M12 Seminar: Economic and Social Problems: Insights from Big Data. Term Paper.

Replication and Extension of "The geographic spread of COVID-19 correlates with the structure of social networks as measured by Facebook"

by

Theresa Kuchler, Dominic Russel and Johannes Stroebel

Anton Koshelev

January 12, 2022

Table of Contents

Research Article Summary

Social Connectedness Index between locations i and j as a central element of the research paper: [1]

$$Social\ Connectedness_{i,j} = \frac{FB\ Connections_{i,j}}{FB\ Users_i * FB\ Users_j} \tag{1}$$

Research question: *Does Social Connectedness Index have a predictive power in the task of future communicable disease spread forecasting?*

- Initial COVID spread analysis (role of social connectedness to early hotspots in US and Italy)
- Social proximity to cases as a predictor of future growth in cases (panel setting)
- Out-of-sample predictive power of SCI-based factors

[1] Where $FB\ Connections_{i,j}$ is the total number of Facebook friendship links between locations i and j.

Data Sets

1. Social Connectedness Indices - *data.humdata.org*
2. COVID-19 daily data (cases and deaths) - Johns Hopkins University (US), Dipartimento della Protezione Civile (Italy)
3. US county-level demographics - American Community Survey and Opportunity Insights
4. EU NUTS3-level demographics - Eurostat
5. Urban-rural US county classification - National Center for Health Statistics
6. County-to-county (US) geographical distances - National Bureau of Economic Research
7. Google searches related to COVID-19 symptoms
8. Smartphone-based Location Exposure Index (LEX)

Identification Strategy - I

Early hotspot analysis:

- Dependent variable - cases per 10,000 people as of 30. March 2020
- Regressors:
  - SCI of locations to the early hotspot (Westchester county in US and Lodi province in Italy) + Controls [2]

Panel regression:

- Dependent variable - cases per 10,000 people in location i in (2-week) period t
- Regressors:
  - Cases per 10,000 people in location i in $t-1$ and $t-2$
  - Social proximity to cases in $t-1$ and $t-2$
  - Share of Facebook friends located within 50 and 150 miles from location i
  - Physical proximity to cases in $t-1$ and $t-2$
  - Controls

[2] Exclusion of nearest locations, geographical distance to the hotspot, median income, population density, rural/urban indicators.

Identification Strategy - II

Out-of-sample prediction (time series cross validation): [3]

- Dependent variable - cases per 10,000 people in location i in (2-week) period t
- Regressors:
  - Baseline explanatory variables [4]
  - Smartphone-based Location Exposure Index (lagged logs)
  - Google searches related to COVID-19 symptoms (lagged logs)
  - Social proximity to cases (lagged logs)

[3] Regression model - random forest (*n_trees=500*)

[4] Population density, median household income, lagged logs of changes in cases in county i and lagged logs of changes in physical proximity to deaths of county i.

Main Findings - Early Hotspot Analysis - USA

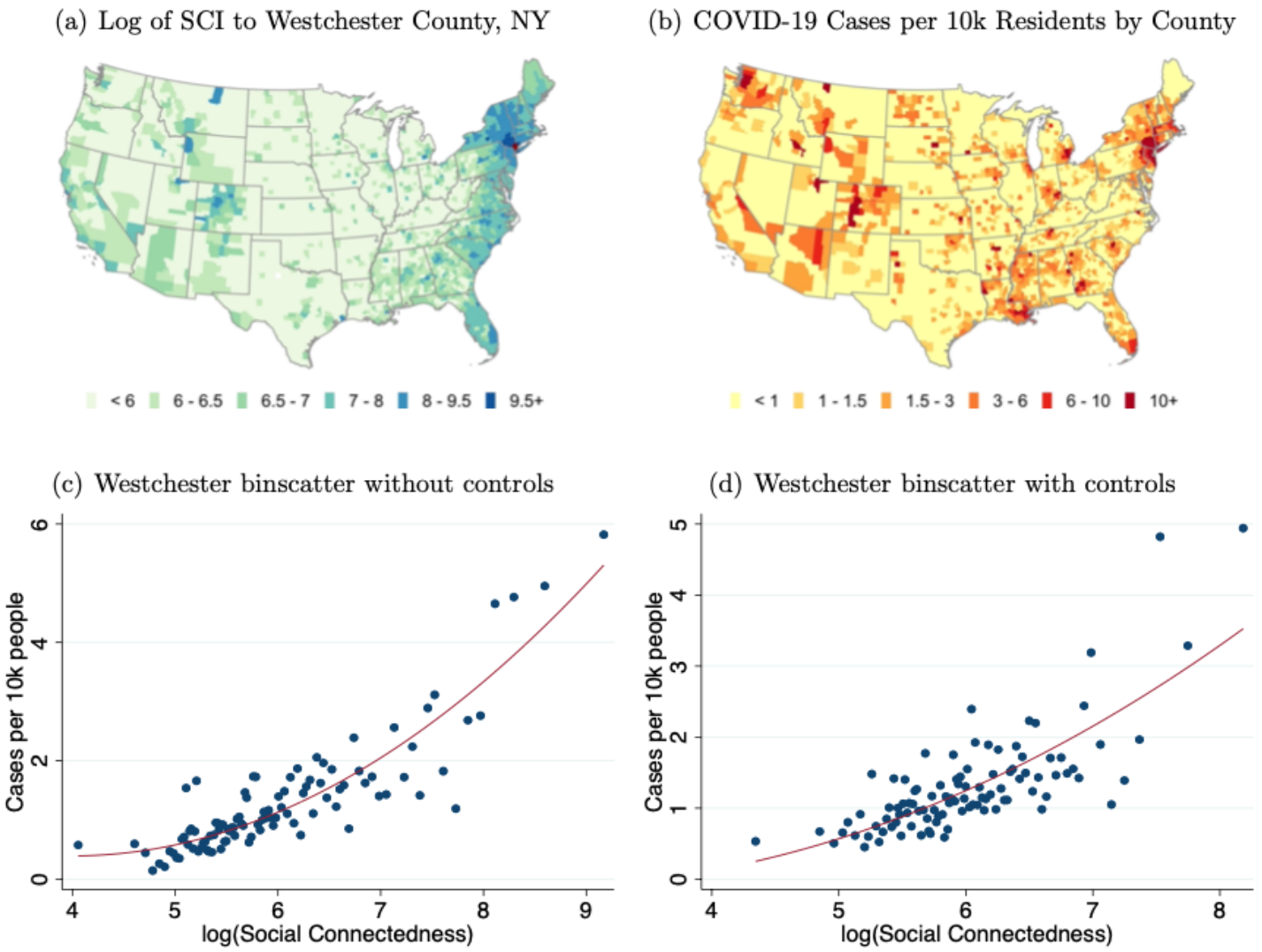

Figure 1: Social Network Distributions from Westchester and COVID-19 Cases in the U.S.

Main Findings - Early Hotspot Analysis - Italy

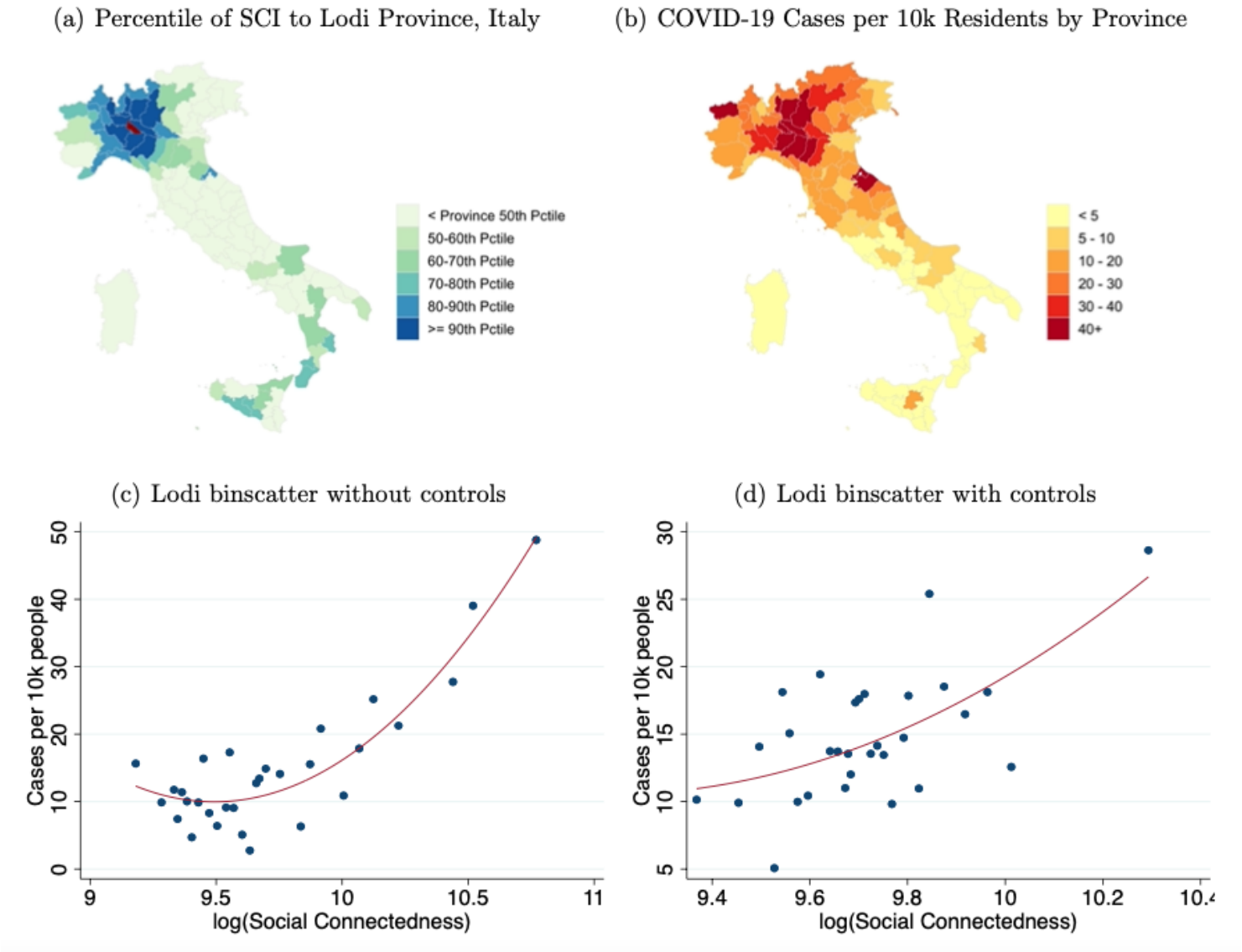

Figure 2: Social Network Distributions of Lodi and COVID-19 Cases in Italy

Main Findings - Panel Regression

Panel A	log(Change in Cases per 10k Residents + 1)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
2 Week Lag: log(Change in Social Proximity to Cases + 1)	0.589*** (0.041)	0.415*** (0.036)					0.414*** (0.041)	0.321*** (0.037)
4 Week Lag: log(Change in Social Proximity to Cases + 1)	-0.124*** (0.037)	-0.080** (0.032)					-0.002 (0.036)	0.010 (0.032)
Share of Friends within 50 Miles			0.096 (0.106)	0.031 (0.086)			0.050 (0.100)	0.076 (0.082)
Share of Friends within 150 Miles			0.018 (0.123)	0.214* (0.113)			-0.256** (0.124)	0.143 (0.109)
2 Week Lag: log(Change in Physical Proximity to Cases + 1)					1.432*** (0.129)	1.754*** (0.184)	1.244*** (0.118)	1.388*** (0.176)
4 Week Lag: log(Change in Physical Proximity to Cases + 1)					-1.208*** (0.131)	-1.433*** (0.196)	-1.037*** (0.121)	-1.225*** (0.187)
2 Week Lag: log(Change in Cases per 10k Residents + 1)	0.317*** (0.022)	0.316*** (0.018)	0.646*** (0.012)	0.526*** (0.011)	0.604*** (0.011)	0.514*** (0.010)	0.372*** (0.022)	0.351*** (0.019)
4 Week Lag: log(Change in Cases per 10k Residents + 1)	0.113*** (0.019)	0.092*** (0.016)	0.077*** (0.009)	0.063*** (0.008)	0.097*** (0.009)	0.072*** (0.008)	0.071*** (0.019)	0.056*** (0.017)
Time x Pop. Density FEs	Y	Y	Y	Y	Y	Y	Y	Y
Time x Median Household Income FEs	Y	Y	Y	Y	Y	Y	Y	Y
Time x State FEs		Y		Y		Y		Y
Sample Mean	2.177	2.177	2.177	2.177	2.177	2.177	2.177	2.177
R-Squared	0.717	0.755	0.706	0.752	0.718	0.754	0.725	0.757
N	47,040	47,025	47,040	47,025	47,040	47,025	47,040	47,025

Figure 3: COVID-19 Case Growth and Prior Proximity to Cases

Main Findings - Out-of-Sample Forecasting

	RMSE: Baseline Model			RMSE: Best Available Model			RMSE: Counties w/ Google + LEX Only		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Without Social Proximity	With Social Proximity to Cases	Diff. from Social Proximity	Without Social Proximity	With Social Proximity to Cases	Diff. from Social Proximity	Without Social Proximity	With Social Proximity to Cases	Diff. from Social Proximity
(1) April 14 - April 27	1.636	1.534	-0.102	1.488	1.387	-0.102	1.399	1.299	-0.100
(2) April 28 - May 11	0.900	0.838	-0.062	0.954	0.889	-0.066	0.887	0.835	-0.053
(3) May 12 - May 25	0.746	0.722	-0.024	0.771	0.746	-0.025	0.671	0.646	-0.025
(4) May 26 - June 8	0.704	0.680	-0.024	0.687	0.675	-0.012	0.584	0.581	-0.003
(5) June 9 - June 22	0.800	0.776	-0.024	0.779	0.766	-0.013	0.669	0.660	-0.010
(6) June 23 - July 6	0.859	0.838	-0.021	0.809	0.798	-0.011	0.665	0.667	0.002
(7) July 7 - July 20	0.793	0.780	-0.013	0.733	0.730	-0.003	0.530	0.526	-0.004
(8) July 21 - Aug. 10	0.755	0.719	-0.036	0.725	0.701	-0.024	0.508	0.509	0.002
(9) Aug. 11 - Aug. 24	0.770	0.740	-0.030	0.741	0.720	-0.022	0.530	0.517	-0.014
(10) Aug. 25 - Sep. 7	0.725	0.719	-0.005	0.728	0.722	-0.006	0.503	0.503	0.000
(11) Sep. 8 - Sep. 21	0.699	0.691	-0.008	0.694	0.686	-0.009	0.495	0.494	-0.001
(12) Sep. 22 - Oct. 5	0.748	0.719	-0.029	0.726	0.705	-0.021	0.513	0.511	-0.002
(13) Oct. 6 - Oct. 19	0.688	0.662	-0.026	0.684	0.658	-0.025	0.475	0.479	0.004
(14) Oct. 20 - Nov. 2	0.667	0.652	-0.015	0.647	0.628	-0.018	0.462	0.455	-0.007

Figure 4: Predicting COVID-19 cases in U.S., with and without Social Proximity to Cases

Strengths and Weaknesses

Strengths:

- SCI - granular and easily accessible variable
- Extensive robustness checks ("transfer learning" approach)
- Out-of-sample forecasting technique

Weaknesses:

- Additional (potentially powerful) factors missing in regressions (e.g. estimated parameters from SIR-type models)

Replication - Early Hotspot Analysis I

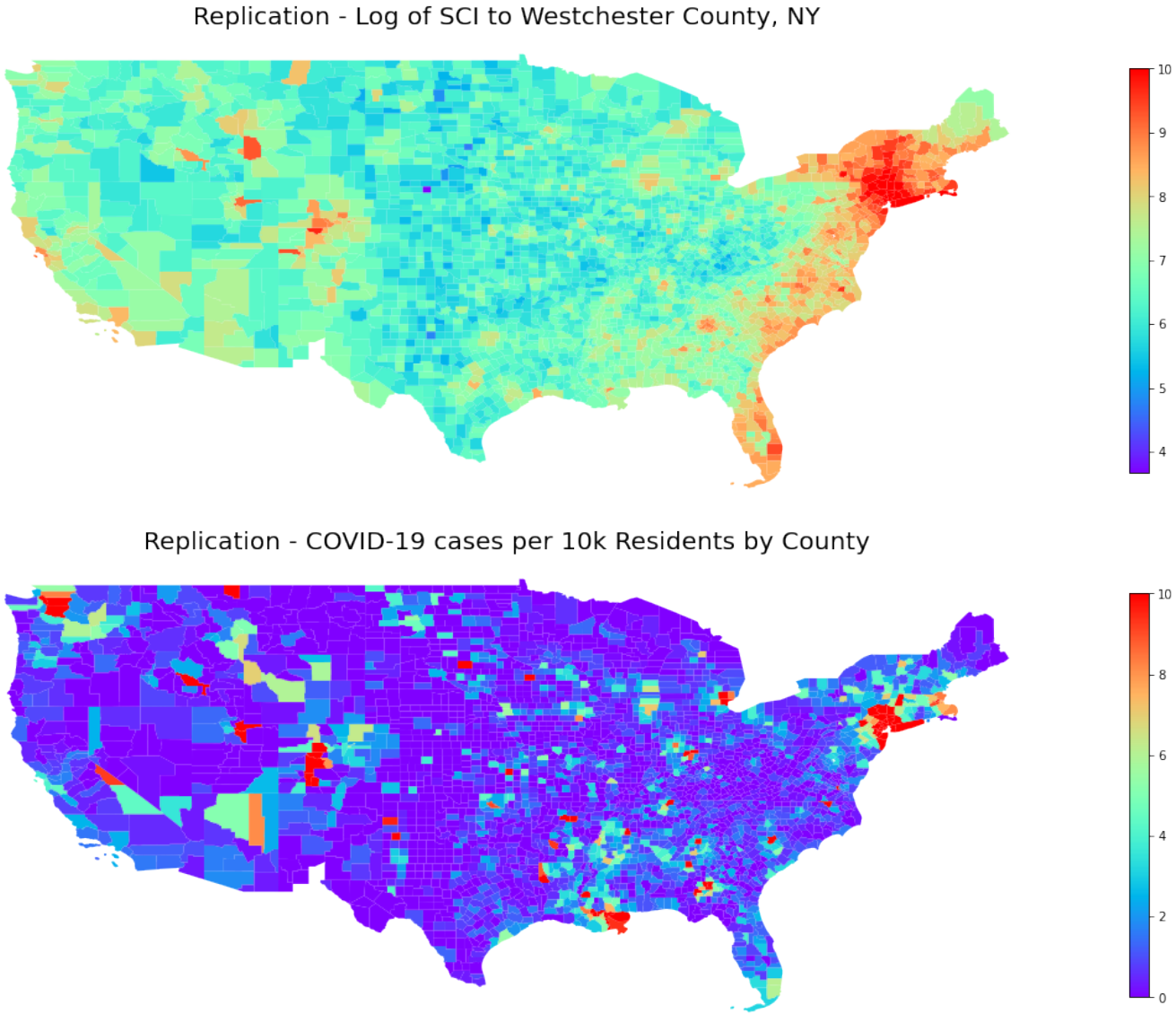

Figure 5: Replication - US heatmaps

Replication - Early Hotspot Analysis II

Figure 6: Replication - Westchester binscatters with (right) and without (left) controls, 95% CI

Replication - Early Hotspot Analysis III

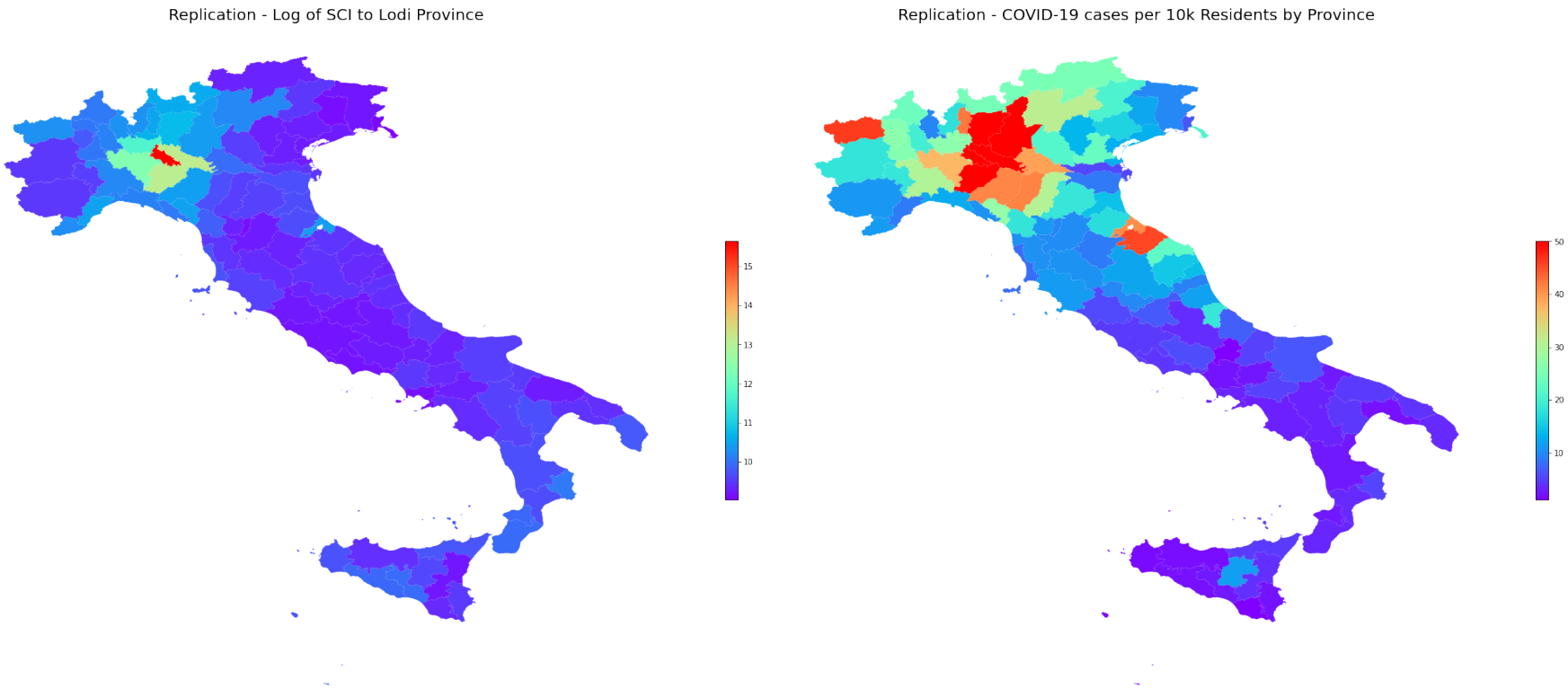

Figure 7: Replication - Replication - Italy heatmaps

Replication - Early Hotspot Analysis IV

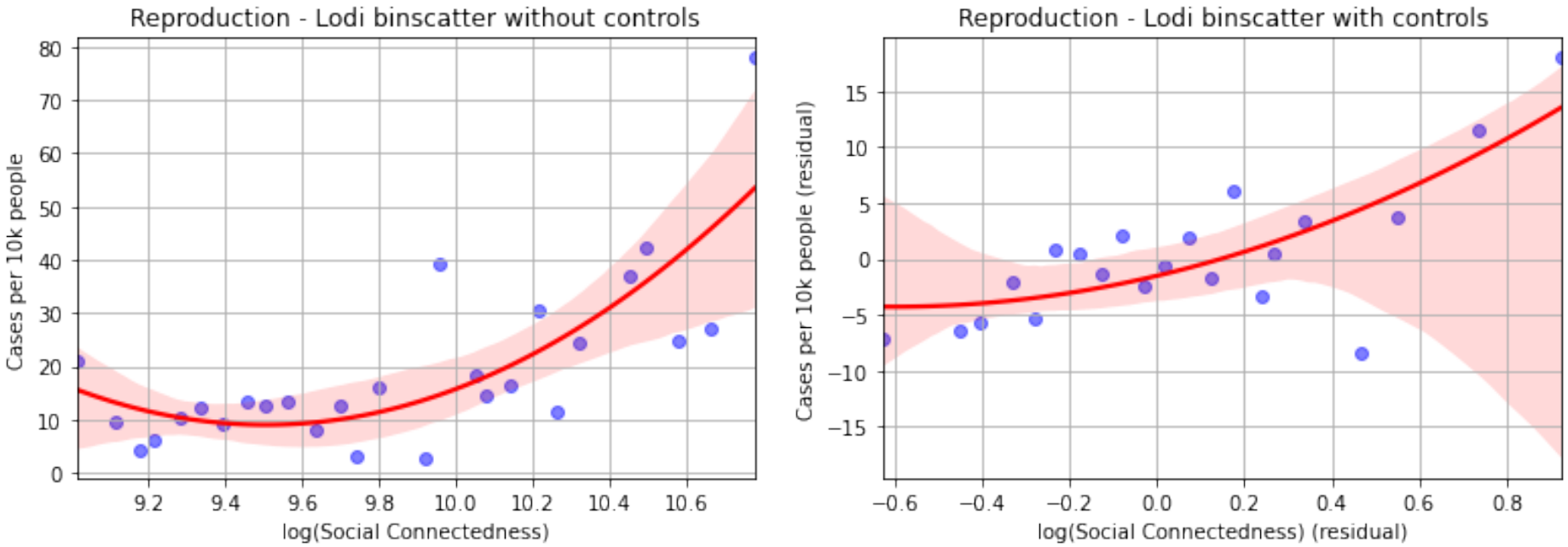

Figure 8: Replication - Lodi binscatters with (right) and without (left) controls, 95% CI

Replication - Panel Regression

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\log(\Delta\text{Social Proximity to Cases})_{i,t-1}$	0.75***	0.737***	-	-	-	-	0.48***	0.48***
$\log(\Delta\text{Social Proximity to Cases})_{i,t-2}$	-0.3***	-0.3***	-	-	-	-	-0.024	-0.024
Share Friends within 50 mi_i	-	-	0.034	0.034	-	-	0.628***	0.628***
Share Friends within 150 mi_i	-	-	0.603***	0.604***	-	-	-0.727***	-0.726***
$\log(\Delta\text{Physical Proximity to Cases})_{i,t-1}$	-	-	-	-	0.358***	0.366***	0.351***	0.351***
$\log(\Delta\text{Physical Proximity to Cases})_{i,t-2}$	-	-	-	-	-0.244***	-0.245***	-0.301***	-0.301***
$\log(\Delta\text{Cases per}10k+1)_{i,t-1}$	0.326***	0.332***	0.716***	0.716***	0.668***	0.666***	0.406***	0.406***
$\log(\Delta\text{Cases per}10k+1)_{i,t-2}$	0.145***	0.147***	0.046***	0.046***	0.084***	0.082***	0.062***	0.062***
Time x Pop.Density FE	X	X	X	X	X	X	X	X
Time x Med.HH.Inc FE	X	X	X	X	X	X	X	X
Time x State FE	-	X	-	X	-	X	-	X

Figure 9: Replication - COVID-19 case growth and prior proximity to cases (panel A). Dependent variable - $\log(\Delta \text{ Cases per } 10k+1)_{i,t}$

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\log(\Delta\text{Social Proximity to Deaths})_{i,t-2}$	0.564***	0.541***	-	-	-	-	0.424***	0.425***
$\log(\Delta\text{Social Proximity to Deaths})_{i,t-4}$	-0.025	-0.026	-	-	-	-	0.04**	0.042**
Share Friends within 50 mi_i	-	-	0.097***	0.1***	-	-	0.133***	0.142***
Share Friends within 150 mi_i	-	-	0.15***	0.145***	-	-	-0.004	-0.016
$\log(\Delta\text{Physical Proximity to Deaths})_{i,t-2}$	-	-	-	-	0.166***	0.168***	0.065***	0.064***
$\log(\Delta\text{Physical Proximity to Deaths})_{i,t-4}$	-	-	-	-	-0.04***	-0.04***	-0.048***	-0.049***
$\log(\Delta\text{Deaths per}10k+1)_{i,t-2}$	0.17***	0.181***	0.511***	0.511***	0.475***	0.475***	0.228***	0.228***
$\log(\Delta\text{Deaths per}10k+1)_{i,t-4}$	-0.005	-0.001	0.018***	0.018***	0.022***	0.021***	-0.022*	-0.023*
Time x Pop.Density FE	X	X	X	X	X	X	X	X
Time x Med.HH.Inc FE	X	X	X	X	X	X	X	X
Time x State FE	-	X	-	X	-	X	-	X

Figure 10: Replication - COVID-19 deaths growth and prior proximity to deaths (Panel B). Dependent variable - $\log(\Delta \text{ Deaths per } 10k+1)_{i,t}$

Replication - Out-of-Sample Prediction

Week #	Baseline	Baseline+LEX+Google
1	0.206488	0.105697
2	-0.000202	-0.016455
3	-0.000089	-0.023757
4	-0.000095	-0.007993
5	-0.001234	-0.007884
6	-0.014858	0.001694
7	-0.019670	-0.001087
8	-0.018778	-0.011043
9	0.011717	0.016560
10	-0.002559	0.000634
11	-0.005591	-0.000101
12	-0.021270	-0.000177
13	-0.015952	-0.000777
14	-0.013399	-0.000221

Table 1: Replication - Predicting COVID-19 cases in U.S., with and without Social Proximity to Cases (RMSE differences after SCI data are included).

Extension I - Gradient Boosting for Out-of-Sample Prediction

Week #	Baseline Data	Baseline+LEX+Google Data
1	-0.081870	0.124046
2	-0.021301	0.086822
3	-0.054955	0.131332
4	-0.022009	0.214086
5	-0.018157	0.003305
6	-0.011928	-0.017522
7	0.028709	-0.081620
8	0.019093	0.013898
9	-0.020993	0.014471
10	-0.007170	-0.003378
11	-0.010163	-0.174044
12	-0.032154	-0.022057
13	-0.002846	-0.211507
14	0.009485	0.060008

Table 2: Gradient Boosting - Predicting COVID-19 cases in U.S., with and without Social Proximity to Cases (RMSE differences after SCI data are included).

Extension II - Gradient Boosting vs Random Forest

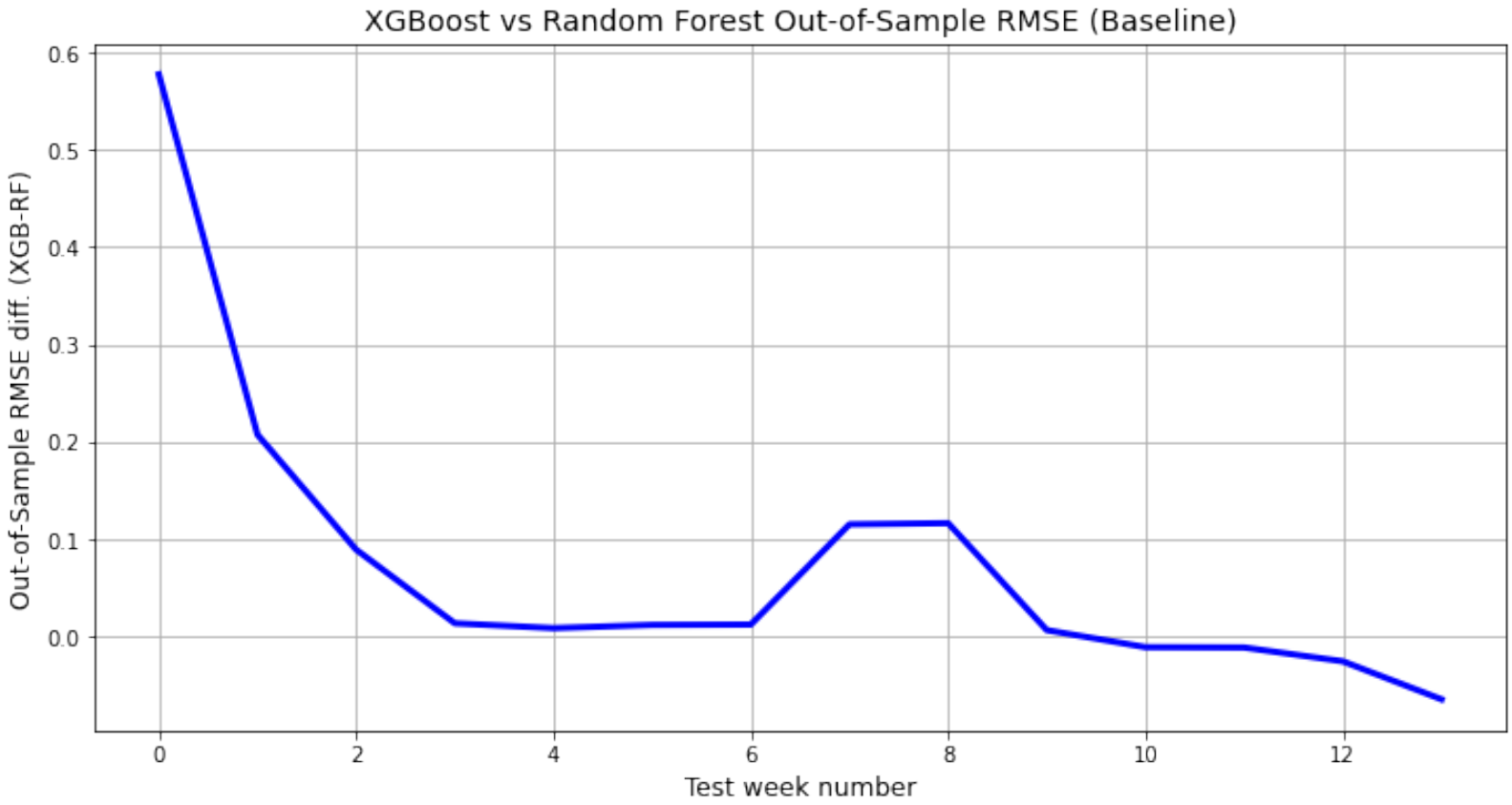

Figure 11: Relative out-of-sample RMSE dynamics, XGBoost vs Random Forest

M12 Seminar: Economic and Social Problems: Insights from Big Data. Term Paper.

Replication and Extension of "The geographic spread of COVID-19 correlates with the structure of social networks as measured by Facebook"

by

Theresa Kuchler, Dominic Russel and Johannes Stroebel

Anton Koshelev

January 12, 2022