The geographic Spread of COVID-19 correlates with the Structure of Social Networks as measured by Facebook

Group No : 25

Presented By :

Mallula Rajesh
Bolisetti Bhanuteja
Karanam Tejendhar
Nandyala Avinash

Introduction

- The agenda of this paper is to show that-
  "Data from Online Social Networks can be useful to forecast the spread of communicable diseases like COVID-19".

- To predict the spread , we need to know, which individuals are more likely to physically interact.
- Social ties shape the pattern of Physical interaction.
- For example, counties with higher levels of social connectedness to New York were more likely to be the destinations for those fleeing the city during the pandemic.
- Even the social connectedness is largely related to travel patterns across regions

Why the study?

- ❖ To show the usefulness to include Social connectedness measure in addition to other factors like geographic distance , income , population etc.
- ❖ Social network data is largely available.

Active body of Research

- ❖ In addition to the study , there are other research studies on “How different aspects of social media and internet-usage patterns can be used for tracking and preventing disease
- ➔ Tracking individual moments by their internet searches and social activity.
- ➔ Using Twitter ,Instagram, facebook posts and likes to predict public health outcomes
- ➔ Using surveys and other crowdsourced info to monitor disease symptoms.

Data Description

- We used de-identified and aggregated snapshot of all active Facebook users and their friendship networks.
- The locations of the users are identified based on their activity on facebook and device information.

Why only FB. Why not Twitter?

- Facebook connections are generally more likely to be between real-world acquaintances than links on many other social networking platforms like Twitter.
- Most of Twitter users have connections to top celebrities rather than actual friends.

Social connectedness measure

❖ Given two locations ‘i’ and ‘j’:

$$Social\ Connectedness_{i,j} = FB\ Connections_{i,j} \,/\, (\ FB\ Users_i * FB\ Users_j\ )$$

❖ SCI actually measures the relative probability of Facebook friendship link between a given Facebook user in location ‘i’ and a given Facebook user on location ‘j’

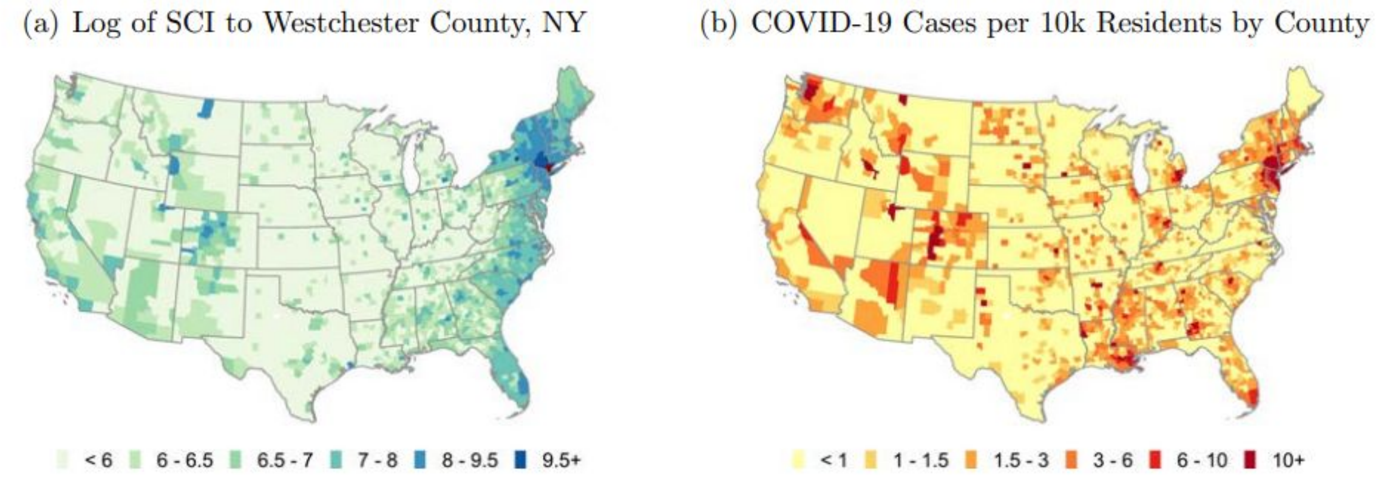
What we try to do

- ❖ We prove the correlation of spread with SCI in the early pandemic time taking covid-19 hotspots as reference
- ❖ We also see the amount/effect of correlation in long time as the pandemic continues (April - July)
- ❖ Later , we provide a naive model for predicting actual cases taking SCI into account
- ❖ Later , we see the limitations and provide the real examples where SCI has no correlation at all

Early Hotspot Analysis

- ❖ Analysing results on a covid hotspot - Westchester , NY , US.
- ❖ The relation of SCI of other regions to Westchester , to covid spread as of March 30, 2020.

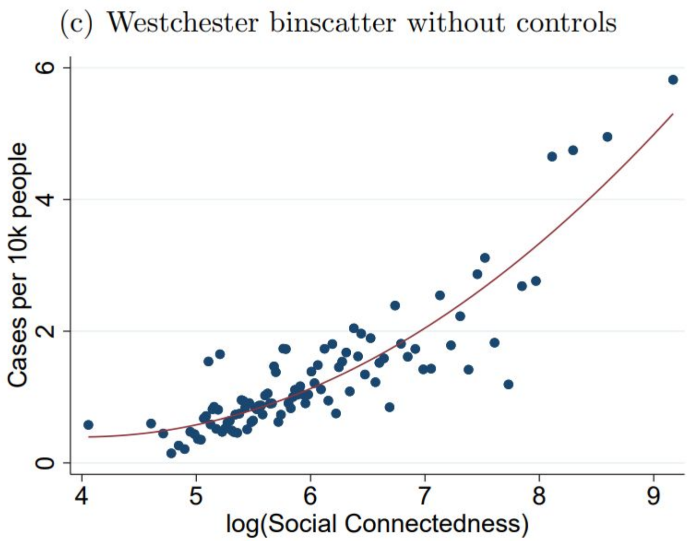
Figure 1: Social Network Distributions from Westchester and COVID-19 Cases in the U.S.



Early Hotspot Analysis

- We analyze the relation in US more formally by using binscatter plots.
- We remove the counties within 50 miles from Westchester , in order to avoid the dependance of geographic distance to covid spread.
- As there are several regions, we group their Log(SCI) values into 100 equal-sized bins and calculate average

Fig. Analysis of US wrt Westchester



Early Hotspot Analysis - Results

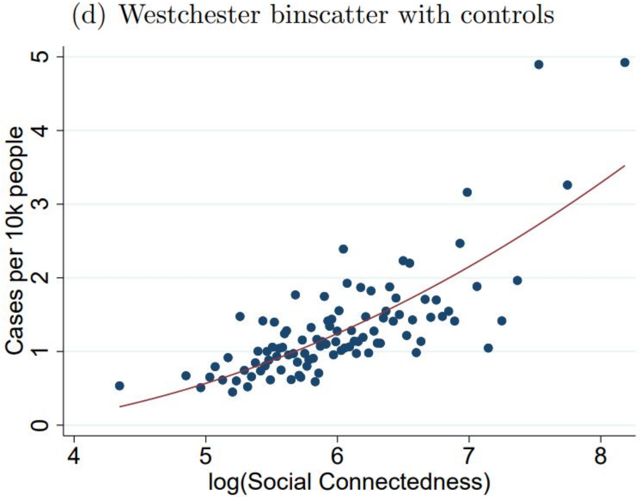
- From the analysis of US- Quantitatively, double the county's social connectedness with Westchester resulted in increase on 0.88 covid-19 cases
- The R-Squared value of this relation is 0.093 i.e 9.3 % region variation in cases can be explained by the region's SCI to Westchester

Controlling other factors

- The above results can not solely explain the importance of SCI , as other factors like Geographic distance , Population Density , Income are also positively correlated with Social Connectedness
- We control the geometric distance, Population Density, Income , GDP and other important effects of a region, so that these factors have no more considerable effect.
- Other factors may also be considered based on country culture , habits etc.

Results with controlling factors

- ❖ Even controlling the other factors, the results show a strong relation of SCI to covid spread.
- ❖ Statistically, double the SCI of a region to Westchester, an increase of about 0.80 cases is observed. The decline can be explained, as the cases also dependent on the controlled factors.
- ❖ A slight increase of about 0.097 (Total - 0.190) in R-Squared explains how influential SCI alone can be in predicting the cases

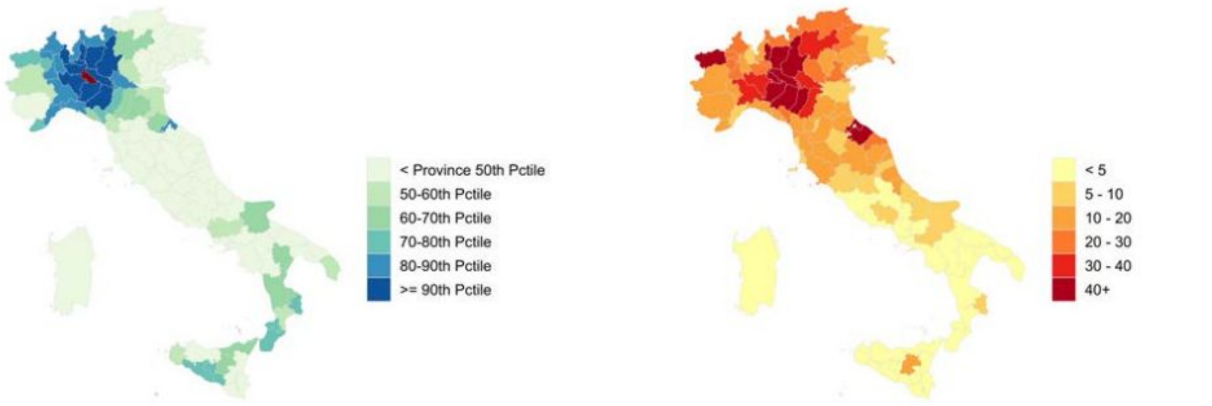


Early Hotspot Analysis

- ❖ Analysis of an Early covid hotspot - Lodi Province , Italy
- ❖ The cases are not disproportionately larger , perhaps reflecting the efforts of Italian authorities to restrict individuals movement.

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

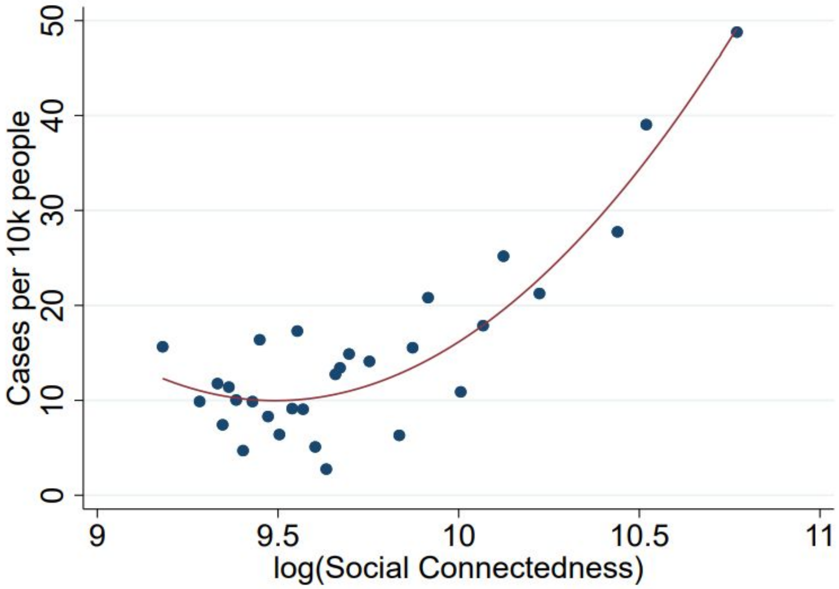
(a) Percentile of SCI to Lodi Province, Italy (b) COVID-19 Cases per 10k Residents by Province



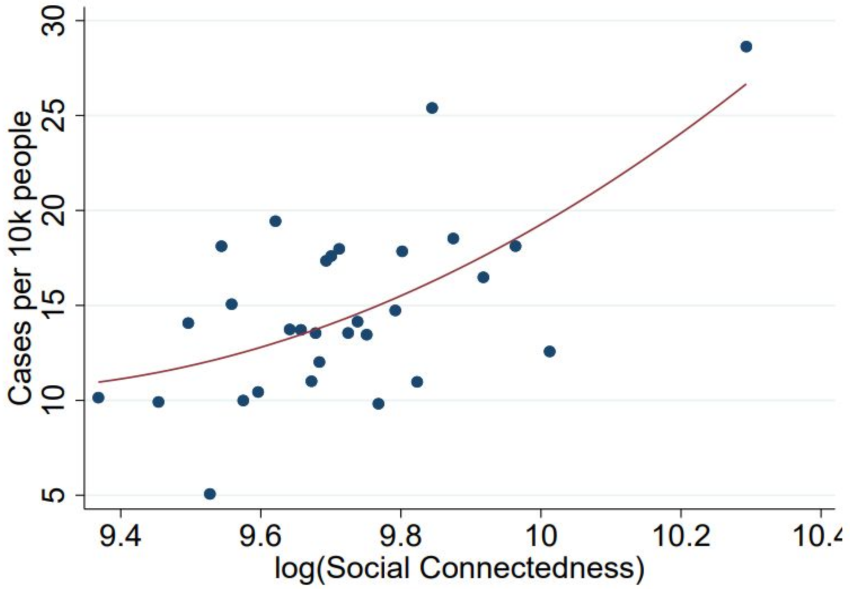
Italy Results

- ❖ Similar explanation to Italy. Analyzed on 30 equal sized bins and controlling GDP, Population, Geographic distance.
- ❖ Incremental R-squared relationship is 0.057 for italy.

(c) Lodi binscatter without controls



(d) Lodi binscatter with controls



Time Series Analysis

- In this section we will exploit how the pandemic is spreading in US.

- More systematically investigating the role of Social Connectedness Index(SCI) in forecasting the spread of covid19.

- The two metrics for this forecast are
  - Social Proximity to Cases:
    - Measure of exposure to COVID cases through social networks.
  - Physical Proximity to Cases:
    - measure of exposure to COVID through physical proximity.

Time Series Analysis

- ❖ The above two measures are related for shorter distance because individuals generally have strong social ties when they are geographically nearby. (source: Journal of Economic Perspectives, 32(3):259–80, 2018b)

- ❖ But when comes to geographically distant places (The Westchester and the East Coast of Florida) , these can have strong social ties and are would not be predicted by physical distance.

- ❖ Many other factors like social ties are also not be predicted by physical distance

- ❖ Here comes the predictive value added by the social connectedness data.

Key variable constructions

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

- The sums j and h are over all counties.
- Cases per 10k$_{j,t}$ is the number of confirmed COVID-19 cases per 10,000 residents in county j as of time t.

Key variable constructions

$$Physical\ Proximity\ to\ Cases_{i,t} = \sum_j Cases\ Per\ 10k_{j,t} * (1 / (1 + Distance_{i,j}))$$

❖ $Distance_{i,j}$ is the physical distance between counties i and j measured in miles

❖ $Cases\ per\ 10k_{j,t}$ is the number of confirmed COVID-19 cases per 10,000 residents in county j as of time t.

Empirical Specification

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

Social connectedness is an important predictor of the path of COVID-19 spread, a lagged measure of social proximity to new cases will have a positive relationship with new case counts in the next period

$X_{i,t}$ are a set of time-specific fixed effects, including population density and median household income.

Empirical Results

Table 1: COVID-19 Case Growth and Prior Proximity to Cases

	log(Change in Cases per 10k Residents + 1)					
2 Week Lag: log(Change in Social Proximity to Cases + 1)	0.592*** (0.071)		0.434*** (0.106)	0.437*** (0.043)		0.325*** (0.054)
4 Week Lag: log(Change in Social Proximity to Cases + 1)	-0.067 (0.050)		0.067 (0.084)	-0.077*** (0.020)		0.020 (0.029)
2 Week Lag: log(Change in Physical Proximity to Cases + 1)		1.266** (0.408)	1.054** (0.372)		1.622*** (0.163)	1.266*** (0.212)
4 Week Lag: log(Change in Physical Proximity to Cases + 1)		-1.170** (0.408)	-1.028** (0.374)		-1.287*** (0.264)	-1.092*** (0.305)
2 Week Lag: log(Change in Cases per 10k Residents + 1)	0.319*** (0.043)	0.635*** (0.022)	0.376*** (0.052)	0.330*** (0.032)	0.549*** (0.025)	0.376*** (0.038)
4 Week Lag: log(Change in Cases per 10k Residents + 1)	0.052 (0.032)	0.069*** (0.016)	0.008 (0.040)	0.079*** (0.012)	0.062*** (0.010)	0.040* (0.017)
Time X Pop Density FEs	Y	Y	Y	Y	Y	Y
Time X Median Household Income FEs	Y	Y	Y	Y	Y	Y
Time X State FEs				Y	Y	Y
Sample Mean	1.593	1.593	1.593	1.593	1.593	1.593
R-Squared	0.641	0.638	0.650	0.684	0.682	0.686
N	25,056	25,056	25,056	25,048	25,048	25,048

- ★ Each observation is a county, two-week period (between March 30 and July 20, 2020).
- ★ The dependent variable in all columns is log of one plus the number of new COVID-19 cases per 10,000 residents.
- ★ Columns 1 and 4 include log of growth in social proximity to cases lagged by two and four weeks (one and two time periods).
- ★ Columns 2 and 5 include analogous measures of physical proximity to cases.
- ★ Columns 3 and 6 include both social and physical measures.
- ★ All columns include controls for two-week and four-week lagged changes in cases, as well as time-specific fixed effects for percentiles of county population density and median household income.

Table:1 Results

- ❖ Growth in social proximity to cases in one period has a strong positive relationship with actual case growth in the next.
- ❖ We can see that both social and physical proximity have good importance on finding cases.
- ❖ Including state fixed effects interacted with week, doubling of social proximity to cases in one period corresponds to a 22.5% increase in actual cases per 10,000 residents in the next period. Which also showing the value of adding this variable in prediction.
- ❖ We can also observe that the physical proximity is more important than social proximity ,even though both contribute some parts each.
- ❖ We next conduct a similar analysis, i.e studying how the relationship between social connections and new COVID-19 cases changes over the course of the pandemic.

Empirical Results

Table 2: COVID-19 Case Growth and Prior Proximity to Cases, by Two-Week Period

	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

- ★ Each observation is a county.
- ★ The dependent variable is log of one plus the number of new COVID-19 cases per 10,000 residents in one two-week period between March 30 and July 20, 2020.
- ★ All columns include log of growth in social and physical proximity to cases, as well as actual cases, lagged by two and four weeks (one and two time periods)
- ★ Significance levels: *(p<0.10), **(p<0.05), ***(p<0.01).

Table:2 Results

- ❖ A one time period lagged measure of social proximity to cases was a statistically significant predictor of actual case growth.

- ❖ The magnitudes of the coefficients suggest that a doubling in social proximity to cases in one two-week period corresponds to between a 9.8% and 50.7% increase in actual cases in the next time period, after controlling for physical proximity to cases.

- ❖ We can observe that how our social connectedness variable varies from march to july and can observe its dependency.

Table:2 Results

- ❖ Columns 1 and 2 describes disease spread in March and the first days of April, the relationship is particularly strong. Because travelling may have been common before public recognition of outbreak.
- ❖ Example: Trips between Westchester and coastal Florida may have been common before public recognition of the outbreak, but relatively infrequent later.
- ❖ In the final four periods (columns 5- 8), the coefficients on social proximity again generally increase, corresponding to the time in which mobility began slowly returning toward baseline levels.
- ❖ Together, these results shows social proximity matters most when there are fewer restrictions on individuals mobility.
- ❖ This provides more evidence that social connectedness is predictive of interactions that spread communicable disease

Table - 3

	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

- Predicted Outcome should be log(1+Δcases per 10k)
- 1-3 show RMSE for Linear Regression and 4-6 show RMSE for Random Forest
- Inputs for Col 1 and Col 4 : Population Density,Median Household income,log(1+ΔPhysical Proximity) for two and four week lags.
- Inputs for Col 2 and Col 5 :They also add log(1+ΔSocial Proximity)
- For Col 3 and Col 6 : They are difference between (Col2-Col1) and (Col5-Col4)

Table-3 Results:

- **For Linear Regression (Col 1-3) :**
  - In the First to periods, Which include the most limited training data , RMSE would be pretty much higher for both models
  - As for Col3 we observe the last 5 rows, the RMSE is lower for including social proximity to the cases, So it make the predictions are improving subsequently.

- **For Random Forest (Col 4-6) :**
  - Here we use 500 regression trees as to know the non - linear relationship of data points.
  - Similarly, in addition ,including measures of Social Proximity leads to improvement of forecasts

Conclusion

- **Advantages by this paper :**
  - ➔ The social connected measures can be used effectively than other factors as there is huge data available.
  - ➔ These measures are robust to change in seasonality and trends in internet.
- **Limitations:**
  - ➔ Our Results should not be interpreted as a fixed model, Instead it only provides a tool for epidemiologists for forecasting future.
  - ➔ Not everything can be explained by SCI.
  - ➔ There are some data points where there is less Social Connectedness Index but more COVID cases .
  - ➔ **Ex:** King County, WA (Seattle) has low social connectedness where as it is early hotspot for COVID.
  - ➔ The geographic structure of social networks is difficult to measure on a national or global scale.