*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 / Sarah Bardin†,1,2 / Joseph Holler3 / Joshua Gilman / Bryant Grady / Megan Seeley / Xin Wang/ Wenxin Yang

**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:** The relationship between environmental, topographic, socioeconomic, behavioral, and demographic factors and county-level COVID-19 incidence in the U.S. exhibit spatial process non-stationarity  **Original test:** Ordinary Least Squares (OLS), Spatial Lagged Model (SLM), Spatial Error Model (SEM), Geographically Weighted Regression (GWR), Multiscale Geographically Weighted Regression (MGWR) were used to identify the best model fit. |

|  |
| --- |
| **Key Findings** |
| * We were able to partially reproduce the original analyses despite not being able to obtain the original data file used for analysis. Original effect estimates for the regression coefficients were generally in the same direction and of the same magnitude. * The authors conclude that local regression models explain a high degree of variance in COVID-19 incidence, as evidenced by the MGWR model fit statistics, and further recommend the continued monitoring of these factors to understand spread of the disease. However, this recommendation ignores both the poor model fit of the OLS specification and the maps of the local R2 from both the GWR and MGWR models which show large numbers of counties with negative R2 values, which combined suggest an underspecification in their model. |

## Original Study Information

**Description:**

Mollalo et al. (2020) fitted a series of spatial models to evaluate how well each explained the variation in county-level COVID-19 incidence using a set of socioeconomic and demographic characteristics as explanatory variables. In total, five regression models were examined, including an ordinary least square (OLS), spatial lag (SLM), spatial error (SEM), geographically weighted regression (GWR), and a multiscale GWR (MGWR). Neither the data nor the code used for the original analysis were made available by the authors, however the data were publicly available.

**Analytical Plan:**

*Sampling Plan and Data Description:* To explore the association between Covid-19 incidence and sociodemographic and health factors, Mollalo et al. obtained county-level data from a variety of publicly available sources. Specifically, information on the total number of Covid-19 cases between January 22, 2020 and April 9, 2020 were obtained from USAFacts in order to calculate the crude Covid-19 incidence rate for each county in the contiguous U.S (N=3,108). A total of 35 potential explanatory variables were considered in the original analysis, but ultimately only four were included in the final models; these variables included median household income, income inequality (ratio of household income at the 80th percentile to income at the 20th percentile), percentage of black females, and number of nurse practitioners. According to Mollalo et al. (2020), information on the median household income and income inequality for each county were obtained from the Small Area Income and Poverty Estimates (SAIPE) of the American Community Survey 5-Year Estimates, however, income inequality as described by the authors is not available from the SAIPE. Instead, we believe the authors likely sourced this information from the 2020 County Health Rankings. Additionally, while the authors described the number of nurse practitioner variable as a percentage, based on their description of how this variable was calculated in Table 1 of the original article, we assumed that the authors created an estimate of the number of nurses in each county.

*Analytical Specification:* ArcGIS Desktop (version 10.7) was used to calculate COVID-19 incidence rates. GeoDa 1.14 (geodacenter.github.io) was used for calibrating all global regression models (OLS, SLM, and SEM), and MGWR 2.2 (<https://sgsup.asu.edu/sparc/mgwr>) was used to calibrate the local GWR and MGWR models. Information on the coordinate and projection systems were not provided by the authors. For the global regression models, a weight matrix was generated based on a first-order Queens’ contiguity matrix.

*Inference Criteria, Results, and Robustness:* To assess which regression framework best fit the county-level COVID-19 incidence rates, the authors compared the adjusted R2 and AICc of each of the models. The adjusted R2 was substantially higher for the two local models in comparison to the global models, and the AICc was substantially smaller for the local models than the global models. These results led the authors to conclude that the local model specifications achieved better model fits than the global models. In the OLS model, multicollinearity was assessed using the variance inflation factor (VIF) for each variable with all variables having a VIF less than 1.5. The autoregressive lag coefficients were found to be strongly significant (P <0.001).

The authors presented the parameter surfaces generated by the GWR and MGWR models but did not indicate statistical significance in these maps. The authors interpreted the parameter surfaces to indicate that the MGWR model was more conservative than the GWR model, however, they did not provide any formal tests to support this assertion.

## Reproduction Procedure

**Protocol:**

We followed the data preparation and analytical procedures of the original study, making as few modifications as possible. The data and code used in the original analyses were not made available by the authors. As a result, we relied on author descriptions of data collection and processing, variable construction, model specification, and analytical procedures in the published article to perform the reproduction. Based on the information available in the published article, we examined five models.

Consistent with the original study, we grouped our analyses into two families (Fig. 1). First, each model tested whether each of the four county-level socioeconomic and demographic characteristics selected had non-zero associations with the incidence rate of COVID-19. The spatial models additionally tested whether additional terms introduced to capture spatial relationships were non-zero, or whether socioeconomic associations varied locally. Second, the authors compared models based on goodness of fit measures without making any formal statistical comparisons.

Diagram

Description automatically generated

**Fig. 1** Reproduction workflow

We conducted all global regression analyses within the R statistical environment using R version 4.0.1 and Rstudio version 1.3.959. Consistent with the original study, we performed our GWR and MGWR models using MGWR 2.2 software.

*Planned Differences from the Original Study:* Using the list of data sources included in Table 1 of the original study, we attempted to recreate the analysis file used by the original authors. Although we were able to obtain the same data on COVID-19 incidence, percent of black females, and median household income, we cannot be certain that the data we obtained on income inequality, or the number of nurse practitioners were identical to that used in the original analysis. Mollalo et al. noted they obtained income inequality data from the Census SAIPE. However, this source to our knowledge does not report income inequality. Based on the authors’ description of the income inequality variable, we were able to identify the County Health Rankings as the likely source of this data. Therefore, we gathered our measure of income inequality from this source.

The authors also did not provide a specific source for their measure of the number of nurse practitioners in each state. We used occupational data on nursing published by the Bureau of Labor Statistics to construct this measure. Despite indicating that their model used the percentage of nurse practitioners as an independent variable, we assumed they instead used the count of nurse practitioners in each county, as the authors were only able to obtain state-level data for this measure and estimated the county-level counts by multiplying the state-level counts by the proportion of the state population residing in each county. Conversion of the number of nurses to a percentage would reduce the variable to a proportional measure of the total population residing in each county.

Because Mollalo et al. did not report any summary statistics for the variables included in their models, we were unable to assess the extent to which our data aligned with that of the original analysis. For this reason, we did not anticipate being able to achieve bitwise reproduction.

The original authors used GeoDa 1.14 to fit their global regression models (OLS, SEM, SLM). We opted to fit these models using the R statistical environment, since working in R allowed us to produce a reproducible code base.

*Assessment Criteria:* As noted earlier, because Mollalo et al. did not provide specific references for all data included in their analysis, we did not anticipate achieving bitwise reproduction - observing/estimating the exact same statistical values. Instead, we assessed the success of the reproduction based on whether we achieved similar estimates in terms of the direction, magnitude, and statistical significance.

Reproduction Results

**Reproduction Results:**

**Associations between COVID-19 Incidence and county-level socioeconomic and demographic characteristics (F1):**

The authors tested the hypothesis that the four county-level socioeconomic and demographic characteristics selected had non-zero associations with the incidence rate of COVID-19. These hypotheses were tested within five different modeling frameworks. We present the results of the reproduction of each of these frameworks separately.

*OLS Model Reproduction Results:*

Consistent with the original findings, our reproductions indicate that income inequality, median household income, the number of nurse practitioners, and percentage of black females in a county are positively associated with the incidence of Covid-19. As expected, the coefficients in the reproduction differ slightly from those in the original analysis, however, the magnitude and direction of the reproduction coefficients align with those reported by Mollalo.

Table

Description automatically generated

**Table 1.** Summary statistics of the OLS model for the original and replication studies: the original study looked at four predictor variables and COVID-19 incidence.

*SLM and SEM Model Reproduction Results:*

Comparisons of the original and reproduced SLM and SEM models (Tables 2 and 3) also show a high degree of similarity. Both the original and reproductions indicate that all covariates included in the model are statistically significant and the same directions of effects are observed, however the magnitude of effects is generally smaller in the reproductions. For instance, the coefficient on income inequality in the SLM model is nearly three times the size of the coefficient on this variable in the reproduction (0.172 compared with 0.068, respectively). Similarly, the coefficient on median household income is almost twice as large in the original study as in the reproduction (0.183 versus 0.106, respectively). These differences may be due to the underlying data used for the analysis, which may differ enough to lead to these changes on the coefficients of the models.

**Table

Description automatically generated**

**Table 2**. Reproduction summary statistics of SLM in modeling COVID-19 counts, continental United States.

Table

Description automatically generated

**Table 3**. Reproduction summary statistics of SEM in modeling COVID-19 counts, continental United States.

*GWR and MGWR Model Reproduction Results:*

As shown in Figs. 2, GWR reproduction generated similar patterns in the local parameter estimates as demonstrated in the original analysis, however the magnitude of the coefficients differed substantially. For instance, both the original and reproduction analyses show clusters of strongly positive parameter estimates in parts of Idaho, Utah, Louisiana, Georgia, and New York. However, these estimates were nearly twice as large in the reproduction than in the original analysis. Similar results are observed in the maps of the local parameter estimates for income inequality. Both the reproduction and original GWR analyses identify locally varying effects on the number of nurse practitioners and percent of black females, where strongly positive relationships (depicted in purple in the reproduction maps) are found nearby strongly negative relationships (depicted in red in the reproduction maps).

**A picture containing diagram

Description automatically generated**

**Map

Description automatically generated**

**Fig. 2**GWR comparison of original and reproduction results

Comparisons of the MGWR reproductions to the original results reveal less consistency. Although the local parameter estimates for the relationship between household income and COVID-19 incidence both identify a cluster of strong positive associations in parts of Idaho and share a general similarity, the maps for the remaining covariates differ substantially from one another. This is most clear in the parameter estimates for the number of nurse practitioners, which reveals a strong regional east-west pattern in the reproduction map, with communities in the northeast exhibiting the strongest relationships, whereas the original analysis found strong relationships in parts of the Great Lakes region and central United States.

Map

Description automatically generated

Chart, map

Description automatically generated

**Fig. 3**MGWR comparison of original and reproduction results

**Assessment of model performance in explaining county-level variation in COVID-19 (F2):**

Mollalo et al. compared the R2 and AICc statistics of each model to assess how well the different types of regression analysis predicted relationships between the selected socioeconomic and demographic variables and COVID-19 incidence. In the original analysis, Mollalo et al. observed that the local regression models, both the GWR and MGWR specifications, achieved substantially lower AICc values than the global regression models, however, the same conclusion cannot be drawn from the reproductions. While the MGWR model has the lowest AICc, the AICc on the GWR model is roughly the same as the AICc achieved by the SEM and SLM models.

Table

Description automatically generated

**Table 4**. Reproduction summary statistics of model performance by model type

## **Unplanned Deviations from the Protocol**

Our investigation of the coefficient estimates found in Mollalo et al. (2020), led us to believe that the authors standardized their variables prior to analysis. The coefficient of the model intercept and the differences in the magnitude of the individual coefficients are all indicators of scaling in the OLS, SLM, and SEM models. However, we were unable to find any direct discussion of how variables were scaled in the original study.

Based on the labels in the result tables in the original paper, we had planned to use the percent of nurses as a covariate in the models, however, because state level data were apportioned to counties based on the share of population, we elected to use the standardized number of nurses in each county rather than the percent. Although the table labels indicate percentage, we believe these labels were used in error, given the translation needed to calculate this variable.

**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**

Despite not having the original data file, we were able to generally reproduce all global regression models examined by Mollalo et al. (2020). Due to issues in obtaining the nurse practitioner variable and uncertainty regarding the source of the income inequality measure, we were not able to fully reproduce the analysis data set that was used in the original analysis. As a result, we were not surprised that our results differed slightly from those presented in the original paper. However, side-by-side comparisons of the original and reproduction model coefficients, p-values, and maps indicate a strong similarity between the two. We consistently observed the same direction and statistical significance of effects as reported by Mollalo et al (2020). Frequently, we also achieved similar coefficient magnitudes and goodness of fit statistics. Importantly, we were unable to reproduce the AICc values for the SLM and SEM models reported in the original analysis. Instead, our reproduction suggests that the GWR model does not meaningfully improve the model fit over the global spatial regression techniques explored, contrary to the conclusion drawn by Mollalo et al. (2020). FInally, although the model fit statistics were generally consistent between the original analysis and the reproduction for the GWR and MGWR models, an investigation of the coefficient surfaces reveals substantial differences. Given the sensitivity of GWR and MGWR to the distribution of predictors across the study region, it is not surprising that the resulting coefficient surfaces differed between the original and reproduction, since we were unable to confirm whether two of our predictors matched those in the original analysis.

**Procedural Concerns**

The lack of published data, metadata, and code from the original study raised procedural concerns regarding the analysis. First, we were unable to evaluate the extent to which our analysis file aligned to the one used by Mollalo et al. As mentioned previously, we could not locate the source file used for calculating income inequality and percent of nurse practitioners variables based on the limited information supplied in the published article. Second, although we surmised that the authors likely scaled their variables for analysis based on the magnitude of the coefficients and standard practices for local modeling, we cannot be certain that the same scaling procedures adopted in the reproductions were used by the original authors. Finally, 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.

**Statistical and Inferential Concerns**

While Mollalo et al. describe their modeling procedures as a mechanism for assessing the comparative explanatory power of regressive and autoregressive spatial models, their discussion abandons this objective and instead focuses on the descriptive interpretability of the MGWR model. While never explicitly stated by Mollalo et al., it appears they focus their attention on the MGWR results given the higher R2 and low AICc achieved by this model specification. They conclude that the combination of variables included in their models explain a high degree of variance in COVID-19 incidence, as evidenced by the MGWR model fit statistics, and further recommend the continued monitoring of these factors to understand spread of the disease. However, this recommendation ignores both the poor model fit of the OLS specification and the maps of the local R2 from both the GWR and MGWR models which show large numbers of counties with negative R2 values, which combined suggest an underspecification in their model. Furthermore, the substantial difference in the goodness of fit between the local and global models is indicative of overfitting in the local models. Given the added complexity introduced by spatial regression models, it is to be expected that the model fit will improve when using these techniques, however, the model fit alone is insufficient for evaluating the usefulness of the model.

Curiously, Mollalo et al. considered 34 variables for inclusion in their analysis but relied on a stepwise forward procedure to reduce their large set of candidate variables to the final set of 4 variables. This data-driven approach to variable selection limits the descriptive utility of their final model, as the stepwise variable reduction procedure does not consider whether the variables in the final model specification adequately account for confounding. With only four variables it is highly probable that the model does not properly control for additional factors that may influence both the independent and dependent variables, and thus, the model coefficients are likely to be biased. If Mollalo et al. were interested in the causal relationship between sociodemographic and environmental factors and COVID-19 incidence, as suggested in their discussion, alternative modeling procedures and a more theory-driven approach to variable selection is warranted.

Additionally, while Mollalo et al. explored multicollinearity within the OLS model, they did not indicate whether similar checks were performed in the local models. Since GWR and MGWR both fit local regressions around each observation in the dataset, it is possible that multicollinearity may exist within these local regressions that are not apparent in a global OLS model. In this situation, the local parameter estimates will be biased, further reducing the interpretability of the estimated associations.