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
Anton.Koshelev@campus.lmu.de

December 22, 2021

Introduction

This term paper comprises a referee report on the paper by Kuchler et al. (2021), which illustrates the use of social connectedness index in context of COVID-19 global outbreak, as well as the paper's results replication and its extension. Authors of the paper show that measures of social proximity to cases/deaths are good determinants of future growth in cases and lethal outcomes. The significance of the SCI-based variables persists even after one includes other, more advanced, proxies of exposure to the virus and various controls. Social proximity to cases does not loose its predictive power even when transition to the out-of-sample forecasting is made, underlining importance of SCI for the future research on epidemiological modelling. All these conclusions can be also made from our replication of the paper, as we get results similar to Kuchler et al. (2021).

The term paper proceeds as follows: section 1 gives a referee report with the discussion of the research question, main findings, ideal experimental setting and data used in the research, as well as identification strategy and briefly reviews relevant literature; section 2 replicates main findings of Kuchler et al. (2021) in three dimensions: early hotspot analysis, panel regression setting and out-of-sample prediction; section 3 makes a modest contribution to the topic and shows the performance of gradient boosting in out-of-sample forecasting, which also improves after inclusion of SCI-based features; section 4 concludes [1].

[1]The data and the original code (R, Stata) is provided by Kuchler et al. (2021) in the project GitHub repository. Replication of the paper's main results (Python) can be found in the following GitHub repository

1 Referee Report

1.1 Research Question and Main Findings

The main research question Kuchler et al. (2021) aim to answer can be formulated as *"Does Social Connectedness Index (SCI) have a predictive power in the task of future communicable disease spread forecasting?"*. In their paper, authors underline that their main objective is not to incorporate SCI in the baseline epidemiological model like SIR and to get state-of-the art forecast results, but rather to show usefulness of measures of social connectedness and bring the attention of broad research community to SCI and its possible applications.

The relevance of this paper and the results Kuchler et al. (2021) receive cannot be underestimated, given current severe epidemiological situation: continuing emergence of new virus variants, failure of many countries in prevention of new disease outbreaks and immense negative impact of COVID-19 on economies and society as a whole.

The paper investigates application of SCI to the task of virus spread forecasting from various perspectives. Firstly, it looks closely at the initial COVID dynamics in Italy provinces and US counties given their social proximity to the epicenter of disease - Lodi province and Westchester County, NY, respectively. Secondly, Kuchler et al. (2021) look at dynamic correlation of new cases and deaths given temporal nature of the virus evolution. Finally, authors conduct step-wise out-of-sample prediction of new cases in both countries and compare SCI contribution in the model performance with two additional advanced predictors - Google symptom searches and Location Exposure Index (LEX). In the first setting (initial cases), social distance to the epicenter appears to be a good predictor of future cases, robust to inclusion of a variety of controls. In the second setting, the coefficients of lagged social proximities to cases (an instrument based on SCI aggregation) are also highly statistically significant, even with the simultaneous inclusion of geographical proximities of cases and other controls. Finally, Kuchler et al. (2021) show that SCI is able to compete with and complement other measures of the population exposure to the virus, given time series temporal cross-validation scheme (step-wise out-of-sample prediction). For the whole history of observations, inclusion of social proximity to cases substantially improves non-linear regression model's quality metric (RMSE). Given wide access to the SCI (in contrast to LEX and Google search history), this instrument is proved to be efficient in the task of the future virus spread prediction.

1.2 Ideal Experiment

Before describing the data and identification strategy of Kuchler et al. (2021), let us consider "ideal" experimental setting needed to answer the research question.

On the one hand, for panel data environment, independent and random assignment of social connectedness would be needed. This would create exogenous variation in "treatment", which is social proximity to cases. Given this exogenous variation, one could determine the impact of social connectedness with regions hit by virus on future case growth in that region. Moreover, representative sample size would be needed, ideally a sample containing all existing countries employing identical procedure of COVID testing and case registration.

Sample homogeneity in terms of virus variant would be also desirable in the ideal setting.

On the other hand, when we switch to the out-of-sample temporal forecasting, assumptions are relaxed and only an accurate treatment of look-ahead bias is needed, as we now focus on the out-of-sample model performance measured by one of the regression/classification quality metrics. This means that one is free to use any information in period t to make a forecast for the period $t+1$, conditional on the requirement that data used do not contain any information from the future periods. In addition, any information that can be potentially unavailable in t should be excluded from the inputs to the model.

1.3 Data Set

Conducting a data-intensive research, Kuchler et al. (2021) use multiple data sources apart from Facebook *Social Connectedness Index*, a measure introduced by Bailey et al. (2018), which estimates the probability that the person from location i is a Facebook friend of a person from location j:

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

Where $Social\ Connectedness_{i,j}$ is the total number of Facebook friendship links between locations i and j.

To acquire 2-week period statistics for *new COVID-cases*, daily data from Johns Hopkins University GitHub repository and Dipartimento della Protezione Civile GitHub repository are used for cases and COVID-related deaths of US counties and Italian provinces, respectively.

In their regression specifications, Kuchler et al. (2021) introduce a number of controls to address endogeneity and omitted variable bias concerns. They use *US county-level demographics* from American Community Survey and Opportunity Insights. Similar demographic information for European NUTS3 regions is acquired from Eurostat. Moreover, authors control for the *urban-rural county classification* provided by National Center for Health Statistics. Additionally, physical proximity to cases and deaths is also considered as an explanatory variable, and researchers construct this measure using *county-to-county distances* from National Bureau of Economic Research.

To measure SCI out-of-sample predictive power, the index is compared to *Google searches related to COVID- 19 symptoms* (cough, fever, fatigue) and *smartphone-based Location Exposure Index* (LEX) introduced by Chevalier et al. (2021). Authors do not provide raw data for LEX index, however, they make preprocessed data public in the project GitHub repository.

The resulting data set is appropriate to address the stated research questions, as it comprises all information necessary to construct discussed social and physical proximity measures as well as additional controls. However, a few concerns arise when a comparison to ideal experimental setting is made. First (and not surprisingly), social connectedness index is not randomly assigned to counties and provinces, and it is logical to assume that social

connectedness is determined by the long-run travel and education patterns within the US and Italy. Second, the issue date of the SCI data is not reported, which raises look-ahead bias and measurement reproducibility issue. It is logical to assume that SCI evolves much slower than virus does, but we cannot rule out SCI temporal evolution completely. Third, Kuchler et al. (2021) do not state that Google trends and LEX data for period $t-1$ are available in period t (while social proximity to cases and deaths can be updated on a daily basis), which underlines the necessity to test respective specifications with exclusion of $t-1$ lags (for Google trends and LEX). Finally, COVID testing procedures and registration of cases/deaths may vary within the USA and Italy, which also violates ideal experimental setting.

1.4 Identification Strategy

The identification in Kuchler et al. (2021) relies on the variation in social proximity to cases across counties and provinces. This measure is constructed using SCI (for location i):

$$Social\ Proximity\ to\ Cases_{i,t} = \sum_j Cases\ Per\ 10k_{j,t} * \frac{Social\ Connectedness_{i,j}}{\sum_h Social\ Connectedness_{i,h}} \quad (2)$$

In the first part of the paper, Kuchler et al. (2021) look at the correlation between social connectedness of a location (county or province) to the early hotspot location, where earliest outbreaks were registered (Westchester County in US and Lodi province in Italy) and a number of cases per 10,000 people as of March 30, 2020. For both countries, strong non-linear correlations are found. Being rightfully concerned with omitted variable bias and endogeneity problem, authors include a number of controls in the regression (exclusion of nearest locations, geographical distance to the hotspot, median income, population density, rural/urban indicators) and run "placebo" regressions, iteratively estimating incremental R^2 of SCI to the Westchester in the regression for each US county. As an additional exercise, Kuchler et al. (2021) run an out-of-sample initial hotspot analysis, where they use lagged Italian data to make a forecast of cases in the USA as of March 30, 2020, and find out that inclusion of information on social connectedness to the early hotspot improves quality metrics of both linear and non-linear regression models.

In the second part of the paper Kuchler et al. (2021) conduct time series analysis in form of a panel regression with cases per 10,000 people in location i in period t being dependent variable and the following variables being regressors:

- cases per 10,000 people in the same location lagged by one and two 2-week periods
- social proximity to cases lagged by one and two 2-week periods
- share of Facebook friends located within 50 and 150 miles from location i
- physical proximity to cases lagged by one and two 2-week periods
- other controls mentioned in the early hotspot analysis part

In all investigated regression specifications with all combinations of included controls, 2-week lagged social proximity to cases is highly statistically significant (this significance

of 2-week lagged social proximity to cases coefficient remains even for iterative regression estimation with 2-week data increment, showing that SCI carries important information on the whole time span of the study). Moreover, to eliminate the concern of heterogeneous registration of new cases across regions, Kuchler et al. (2021) use the same regression specifications, but change cases to deaths, and the result does not change qualitatively.

In the third part of the research, authors continue testing predictive power of SCI-based explanatory variables and bring regressions closer to the real world setting. Namely, they conduct temporal out-of-sample prediction (also known as time series cross validation) with incremental growth of data available for the regression model (random forest in this case). This time, apart from social proximity to cases, Kuchler et al. (2021) use state-of-the-art predictors of COVID-19 spread - LEX index and Google search trends [2]. In almost all regression specifications inclusion of social proximity to cases improves regression quality in terms of RMSE metric for all consecutive 2-week periods (when both LEX and Google symptom search data are included, social proximity to cases data improves quality in 10 out of 14 periods).

To sum up the identification strategy used in the Kuchler et al. (2021) paper, one should mention that this strategy is the most plausible given the underlying nature of real world time series data. Authors are concerned with possible biases and other identification problems and devote substantial part of the paper to robustness checks, and include additional variables that could possibly add explanatory power to the model. Social connectedness component appears to be relevant and informative in all researched specifications and settings, which convinces that social network structure proxied by SCI does explain part of variation in communicable disease spread across regions.

1.5 Literature Review

The discussed Kuchler et al. (2021) paper makes an important contribution to the evolving literature covering application of broad network theory to the spatial epidemiological modelling. Key questions in this literature depend on the dimension of epidemiological modelling under study, but it seems that Kuchler et al. (2021) study fits well in the domain of future disease spread forecasting, which has been enjoying extended demand since the COVID-19 global outbreak. However, according to Kuchler et al. (2021), previous research did not provide network measurements of such spatial coverage and granularity.

More precisely, the paper adds to the series of papers that investigate how social media can be used in explaining and predicting spread of communicable diseases. As it is mentioned by authors, one strand of this literature focuses on predictive power of people behaviour in social media: posts, likes, searches of specific information (disease symptoms). However, these papers have to cope with rapid change of trends in internet behaviour and seasonality. Kuchler et al. (2021) also mention, that such approach can track the spread only in ex-post fashion, as internet searches and posts imply that a person already has first disease symptoms. SCI-based measures, on the contrary, are able to spot locations at risk before they

[2] As of December 2021, Google provides search trend data only up to February 2021, which limits its usability in the real world setting.

experience virus outbreak.

Another strand of literature exploits individuals' geolocation data to track their movement and interaction, which could facilitate the spread of virus. While being very precise and granular, such information is highly private and may be limited in terms of global coverage. SCI, in contrast, is very aggregated and provided publicly, which makes it easily accessible for all interested researchers. It is important to add that reproducibility plays an important role in the field of epidemiological forecasting, where data should be updated on a weekly or even daily basis. This time constraint makes a number of explanatory variables irrelevant for the use in the real world setting - as mentioned above, Goggle search trends, for example, are currently provided until February 2021. SCI-based features of a social network, in contrast, can be easily updated daily, as SCI across locations is more likely to evolve at a much slower pace (decades) than the virus itself (for which daily data is available and regularly updated by policymakers).

Finally, Kuchler et al. (2021) continue the line of research on social connectedness index and its application to the modern social and economic problems. Introduced by Bailey et al. (2018), SCI idea has been already used to explore urban social connectedness and travel patters (Bailey et al., 2020a), gain insight of international trade's social determinants (Bailey et al., 2021) and explain differences in travel across Europe (Bailey et al., 2020b).

2 Replication

2.1 Early Hotspot Analysis

We start our replication exercise with an analysis of early spread of COVID-19 among counties and provinces with heterogeneous social ties to the virus hotspot locations - Westchester county in the USA and Lodi province in Italy.

First, in figure 1 we plot heatmaps depicting SCI of each US location to Westchester and a number of cases in these counties per 10,000 residents as of March 30, 2020, using geographical data hosted by Kuchler et al. (2021). Two counties from initial cases table cannot be matched to corresponding geographical data (District of Columbia, District of Columbia; Doca Ana County, New Mexico), probably due to county name differences between these two data sources. Following authors, we also drop Hawaii and Alaska states from observation in heatmaps, but include them in the subsequent regressions.

Apart from heatmaps, authors construct binscatterplots with $log(Social\,Connectedness)$ to the virus epicenter on the x-axis and $Cases per 10k people$ on the y-axis, and fit a quadratic trend on these data. Binscatters represent the correlation of the two variables without controls (apart from the exclusion of regions that are situated closer than 50 miles/50 kilometers from Westchester/Lodi, which is a basic constraint for any regression specification) and including additional controls. According to Kuchler et al. (2021), controls for the US data are constructed in the following fashion:

1. $log(SCI)$ and $Cases\ per\ 10k\ people$ variables are first regressed on a set of control variables:

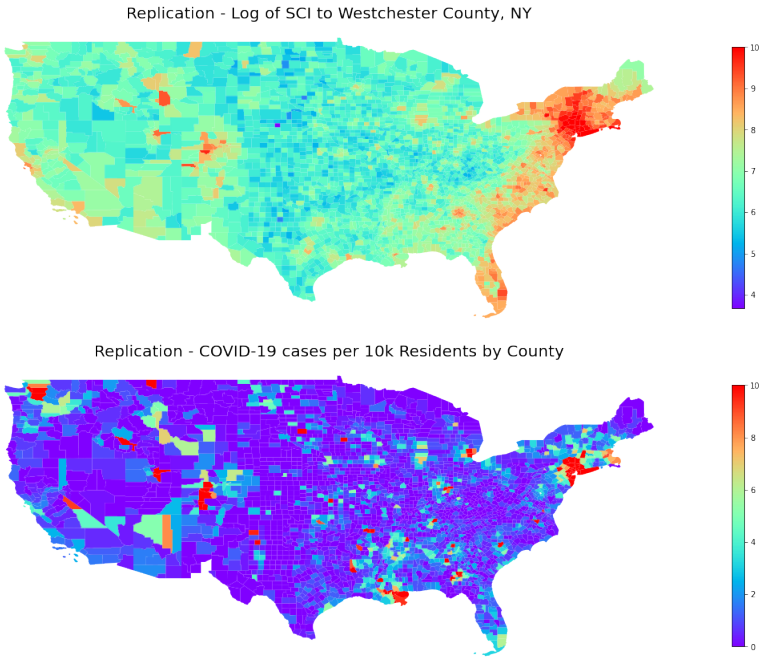

Figure 1: Replication - US heatmaps

- 100 dummies for the percentile of the county's geographic distance to Westchester
- population density
- median household income
- dummies for the six National Center for Health Statistics Urban-Rural county classifications

Residual values of $log(SCI)$ are then grouped into 100 equal-sized bins and average residual $log(SCI)$ of the bins are plotted against average residual $Cases\ per\ 10k\ people$ of the bins. For the specification without additional controls, binning procedure is the same, but original values of $log(SCI)$ and $Cases\ per\ 10k\ people$ are averaged across bins (figure 2).

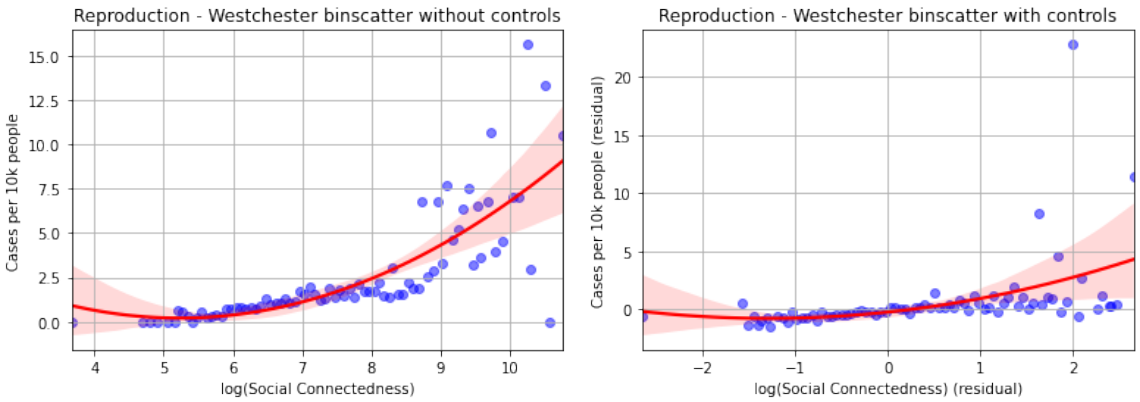

Figure 2: Replication - Westchester binscatters with (right) and without (left) controls

At this step, differences from the results in Kuchler et al. (2021) can be clearly seen. First, reconstructed binscatters show more variation of $Cases\ per\ 10k\ people$ at the right tail of the

$log(SCI)$ and residual $log(SCI)$ distributions. Secondly, specification with additional controls implies much less pronounced linear/quadratic correlation between residual $log(SCI)$ and residual $Cases\ per\ 10k\ people$ in contrast to the results in Kuchler et al. (2021). It should be mentioned that even after making a transition to residuals of $log(SCI)$, authors' binscatter x-axis hardly changes, while it is logical to assume that OLS residuals will be clustered around origin (which is the case in our replication results), rather than being close to initial values of $log(SCI)$. This concern underlines the necessity to double-check authors' R code (replication exercise is conducted in Python).

However, if we left unchanged $log(SCI)$ values on the x-axis and plot residualized $Cases\ per\ 10k\ people$ on the y-axis (figure 3), the correlation between these two variables remains at the lower level in comparison with Kuchler et al. (2021) findings.

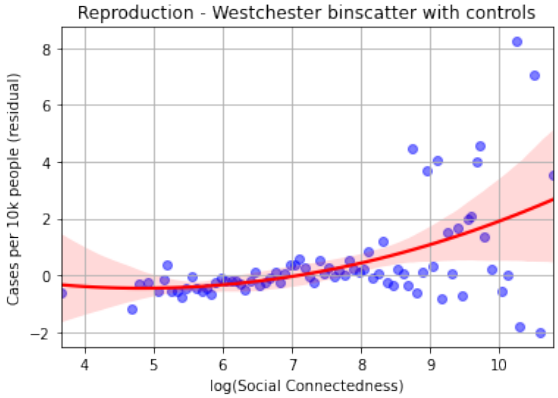

Figure 3: Replication - Westchester binscatters with controls, original x-axis

The same procedure is done for Italian provinces (figure 4), where Lodi serves as a source of virus spread across the country, the only difference is binning, that is now performed using 30 equal-sized bins, due to the smaller data size. Here, only "ITG2" province cannot be matched to the geographical data.

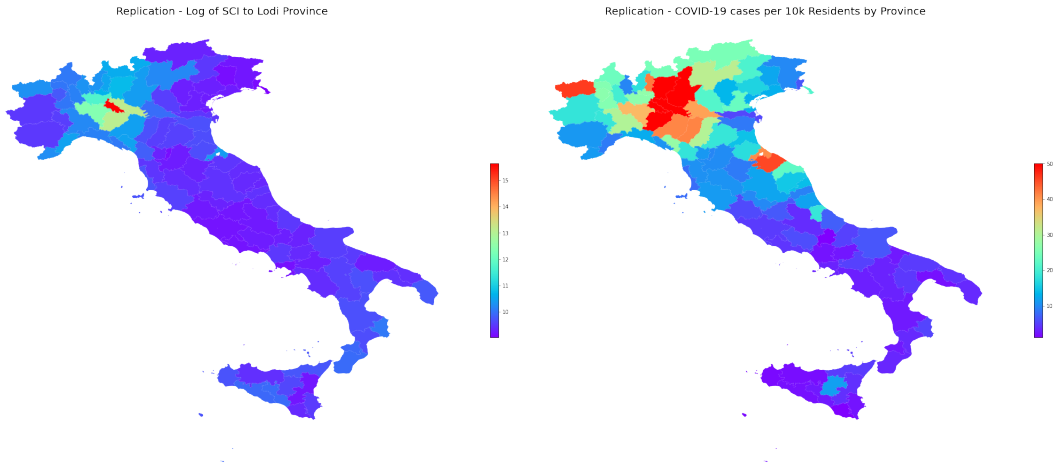

Figure 4: Replication - Italy heatmaps

The same concern about correlation estimation robustness in presence of controls remains for the Italian data. As we introduce residualized $log(SCI)$ and $Cases\,per\,10k\,people$ values, x-axis position shifts towards the origin (figure 5). The variation of $Cases\,per\,10k\,people$ in the specification without controls also gets higher (in contrast to the original results) with the growth of $log(SCI)$, as can be seen in figure 5 .

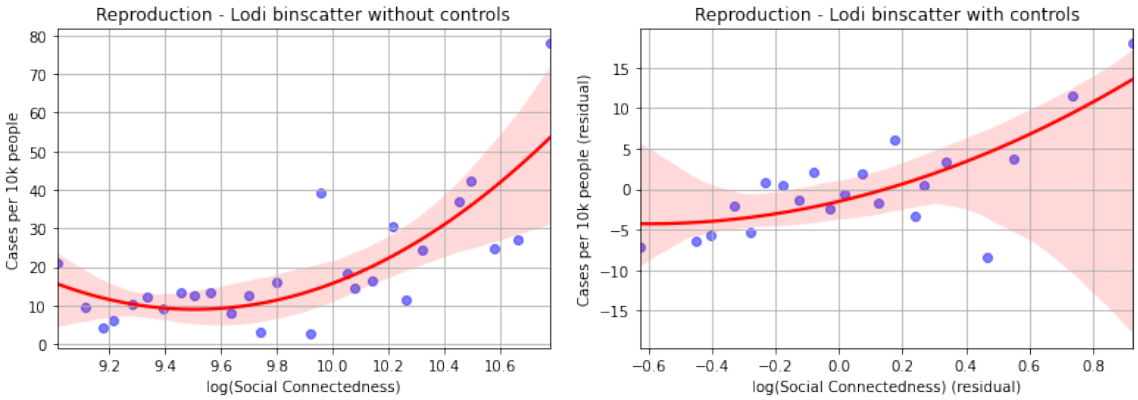

Figure 5: Replication - Lodi binscatters with (right) and without (left) controls

2.2 Time Series Analysis - Panel Regression

At the next step of their analysis, Kuchler et al. (2021) move from initial spread of the disease across countries' locations to the temporal evolution of COVID-19 and its expansion across US counties. In this time series analysis setting, authors measure influence of lagged social proximity to cases on current morbidity in presence of other competing measures of closeness to cases: lagged increase in cases in the same region, share of friends within 50 and 150 miles and physical proximity to cases (equation 3).

$$Physical\ Proximity\ to\ Cases_{i,t} = \sum_j Cases\ Per\ 10k_{j,t} * \frac{1}{1 + Distance_{i,j}} \tag{3}$$

General panel regression specification of Kuchler et al. (2021) can be written as:

$$\begin{aligned}
\log(\Delta \text{ Cases per } 10k+1)_{i,t} = {} & \beta_1 * \log(\Delta \text{ Cases per } 10k+1)_{i,t-1} \\
& + \beta_2 * \log(\Delta \text{ Cases per } 10k+1)_{i,t-2} \\
& + \beta_3 * \log(\Delta \text{ Social Proximity to Cases })_{i,t-1} \\
& + \beta_4 * \log(\Delta \text{ Social Proximity to Cases })_{i,t-2} \\
& + \beta_5 * \text{ Share Friends within 50mi}_i \\
& + \beta_6 * \text{ Share Friends within 150mi}_i \\
& + \beta_7 * \log(\Delta \text{ Physical Proximity to Cases })_{i,t-1} \\
& + \beta_8 * \log(\Delta \text{ Physical Proximity to Cases })_{i,t-2} \\
& + X_{i,t} + \epsilon_{i,t},
\end{aligned} \tag{4}$$

where $X_{i,t}$ denotes fixed effects of time and population density, time and median household income, and time and state. Following Kuchler et al. (2021), we do not include a constant in regressions and select observations without missing values between March 30, 2020 and

	(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

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

November 2, 2020 (April 28, 2020 and November 2, 2020 for data on deaths). For columns 1-6, estimated coefficients are close to those reported in the paper, but social proximity influence (columns 1-2) on the outcome variable is higher than in Kuchler et al. (2021). Furthermore, for specifications 7 and 8 (panel A), which include all the competing regressors, not only social proximity to cases becomes more important, but other variables' coefficients (share of friends within 50 and 150 miles) gain statistical significance too.

In panel B, Kuchler et al. (2021) change cases to deaths and estimate influence of prior social proximity to deaths on current growth of COVID-related lethal cases. In this setting, strong positive effect of social proximity can also be clearly seen. In our replication, influence of SCI-based variables is also higher than in the original paper, and in final regression specifications (columns 7-8 of panel B), some coefficients become statistically significant, but do not change qualitatively.

	(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

Table 2: Replication - COVID-19 deaths growth and prior proximity to deaths (Panel B). Dependent variable - $\log(\Delta$ Deaths per $10k+1)_{i,t}$

Despite some (minor) differences in estimations, overall magnitudes of regressors under study and point estimates of their coefficients are consistent with Kuchler et al. (2021). For both cases (panel A) and deaths (panel B), prior social proximity to the disease is crucial in explaining current cases/deaths growth rates. We can conclude that SCI-based indicators indeed have strong predictive power, at least in this research setting. Next, we move forward to iterative out-of-sample prediction making of future cases.

2.3 Time Series Analysis - Out-of-Sample Prediction

In this section we reproduce final regression exercise undertaken by Kuchler et al. (2021). Authors test predictive power of social connectedness index in the real-world setting, in which information builds up iteratively and one can only use information from periods $0, ..., t-1$ to make a forecast for period t.

In this out-of-sample forecast section, four sets of information are used:

- Baseline explanatory variables - 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
- LEX data - lagged logs of changes in LEX-weighted proximity to cases
- Google search data - lagged changes in number of Google searches on one of the three COVID-19 symptoms (fever, cough, fatigue) in county i
- SCI data - lagged logs of changes in social proximity to cases, which served as central variables of interest in the previous sections

LEX and Google search data are two "state-of-the-art" measures of COVID-19 exposure and need large amounts of private information for construction. While number of COVID-related Google requests is a straightforward variable, LEX-weighted proximity to cases needs additional explanation.

$LEX_{i,j}$ index shows a number of phones that pinged in location i today and pinged in location j within the last 14 days. By construction, this indicator is able to show the inflow of people from different regions hit by virus, that could potentially increase the number of cases within the forthcoming weeks. LEX-weighted proximity to cases is calculated the following way:

$$LEX\ Proximity\ to\ Cases_{i,t} = \sum_j Cases\ Per\ 10k_{j,t} * \frac{ELX_{i,j,t}}{\sum_h LEXi, h, t} \tag{5}$$

To test the value of SCI data, Kuchler et al. (2021) estimate regressions with and without additional SCI information (with baseline variables and baseline + LEX + Google regressors). At each of consecutive 14 weeks (with 2-week interval), authors use all the available information up to "current" week (ω) to fit a random forest regression model (number of trees = 500) and make predictions for each US county in the week ω (target variable is $\log(\Delta\text{Cases per } 10k + 1)_{i,\omega}$). To measure out-of-sample quality of the model, root mean squared error (RMSE) metric is used. Unfortunately, authors didn't specify any of random forest hyper-parameters, except the number of trees (and also did not fix random state of the model, which could have made the experiment fully reproducible). In our reproduction exercise we set the following parameters of random forest regression (from *scikit-learn* python library) - number of trees = 500, maximum tree depth = 3 (remaining parameters are set at default values by *scikit-learn* library). As in Kuchler et al. (2021), we test the model on baseline data, baseline + SCI data, baseline + LEX + Google data and baseline + LEX + Google + SCI data, to be able to spot any RMSE improvement explained by inclusion of

social proximity information.

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

In table 3 we replicate columns 3 and 9 of table 2 in Kuchler et al. (2021). These columns show a reduction in random forest out-of-sample RMSE after SCI-based features (regressors) are added. Negative values represent an improvement in the model quality on the test subset of data.

As in Kuchler et al. (2021), introduction of new SCI information to the model results in decreasing RMSE, which once again signals that social connectedness to cases has unique predictive power and could be used to make more informed and accurate inferences regarding growth of cases in the near future. However, our RMSE improvements are somewhat smaller than those reported by Kuchler et al. (2021). This could be due to differences in random forest optimization algorithms in R (used in the original paper) and python's *scikit-learn*. Furthermore, while authors' random forest quality on baseline data improves in 100% of weeks when social connectedness as added, our algorithm performs better in 12 out of 14 weeks. Regarding more advanced baseline + LEX + Google specification, both Kuchler et al. (2021) model and our replication show a slight improvement in 10 out of 14 weeks. Next, we illustrate the dynamics of out-of-sample RMSE metric (figure 6). Intuitively, as the model receives more data, its general quality grows, and levels by the 8th week of out-of-sample forecasting. In the baseline specification, social proximity data is still able to significantly shift RMSE downwards, while in the baseline + LEX + Google case, only a marginal improvement is seen.

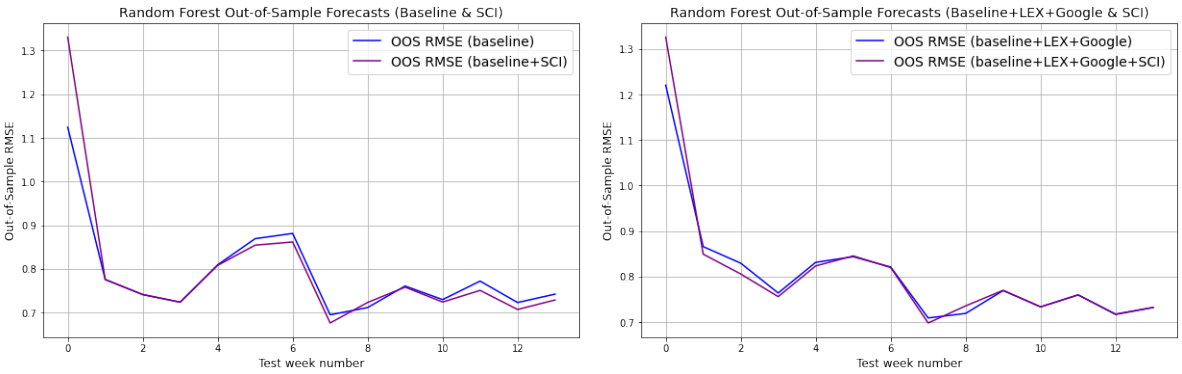

Figure 6: Out-of-sample RMSE dynamics, Random Forest

3 Extension

There are several dimensions along which the original paper can be possibly extended: alternative social proximity to cases measures, investigation of used specifications at the ZIP code level data for the USA, study of countries' social connectedness and subsequent international spread of COVID-19, etc. In this part of the term paper, we try gain more insights about out-of-sample prediction ability of classic machine-learning algorithms on the set of data proposed by Kuchler et al. (2021). One substitution for a random forest algorithm could be gradient boosting model, which comprises a set of "nested" decision trees, which learn to predict the error of previous trees. This model, described in Chen and Guestrin (2016), is still considered to be one of the best algorithms for tabular data (Gorishniy et al., 2021).

First, we fit gradient boosting model on a data set used in the previous section (baseline and baseline + LEX + Google data) and study the effect of inclusion of SCI-data in the set of features, used for inference [3]. Resulting table 4 indicates that social proximity data is also able to significantly enhance the performance of gradient boosting given the baseline specification (lower RMSE in 11 out of 14 weeks). However, when LEX and Google search trends are included and passed to XGBoost, this positive effect of SCI fades, as we can see an increase in performance for only 6 iterations out of 14.

Next, we compare performance of gradient boosting model in comparison with random forest on a baseline data set. As depicted in the figure 7, relative performance of XGBoost increases over time, meaning that this algorithm tens to leverage larger amounts of data better than random forest (proposed by Kuchler et al. (2021)) and could be used in the real world setting. However, one should be concerned about the overfit problem in both random forest and gradient boosting algorithms (although, the latter is more prone to this issue in our setting), which can be described as a dramatic increase in quality metrics on "train" data and substantial deterioration of those metrics on out-of-sample or "test" data. Because of the overfit problem, we use shallow trees of maximum depth 3 for both algorithms.

[3]Here, the parameters of XGBoost are: *n estimators=500, max depth=3*, random state is fixed.

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

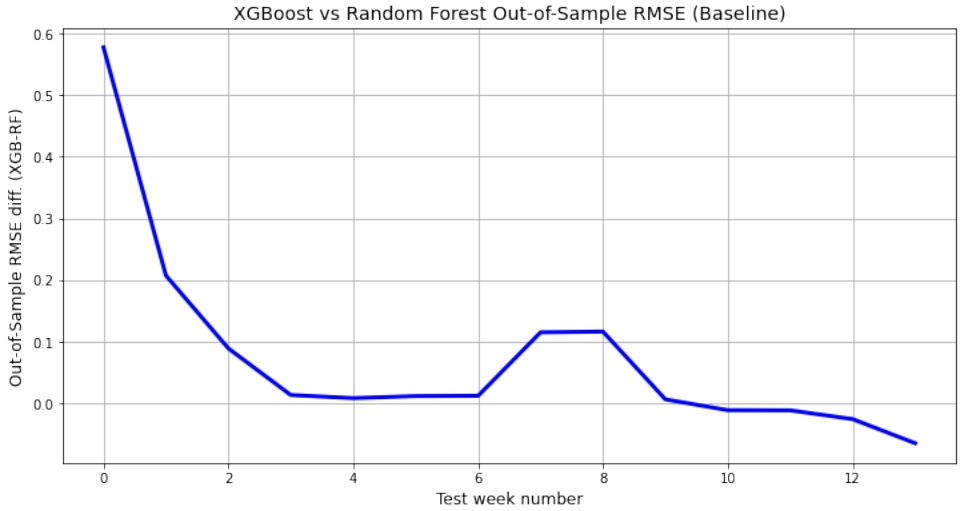

Figure 7: Out-of-sample RMSE dynamics, Random Forest

4 Conclusion

This term paper comprises a referee report on the paper by Kuchler et al. (2021), which brings our attention to the Social Connectedness Index, developed by Bailey et al. (2018), in context of global pandemics. Using high resolution spatio-temporal data for US counties and Italian provinces, authors of the paper illustrate that social connectedness to the early virus hotspots served as an important determinant of counties' and provinces' number of COVID-19 cases as of March 30, 2020. Moreover, Kuchler et al. (2021) show that measures of previous social proximity to cases/deaths of a county determined the subsequent growth of cases/deaths in that county, and these results hold across counties and time. Finally, Kuchler et al. (2021) illustrate how SCI can be used to make better informed out-of-sample forecasts of future cases, using classic random forest algorithm and bringing their research closer to the real world setting.

Replication of the paper confirms main conclusions of Kuchler et al. (2021) at each step of their research. Indeed, reconstructing original calculations using the data provided by authors, we can also infer that social closeness to the initial hotspot made US counties and Italian provinces more susceptible to virus spread. Following Kuchler et al. (2021) in their time series analysis, we come to the results that are close to the original paper. More precisely, we document significant influence of prior social proximity to cases/deaths on future growth in cases and lethal outcomes. Finally, we also show that inclusion of SCI-based indicators of virus exposure can substantially increase predictive power of non-linear regression algorithm (random forest) on out-of-sample data, even when more advanced proxies of COVID exposure and risk are also included in the specification.

To make a modest contribution to the discussed topic, we show that other machine learning algorithms, such as gradient boosted decision trees, are able to show even greater predictive power in comparison with random forest regression, given large enough amount of spatio-temporal data. We, in turn, also witness an improvement in algorithm out-of-sample quality when social proximity features are added to the baseline regression specification, which confirms that SCI application to epidemiological forecasting should be researched further.

It is also worth mentioning that this topic has a lot of potential improvements. One way further could be transition from tabular data to graph data, as SCI gives a unique opportunity for the construction of spatio-temporal graphs with SCI serving as a weight of an edge connecting two graph nodes (locations). This transition would make it possible to use emerging graph machine learning approach (graph neural networks) to make forecasts of the communicable disease spread.

References

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., and Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3):259–80.

Bailey, M., Farrell, P., Kuchler, T., and Stroebel, J. (2020a). Social connectedness in urban areas. *Journal of Urban Economics*, 118:103264.

Bailey, M., Gupta, A., Hillenbrand, S., Kuchler, T., Richmond, R., and Stroebel, J. (2021). International trade and social connectedness. *Journal of International Economics*, 129:103418.

Bailey, M., Johnston, D., Kuchler, T., Russel, D., Stroebel, J., et al. (2020b). The determinants of social connectedness in europe. In *International Conference on Social Informatics*, pages 1–14. Springer.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chevalier, J. A., Schwartz, J. L., Su, Y., Williams, K. R., et al. (2021). Measuring movement and social contact with smartphone data: A real-time application to covid-19. Technical report, Cowles Foundation for Research in Economics, Yale University.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting deep learning models for tabular data. *arXiv preprint arXiv:2106.11959*.

Kuchler, T., Russel, D., and Stroebel, J. (2021). Jue insight: The geographic spread of covid-19 correlates with the structure of social networks as measured by facebook. *Journal of Urban Economics*, page 103314.