Social Connectedness and the Geographic Spread of COVID-19

Theresa Kuchler (NYU Stern, CEPR)
Dominic Russel (NYU Stern)
Johannes Stroebel (NYU Stern, CEPR, NBER)

August 28, 2020

Motivation

- To determine a region's risk of outbreak, its valuable to know which individuals are likely to physically interact (Piontti et al., 2018)

- **Problem:** The geographic structure of social networks is difficult to measure on a national or global scale

- **Solution:** Aggregated measure of connections between region pairs from de-identified Facebook social graph
  - Facebook global social network = 2.5 billion monthly active users
  - Limit on friends & required consent of both parties $\rightarrow$ more likely to capture real-world connections than other online networks

Social Connectedness Index

- Social Connectedness Index (Bailey et al., 2018)

$$Social\ Connectedness_{i,j} = \frac{FB\ Connections_{i,j}}{FB\ Users_i * FB\ Users_j}$$

- Normalized number of Facebook friendship links between regions
- Captures relative probability of friendship between Facebook users in regions i and j
- Higher county SCI to NYC $\rightarrow$ increase in likelihood of being destination for those fleeing pandemic (Coven and Gupta, 2020)
- Data are widely available to other researchers!

sci_data@fb.com

"Hotspot" Case Studies

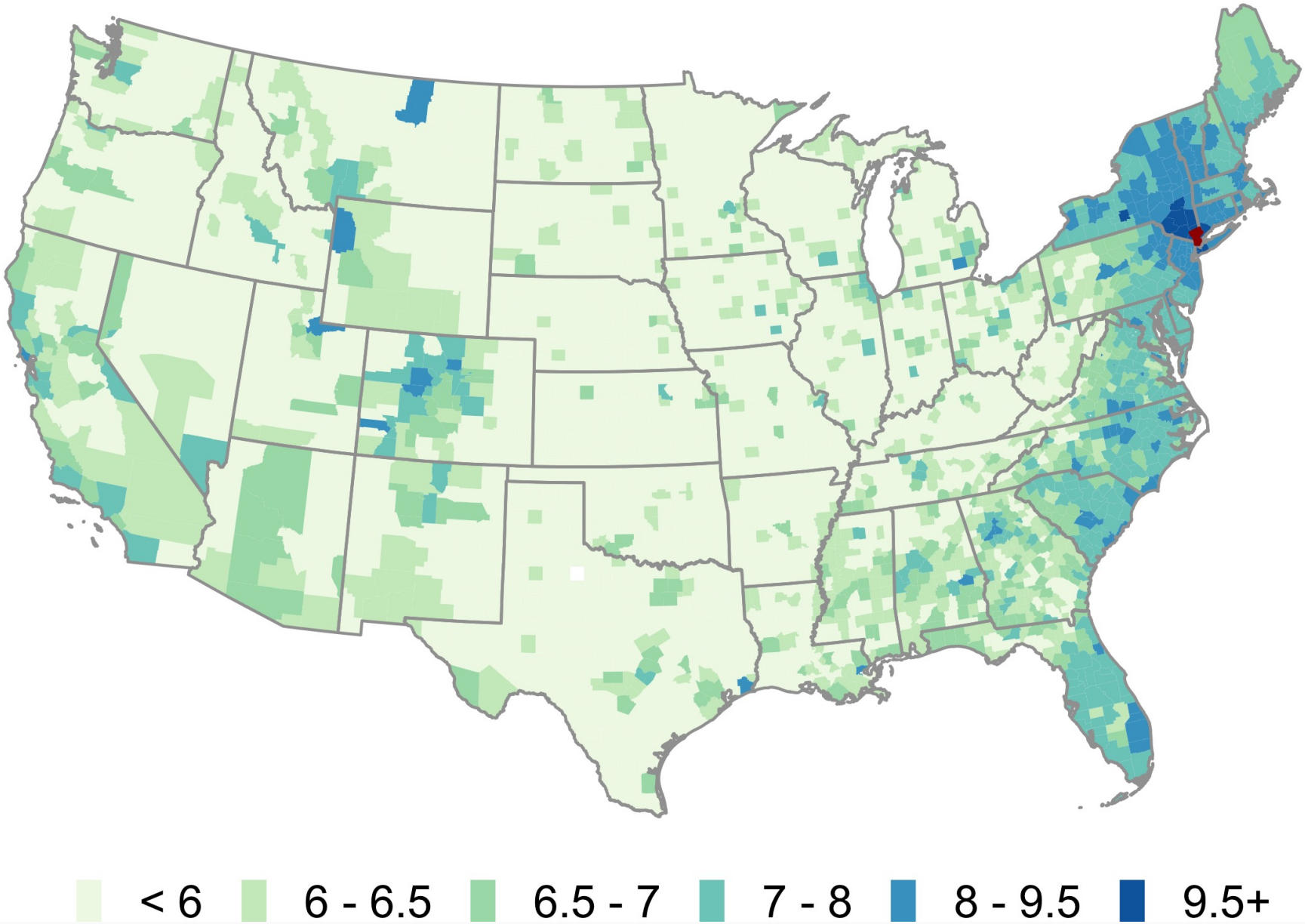
- **Question:** Do areas with stronger social ties to early COVID-19 "hotspots" have more confirmed cases by end of March?

- **Case 1:** Westchester County, USA
  - Contains New Rochelle, highly publicized early U.S. outbreak
  - Home of many well-heeled residents who might have fled

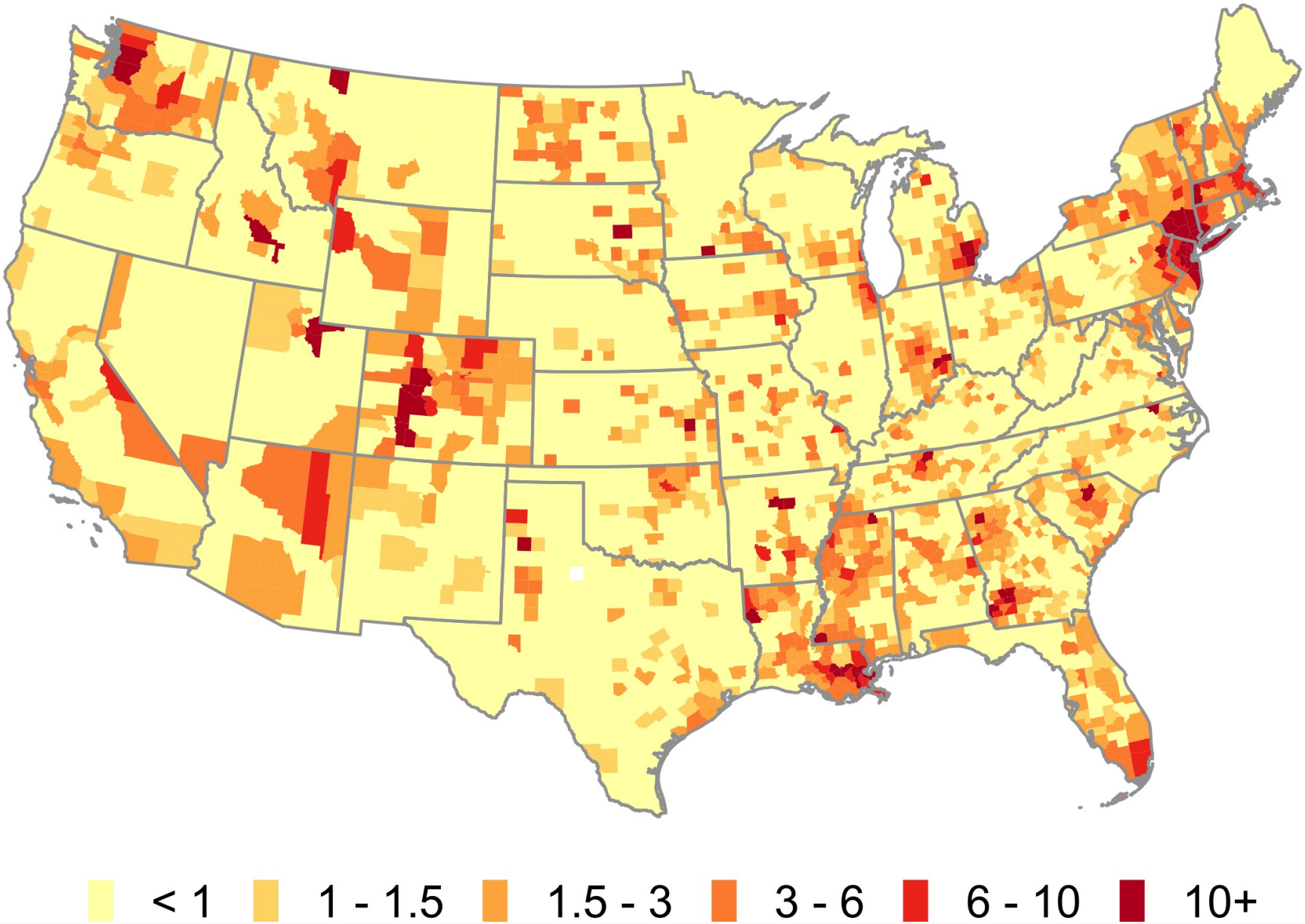
- **Case 2:** Lodi Province, Italy
  - Contains Codogno, one of earliest Italian outbreak centers
  - Part of Lombardy, a destination for workers/students from south Italy

- Importantly: control for geographic distance, population density, and median household income; exclude regions within 50mi/km

Case Study 1/2: log(SCI) to Westchester County, USA

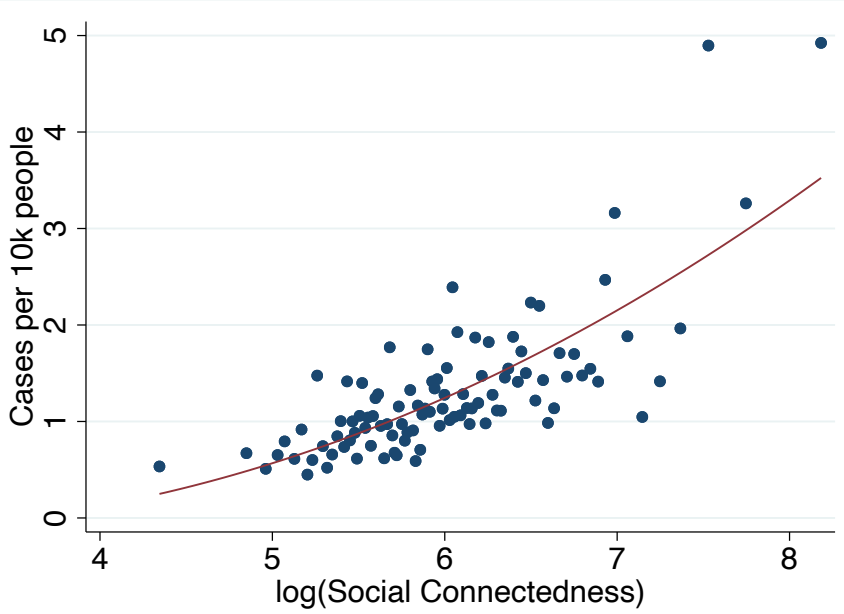


Case Study 1/2: COVID Cases per 10k Residents, March 30

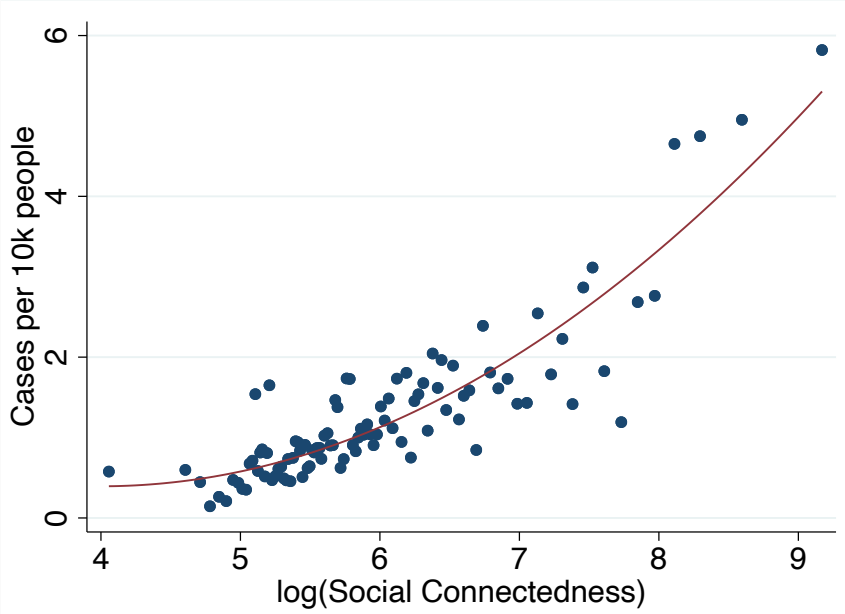


Case Study 1/2: SCI vs COVID Binscatters, Westchester

(a) Without controls

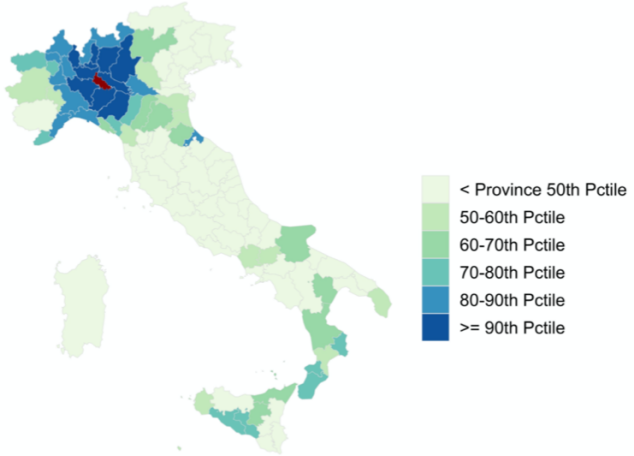


(b) With controls

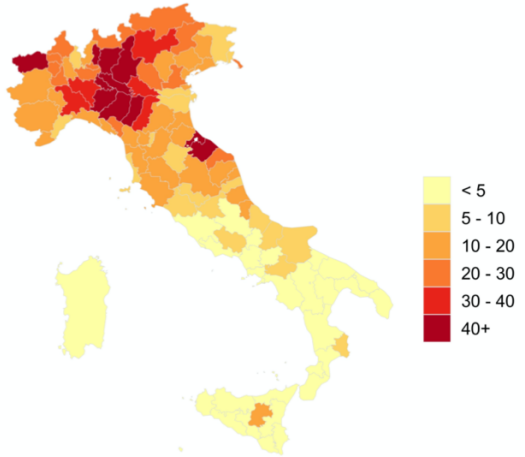


Case Study 2/2: Lodi Province, Italy

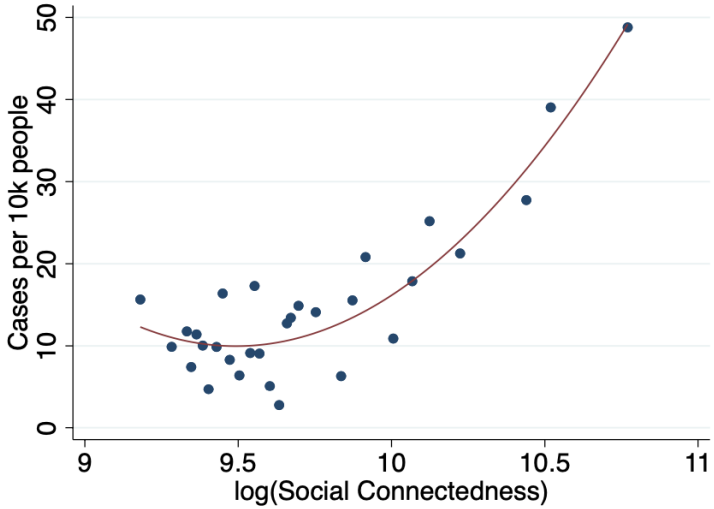
(a) Percentile of SCI to Lodi Province, Italy



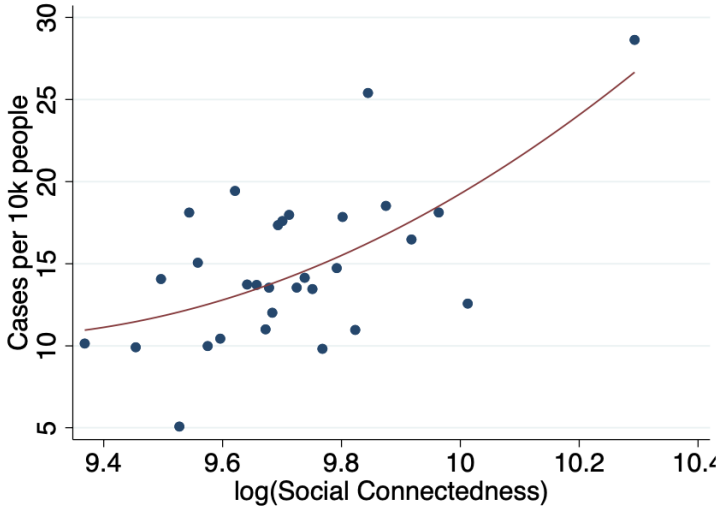
(b) COVID-19 Cases per 10k Residents by Province



(c) Lodi binscatter without controls



(d) Lodi binscatter with controls



County Time Series

- **Question:** Does the *Social Connectedness Index* have systematic predictive power above geographic distance?

- Construct two measures:

$$(1)\ \textit{Social Proximity to Cases}_{i,t} = \sum_j \textit{Cases Per 10k}_{j,t} * \frac{\textit{Social Connectedness}_{i,j}}{\sum_h \textit{Social Connectedness}_{i,h}}$$

$$(2)\ \textit{Physical Proximity to Cases}_{i,t} = \sum_j \textit{Cases Per 10k}_{j,t} * \frac{1}{1 + \textit{Distance}_{i,j}}$$

County Time Series - Regression Framework

$$
\begin{aligned}
log(\Delta\ Cases\ per\ 10k + 1)_{i,t} &= \beta_1 * log(\Delta Cases\ per\ 10k + 1)_{i,t-1} \\
&+ \beta_2 * log(\Delta Cases\ per\ 10k + 1)_{i,t-2} \\
&+ \beta_3 * log(\Delta Social\ Proximity\ to\ Cases)_{i,t-1} \\
&+ \beta_4 * log(\Delta Social\ Proximity\ to\ Cases)_{i,t-2} \\
&+ \beta_5 * log(\Delta Physical\ Proximity\ to\ Cases)_{i,t-1} \\
&+ \beta_6 * log(\Delta Physical\ Proximity\ to\ Cases)_{i,t-2} \\
&+ X_i + \epsilon_{i,t}
\end{aligned}
$$

- X_i includes controls for population density, median household income, and state

County Times Series - Regression Results

	log(Change in Cases per 10k Residents + 1)							
	March 31 - April 13	April 14 - April 27	April 28 - May 11	May 12 - May 25	May 26 - June 8	June 9 - June 22	June 23 - July 6	July 7- July 20
2 Week Lag: log(Change in Social Proximity to Cases + 1)	0.731*** (0.093)	0.379*** (0.087)	0.141** (0.059)	0.189*** (0.061)	0.577*** (0.062)	0.182** (0.073)	0.320*** (0.057)	0.259*** (0.070)
4 Week Lag: log(Change in Social Proximity to Cases + 1)	0.384 (0.449)	-0.224* (0.129)	0.137* (0.082)	0.023 (0.060)	-0.111* (0.061)	0.208*** (0.074)	0.046 (0.057)	0.101 (0.063)
2 Week Lag: log(Change in Physical Proximity to Cases + 1)	1.259*** (0.182)	0.699* (0.395)	2.105*** (0.283)	1.232*** (0.261)	-0.074 (0.314)	2.270*** (0.434)	1.361*** (0.350)	2.025*** (0.427)
4 Week Lag: log(Change in Physical Proximity to Cases + 1)	-2.425*** (0.745)	-0.273 (0.463)	-1.593*** (0.291)	-0.892*** (0.282)	0.412 (0.288)	-2.742*** (0.443)	-1.556*** (0.329)	-1.871*** (0.403)
2 Week Lag: log(Change in Cases per 10k Residents + 1)	0.174*** (0.059)	0.403*** (0.050)	0.556*** (0.036)	0.466*** (0.036)	0.278*** (0.035)	0.365*** (0.041)	0.306*** (0.033)	0.320*** (0.037)
4 Week Lag: log(Change in Cases per 10k Residents + 1)	-0.136 (0.256)	0.136* (0.076)	-0.019 (0.047)	0.068* (0.037)	0.126*** (0.035)	-0.017 (0.039)	0.005 (0.033)	0.021 (0.034)
Pop Density FEs	Y	Y	Y	Y	Y	Y	Y	Y
Median Household Income FEs	Y	Y	Y	Y	Y	Y	Y	Y
State FEs	Y	Y	Y	Y	Y	Y	Y	Y
Sample Mean	1.234	1.253	1.331	1.369	1.422	1.579	2.031	2.524
R-Squared	0.600	0.571	0.642	0.647	0.667	0.621	0.678	0.706
N	3,131	3,131	3,131	3,131	3,131	3,131	3,131	3,131

- Two-week lagged SCI-weighted cases is always significant predictor

County Times Series - Regression Results

	log(Change in Cases per 10k Residents + 1)							
	March 31 - April 13	April 14 - April 27	April 28 - May 11	May 12 - May 25	May 26 - June 8	June 9 - June 22	June 23 - July 6	July 7- July 20
2 Week Lag: log(Change in Social Proximity to Cases + 1)	0.731*** (0.093)	0.379*** (0.087)	0.141** (0.059)	0.189*** (0.061)	0.577*** (0.062)	0.182** (0.073)	0.320*** (0.057)	0.259*** (0.070)
4 Week Lag: log(Change in Social Proximity to Cases + 1)	0.384 (0.449)	-0.224* (0.129)	0.137* (0.082)	0.023 (0.060)	-0.111* (0.061)	0.208*** (0.074)	0.046 (0.057)	0.101 (0.063)
2 Week Lag: log(Change in Physical Proximity to Cases + 1)	1.259*** (0.182)	0.699* (0.395)	2.105*** (0.283)	1.232*** (0.261)	-0.074 (0.314)	2.270*** (0.434)	1.361*** (0.350)	2.025*** (0.427)
4 Week Lag: log(Change in Physical Proximity to Cases + 1)	-2.425*** (0.745)	-0.273 (0.463)	-1.593*** (0.291)	-0.892*** (0.282)	0.412 (0.288)	-2.742*** (0.443)	-1.556*** (0.329)	-1.871*** (0.403)
2 Week Lag: log(Change in Cases per 10k Residents + 1)	0.174*** (0.059)	0.403*** (0.050)	0.556*** (0.036)	0.466*** (0.036)	0.278*** (0.035)	0.365*** (0.041)	0.306*** (0.033)	0.320*** (0.037)
4 Week Lag: log(Change in Cases per 10k Residents + 1)	-0.136 (0.256)	0.136* (0.076)	-0.019 (0.047)	0.068* (0.037)	0.126*** (0.035)	-0.017 (0.039)	0.005 (0.033)	0.021 (0.034)
Pop Density FEs	Y	Y	Y	Y	Y	Y	Y	Y
Median Household Income FEs	Y	Y	Y	Y	Y	Y	Y	Y
State FEs	Y	Y	Y	Y	Y	Y	Y	Y
Sample Mean	1.234	1.253	1.331	1.369	1.422	1.579	2.031	2.524
R-Squared	0.600	0.571	0.642	0.647	0.667	0.621	0.678	0.706
N	3,131	3,131	3,131	3,131	3,131	3,131	3,131	3,131

- Two-week lagged SCI-weighted cases is always significant predictor

County Time Series - Prediction Exercise

- Previous result provides evidence that *Social Connectedness Index* is predictive of spread of COVID-19
- Next we test by making actual predictions
- For each time period, build models with same setup (one & two period lags) trained using predictions from all previous periods
- Two model specifications:
  1. Simple linear regression
  2. Random forest

County Time Series - Prediction Results

	RMSE: Linear Regression			RMSE: Random Forest		
	Without Social Proximity to Cases	With Social Proximity to Cases	Diff. from Social Proximity to Cases	Without Social Proximity to Cases	With Social Proximity to Cases	Diff. from Social Proximity to Cases
(1) April 14 - April 27	2.523	2.598	0.075	1.597	1.497	-0.099
(2) April 28 - May 11	1.082	1.168	0.086	0.922	0.845	-0.077
(3) May 12 - May 25	0.742	0.729	-0.014	0.754	0.726	-0.028
(4) May 26 - June 8	0.742	0.716	-0.026	0.701	0.678	-0.024
(5) June 9 - June 22	0.826	0.798	-0.027	0.795	0.770	-0.025
(6) June 23 - July 6	0.886	0.865	-0.022	0.862	0.840	-0.022
(7) July 7 - July 20	0.813	0.792	-0.020	0.802	0.786	-0.016

- Social proximity to cases reduces root mean squared error of predictions, over and above physical proximity to cases

Conclusion

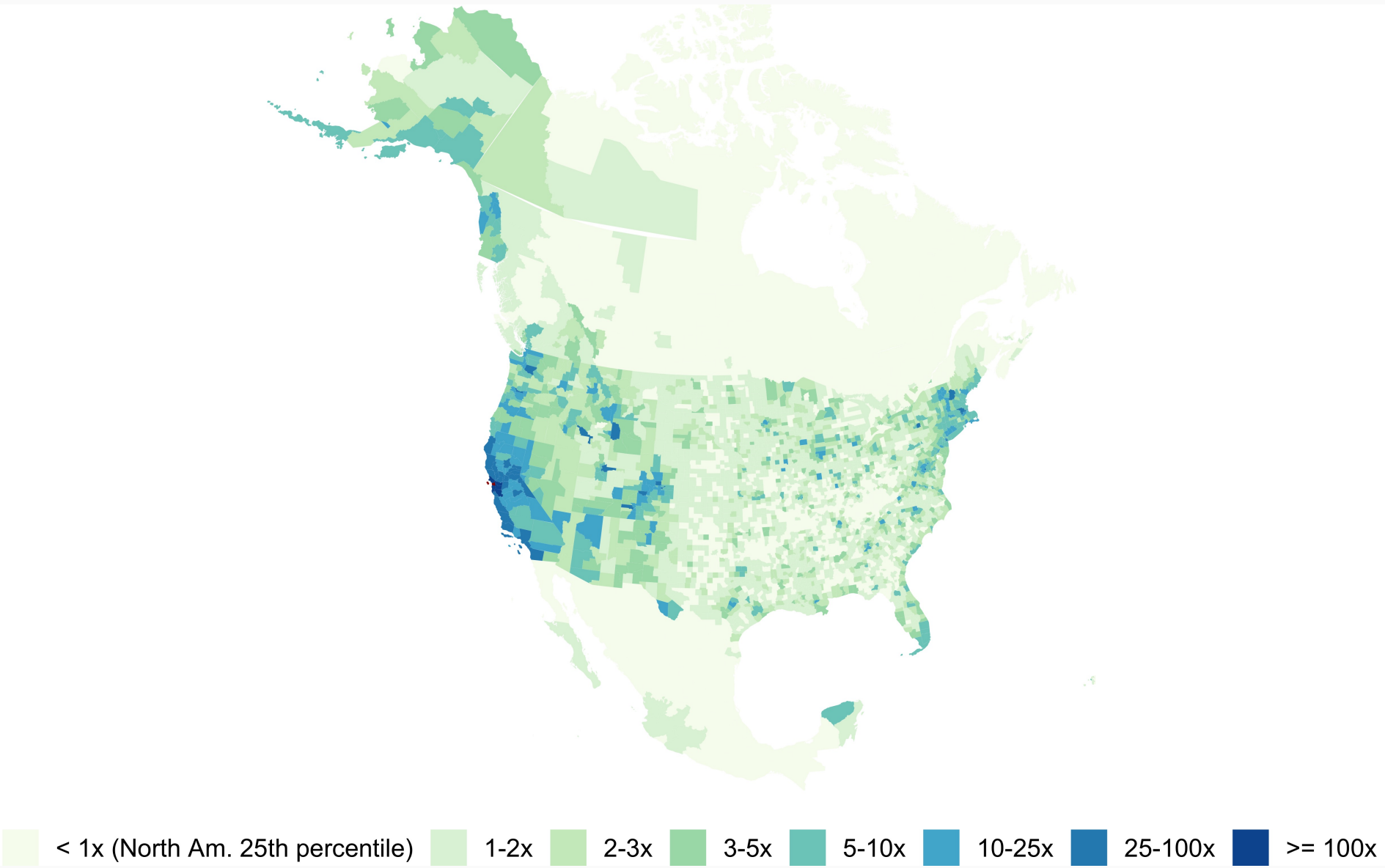
- *Social Connectedness Index* = unique measure to quantify strength of social connections that are important for disease spread
- Regions more connected to early hotspots in U.S. and Italy have, on average, more cases by March 30
- Inclusion of social proximity to cases improves predictions of COVID-19 spread in U.S., *above* physical proximity to cases
- Many opportunities for future research
  - Data available for U.S. counties; Europe NUTS3; GADM1 or GADM2 in much of the rest of the world
  - 1-page proposal to sci_data@fb.com (already 500+ research teams)

Other SCI examples: San Francisco & Kern Counties

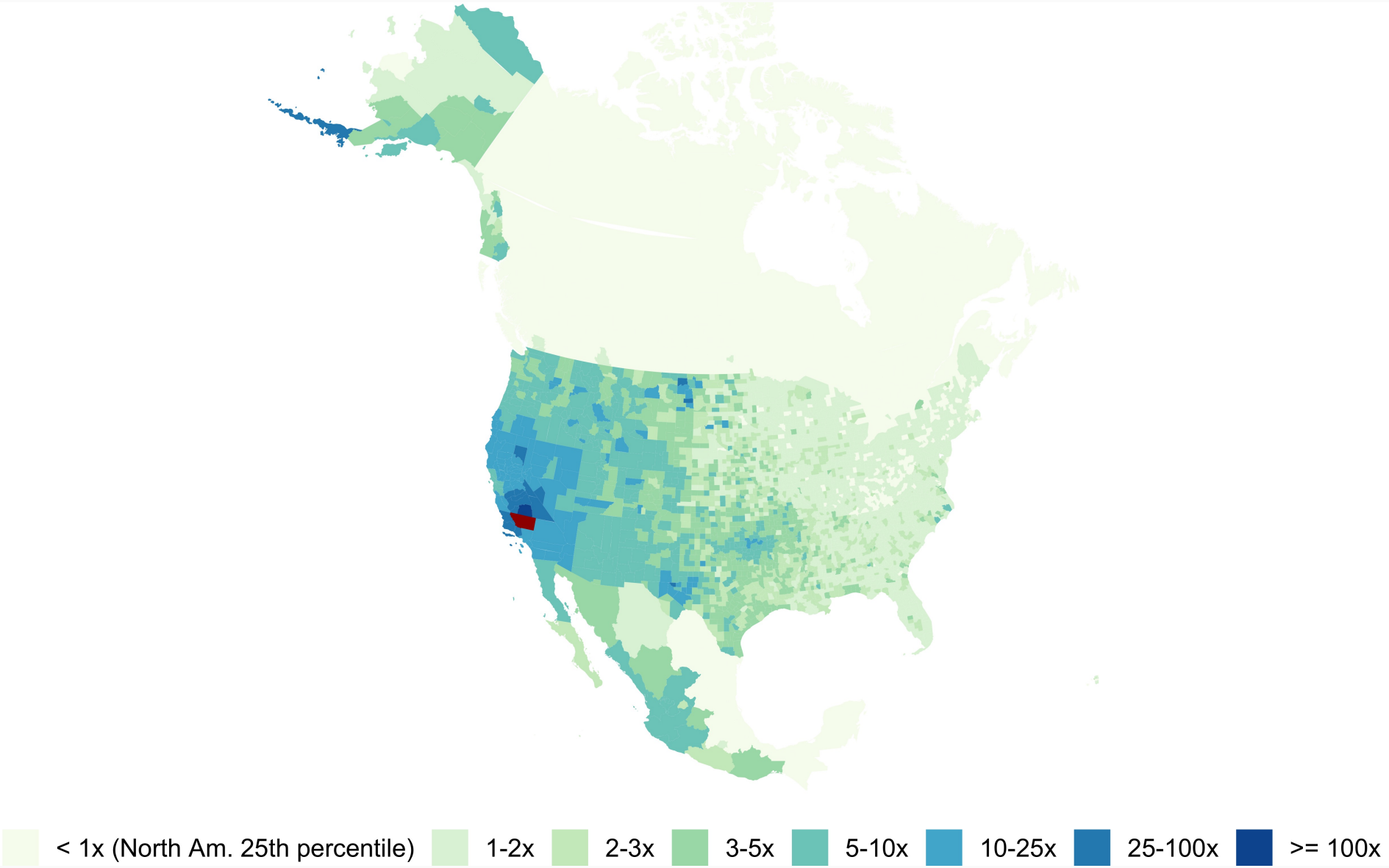
- San Francisco County, CA
  - Median Household Income: $72,947
  - Median age: 39 years
  - Share non-Hispanic White: 41.9%
  - Share Hispanic: 15.1%
  - Share Black: 6.1%
  - Share Asian: 33.3%

- Kern County, CA
  - Median Household Income: $48,021
  - Median Age: 32 years
  - Share non-Hispanic White: 49.5%
  - Share Hispanic: 38.4%
  - Share Black: 6.0%
  - Share Asian: 3.4%

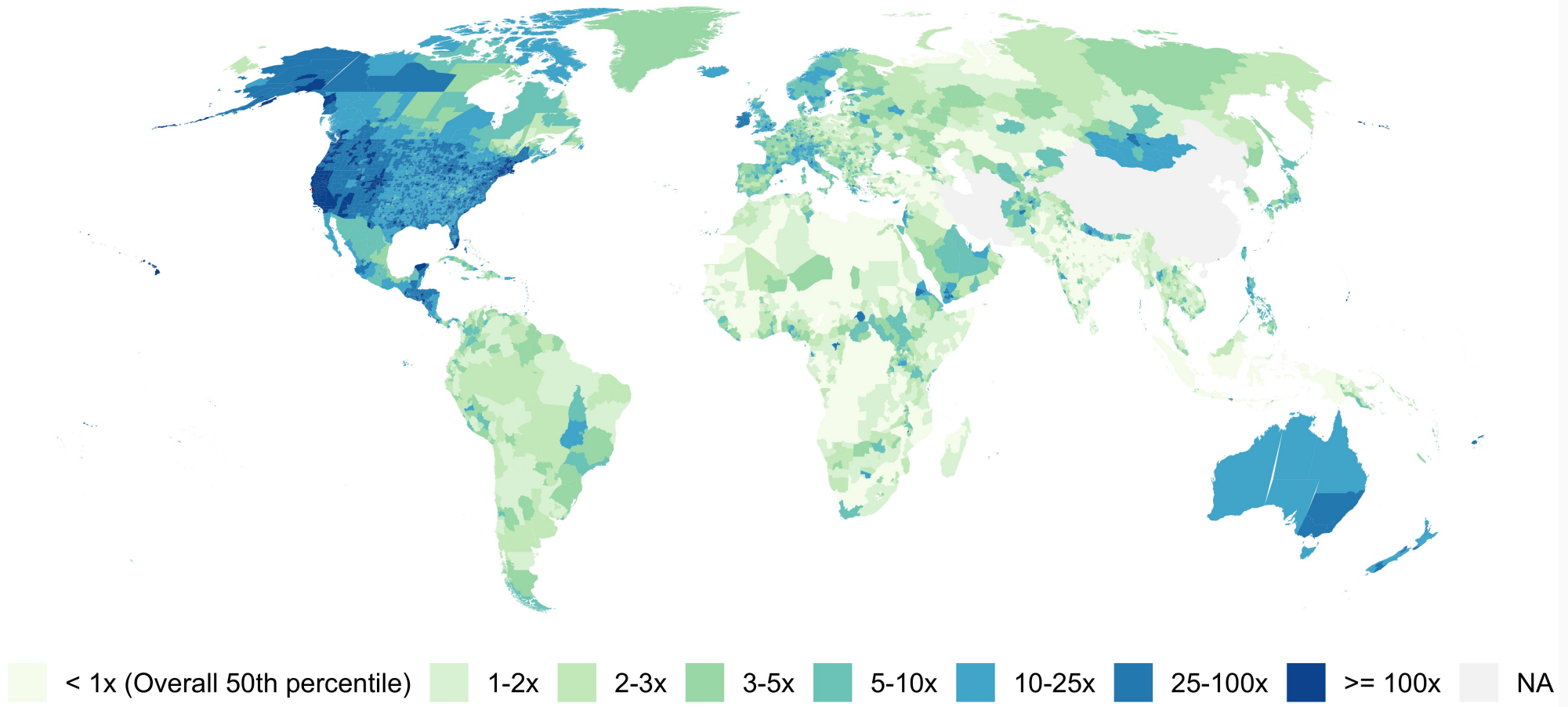
Other SCI examples: San Francisco County to North America



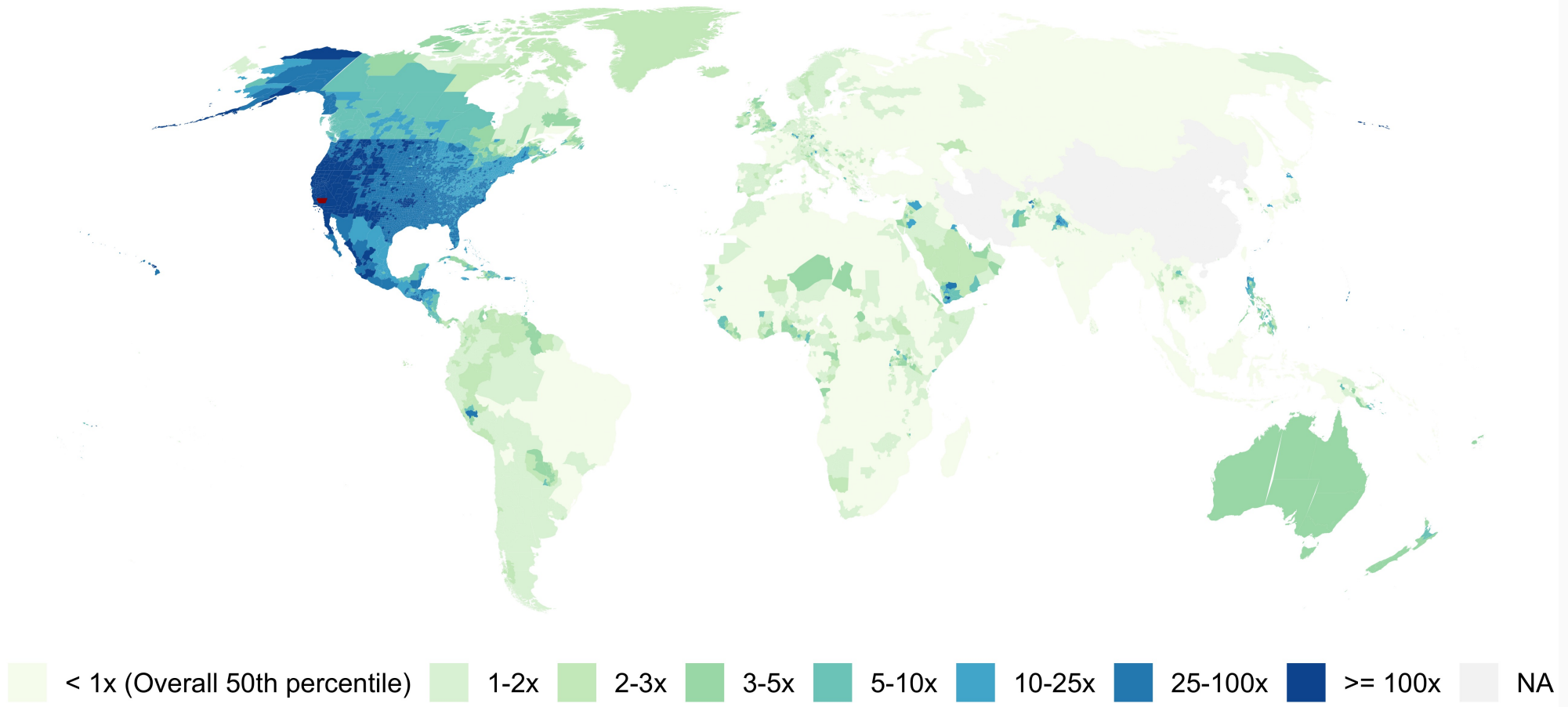
Other SCI examples: Kern County to North America



Other SCI examples: San Francisco County to the World



Other SCI examples: Kern County to the World



Other SCI examples: SCI in San Francisco & Kern Counties

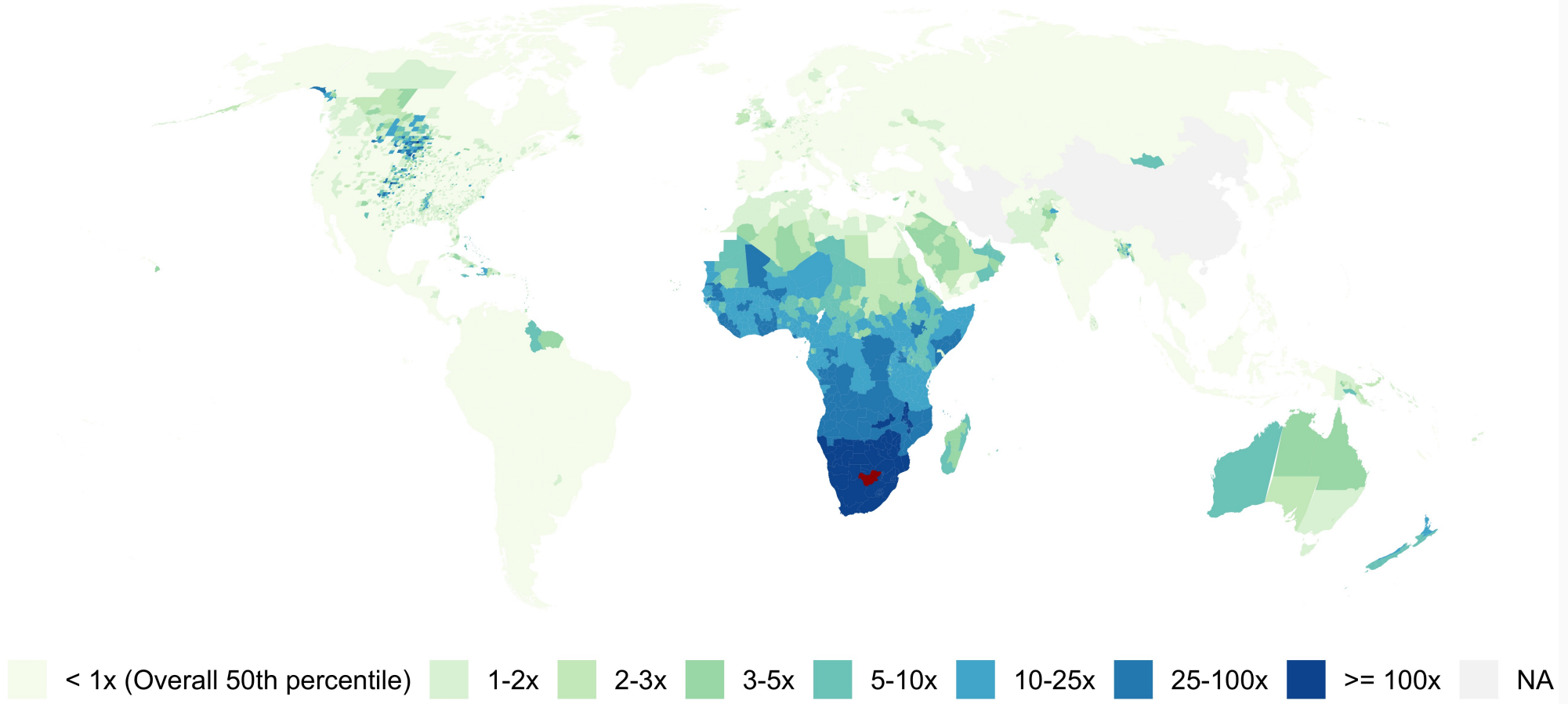
- San Francisco County, CA
  - Stronger connections to US east coast, western Europe (esp. Ireland), Australia, and Mongolia

- Kern County, CA
  - Stronger connections to western Mexico (consistent with large Hispanic population) and close-by areas in California
  - Connections to Oklahoma (Dust Bowl migration) and North Dakota (oil boom)
  - Generally less connected to rest of US and world

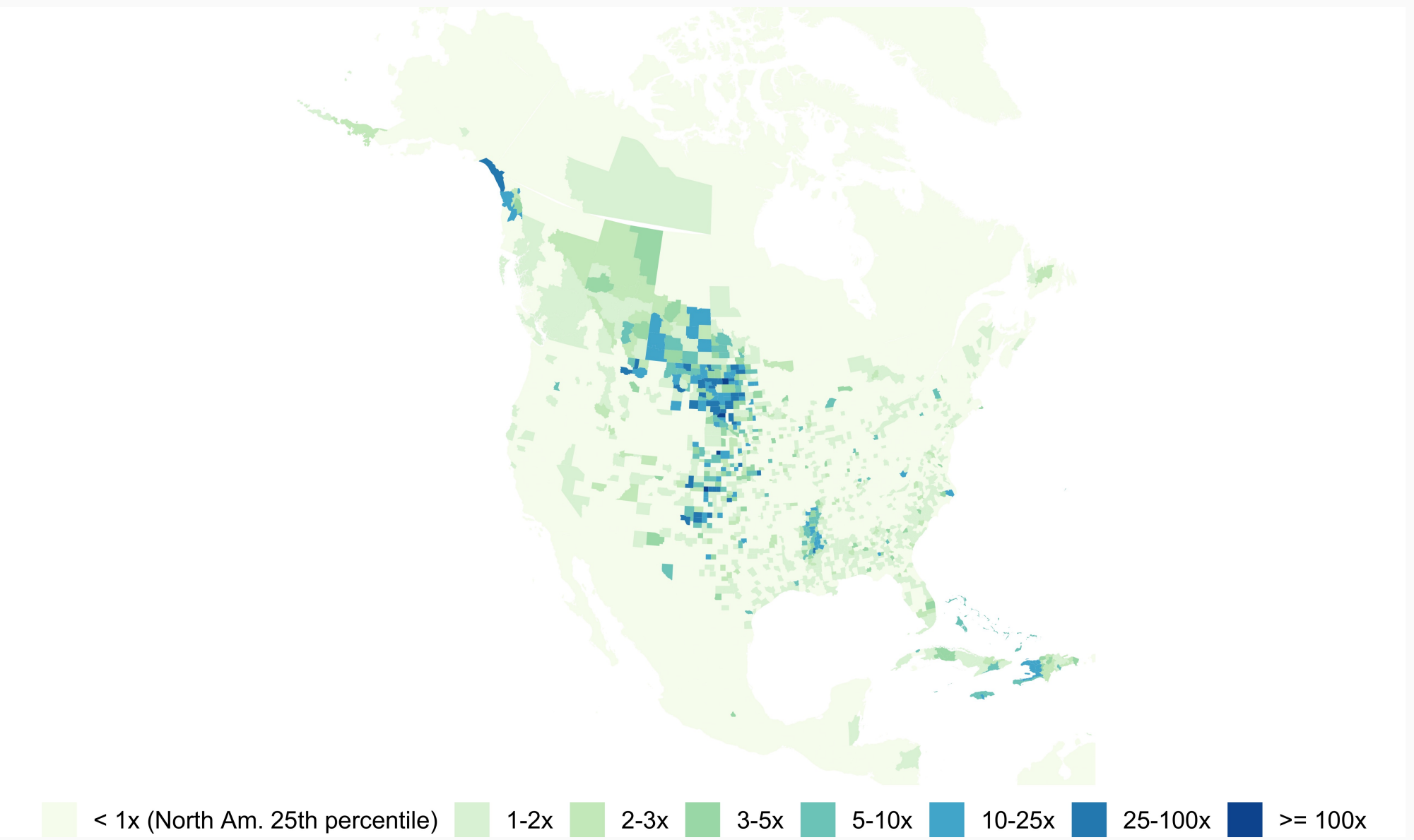
Other SCI examples: Gauteng, South Africa

- Includes Johannesburg and Pretoria
- 1% of South Africa land area, but 1/3 of GDP
- Major center of oil and gas production

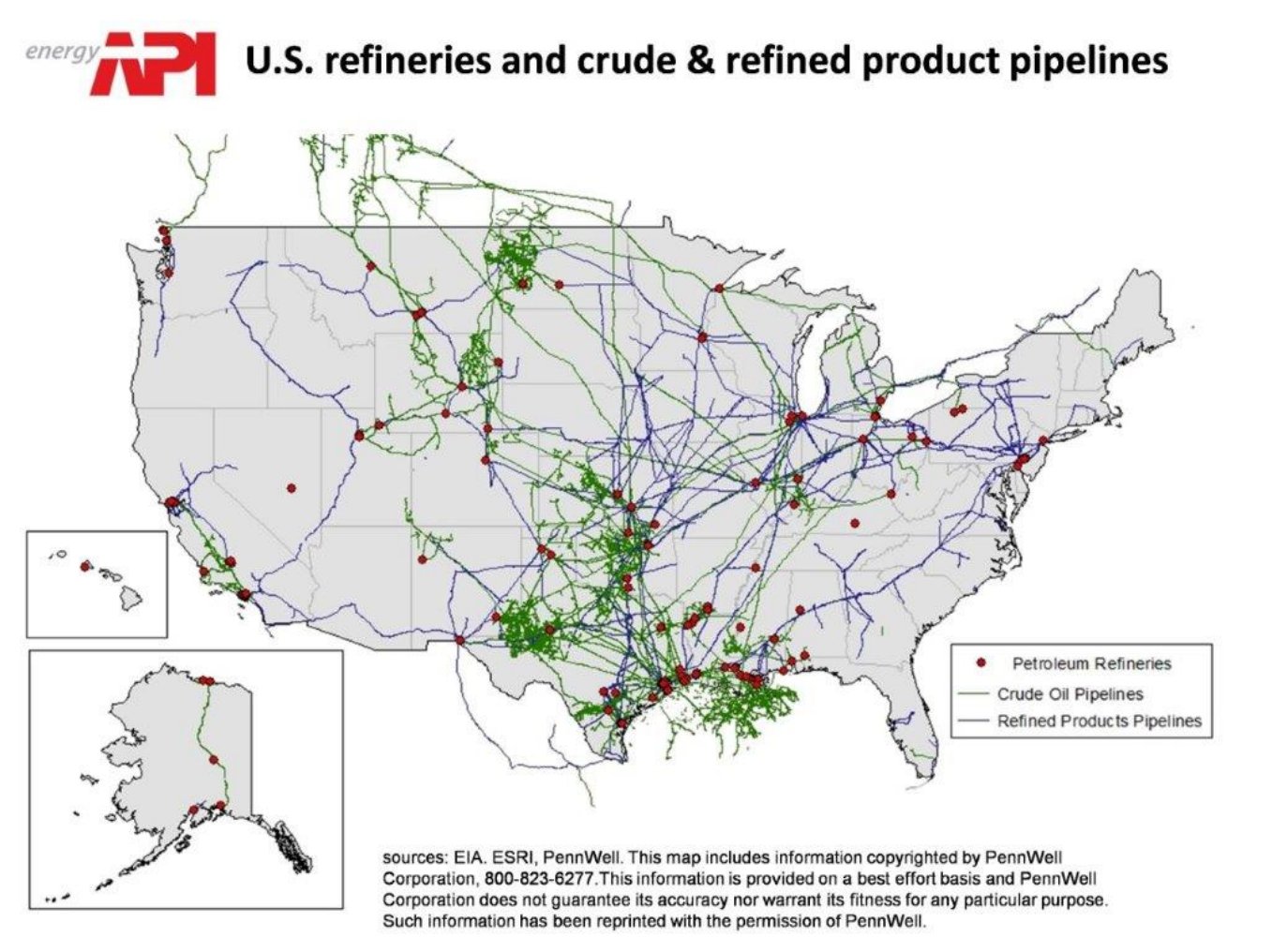
Other SCI examples: Gauteng, South Africa to the World



Other SCI examples: Gauteng, South Africa to North America



Other SCI examples: Oil Production in the United States



References

References


Bailey, Michael and Cao, Rachel and Kuchler, Theresa and Stroebel, Johannes and Wong, Arlene
Social connectedness: Measurements, determinants, and effects.
Journal of Economic Perspectives, 2018.


Coven, Joshua and Gupta, Arpit
Disparities in Mobility Responses to COVID-19.
Working Paper, 2020.


Piontti, Ana Pastore and Perra, Nicola and Rossi, Luca and Samay, Nicole and Vespignani, Alessandro
Charting the Next Pandemic: Modeling Infectious Disease Spreading in the Data Science Age.
Springer, 2018.