Stage 1 Registered Report: Spatiotemporal patterns of the COVID-19 epidemic in Mexico at the municipality level (revised version R1)

Jean-Francois Mas [Corresp. 1]

[1] Laboratorio de análisis espacial, Centro de Investigaciones en Geografía Ambiental, Universidad Nacional Autónoma de México, Morelia, Michoacán, Mexico

Corresponding Author: Jean-Francois Mas
Email address: jfmas@ciga.unam.mx

In this stage 1 Registered Report, we propose an analysis of the spatio-temporal patterns of the COVID- 19 epidemic in Mexico using the georeferenced confirmed cases aggregated at the municipality level. We will compute weekly Moran index to assess spatial autocorrelation over time and identify clusters of the disease using the "flexibly shaped spatial scan" approach. Finally, different distance models will be compared to select the best suited to predict inter-municipality contagion. This study will help us understand the spread of the epidemic over the Mexican territory and give insights to model and predict the epidemic behavior.

Stage 1 Registered Report: Spatiotemporal patterns of the COVID-19 epidemic in Mexico at the municipality level

Jean-François Mas[1]

[1]**Laboratorio de análisis espacial, Centro de Investigaciones en Geografía Ambiental, Universidad Nacional Autónoma de México (UNAM), Antigua Carretera a Pátzcuaro No. 8701, Col. Ex-Hacienda de San José de La Huerta, C.P. 58190 Morelia Michoacán MÉXICO**

Corresponding author:

Jean-François Mas[1]

Email address: jfmas@ciga.unam.mx

ABSTRACT

In this stage 1 Registered Report, we propose an analysis of the spatio-temporal patterns of the COVID-19 epidemic in Mexico using the georeferenced confirmed cases aggregated at the municipality level. We will compute weekly Moran index to assess spatial autocorrelation over time and identify clusters of the disease using the "flexibly shaped spatial scan" approach. Finally, different distance models will be compared to select the best suited to predict inter-municipality contagion. This study will help us understand the spread of the epidemic over the Mexican territory and give insights to model and predict the epidemic behavior.

INTRODUCTION

At this date (October 15), the most affected WHO region by COVID-19 pandemic, with 18,090,384 confirmed cases, is the Americas. In this region, the countries that present the most substantial number of deaths are the USA, Brazil and, Mexico.

According to the Mexican Health Secretary's records, the first confirmed cases of COVID-19 in Mexico were reported in February 2020, and community transmission started at the end of March. Since this date, COVID-19 has spread over the Mexican territory, with over 821,000 accumulated confirmed cases and 83,900 deaths by October 15, 2020 (https://www.gob.mx/salud/documentos/datos-abiertos-152127). Measures for containing virus spreading included a national campaign to promote social distancing, called "*Jornada Nacional de Sana Distancia*" (National Workday of Healthy Distance) and the closing of non-essential economic activities from March 23 to May 30, 2020. After this lockdown period, gradual reactivation of economic activities was initiated and tuned at the state level using colour-coded restriction levels (Acuña-Zegarra et al., 2020). The restrictions are determined taking into account hospital occupancy and its trend, and the incidence rates of each state and its neighbours. Schools, colleges and universities suspended classes and activities are carried out remotely.

An integral comprehension of this epidemic, from different perspectives, is needed to prevent and control it. Presently, a large body of research has been accomplished, principally in the field of medicine. However, it is essential to improve our understanding of how COVID-19 spreads over a territory. For instance, the description of the epidemic characteristics, including spatio-temporal distribution and association with other features, may be useful for identifying covariates, modelling epidemic behavior, and providing reliable information for decision-making. Nevertheless, there are very few studies concerning the spatio-temporal dynamics of COVID-19 in Mexico.

This study aims at investigating the spatial distribution patterns and dynamics of the COVID-19 in Mexico from the beginning of the epidemic, totalizing six months by carrying spatial analysis techniques as the computing of spatial autocorrelation and the identification of clustering patterns of the confirmed cases over time.

This Stage 1 Registered Report does not follow a standard report because this research is not based on an experimental design in which independent variables are manipulated and their effect on the dependent variable(s) is evaluated. However, the proposed study is based on robust geo-statistical methods and the statistical significance of the results will be tested. We will analyze the spatio-temporal patterns of the epidemic to test the following hypothesis:

- COVID-19 pandemic in Mexico is characterized by different clusters evolving in space and time as parallel epidemics.

- Connectivity based on a gravitational model allows explaining intermunicipality contagion better than other distances such as Euclidean, least-cost, and resistance distances.

STUDY AREA AND DATASET

The data sets we will use for this study are composed of the epidemiological and geographical auxiliary data:

- The daily numbers of confirmed COVID-19 cases and deaths for the period from January 1 to October 15, 2020, reported by the Secretary of Health were obtained from the Mexican federal government open data platform (https://www.gob.mx/salud/documentos/datos-abiertos-152127).

- Country-based COVID-19 statistics (Total COVID-19 deaths per million people and biweekly cases per million people) from https://ourworldindata.org (University of Oxford).

- Digital maps of municipality boundaries, human settlements, and road networks from the National Institute of Geography and Statistics (INEGI),

- Population projection produced by the National Population Council (CONAPO) to estimate the current population (CONAPO, 2014).

The records from the Secretary of Health contain additional information such as age, sex, foreign and immigrant status, co-morbidities as well as the state and municipality of residence, and the state of the health unit where consultation occurred. In the present study, focused mainly on spatial patterns, we will aggregate data at the municipal level.

Due to the existence of many asymptomatic cases and the reduced number of diagnostic tests, the epidemiological surveillance of confirmed cases represents only a proportion of all infections. A seroepidemiological survey based on a robust sampling can provide precise estimates of seroprevalence in the population than analysis based on confirmed cases (Pollán et al., 2020). However, Mexico carried out a reduced number of tests and confirmed cases dataset is the only available source at the national level. Limitations of this information are that confirmatory testing is strongly biased towards symptomatic cases, and each state makes its own data collection. However, since the federal government (Secretary of Health) mediates the data aggregation, there is some methodological consistency that allows for homogeneity and reasonable comparisons between states and municipalities.

Since there is a delay in case reporting and confirmation, we will limit this analysis to the period between Jan 1st and October 10, 2020. We also will discard records in which the state of residence was different from the state where the patient did the consultation because these records correspond likely to patients who get infected outside of their municipality of residence. These records represent about 6% of the confirmed cases.

All the analyses will be carried out using the R program (version 4.0.2) (R Core Team, 2020), in particular the packages FlexScan 0.2.0 (Tango and Takahashi, 2012), gdistance 1.3-1 (van Etten, 2017),

95 rflexscan 0.3.1 (Otani and Takahashi, 2020), sf 0.9-6 (Pebesma, 2016), and spdep 1.1-5 (Bivand et al.,
96 2008).

97

98 The geographical database is based on the Lambert conformal conic projection, a conic map projection
99 established on two standard parallels, which minimizes deviation from the unit scale within a region
100 comprising the two standard parallels. Its elaboration will be presented in Mas (2020). The dataset and
101 the scripts are available at Mendeley Data:

102 • Repository name: Data_Mexico_COVID19

103 • Data identification number: 10.17632/mc37xdzw74.1 (DOI)

104 • Direct URL to data: https://data.mendeley.com/datasets/mc37xdzw74/1

METHODS

106 For comparison purposes, we will plot the cumulative deaths per million inhabitants of Brazil, France,
107 Mexico, and the US using the country-based data from the University of Oxford. As the epidemic began
108 at different dates in these countries, the origin (day one) in the time axis will be defined as the day when
109 one death per million inhabitants was reached.

110

111 To simplify the data and avoid the day of the week bias, the daily information from the Secretary of
112 Health will be aggregated to the weekly level. The number of new confirmed cases and the number of
113 deaths will be computed for each municipality per week during the period the beginning of community
114 transmission (End of March) and October 10, 2020.

115

116 Spatial autocorrelation analysis enables users to test whether the observed value of a nominal or
117 ordinal variable in one place is independent of values of the same variable in neighbouring areas (Sokal
118 and Oden, 1978). In this case, the spatial autocorrelation assessment will enable us to evaluate if the
119 municipalities with high or low infection rates tend to be spatially aggregated and form clusters. Moran's I
120 index is a measure of spatial autocorrelation that can be used to explore the spatial distribution of diseases
121 (Lawson, 2013). It has been used in various studies concerning COVID-19 (Barrantes Sotela and Solano
122 Mayorga, 2020; Cordes and Castro, 2020; Kang et al., 2020; Yang et al., 2020). To depict the spatial
123 association of COVID-19 cases over time, Moran's I statistic will be computed for each week (Equation
124 1).

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{1}$$

125 where x_i is the number of cases at location (municipality) i, n the number of municipalities, w_{ij} is the
126 weight between observation i and j, and S_0 is the sum of all w_{ij}'s:

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}.$$

127 In its simplest form, the weights take values 1 for close neighbours, and 0 otherwise. We will also set
128 $w_{ii} = 0$ because a region cannot be adjacent to itself.

129

130 During an epidemic, it is crucial to implement spatio-temporal surveillance that can prioritize locations
131 for specific interventions, rapid tests, and resource allocation. One such method is space-time exploration
132 statistics (Kulldorff, 1997), widely used for different types of diseases (Coleman et al., 2009; Zheng
133 et al., 2014) including COVID-19 (Ballesteros et al., 2020; Desjardins et al., 2020; Hohl et al., 2020). As
134 the following step, spatial scan statistic will be applied to detect and evaluate disease clusters using the

135 "flexibly shaped spatial scan" approach proposed by Tango and Takahashi (2005). This algorithm can
136 detect irregularly shaped clusters such as those along with a linear feature as a road, while algorithms
137 based on a circular window have difficulty in accurately detecting non-circular clusters and tends to define
138 a more extensive cluster than the true one by incorporating surrounding regions (Tango and Takahashi,
139 2005, 2012). In epidemic monitoring, the size of a cluster cannot be known a priori, and the population at
140 risk is not evenly distributed. For instance, under the null hypothesis of equal risk of disease inside and
141 outside the cluster, we expect more cases in an urban area compared to a rural area of similar size, due to
142 the higher urban population density. No analytical solutions have been found to obtain the probabilities in
143 these tricky conditions and the algorithm uses Monte Carlo hypothesis test to get the p-values (Kulldorff,
144 1999).

146 A large number of candidate clusters is obtained through the creation of irregularly shaped windows
147 on each region (e.g. municipality) by connecting adjacent regions. For each candidate cluster, the number
148 of observed cases will be compared to the number of expected cases of COVID-19, assuming that the
149 COVID-19 cases follow a Poisson distribution. The null hypothesis is that the incidence of COVID-19 is
150 randomly distributed over space, and the alternative hypothesis is that the incidence increases inside the
151 cluster. In order to test whether the clusters are statistically significant, the log likelihood ratio (LLR) will
152 be estimated by Monte Carlo randomization with 999 replications. The p-value is estimated by comparing
153 the rank of the likelihood from the real data set with the likelihood values from the randomized data
154 sets. If this rank is R, then $p = R/(1 + N_s)$, where N_s is the number of simulations. The non overlapping
155 statistically significant clusters will be retained ($p \leq 0.05$). Clusters will be estimated for each week,
156 allowing to analyze their temporal evolution by computing the date of their first occurrence and their
157 duration over time.

159 For each cluster, the relative risk (RR) will be computed (equation 2). RR is the risk inside a cluster
160 divided by the risk outside the cluster.

$$RR = \frac{c/e}{(C-c)/(C-e)}, \qquad (2)$$

161 where RR is the relative risk, c is the total number of cases in a cluster, e is the total number of expected
162 cases in a cluster, and C is the total number of cases in the country.

164 Finally, an attempt will be made to determine which type of distance measure is better suited to
165 explain disease spreading. For that, the distance between municipalities will be calculated using different
166 approaches: 1) Euclidean distance, 2) least-cost distance, 3) resistance distance, and 4) gravity model
167 interaction.

169 Euclidean distance is a straight line between two locations based only on their coordinates. The other
170 distances, least-cost distance, and resistance distance are based on graph theory. Graphs are obtained by
171 connecting each cell center with its nearest neighbours, which become the nodes of the graph. Weights
172 are associated with each edge and express the conductance (inverse of the resistance). In the present
173 study, the road network map will be rasterized using the speed limit as the value of conductance and
174 a spatial resolution of one kilometer. Cells will be connected with their eight orthogonal and diagonal
175 nearest neighbours (Moore neighbourhood). The least-cost distance is the least costly path to travel
176 from one point to another, taking into account the cost associated with moving through space. The
177 distance is expressed in cost units, in this case, time (hours). The resistance distance allows the incor-
178 poration of multiple pathways, for instance, the least cost route and alternative ones using secondary roads.

The gravity model enables geographers to model the amount of interaction between two places
(Flowerdew and Aitkin, 1982). The expected interaction between two places is proportional to the size
of their populations P_i and P_j, and inversely proportional to the square of the distance between them
(equation 3):

$$I_{ij} = k \frac{P_i P_j}{d_{ij}^2} \qquad (3)$$

where I_{ij} is the interaction between places i and j, k is a constant, and P_i, and P_j are the population sizes of places i and j and, d_{ij} the distance between these two places.

To determinate the geographical position of each region, instead of using the coordinates of the municipality centroid, the coordinates of the municipality head cities will be used. This decision is based on the fact that, at this stage of the epidemic, most of the confirmed cases are found in largest cities (Villerías Salinas et al., 2020), and the distance between head cities represents the connection between municipalities better than the distance between centroids. Additionally, it is worth noting that the gravitational model will enable us to enhance the relationship between the largest cities.

For each week, the municipality newly incorporated into a cluster will be identified (orange regions in Figure 1). The first set of distances includes the distances between these municipalities and the municipalities already belonging to a cluster in the previous week (time t-1). In contrast, the second group comprises distances between the municipalities which remained outside of any cluster (green regions in Figure 1) and the municipalities already belonging to a cluster (red region). The distances will be calculated using the four approaches described previously. The type of distance that explains the spread of the epidemic better is expected to express a larger closeness between the pre-existing clusters and the newly incorporated municipalities. Therefore, it will increase the difference between the mean distance of the two sets (equation 4).

$$RD = \frac{(D_c - D_{nc})}{D_{nc}} \tag{4}$$

where RD is the relative difference between the two sets of distances, D_c is the mean distance between municipalities that belong to a cluster at weeks t-1 and municipalities incorporated into clusters at time t. D_{nc} is the mean distance between municipalities that belong to a cluster a week t-1 and municipalities that not belong to any cluster at week t.
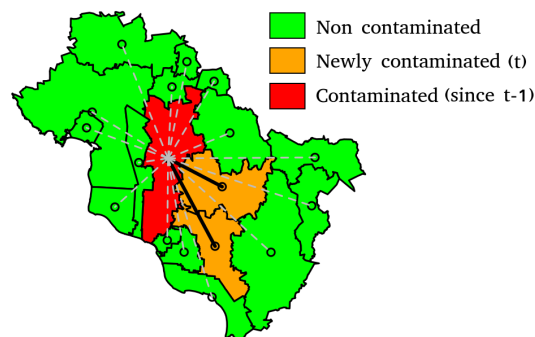
Figure 1. Euclidian Distances between contaminated municipalities a time t-1 and contaninated and non contaminated municipalities at time t. Other types of distance will be calculated using the same sets of municipalities.

REFERENCES

Acuña-Zegarra, M. A., Santana-Cibrian, M., and 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.

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

Barrantes Sotela, O. and Solano Mayorga, M. A. (2020). Comportamiento espacial de la pandemia COVID-19 en Costa Rica durante los meses de marzo y abril de 2020 mediante un análisis de autocorrelación espacial. *Posición*, 3:1–17.

212 Bivand, R. S., Pebesma, E. J., and Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*.
213 Springer, New York.

214 Coleman, M., Coleman, M., Mabuza, A. M., Kok, G., Coetzee, M., and Durrheim, D. N. (2009). Using
215 the SaTScan method to detect local malaria clusters for guiding malaria control programmes. *Malaria*
216 *Journal*, 8(1):68.

217 CONAPO (2014). Proyecciones de la población de los municipios de México. Technical report, Consejo
218 Nacional de Población, Mexico City.

219 Cordes, J. and Castro, M. C. (2020). Spatial analysis of covid-19 clusters and contextual factors in new
220 york city. *Spatial and Spatio-temporal Epidemiology*, 34:100355.

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

224 Flowerdew, R. and Aitkin, M. (1982). A Method of Fitting the Gravity Model Based on the Poisson
225 Distribution. *Journal of Regional Science*, 22(2):191–202.

226 Hohl, A., Delmelle, E. M., Desjardins, M. R., and Lan, Y. (2020). Daily surveillance of covid-19 using
227 the prospective space-time scan statistic in the united states. *Spatial and Spatio-temporal Epidemiology*,
228 34:100354.

229 Kang, D., Choi, H., Kim, J. H., and Choi, J. (2020). Spatial epidemic dynamics of the COVID-19 outbreak
230 in China. *International Journal of Infectious Diseases*, 94:96–102.

231 Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*,
232 26(6):1481–1496.

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

235 Lawson, A. B. (2013). *Statistical Methods in Spatial Epidemiology, Second Edition*. Wiley Series in
236 Probability and Statistics. Wiley, 2nd editio edition.

237 Mas, J. F. (2020). Spatio-temporal dataset of COVID-19 outbreak in Mexico. *Data in brief*.

238 Otani, T. and Takahashi, K. (2020). *rflexscan: The Flexible Spatial Scan Statistic*.

239 Pebesma, E. (2016). sf: Simple Features for R.

240 Pollán, M., Pérez-Gómez, B., Pastor-Barriuso, R., Oteo, J., Hernán, M. A., Pérez-Olmeda, M., Sanmartín,
241 J. L., Fernández-García, A., Cruz, I., Fernández de Larrea, N., Molina, M., Rodríguez-Cabrera, F.,
242 Martín, M., Merino-Amador, P., León Paniagua, J., Muñoz-Montalvo, J. F., Blanco, F., Yotti, R., Blanco,
243 F., Gutiérrez Fernández, R., Martín, M., Mezcua Navarro, S., Molina, M., Muñoz-Montalvo, J. F.,
244 Salinero Hernández, M., Sanmartín, J. L., Cuenca-Estrella, M., Yotti, R., León Paniagua, J., Fernández
245 de Larrea, N., Fernández-Navarro, P., Pastor-Barriuso, R., Pérez-Gómez, B., Pollán, M., Avellón, A.,
246 Fedele, G., Fernández-García, A., Oteo Iglesias, J., Pérez Olmeda, M. T., Cruz, I., Fernandez Martinez,
247 M. E., Rodríguez-Cabrera, F. D., Hernán, M. A., Padrones Fernández, S., Rumbao Aguirre, J. M.,
248 Navarro Marí, J. M., Palop Borrás, B., Pérez Jiménez, A. B., Rodríguez-Iglesias, M., Calvo Gascón,
249 A. M., Lou Alcaine, M. L., Donate Suárez, I., Suárez Álvarez, O., Rodríguez Pérez, M., Cases Sanchís,
250 M., Villafáfila Gomila, C. J., Carbo Saladrigas, L., Hurtado Fernández, A., Oliver, A., Castro Feliciano,
251 E., González Quintana, M. N., Barrasa Fernández, J. M., Hernández Betancor, M. A., Hernández
252 Febles, M., Martín Martín, L., López López, L.-M., Ugarte Miota, T., De Benito Población, I., Celada
253 Pérez, M. S., Vallés Fernández, M. N., Maté Enríquez, T., Villa Arranz, M., Domínguez-Gil González,
254 M., Fernández-Natal, I., Megías Lobón, G., Muñoz Bellido, J. L., Ciruela, P., Mas i Casals, A., Doladé
255 Botías, M., Marcos Maeso, M. A., Pérez del Campo, D., Félix de Castro, A., Limón Ramírez, R., Elías
256 Retamosa, M. F., Rubio González, M., Blanco Lobeiras, M. S., Fuentes Losada, A., Aguilera, A., Bou,
257 G., Caro, Y., Marauri, N., Soria Blanco, L. M., del Cura González, I., Hernández Pascual, M., Alonso
258 Fernández, R., Merino-Amador, P., Cabrera Castro, N., Tomás Lizcano, A., Ramírez Almagro, C.,
259 Segovia Hernández, M., Ascunce Elizaga, N., Ederra Sanz, M., Ezpeleta Baquedano, C., Bustinduy
260 Bascaran, A., Iglesias Tamayo, S., Elorduy Otazua, L., Benarroch Benarroch, R., Lopera Flores, J., and
261 Vázquez de la Villa, A. (2020). Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide,
262 population-based seroepidemiological study. *The Lancet*, 396(10250):535–544.

263 R Core Team (2020). *A Language and Environment for Statistical Computing*. R Foundation for Statistical
264 Computing, Vienna, Austria.

265 Sokal, R. R. and Oden, N. L. (1978). Spatial autocorrelation in biology: 1. Methodology. *Biological*
266 *Journal of the Linnean Society*, 10(2):199–228.

267 Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters.
268   *International Journal of Health Geographics*, 4(1):11.
269 Tango, T. and Takahashi, K. (2012). A flexible spatial scan statistic with a restricted likelihood ratio for
270   detecting disease clusters. *Statistics in Medicine*, 31(30):4207–4218.
271 van Etten, J. (2017). R package gdistance: Distances and routes on geographical grids. *Journal of*
272   *Statistical Software*, 76(1):1–21.
273 Villerías Salinas, S., Nochebuena, G., and Uriostegui Flores, A. (2020). Análisis espacial de vulnerabilidad
274   y riesgo en salud por COVID-19 en el estado de Guerrero, MéxicoNo Title. *Posicion*, 3:1–14.
275 Yang, W., Deng, M., Li, C., and Huang, J. (2020). Spatio-Temporal Patterns of the 2019-nCoV Epidemic
276   at the County Level in Hubei Province, China. *International Journal of Environmental Research and*
277   *Public Health*, 17(7):2563.
278 Zheng, S., Cao, C., Cheng, J., Wu, Y., Xie, X., and Xu, M. (2014). Epidemiological features of
279   hand-foot-and-mouth disease in Shenzhen, China from 2008 to 2010. *Epidemiology and Infection*,
280   142(8):1751–1762.