EFFECT OF THE DELAY IN THE REPORTS OF COVID-19 CASES ON NEAR REAL-TIME CLUSTERS DETECTION

J-F. Mas[1]*, A. Pérez Vega[2], A. Ghilardi[1]

[1] Centro de Investigaciones en Geografía Ambiental, Universidad Nacional Autónoma de México, Morelia, Mexico - (jfmas, adrian)@ciga.unam.mx
[2] Departamento de Geomática e Hidráulica, Universidad de Guanajuato, Guanajuato, Mexico - azupv@ugto.mx

KEY WORDS: COVID-19, clusters, reporting delays, scan statistic.

ABSTRACT:

The COVID-19 pandemic has strongly impacted the vast majority of countries in the world. As of today (April 12th, 2023), more than 762 million confirmed cases and nearly 6.9 million deaths are considered widely underestimated. During a pandemic, detecting clusters of patients is crucial to allocate resources and aiding decision-making better as emergent outbreaks continue to grow. However, delays in reporting suspected or confirmed cases can affect the detection of clusters in near real-time. This study aimed to assess whether the delays in reporting COVID-19 in Mexico presented specific Spatiotemporal patterns and whether they significantly affected the detection of clusters. To do this, we used the daily records of the Mexican Ministry of Health for three dates at the beginning and during the increase in cases of the fourth wave (January 2022). We compared the clusters obtained using the data available on the same date and during the following days, including delayed data. We carried out cluster detection using the flexible spatial scan statistic (FlexScan) on the R platform. The results indicate that the spatial distribution of delays was heterogeneous and that delays affect cluster detection.

1. INTRODUCTION

In December 2019, a new type of coronavirus called SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) was initially documented in Wuhan, located in China. Subsequently, the Coronavirus Disease (COVID-19) pandemic has rapidly disseminated across the world, leading to 763 million confirmed cases and approximately 6.9 million recorded fatalities worldwide, according to the World Health Organization (https://covid19.who.int/). America was a highly impacted region America, particularly the USA, Brazil and Mexico, which presented the most substantial number of deaths.

According to the Ministry of Health of Mexico records, the first confirmed cases of COVID-19 in Mexico were reported in February 2020. COVID-19 has spread throughout the territory, with about 7,553,646 cumulative confirmed cases and 333,596 deaths. Without an effective treatment or vaccine, the Mexican Government declared a health emergency and put in place various sanitary measures to control the spread of the virus.

These measures encompassed a nationwide campaign to encourage social distancing, increased healthcare spending, and the closure of non-essential economic activities from March 23 to May 30, 2020. Following this lockdown period, a phased reopening of economic activities commenced and was adjusted at the state level using color-coded restriction levels (Acuña-Zegarra et al., 2020). These restrictions were determined based on hospital occupancy rates, trends, and the incidence of cases in each state and neighboring areas.

In December 2020, a national vaccination plan against COVID-19 was initiated, initially prioritizing healthcare workers dealing with COVID-19 from December 2020 to February 2021, followed by other healthcare workers and older people from February to April 2021. As of October 23, 2021, the vaccination rate had reached 84.5 doses per 100 people, with 40.7% of the population fully vaccinated and 54.1% receiving at least one dose. By the end of December 2022, approximately 76% of the country's population had received at least one dose of the vaccine.

The Ministry of Health collected the cases of COVID-19 daily and made them available to the public on an open data platform (https://www.gob.mx/salud/documentos/datos-abiertos-152127, accessed September 26, 2023), allowing researchers to use these data.

During an epidemic, it is crucial to carry out spatiotemporal surveillance to identify unusual aggregations of cases in space and time or "clusters". Identifying these clusters allows for prioritizing areas for specific interventions and resource allocation. Spatiotemporal scanning statistics (Kulldorff, 1999) have been widely used for various diseases (Coleman et al., 2009; Zheng et al., 2014), including COVID-19 in various countries (Andersen et al., 2021; Ballesteros et al., 2020; Desjardins et al., 2020; Greene et al., 2021; Hohl et al., 2020; Mas and Pérez-Vega, 2021; Rosillo et al., 2021).

The detection of active outbreaks, which seeks to identify emerging clusters in almost real-time, allows decisions to be made based on the development of the epidemic, , allowing for a rapid and focused response (Desjardins et al., 2020). However, it depends on the quality and updating of the records. In particular, we can assume that spatial biases in the reporting rate and delays in case confirmation and reporting affect the detection of clusters.

This study aims to evaluate the reporting delays of confirmed cases and their impact on the detection of clusters during the COVID-19 pandemic in Mexico.

* Corresponding author

2. MATERIALS

We used both epidemiological and geographical auxiliary data:

- Records of confirmed cases of COVID-19 obtained from the open data platform of the Ministry of Health. The records contain information on the date of admission of the patient to a health unit, the date of confirmation of the case, and the municipality of residence of the patient, allowing data aggregation at the municipal level.
- Digital maps of municipality boundaries and municipal capital cities from the National Institute of Geography and Statistics (INEGI), available at http://en.www.inegi.org.mx/datos/.
- 2020 Population and Housing Census (https://censo2020.mx/) (INEGI, 2021).

To construct the records of confirmed cases of COVID-19, information from each health center or hospital was concentrated by the Health Secretary and put online. However, there are large differences between health centers (for instance, between rural and urban areas) concerning access to the internet, administrative capacity, etc.) and a delay between the date of admission of a patient in the hospital, the confirmation of the case and the integration on the national database (Figure 1).

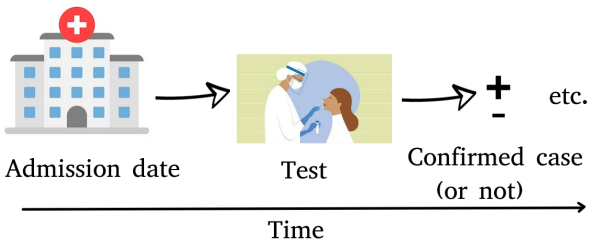

Figure 1. Collecting and processing of COVID-19 confirmed cases information.

We carried out all the analyzes using the R program (R Core Team, 2021), in particular the packages FlexScan (Tango and Takahashi, 2012), gdistance (van Etten, 2017), rflexscan (Otani and Takahashi, 2020), sf (Pebesma, 2018), and spdep (Bivand et al., 2013).

The geographical database relies on the Lambert conformal conic projection, which is a type of map projection using two standard parallels to minimize scale distortion within a specific region. False easting and northing are used to ensure that coordinate values are expressed in meters with positive values. A detailed explanation is presented in Mas (2021). To support practical reproducibility (Nüst and Pebesma, 2021), both the dataset and the associated scripts are accessible on Mendeley Data (DOI: 10.17632/mc37xdzw74.1, https://data.mendeley.com/datasets/mc37xdzw74).

3. METHODS

We obtained the daily databases between January 1 and March 31, 2022. Based on these data, we constructed a three-dimensional array, each dimension representing the municipality, the date of admission of the patient, and the date of registration in the database, respectively. The number of cases and the reporting date corresponding to a specific date of admission can be easily extracted from the array, allowing us to observe the delay between the date of entry of a patient and the date on which the case was reported in the records. We analyzed the admission dates of January 5, 15, and 31, 2002. These dates correspond to the beginning, midpoint, and peak of Mexico's fourth wave of COVID-19 (Figure 2).

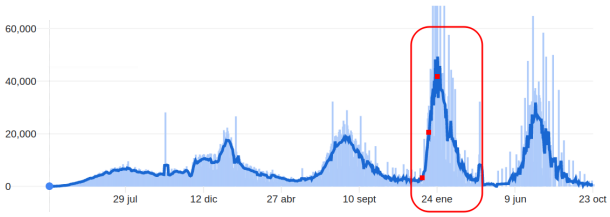

Figure 2. Beginning, midpoint, and peak of Mexico's fourth wave of COVID-19.

We calculated a delay index DI for each date of admission and each municipality, which is the sum of the proportion of cases p_d weighted by the number of days of delay d as shown in equation (1).

$$DI = \frac{\sum_{d=1}^{n} d.p_d}{\sum_{d=1}^{n} d} \quad (1)$$

where DI is the delay index,
p_d is the proportion of cases in the reports,
d is the duration of the delay (number of days).

We examined 1) the relationship between the delay index and the number of cases to assess whether the delay was related to the saturation of health services and 2) the relationship between the delays observed on different dates by calculating the Pearson correlation coefficient.

In the next step, the clusters of a specific date were identified with the data available the next day and subsequent days to assess whether the delays impacted the detection of clusters (Figure 3). For this, we used the "flexibly shaped spatial scanning" (FlexScan) approach proposed by Tango and Takahashi (2005), which allows the detection of irregularly shaped conglomerates (Tango and Takahashi, 2005).

This algorithm can identify irregularly shaped clusters, such as those resembling linear features like communication network. In contrast, algorithms relying on circular windows struggle to identify non-circular clusters accurately. They tend to encompass larger areas than the actual cluster by including surrounding regions (Tango and Takahashi, 2005, 2012). In the context of monitoring epidemics, it is challenging to predict the size of a cluster in advance, especially when the population at risk is not evenly distributed. For instance, when considering the null hypothesis that disease risk is equal inside and outside a cluster, urban areas are expected to have more cases than similarly sized rural areas due to higher population density. Analytical solutions for obtaining probabilities in such complex scenarios have not been found. Therefore, the algorithm utilizes Monte Carlo hypothesis testing to determine p-values (Kulldorff, 1999).

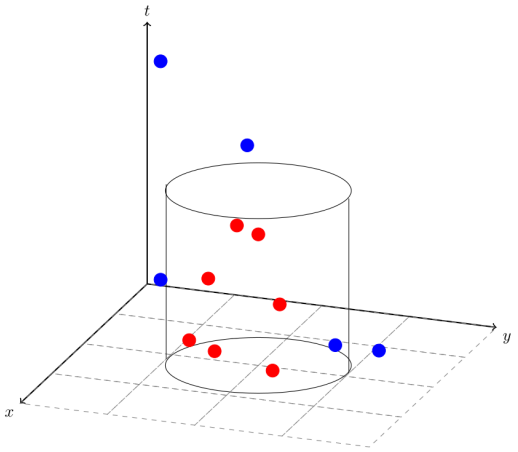

Figure 3. Spatio-temporal scanning based on a space-defined by a cylinder (x and y axis) and time (t axis). The cases inside the cylinder define a cluster (red points) when they are significantly more frequent than outside the cylinder. In Flexscan, an irregular shape is used instead of the ellipse.

The algorithm generates a series of irregularly shaped candidate clusters for each region, such as municipalities, by progressively including connected areas. In essence, it creates numerous distinct but overlapping irregular windows. For each candidate cluster, the algorithm compares the observed number of COVID-19 cases to the expected number. The assumption here is that COVID-19 cases follow a Poisson distribution, consistent with previous spatial COVID-19 studies that employed the spatial scan statistic (Andersen et al., 2021; Ballesteros et al., 2020; Desjardins et al., 2020; Greene et al., 2021; Hohl et al., 2020; Rosillo et al., 2021).

The null hypothesis posits that the distribution of COVID-19 incidence across space is random, while the alternative hypothesis suggests that incidence increases within the cluster. To assess the statistical significance of these clusters, the algorithm estimates the log likelihood ratio (LLR) using Monte Carlo randomization with 999 replications. The p-value is derived by comparing the rank of LLR values from actual data with those from randomized data sets. If this rank is denoted as R, then the p-value can be calculated as equation (2).

$$p = \frac{R}{1 + N_s} \tag{2}$$

where p is the p-value and,
N_s represents the number of simulations.

Statistically significant clusters that do not overlap are retained ($p \leq 0.05$). The sets of municipalities belonging to detected clusters were compared with the data available immediately after the admission date and during the subsequent days. The comparison was made through the Jaccard index, which allows the evaluation of the similarity between the elements of two lists (Real and Vargas, 1996). The index was developed to compare lists of species at various sampling sites and varies between 0 (no elements in common) and 1 (same elements in both lists).

4. RESULTS

Figure 4 shows the number of cases with entry (admission) date of January 15th 2022 reported on the following dates for some municipalities with more than 50 selected cases. We can observe that the curves saturate before day 60, indicating that all entries were recorded within this period. However, in some cases, a significant proportion of the patients take up to 40 days to be integrated into the database. We observed similar patterns for the entry dates of January 5th and 31st (Figures 5 and 6).

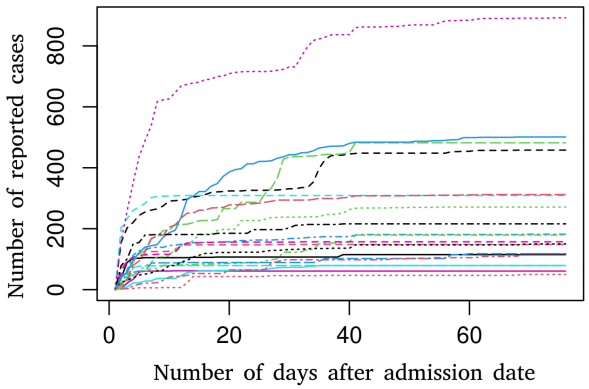

Figure 4. Number of cases of 1/15/2022 reported during the 60 following days for a set of randomly selected municipalities.

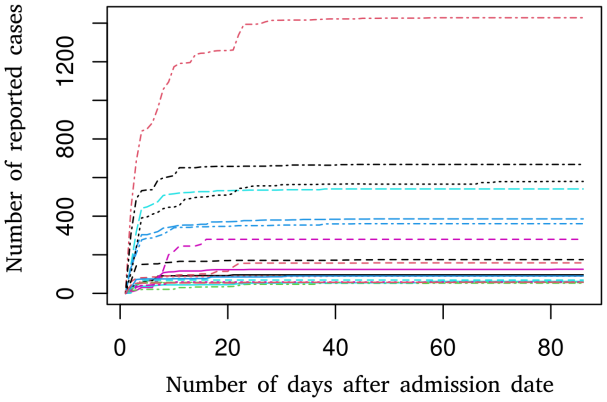

Figure 5. Number of cases of 1/05/2022 reported during the 60 following days for a set of randomly selected municipalities.

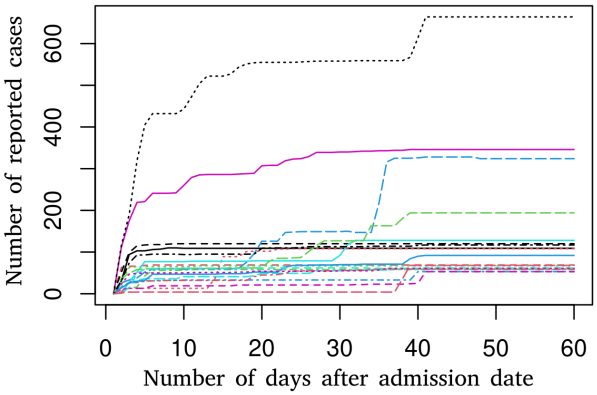

Figure 6. Number of cases of 1/31/2022 reported during the 60 following days for a set of randomly selected municipalities.

Figures 7, 8 and 9 represent the delay index by municipality for January 5th, 2022, January 15th, 2022 and January 31st, 2022, respectively.

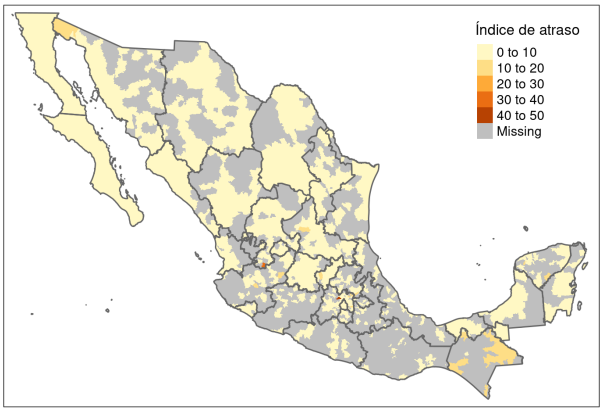

Figure 7. Delay index (January 5th, 2022)

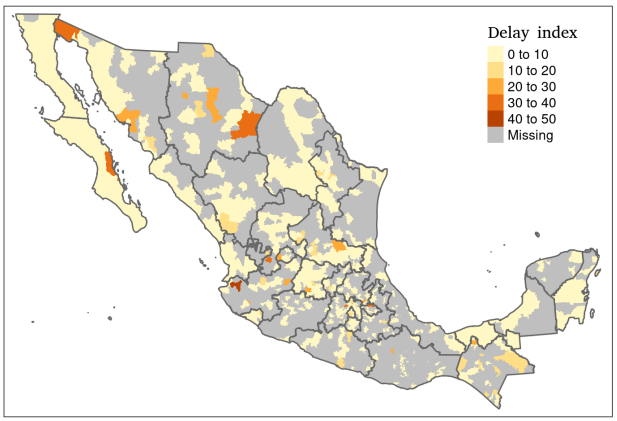

Figure 8. Delay index (January 15th, 2022)

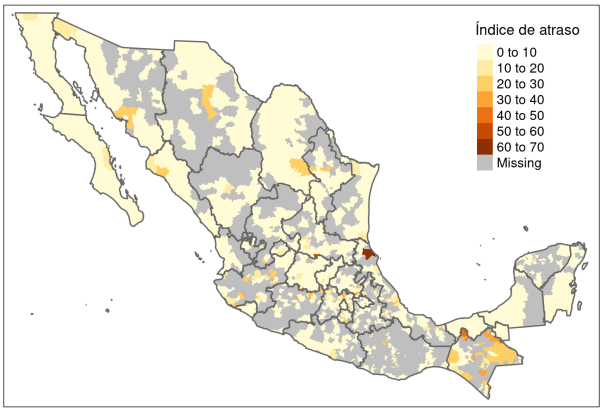

Figure 9. Delay index (January 31th, 2022)

The delay index shows a Pearson correlation coefficient of 0.14 and 0.11 with the number of cases per municipality for the 15th and 31st of January 2022, respectively. For the date of January 5th (beginning of the wave) and considering the proportion of the municipal population infected, the coefficient values are close to zero. On the other hand, the correlation between the delay indices observed for the three dates is between 0.29 and 0.41, which shows that the same municipalities tend to present delays on the three dates.

Date	Number of cases	Case Proportion
January 5th	0.04	-0.03
January 15th	0.14	0.04
January 31st	0.11	0.00

Table 1. Correlation between the delay index and the number of cases and the proportion of cases in population.

	January 15th	January 31s
January 5th	0.41	0.29
January 15th	-	0.40

Table 2. Correlation between delay index of different dates.

Finally, we applied the ”flexibly spatial scanning” to the confirmed cases reported daily after the studied entry dates to detect clusters. Figure 10 presents the value of the Jaccard index among the municipalities that were integrated into a cluster using the complete data (those from the records 60 days after the date of entry) and those obtained between 1 and 10 days after the admission date. We can observe that for the three dates, the Jaccard index reaches 0.8 with data available one week after the entry date, indicating that the detected clusters are very similar to those obtained with the complete data.

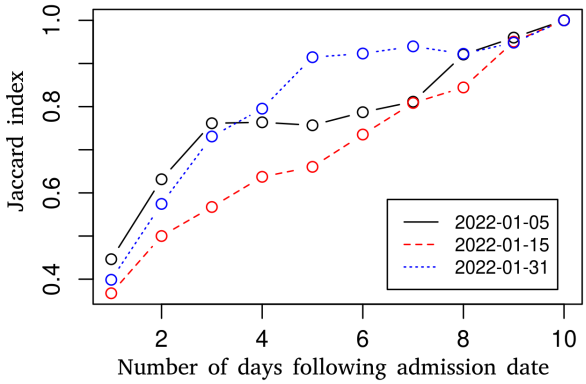

Figure 10. Similarity (Jaccard index) between the clusters obtained with the complete data and the data available between 1 and 10 days after the admission date.

5. DISCUSSION

This results shows that the delay in daily confirmed case registrations affects cluster detection in near real-time but this effect decreases importantly after a five or six days. However, there are additional limitations that this study does not consider as it supposes that the data are complete after a delay of 60 days. We should consider that confirmatory tests are strongly biased towards symptomatic patients. The epidemiological information about confirmed cases represents only a relatively small proportion of all infections (Pullano et al., 2021; Wu et al., 2020). Asymptomatic people account for approximately 30 to 45 per cent of SARS-CoV-2 infections (Oran and Topol, 2020).

When working with aggregated data, an important consideration is the potential impact of aggregation on statistical analysis. This phenomenon, the Modifiable Area Unit Problem

(MAUP), is a type of ecological fallacy (Openshaw, 1984). Interestingly, it has received minimal attention in studies examining the spatial aspects of the COVID-19 pandemic (Wang and Di, 2020). They demonstrated that MAUP could affect the relationship between COVID-19 and atmospheric NO_2.

In our current investigation, we have minimized the influence of MAUP since our primary objective was not to establish correlations between variables, such as the incidence rate versus risk factors. Nonetheless, it is essential to acknowledge that using the total population of a municipality simplifies the analysis and overlooks the population's distribution within that municipality. The contagion patterns are likely to differ significantly between municipalities where most of the population resides in the capital city and those where the population is spread across numerous small settlements (Garland et al., 2020).

6. CONCLUSIONS

This study shows that the delay in daily confirmed case registrations affects cluster detection in near real-time but this effect decreases importantly after a week. However, it would be helpful to analyze the historical data further to assess whether using the unconfirmed cases or the active or weekly cases allows better detection of the clusters. In addition, we should consider that confirmatory tests are strongly biased towards symptomatic patients. Aggregation at the municipality level is also a limitation for decision-making, and data aggregated at the neighbourhood or zip code level would be more appropriate.

References

Acuña-Zegarra, M. A., Santana-Cibrian, M., Velasco-Hernandez, J. X., 2020. Modeling behavioral change and COVID-19 containment in Mexico: A trade-off between lockdown and compliance. *Mathematical Biosciences*, 325, 108370. https://doi.org/10.1016/j.mbs.2020.108370.

Andersen, L. M., Harden, S. R., Sugg, M. M., Runkle, J. D., Lundquist, T. E., 2021. Analyzing the spatial determinants of local COVID-19 transmission in the United States. *Science of The Total Environment*, 754, 142396. https://doi.org/10.1016/j.scitotenv.2020.142396.

Ballesteros, P., Salazar, E., Sánchez, D., Bolanos, C., 2020. Spatial and spatiotemporal clustering of the COVID-19 pandemic in Ecuador. *Revista de la Facultad de Medicina*, 69(1).

Bivand, R. S., Pebesma, E., Gómez-Rubio, V., 2013. *Applied Spatial Data Analysis with R: Second Edition*. Springer, New York.

Coleman, M., Coleman, M., Mabuza, A. M., Kok, G., Coetzee, M., Durrheim, D. N., 2009. Using the SaTScan method to detect local malaria clusters for guiding malaria control programmes. *Malaria Journal*, 8(1), 68. https://doi.org/10.1186/1475-2875-8-68.

Desjardins, M. R., Hohl, A., Delmelle, E. M., 2020. Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: Detecting and evaluating emerging clusters. *Applied Geography*, 118, 102202.

Garland, P., Babbitt, D., Bondarenko, M., Sorichetta, A., Tatem, A. J., Johnson, O., 2020. The COVID-19 pandemic as experienced by the individual. *arXiv:2005.01167v3 (preprint)*. https://arxiv.org/abs/2005.01167.

Greene, S. K., Peterson, E. R., Balan, D., Jones, L., Culp, G. M., Fine, A. D., Kulldorff, M., 2021. Detecting COVID-19 Clusters at High Spatiotemporal Resolution, New York City, New York, USA. *Emerging Infectious Diseases*, 27(5), 1500–1504. https://wwwnc.cdc.gov/eid/article/27/5/20-3583_article.

Hohl, A., Delmelle, E. M., Desjardins, M. R., Lan, Y., 2020. Daily surveillance of COVID-19 using the prospective space-time scan statistic in the United States. *Spatial and Spatio-temporal Epidemiology*, 34, 100354. https://doi.org/10.1016/j.sste.2020.100354.

INEGI, 2021. En México somos 126 014 024 habitantes: censo de población y vivienda 2020.

Kulldorff, M., 1999. An isotonic spatial scan statistic for geographical disease surveillance. *Journal of the National Institute of Public Health*, 48(2), 94–101.

Mas, J.-F., 2021. Spatio-temporal dataset of COVID-19 outbreak in Mexico. *Data in Brief*, 35(106843), 106843. https://doi.org/10.1016/j.dib.2021.106843.

Mas, J. F., Pérez-Vega, A., 2021. Spatiotemporal patterns of the COVID-19 epidemic in Mexico at the municipality level. *PeerJ*, 9, 1-24. https://peerj.com/articles/12685/.

Nüst, D., Pebesma, E., 2021. Practical Reproducibility in Geography and Geosciences. *Annals of the American Association of Geographers*, 111(5), 1300–1310. https://doi.org/10.1080/24694452.2020.1806028.

Openshaw, S., 1984. *The modifiable areal unit problem*. GeoBooks, Norwich, England.

Oran, D. P., Topol, E. J., 2020. Prevalence of Asymptomatic SARS-CoV-2 Infection. *Annals of Internal Medicine*, 173(5), 362-367. https://doi.org/10.7326/M20-3012. PMID: 32491919.

Otani, T., Takahashi, K., 2020. rflexscan: The Flexible Spatial Scan Statistic. Technical report.

Pebesma, E., 2018. sf: Simple Features for R.

Pullano, G., Di Domenico, L., Sabbatini, C. E., Valdano, E., Turbelin, C., Debin, M., Guerrisi, C., Kengne-Kuetche, C., Souty, C., Hanslik, T., Blanchon, T., Boëlle, P. Y., Figoni, J., Vaux, S., Campèse, C., Bernard-Stoecklin, S., Colizza, V., 2021. Underdetection of cases of COVID-19 in France threatens epidemic control. *Nature*, 590(7844), 134–139. https://doi.org/10.1038/s41586-020-03095-6.

R Core Team, 2021. R: A language and environment for statistical computing. Technical report, Vienna, Austria.

Real, R., Vargas, J. M., 1996. The Probabilistic Basis of Jaccard's Index of Similarity. *Systematic Biology*, 45(3), 380-385. https://doi.org/10.1093/sysbio/45.3.380.

Rosillo, N., Del-Águila-Mejía, J., Rojas-Benedicto, A., Guerrero-Vadillo, M., Peñuelas, M., Mazagatos, C., Segú-Tell, J., Ramis, R., Gómez-Barroso, D., 2021. Real time surveillance of COVID-19 space and time clusters during the summer 2020 in Spain. *BMC Public Health*, 21(1), 961. https://doi.org/10.1186/s12889-021-10961-z.

Tango, T., Takahashi, K., 2005. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4(1), 11. https://doi.org/10.1186/1476-072X-4-11.

Tango, T., Takahashi, K., 2012. A flexible spatial scan statistic with a restricted likelihood ratio for detecting disease clusters. *Statistics in Medicine*, 31(30), 4207–4218. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5478.

van Etten, J., 2017. R package gdistance: Distances and routes on geographical grids. *Journal of Statistical Software*, 76(1), 1–21. https://www.jstatsoft.org/v076/i13.

Wang, Y., Di, Q., 2020. Modifiable areal unit problem and environmental factors of COVID-19 outbreak. *Science of the Total Environment*, 740.

Wu, S. L., Mertens, A. N., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., Seth, A., Hsiang, M. S., Colford, J. M., Reingold, A., Arnold, B. F., Hubbard, A., Benjamin-Chung, J., 2020. Substantial underestimation of SARS-CoV-2 infection in the United States. *Nature Communications*, 11(1), 4507. https://doi.org/10.1038/s41467-020-18272-4.

Zheng, S., Cao, C. X., Cheng, J. Q., Wu, Y. S., Xie, X., Xu, M., 2014. Epidemiological features of hand-foot-and-mouth disease in Shenzhen, China from 2008 to 2010. *Epidemiology and Infection*, 142(8), 1751–1762. http://www.scopus.com/.