Testing Coefficient Variability in Spatial Regressions

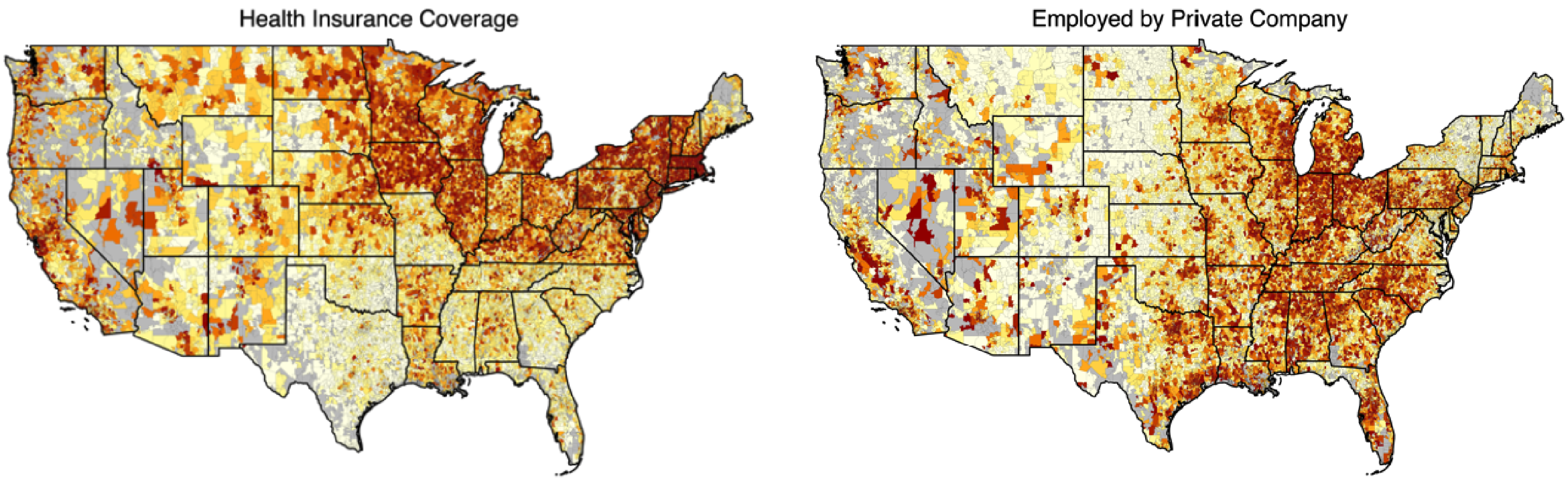
Ulrich Müller and Mark Watson

Princeton University

July 9, 2024: NBER-SI

A Motivating Example: Health insurance and employment across 21k zip codes

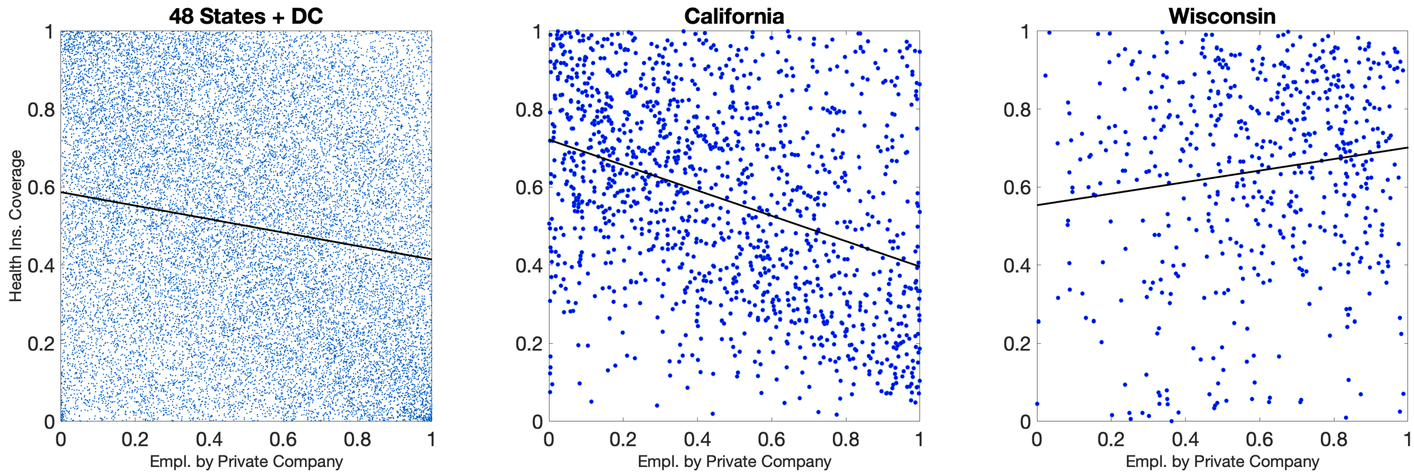
Data by zip code: Levels



- Units: Variables are measured in percentiles across the 21k zip codes.

A Motivating Example (continued)

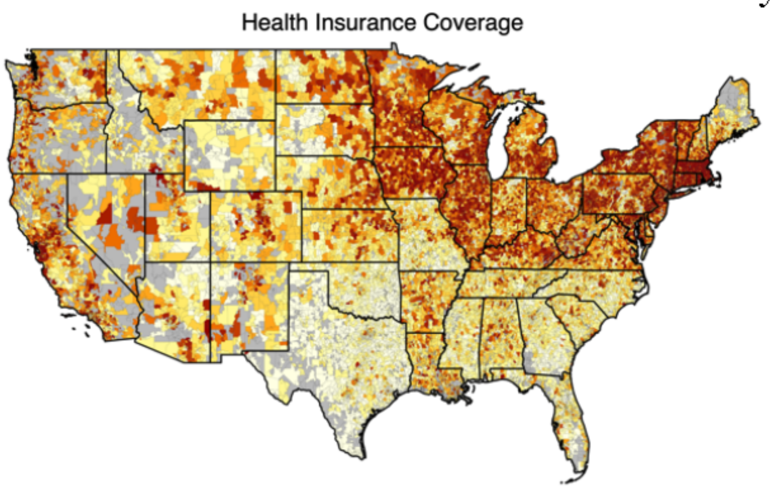
Scatter Plots and OLS estimates: Levels

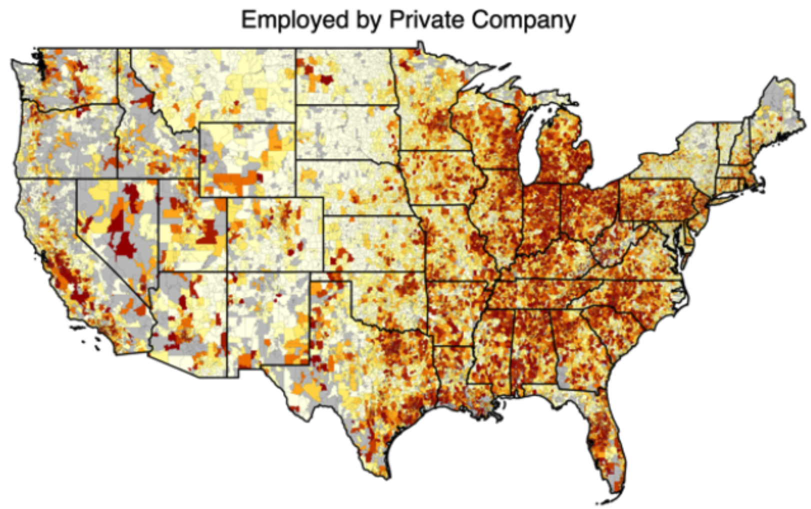


$\hat{\beta}$ = -0.17 $\hat{\beta}$ = -0.33 $\hat{\beta}$ = 0.15

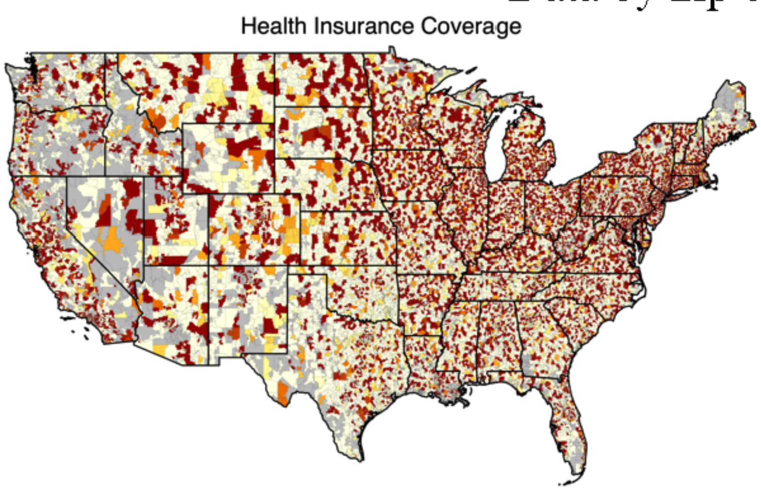
A Motivating Example (continued)

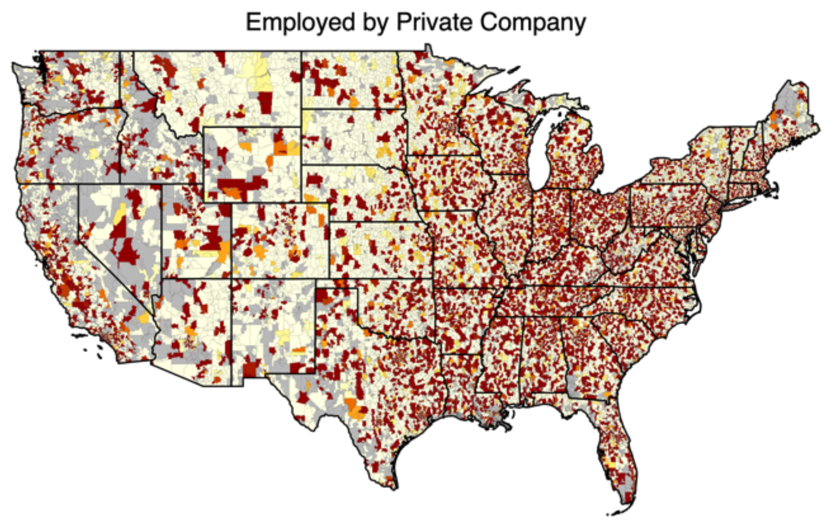
Data by zip code: Levels



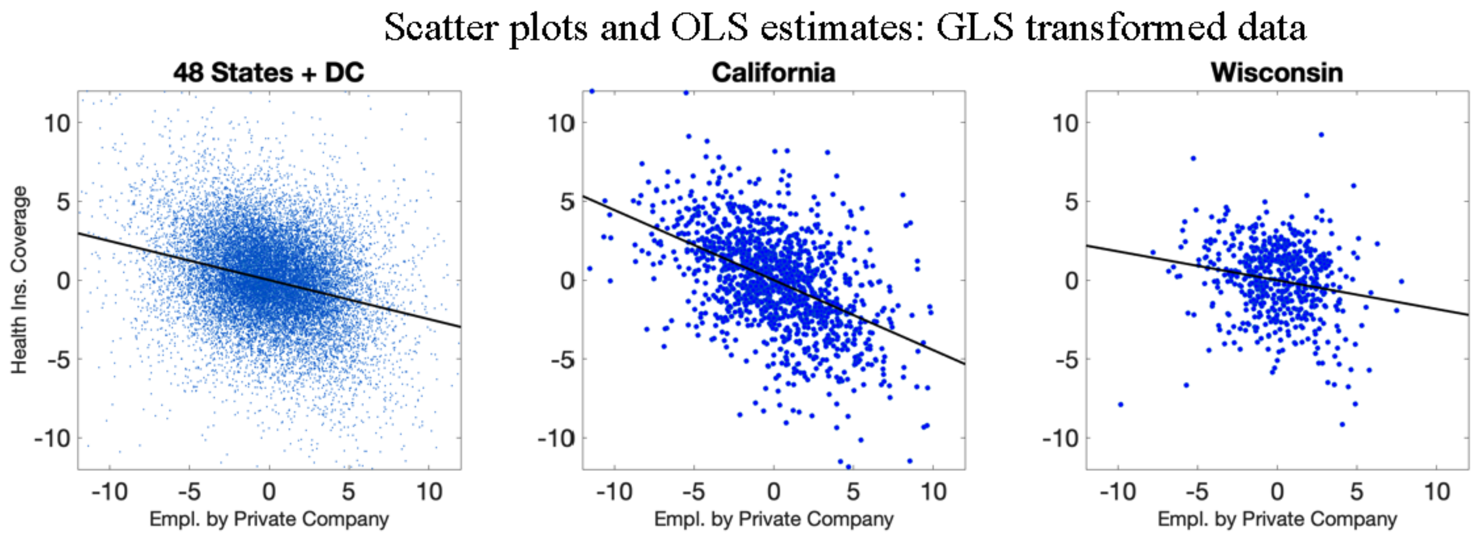


Data by zip code: GLS transformed





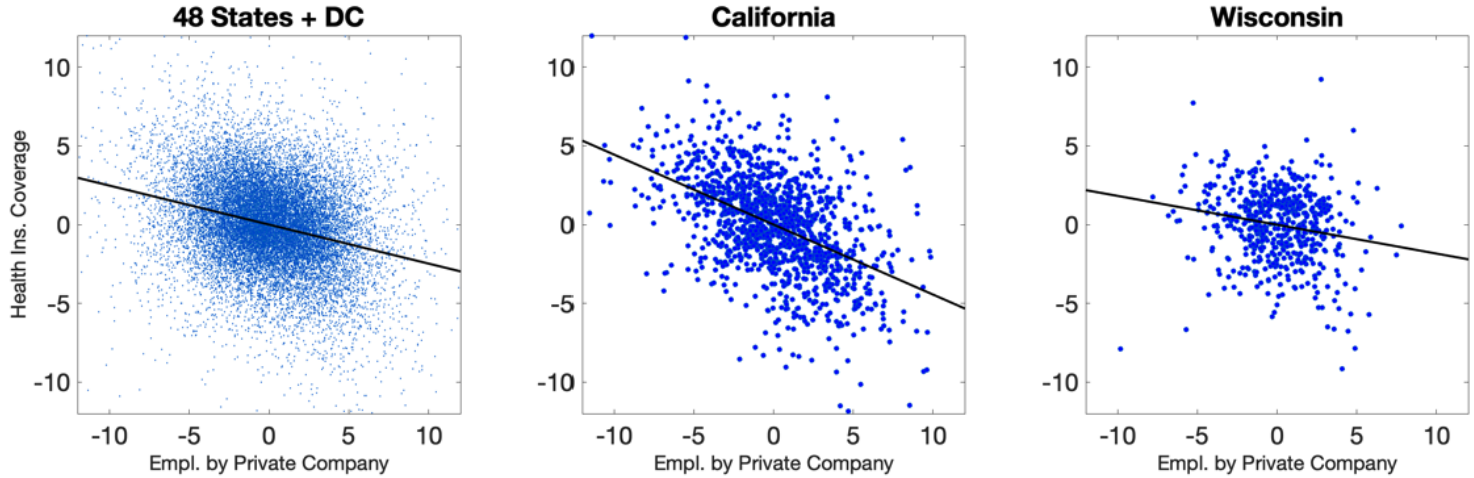
A Motivating Example (continued)



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

Spatial Variation in the Regression Coefficient?

Scatter plots and OLS estimates: GLS transformed data



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

Familiar Issues: (a) Why California and Wisconsin? (b) Other states? (c) How does the coefficient change (if at all) across the U.S.?

Time Series Regression Precendents?

- Discrete Breaks: Chow (1960), Quandt (1960), Andrews (1993), ...
- Martingale Variation: Nyblom (1989), ...
- Inference refinements: Serial Correlation (Newey-West (1987), 2nd-moment heterogeneity (Hansen (2000)), ...
- Lots on estimation and modelling ...

Spatial Regressions:

- Chow tests (with spatially correlated errors): Anselin (1990), ...
- Local Spatial Regressions: Fotheringham et al (2002, 2024), ... (inference assume *iid* observations)

This Paper:

- Nyblom-like test spatial variation in coefficients (best local Lévy-Brownian motion variation)
  - Size under general distributional assumptions and spatial correlation (like Andrews (1993))
  - Accomodates spatially varying second moments (like Hansen (2000))

Outline:

1. Canonical Gaussian model $\Rightarrow$ test statistic
2. Validity of test under more general assumptions: distribution, spatial correlation, 2nd moments, etc.
3. Power under different local alternatives
4. Details: Computing the test statistic. Computing other statistics measuring spatial variation.
5. Simulation experiments (Empirical calibration)
6. Instabillity over the U.S. in 1514 regressions involving 62 socio-economic variables.

Canonical Gaussian Model:

$$y_l = x_l\beta_l + \ldots(z_l'\alpha)\ldots + u_l,\ \ l = 1, \ldots, n,$$

$$= x_l\beta + e_l \text{ with } e_l = u_l + x_l(\beta_l - \beta)$$

- y_l, x_l, u_l are scalars
- (y_l, x_l) are associated with known spatial locations $s_l \in \mathcal{S} \subset \mathbb{R}^d$, for $d \geq 1$
- $u_l \sim iid\mathcal{N}(0,1)$ and $\{x_l\}$ is nonstochastic.
- Null and alternative:

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

- Invariance: $y \to y + xb$. (Test will be based on the OLS residuals $\hat{e}_l$.)

Best Local Test:

Best test against alternative $\{\beta_l\}_{l=1}^n = \{\beta_l^1\}_{l=1}^n$ rejects the null hypothesis for large values of

$$\sum_{l=1}^{n} \beta_l^1 x_l \hat{e}_l$$

- But ... what value of $\{\beta_l\}_{l=1}^n = \{\beta_l^1\}_{l=1}^n$ should one use?
  - Standard Suggestion: Consider many possible values of $\{\beta_l\}_{l=1}^n$ and evaluate tests based on weighted average power
    - Same as using alternative with a stochastic model for $\{\beta_l\}_{l=1}^n$ using the weight function as pdf.

- We use:

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

  where $L(s)$ is Lévy-Brownian motion (LBM) and κ is a scale.
  - Lévy-Brownian motion: Spatial generalization of Brownian motion.
    - Gaussian process with $\mathbb{E}[L(s)L(r)] = \frac{1}{2}(||r|| + ||s|| - ||s - r||)$, etc.

- Best local test is Score/LM test for $\kappa = 0$ versus $\kappa > 0$.

Best Local Test (continued)

Best local test: Reject for large values of quadratic form of $\{x_l \hat{e}_l\}$ around LBM covariance matrix:

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

where $\bar{\Sigma}_L$ is the covariance matrix of (demeaned) L at the sample locations $(s_1, ..., s_n)$ and $D_x = \text{diag}(x_l)$

Best Local Test (continued)

- A useful re-writing of ξ^*
  - Write the spectral decomposition of $\bar{\Sigma}_L = R\Lambda R'$ where columns of R are eigenvectors and Λ is diagonal with eigenvalues (ordered from largest to smallest) on diagonal.
  - Then

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

- A cheat (facilitates large-sample analysis in more general model)

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

Moving beyond the canonical model:

- Model:

$$y_l = x_l \beta_l + u_l, \;\; l = 1, \ldots, n$$

- Test statistic:

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

- Large-n assumptions:

  – Locations: $(\lambda_{j,n}, r_{j,l,n})$ ... $\{s_l\}$ are non-stochastic with empirical CDF $G_n \to G$ with density g ... then $(\lambda_{j,n}, r_{j,l,n})$ converges to eigenvalues and eigenfunctions of covariance kernel of demeaned L.

  – CLT and LLN allowing for spatial correlation and spatially varying 2nd moments:

    * CLT for $a_n^{-1/2} n^{-1/2} \sum_l h(s_l) x_l u_l$
    * LLN for $n^{-1} \sum_l h(s_l) x_l^2$

  – Local alternatives: $\beta_l - \beta = \kappa_n b(s_l), \kappa_n = a_n^{1/2} n^{-1/2}$ where b is a continuous function (could be $L(s)$).

Large-n results:

- From above

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

- So (with assumptions)

$$Y_n \Rightarrow Y \sim N(0, V_0 + V_1)$$

  - (Deterministic $b(s)$ replaces V_1 with non-zero mean.)

- Estimators

$$\hat{V}_{0,i,j} = n^{-1} \sum_{l,\ell} (\tilde{r}_{j,l,n} x_l \hat{e}_l) k_c(s_l, s_\ell)(\tilde{r}_{i,\ell,n} x_\ell \hat{e}_\ell)$$

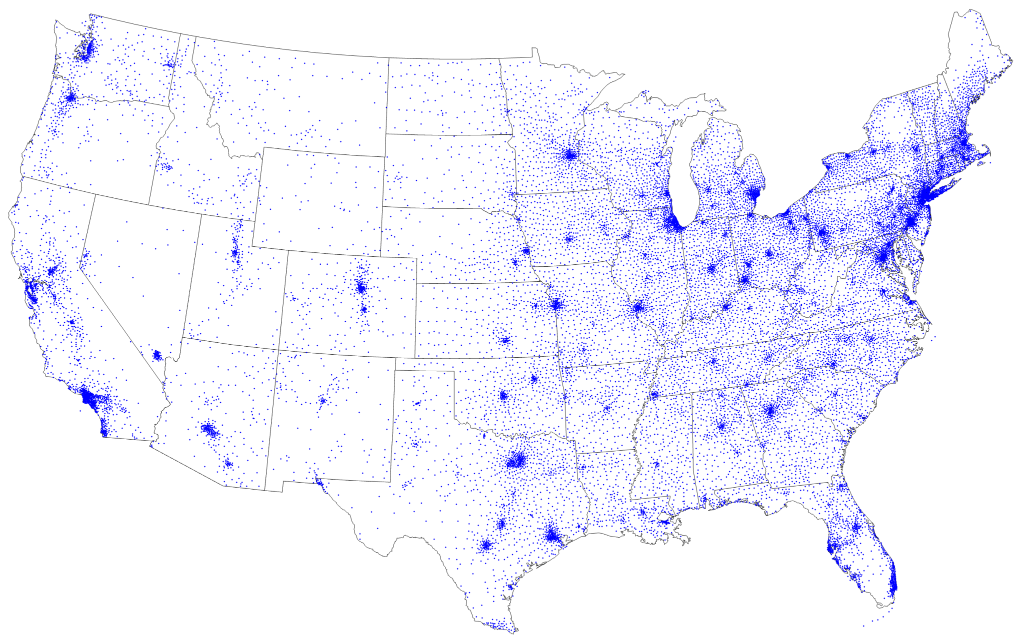
with $k_c(s, r) = \exp(-c||s - r||)$. ($c$ is bandwidth parameter.)

  - Consistency as $c \to 0$

- $\hat{V}_1$: see paper

That summarizes much of the theory. Now on to a dataset:

- 62 Socioeconomic variables (population, educational attainment, income, employment, race, citizenship, health, marital status, mobility, ...) from ACS. 5-year averages from 2018-2022.
- GLS transform applied to all of the variables
- $n = 21,194$ zip codes in 48-states + DC.



- 1,514 Bivariate regressions using the 62 variables.

Simulation Results: (see paper)

- Use data to calibrate a variety of DGPs (under null and alternative, including discrete breaks)
- Issue: effect of bandwidth choice for covariance matrix estimate on size and power.

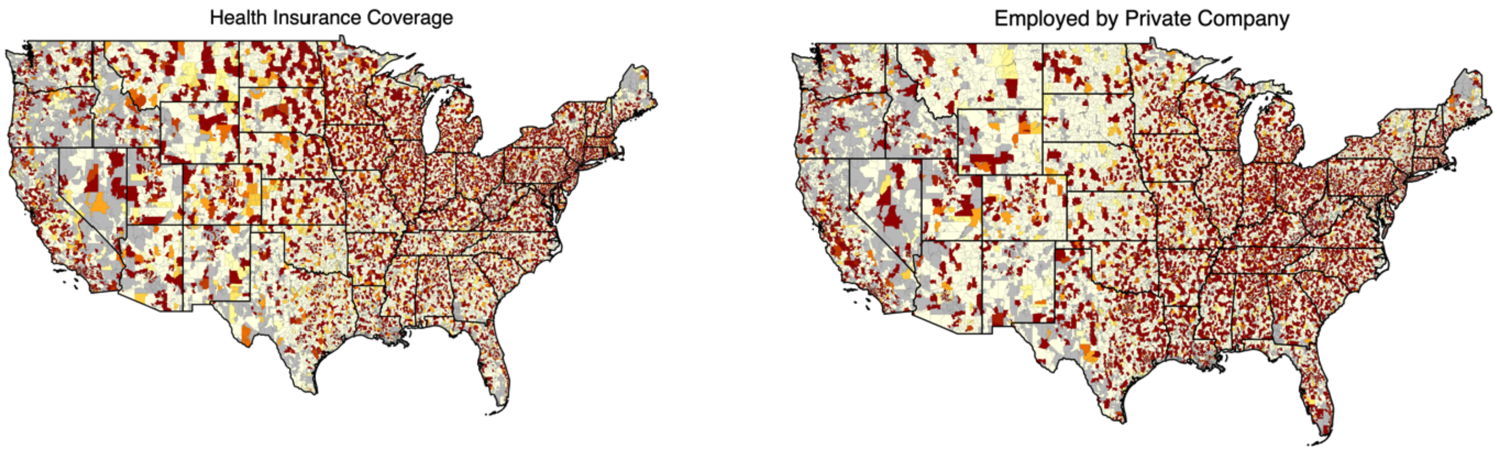
Empirical Results (1514 bivariate regressions)

- Units: Measured in percentiles across the 21k zip codes.
- Results Summary:

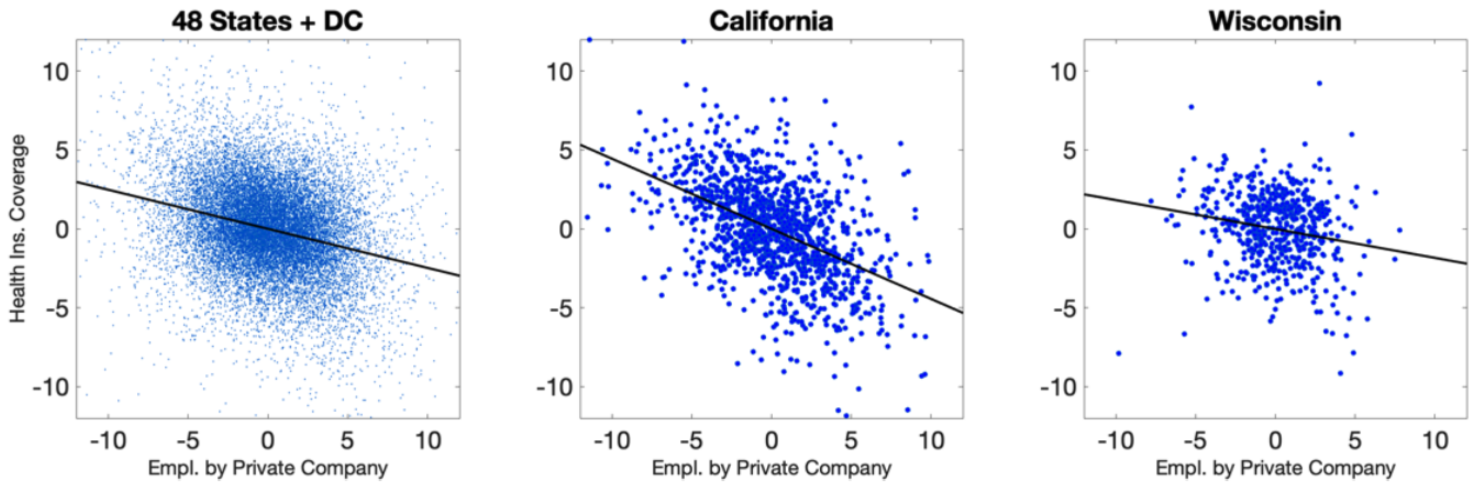
		Quantile (across 1,514 regressions)				
		0.05	0.25	0.50	0.75	0.95
		(a) OLS estimates				
$\lvert t_{\hat{\beta}}\rvert$		0.63	3.75	**8.28**	14.60	29.36
$\lvert\hat{\beta}\rvert$		0.01	0.05	**0.11**	0.22	0.45
		(b) Spatial variation in β				
ξ_{15} p-value		0.00	0.02	**0.07**	0.20	0.52
$\sigma_{\Delta^{1000km}}$ $(\hat{\kappa}^{MU})$		0.00	0.03	**0.05**	0.09	0.18

A Motivating Example (again)

Data by zip code: GLS transformed

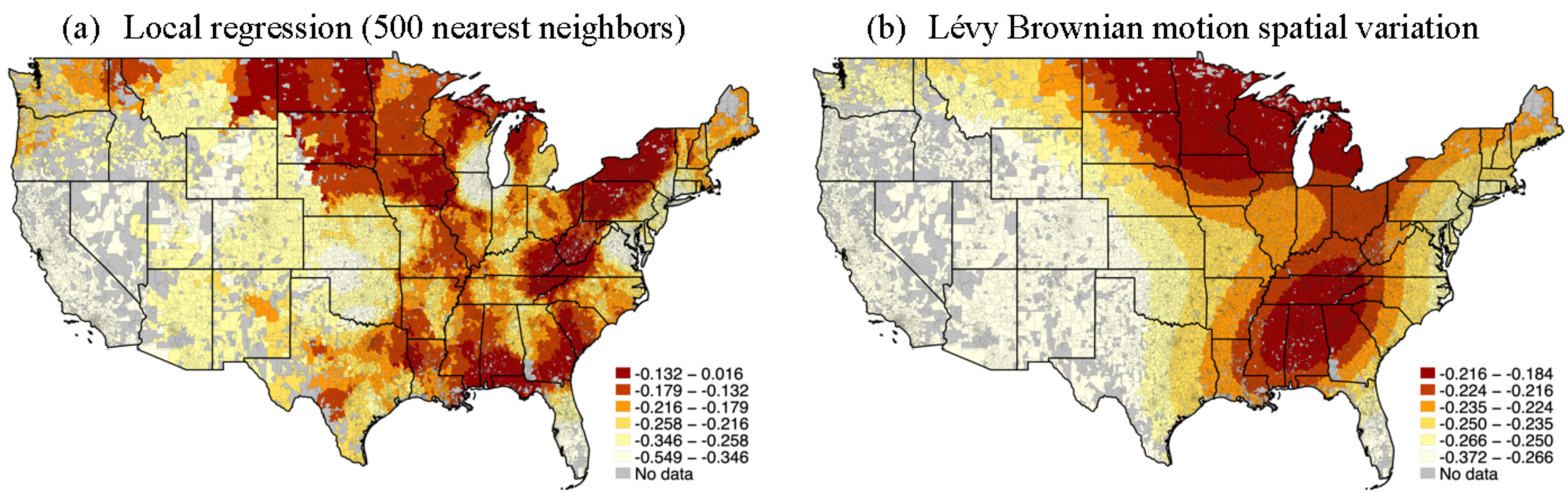


Scatter plots and OLS estimates: GLS transformed data



A Motivating Example (again)

Estimates of β in the HIC-PCE regression



Concluding Slide