.f. equal to weighting function

  $\Rightarrow$ We use

$$H_a^* : \beta_l - \beta = \kappa L(s_l),\ l = 1, \ldots, n$$

  where $L(\cdot)$ is Lévy-Brownian motion (LBM), i.e. $\mathbb{E}[L(s)L(r)] = \frac{1}{2}(||s|| + ||r|| - ||s - r||)$

- Locally best invariant test of $\kappa = 0$ against $\kappa > 0$ rejects for large values of

$$\xi^* = n^{-1}\hat{e}' D_x \bar{\mathbf{\Sigma}}_L D_x \hat{e}$$

  where $D_x = \mathrm{diag}(x_1, \ldots, x_n)$ and $\bar{\mathbf{\Sigma}}_L$ is covariance matrix of (demeaned) LBM evaluated at $s_1, \ldots, s_n$

Sample Realizations of LBM for $d = 2$

Martingale-like variation in space: $L(a + bs) - L(a) \sim W(s)$ for $s \in \mathbb{R}$ and $a, b \in \mathbb{R}^d$ with $||b|| = 1$

Rewriting the Locally Best Test

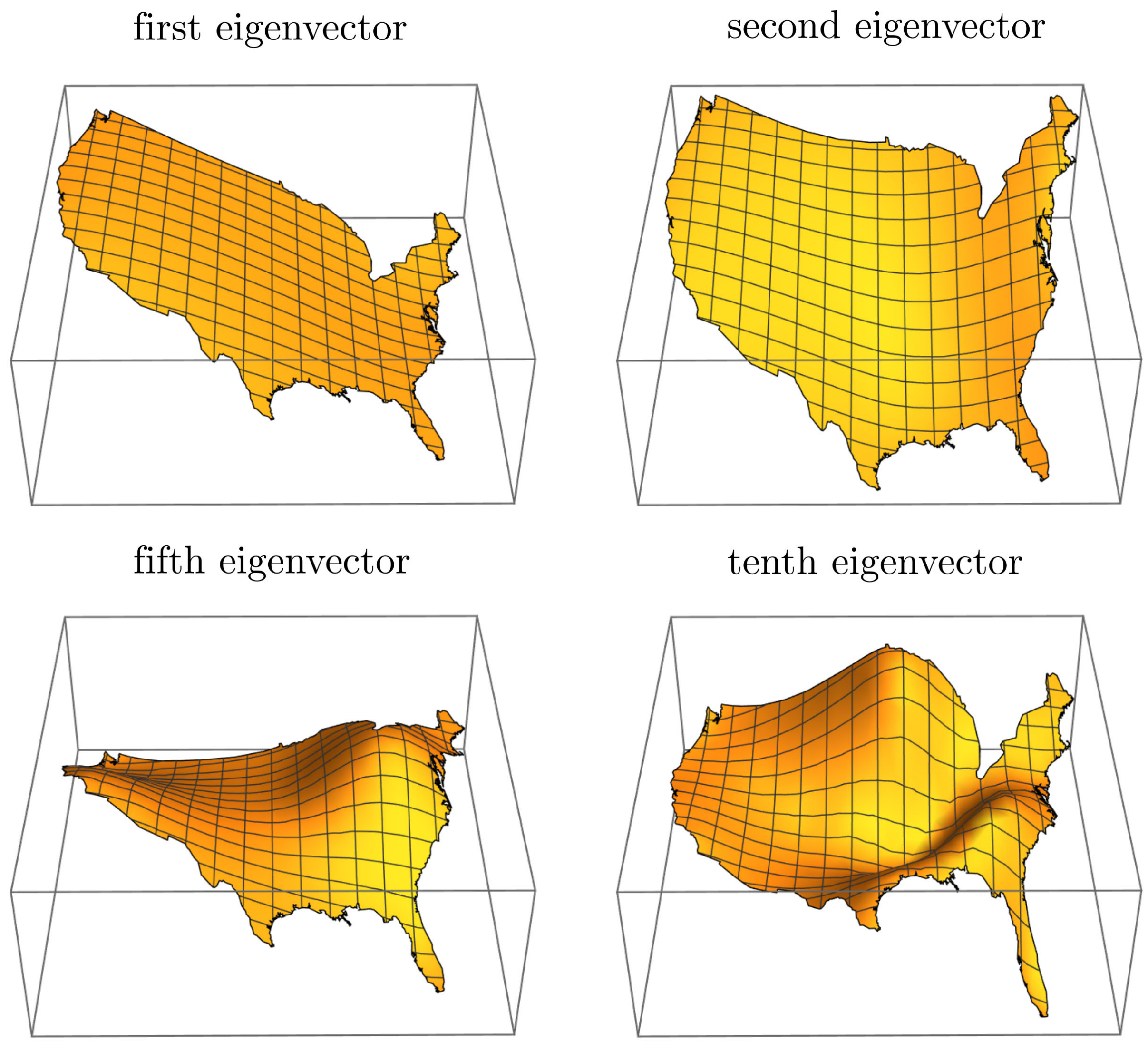
- With spectral decomposition $\bar{\Sigma}_L = R\Lambda R'$ we have

$$\begin{aligned}
\xi^* &= n^{-1}\hat{e}' D_x \bar{\Sigma}_L D_x \hat{e} \\
&= \sum_{j=1}^{n} \lambda_j \left(n^{-1/2} \sum_{l=1}^{n} r_{j,l} x_l \hat{e}_l \right)^2 \\
&= \sum_{j=1}^{n} \lambda_j Y_j^2 \quad \text{with} \quad Y_j = n^{-1/2} \sum_{l=1}^{n} r_{j,l} x_l \hat{e}_l
\end{aligned}$$

- Convenient for asymptotics: truncate at largest q eigenvalues

$$\xi = \sum_{j=1}^{q} \lambda_j Y_j^2 \quad \approx \quad \sum_{j=1}^{n} \lambda_j Y_j^2 = \xi^*$$

Eigenvectors for ACS Data



Asymptotics in General Model

- Model: $y_l = x_l\beta_l + u_l = x_l\beta + e_l, \quad e_l = u_l + x_l(\beta_l - \beta)$

- Test statistic:

$$\xi_n = \sum_{j=1}^{q} \lambda_j Y_{n,j}^2 \quad \text{with} \quad Y_{n,j} = n^{-1/2}\sum_{l=1}^{n} r_{j,l}x_l\hat{e}_l = n^{-1/2}\sum_{l=1}^{n} \tilde{r}_{j,l}x_l e_l$$

- Assumptions

  1. $\{s_l\}_{l=1}^n \subset \mathcal{S} \subset \mathbb{R}^d$ with empirical c.d.f. $G_n \to G$ with bounded density g on $\mathcal{S}$
  2. For uniformly converging $h_n : \mathcal{S} \mapsto \mathbb{R}$, $n^{-1/2}\sum_{l=1}^n h_n(s_l)x_l u_l \Rightarrow \mathcal{N}\left(0, \int h(s)^2\Omega_{xu}(s)dG(s)\right)$ and $n^{-1}\sum_{l=1}^n x_l^2 h(s_l) \xrightarrow{p} \int \Omega_{xx}(s)h(s)dG(s)$ for some functions $\Omega_{xu}, \Omega_{xx} : \mathcal{S} \mapsto \mathbb{R}$
  3. $\beta_l - \beta = n^{-1/2}b(s_l)$

Limit Distribution of ξ

- From above, exploiting that $e_l = u_l + x_l(\beta_l - \beta)$,

$$\begin{aligned} Y_{n,j} &= n^{-1/2} \sum_{l=1}^{n} r_{j,l} x_l \hat{e}_l \\ &= n^{-1/2} \sum_{l=1}^{n} \tilde{r}_{j,l} x_l e_l \\ &= n^{-1/2} \sum_{l=1}^{n} \tilde{r}_{j,l} x_l u_l + n^{-1/2} \sum_{l=1}^{n} \tilde{r}_{j,l} x_l^2 (\beta_l - \beta) \end{aligned}$$

- Using the assumption, we get

$$Y_n \Rightarrow Y \sim \mathcal{N}(B, V)$$

where $B = 0$ under null hypothesis, and

$$\xi \Rightarrow Y' \Lambda_q Y$$

where Λ_q collects the largest q eigenvalues of a covariance kernel of demeaned LBM on $\mathcal{S}$

Feasible Inference

- Estimate V by spatial kernel estimator HAC estimator with elements

$$\hat{V}_{i,j} = n^{-1} \sum_{l,\ell=1}^{n} \hat{v}_{l,i} \,\mathsf{exp}(-c_V||s_l - s_\ell||)\hat{v}_{\ell,j}, \quad \hat{v}_{l,j} = \tilde{r}_{j,l} x_l \hat{e}_l$$

for large c_V ($c_V \to \infty$ induces consistency)

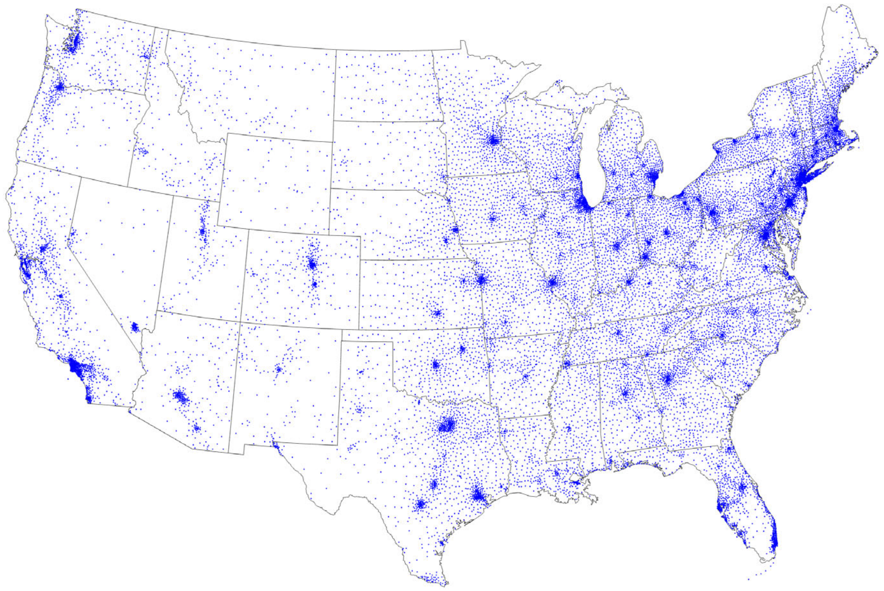
American Community Survey

- 62 socioeconomic variables (education, income, employment, race, health, marital status, . . .), 5 year averages 2018-2022, variables measured in percentiles across the 21k zip codes

  $\Rightarrow$ 1514 bivariate regressions

- GLS difference transform applied to all variables

- $n = 21,194$ zip codes in 48 states + DC



Monte Carlo Simulations

- Same locations $\{s_l\}_{l=1}^{n}$ as ACS data

- Let $\eta \sim \mathcal{G}_c$ be mean-zero Gaussian $n \times 1$ vector with $\mathbb{E}[\eta_l \eta_\ell] = \exp[-c||s_l - s_\ell||]$. DGPs:

  1. x_l and u_l are generated by independent $\mathcal{G}_c$ processes
  2. x_l is randomly selected from the 62 variables and u_l follows a $\mathcal{G}_c$ process
  3. $\{y_l^o, x_l^o\}$ are a pair of series from 1,514 bivariate regressions, $x_l = x_l^o \eta_{x,l}$ and $u_l = y_l^o \eta_{u,l}$ where $\eta_x, \eta_u \sim iid\mathcal{G}_c$

- Results:
  - Generally good size control even under heteroskedasticity
  - Familiar relationship between HAC bandwidth c_V and degree of robustness against spatial correlation, and a corresponding trade-off in power

Empirical Results in 1514 Bivariate Regression

	Quantile (across 1514 regressions)				
	0.05	0.25	0.50	0.75	0.95
	OLS estimates with HAC SE				
$\lvert t_{\hat{\beta}} \rvert$	0.63	3.75	8.28	14.6	29.4
$\lvert \hat{\beta} \rvert$	0.01	0.05	0.11	0.22	0.45
	Spatial Variation in β				
ξ_{15} p-value	0.00	0.02	0.07	0.20	0.52
$\sigma_{\Delta^{1000\text{km}}}(\hat{\kappa}^{MU})$	0.03	0.03	0.05	0.09	0.18

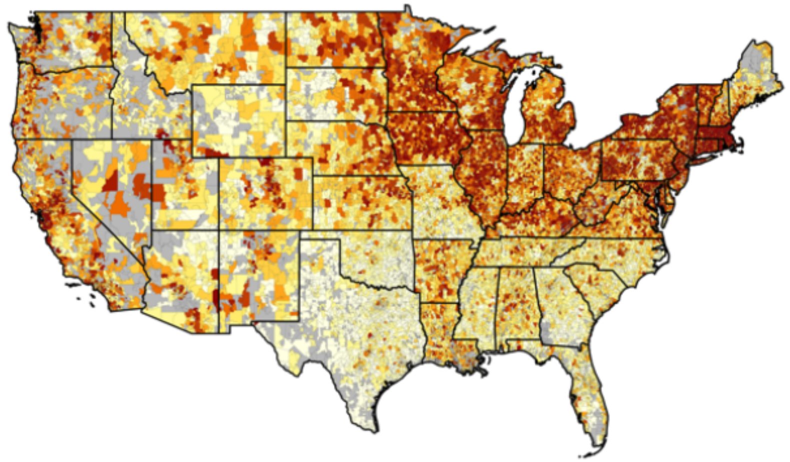
Notes:

- $\hat{\kappa}^{MU}$ is median unbiased estimate of κ in $\beta_l - \beta \sim \kappa L(s_l)$ based on ξ_{15}
- $\sigma_{\Delta^{1000\text{km}}}(\hat{\kappa}^{MU})$ is implied standard deviation over 1000km

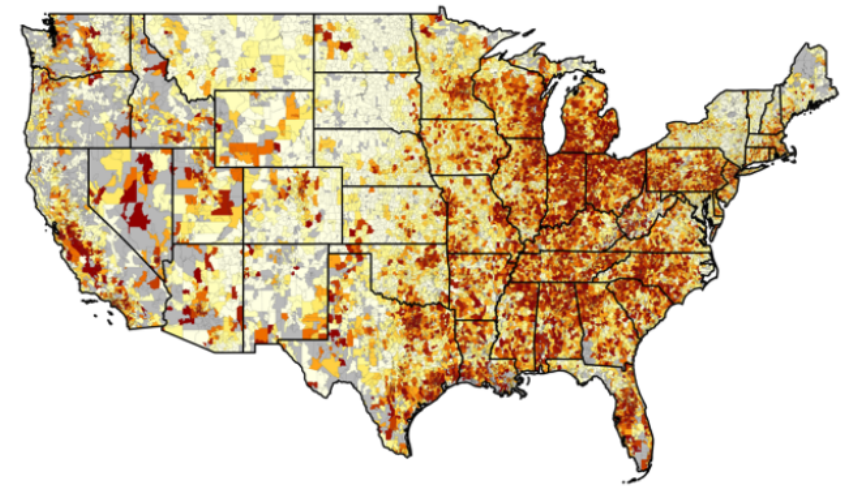
Motivating Example Revisited

Data by zip code: Levels

Health Insurance Coverage

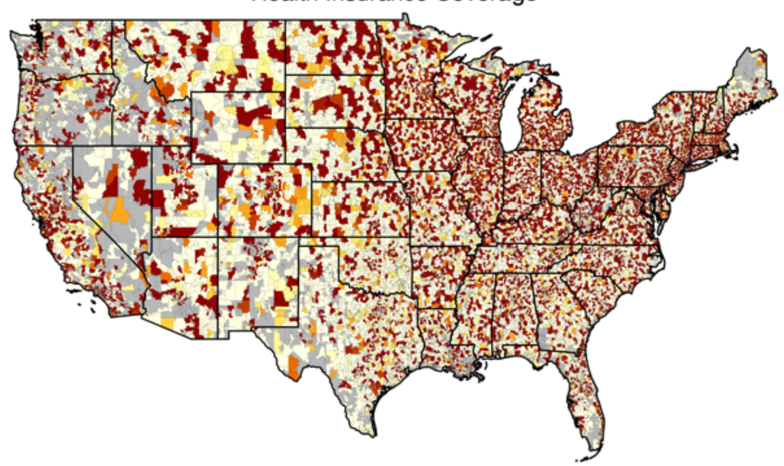


Employed by Private Company

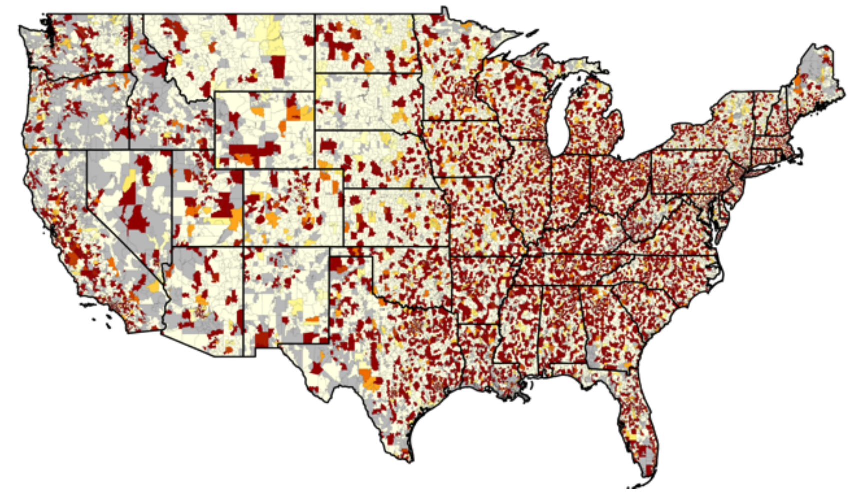


Data by zip code: GLS transformed

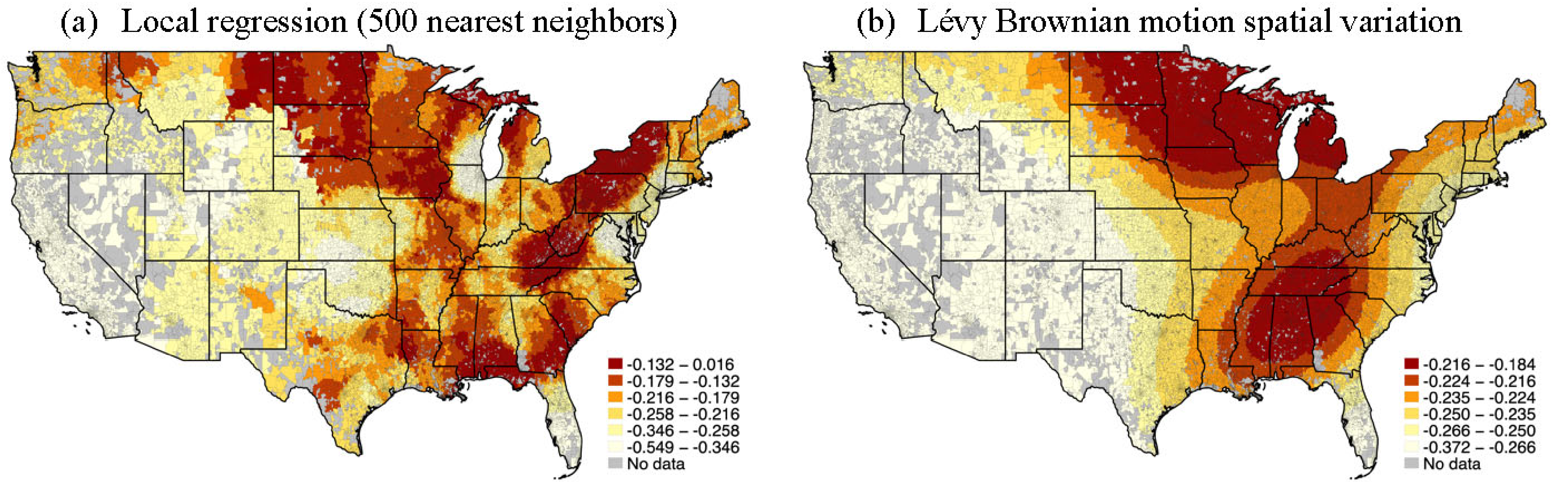
Health Insurance Coverage



Employed by Private Company



Estimate of Spatial Variation



$\Rightarrow$ Right panel exploits (approximately) jointly normal distribution of q weighted averages Y_n and LBM variation $\beta_l - \beta = \kappa L(s_l)$ with $\kappa = \hat{\kappa}^{MU}$

Conclusion

- Generalize standard time series tool of generic regression stability testing to spatial case
- Allow for spatially correlated errors and heterogeneity in second moments
- Empirical finding of widespread instability in bivariate regressions using ACS data
- Interpretability under instability?
  - Coefficient no longer BLP given location (could use smoothed estimates)
  - Extrapolationt to other regions highly questionable

Thank you!