

1 The importance of reproducibility in geographical
2 research: the case of population density and the spread
3 of COVID-19

Author^{*,a}

^aDepartment, Street, City, State ZIP

6 Abstract

The emergence of the novel SARS-CoV-2 coronavirus and the global COVID-19 pandemic has led to explosive growth in scientific research. Of interest has been the associations between population density and the spread of the pandemic. In this paper, population density and the basic reproductive number of SARS-CoV-2 are examined in an example of reproducible research. Given the high stakes of the situation, it is essential that scientific activities, on which good policy depends, are as transparent and reproducible as possible. Reproducibility is key for the efficient operation of the self-correction mechanisms of science. Transparency and openness means that the same problem can, with relatively modest efforts, be examined by independent researchers who can verify findings, and bring to bear different perspectives, approaches, and methods—sometimes with consequential changes in the conclusions, as the empirical example of the spread of COVID-19 in the US shows.

*Corresponding Author
Email address: author@institution.edu (Author)

⁷ **Introduction**

⁸ The emergence of the novel SARS-CoV-2 coronavirus in 2019, and the global
⁹ pandemic that followed in its wake, led to an explosive growth of research around
¹⁰ the globe. According to Fraser et al. (2021), over 125,000 COVID-19-related
¹¹ papers were released in the first ten months from the first confirmed case of
¹² the disease. Of these, more than 30,000 were shared in pre-print servers, the
¹³ use of which also exploded in the past year (Añazco et al., 2021; Kwon, 2020;
¹⁴ Vlasschaert et al., 2020).

¹⁵ Given the heavy human and economic cost of the pandemic, there has been a
¹⁶ natural tension in the scientific community between the need to publish research
¹⁷ results quickly and the imperative to maintain consistently high quality standards
¹⁸ in scientific reporting; indeed, a call for maintaining the standards in published
¹⁹ research has termed this deluge of publications a “carnage of substandard research”
²⁰ (Bramstedt, 2020). Part of the challenge of maintaining quality standards
²¹ in published research is that, despite an abundance of recommendations and
²² guidelines (Broggini et al., 2017; Brunsdon and Comber, 2020; Ince et al., 2012;
²³ Ioannidis et al., 2014), in practice reproducibility has remained a lofty and
²⁴ somewhat aspirational goal (Konkol et al., 2019; Konkol and Kray, 2019). As
²⁵ reported in the literature, only a woefully small proportion of published research
²⁶ was actually reproducible before the pandemic (Iqbal et al., 2016; Stodden et
²⁷ al., 2018), and the situation does not appear to have changed substantially since
²⁸ (Gustot, 2020; Sumner et al., 2020).

²⁹ The push for open data and software, along with more strenuous efforts
³⁰ towards open, reproducible research, is simply a continuation of long-standing
³¹ scientific practices of independent verification. Despite the (at times dispro-
³² portionate) attention that high profile scandals in science tend to elicit in the
³³ media, science as a collective endeavor is remarkable for being a self-correcting

34 enterprise, one with built-in mechanisms and incentives to weed out erroneous
35 ideas. Over the long term, facts tend to prevail in science. At stake is the
36 shorter-term impacts that research may have in other spheres of economic and
37 social life. The case of economists Reinhart and Rogoff comes to mind: by the
38 time the inaccuracies and errors in their research were uncovered (see Herndon
39 et al., 2014), their claims about debt and economic growth had already been
40 seized by policy-makers on both sides of the Atlantic to justify austerity policies
41 in the aftermath of the Great Recession of 2007-2009¹. As later research has
42 demonstrated, those policies cast a long shadow, and their sequels continued to
43 be felt for years (Basu et al., 2017).

44 In the context of COVID-19, a topic that has grabbed the imagination of
45 numerous thinkers has been the prospect of life in cities after the pandemic
46 (Florida et al., 2020); the implications are the topic of ongoing research (Sharifi
47 and Khavarian-Garmsir, 2020). The fact that the worst of the pandemic was
48 initially felt in dense population centers such as Wuhan, Milan, Madrid, and New
49 York, brought a torrent of research into the associations between density and the
50 spread of the pandemic. Some important questions hang on the results of these
51 research efforts. For example, are lower density regions safer from the pandemic?
52 Are de-densification policies warranted, even if just in the short term? And in
53 the longer term, will the risks of life in high density regions presage a flight from
54 cities? What are the implications of the pandemic for future urban planning and
55 practice? Over the past year, numerous papers have sought to throw light into
56 the underlying issue of density and the pandemic; nonetheless the results, as will
57 be detailed next, remain mixed. Further, to complicate matters, precious few of

¹Nobel Prize in Economics Paul Krugman noted that “Reinhart–Rogoff may have had more immediate influence on public debate than any previous paper in the history of economics” <https://www.nybooks.com/articles/2013/06/06/how-case-austerity-has-crumbled/?pagination=false>

58 these studies appear to be sufficiently open to support independent verification.

59 The objective of this paper is to illustrate the importance of reproducibility
60 in research in the context of the flood of COVID-19 papers. To this end,
61 a recent study by Sy et al. (2021) is chosen as an example of reproducible
62 research. The objective is not to malign the analysis of these researchers, but
63 rather to demonstrate the value of openness to allow for independent verification
64 and further analysis. Open data and open code mean that an independent
65 researcher can, with only modest efforts, not only verify the findings reported,
66 but also examine the same data from a perspective which may not have been
67 available to the original researchers due to differences in disciplinary perspectives,
68 methodological traditions, and/or training, among other possible factors. The
69 example, which shows consequential changes in the conclusions reached by
70 different analyses, should serve as a call to researchers to redouble their efforts
71 to increase transparency and reproducibility in research. This paper, in addition,
72 aims to show how data can be packaged in well-documented, shareable units,
73 and code can be embedded into self-contained documents suitable for review and
74 independent verification. The source for this paper is an R Markdown document
75 which, along with the data package, will be available in a public repository².

76 **Background: the intuitive relationship between density and spread of
77 contagious diseases**

78 The concern with population density and the spread of the virus during the
79 COVID-19 pandemic was fueled, at least in part, by dramatic scenes seen in
80 real-time around the world from large urban centers such as Wuhan, Milan,
81 Madrid, and New York. In theory, there are good reasons to believe that higher

²For peer-review purposes, the data package and code are currently in an anonymous Drive folder: <https://drive.google.com/drive/folders/1cT6tcUc1pJ4aT5ajQ0emO0lyS46P8Ige?usp=sharing>

density may have a positive association with the transmission of a contagious virus. It has long been known that the potential for inter-personal contact is greater in regions with higher density (see for example the research on urban fields and time-geography, including Farber and Páez, 2011; Moore, 1970; Moore and Brown, 1970). Mathematically, models of exposure and contagion indicate that higher densities can catalyze the transmission of contagious diseases (Li et al., 2018; Rocklöv and Sjödin, 2020). The idea is intuitive and likely at the root of messages, by some figures in positions of authority, that regions with sparse population densities faced lower risks from the pandemic³.

As Rocklöv and Sjödin (Rocklöv and Sjödin, 2020) note, however, mathematical models of contagion are valid at small-to-medium spatial scales (and presumably, small temporal scales too, such as time spent in restaurants, concert halls, cruises), and the results do not necessarily transfer to larger spatial units and different time scales. There are solid reasons for this: while in a restaurant, one can hardly avoid being in proximity to other customers-however, a person can choose to (or be forced to as a matter of policy) not go to a restaurant in the first place. Nonetheless, the idea that high density correlates with high transmission is so seemingly reasonable that it is often taken for granted even at larger scales (e.g., Cruz et al., 2020; Micallef et al., 2020). At larger scales, however, there exists the possibility of behavioral adaptations, which are difficult to capture in the mechanistic framework of differential equations (or can be missing in agent-based models, e.g., Gomez et al., 2021); these adaptations, in fact, can be a key aspect of disease transmission.

A plausible behavioral adaptation during a pandemic, especially one broadcast

³Governor Kristi Noem of South Dakota, for example, claimed that sparse population density allowed her state to face the pandemic down without the need for strict policy interventions <https://www.inforum.com/lifestyle/health/5025620-South-Dakota-is-not-New-York-City-Noem-defends-lack-of-statewide-COVID-19-restrictions>

¹⁰⁶ as widely and intensely as COVID-19, is risk compensation. Risk compensation
¹⁰⁷ is a process whereby people adjust their behavior in response to their *perception*
¹⁰⁸ of risk (Noland, 1995; Phillips et al., 2011; Richens et al., 2000). In the case of
¹⁰⁹ COVID-19, Chauhan et al. (Chauhan et al., 2021) have found that perception of
¹¹⁰ risks in the US varies between rural, suburban, and urban residents, with rural
¹¹¹ residents in general expressing less concern about the virus. It is possible that
¹¹² people who listened to the message of leaders saying that they were safe because
¹¹³ of low density may not have taken adequate precautions against the virus. People
¹¹⁴ in dense places who could more directly observe the impact of the pandemic
¹¹⁵ may have become overly cautious. Both Paez et al. (2020) and Hamidi et al.
¹¹⁶ (2020b) posit this mechanism (i.e., greater compliance with social distancing in
¹¹⁷ denser regions) to explain the results of their analyses. The evidence available
¹¹⁸ does indeed show that there were important changes in behavior with respect
¹¹⁹ to mobility during the pandemic (Harris and Branon-Calles, 2021; Jamal and
¹²⁰ Paez, 2020; Molloy et al., 2020); furthermore, shelter in place orders may have
¹²¹ had greater buy-in from the public in higher density regions (Feyman et al.,
¹²² 2020; Hamidi and Zandiatashbar, 2021), and the associated behavior may have
¹²³ persisted beyond the duration of official social-distancing policies (Prahraj et
¹²⁴ al., 2020). In addition, there is evidence that changes in mobility correlated with
¹²⁵ the trajectory of the pandemic (Noland, 2021; Paez, 2020). Given the potential
¹²⁶ for behavioral adaptation, the question of density becomes more nuanced: it
¹²⁷ is not just a matter of proximity, but also of human behavior, which is better
¹²⁸ studied using population-level data and models.

¹²⁹ **Background: but what does the literature say?**

¹³⁰ When it comes to population density and the spread of COVID-19, the
¹³¹ international literature to date remains inconclusive.

132 On the one hand, there are studies that report positive associations between
133 population density and various COVID-19-related outcomes. Bhadra (2021),
134 for example, reported a moderate positive correlation between the spread of
135 COVID-19 and population density at the district level in India, however their
136 analysis was bivariate and did not control for other variables, such as income.
137 Similarly, Kadi and Khelfaoui (2020) found a positive and significant correlation
138 between number of cases and population density in cities in Algeria in a series
139 of simple regression models (i.e., without other controls). A question in these
140 relatively simple analyses is whether density is not a proxy for other factors.
141 Other studies have included controls, such as Pequeno et al. (2020), a team
142 that reported a positive association between density and cumulative counts
143 of confirmed COVID-19 cases in state capitals in Brazil after controlling for
144 covariates, including income, transport connectivity, and economic status. In
145 a similar vein, Fielding-Miller et al. (2020) reported a positive relationship
146 between the absolute number of COVID-19 deaths and population density (rate)
147 in rural counties in the US. Roy and Ghosh (2020) used a battery of machine
148 learning techniques to find discriminatory factors, and a positive and significant
149 association between COVID-19 infection and death rates in US states. Wong and
150 Li (2020) also found a positive and significant association between population
151 density and number of confirmed COVID-19 cases in US counties, using both
152 univariate and multivariate regressions with spatial effects. More recently, Sy
153 et al. (2021) reported that the basic reproductive number of COVID-19 in US
154 counties tended to increase with population density, but at a decreasing rate at
155 higher densities.

156 On the flip side, a number of studies report non-significant or negative
157 associations between population density and COVID-19 outcomes. This includes
158 the research of Sun et al. (2020) who did not find evidence of significant

correlation between population density and confirmed number of cases per day *in conditions of lockdown* in China. This finding echoes the results of Paez et al. (2020), who in their study of provinces in Spain reported non-significant associations between population density and infection rates in the early days of the first wave of COVID-19, and negative significant associations in the later part of the first lockdown. Similarly, (2020) found zero or negative associations between population density and infection numbers/deaths by country. Fielding-Miller et al. (2020) contrast their finding about rural counties with a negative relationship between COVID-19 deaths and population density in urban counties in the US. For their part, in their investigation of doubling time, White and Hébert-Dufresne (2020) identified a negative and significant correlation between population density and doubling time in US states. Likewise, (2021) found a small negative (and significant) association between population density and COVID-19 morbidity in districts in Tehran. Finally, two of the most complete studies in the US [by Hamidi et al.; (2020a); (2020b)] used an extensive set of controls to find negative and significant correlations between density and COVID-19 cases and fatalities at the level of counties in the US.

As can be seen, these studies are implemented at different scales in different regions of the world. They also use a range of techniques, from correlation analysis, to multivariate regression, spatial regressions, and machine learning techniques. This is natural and to be expected: individual researchers have only limited time and expertise. This is why reproducibility is important. To pick an example (which will be further elaborated in later sections of this paper), the study of Sy et al. [(2021); hereafter SWN] would immediately grab the attention of a researcher with a somewhat different toolbox.

184 **Reproducibility of research**

185 SWN investigated the basic reproductive number of COVID-19 in US counties,
186 and its association with population density, median household income, and
187 prevalence of private mobility. For their multivariate analysis, SWN used mixed
188 linear models. This is a reasonable modelling choice: R_0 is an interval-ratio
189 variable that is suitably modeled using linear regression; further, as SWN note
190 there is a likelihood that the process is not independent “among counties
191 within each state, potentially due to variable resource allocation and differing
192 health systems across states” (p. 3). A mixed linear model accounts for this
193 by introducing random components (in the case of SWN, random intercepts at
194 the state level). SWN estimated various models with different combinations
195 of variables, including median household income and prevalence of travel by
196 private transportation. These are sensible controls, given potential variations
197 in behavior: people in more affluent counties may have greater opportunities
198 to work from home, and use of private transportation reduces contact with
199 strangers. Moreover, they also conducted various sensitivity analyses. After
200 these efforts, SWN conclude that there is a positive association between the
201 basic reproductive number and population density at the level of counties in the
202 US.

203 One salient aspect of the analysis in SWN is that the basic reproductive
204 number can only be calculated reliably with a minimum number of cases, and a
205 large number of counties did not meet such threshold. As researchers do, SWN
206 made modelling decisions, in this case basing their analysis only on counties
207 with valid observations. A modeler with expertise in different methods would
208 likely ask some of the following questions on reading SWN’s paper: how were
209 missing counties treated? What are the implications of the spatial sampling
210 framework used in the analysis? Is it possible to spatially interpolate the missing

211 observations? These questions are relevant and their implications important.
212 Fortunately, SWN are an example of a reasonably open, reproducible research
213 product: their paper is accompanied by (most of) the data and (most of) the
214 code used in the analysis. This means that an independent expert can, with only
215 a moderate investment of time and effort, reproduce the results in the paper, as
216 well as ask additional questions.

217 Alas, reproducibility is not necessarily the norm in the relevant literature.
218 There are various reasons why a project can fail to be reproducible. In some
219 cases, there might be legitimate reasons to withhold the data, perhaps due to
220 confidentiality and privacy reasons (e.g., Lee et al., 2020). But in many other
221 cases the data are publicly available, which in fact has commonly been the case
222 with population-level COVID-19 information. Typically the provenance of the
223 data is documented, but in numerous studies the data themselves are not shared
224 (Amadu et al., 2021; Bhadra et al., 2021; Cruz et al., 2020; Feng et al., 2020;
225 Fielding-Miller et al., 2020; Hamidi et al., 2020a, 2020b; Inbaraj et al., 2021;
226 Souris and Gonzalez, 2020). As any researcher can attest, whether a graduate
227 student or a seasoned scientist, collecting, organizing, and preparing data for a
228 project can take a substantial amount of time. Pointing to the sources of data,
229 even when these sources are public, is a small step towards reproducibility-but
230 only a very small one. Faced with the prospect of having to recreate a data set
231 from raw sources is probably sufficient to dissuade all but the most dedicated
232 (or stubborn) researcher from independent verification. This is true even if part
233 of the data are shared (e.g., Wong and Li, 2020). In other cases, data are shared,
234 but the processes followed in the preparation of the data are not fully documented
235 (Ahmad et al., 2020; Skórka et al., 2020). These processes matter, as shown
236 by the errors in the spreadsheets of Reinhart and Rogoff (Herndon et al., 2014)
237 and the data of biologist Jonathan Pruitt that led to an “avalanche” of paper

²³⁸ retractions⁴. Another situation is when papers share well-documented data, but
²³⁹ fail to provide the code used in the analysis (Noury et al., 2021; Pequeno et
²⁴⁰ al., 2020; Wang et al., 2021). Making code available only “on demand” (e.g.,
²⁴¹ Brandtner et al., 2021) is an unnecessary barrier when most journals offer the
²⁴² facility to share supplemental materials online. Then there are those papers that
²⁴³ more closely comply with reproducibility standards, and share well-documented
²⁴⁴ processes and data, as well as the code used in any analyses reported (Feyman
²⁴⁵ et al., 2020; Paez et al., 2020; Stephens et al., 2021; Sy et al., 2021; White and
²⁴⁶ Hébert-Dufresne, 2020).

²⁴⁷ In the following sections, the analysis of RWN is reproduced, some relevant
²⁴⁸ questions from the perspective of an independent researcher are asked, and the
²⁴⁹ data are reanalyzed.

²⁵⁰ Reproducing SWN

²⁵¹ SWN examined the association between the basic reproductive number of
²⁵² COVID-19 and population density. The basic reproductive number R_0 is a
²⁵³ summary measure of contact rates, probability of transmission of a pathogen,
²⁵⁴ and duration of infectiousness. In rough terms, R_0 measures how many new
²⁵⁵ infections each infections begets. Infectious disease outbreaks generally tend
²⁵⁶ to die out when $R_0 < 1$, and to grow when $R_0 > 1$. Reliable calculation of
²⁵⁷ R_0 requires a minimum number of cases to be able to assume that there is
²⁵⁸ community transmission of the pathogen. Accordingly, SWN based their analysis
²⁵⁹ only on counties that had at least 25 cases or more at the end of the exponential
²⁶⁰ growth phase (see Fig. 1). Their final sample included 1,151 counties in the US,
²⁶¹ including in Alaska, Hawaii, Puerto Rico, and island territories.

²⁶² Table 1 reproduces the first three models of SWN (the fourth model did

⁴<https://doi.org/10.1038/d41586-020-00287-y>

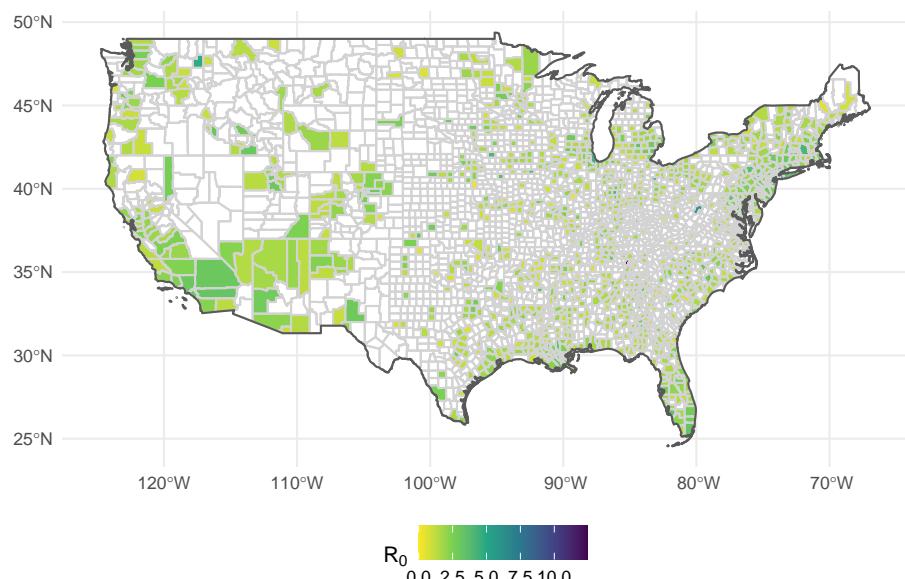


Figure 1: Basic reproductive rate in US counties (Alaska, Hawaii, Puerto Rico, and territories not shown).

Table 1: Reproducing SWN: Models 1-3

Variable	Model 1		Model 2		Model 3	
	beta	95% CI	beta	95% CI	beta	95% CI
Intercept	2.274	[2.167, 2.381]	3.347	[2.676, 4.018]	3.386	[2.614, 4.157]
Log of population density	0.162	[0.133, 0.191]	0.145	[0.115, 0.176]	0.147	[0.113, 0.18]
Percent of private transportation			-0.013	[-0.02, -0.005]	-0.013	[-0.021, -0.005]
Median household income (\$10,000)					-0.003	[-0.033, 0.026]
Standard deviation (Intercept)	0.166	[0.108, 0.254]	0.136	[0.081, 0.229]	0.137	[0.081, 0.232]
Within-group standard error	0.665	[0.638, 0.693]	0.665	[0.638, 0.693]	0.665	[0.638, 0.694]

not have any significant variables; see Table 1 in SWN). It is possible to verify that the results match, with only the minor (and irrelevant) exception of the magnitude of the coefficient for travel by private transportation, which is due to a difference in the input (here the variable is one percent units, whereas in SWN it was ten percent units). The mixed linear model gives random intercepts (i.e., the intercept is a random variable), and the standard deviation is reported in the fourth row of Table 1. It is useful to map the random intercepts: as seen in Figure 2, other things being equal, counties in Texas tend to have somewhat lower values of R_0 (i.e., a negative random intercept), whereas counties in South Dakota tend to have higher values of R_0 . The key of the analysis, after extensive sensitivity analysis, is a robust finding that population density has a positive association with the basic reproductive number. But does it?

Expanding on SWN

The preceding section shows that thanks to the availability of code and data, it is possible to verify the results reported by SWN. As noted earlier, though, an independent researcher might have wondered about the implications of the spatial sampling procedure used by SWN. The decision to use a sample of counties with reliable basic reproductive numbers, although apparently sensible, results in a non-random spatial sampling scheme. Turning our attention back to Figure 1, we form the impression that many counties without reliable values

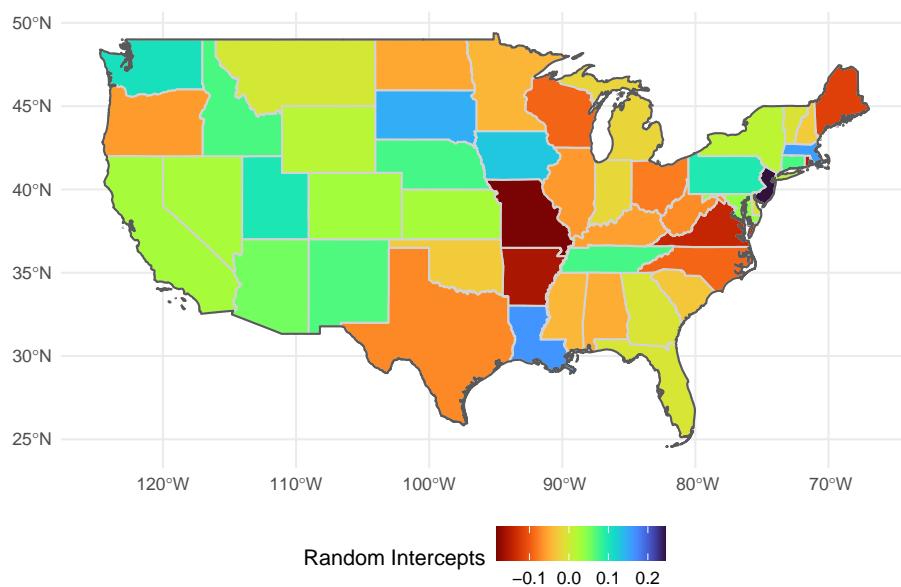
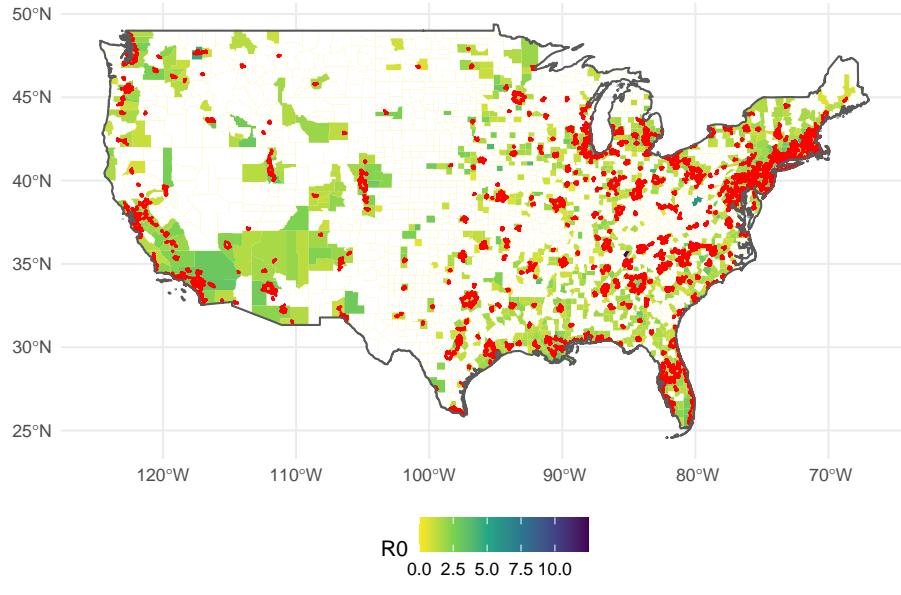


Figure 2: Random intercepts of Model 3 (Alaska, Hawaii, Puerto Rico, and territories not shown).



Note: boundaries of urbanized areas with population > 50,000 are shown in red

Figure 3: Urban areas with population > 50,000 (Alaska, Hawaii, Puerto Rico, and territories not shown).

of R_0 are in more rural, less dense parts of the United States. This impression is reinforced when we overlay the boundaries of urban areas with population greater than 50,000 on the counties with valid values of R_0 (see Figure 3). The fact that R_0 could not be accurately computed in many counties without large urban areas does not mean that there was no transmission of the virus: it simply means that we do not know with precision whether that was the case. The low number of cases may be related to low population and/or low population density. This is intriguing, to say the least: by excluding cases based on the ability to calculate R_0 we are potentially *censoring* the sample in a non-random way.

A problematic issue with non-random sample selection is that parameter estimates can become unreliable, and numerous techniques have been developed over time to address this. A model useful for sample selection problems is

²⁹⁵ Heckman's selection model (see Maddala, 1983). The selection model is in fact a
²⁹⁶ system of two equations, as follows:

$$y_i^{S*} = \beta^{S'} x_i^S + \epsilon_i^S$$

$$y_i^{O*} = \beta^{O'} x_i^O + \epsilon_i^O$$

²⁹⁷ where y_i^{S*} is a latent variable for the sample selection process, and y_i^{O*} is
²⁹⁸ the latent outcome. Vectors x_i^S and x_i^O are explanatory variables (with the
²⁹⁹ possibility that $x_i^S = x_i^O$). Both equations include random terms (i.e., ϵ_i^S and
³⁰⁰ ϵ_i^O) The first equation is designed to model the *probability* of sampling, and the
³⁰¹ second equation the outcome of interest (say R_0). The random terms are jointly
³⁰² distributed and correlated with parameter ρ .

What the analyst observes is the following:

$$y_i^S = \begin{cases} 0 & \text{if } y_i^{S*} < 0 \\ 1 & \text{otherwise} \end{cases}$$

and:

$$y_i^O = \begin{cases} 0 & \text{if } y_i^S = 0 \\ y_i^{O*} & \text{otherwise} \end{cases}$$

³⁰³ In other words, the outcome of interest is observed *only* for certain cases
³⁰⁴ ($y_i^S = 1$, i.e., for sampled observations). The probability of sampling depends on
³⁰⁵ x_i^S . For the cases observed, the outcome y_i^O depends on x_i^O .

³⁰⁶ A sample selection model is estimated using the same selection of variables as
³⁰⁷ SWN Model 3. This is Sample Selection Model 1 in Table 2. The first thing to
³⁰⁸ notice about this model is that the sample selection process and the outcome are
³⁰⁹ not independent ($\rho \neq 0$ with 5% of confidence). The selection equation indicates
³¹⁰ that the probability of a county to be in the sample increases with population
³¹¹ density (but at a decreasing rate due to the log-transformation), when travel by

312 private modes is more prevalent, and as median household income in the county
313 is higher. This is in line with the impression left by Figure 3 that counties with
314 reliable values of R_0 tended to be those with larger urban centers. Once that the
315 selection probabilities are accounted for in the model, several things happen with
316 the outcomes model. First, the coefficient for population density is still positive,
317 but the magnitude changes: in effect, it appears that the effect of density is more
318 pronounced than what SWN Model 3 indicated. The coefficient for percent of
319 private transportation changes signs. And the coefficient for median household
320 income is now significant.

321 The second model in Table 2 (Selection Model 2) changes the way the
322 variables are entered into the model. The log-transformation of density in SWN
323 and Selection Model 1 assumes that the association between density and R_0 is
324 monotonically increasing (if the sign of the coefficient is positive) or decreasing
325 (if the sign of the coefficient is negative). There are some indications that the
326 relationship may actually not be monotonical. For example, Paez et al. (2020)
327 found a positive (if non-significant) relationship between density and incidence
328 of COVID-19 in the provinces of Spain at the beginning of the pandemic. This
329 changed to a negative (and significant) relationship during the lockdown. In
330 the case of the US, Fielding-Miller et al. (2020) found that the association
331 between COVID-19 deaths and population density was positive in rural counties,
332 but negative in urban counties. A variable transformation that allows for non-
333 monotonic changes in the relationship is the square of the density.

334 As seen in the table, Selection Model 2 replaces the log-transformation of
335 population density with a quadratic expansion. The results of this analysis
336 indicate that with this variable transformation, the selection and outcome
337 processes are still not independent ($\rho \neq 0$ with 5% of confidence). But a few
338 interesting things emerge. When we examine the outcomes model, we see that

Table 2: Estimation results of sample selection models

Variable	Selection Model 1		Selection Model 2	
	β	95% CI	β	95% CI
Sample Selection Model				
Intercept	-2.237	[-3.109, -1.365]	-7.339	[-8.381, -6.297]
Log of population density	0.385	[0.352, 0.418]		
Density (1,000 per sq.km)			2.484	[2.13, 2.838]
Density squared			-0.387	[-0.473, -0.3]
Percent of private transportation	0.025	[0.016, 0.034]	0.057	[0.046, 0.067]
Median household income (10,000)	0.202	[0.168, 0.235]	0.32	[0.283, 0.357]
Outcome Model				
Intercept	0.605	[-0.257, 1.466]	2.784	[1.652, 3.915]
Log of population density	0.39	[0.354, 0.426]		
Density (1,000 per sq.km)			0.758	[0.509, 1.008]
Density squared			-0.132	[-0.187, -0.077]
Percent of private transportation	0.01	[0.001, 0.018]	-0.011	[-0.021, -0.001]
Median household income (\$10,000)	0.126	[0.094, 0.159]	0.002	[-0.033, 0.037]
σ	0.954	[0.904, 1.003]	0.684	[0.652, 0.716]
ρ	0.971	[0.961, 0.98]	-0.199	[-0.377, -0.022]

339 the quadratic expansion has a positive coefficient for the first order term, but a
 340 negative coefficient for the second order term. This indicates that R_0 initially
 341 tends to increase with higher density, but only up to a point, after which the
 342 negative second term (which grows more rapidly due to the square), becomes
 343 increasingly dominant. Secondly, the sign of the coefficient for travel by private
 344 transportation becomes negative again. This, of course, makes more sense
 345 than the positive sign of Selection Model 1: if people tend to travel in private
 346 transportation, the potential for contact should be lower instead of higher. And
 347 finally median household income is no longer significant.

348 How relevant is the difference between these different model specifications?
 349 Figure 4 shows the relationship between density and R_0 implied by SWN Model
 350 1 and Selection Model 2. The left panel of the figure shows the non-linear but
 351 monotonic relationship implied by SWN Model 1. The conclusion is that at
 352 higher densities, R_0 is *always* higher. The right panel, in contrast, shows that,
 353 according to Selection Model 2, R_0 is zero when density is zero (as expected),

³⁵⁴ and then it tends to increase at higher densities. This continues until a density
³⁵⁵ of approximately 2.9 (1,000 people per sq.km). At higher densities than that R_0
³⁵⁶ begins to decline, and the relationship becomes negative at densities higher than
³⁵⁷ approximately 5.7 (1,000 people per sq.km).

³⁵⁸ Thus, other things being equal, the effect of density in a county like Char-
³⁵⁹ lottesville in Virginia (density ~1,639 people per sq.km) is roughly the same as
³⁶⁰ that in a county like Philadelphia (density ~4,127 people per sq.km). In contrast,
³⁶¹ the effect of density on R_0 in a county like Arlington in Virginia (density ~3,093
³⁶² people per sq.km) is *stronger* than either of the previous two examples. Lastly,
³⁶³ the density of counties like San Francisco in California, or Queens and Bronx in
³⁶⁴ NY, which are among the densest in the US, contributes even less to R_0 than
³⁶⁵ even the most rural counties in the country.

³⁶⁶ It is worth at this point to recall Cressie's dictum about modelling: “[w]hat
³⁶⁷ is one person's mean structure could be another person's correlation structure”
³⁶⁸ (Cressie, 1989, p. 201). There are almost always multiple ways to approach a
³⁶⁹ modelling situation. In the present case, we would argue that spatial sampling
³⁷⁰ is an important aspect of the modeling process, but one that perhaps required
³⁷¹ different technical skills than those available to SWN. There is nothing wrong
³⁷² with that. What matters is that, by adopting relatively high reproducibility
³⁷³ standards, these researchers made a valuable and honest contribution to the
³⁷⁴ collective enterprise of seeking knowledge. Their effort, and subsequent efforts
³⁷⁵ to validate and expand on their work, can potentially contribute to provide
³⁷⁶ clarity to ongoing conversations about the relevance of density and the spread of
³⁷⁷ COVID-19.

³⁷⁸ In particular, it is noteworthy that a sample selection model with a different
³⁷⁹ variable transformation does not lend support to the thesis that higher density is
³⁸⁰ *always* associated with a greater risk of spread of the virus [as put by Wong and

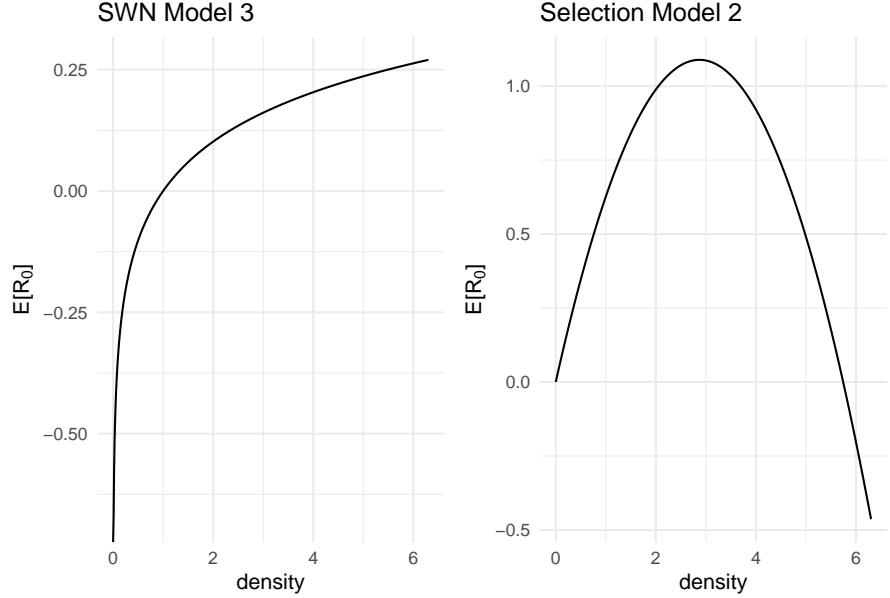


Figure 4: Effect of density according to SWN Model 3 and Sample Selection Model 2.

381 Li, “‘Density is destiny’ is probably an overstatement”, (2020)]. At the same
 382 time, this also stands in contrast to the findings of Hamidi et al., who found that
 383 higher density was either not significantly associated with the rate of the virus in
 384 a cross-sectional study (Hamidi et al., 2020b), or was negatively associated with
 385 in a longitudinal setting [Hamidi et al. (2020a). In this sense, the conclusion that
 386 density does not aggravate the pandemic may have been somewhat premature;
 387 instead, reanalysis of the data of SWN suggests that Fielding-Miller et al. (2020)
 388 might be onto something with respect to the difference between rural and urban
 389 counties. More generally, in population-level studies, density is indicative of
 390 proximity, no doubt about that, but also for adaptive behavior. And it is possible
 391 that the determining factor during COVID-19, at least in the US, has been
 392 variations in perceptions of the risks associated with contagion (Chauhan et al.,
 393 2021), and subsequent compensations in behavior in more and less dense regions.

394 **Conclusion**

395 The tension between the need to publish research potentially useful in dealing
396 with a global pandemic, and a “carnage of substandard research” (Bramstedt,
397 2020), highlights the importance of efforts to maintain the quality of scientific
398 outputs during COVID-19. An important part of quality control is the ability of
399 independent researchers to verify and examine the results of materials published
400 in the literature. As previous research illustrates, reproducibility in scientific
401 research remains an important but elusive goal (Gustot, 2020; e.g., Iqbal et al.,
402 2016; Stodden et al., 2018; Sumner et al., 2020). This idea is reinforced by the
403 review conducted for this paper in the context of research about population
404 density and the spread of COVID-19.

405 Taking one recent example from the literature [Sy et al., Sy et al. (2021);
406 SWN], the present paper illustrates the importance of good reproducibility
407 practices. Sharing data and code can catalyze research, by allowing independent
408 verification of findings, as well as additional research. After verifying the results of
409 SWN, experiments with sample selection models and variations in the definition
410 of model inputs, lead to an important reappraisal of the conclusion that high
411 density is associated with greater spread of the virus. Instead, the possibility
412 of a non-monotonical relationship between population density and contagion is
413 raised.

414 In the spirit of openness, this paper is prepared as an R Markdown document,
415 an a companion data package is provided. The data package contains the relevant
416 documentation of the data, and all data pre-processing is fully documented.
417 Hopefully this, and similar reproducible papers, will continue to encourage others
418 to adopt reproducible standards in their research.

419 **References**

- 420 Ahmad, K., Erqou, S., Shah, N., Nazir, U., Morrison, A.R., Choudhary, G.,
421 Wu, W.-C., 2020. Association of poor housing conditions with COVID-
422 19 incidence and mortality across US counties. PLOS ONE 15, e0241327.
423 doi:10.1371/journal.pone.0241327
- 424 Amadu, I., Ahinkorah, B.O., Afitiri, A.-R., Seidu, A.-A., Ameyaw, E.K., Hagan,
425 J.E., Duku, E., Aram, S.A., 2021. Assessing sub-regional-specific strengths of
426 healthcare systems associated with COVID-19 prevalence, deaths and recov-
427 eries in africa. PLOS ONE 16, e0247274. doi:10.1371/journal.pone.0247274
- 428 Añazco, D., Nicolalde, B., Espinosa, I., Camacho, J., Mushtaq, M., Gimenez, J.,
429 Teran, E., 2021. Publication rate and citation counts for preprints released
430 during the COVID-19 pandemic: The good, the bad and the ugly. PeerJ 9,
431 e10927. doi:10.7717/peerj.10927
- 432 Basu, S., Carney, M.A., Kenworthy, N.J., 2017. Ten years after the financial
433 crisis: The long reach of austerity and its global impacts on health. Social
434 Science & Medicine 187, 203–207. doi:10.1016/j.socscimed.2017.06.026
- 435 Bhadra, A., Mukherjee, A., Sarkar, K., 2021. Impact of population density on
436 covid-19 infected and mortality rate in india. Modeling Earth Systems and
437 Environment 7, 623–629. doi:10.1007/s40808-020-00984-7
- 438 Bramstedt, K.A., 2020. The carnage of substandard research during the COVID-
439 19 pandemic: A call for quality. Journal of Medical Ethics 46, 803–807.
440 doi:10.1136/medethics-2020-106494
- 441 Brandtner, C., Bettencourt, L.M.A., Berman, M.G., Stier, A.J., 2021. Crea-
442 tures of the state? Metropolitan counties compensated for state inaction
443 in initial u.s. Response to COVID-19 pandemic. PLOS ONE 16, e0246249.
444 doi:10.1371/journal.pone.0246249
- 445 Broggini, F., Dellinginger, J., Fomel, S., Liu, Y., 2017. Reproducible research: Geo-
446 physics papers of the future - introduction. Geophysics 82. doi:10.1190/geo2017-

- 447 0918-spseintro.1
- 448 Brunsdon, C., Comber, A., 2020. Opening practice: Supporting reproducibil-
- 449 ity and critical spatial data science. *Journal of Geographical Systems*.
- 450 doi:10.1007/s10109-020-00334-2
- 451 Chauhan, R.S., Capasso da Silva, D., Salon, D., Shamshiripour, A., Rahimi,
- 452 E., Sutradhar, U., Khoeini, S., Mohammadian, A.(Kouros)., Derrible, S.,
- 453 Pendyala, R., 2021. COVID-19 related attitudes and risk perceptions
- 454 across urban, rural, and suburban areas in the united states. *Findings*.
- 455 doi:10.32866/001c.23714
- 456 Cressie, N., 1989. Geostatistics. *The American Statistician* 43, 197. doi:10.2307/2685361
- 457 Cruz, C.J.P., Ganly, R., Li, Z., Gietel-Basten, S., 2020. Exploring the young de-
- 458 mographic profile of COVID-19 cases in hong kong: Evidence from migration
- 459 and travel history data. *PLOS ONE* 15, e0235306. doi:10.1371/journal.pone.0235306
- 460 Farber, S., Páez, A., 2011. Running to stay in place: The time-use implications
- 461 of automobile oriented land-use and travel. *Journal of Transport Geography*
- 462 19, 782–793. doi:10.1016/j.jtrangeo.2010.09.008
- 463 Feng, Y., Li, Q., Tong, X., Wang, R., Zhai, S., Gao, C., Lei, Z., Chen, S., Zhou,
- 464 Y., Wang, J., Yan, X., Xie, H., Chen, P., Liu, S., Xv, X., Liu, S., Jin, Y.,
- 465 Wang, C., Hong, Z., Luan, K., Wei, C., Xu, J., Jiang, H., Xiao, C., Guo, Y.,
- 466 2020. Spatiotemporal spread pattern of the COVID-19 cases in china. *PLOS*
- 467 *ONE* 15, e0244351. doi:10.1371/journal.pone.0244351
- 468 Feyman, Y., Bor, J., Raifman, J., Griffith, K.N., 2020. Effectiveness of COVID-19
- 469 shelter-in-place orders varied by state. *PLOS ONE* 15, e0245008. doi:10.1371/journal.pone.0245008
- 470 Fielding-Miller, R.K., Sundaram, M.E., Brouwer, K., 2020. Social determinants
- 471 of COVID-19 mortality at the county level. *PLOS ONE* 15, e0240151.
- 472 doi:10.1371/journal.pone.0240151
- 473 Florida, R., Glaeser, E., Sharif, M., Bedi, K., Campanella, T., Chee, C., Doctoroff,

- 474 D., Katz, B., Katz, R., Kotkin, J., 2020. How life in our cities will look after
475 the coronavirus pandemic. *Foreign Policy* 1.
- 476 Fraser, N., Brierley, L., Dey, G., Polka, J.K., Pálfy, M., Nanni, F., Coates,
477 J.A., 2021. The evolving role of preprints in the dissemination of COVID-19
478 research and their impact on the science communication landscape. *PLOS*
479 *Biology* 19, e3000959. doi:10.1371/journal.pbio.3000959
- 480 Gomez, J., Prieto, J., Leon, E., Rodríguez, A., 2021. INFEKTA—an agent-based
481 model for transmission of infectious diseases: The COVID-19 case in bogotá,
482 colombia. *PLOS ONE* 16, e0245787. doi:10.1371/journal.pone.0245787
- 483 Gustot, T., 2020. Quality and reproducibility during the COVID-19 pandemic.
484 *JHEP Rep* 2, 100141. doi:10.1016/j.jhepr.2020.100141
- 485 Hamidi, S., Ewing, R., Sabouri, S., 2020a. Longitudinal analyses of the rela-
486 tionship between development density and the COVID-19 morbidity and
487 mortality rates: Early evidence from 1,165 metropolitan counties in the united
488 states. *Health & Place* 64, 102378. doi:10.1016/j.healthplace.2020.102378
- 489 Hamidi, S., Sabouri, S., Ewing, R., 2020b. Does density aggravate the COVID-
490 19 pandemic? *Journal of the American Planning Association* 86, 495–509.
491 doi:10.1080/01944363.2020.1777891
- 492 Hamidi, S., Zandiatashbar, A., 2021. Compact development and adherence
493 to stay-at-home order during the COVID-19 pandemic: A longitudinal
494 investigation in the united states. *Landscape and Urban Planning* 205,
495 103952. doi:<https://doi.org/10.1016/j.landurbplan.2020.103952>
- 496 Harris, M.A., Branić-Calles, M., 2021. Changes in commute mode attributed to
497 COVID-19 risk in canadian national survey data. *Findings*. doi:10.32866/001c.19088
- 498 Herndon, T., Ash, M., Pollin, R., 2014. Does high public debt consistently stifle
499 economic growth? A critique of reinhart and rogoff. *Cambridge Journal of*
500 *Economics* 38, 257–279. doi:10.1093/cje/bet075

- 501 Inbaraj, L.R., George, C.E., Chandrasingh, S., 2021. Seroprevalence of COVID-19
502 infection in a rural district of south india: A population-based seroepidemiological study. PLOS ONE 16, e0249247. doi:10.1371/journal.pone.0249247
- 503 504 Ince, D.C., Hatton, L., Graham-Cumming, J., 2012. The case for open computer
505 programs. Nature 482, 485–488. doi:10.1038/nature10836
- 506 507 Ioannidis, J.P.A., Greenland, S., Hlatky, M.A., Khoury, M.J., Macleod, M.R.,
508 Moher, D., Schulz, K.F., Tibshirani, R., 2014. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 383, 166–175.
509 doi:10.1016/s0140-6736(13)62227-8
- 510 511 Iqbal, S.A., Wallach, J.D., Khoury, M.J., Schully, S.D., Ioannidis, J.P.A., 2016.
512 Reproducible research practices and transparency across the biomedical
literature. Plos Biology 14. doi:10.1371/journal.pbio.1002333
- 513 514 Jamal, S., Paez, A., 2020. Changes in trip-making frequency by mode during
COVID-19. Findings. doi:10.32866/001c.17977
- 515 516 Kadi, N., Khelfaoui, M., 2020. Population density, a factor in the spread of
COVID-19 in algeria: Statistic study. Bulletin of the National Research
517 Centre 44. doi:10.1186/s42269-020-00393-x
- 518 519 Khavarian-Garmsir, A.R., Sharifi, A., Moradpour, N., 2021. Are high-density
districts more vulnerable to the COVID-19 pandemic? Sustainable Cities
and Society 70, 102911. doi:10.1016/j.scs.2021.102911
- 520 521 Konkol, M., Kray, C., 2019. In-depth examination of spatiotemporal figures
in open reproducible research. Cartography and Geographic Information
522 Science 46, 412–427. doi:10.1080/15230406.2018.1512421
- 523 524 Konkol, M., Kray, C., Pfeiffer, M., 2019. Computational reproducibility in
geoscientific papers: Insights from a series of studies with geoscientists and
525 a reproduction study. International Journal of Geographical Information
526 Science 33, 408–429. doi:10.1080/13658816.2018.1508687

- 528 Kwon, D., 2020. How swamped preprint servers are blocking bad coronavirus
529 research. *Nature* 581, 130–132.
- 530 Lee, M., Zhao, J., Sun, Q., Pan, Y., Zhou, W., Xiong, C., Zhang, L., 2020. Human
531 mobility trends during the early stage of the COVID-19 pandemic in the
532 united states. *PLOS ONE* 15, e0241468. doi:10.1371/journal.pone.0241468
- 533 Li, R., Richmond, P., Roehner, B.M., 2018. Effect of population density on
534 epidemics. *Physica A: Statistical Mechanics and its Applications* 510, 713–724.
535 doi:10.1016/j.physa.2018.07.025
- 536 Maddala, G.S., 1983. Limited-dependent and qualitative variables in economet-
537 rics. Cambridge University Press, Cambridge.
- 538 Micallef, S., Piscopo, T.V., Casha, R., Borg, D., Vella, C., Zammit, M.-A.,
539 Borg, J., Mallia, D., Farrugia, J., Vella, S.M., Xerri, T., Portelli, A., Fenech,
540 M., Fsadni, C., Mallia Azzopardi, C., 2020. The first wave of COVID-
541 19 in malta; a national cross-sectional study. *PLOS ONE* 15, e0239389.
542 doi:10.1371/journal.pone.0239389
- 543 Molloy, J., Tchervenkov, C., Hintermann, B., Axhausen, K.W., 2020. Trac-
544 ing the sars-CoV-2 impact: The first month in switzerland. Findings.
545 doi:10.32866/001c.12903
- 546 Moore, E.G., 1970. Some spatial properties of urban contact fields. *Geographical
547 Analysis* 2, 376–386.
- 548 Moore, E.G., Brown, L.A., 1970. Urban acquaintance fields: An evaluation of a
549 spatial model. *Environment and Planning* 2, 443–454.
- 550 Noland, R.B., 1995. PERCEIVED RISK AND MODAL CHOICE - RISK
551 COMPENSATION IN TRANSPORTATION SYSTEM. *Accident Analysis
552 and Prevention* 27, 503–521. doi:10.1016/0001-4575(94)00087-3
- 553 Noland, R.B., 2021. Mobility and the effective reproduction rate of COVID-19.
554 *Journal of Transport & Health* 20, 101016. doi:<https://doi.org/10.1016/j.jth>.

555 2021.101016

556 Noury, A., François, A., Gergaud, O., Garel, A., 2021. How does COVID-19
557 affect electoral participation? Evidence from the french municipal elections.

558 PLOS ONE 16, e0247026. doi:10.1371/journal.pone.0247026

559 Paez, A., 2020. Using google community mobility reports to investigate the
560 incidence of COVID-19 in the united states. Findings. doi:<https://doi.org/10.32866/001c.12976>

561 Paez, A., Lopez, F.A., Menezes, T., Cavalcanti, R., Pitta, M.G. da R., 2020.
562 A spatio-temporal analysis of the environmental correlates of COVID-19
563 incidence in spain. Geographical Analysis n/a. doi:10.1111/gean.12241

564 Pequeno, P., Mendel, B., Rosa, C., Bosholn, M., Souza, J.L., Baccaro, F.,
565 Barbosa, R., Magnusson, W., 2020. Air transportation, population density
566 and temperature predict the spread of COVID-19 in brazil. PeerJ 8, e9322.
567 doi:10.7717/peerj.9322

568 Phillips, R.O., Fyhri, A., Sagberg, F., 2011. Risk compensation and bicycle
569 helmets. Risk Analysis 31, 1187–1195. doi:10.1111/j.1539-6924.2011.01589.x

570 Praharaj, S., King, D., Pettit, C., Wentz, E., 2020. Using aggregated mobility
571 data to measure the effect of COVID-19 policies on mobility changes in
572 sydney, london, phoenix, and pune. Findings. doi:10.32866/001c.17590

573 Richens, J., Imrie, J., Copas, A., 2000. Condoms and seat belts: The parallels
574 and the lessons. Lancet 355, 400–403. doi:10.1016/s0140-6736(99)09109-6

575 Rocklöv, J., Sjödin, H., 2020. High population densities catalyse the spread of
576 COVID-19. Journal of Travel Medicine 27. doi:10.1093/jtm/taaa038

577 Roy, S., Ghosh, P., 2020. Factors affecting COVID-19 infected and death
578 rates inform lockdown-related policymaking. PLOS ONE 15, e0241165.
579 doi:10.1371/journal.pone.0241165

580 Sharifi, A., Khavarian-Garmsir, A.R., 2020. The COVID-19 pandemic: Impacts

582 on cities and major lessons for urban planning, design, and management.

583 Science of The Total Environment 749, 142391. doi:<https://doi.org/10.1016/j.scitotenv.2020.142391>

585 Skórka, P., Grzywacz, B., Moroń, D., Lenda, M., 2020. The macroecology of
586 the COVID-19 pandemic in the anthropocene. PLOS ONE 15, e0236856.
587 doi:[10.1371/journal.pone.0236856](https://doi.org/10.1371/journal.pone.0236856)

588 Souris, M., Gonzalez, J.-P., 2020. COVID-19: Spatial analysis of hospital case-
589 fatality rate in france. PLOS ONE 15, e0243606. doi:[10.1371/journal.pone.0243606](https://doi.org/10.1371/journal.pone.0243606)

590 Stephens, K.E., Chernyavskiy, P., Bruns, D.R., 2021. Impact of altitude on
591 COVID-19 infection and death in the united states: A modeling and obser-
592 vational study. PLOS ONE 16, e0245055. doi:[10.1371/journal.pone.0245055](https://doi.org/10.1371/journal.pone.0245055)

593 Stodden, V., Seiler, J., Ma, Z.K., 2018. An empirical analysis of journal pol-
594 icy effectiveness for computational reproducibility. Proceedings of the Na-
595 tional Academy of Sciences of the United States of America 115, 2584–2589.
596 doi:[10.1073/pnas.1708290115](https://doi.org/10.1073/pnas.1708290115)

597 Sumner, J., Haynes, L., Nathan, S., Hudson-Vitale, C., McIntosh, L.D., 2020.
598 Reproducibility and reporting practices in COVID-19 preprint manuscripts.
599 medRxiv 2020.03.24.20042796. doi:[10.1101/2020.03.24.20042796](https://doi.org/10.1101/2020.03.24.20042796)

600 Sun, Z., Zhang, H., Yang, Y., Wan, H., Wang, Y., 2020. Impacts of geographic
601 factors and population density on the COVID-19 spreading under the lock-
602 down policies of china. Science of The Total Environment 746, 141347.
603 doi:[10.1016/j.scitotenv.2020.141347](https://doi.org/10.1016/j.scitotenv.2020.141347)

604 Sy, K.T.L., White, L.F., Nichols, B.E., 2021. Population density and basic
605 reproductive number of COVID-19 across united states counties. PLOS ONE
606 16, e0249271. doi:[10.1371/journal.pone.0249271](https://doi.org/10.1371/journal.pone.0249271)

607 Vlasschaert, C., Topf, J.M., Hiremath, S., 2020. Proliferation of papers and
608 preprints during the coronavirus disease 2019 pandemic: Progress or prob-

- 609 lems with peer review? *Advances in Chronic Kidney Disease* 27, 418–426.
- 610 doi:10.1053/j.ackd.2020.08.003
- 611 Wang, F., Tan, Z., Yu, Z., Yao, S., Guo, C., 2021. Transmission and control pres-
- 612 sure analysis of the COVID-19 epidemic situation using multisource spatio-
- 613 temporal big data. *PLOS ONE* 16, e0249145. doi:10.1371/journal.pone.0249145
- 614 White, E.R., Hébert-Dufresne, L., 2020. State-level variation of initial COVID-19
- 615 dynamics in the united states. *PLOS ONE* 15, e0240648. doi:10.1371/journal.pone.0240648
- 616 Wong, D.W.S., Li, Y., 2020. Spreading of COVID-19: Density matters. *PLOS*
- 617 ONE 15, e0242398. doi:10.1371/journal.pone.0242398