

¹ Reproducibility of research during COVID-19:
² examining the case of population density and the basic
³ reproductive rate from the perspective of spatial analysis

⁴ Author^{*,a}

⁵ *^aDepartment, Street, City, State ZIP*

⁶ **Abstract**

The emergence of the novel SARS-CoV-2 coronavirus and the global COVID-19 pandemic in 2019 led to explosive growth in scientific research. Alas, much of the research in the literature lacks conditions to be reproducible, and recent publications on the association between population density and the basic reproductive number of SARS-CoV-2 are no exception. Relatively few papers share code and data sufficiently, which hinders not only verification but additional experimentation. In this paper, an example of reproducible research shows the potential of spatial analysis for epidemiology research during COVID-19. Transparency and openness means that independent researchers can, with only modest efforts, verify findings and use different approaches as appropriate. Given the high stakes of the situation, it is essential that scientific findings, on which good policy depends, are as robust as possible; as the empirical example shows, reproducibility is one of the keys to ensure this.

*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 ruinous 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 termed the deluge of COVID-19 publications a “carnage of substandard
²⁰ research” (Bramstedt, 2020). Part of the challenge of maintaining quality
²¹ standards in published research is that, despite an abundance of recommendations
²² and guidelines (e.g., Broggini et al., 2017; Brunsdon and Comber, 2020; Ince et
²³ al., 2012; Ioannidis et al., 2014), in practice reproducibility has remained a lofty
²⁴ and 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 software and data (Arribas-Bel et al., 2021; e.g., Bivand,
³⁰ 2020), along with more strenuous efforts towards open, reproducible research,
³¹ is simply a continuation of long-standing scientific practices of independent
³² verification. Despite the (at times disproportionate) attention that high profile
³³ scandals in science tend to elicit in the media, science as a collective endeavor is
³⁴ remarkable for being a self-correcting enterprise, one with built-in mechanisms
³⁵ and incentives to weed out erroneous ideas. Over the long term, facts tend to
³⁶ prevail in science. At stake is the shorter-term impacts that research may have
³⁷ in other spheres of economic and social life. The case of economists Reinhart and
³⁸ Rogoff comes to mind: by the time the inaccuracies and errors in their research
³⁹ were uncovered (see Herndon et al., 2014), their claims about debt and economic
⁴⁰ growth had already been seized by policy-makers on both sides of the Atlantic to

41 justify austerity policies in the aftermath of the Great Recession of 2007-2009¹.
42 As later research has demonstrated, those policies cast a long shadow, and their
43 sequels continued to 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 (e.g.,
46 Florida et al., 2020); as a result, the implications of the pandemic for urban
47 planning, design, and management are the topic of ongoing research (e.g., Sharifi
48 and Khavarian-Garmsir, 2020). The fact that the worst of the pandemic was
49 initially felt in dense population centers such as Wuhan, Milan, Madrid, and
50 New York, unleashed a torrent of research into the associations between density
51 and the spread of the pandemic. The answers to some important questions hang
52 on the results of these research efforts. For example, are lower density regions
53 safer from the pandemic? Are de-densification policies warranted, even if just in
54 the short term? In the longer term, will the risks of life in high density regions
55 presage a flight from cities? And, what are the implications of the pandemic
56 for future urban planning and practice? Over the past year, numerous papers
57 have sought to throw light on the underlying issue of density and the pandemic;
58 nonetheless the results, as will be detailed next, remain mixed. Further, to
59 complicate matters, precious few of these studies appear to be sufficiently open
60 to support independent verification.

61 The objective of this paper is to illustrate the importance of reproducibility
62 in research in the context of the flood of COVID-19 papers. For this, I focus
63 on a recent study by Sy et al. (2021) that examined the correlation between
64 the basic reproductive number of COVID-19, R_0 , and population density. The
65 basic reproductive number is a summary measure of contact rates, probability
66 of transmission of a pathogen, and duration of infectiousness. In rough terms,
67 it measures how many new infections each infections begets. The paper of Sy
68 et al. (2021) was selected for being, in the literature examined, almost alone
69 in supporting reproducible research. Accordingly, I wish to be clear that my
70 objective in singling their work for discussion is not to malign their efforts, but
71 rather to demonstrate how open and reproducible research efforts can greatly
72 help to accelerate discovery. More concretely, open data and open code mean

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>

73 that an independent researcher can, with only modest efforts, not only verify
74 the findings reported, but also examine the same data from a perspective which
75 may not have been available to the original researchers due to differences in
76 disciplinary perspectives, methodological traditions, and/or training, among
77 other possible factors. The example, which shows consequential changes in the
78 conclusions reached by different analyses, should serve as a call to researchers
79 to redouble their efforts to increase transparency and reproducibility in their
80 research. In this spirit, the present paper also aims to show how data can be
81 packaged in well-documented, shareable units, and code can be embedded into
82 self-contained documents suitable for review and independent verification. The
83 source for this paper is an R Markdown document which, along with the data
84 package, are available in a public repository².

85 **Background: the intuitive relationship between density and spread of
86 contagious diseases**

87 The concern with population density and the spread of the virus during the
88 COVID-19 pandemic was fueled, at least in part, by dramatic scenes seen in
89 real-time around the world from large urban centers such as Wuhan, Milan,
90 Madrid, and New York. In theory, there are good reasons to believe that higher
91 density could have a positive association with the transmission of a contagious
92 virus. It has long been known that the potential for inter-personal contact is
93 greater in regions with higher density (see for example the research on urban
94 fields and time-geography, including Farber and Páez, 2011; Moore, 1970; Moore
95 and Brown, 1970). Mathematically, models of exposure and contagion indicate
96 that higher densities can catalyze the transmission of contagious diseases (Li et
97 al., 2018; Rocklöv and Sjödin, 2020). The idea is intuitive and likely at the root
98 of messages, by some figures in positions of authority, that regions with sparse
99 population densities faced lower risks from the pandemic³.

100 As Rocklöv and Sjödin (Rocklöv and Sjödin, 2020) note, however, mathematical
101 models of contagion are valid at small-to-medium spaces (and presumably,
102 smaller time intervals too, such as time spent in restaurants, concert halls,

²<https://github.com/paezha/Reproductive-Rate-and-Density-US-Reanalyzed>

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>

103 cruises), and the results do not necessarily transfer to larger spatial units and
104 longer time periods. There are solid reasons for this: while in a restaurant, one
105 can hardly avoid being in proximity to other customers. On the other hand, a
106 person can choose to (or be forced to as a matter of policy) not go to a restaurant
107 in the first place. Nonetheless, the idea that high density correlates with high
108 transmission is so seemingly sensible that it is often taken for granted even at the
109 scale of large spaces (e.g., Cruz et al., 2020; Micallef et al., 2020). In such condi-
110 tions, however, there exists the possibility of behavioral adaptations, which are
111 difficult to capture in the mechanistic framework of differential equations (or can
112 be missing in agent-based models, e.g., Gomez et al., 2021); these adaptations,
113 in fact, can be a key aspect of disease transmission.

114 A plausible behavioral adaptation during a pandemic, especially one broadcast
115 as widely and intensely as COVID-19, is risk compensation. Risk compensation
116 is a process whereby people adjust their behavior in response to their *perception*
117 of risk (Noland, 1995; Phillips et al., 2011; Richens et al., 2000). In the case of
118 COVID-19, Chauhan et al. (Chauhan et al., 2021) have found that perception
119 of risks in the US varies between rural, suburban, and urban residents, with
120 rural residents in general expressing less concern about the virus. It is possible
121 that people who listened to the message of leaders saying that they were safe
122 from the virus because of low density may not have taken adequate precautions.
123 Conversely, people in dense places who could more directly observe the impact
124 of the pandemic may have become overly cautious. Both Paez et al. (2020) and
125 Hamidi et al. (2020b) posit this mechanism (i.e., greater compliance with social
126 distancing in denser regions) to explain the results of their analyses. The evidence
127 available does indeed show that there were important changes in behavior with
128 respect to mobility during the pandemic (Harris and Braniion-Calles, 2021; Jamal
129 and Paez, 2020; Molloy et al., 2020); furthermore, shelter in place orders may
130 have had greater buy-in from the public in higher density regions (Feyman et al.,
131 2020; Hamidi and Zandiataashbar, 2021), and the associated behavior may have
132 persisted beyond the duration of official social-distancing policies (Prahraj et
133 al., 2020). In addition, there is evidence that changes in mobility correlated with
134 the trajectory of the pandemic (Noland, 2021; Paez, 2020). Given the potential
135 for behavioral adaptation, the question of density becomes more nuanced: it
136 is not just a matter of proximity, but also of human behavior, which is better
137 studied using population-level data and models.

¹³⁸ **Background: but what does the literature say?**

¹³⁹ When it comes to population density and the spread of COVID-19, the
¹⁴⁰ international literature to date remains inconclusive.

¹⁴¹ On the one hand, there are studies that report positive associations between
¹⁴² population density and various COVID-19-related outcomes. Bhadra (2021),
¹⁴³ for example, reported a moderate positive correlation between the spread of
¹⁴⁴ COVID-19 and population density at the district level in India, however their
¹⁴⁵ analysis was bivariate and did not control for other variables, such as income.
¹⁴⁶ Similarly, Kadi and Khelfaoui (2020) found a positive and significant correlation
¹⁴⁷ between number of cases and population density in cities in Algeria in a series
¹⁴⁸ of simple regression models (i.e., without other controls). A question in these
¹⁴⁹ relatively simple analyses is whether density is not a proxy for other factors.
¹⁵⁰ Other studies have included controls, such as Pequeno et al. (2020), a team
¹⁵¹ that reported a positive association between density and cumulative counts
¹⁵² of confirmed COVID-19 cases in state capitals in Brazil after controlling for
¹⁵³ covariates, including income, transport connectivity, and economic status. In
¹⁵⁴ a similar vein, Fielding-Miller et al. (2020) reported a positive relationship
¹⁵⁵ between the absolute number of COVID-19 deaths and population density (rate)
¹⁵⁶ in rural counties in the US. Roy and Ghosh (2020) used a battery of machine
¹⁵⁷ learning techniques to find discriminatory factors, and a positive and significant
¹⁵⁸ association between COVID-19 infection and death rates in US states. Wong and
¹⁵⁹ Li (2020) also found a positive and significant association between population
¹⁶⁰ density and number of confirmed COVID-19 cases in US counties, using both
¹⁶¹ univariate and multivariate regressions with spatial effects. More recently, Sy
¹⁶² et al. (2021) reported that the basic reproductive number of COVID-19 in US
¹⁶³ counties tended to increase with population density, but at a decreasing rate at
¹⁶⁴ higher densities.

¹⁶⁵ On the flip side, a number of studies report non-significant or negative
¹⁶⁶ associations between population density and COVID-19 outcomes. This includes
¹⁶⁷ the research of Sun et al. (2020) who did not find evidence of significant
¹⁶⁸ correlation between population density and confirmed number of cases per day
¹⁶⁹ *in conditions of lockdown* in China. This finding echoes the results of Paez et
¹⁷⁰ al. (2020), who in their study of provinces in Spain reported non-significant
¹⁷¹ associations between population density and infection rates in the early days
¹⁷² of the first wave of COVID-19, and negative significant associations in the
¹⁷³ later part of the first lockdown. Similarly, Skórka et al. (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, Khavarian-Garmsir et al. (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) and Hamidi et al. (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 referred to as SWN, would immediately grab the attention of a researcher with expertise in spatial analysis.

Reproducibility of research

SWN investigated the basic reproductive number of COVID-19 in US counties, and its association with population density, median household income, and prevalence of private mobility. For their multivariate analysis, SWN used mixed linear models. This is an appropriate modelling choice: R_0 is an interval-ratio variable that is suitably modeled using linear regression; further, as SWN note there is a likelihood that the process is not independent “among counties within each state, potentially due to variable resource allocation and differing health systems across states” (p. 3). A mixed linear model accounts for this by introducing random components; in the case of SWN, these are random intercepts at the state level. SWN estimated various models with different combinations of variables, including median household income and prevalence of travel by private transportation. These controls help to account for potential variations in behavior: people in more affluent counties may have greater opportunities to work from home, and use of private transportation reduces contact with

209 strangers. Moreover, they also conducted various sensitivity analyses. After
210 these efforts, SWN concluded that there is a positive association between the
211 basic reproductive number and population density at the level of counties in the
212 US.

213 One salient aspect of the analysis in SWN is that the basic reproductive
214 number can only be calculated reliably with a minimum number of cases, and a
215 large number of counties did not meet such threshold. As researchers do, SWN
216 made modelling decisions, in this case basing their analysis only on counties
217 with valid observations. A modeler with expertise in spatial analysis would
218 likely ask some of the following questions on reading SWN's paper: how were
219 missing counties treated? What are the implications of the spatial sampling
220 framework used in the analysis? Is it possible to spatially interpolate the missing
221 observations? Was there spatial residual autocorrelation in the models, or was the
222 use of mixed models sufficient to capture spatial dependencies? These questions
223 are relevant and their implications important. Fortunately, SWN are an example
224 of a reasonably open, reproducible research product: their paper is accompanied
225 by (most of) the data and (most of) the code used in the analysis. This means
226 that an independent researcher can, with only a moderate investment of time
227 and effort, reproduce the results in the paper, as well as ask additional questions.

228 Alas, reproducibility is not necessarily the norm in the relevant literature.

229 There are various reasons why a project can fail to be reproducible. In some
230 cases, there might be legitimate reasons to withhold the data, perhaps due to
231 confidentiality and privacy reasons (e.g., Lee et al., 2020). But in many other
232 cases the data are publicly available, which in fact has commonly been the case
233 with population-level COVID-19 information. Typically the provenance of the
234 data is documented, but in numerous studies the data themselves are not shared
235 (Amadu et al., 2021; Bhadra et al., 2021; Cruz et al., 2020; Feng et al., 2020;
236 Fielding-Miller et al., 2020; Hamidi et al., 2020a, 2020b; Inbaraj et al., 2021;
237 Souris and Gonzalez, 2020). As any researcher can attest, collecting, organizing,
238 and preparing data for a project can take a substantial amount of time. Pointing
239 to the sources of data, even when these sources are public, is a small step towards
240 reproducibility-but only a very small one. Faced with the prospect of having to
241 recreate a data set from raw sources is probably sufficient to dissuade all but the
242 most dedicated (or stubborn) researcher from independent verification. This is
243 true even if part of the data are shared (e.g., Wong and Li, 2020). In other cases,
244 data are shared, but the processes followed in the preparation of the data are
245 not fully documented (Ahmad et al., 2020; Skórka et al., 2020). These processes

matter, as shown by the errors in the spreadsheets of Reinhart and Rogoff (see Herndon et al., 2014 for the discovery of these errors), as well as by the data of biologist Jonathan Pruitt that led to an “avalanche” of paper retractions (see Viglione, 2020). Another situation is when papers share well-documented data, but fail to provide the code used in the analysis (Noury et al., 2021; Pequeno et al., 2020; Wang et al., 2021). Making code available only “on demand” (e.g., Brandtner et al., 2021) is an unnecessary barrier when most journals offer the facility to share supplemental materials online. Then there are those papers that more closely comply with reproducibility standards, and share well-documented processes and data, as well as the code used in any analyses reported (Feyman et al., 2020; Paez et al., 2020; Stephens et al., 2021; Sy et al., 2021; White and Hébert-Dufresne, 2020). Even in this case, the pressure to publish “new findings” instead of replication studies can act as a deterrent, perhaps particularly for younger researchers⁴.

In the following sections, the analysis of SWN is reproduced, some relevant questions from the perspective of an independent researcher with expertise in spatial analysis are asked, and the data are reanalyzed.

Reproducing SWN

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

⁴The present paper was desk rejected by three journals that had previously published research on population density and the spread of COVID-19; in one case, the paper was too opinionated for the journal, in the other two cases, the paper was not a “good fit” despite dealing with a nearly identical issue as papers previously published in said journals.

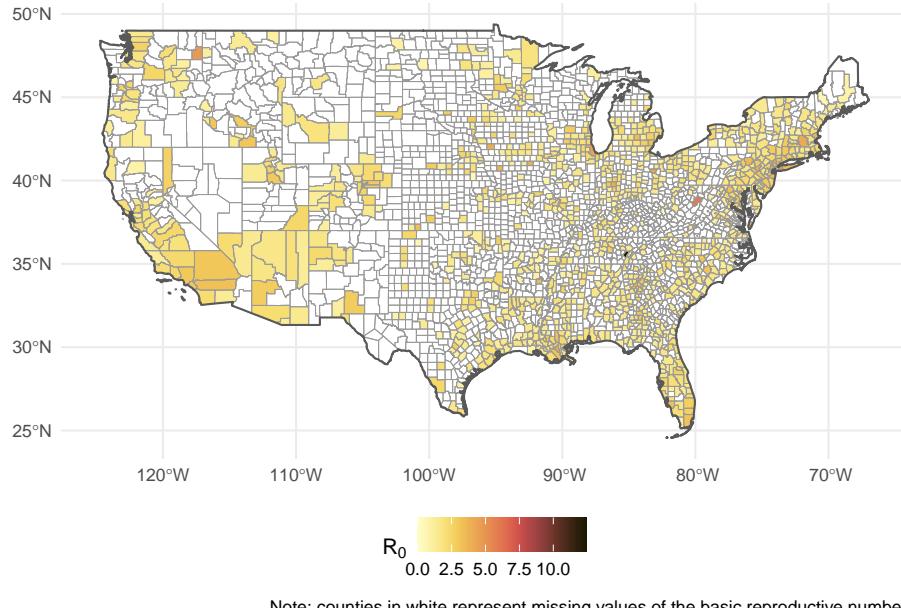


Figure 1: Basic reproductive rate in US counties (Alaska, Hawaii, Puerto Rico, and territories not shown).

versioning) in a GitHub repository⁵. For each county, SWN assumed that the exponential growth period began one week prior to the second daily increase in cases, and assumed that the period of exponential growth lasted approximately 18 days.

Table 1 reproduces the first three models of SWN (the fourth model did 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 changed to one percent units, instead of the ten percent units used by SWN). The mixed linear model gives random intercepts (i.e., the intercept is a random variable), and the standard deviation is reported in the fifth 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

⁵<https://github.com/nytimes/covid-19-data>

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]

290 counties in South Dakota tend to have higher values of R_0 . The key of the
 291 analysis, after extensive sensitivity analysis, is a robust finding that population
 292 density has a positive association with the basic reproductive number. But does
 293 it?

294 Expanding on SWN

295 The preceding section shows that thanks to the availability of code and
 296 data, it is possible to verify the results reported by SWN. As noted earlier,
 297 though, an independent researcher might have wondered about the implications
 298 of the spatial sampling procedure used by SWN. The decision to use a sample of
 299 counties with reliable basic reproductive numbers, although apparently sensible,
 300 results in a non-random spatial sampling scheme. Turning our attention back
 301 to Figure 1, we form the impression that many counties without reliable values
 302 of R_0 are in more rural, less dense parts of the United States. This impression
 303 is reinforced when we overlay the boundaries of urban areas with population
 304 greater than 50,000 on the counties with valid values of R_0 (see Figure 3). The
 305 fact that R_0 could not be accurately computed in many counties without large
 306 urban areas does not mean that there was no transmission of the virus: it simply
 307 means that we do not know with sufficient precision to what extent that was
 308 the case. The low number of cases may be related to low population and/or
 309 low population density. This is intriguing, to say the least: by excluding cases
 310 based on the ability to calculate R_0 we are potentially *selecting* the sample in a
 311 non-random way.

A problematic issue with non-random sample selection is that parameter estimates can become unreliable, and numerous techniques have been developed to address this. A model useful for sample selection problems is Heckman's selection model (see Maddala, 1983). The selection model is in fact a system of

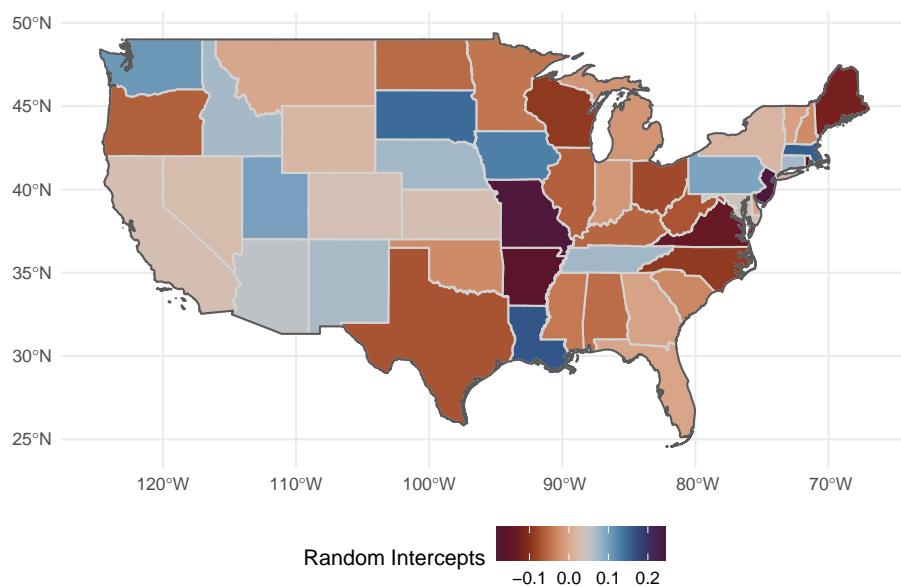


Figure 2: Random intercepts of Model 3 (Alaska, Hawaii, Puerto Rico, and territories not shown).

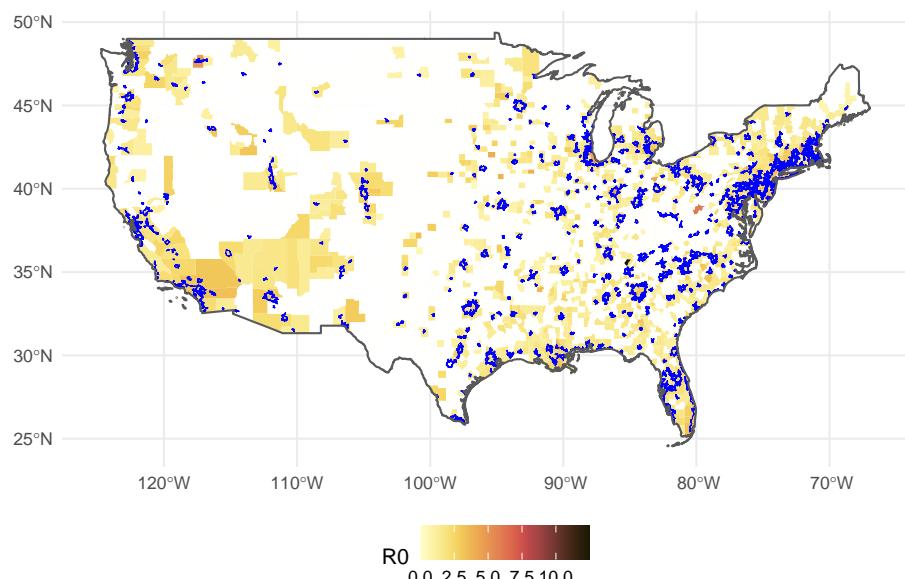


Figure 3: Urban areas with population > 50,000 (Alaska, Hawaii, Puerto Rico, and territories not shown).

two equations, as follows:

$$\begin{aligned} y_i^{S*} &= \beta^{S'} x_i^S + \epsilon_i^S \\ y_i^{O*} &= \beta^{O'} x_i^O + \epsilon_i^O \end{aligned}$$

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

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

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

A sample selection model is estimated using the same selection of variables as SWN Model 3. This is Sample Selection Model 1 in Table 2. The first thing to notice about this model is that the sample selection process and the outcome are correlated ($\rho \neq 0$ with 5% of confidence). The selection equation indicates that the probability of a county to be in the sample increases with population density (but at a decreasing rate due to the log-transformation), when travel by private modes is more prevalent, and as median household income in the county is higher. This is in line with the impression made by Figure 3 that counties with reliable values of R_0 tended to be those with larger urban centers. Once that the selection probabilities are accounted for in the model, several things happen with the outcomes model. First, the coefficient for population density is still positive, but the magnitude changes: in effect, it appears that the effect of density is more pronounced than what SWN Model 3 indicated. The coefficient for percent of private transportation changes signs. And the coefficient for median household income is now significant.

336 The second model in Table 2 (Selection Model 2) changes the way the
337 variables are entered into the model. The log-transformation of density in SWN
338 and Selection Model 1 assumes that the association between density and R_0 is
339 monotonically increasing (if the sign of the coefficient is positive) or decreasing
340 (if the sign of the coefficient is negative). There are some indications that the
341 relationship may actually not be monotonical. For example, Paez et al. (2020)
342 found a positive (if non-significant) relationship between density and incidence
343 of COVID-19 in the provinces of Spain at the beginning of the pandemic. This
344 changed to a negative (and significant) relationship during the lockdown. In
345 the case of the US, Fielding-Miller et al. (2020) found that the association
346 between COVID-19 deaths and population density was positive in rural counties,
347 but negative in urban counties. A variable transformation that allows for non-
348 monotonic changes in the relationship is the square of the density.

349 As seen in the table, Selection Model 2 replaces the log-transformation of
350 population density with a quadratic expansion. The results of this analysis
351 indicate that with this variable transformation, the selection and outcome
352 processes are still correlated ($\rho \neq 0$ with 5% of confidence). But a few other
353 interesting things emerge. When we examine the outcomes model, we see that
354 the quadratic expansion has a positive coefficient for the first order term, but a
355 negative coefficient for the second order term. This indicates that R_0 initially
356 tends to increase as density grows, but only up to a point, after which the
357 negative second term (which grows more rapidly due to the square), becomes
358 increasingly dominant. Secondly, the sign of the coefficient for travel by private
359 transportation becomes negative again. This, of course, makes more sense
360 than the positive sign of Selection Model 1: if people tend to travel in private
361 transportation, the potential for contact should be lower instead of higher. And
362 finally median household income is no longer significant, similar to SWN Model
363 3.

364 **Proceed with caution: spatial effects ahead**

365 The results of the selection models, in particular Selection Model 2, make
366 us reassess the original conclusion that density has a positive association with
367 the basic reproductive number of COVID-19. A spatial analyst might still
368 wonder about spatial residual autocorrelation. A challenge here is that spatial
369 models tend to be technically more demanding, and although spatial models
370 for qualitative variables exist, a spatial implementation of the sample selection

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]

371 model does not appear to exist. It might be argued that a reproducible research
 372 project can also allow a researcher to be more adventurous with their modeling
 373 decisions: since data and code are shared, other researchers can promptly and
 374 with relative ease poke the methods and see if they appear to be sound.

375 In the present case, it appears that an application of spatial filtering (see
 376 Getis and Griffith, 2002; Griffith, 2004; Paez, 2019) can help. Spatial filtering
 377 provides an elegant solution to regression problems that may have difficulties
 378 handling the spatial structures of spatial statistical and econometric models
 379 (Griffith, 2000). A key issue in the present example is the fact that there are
 380 numerous missing observations, which prevents the calculation of autocorrelation
 381 statistics, let alone the estimation of models with spatial components.

382 The following is an unorthodox, but potentially effective use of filters in a
 383 sample selection model:

- 384 1. Estimate a sample selection model and retrieve the residuals of the outcome.
 This will be a vector with missing values for locations that were not sampled.
- 385 2. Fit a spatial filter to the residuals. This is done by regressing the estimated
 residuals of the *observed* data on the corresponding values of the Moran
 eigenvectors.
- 386 3. The resulting filter will correlate highly with the known residuals, and will
 provide information in non-sampled locations that is consistent with the

391 spatial pattern of the known residuals.
 392 4. Test the filter for spatial autocorrelation:
 393 4.1 If significant spatial autocorrelation is detected, this would be indicative
 394 of residual spatial pattern. Introduce the filter as a covariate in the outcome
 395 model of the sample selection model and return to step 1.
 396 4.2 If no significant spatial autocorrelation is detected, this would be
 397 indicative of random residual pattern. Stop.
 398 This procedure is implemented using a stopping criterion whereby the search
 399 for the filter only stops when the p-value of Moran's Coefficient of the filter
 400 fitted to the residuals is greater than 0.25, which was chosen as a sufficiently
 401 conservative value for testing for autocorrelation. The correlation of the known
 402 residuals with the corresponding elements of the filter is consistently high (the
 403 correlation coefficient typically is greater than 0.9). The results of implementing
 404 this procedure appear in Table 3 as Selection Model 3. The results are consistent
 405 with Selection Model 2, with two intriguing differences: 1) the variance of
 406 Sample Model 3 is smaller; and 2) the sample and outcome processes are
 407 no longer correlated (the confidence interval of ρ includes zero). It appears
 408 that by capturing the spatial pattern of the residuals, which is likely strongly
 409 determined by the non-random sampling framework, the outcome model is not
 410 only substantially more precise, but also appears to be independent from the
 411 selection process.

Table 3: Estimation results of sample selection model with spatial filter

Variable	Selection Model 3	
	β	95% CI
Sample Selection Model		
Intercept	-7.304	[-8.346, -6.262]
Density (1,000 per sq.km)	2.445	[2.089, 2.802]
Density squared	-0.380	[-0.468, -0.292]
Percent of private transportation	0.056	[0.046, 0.067]
Median household income (10,000)	0.318	[0.28, 0.356]
Outcome Model		
Intercept	2.563	[2.497, 2.629]
Density (1,000 per sq.km)	0.760	[0.746, 0.774]
Density squared	-0.133	[-0.135, -0.13]
Percent of private transportation	-0.011	[-0.012, -0.011]
Median household income (\$10,000)	0.002	[-0.001, 0.004]
Spatial filter	1.000	[0.998, 1.001]
σ	0.017	[0.015, 0.019]
ρ	-0.304	[-0.957, 0.349]

412 Clearly, the various models display some intriguing differences; but how
413 relevant are said differences from a more substantive standpoint? Figure 4 shows
414 the relationship between density and R_0 implied by SWN Model 3, Selection
415 Model 2, and Selection Model 3. The left panel of the figure shows the non-linear
416 but monotonic relationship implied by SWN Model 1. The conclusion is that at
417 higher densities, R_0 is *always* higher. The two panels on the right, in contrast,
418 shows that Selection Model 2 and Selection Model 3 coincide that R_0 tends
419 to increase as density grows. This continues until a density of approximately
420 2.9 (1,000 people per sq.km). At higher densities than that the relationship
421 between density and R_0 begins to weaken, and the relationship becomes negative
422 at densities higher than approximately 5.7 (1,000 people per sq.km).

423 To put this into context, other things being equal, the effect of density in
424 a county like Charlottesville in Virginia (density ~1,639 people per sq.km) is
425 roughly the same as that in a county like Philadelphia (density ~4,127 people
426 per sq.km). In contrast, the effect of density on R_0 in a county like Arlington in
427 Virginia (density ~3,093 people per sq.km) is *stronger* than either of the previous
428 two examples. Lastly, the density of counties like San Francisco in California, or
429 Queens and Bronx in NY, which are among the densest in the US, contributes
430 even less to R_0 than even the most rural counties in the country.

431 Discussion

432 It is worth at this point to recall Cressie's dictum about modelling: "[w]hat
433 is one person's mean structure could be another person's correlation structure"
434 (Cressie, 1989, p. 201). There are almost always multiple ways to approach a
435 modelling situation, as lively illustrated by a recent paper that reports the results
436 of a crowdsourced modelling experiment (Schweinsberg et al., 2021). In the
437 present case, we would argue that spatial sampling is an important aspect of the
438 modeling process. Importantly, by adopting high reproducibility standards, SWN
439 made a valuable contribution to the collective enterprise of seeking knowledge.
440 Their effort, and subsequent efforts to validate and expand on their work, can
441 potentially contribute to provide clarity to ongoing conversations about the
442 relevance of density and the spread of COVID-19.

443 In particular, it is noteworthy that a sample selection model with a different
444 variable transformation does not lend support to the thesis that higher density
445 is *always* associated with a greater risk of spread of the virus [in Wong and
446 Li's words, "‘Density is destiny’ is probably an overstatement"; (2020)]. At

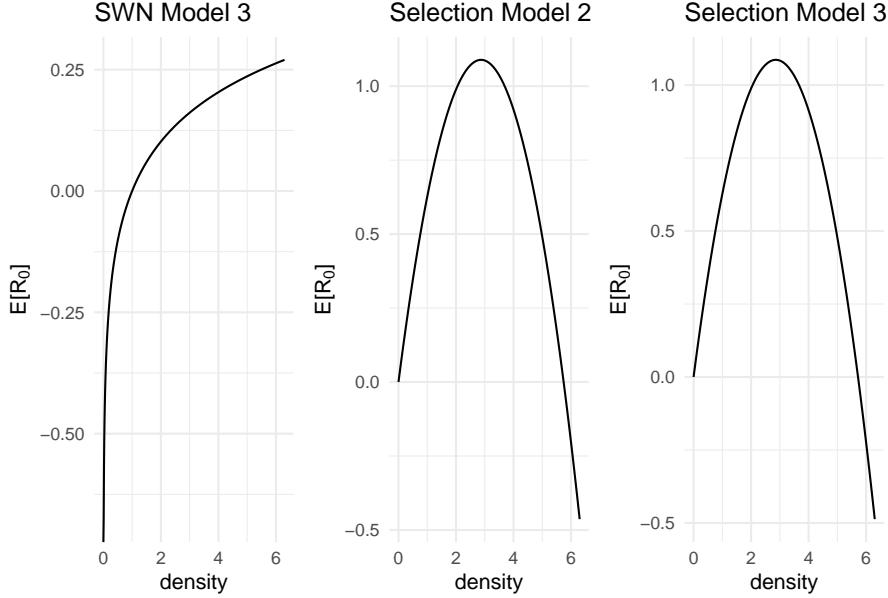


Figure 4: Effect of density according to SWN Model 3 and Sample Selection Model 2.

447 the same time, the results presented here also stand in contrast to the findings
 448 of Hamidi et al., who found that higher density was either not significantly
 449 associated with the rate of the virus in a cross-sectional study (Hamidi et al.,
 450 2020b), or was negatively associated with it in a longitudinal setting [Hamidi et
 451 al. (2020a)]. In this sense, the conclusion that density does not aggravate the
 452 pandemic may have been somewhat premature; instead, reanalysis of the data of
 453 SWN suggests that Fielding-Miller et al. (2020) might be onto something with
 454 respect to the difference between rural and urban counties. More generally, there
 455 is no doubt that in population-level studies density is indicative of proximity,
 456 but it also potentially is a proxy for adaptive behavior. And it is possible that
 457 the determining factor during COVID-19, at least in the US, has been variations
 458 in perceptions of the risks associated with contagion (Chauhan et al., 2021), and
 459 subsequent compensations in behavior in more and less dense regions.

460 **Conclusion**

461 The tension between the need to publish research potentially useful in dealing
 462 with a global pandemic, and a potential “carnage of substandard research”

(Bramstedt, 2020), highlights the importance of efforts to maintain the quality of scientific outputs during COVID-19. An important part of quality control is the ability of independent researchers to verify and examine the results of materials published in the literature. As previous research illustrates, reproducibility in scientific research remains an important but elusive goal (Gustot, 2020; e.g., Iqbal et al., 2016; Stodden et al., 2018; Sumner et al., 2020). This idea is reinforced by the review conducted for this paper in the context of research about population density and the spread of COVID-19.

Taking one recent example from the literature [Sy et al., Sy et al. (2021); SWN], the present paper illustrates the importance of good reproducibility practices. Sharing data and code can catalyze research, by allowing independent verification of findings, as well as additional research. After verifying the results of SWN, experiments with sample selection models and variations in the definition of model inputs, lead to an important reappraisal of the conclusion that high density is associated with greater spread of the virus. Instead, the possibility of a non-monotonical relationship between population density and contagion is raised. I do not claim that the analysis presented here is the last word on the topic of density and the spread of COVID-19, and there is always the possibility that someone else will be better equipped to analyze these data with greater competence. By opening up the analysis, documenting the way data were pre-processed, and by sharing analysis ready data, my hope would be that others will be able to discover the limitations of my own analysis and improve on it, as appropriate.

More generally, my hope is that the research of Sy et al. (2021), the present paper, and similar reproducible publications, will continue to encourage others to adopt higher reproducibility standards in their research.

Acknowledgments

The analysis reported in this paper was conducted in the R computing statistical language (R Core Team, 2021). The source document is an Rmark-down document (Xie et al., 2020, 2018) processed using `knitr` (Xie, 2015, 2014). The following packages were used in the analysis, and I wish to acknowledge their creators for their generous efforts: `adespatial` (Dray et al., 2021), `censReg` (Henningsen, 2020), `dplyr` (Wickham et al., 2021), `forcats` (Wickham, 2021a), `ggplot2` (Wickham, 2016), `gmm` (Chaussé, 2010), `kableExtra` (Zhu, 2021), `Matrix` (Bates and Maechler, 2021), `maxLik` (Henningsen and Toomet,

498 2011), **miscTools** (Henningsen and Toomet, 2019), **mvtnorm** (Genz and Bretz,
499 **nlme** (Pinheiro et al., 2021), **patchwork** (Pedersen, 2020), **purrr** (Henry
500 and Wickham, 2020), **readr** (Wickham and Hester, 2021), **sampleSelection**
501 (Toomet and Henningsen, 2008), **sandwich** (Zeileis et al., 2020; Zeileis, 2006,
502 2004), **scico** (Pedersen and Cramer, 2020), **sf** (Pebesma, 2018), **sp** (Bivand
503 et al., 2013a; Pebesma and Bivand, 2005), **spatialprobit** (Wilhelm and de
504 Matos, 2015), **spData** (Bivand et al., 2021), **spdep** (Bivand et al., 2013b; Bivand
505 and Wong, 2018), **stringr** [R-stringr], **tibble** (Müller and Wickham, 2021),
506 **tidycensus** (Walker and Herman, 2021), **tidyrr** (Wickham, 2021b), **tidyverse**
507 (Wickham et al., 2019), **tmvtnorm** (Manjunath B G <bgmanjunath@gmail.com>,
508 2015), **units** (Pebesma et al., 2016). This research was not supported by
509 Canada's Research Councils.

510 **References**

- 511 Ahmad, K., Erqou, S., Shah, N., Nazir, U., Morrison, A.R., Choudhary, G.,
512 Wu, W.-C., 2020. Association of poor housing conditions with COVID-
513 19 incidence and mortality across US counties. PLOS ONE 15, e0241327.
514 doi:10.1371/journal.pone.0241327
- 515 Amadu, I., Ahinkorah, B.O., Afitiri, A.-R., Seidu, A.-A., Ameyaw, E.K., Hagan,
516 J.E., Duku, E., Aram, S.A., 2021. Assessing sub-regional-specific strengths of
517 healthcare systems associated with COVID-19 prevalence, deaths and recov-
518 eries in africa. PLOS ONE 16, e0247274. doi:10.1371/journal.pone.0247274
- 519 Añazco, D., Nicolalde, B., Espinosa, I., Camacho, J., Mushtaq, M., Gimenez, J.,
520 Teran, E., 2021. Publication rate and citation counts for preprints released
521 during the COVID-19 pandemic: The good, the bad and the ugly. PeerJ 9,
522 e10927. doi:10.7717/peerj.10927
- 523 Arribas-Bel, D., Green, M., Rowe, F., Singleton, A., 2021. Open data products: A
524 framework for creating valuable analysis-ready data. Journal of Geographical
525 Systems.
- 526 Basu, S., Carney, M.A., Kenworthy, N.J., 2017. Ten years after the financial
527 crisis: The long reach of austerity and its global impacts on health. Social
528 Science & Medicine 187, 203–207. doi:10.1016/j.socscimed.2017.06.026
- 529 Bates, D., Maechler, M., 2021. Matrix: Sparse and dense matrix classes and
530 methods.
- 531 Bhadra, A., Mukherjee, A., Sarkar, K., 2021. Impact of population density on
532 covid-19 infected and mortality rate in india. Modeling Earth Systems and

- 533 Environment 7, 623–629. doi:10.1007/s40808-020-00984-7
- 534 Bivand, R., Nowosad, J., Lovelace, R., 2021. spData: Datasets for spatial
535 analysis.
- 536 Bivand, R.S., 2020. Progress in the r ecosystem for representing and handling
537 spatial data. Journal of Geographical Systems. doi:10.1007/s10109-020-
538 00336-0
- 539 Bivand, R.S., Pebesma, E., Gomez-Rubio, V., 2013a. Applied spatial data
540 analysis with R, second edition. Springer, NY.
- 541 Bivand, R.S., Pebesma, E., Gomez-Rubio, V., 2013b. Applied spatial data
542 analysis with R, second edition. Springer, NY.
- 543 Bivand, R., Wong, D.W.S., 2018. Comparing implementations of global and
544 local indicators of spatial association. TEST 27, 716–748.
- 545 Bramstedt, K.A., 2020. The carnage of substandard research during the COVID-
546 19 pandemic: A call for quality. Journal of Medical Ethics 46, 803–807.
547 doi:10.1136/medethics-2020-106494
- 548 Brandtner, C., Bettencourt, L.M.A., Berman, M.G., Stier, A.J., 2021. Creatures
549 of the state? Metropolitan counties compensated for state inaction
550 in initial u.s. Response to COVID-19 pandemic. PLOS ONE 16, e0246249.
551 doi:10.1371/journal.pone.0246249
- 552 Broggini, F., Dellinger, J., Fomel, S., Liu, Y., 2017. Reproducible research: Geo-
553 physics papers of the future - introduction. Geophysics 82. doi:10.1190/geo2017-
554 0918-spseintro.1
- 555 Brunsdon, C., Comber, A., 2020. Opening practice: Supporting reproducibil-
556 ity and critical spatial data science. Journal of Geographical Systems.
557 doi:10.1007/s10109-020-00334-2
- 558 Chauhan, R.S., Capasso da Silva, D., Salon, D., Shamshiripour, A., Rahimi,
559 E., Sutradhar, U., Khoeini, S., Mohammadian, A.(Kouros)., Derrible, S.,
560 Pendyala, R., 2021. COVID-19 related attitudes and risk perceptions
561 across urban, rural, and suburban areas in the united states. Findings.
562 doi:10.32866/001c.23714
- 563 Chaussé, P., 2010. Computing generalized method of moments and generalized
564 empirical likelihood with R. Journal of Statistical Software 34, 1–35.
- 565 Cressie, N., 1989. Geostatistics. The American Statistician 43, 197. doi:10.2307/2685361
- 566 Cruz, C.J.P., Ganly, R., Li, Z., Gietel-Basten, S., 2020. Exploring the young de-
567 mographic profile of COVID-19 cases in hong kong: Evidence from migration
568 and travel history data. PLOS ONE 15, e0235306. doi:10.1371/journal.pone.0235306

- 569 Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guenard, G.,
570 Jombart, T., Larocque, G., Legendre, P., Madi, N., Wagner, H.H., 2021.
571 Adespatial: Multivariate multiscale spatial analysis.
- 572 Farber, S., Páez, A., 2011. Running to stay in place: The time-use implications
573 of automobile oriented land-use and travel. *Journal of Transport Geography*
574 19, 782–793. doi:10.1016/j.jtrangeo.2010.09.008
- 575 Feng, Y., Li, Q., Tong, X., Wang, R., Zhai, S., Gao, C., Lei, Z., Chen, S., Zhou,
576 Y., Wang, J., Yan, X., Xie, H., Chen, P., Liu, S., Xv, X., Liu, S., Jin, Y.,
577 Wang, C., Hong, Z., Luan, K., Wei, C., Xu, J., Jiang, H., Xiao, C., Guo, Y.,
578 2020. Spatiotemporal spread pattern of the COVID-19 cases in china. *PLOS*
579 ONE 15, e0244351. doi:10.1371/journal.pone.0244351
- 580 Feyman, Y., Bor, J., Raifman, J., Griffith, K.N., 2020. Effectiveness of COVID-19
581 shelter-in-place orders varied by state. *PLOS ONE* 15, e0245008. doi:10.1371/journal.pone.0245008
- 582 Fielding-Miller, R.K., Sundaram, M.E., Brouwer, K., 2020. Social determinants
583 of COVID-19 mortality at the county level. *PLOS ONE* 15, e0240151.
584 doi:10.1371/journal.pone.0240151
- 585 Florida, R., Glaeser, E., Sharif, M., Bedi, K., Campanella, T., Chee, C., Doctoroff,
586 D., Katz, B., Katz, R., Kotkin, J., 2020. How life in our cities will look after
587 the coronavirus pandemic. *Foreign Policy* 1.
- 588 Fraser, N., Brierley, L., Dey, G., Polka, J.K., Pálfy, M., Nanni, F., Coates,
589 J.A., 2021. The evolving role of preprints in the dissemination of COVID-19
590 research and their impact on the science communication landscape. *PLOS*
591 Biology 19, e3000959. doi:10.1371/journal.pbio.3000959
- 592 Genz, A., Bretz, F., 2009. Computation of multivariate normal and t probabilities,
593 Lecture notes in statistics. Springer-Verlag, Heidelberg.
- 594 Getis, A., Griffith, D.A., 2002. Comparative spatial filtering in regression analysis.
595 *Geographical Analysis* 34, 130–140.
- 596 Gomez, J., Prieto, J., Leon, E., Rodríguez, A., 2021. INFEKTA—an agent-based
597 model for transmission of infectious diseases: The COVID-19 case in bogotá,
598 colombia. *PLOS ONE* 16, e0245787. doi:10.1371/journal.pone.0245787
- 599 Griffith, D.A., 2000. A linear regression solution to the spatial autocorrelation
600 problem. *Journal of Geographical Systems* 2, 141–156.
- 601 Griffith, D.A., 2004. A spatial filtering specification for the autologistic model.
602 *Environment and Planning A* 36, 1791–1811.
- 603 Gustot, T., 2020. Quality and reproducibility during the COVID-19 pandemic.
604 *JHEP Rep* 2, 100141. doi:10.1016/j.jhepr.2020.100141

- 605 Hamidi, S., Ewing, R., Sabouri, S., 2020a. Longitudinal analyses of the rela-
606 tionship between development density and the COVID-19 morbidity and
607 mortality rates: Early evidence from 1,165 metropolitan counties in the united
608 states. *Health & Place* 64, 102378. doi:10.1016/j.healthplace.2020.102378
- 609 Hamidi, S., Sabouri, S., Ewing, R., 2020b. Does density aggravate the COVID-
610 19 pandemic? *Journal of the American Planning Association* 86, 495–509.
611 doi:10.1080/01944363.2020.1777891
- 612 Hamidi, S., Zandiatashbar, A., 2021. Compact development and adherence
613 to stay-at-home order during the COVID-19 pandemic: A longitudinal
614 investigation in the united states. *Landscape and Urban Planning* 205,
615 103952. doi:<https://doi.org/10.1016/j.landurbplan.2020.103952>
- 616 Harris, M.A., Bramion-Calles, M., 2021. Changes in commute mode attributed to
617 COVID-19 risk in canadian national survey data. *Findings*. doi:10.32866/001c.19088
- 618 Henningsen, A., 2020. censReg: Censored regression (tobit) models.
- 619 Henningsen, A., Toomet, O., 2011. maxLik: A package for maximum likelihood
620 estimation in R. *Computational Statistics* 26, 443–458. doi:10.1007/s00180-
621 010-0217-1
- 622 Henningsen, A., Toomet, O., 2019. miscTools: Miscellaneous tools and utilities.
- 623 Henry, L., Wickham, H., 2020. Purrr: Functional programming tools.
- 624 Herndon, T., Ash, M., Pollin, R., 2014. Does high public debt consistently stifle
625 economic growth? A critique of reinhart and rogoff. *Cambridge Journal of
626 Economics* 38, 257–279. doi:10.1093/cje/bet075
- 627 Inbaraj, L.R., George, C.E., Chandrasingh, S., 2021. Seroprevalence of COVID-19
628 infection in a rural district of south india: A population-based seroepidemiolog-
629 ical study. *PLOS ONE* 16, e0249247. doi:10.1371/journal.pone.0249247
- 630 Ince, D.C., Hatton, L., Graham-Cumming, J., 2012. The case for open computer
631 programs. *Nature* 482, 485–488. doi:10.1038/nature10836
- 632 Ioannidis, J.P.A., Greenland, S., Hlatky, M.A., Khoury, M.J., Macleod, M.R.,
633 Moher, D., Schulz, K.F., Tibshirani, R., 2014. Increasing value and reduc-
634 ing waste in research design, conduct, and analysis. *Lancet* 383, 166–175.
635 doi:10.1016/s0140-6736(13)62227-8
- 636 Iqbal, S.A., Wallach, J.D., Khoury, M.J., Schully, S.D., Ioannidis, J.P.A., 2016.
637 Reproducible research practices and transparency across the biomedical
638 literature. *Plos Biology* 14. doi:10.1371/journal.pbio.1002333
- 639 Jamal, S., Paez, A., 2020. Changes in trip-making frequency by mode during
640 COVID-19. *Findings*. doi:10.32866/001c.17977

- 641 Kadi, N., Khelfaoui, M., 2020. Population density, a factor in the spread of
642 COVID-19 in algeria: Statistic study. Bulletin of the National Research
643 Centre 44. doi:10.1186/s42269-020-00393-x
- 644 Khavarian-Garmsir, A.R., Sharifi, A., Moradpour, N., 2021. Are high-density
645 districts more vulnerable to the COVID-19 pandemic? Sustainable Cities
646 and Society 70, 102911. doi:10.1016/j.scs.2021.102911
- 647 Konkol, M., Kray, C., 2019. In-depth examination of spatiotemporal figures
648 in open reproducible research. Cartography and Geographic Information
649 Science 46, 412–427. doi:10.1080/15230406.2018.1512421
- 650 Konkol, M., Kray, C., Pfeiffer, M., 2019. Computational reproducibility in
651 geoscientific papers: Insights from a series of studies with geoscientists and
652 a reproduction study. International Journal of Geographical Information
653 Science 33, 408–429. doi:10.1080/13658816.2018.1508687
- 654 Kwon, D., 2020. How swamped preprint servers are blocking bad coronavirus
655 research. Nature 581, 130–132.
- 656 Lee, M., Zhao, J., Sun, Q., Pan, Y., Zhou, W., Xiong, C., Zhang, L., 2020. Human
657 mobility trends during the early stage of the COVID-19 pandemic in the
658 united states. PLOS ONE 15, e0241468. doi:10.1371/journal.pone.0241468
- 659 Li, R., Richmond, P., Roehner, B.M., 2018. Effect of population density on
660 epidemics. Physica A: Statistical Mechanics and its Applications 510, 713–724.
661 doi:10.1016/j.physa.2018.07.025
- 662 Maddala, G.S., 1983. Limited-dependent and qualitative variables in economet-
663 rics. Cambridge University Press, Cambridge.
- 664 Manjunath B G <bgmanjunath@gmail.com>, S.W. with contributions from,
665 2015. Tmvtnorm: Truncated multivariate normal and student t distribution.
- 666 Micallef, S., Piscopo, T.V., Casha, R., Borg, D., Vella, C., Zammit, M.-A.,
667 Borg, J., Mallia, D., Farrugia, J., Vella, S.M., Xerri, T., Portelli, A., Fenech,
668 M., Fsadni, C., Mallia Azzopardi, C., 2020. The first wave of COVID-
669 19 in malta; a national cross-sectional study. PLOS ONE 15, e0239389.
670 doi:10.1371/journal.pone.0239389
- 671 Molloy, J., Tchervenkov, C., Hintermann, B., Axhausen, K.W., 2020. Trac-
672 ing the sars-CoV-2 impact: The first month in switzerland. Findings.
673 doi:10.32866/001c.12903
- 674 Moore, E.G., 1970. Some spatial properties of urban contact fields. Geographical
675 Analysis 2, 376–386.
- 676 Moore, E.G., Brown, L.A., 1970. Urban acquaintance fields: An evaluation of a
677 spatial model. Environment and Planning 2, 443–454.

- 678 Müller, K., Wickham, H., 2021. Tibble: Simple data frames.
- 679 Noland, R.B., 1995. PERCEIVED RISK AND MODAL CHOICE - RISK
680 COMPENSATION IN TRANSPORTATION SYSTEM. Accident Analysis
681 and Prevention 27, 503–521. doi:10.1016/0001-4575(94)00087-3
- 682 Noland, R.B., 2021. Mobility and the effective reproduction rate of COVID-19.
683 Journal of Transport & Health 20, 101016. doi:<https://doi.org/10.1016/j.jth.2021.101016>
- 685 Noury, A., François, A., Gergaud, O., Garel, A., 2021. How does COVID-19
686 affect electoral participation? Evidence from the french municipal elections.
687 PLOS ONE 16, e0247026. doi:10.1371/journal.pone.0247026
- 688 Paez, A., 2019. Using spatial filters and exploratory data analysis to en-
689 hance regression models of spatial data. Geographical Analysis 51, 314–338.
690 doi:10.1111/gean.12180
- 691 Paez, A., 2020. Using google community mobility reports to investigate the
692 incidence of COVID-19 in the united states. Findings. doi:<https://doi.org/10.32866/001c.12976>
- 694 Paez, A., Lopez, F.A., Menezes, T., Cavalcanti, R., Pitta, M.G. da R., 2020.
695 A spatio-temporal analysis of the environmental correlates of COVID-19
696 incidence in spain. Geographical Analysis n/a. doi:10.1111/gean.12241
- 697 Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial
698 Vector Data. The R Journal 10, 439–446. doi:10.32614/RJ-2018-009
- 699 Pebesma, E.J., Bivand, R.S., 2005. Classes and methods for spatial data in R.
700 R News 5, 9–13.
- 701 Pebesma, E., Mailund, T., Hiebert, J., 2016. Measurement units in R. R Journal
702 8, 486–494. doi:10.32614/RJ-2016-061
- 703 Pedersen, T.L., 2020. Patchwork: The composer of plots.
- 704 Pedersen, T.L., Crameri, F., 2020. Scico: Colour palettes based on the scientific
705 colour-maps.
- 706 Pequeno, P., Mendel, B., Rosa, C., Bosholn, M., Souza, J.L., Baccaro, F.,
707 Barbosa, R., Magnusson, W., 2020. Air transportation, population density
708 and temperature predict the spread of COVID-19 in brazil. PeerJ 8, e9322.
709 doi:10.7717/peerj.9322
- 710 Phillips, R.O., Fyhri, A., Sagberg, F., 2011. Risk compensation and bicycle
711 helmets. Risk Analysis 31, 1187–1195. doi:10.1111/j.1539-6924.2011.01589.x
- 712 Pinheiro, J., Bates, D., R-core, 2021. Nlme: Linear and nonlinear mixed effects
713 models.

- 714 Praharaj, S., King, D., Pettit, C., Wentz, E., 2020. Using aggregated mobility
715 data to measure the effect of COVID-19 policies on mobility changes in
716 sydney, london, phoenix, and pune. Findings. doi:10.32866/001c.17590
- 717 R Core Team, 2021. R: A language and environment for statistical computing.
718 R Foundation for Statistical Computing, Vienna, Austria.
- 719 Richens, J., Imrie, J., Copas, A., 2000. Condoms and seat belts: The parallels
720 and the lessons. Lancet 355, 400–403. doi:10.1016/s0140-6736(99)09109-6
- 721 Rocklöv, J., Sjödin, H., 2020. High population densities catalyse the spread of
722 COVID-19. Journal of Travel Medicine 27. doi:10.1093/jtm/taaa038
- 723 Roy, S., Ghosh, P., 2020. Factors affecting COVID-19 infected and death
724 rates inform lockdown-related policymaking. PLOS ONE 15, e0241165.
725 doi:10.1371/journal.pone.0241165
- 726 Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O.R., van Aert,
727 R.C.M., van Assen, M.A.L.M., Liu, Y., Althoff, T., Heer, J., Kale, A.,
728 Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson,
729 E., Snellman, K., Amy Sommer, S., Otner, S.M.G., Robinson, D., Madan, N.,
730 Silberzahn, R., Goldstein, P., Tierney, W., Murase, T., Mandl, B., Viganola,
731 D., Strobl, C., Schaumans, C.B.C., Kelchtermans, S., Naseeb, C., Mason
732 Garrison, S., Yarkoni, T., Richard Chan, C.S., Adie, P., Alaburda, P., Albers,
733 C., Alspaugh, S., Alstott, J., Nelson, A.A., Ariño de la Rubia, E., Arzi, A.,
734 Bahník, Š., Baik, J., Winther Balling, L., Bunker, S., AA Baranger, D., Barr,
735 D.J., Barros-Rivera, B., Bauer, M., Blaise, E., Boelen, L., Bohle Carbonell,
736 K., Briers, R.A., Burkhard, O., Canela, M.-A., Castrillo, L., Catlett, T.,
737 Chen, O., Clark, M., Cohn, B., Coppock, A., Cugueró-Escofet, N., Curran,
738 P.G., Cyrus-Lai, W., Dai, D., Valentino Dalla Riva, G., Danielsson, H., F. S.
739 M. Russo, R. de, de Silva, N., Derungs, C., Dondelinger, F., Duarte de Souza,
740 C., Tyson Dube, B., Dubova, M., Mark Dunn, B., Adriaan Edelsbrunner,
741 P., Finley, S., Fox, N., Gnambs, T., Gong, Y., Grand, E., Greenawalt,
742 B., Han, D., Hanel, P.H.P., Hong, A.B., Hood, D., Hsueh, J., Huang, L.,
743 Hui, K.N., Hultman, K.A., Javaid, A., Ji Jiang, L., Jong, J., Kamdar, J.,
744 Kane, D., Kappler, G., Kaszubowski, E., Kavanagh, C.M., Khabsa, M.,
745 Kleinberg, B., Kouros, J., Krause, H., Kryptotos, A.-M., Lavbič, D., Ling
746 Lee, R., Leffel, T., Yang Lim, W., Liverani, S., Loh, B., Lønsmann, D.,
747 Wei Low, J., Lu, A., MacDonald, K., Madan, C.R., Hjorth Madsen, L.,
748 Maimone, C., Mangold, A., Marshall, A., Ester Matskewich, H., Mavon, K.,
749 McLain, K.L., McNamara, A.A., McNeill, M., Mertens, U., Miller, D., Moore,
750 B., Moore, A., Nantz, E., Nasrullah, Z., Nejkovic, V., Nell, C.S., Arthur

- 751 Nelson, A., Nilsonne, G., Nolan, R., O'Brien, C.E., O'Neill, P., O'Shea, K.,
752 Olita, T., Otterbacher, J., Palsetia, D., Pereira, B., Pozdniakov, I., Protzko,
753 J., Reyt, J.-N., Riddle, T., (Akmal) Ridhwan Omar Ali, A., Ropovik, I.,
754 Rosenberg, J.M., Rothen, S., Schulte-Mecklenbeck, M., Sharma, N., Shotwell,
755 G., Skarzynski, M., Stedden, W., Stodden, V., Stoffel, M.A., Stoltzman,
756 S., Subbaiah, S., Tatman, R., Thibodeau, P.H., Tomkins, S., Valdivia, A.,
757 Druijff-van de Woestijne, G.B., Viana, L., Villesèche, F., Duncan Wadsworth,
758 W., Wanders, F., Watts, K., Wells, J.D., Whelpley, C.E., Won, A., Wu, L.,
759 Yip, A., Youngflesh, C., Yu, J.-C., Zandian, A., Zhang, L., Zibman, C., Luis
760 Uhlmann, E., 2021. Same data, different conclusions: Radical dispersion
761 in empirical results when independent analysts operationalize and test the
762 same hypothesis. *Organizational Behavior and Human Decision Processes*
763 165, 228–249. doi:<https://doi.org/10.1016/j.obhdp.2021.02.003>
- 764 Sharifi, A., Khavarian-Garmsir, A.R., 2020. The COVID-19 pandemic: Impacts
765 on cities and major lessons for urban planning, design, and management.
766 *Science of The Total Environment* 749, 142391. doi:<https://doi.org/10.1016/j.scitotenv.2020.142391>
- 767 Skórka, P., Grzywacz, B., Moroń, D., Lenda, M., 2020. The macroecology of
768 the COVID-19 pandemic in the anthropocene. *PLOS ONE* 15, e0236856.
769 doi:[10.1371/journal.pone.0236856](https://doi.org/10.1371/journal.pone.0236856)
- 770 Souris, M., Gonzalez, J.-P., 2020. COVID-19: Spatial analysis of hospital case-
771 fatality rate in france. *PLOS ONE* 15, e0243606. doi:[10.1371/journal.pone.0243606](https://doi.org/10.1371/journal.pone.0243606)
- 772 Stephens, K.E., Chernyavskiy, P., Bruns, D.R., 2021. Impact of altitude on
773 COVID-19 infection and death in the united states: A modeling and obser-
774 vational study. *PLOS ONE* 16, e0245055. doi:[10.1371/journal.pone.0245055](https://doi.org/10.1371/journal.pone.0245055)
- 775 Stodden, V., Seiler, J., Ma, Z.K., 2018. An empirical analysis of journal pol-
776 icy effectiveness for computational reproducibility. *Proceedings of the Na-
777 tional Academy of Sciences of the United States of America* 115, 2584–2589.
778 doi:[10.1073/pnas.1708290115](https://doi.org/10.1073/pnas.1708290115)
- 779 Sumner, J., Haynes, L., Nathan, S., Hudson-Vitale, C., McIntosh, L.D., 2020.
780 Reproducibility and reporting practices in COVID-19 preprint manuscripts.
781 medRxiv 2020.03.24.20042796. doi:[10.1101/2020.03.24.20042796](https://doi.org/10.1101/2020.03.24.20042796)
- 782 Sun, Z., Zhang, H., Yang, Y., Wan, H., Wang, Y., 2020. Impacts of geographic
783 factors and population density on the COVID-19 spreading under the lock-
784 down policies of china. *Science of The Total Environment* 746, 141347.
785 doi:[10.1016/j.scitotenv.2020.141347](https://doi.org/10.1016/j.scitotenv.2020.141347)

- 787 Sy, K.T.L., White, L.F., Nichols, B.E., 2021. Population density and basic
788 reproductive number of COVID-19 across united states counties. PLOS ONE
789 16, e0249271. doi:10.1371/journal.pone.0249271
- 790 Toomet, O., Henningsen, A., 2008. Sample selection models in R: Package
791 sampleSelection. Journal of Statistical Software 27.
- 792 Viglione, G., 2020. 'Avalanche'of spider-paper retractions shakes behavioural-
793 ecology community. Nature 578, 199–201.
- 794 Vlasschaert, C., Topf, J.M., Hiremath, S., 2020. Proliferation of papers and
795 preprints during the coronavirus disease 2019 pandemic: Progress or prob-
796 lems with peer review? Advances in Chronic Kidney Disease 27, 418–426.
797 doi:10.1053/j.ackd.2020.08.003
- 798 Walker, K., Herman, M., 2021. Tidycensus: Load US census boundary and
799 attribute data as tidyverse and sf-ready data frames.
- 800 Wang, F., Tan, Z., Yu, Z., Yao, S., Guo, C., 2021. Transmission and control pres-
801 sure analysis of the COVID-19 epidemic situation using multisource spatio-
802 temporal big data. PLOS ONE 16, e0249145. doi:10.1371/journal.pone.0249145
- 803 White, E.R., Hébert-Dufresne, L., 2020. State-level variation of initial COVID-19
804 dynamics in the united states. PLOS ONE 15, e0240648. doi:10.1371/journal.pone.0240648
- 805 Wickham, H., 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag
806 New York.
- 807 Wickham, H., 2021a. Forcats: Tools for working with categorical variables
808 (factors).
- 809 Wickham, H., 2021b. Tidyr: Tidy messy data.
- 810 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François,
811 R., Golemünd, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen,
812 T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel,
813 D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H.,
814 2019. Welcome to the tidyverse. Journal of Open Source Software 4, 1686.
815 doi:10.21105/joss.01686
- 816 Wickham, H., François, R., Henry, L., Müller, K., 2021. Dplyr: A grammar of
817 data manipulation.
- 818 Wickham, H., Hester, J., 2021. Readr: Read rectangular text data.
- 819 Wilhelm, S., de Matos, M.G., 2015. Spatialprobit: Spatial probit models.
- 820 Wong, D.W.S., Li, Y., 2020. Spreading of COVID-19: Density matters. PLOS
821 ONE 15, e0242398. doi:10.1371/journal.pone.0242398
- 822 Xie, Y., 2014. Knitr: A comprehensive tool for reproducible research in R,
823 in: Stodden, V., Leisch, F., Peng, R.D. (Eds.), Implementing Reproducible

- 824 Computational Research. Chapman; Hall/CRC.
- 825 Xie, Y., 2015. Dynamic documents with R and knitr, 2nd ed. Chapman;
826 Hall/CRC, Boca Raton, Florida.
- 827 Xie, Y., Allaire, J.J., Grolemund, G., 2018. R markdown: The definitive guide.
828 Chapman; Hall/CRC, Boca Raton, Florida.
- 829 Xie, Y., Dervieux, C., Riederer, E., 2020. R markdown cookbook. Chapman;
830 Hall/CRC, Boca Raton, Florida.
- 831 Zeileis, A., 2004. Econometric computing with HC and HAC covariance matrix
832 estimators. Journal of Statistical Software 11, 1–17. doi:10.18637/jss.v011.i10
- 833 Zeileis, A., 2006. Object-oriented computation of sandwich estimators. Journal
834 of Statistical Software 16, 1–16. doi:10.18637/jss.v016.i09
- 835 Zeileis, A., Köll, S., Graham, N., 2020. Various versatile variances: An object-
836 oriented implementation of clustered covariances in R. Journal of Statistical
837 Software 95, 1–36. doi:10.18637/jss.v095.i01
- 838 Zhu, H., 2021. kableExtra: Construct complex table with kable and pipe syntax.