*Reproduction of*

HE&G ReScience

[RP] Report

**Beyond the 405 and the 5: Geographic Variations and**

**Factors Associated With Severe Acute Respiratory**

**Syndrome Coronavirus 2 (SARS-CoV-2) Positivity Rates in**

**Los Angeles County**

*by* Vijayan T., Shin M., Adamson P.C., Harris C., Seeman T., Norris K.C., Goodman-Meza D.

**Created**

18 February 2021

*Reproduction Authors:*

**Created**

18 February 2021

Peter Kedron1,2 / Joseph Holler3 / Sarah Bardin†,1,2 / Summer Cliff / Kimberly Fuller / Joshua Gilman / Bryant Grady / Megan Seeley / Addison Van Zanbergen / Wenxin Wang / Xin Wang

**Revised**

20 May 2021

**Revised**

05 July 2021

1Arizona State University, School of Geographical Sciences and Urban Planning, Tempe, AZ, USA 2Arizona State University, Spatial Analysis Research Center (SPARC), Tempe, AZ, USA 3Middlebury College, Department of Geography, Middlebury, VT, USA

†Graduate Student Author

*Replication Materials Available at:*

**Pre-registered Plan** – <https://github.com/HEGSRR/RP-Vijayan-2020/tree/main/docs/report>

**Data** –<https://github.com/HEGSRR/RP-Vijayan-2020/tree/main/data/private>

**Code** – <https://github.com/HEGSRR/RP-Vijayan-2020/tree/main/procedure/code>

|  |
| --- |
| **Research Hypotheses to Reproduce** |
| **H1:** There is a difference in mean values of key socioeconomic and demographic variables by positivity rate groupings of low, medium, and high areas.  **Original test:** One-way analysis of variance (ANOVA) indicated that all variables, except the percentage black, exhibited statistically significant differences among the three subgroups (Table 1).  **H2:** (a) COVID-19 age-adjusted testing rate, (b) age-adjusted diagnosis rate, and (c) crude positivity rate were non-randomly distributed throughout LA County.  **Original test:** Local indicators of spatial association (LISA) identified elevated values of each variable around the center of LA, and depressed values around the edges of the county (Fig. 1)  **H3:** Socio-structural characteristics of LAC have non-zero association with crude positivity rate.  **Original test:** The authors used a regression model with a spatially lagged response was to identify significant positive associations between crude positivity and (i) proportion of population over 65, (ii) proportion Latino, proportion living in poverty, and (iii) housing density. |

|  |
| --- |
| **Key Findings** |
| * We were able to reproduce the original analyses after contacting the authors to obtain their data file. Original effect estimates for the regression coefficients fell within the 95% confidence intervals of our reproduction. * The lack of details concerning how the hexagons that were used as the unit of analysis by the authors were created, or how data from census tracts were aggregated to those hexagonal units inhibited our ability to independently reproduce the original results. Moreover, these same issues made it difficult to assess what the predictor and response variables were in fact measuring as this spatial aggregation potentially misaligned several variables with the conventional definitions. * The authors did not provide equations or formulations related to their implementation of the SLM analyses, it is therefore difficult to assess how the models should be properly interpreted. The language used in the original discussion by the authors imprecisely describes the coefficients reported and ignores the fact that these are based on standardized variables and that the model intercept was omitted from the analysis. These issues raise questions about how the estimated effects should be interpreted. |

## Original Study Information

**Description:**

Vijayan et al. (2020) examined whether spatial patterns existed in SARS-CoV-2 age-adjusted testing rates, age-adjusted diagnosis rates, and crude positivity rates in Los Angeles County (LAC), and used a spatial regression model to explore associations between COVID-19 crude positivity rates and a series of predictor variables. The original analyses are retrospective and use observational data collected from federal and private sources. Although not publicly available, we were able to obtain the original study data after contacting the authors. However, the analysis code was not made available, nor was information about the computational environment used.

**Analytical Plan:**

*Sampling Plan and Data Description:* Vijayan et al. collected data from online, publicly available datasets.

Predictor variables were obtained from the 2018 5-year American Community Survey at the census tract level. Syndrome Coronavirus 2 (SARS-CoV-2) or COVID-19 testing and diagnosis data were obtained from the LAC Department of Public Health COVID-19 surveillance dashboard (http://dashboard.publichealth. lacounty.gov/covid19\_surveillance\_dashboard/) through 30 June 2022.

Prior to analysis, the data were reapportioned to a hexagonal grid created by the authors. However, the authors provided limited details about how the hexagonal grid was generated. The authors did not provide any details about the algorithm or parameters used to create the grid. Each hexagon in the grid encompassed an area of 10 square kilometers. Once created, the grid was overlaid onto the centroids of city, community, and census tract boundaries within LAC, and all data were summarized and joined to the hexagon layer by location. Hexagons that either contained missing COVID-19 data, had a population of less than 1,000 people, or did not have contiguous neighbors were excluded from the analysis. The final analysis sample contained 184 hexagons.

*Variables:* Predictor variables used in the spatial regression analysis included measures of: race/ethnicity, poverty, insurance status, educational status, population density and household density. The percentage of those under the age of 18 and above the age of 65 were also assessed in relation to COVID-19 positivity rates.

* *Race/ethnicity* was expressed as a percentage of the population and segmented into four categories - non-Hispanic white, non-Hispanic Black, Asian, and Hispanic or Latino/a groups.
* *Poverty* was measured as the percentage of households living below the federal family poverty threshold.
* *Educational status* was measured as the percentage who completed a bachelor’s degree or higher.
* *Population density* was calculated by dividing the total population within each hexagon by the area of the hexagon. Although referred to as household density by Vijayan et al., this variable was calculated by dividing the total population within the hexagon by the number of reported households in the hexagon, and therefore captures the average household size within each hexagon as opposed to density.

Because the Census data were originally obtained at the tract level, a spatial transformation was needed to convert the tract data to the hexagon level. According to the original paper, this transformation was achieved by associating tracts to the hexagon in which the centroid was located, however, the authors do not provide details regarding whether additional manipulation of the raw data occurred (e.g. averaging the values of characteristics across multiple tracts located within a single hexagon).

The response variables examined by the original authors were 1) COVID-19 age-adjusted testing rates, 2) age-adjusted diagnosis rates, and 3) crude positivity rates. The crude positivity rate was calculated by dividing the count of positive COVID-19 cases by the number of tests conducted and multiplying by 100.

*Analytical Specification:* The authors did not specify the computational environment used during their analysis. No coordinate system or projections information were specified. Edge effects and multiple testing were also not addressed.

Vijayan et al. perform two descriptive statistical analysis before moving to a subsequent spatial regression analysis. The authors claim to perform a set of correlational analyses as an intermediate step to assess differences in the means of each of the predictor variables by low, medium, and high crude positivity rate subgroups. However, upon inspection of the paper it appears that the authors in fact performed a one-way ANOVA analysis without post-hoc testing. The authors used a LISA statistic to explore spatial patterning of the outcome variables. A conditional permutation test was used to determine significance with an alpha of 0.05, however the number of simulations performed was not specified.

The results from the descriptive analysis were used to inform several spatial regression models, which included a spatially lagged dependent variable. The spatial lag was set to be equivalent to the average values of SARS-CoV-2 outcomes in adjacent hexagons. The authors reported direct, indirect, and total effects estimates.

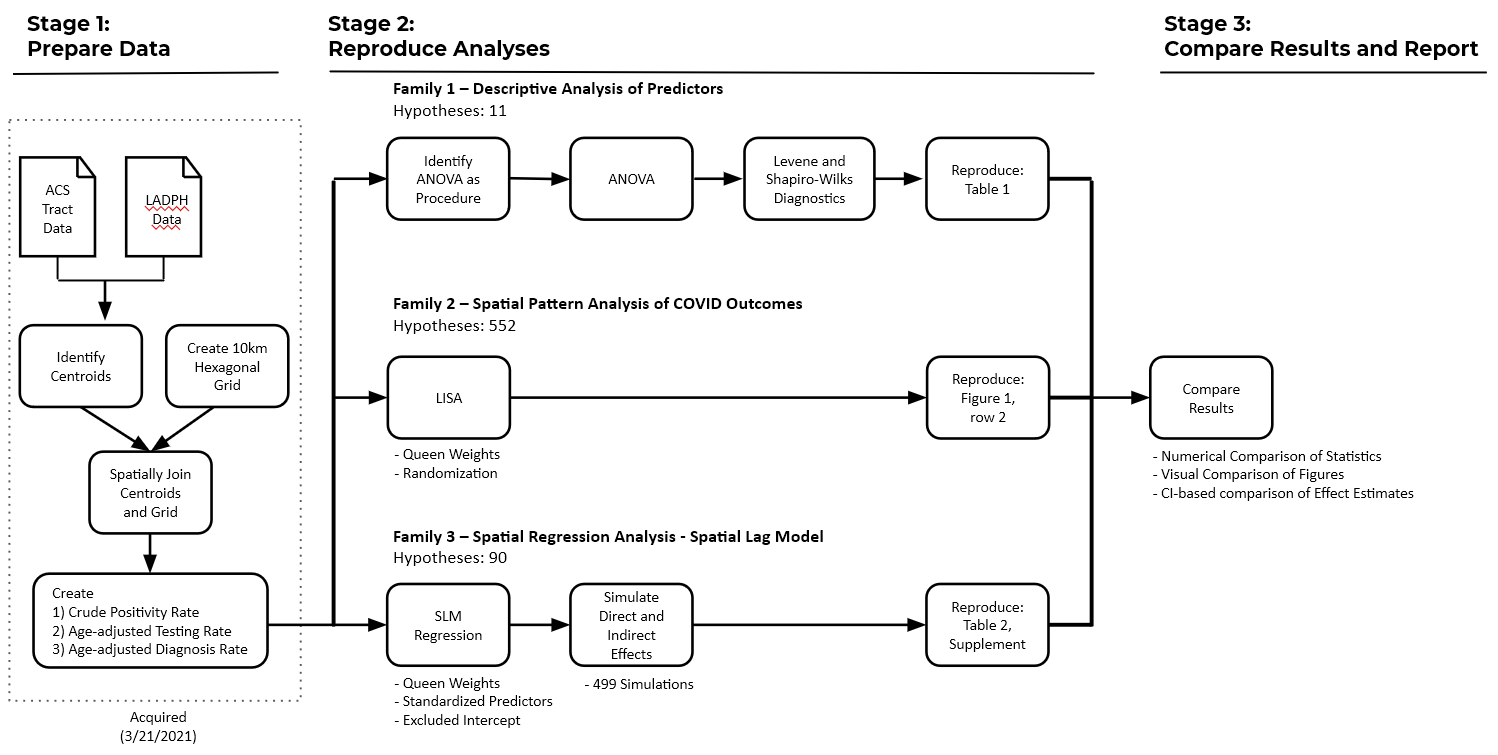
*Inference Criteria, Results, and Robustness:* The original authors conducted all hypothesis tests using the *p*=0.05 significance threshold. Local spatial statistics were also mapped. Direct, indirect, and total effects were presented in table form (Table 2) and similarly assessed using the *p*=0.05 threshold.

## Reproduction Procedure

**Protocol:**

We followed the data preparation and analytical procedures of the original study, making as few modifications as possible. We obtained data on both the response and predictor variables from the same sources as the original authors. Initially, we attempted to reconstruct the hexagonal grid used in the original study following the methods described. However, we were unable to create a matching grid with the details provided in the published paper, and by extension could not spatially join the predictor and response data to the grid for analysis. Because we were unable to independently reproduce the spatial data needed for further analysis, we contacted the original authors who shared their spatial data files. We used these files for our analyses. However, we independently gathered all details of the authors’ analytical workflow from the original article and accompanying supplemental material.

Following the original authors, we grouped our analyses into three families (Fig. 1). First,we performed one-way analysis of variance (ANOVA) tests to assess differences in the mean values of key predictors across three subgroups. These subgroups included areas with low crude positivity rates (rates < 0.5%), medium crude positivity rates (rates between 5% and 10%), and high crude positivity rates (rates of 10% or higher). While Vijayan et al. only referred to these tests as correlational analyses, the presentation of the findings in Table 1 suggested that an ANOVA was likely performed. Second, we used univariate LISA to test the null hypothesis that the geographic distribution of SARS-CoV-2 outcomes are randomly distributed across LA County. Finally, we used a spatial lag model (SLM) to explore associations between each response measure and a series of socioeconomic and demographic predictors to test the null hypothesis that there is no association between the predictor variable and the outcome.



**Fig. 1** Reproduction workflow

*Notes:* Hashed lines indicate steps that were completed by original author and supplied upon request for purposes of the reproduction.

*Planned Differences from the Original Study:* Vijiyan et al. (2020) did not specify the computational environment or software used during their original study. In the absence of this information, we decided to implement these analyses in R v4.0.5 and Rstudio v1.4.1106 using the following packages: tidyverse, sp, rgdal, spatialreg, spdep. Although Vijayan et al. described using a permutation approach for identifying statistically significant clusters in their LISA analyses, they did not detail the number of simulations conducted, nor did they indicate how statistical significance was calculated for the SLM models. We elected to implement a permutation approach which used 499 simulations in order to calculate our p-values. Because there is inherent randomness in the permutation approach, when a seed is not set, and we cannot be certain of the number of simulations performed, we did not expect to be able to fully reproduce the exact p-values reported in the paper, however we expected that the direction and magnitude of the results would be consistent between the original analysis and the reproductions.

*Assessment Criteria:* As noted earlier, we did not anticipate being able to achieve bitwise reproduction of any of the LISA or spatial regression results with those from the original study. As a result, we compared the direction, magnitude, and significance of our results with those of the original authors.

## Reproduction Results

**Reproduction Results for the Descriptive Statistical Analyses of Predictors (H1):**

The first set of hypotheses examined by Vijayan et al. compared the distribution of a set of predictor variables across three subgroups of low, medium, and high positivity rates. Vijayan et al. did not specify the type of correlational analysis performed, so we performed one-way ANOVA tests. As shown in Table 1 below, we achieved bitwise reproduction of the original analysis results. That is, all means, standard deviations, and reported p-values from the original study were identically reproduced.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Original** | | | | **Reproduction** | | | |
|  | **< 5%** | **5-9.9%** | **≥ 10%** | ***p*** | **< 5%** | **5-9.9%** | **≥ 10%** | ***p*** |
| Sample size | 44 | 74 | 66 |  | 44 | 74 | 66 |  |
| % Age below 18 years | 18.8 (4.80) | 19.9 (3.73) | 24.5 (4.08) | <0.001 | 18.8 (4.80) | 19.9 (3.73) | 24.5 (4.08) | <0.001 |
| % Age above 65 years | 17.1 (5.31) | 15.5 (4.40) | 11.5 (3.70) | <0.001 | 17.1 (5.31) | 15.5 (4.40) | 11.5 (3.7) | <0.001 |
| % Latino/a | 68 (14.33) | 48.8 (19.2) | 46.1 (14.1) | <0.001 | 68 (14.33) | 48.8 (19.23) | 46.1 (14.09) | <0.001 |
| % White | 4.94 (7.64) | 8.55 (12.5) | 8.4 (13.24) | 0.223 | 4.9 (7.64) | 8.6 (12.48) | 8.4 (13.24) | 0.223 |
| % Black | 15.3 (9.91) | 20.9 (17.9) | 10.8 (12.01) | <0.001 | 15.3 (9.91) | 20.9 (17.90) | 10.8 (12.01) | <0.001 |
| % Asian | 17.3 (13.54) | 37.8 (18.7) | 69.3 (18.50) | <0.001 | 17.3 (13.54) | 37.8 (18.73) | 69.3 (18.50) | <0.001 |
| % Poverty | 4.70 (2.75) | 9.93 (5.08) | 16.4 (7.74) | <0.001 | 4.7 (2.75) | 9.9 (5.08) | 16.4 (7.74) | <0.001 |
| % Uninsured | 5.36 (2.71) | 8.84 (3.73) | 13.6 (4.85) | <0.001 | 5.4 (2.71) | 8.8 (3.73) | 13.6 (4.85) | <0.001 |
| % Bachelor’s degree or higher | 58.3 (14.69) | 35.4 (14.6) | 17.4 (11.0) | <0.001 | 58.3 (14.69) | 35.4 (14.56) | 17.4 (11.01) | <0.001 |
| Population density per square kilometer | 2248 (1547) | 3115 (1735) | 4381 (2363) | <0.001 | 2248.2 (1548) | 3115.7 (1735) | 4381.5 (2363) | <0.001 |
| Household density per hexagon | 2.36 (0.48) | 2.84 (0.49) | 3.54 (0.56) | <0.001 | 2.4 (0.48) | 2.8 (0.49) | 3.5 (0.56) | <0.001 |

**Table. 1** Distribution of independent variables by positivity rate subgroups in original analysis and reproduction analysis

*Note:* All data are presented as percentages (SD), except for population density which is the total population per square kilometer. P-values are based on one-way ANOVA.

*Abbreviation:* ANOVA = analysis of variance; SD = standard deviation

Although we were able to fully reproduce Vijayan et al.’s subgroup analyses, it is worth noting that diagnostic checks for the ANOVA tests indicated that most variables did not meet the normality and homoscedasticity assumptions needed for these tests. Only the distributions for the percentage of adults under the age of 18 and household density met these assumptions, while all other variables failed to meet at least one of the assumptions of ANOVA, based on post-estimation Levene and Shapiro-Wilks diagnostic tests. These findings suggest that a non-parametric test, such as a Kruskal-Wallis test, would have been more appropriate for assessing subgroup differences for the majority of predictor variables. We executed these tests for all variables. These tests produced *p*-values less than 0.001 for all variables other than percent white. For that variable the Kruskal-Wallis value was 0.63.

**Reproduction Results for Spatial Pattern Analysis of SARS-CoV-2 outcomes (H2):**

Consistent with the original analysis, reproductions of the LISA analyses indicate that the COVID-19 testing rates, diagnosis rates, and positivity rates are not spatially random across LA county (Fig 2).

Map

Description automatically generated

**Map

Description automatically generated**

**Fig 2.** LISA maps of SARS-CoV-2 testing, diagnosis and positivity rates across LAC in original analysis and reproduction analysis

Because a permutation-based approach was used for determining statistical significance and the Vijayan et al. did not provide requisite information for fully reproducing their analysis, it is not surprising that we were unable to achieve bitwise reproduction of these results. However, our reproductions for the diagnosis rate clusters and positivity rate clusters show clear similarities with the original analysis. In both cases, we find a large high-high cluster of hexagons in central LAC with low-low clusters along the western coast of LAC. Consistent with Vijayan et al.’s results, we also identified low-low diagnosis rate clusters in portions of eastern LAC. While these two reproductions generally align with the original analysis, the same can not be said of the testing rate clusters. In the original analysis, Vijayan et al. identified high-high clusters in both central and northwestern positions of LAC and low-low clusters in the southern and eastern portions of the county. In contrast, our reproductions identified a small number of high-high clusters in central LAC and no low-low clusters. Importantly, Vijayan et al. only reported high-high and low-low clusters for each of these outcomes, while our reproductions indicated that both low-high and high-low clusters were present. It is not clear whether Vijayan et al. chose to omit these findings and instead focus on high-high and low-low clusters exclusively, or if their analyses failed to identify any of these hybrid clusters.

**Reproduction Results for the Spatial Lag Model Regression Testing for Predictors of SARS-CoV-2 and the Presence of a Spatial Diffusion Process in the spread of SARS-CoV-2 (H3):**

The final set of hypotheses examined by Vijayan et al. assessed the relationship between a set of predictors and three response variables: the crude positivity rate, the age-adjusted diagnosis rate, and the age-adjusted testing rate. Consistent with Vijayan et al.’s original findings, results from the spatial lag model reproduction of the positivity rate model indicate a positive rho value, which was statistically significant at a p-value of 0.05, indicating the presence of a spatial diffusion process. The original *rho* value obtained by Vijayan et al. was 0.212, which was more than 4 times larger than the rho obtained during the reproduction (*rho* = 0.046).

The coefficient estimates of the total effects of all predictors reported in the original analysis fall within the 95% confidence interval of the reproduction analysis, further indicating strong agreement between the reproduction and original analysis. We next used 499 simulations to estimate the direct and indirect effects of each predictor. To compare our estimates for each effect from each predictor to those of Vijayan et al., we then examined whether each effect reported by Vijayan et al. fell within our simulated 95% confidence interval. All of the original findings fall within the 2.5% and 97.5% thresholds of the effect estimate distributions for each coefficient from our simulations. This result indicates that we were able to reproduce the results.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Original** | | | | | | **Reproduction** | | | | | |
|  | **Estimate** | **SE** | **Direct** | **Indirect** | **Total** | **P** | **Estimate** | **SE** | **Direct** | **Indirect** | **Total** | **P** |
| % Age below 18 years | −0.042 | 0.075 | −0.042 | −0.011 | −0.053 | .575 | -0.04 | 0.077 | -0.036 | -0.007 | -0.044 | 0.638 |
| % Age above 65 years | 0.161 | 0.068 | 0.164 | 0.041 | 0.204 | .018 | 0.178 | 0.071 | 0.179 | 0.036 | 0.215 | 0..012 |
| % Latino/a | 0.283 | 0.162 | 0.288 | 0.071 | 0.359 | .081 | 0.310 | 0.163 | 0.313 | 0.063 | 0.375 | 0.058 |
| % White | 0.094 | 0.131 | 0.095 | 0.024 | 0.119 | .474 | 0.080 | 0.134 | 0.081 | 0.016 | 0.097 | 0.551 |
| % Black | 0.033 | 0.104 | 0.033 | 0.008 | 0.042 | .753 | 0.030 | 0.107 | 0.030 | 0.006 | 0.036 | 0.781 |
| % Asian | −0.019 | 0.127 | −0.019 | −0.005 | −0.024 | .882 | -0.012 | 0.130 | -0.012 | -0.002 | -0.014 | 0.927 |
| % Poverty | 0.293 | 0.100 | 0.298 | 0.074 | 0.371 | .004 | 0.198 | 0.108 | 0.199 | 0.040 | 0.239 | 0.066 |
| % Uninsured | −0.015 | 0.117 | −0.015 | −0.004 | −0.019 | .898 | 0.067 | 0.118 | 0.068 | 0.014 | 0.081 | 0.570 |
| % Bachelor’s degree or higher | −0.006 | 0.131 | −0.006 | −0.001 | −0.008 | .964 | -0.009 | 0.135 | -0.009 | -0.002 | -0.011 | 0.946 |
| Population density per square kilometer | 0.101 | 0.069 | 0.103 | 0.026 | 0.129 | .145 | 0.151 | 0.072 | 0.152 | 0.030 | 0.182 | 0.036 |
| Household density per hexagon | 0.329 | 0.110 | 0.334 | 0.083 | 0.417 | .003 | 0.298 | 0.114 | 0.300 | 0.060 | 0.360 | 0.009 |
| rho, Spatial lag | 0.212 | 0.070 |  |  |  | .00 | 0.046 | 0.018 |  |  |  | 0.012 |

**Table 2.** Reproduction of Spatial Lag Model of Socio-structural Correlates of SARS-CoV-2 Crude Positivity Rates in LAC

Although not presented in the main paper, Vijayan et al. provided supplementary results for SLM analyses using what the authors label age-adjusted diagnosis rates and age-adjusted testing rates, respectively. All coefficient estimates except for the percent uninsured variable for the age-adjusted diagnosis rate model reported in Supplementary Table 1 of the original paper fall within the 95% confidence interval of the reproduction SLM analysis (Table 3). Similarly, all coefficient estimates from Supplementary Table 2 fall within the 95% confidence interval of our reproduction SLM analysis (Table 4). See the discussion section for further discussion of concerns related to the construction of these age-adjusted rates.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Original** | | | | | | **Reproduction** | | | | | |
|  | **Estimate** | **SE** | **Direct** | **Indirect** | **Total** | ***p*** | **Estimate** | **SE** | **Direct** | **Indirect** | **Total** | ***p*** |
| % Latino | 0.301 | 0.139 | 0.307 | 0.085 | 0.392 | 0.030 | -0.183 | 0.255 | -0.183 | 0.020 | -0.163 | 0.475 |
| % White | 0.195 | 0.113 | 0.199 | 0.055 | 0.254 | 0.083 | 0.042 | 0.211 | 0.042 | 0.005 | -0.038 | 0..841 |
| % Black | 0.069 | 0.090 | 0.070 | 0.019 | 0.090 | 0.444 | -0.092 | 0.167 | -0.092 | 0.010 | -0.082 | 0.582 |
| % Asian | 0.007 | 0.104 | 0.008 | 0.002 | 0.010 | 0.943 | -0.060 | 0.197 | -0.060 | 0.006 | -0.054 | 0.761 |
| % Poverty | 0.451 | 0.083 | 0.460 | 0.127 | 0.587 | <0.001 | -0.028 | 0.153 | -0.028 | 0.003 | -0.025 | 0.853 |
| % Uninsured | -0.100 | 0.099 | -0.102 | -0.028 | -0.130 | 0.314 | 0.450 | 0.184 | 0.450 | -0.048 | 0.402 | 0.015 |
| % Bachelor’s degree or higher | 0.083 | 0.110 | 0.084 | 0.023 | 0.107 | 0.455 | 0.095 | 0.208 | 0.095 | -0.010 | 0.085 | 0.650 |
| Population density per square kilometer | 0.124 | 0.058 | 0.126 | 0.035 | 0.161 | 0.032 | 0.205 | 0.110 | 0.205 | -0.022 | 0.183 | 0.063 |
| Household density per hexagon | 0.201 | 0.075 | 0.205 | 0.056 | 0.261 | 0.008 | 0.136 | 0.140 | 0.136 | -0.015 | 0.121 | 0.336 |
| rho, Spatial lag | 0.231 | 0.063 |  |  |  | 0.001 | -0.033 | 0.026 |  |  |  | 0.203 |

**Table 3.** Reproduction of Spatial Lag Model of Socio-structural Correlates of SARS-CoV-2 Age-Adjusted Diagnosis Rates in LAC

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Original** | | | | | | **Reproduction** | | | | | |
|  | **Estimate** | **SE** | **Direct** | **Indirect** | **Total** | ***p*** | **Estimate** | **SE** | **Direct** | **Indirect** | **Total** | ***p*** |
| % Latino | 0.182 | 0.217 | 0.193 | 0.102 | 0.295 | 0.402 | -0.080 | 0.278 | -0.080 | -0.005 | -0.084 | 0.775 |
| % White | 0.109 | 0.179 | 0.116 | 0.061 | 0.177 | 0.541 | 0.056 | 0.229 | 0.056 | 0.003 | 0.059 | 0..808 |
| % Black | 0.051 | 0.142 | 0.054 | 0.029 | 0.083 | 0.718 | -0.057 | 0.182 | -0.057 | -0.003 | -0.061 | 0.752 |
| % Asian | -0.157 | 0.168 | -0.166 | -0.088 | -0.253 | 0.350 | -0.167 | 0.215 | -0.167 | -0.010 | -0.177 | 0.437 |
| % Poverty | 0.181 | 0.132 | 0.192 | 0.101 | 0.293 | 0.168 | 0.165 | 0.166 | 0.165 | 0.010 | 0.175 | 0.319 |
| % Uninsured | -0.067 | 0.160 | -0.071 | -0.038 | -0.109 | 0.674 | -0.107 | 0.200 | -0.107 | -0.006 | -0.113 | 0.595 |
| % Bachelor’s degree or higher | 0.271 | 0.179 | 0.287 | 0.152 | 0.438 | 0.131 | -0.056 | 0.227 | -0.056 | -0.003 | -0.059 | 0.806 |
| Population density per square kilometer | 0.137 | 0.093 | 0.145 | 0.076 | 0.221 | 0.140 | 0.050 | 0.119 | 0.050 | -0.003 | 0.053 | 0.673 |
| Household density per hexagon | -0.158 | 0.120 | -0.167 | -0.088 | -0.255 | 0.188 | 0.003 | 0.153 | 0.003 | 0.000 | 0.003 | 0.987 |
| rho, Spatial lag | 0.382 | 0.070 |  |  |  | <0.001 | 0.016 | 0.026 |  |  |  | 0.592 |

**Table 4.** Reproduction of Spatial Lag Model of Socio-structural Correlates of SARS-CoV-2 Age-Adjusted Testing Rates in LAC

## 

## **Unplanned Deviations from the Protocol**

Although we were able to obtain the original study data from the authors, this dataset was not restricted to the final sample size of 184 hexagons. As a result, we needed to implement the exclusion criteria described in the paper to restrict the data to the analytic sample. After removing hexagons with populations under 1,000, hexagons with missing response or predictor variable values, and those without contiguous neighbors, we were left with 184 hexagons, which matched the sample size reported in the original paper. Furthermore, our reproduction of the summary statistics in Table 1 indicated that we achieved the same set of hexagons in our analysis data set as the original authors.

In order to reproduce Table 1, we needed to test for correlation between positivity rates and various predictor variables. Nowhere in the text did the authors specify the type of correlational analysis performed. Based on the description of the results, we initially assumed that either a pearson or spearman correlation was performed. However, the presentation of the results in Table 1 suggested that the authors used ANOVA to explore differences in the mean values across the three subgroups of positivity rates, as opposed to the correlation between positivity rate and each variable. Given that the authors did not specify the hypotheses they were testing for these analyses, we performed all three analyses and determined that ANOVA was likely used as the p-value obtained for the percent Black variable matched identically between the reproduction and the original analysis.

Finally, Vijayan et al. do not report the intercept of their SLM analyses, which may suggest that they dropped the intercept from their regression models. Although this was not explicitly stated in their methods section, we opted to exclude an intercept in our SLM reproductions for consistency with how Vijayan et al. report their results.

## **Discussion**

Overall our reproductions support several of the high-level conclusions presented in the original paper. Specifically, our results suggest that geographic clusters of high positivity, diagnosis, and testing rates can be found in the central part of LA County, consistent with the findings from the original study. Similarly, results from the spatial lag model indicate that the crude positivity rate is associated with the proportions of Latino/a individuals in an area, poverty rate, and household density. Additionally the spatial lag term rho was found to be statistically significant in both the original analysis and in the reproduction.

While our results generally support the main findings presented by Vijayan et al., there are several steps that the original authors could have taken to improve the reproducibility of their work. Below we outline a handful of procedural and statistical critiques that could have strengthened the overall quality of the analysis conducted.

**Procedural Concerns**

A primary procedural concern with the analysis is rooted in the construction of the three response variables: the age-adjusted diagnosis rate, the age-adjusted testing rate, and the crude positivity rate. Although the age-adjusted rates were obtained directly from the LA County Department of Health, these rates were provided at multiple geographic units, including the tract, city, and county levels. Rather than using the geographic units in which the COVID-19 data were originally reported as the unit of analysis, Vijayan et al. instead translated these data into a standardized hexagonal grid. This translation required the authors to make assumptions as to how the original data corresponded to the new unit of analysis. While the authors indicate that they intersected the centroids of the geographic units in which the COVID-19 data were reported with the hexagonal grid, they do not mention whether the centroids from multiple tracts, cities, or counties intersected the same hexagonal unit within the grid or the degree of overlap between the hexagonal grid and original geographic boundaries. Given that the age-adjustment was not based on the population within the hexagonal units developed by Vijayan et al., these response variables are no longer accurately age-adjusted, and it is therefore misleading to refer to these response variables as such.

In addition to these concerns, the authors failed to explicitly describe the hypotheses tested in their “correlational analysis”. As a result, we cannot be confident as to whether the procedures we implemented and the hypotheses we tested in our reproductions are consistent with those presented by Vijayan et al. While it appears that ANOVA was used for these analyses, it is also possible that a non-parametric test, such as Kruskal-Wallis, was used instead, and given the distributions of several predictor variables, our reproductions suggest that a combination of ANOVA and Kruskal-Wallis should have been used to analyze these subgroup differences.

Finally, because we do not know the specific software used to conduct the original analyses, it is possible that the R packages used to implement the analyses rely on different default settings or underlying modeling mechanisms, which may affect the estimates produced and hamper our ability to fully reproduce the analysis. By providing this additional level of detail regarding the implementation of their methods, we would be more confident that departures in our results were not partly due to differences in the software used.

**Statistical and Inferential Concerns**

Through the process of reproducing Vijayan et al.’s analyses, we noticed several analytic decisions that may impact the validity and reliability of the results presented. Specifically, Vijayan et al. spend little time justifying the analytic unit of analysis chosen and fail to note the additional assumptions needed in order to translate the raw data provided at multiple spatial scales into a single scale. The authors do not explain why they elected to aggregate all data to the 10km hexagonal-level, however, because of well known issues related to the modifiable areal unit problem, the arbitrariness of this unit of analysis directly affects the relationships being modeled. Had an alternative shape or size of these aggregation units been selected, the results from the analyses would likely be different. As a result, the authors should provide greater justification for why this particular unit of analysis is appropriate for the processes being modeled.

Relatedly, in order to standardize all data to the hexagonal level, the authors ignore the degree of geographic overlap between the hexagonal grid and the source data. Instead, they associate the raw data to hexagons based solely on the location of the centroid of the input data. Because some raw data is provided at a coarser geographic scale (e.g. community-level) than other data, this simplified approach may assign COVID-19 data to a hexagon that is not representative of the broader population included in the testing data.

Although we did not reproduce the quintile maps included by the authors as the first row of Figure 1, we wish to emphasize that these maps contain more than the 184 hexagons used in cluster analyses presented in row two of the same figure. This difference is difficult to identify without deeper inspection, because the original authors use white to denote both “not statistically significant” and “not in sample” in the second row. Therefore, it is not immediately obvious to readers that the two rows have different hexagon counts that will lead to misleading visual comparisons. Our reproductions of the authors’ LISA analysis found low-high and high-low clusters either not identified or not originally reported. Without further information from the original authors, we are not able to determine whether they did not find these cluster types or omitted them for some other reason. If these clusters were purposefully omitted, this decision represents a cartographic form of observed selective inference.

In addition to these broader issues related to the unit of analysis, there are also specific variable construction concerns with respect to the calculation of population density, which may affect the interpretation of the SLM results. Since LA County is adjacent to a large body of water, several of the hexagons located along the coast line include uninhabitable areas. Because the authors calculate population density by dividing the total population by the total area of the hexagon, as opposed to the inhabitable area contained within the hexagon, they likely underestimate the true population density in these coastal areas. By imprecisely estimating population density for several hexagons, the relationships that are estimated between population density and each outcome may be biased.

These issues are independent of the decision to use tract-level data as the source for all other predictor variables. Because Census tracts tend to have small populations, there is increased error in the estimates provided in the American Community Survey for data at this level. Although the authors appear to try to correct for this by excluding hexagons with populations less than 1,000, by not examining the margins of error associated with the tract-level data retained in the analysis, it is not possible to assess the reliability of the estimates used.

Finally, Vijayan et al. noted that they standardized all variables prior to implementing the spatial regression models in a the caption of original Table 2, however, it is unclear whether the response variables in addition to the predictor variables were standardized. Since the authors do not provide any equation or formulations related to their implementation of the SLM analyses, it is difficult to assess how the models should be properly interpreted. Even so, the language used in the original discussion by the authors imprecisely describes the coefficients reported, ignoring the fact that these are based on standardized variables and that the model intercept was omitted from the analysis.