Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Health Sentinel: A mobile crowdsourcing platform for self-reported surveys provides early detection of COVID-19 clusters in San Luis Potosí, Mexico

Salvador Ruiz-Correa a, *, Rubén López-Revilla b,1, Fernando Díaz-Barriga c, Francisco Marmolejo-Cossío d, Viridiana del Carmen Robledo-Valero e, Emilio Ernesto Hernández-Huérfano a, Leonardo Álvarez-Rivera a, Mónica Liliana Rangel-Martínez e, Miguel Ángel Lutzow-Steiner e, Luis Alfredo Ortiz-Vázquez a, Andrea Rebeca Mendoza-Lara a, Montserrat Olivo-Ramírez a, Marco Sebastián Galván-Ramírez a, Ángel Emanuel Morales-Neri a, Víctor Uriel Martínez-Donjuan a, Massiel Isabella Cervantes-Irurzo a, Andreu Comas-García c, Fernando Hernández-Maldonado c, Carlos Aguilar-Acosta c

a Youth Innovation Laboratory & Grupo de Ciencia e Ingeniería Computacionales, Centro Nacional de Supercomputación, Instituto Potosino de Investigación Científica y Tecnológica, San Luis Potosí, S.L.P., 78216, Mexico  
b Division de Biología Molecular, Instituto Potosino de Investigación Científica y Tecnológica, San Luis Potosí, S.L.P., 78216, Mexico  
c Centro de Investigación Aplicada en Ambiente y Salud, Facultad de Medicina UASLP, San Luis Potosí, S.L.P., 78210, Mexico  
dBalliol College, Oxford University, Oxford, OX1 3BJ, United Kingdom  
e Servicios de Salud de San Luis Potosí, San Luis Potosí, S.L.P., 78200, Mexico  

* Corresponding author at: Youth Innovation Laboratory, Centro Nacional de Supercomputación, Instituto Potosino de Investigación Científica y Tecnológica, San Luis Potosí, S.L.P., 78216, Mexico.

E-mail addresses: salvador.ruiz@ipicyt.edu.mx (S. Ruiz-Correa), rlopez@ipicyt.edu.mx (R. López-Revilla), fdia@uaslp.mx (F. Díaz-Barriga), francisco.marmolejo@es.ox.ac.uk (F. Marmolejo-Cossío), viridiana.robledo@ipicyt.edu.mx (V. del Carmen Robledo-Valero), emilio.hernandez@ipicyt.edu.mx (E.E. Hernández-Huérfano), leonardo.alvarez@ipicyt.edu.mx (L. Álvarez-Rivera), al00285258_slp@hotmail.com (M.L. Rangel-Martínez), mah00@gmail.com (M.A. Lutzow-Steiner), luis.alfredo.ortiz.vazquez@gmail.com (L.A. Ortiz-Vázquez), andreia.mendoza73@outlook.com (A.R. Mendoza-Lara), montse.olivo9@gmail.com (M. Olivo-Rodríguez), marco.zules@gmail.com (M.S. Galván-Ramírez), anemone20@gmail.com (À.E. Morales-Neri), urielmd00@gmail.com (V.U. Martínez-Donjuan), micirurzo@gmail.com (M.I. Cervantes-Irurzo), andreu.comas@uaslp.mx (A. Comas-García), psaerug@hotmail.com (F. Hernández-Maldonado), carlosagalar@acosta@hotmail.com (C. Aguilar-Acosta).

1 Both authors contributed equally to this study.

https://doi.org/10.1016/j.ijmedinf.2021.104508

Received 18 February 2021; Received in revised form 13 May 2021; Accepted 25 May 2021
Available online 28 May 2021

© 2021 Elsevier B.V. All rights reserved.
1. Introduction

The coronavirus disease 2019 (COVID-19) epidemic in Mexico is part of the pandemic caused by the Severe Acute Respiratory Syndrome Coronavirus Virus 2, SARS-CoV-2 [1]. Mexican health authorities confirmed the first COVID-19 case on February 27, 2020 [2] and declared a health emergency on April 9, 2020. Mitigation measures implemented in early March 2020 to avoid healthcare saturation, included voluntary home confinement, maintaining a safe distance between people, suspension of non-critical school and work activities, and frequent hand hygiene [3]. At the time of writing (January 27, 2021), there had been 1,788,905 confirmed COVID-19 cases, and a case fatality ratio (CFR) of 8.5 % in Mexico, the country with the largest CFR [4]. COVID-19 features in Mexico are summarized in Table 1. The disease quickly spread across Mexican cities and on March 13 the first case was confirmed in the San Luis Potosí Metropolitan Area (SLPMA); at the time of writing, the state of San Luis Potosí had 48,486 confirmed cases, and 3632 deaths.

COVID-19 spreads in clusters [5], as individuals pass the virus to their contacts, resulting in clusters of geographically localized infections [6–8]. Cluster mapping has been useful to explore transmission characterized by spatial features distinctive of early infection scattering, community spread, and full-scale outbreak [9,10]. Epidemiological surveillance of COVID-19 cases in Mexico is carried out through sentinel surveillance for early detection of suspected cases through their most frequent signs and symptoms and epidemiological antecedents [11]. Only 0.4 qRT-PCR COVID-19 tests per 1000 inhabitants have been presented here address the research questions listed in Table 2.

COVID-19 apps are informational, self-assessment/medical-reporting, contact tracing, multipurpose, and others; multipurpose apps combine the functionality of at least two of the first three categories [12]. The limited resources to conduct diagnostic tests for early detection of SARS-CoV-2 in San Luis Potosí prompted us to develop the Health Sentinel (CDS), a mobile crowdourcing platform with a multipurpose app (Table S 1) for collecting real-time surveys of self-reported demographic data, signs and symptoms, and vulnerabilities related to COVID-19. The CDS app also serves as a public surveillance tool, to educate users about their risk, and as a community triage to support the Epidemiological Intelligence task force of the UASLP School of Medicine to address the COVID-19 pandemic. CDS data collected from April 1 to July 31, 2020, presented here address the research questions listed in Table 2.

2. Methods

2.1. Study design

This study stemmed from the local VIRUS Program, inspired by the United Nations Human Security methodology [13], focused on collective actions to address community threats and comorbidities associated with COVID-19. It allowed our research team to evaluate the potential contribution of the CDS app for epidemiological surveillance of COVID-19 in the SLPMA, through a descriptive analysis of the users (demographics, risk categories, geographic location, and clusters), and by statistically comparing their spatial and temporal distribution with confirmed COVID-19 cases as a way of validating the CDS app (Table 2).

2.2. Setting and participants

The SLPMA includes the San Luis Potosí City and Soledad de Gra- ciano Sánchez municipalities, subdivided into primary geostatistical areas (AGEBs) [15]. Data collected in the SLPMA consists of two distinct data sets: one originating from survey responses of volunteer app users, and another one of confirmed COVID-19 cases provided by the local health department. Out of the 9971 anonymized COVID-19 cases confirmed through qRT-PCR tests between March 12 and August 3 in the SLPMA, only 9453 (94.8 %) with known postal addresses were included. Privacy laws did not allow app users to indicate if they had been tested and included in the COVID-19 confirmed cases database.

The Mexican Health Ministry established that COVID-19 testing was required for people with suspected COVID-19 symptoms [11]. The overall spectrum of symptom severity is included in the confirmed cases database.

Federal regulations require subjects manifesting COVID-19 symp- toms to provide consent for the use of their data. SSSLP records were stored in the National Supercomputing Center data center at IPICYT, a tier 3 facility designed to assure data integrity and privacy according to Mexican law. Only authorized personnel had access to the SSSLP data.

The CDS requires 11.76 MB of data space and uses state-of-the-art mobile crowdsourcing [14] to collect self-assessment survey data; it was deployed on April 1, 2020, in the SLPMA, and became available through Google Play Store on May 1. CDS app users do not require a personal account and their data are sent to the cloud through an open Wi-Fi connection to prevent the use of their data; users can participate multiple times. The local health department registered the CDS app after checking its compliance with data privacy, security, and ethical requirements. To activate the app, users had to read and agree with the terms of use and consent form. CDS architecture is described in Ap- pendix A. The CDS app was available for Android OS only; 68.3 % of Mexico’s population has access to a smartphone, with 82.5 % Android OS users, and 17.1 % iOS users.

2.3. Personal COVID-19 risk index assessment

A group of medical and public health experts from UASLP School of Medicine, SSSLP, and IPICYT developed the algorithm to compute the individual risk assessment. Risk calculation combines information about the risk of infection and risk of disease severity into one category. This category is most useful at the individual level to help users decide on an informed course of action by following official recommendations from local health authorities, based on the COVID-19 operational definition of the World Health Organization [16]. This category is also helpful to understand the risk of infection across the SLPMA area, but its use to investigate the risk of severity is limited.

The personal profile includes the user’s gender, age, and occupation. Risk index calculation depends on three categories of variables: contact, signs-and-symptoms, and comorbidity. The CDS app enables the administration of the eight questions on risk items (Table 3), conforming to the COVID-19 guidelines of the World Health Organization [17] and the Mexican System for Epidemiological Surveillance [18].

Contact considers whether the user recently traveled to a country with a COVID-19 epidemic, had close contact with someone recently returned from abroad, or has suspected or confirmed infection. Signs and symptoms include cough and fever above 37.5 °C, six general symptoms, and two respiratory symptoms. Comorbidity includes risk factors that can exacerbate the disease, and whether the user belongs to one or more risk groups or has one or more concomitant illnesses that can worsen COVID-19. Risk groups are pregnant women and people aged 65 years or more. Concomitant diseases include diabetes, hypertension, heart problems, liver disease, asthma, chronic pulmonary obstructive disease (COPD), HIV/AIDS, cancer, chronic kidney failure, obesity, and smoking which increase the risk of fatal outcomes associated with COVID-19 [19]. Two questions address social vulnerabilities:
Table 1
Main COVID-19 features in Mexico [2,31,32].

| Feature | Details |
|---------|---------|
| 1 | Around 80% of patients infected by SARS-CoV-2 recover without needing hospital admission. |
| 2 | Around one out of five people who get COVID-19 become seriously ill and develop difficulty breathing. |
| 3 | Obesity is the strongest predictor for COVID-19, followed by diabetes and hypertension in both sexes, and chronic renal failure in women only. |
| 4 | Active smoking is associated with an increased risk of COVID-19. |
| 5 | Men are more likely to be hospitalized and to die than women. |
| 6 | Individuals more than 50 years old are more likely to be hospitalized and die. |
| 7 | Individuals with hypertension, obesity, and diabetes are more likely to be hospitalized and die than those without these morbidities. |
| 8 | Individuals infected by SARS-CoV-2 have a higher chance of dying if they are simultaneously affected by two or more chronic conditions. |
| 9 | There is a high risk of fatal complications from COVID-19 due to the high prevalence of chronic non-degenerative diseases in Mexicans. |

Table 2
Research questions addressed in this study.

| Research Question | Details |
|------------------|---------|
| RQ1: | What are the sociodemographic features of the app users? |
| RQ2: | How are the risk indexes of the app users geographically distributed? |
| RQ3: | Does the point-pattern distribution of risk cases and confirmed COVID-19 cases show statistically significant and consistent clustering features? |
| RQ4: | At what spatial scale are the clustering features of risk cases statistically significant; do they change with time and correlate with the clustering features of COVID-19 confirmed cases? |

1) Can you stay home regardless of developing COVID-19 symptoms? and 2) Do you have someone (relative, friend, neighbor) who can help you in case you get ill?; these questions are motivated by the fact that a significant proportion of Mexicans have informal employment and cannot stay at home [20].

Personal risk indexes are, in order of increasing severity: risk-free, low-risk, moderate-risk, high-risk, severe-risk. Based on their risk index, users were subdivided into risk-free cases, risk cases (low- to moderate-risk, moderate-risk, high-risk, severe-risk), and respiratory-risk cases (Table 4). Since the risk index is a triage tool, respiratory-risk cases without evidence of COVID-19 are treated separately, atypical pneumonia being one of the possible diagnoses [21].

2.4. Georeferenced database

Data collected through CDS includes the georeferenced users’ contributions, and a web application allows data visualization through a dashboard (Fig. 1) as soon as the user completes the survey.

During the survey collection period, health authorities issued a home confinement recommendation. Visual inspection of the GPS data shows that most users completed and sent their surveys from residential areas, probably from their place of residence. These areas also encompass social and recreational activities.

2.5. Spatial data analysis

Coronavirus infection spreads in clusters, whose early identification is essential for slowing down SARS-CoV-2 transmission [9,22]. The spatial analysis allowed us to determine whether the geographical features of the data collected with the CDS app suggest the presence of COVID-19 clusters across geographical space and whether they coincide with those of confirmed COVID-19 case clusters. Two complementary statistical tools enabled quantification of spatial features: computation of point process descriptive statistics, and spatial correlation modeling. We emphasize that the risk index calculation combines information about risk of infection and risk of disease severity into one category. Therefore, clustering information derived from the index is useful to understand risk of infection, but not about risk of disease severity. This is important to interpret the data presented in the results section.

2.5.1. Point process descriptive statistics

We used kernel density estimation tools to estimate point process intensity functions [23] and visually explore risk-free and risk cases. We used Ripley’s K-function and Monte Carlo simulations to describe how point patterns occur over a given area and to determine whether they are dispersed, clustered, or randomly distributed. Point process statistics computation is described in Appendix B.

2.5.2. Spatial correlation modeling

Point process statistics allowed us to establish clustering features in CDS geographic data from observed point pattern correlations. An approach to account for these correlations is the log-Gaussian Cox process model [24], which describes the spatial distribution of symptoms as a function of explanatory variables and estimates a residual process independent of these variables (Appendix C).

3. Results

3.1. Socio-demographic users’ features (RQ1)

Contributors to CDS surveys resided in Mexico (n = 2125) and other countries (United States, Bolivia, Chile, Spain, Argentina, France, and Bangladesh; n = 64). Only SLPMA users (n = 1554) were included in this study, whose overall mean age ± SD was 33.5 ± 11.5 years, with a median of 30.0 years (Table 5). The predominant age group was that of 20–29 years (44.0%). The three largest groups, spanning 20–49 years, comprised 85.1% of the total. Female and male contributors represented 49.6% and 50.4%, respectively.

3.2. Spatial analysis of the risk cases

3.2.1. Geographic distribution of risk indexes (RQ2)

The SLPMA has 1,094,177 inhabitants distributed over 1,787.72 km². This area is subdivided into AGEBs with a mean ± SD size of 3.80 ±
5.88 km², and a mean ± SD population of 2348 ± 3098 inhabitants.

A population density map is shown in Fig. 2, and the AGEBs heat-map of risk cases shown in Fig. 3 depicts the normalized number of risk reports per AGEB, conditioned to a given risk index, to provide information on the conditional probability of the number of reports. This figure intends to depict the normalized number of risk reports per AGEB, conditioned to a given risk category (low-risk, medium-risk, etc.); namely the heat maps provide information on the conditional probability of the number of reports. This figure intends to depict the normalized number of risk reports per AGEB, conditioned to a given risk index (low-risk, medium-risk, etc.).

### Table 4
Calculation of the users’ risk index. (For interpretation of the references to color in this table, the reader is referred to the web version of this article).

| Contact (1-3 points) | Signs and symptoms (4-6 points) | Vulnerability (7-8 points) | Color-coded risk index |
|----------------------|---------------------------------|-----------------------------|------------------------|
| No                   | No                              | No                          | Risk-free              |
| Yes                  | No                              | Yes                         | Low risk               |
| Yes                  | Yes                              | No                          | Moderate risk           |
| Yes                  | Yes                              | Yes                         | High risk               |
| Yes                  | Yes                              | Yes                         | Severe risk             |
| Yes                  | No                              | Yes                         | Respiratory risk        |

*Based on the answers to the survey questions of the three risk categories enlisted in Table 3.

Fig. 1. CDS platform front-end dashboard. The web application allows epidemiologists and public health specialists to visualize real-time data through a dashboard. The geographic information system displays several data layers, including population density, poverty index, overcrowding, medical center locations, business locations, and contributing users' locations. The dashboard also shows descriptive statistics of all the variables of the CDS survey. Color codes: green (risk-free), low-risk (yellow), medium-risk (orange), high-risk (red), very high risk (maroon), respiratory signs only (purple). The image on the top right shows the location of the SLPMA (population 1.4 million) within Mexico (population 135 million). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).
the number of reports. The following data are presented as supplementary material: personal risk index distribution by gender and age in Table S 2, occupation distribution of risk cases by gender in Table S 3, personal risk index distribution by gender in Table S 4, and overall distribution of risk factors in Table S 5.

Point process analysis of the risk cases generated darker green areas indicating higher intensity. Fig. 4 depicts the spatial distribution of all survey contributions, no-risk cases, and risk (low- to severe-risk) cases. This figure aims to provide an alternative visualization of raw data collected and how they segregate as a function of all users’ risk and no-risk contributions. These functions consider kernel density estimation techniques with a bandwidth $h = 1200$ m. Lower bandwidth values are more likely to overfit data. Note that Fig. 4c summarizes raw data from Fig. 3a–d.

Most CDS contributions corresponded to non-risk cases (Fig. 4b); risk cases amounted to 15.77 %, and their maximum intensity was located in an area of high mobility and social interaction (Fig. 4c) which is also a populated residential area.

### 3.2.2. Point pattern clustering features (RQ3)

#### 3.2.2.1. Point process descriptive statistics

Ripley’s $L$ function was computed using Procedure I of Appendix B to test whether the point patterns of risk cases fulfill the nonhomogeneous CSR hypothesis. The reference probability measure was given by the AGEBs population (Fig. 2). $L$ functions of Periods I (April-May) and II (April-July) are shown in Fig. 5.

The $L$-function (solid line) and the corresponding upper and lower envelopes (dashed lines) for risk cases computed with Period I and Period II data are shown in Fig. 5a and c, respectively. The $L$ function is above the upper envelope at scales between 0 and 600 m (0 and 730 m), suggesting that accumulated risk cases do not fulfill the CRS hypothesis and exhibit clustering features in this spatial range ($L > L_{up} > 0, p < 0.001$). Fig. 5b and d show the corresponding $p$-values, where the red dashed line represents $p = 0.01$.

### Table 5

Age and gender distribution of CDS data

| Age group (years) | Male (M) | Female (F) | All | M/F ratio |
|-------------------|----------|------------|-----|-----------|
| 10–19             | 83       | 21         | 55  | 1.50      |
| 20–29             | 316      | 132        | 684 | 1.02      |
| 30–39             | 102      | 30         | 132 | 1.11      |
| 40–49             | 53       | 16         | 69  | 0.84      |
| 50–59             | 21       | 6          | 27  | 0.75      |
| Total             | 783      | 500        | 1554| 100.0%    |

* In the SLPMA the predominant age groups are, in descending order, those of 10–19 years (18.65 %), 20–29 years (17.54 %), 0–9 years (16.65 %), and 40–49-years (13.29 %). Our study did not consider children under 10 years of age.

In the SLPMA the predominant age groups are, in descending order, those of 10–19 years (18.65 %), 20–29 years (17.54 %), 0–9 years (16.65 %), and 40–49-years (13.29 %). Our study did not consider children under 10 years of age.
Fig. 3. Heat-map of risk cases per AGEB. Geographic distribution of the risk indexes of the users’ contributions. Each AGEB is color-coded according to the number of contributions recorded within it, divided by the total number of AGEB contributions. Color bars show the number of normalized counts as a percentage. The darker color in a given AGEB indicates higher count contributions. The color palette used for each map indicates a specific risk level: (a) low-risk, yellow; (b) medium-risk, orange; (c) high-risk, red; (d) very high-risk, brown; (e) respiratory symptoms, purple; (f) no-risk/no-symptoms, green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Fig. 4. Intensity functions in the geographic space. Estimated sample intensity functions are shown as green-shaded areas in the maps for the point processes representing app users’ contributions: (a) all users’ contributions; (b) no-risk cases; (c) risk cases (those between low- and severe-risk). Black plus signs represent app users’ contributions. Darker green areas represent higher intensity. Locations with maximum empirical intensity values are indicated with a red plus sign. Kernel density estimation techniques (with kernel bandwidth $h = 1200$ m) enabled the calculation of intensity functions. The maximum intensity value for risk cases (Fig. 4c) is located across the street of the City’s ISSSTE Hospital General, a few meters away from one of the oldest tree-lined pedestrian streets (Calzada de Guadalupe). People use this avenue for recreational/religious purposes, such as visiting the Nuestra Señora de Guadalupe Sanctuary. Close to this location is a children’s park, the Center for the Arts, the Red Cross, and a military garrison. Mobility and social interactions in this area were significant during the study period despite the mitigation measures recommended by health authorities. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).
Procedure 2 of Appendix B enabled local tests of homogeneous CSR hypotheses. Fig. 6 shows color-labeled AGEBs of the SLPMA with p-values < 0.05 calculated at scales of 300, 450, and 600 m for Period I and Period II. Several AGEBs exhibited statistically significant clusters (Fig. 6 d–f) at different scales in the 0–730 m range. Risk-cases clusters from Period II are those statistically significant at a 730 m scale. Period I has similar results at different scales in the 0–600 m range (Fig. 6 a–c) with risk case clusters as those statistically significant at a 600 m scale. Some risk clusters computed in Period I intersect those computed in Period II (compare Fig. 6 a–c and d–f plots).

3.2.2.2. Spatial correlation model. We conducted Bayesian inference of a log-Gaussian Cox model of the risk cases data [25] considering seven covariates: population density, male proportion, age, travel abroad or contact with someone who traveled abroad, contact with a suspect COVID-19 case, contact with a confirmed COVID-19 case, risk conditions, and comorbidities. Calculations were conducted over a 128 × 128 grid. The Markov chain Monte Carlo inference algorithm was run with 2, 200,000 iterations, an initial burn-in of 200,000 iterations followed by 2,000,000 further iterations from which every 1000th sample was stored. Markov chain mixing and convergence were conducted through standard techniques [24]. To reveal SLPMA areas with a high incidence of risk symptoms (Fig. 7) we also computed exceedance probabilities (Appendix C) with the variance $\sigma^2$, scale $\phi$, and other latent field parameters enlisted in Table 6.

3.2.3. Distribution of risk-clusters and confirmed COVID-19-clusters (RQ4)

We compared the distribution of clusters identified from the 9971 SLPMA COVID-19 cases confirmed between March 12 and August 3, with the distribution of risk case clusters identified from the April-July surveys (Period II). Only 9453 confirmed COVID-19 patients (94.8 %) with known postal addresses were included in the comparison. Individual records included demographics, date when symptoms started, and date when the diagnosis was confirmed. Several neighborhoods had clusters of COVID-19 cases confirmed during Periods I and II (Table 7). A summary of the findings related to RQ4 are listed in Table 10.

4. Discussion

Early detection of COVID-19 hotspots has been achieved in Israel through one-minute population-wide surveys [9]. In the United Kingdom, a mobile app for real-time tracking of self-reported symptoms predicted potential COVID-19 [26]. CDS is a multi-use platform allowing early detection of COVID-19 hotspots that may be of use to prevent spreading, especially at the beginning of the epidemic, when data of confirmed cases are scarce.

The sociodemographic features of CDS users result in a sampling bias whose sources, listed in Table 9, restrict the assessment to the surveyed

Fig. 5. Spatial distribution patterns of risk cases. Period I: (a) L-function (solid line) and upper and lower envelopes (dashed lines) from Monte Carlo simulations for risk cases. The L function is above the upper envelope at scales between 0 and 600 m, suggesting significant clustering features ($L > L_{up} > 0$, $p < 0.001$) at these scales. (b) The corresponding p-values for each h in the analysis range. The red dashed line represents a p-value of 0.0. Period II: (c) L-function (solid line) and upper and lower envelopes (dashed lines) resulting from MC simulations for the risk cases. (d) The L function is above the upper envelope at scales between 0 and 730 m, suggesting significant clustering features ($L > L_{up} > 0$, $p < 0.001$) at these scales. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).
Fig. 6. Risk-case clusters. Local tests of homogeneous CSR hypotheses. These tests were conducted to reveal point clusters at different spatial scales ($h = 300, 450,$ and $600$ m). Point clusters are defined as geographical areas where $p$-values $< 0.05$. (a-c) Period I (April-May). (d-f) Period II (April-July).

Fig. 7. Areas of a high incidence of risk symptoms for Period II data. The posterior exceedance probabilities $P(\exp(Y) > k |X)$ computed for several values of $k$ identified areas of the SLPMA with a particularly high incidence of risk symptoms. The figure shows the exceedence probability for $k = 1$. 
population [27].

Most CDS contributions corresponded to non-risk cases (Fig. 4b); risk cases amounted to 15.77 % and their maximum intensity was located in an area of high mobility and social interaction (Fig. 4c). This urban space is also a populated residential area.

Risk cases from Period II (April-July) had statistically significant spatial clustering independent of population density across geographical scales from 0 to 730 m (Fig. 6D-f), and risk clusters were computed at a scale of 730 m. Independence of clustering from population density is further supported by the fact that the Pearson’s correlation coefficient between density and number of risk cases per AGEB is low ($\rho = 0.17$) as show in Fig. 2. Similar results were obtained for clusters of COVID-19 cases confirmed in Period I (April-May) at a scale of 600 m (Fig. 6a-c). Many Period II clusters have significant overlap, are more prominent, and expand several Period I clusters (Fig. 8).

Interestingly, Period II clusters also overlap city areas where the exceedance probabilities are close to one (Fig. 7). Risk case clusters in areas with a high incidence of risk of symptoms in Period I evolved to become the larger confirmed COVID-19 clusters of Period II (Figs. 7 and 8).

### Table 6
Summary of the latent field parameters.

| Parameter | Significant effect | Relative risk | Median (95 % Confidence Interval) |
|-----------|--------------------|---------------|----------------------------------|
| $\sigma$  | –                  | –             | 1.66 (1.39–1.90)                  |
| $\psi$    | –                  | –             | 153.20 (99.51–240.60)             |
| Population density | Yes        | Increase     | 2.32 (1.90–2.90)                  |
| Travel abroad | Yes        | Increase     | 9.50 (4.92–19.59)                 |
| Contact with suspect COVID-19 patient | Yes      | Increase     | 7.84 (4.27–14.35)                 |
| Contact with confirmed COVID-19 patient | Yes      | Increase     | 2.55 (1.38–4.84)                  |
| Risk conditions | Yes      | Increase     | 2.05 (1.13–3.83)                  |
| Age       | No                 | Reduction    | 0.98 (0.97–2.46)                  |
| Male proportion | No        | Increase     | 1.46 (0.83–2.46)                  |
| Comorbidities | No             | Increase     | 1.23 (0.75–2.19)                  |

### Table 7
Top five SLPMA neighborhoods with clusters of confirmed COVID-19 cases during March 12 and August 3, 2020.

| Month       | Accumulated neighborhoods | Accumulated cases | The top five COVID-19 case clusters per neighborhood |
|-------------|---------------------------|-------------------|------------------------------------------------------|
| March (M1)  | 27                        | 42                | Los Angeles 5 11.9                                    |
| April (M2)  | 85                        | 147               | Los Angeles 5 11.9                                    |
| May (M3)    | 287                       | 616               | Los Angeles 5 11.9                                    |
| June (M4)   | 587                       | 1827              | Los Angeles 5 11.9                                    |
| July (M5)   | 1076                      | 6180              | Los Angeles 5 11.9                                    |
| August (M6) | 1315                      | 8592              | Los Angeles 5 11.9                                    |

The exponential growth of COVID-19 cases in the SLPMA began in June when public activities partially reopened, and SARS-CoV-2 hotspots increased significantly between May and August (Tables 7 and 8). The top infected neighborhood reached 161 confirmed COVID-19 cases with an 82 % increase between June and August (Table 8).

Analysis of the epidemiological curve and the trend line of risk-clusters by neighborhood shed light on the associations between risk clusters, symptoms, and actual infection hotspots (Fig. 9). Analysis of cluster A (Fig. 8b and c) showed that risk cases had symptoms between April and May (red line of Fig. 9) before confirmed COVID-19 cases started to appear during the transition to the exponential phase of the
Fig. 8. Socio-urban analysis of statistically significant point-pattern clusters. Risk-case clusters are shown by the shaded areas labeled A–E. Cluster A (294 confirmed COVID-19 cases), intersects several of the neighborhoods with the highest incidence of COVID-19 cases listed in Tables 6 and 7. Cluster B (138 confirmed COVID-19 cases) intersects the Tequisquiapan area that ranks second among the confirmed SARS-CoV-2 clusters. Cluster C (37 confirmed COVID-19 cases) intersects neighborhoods adjacent to the Tequisquiapan area. Cluster D (56 confirmed COVID-19 cases) includes the largest city park surrounded by a trail for strolling/jogging/biking. Cluster E (68 COVID-19 cases), on the northwest side of the park, intersect five neighborhoods. Cluster F (77 confirmed COVID-19 cases) intersects Graciano Sanchez municipality, which includes seven neighborhoods. H and J are groups of risk cases that also intersect areas with a high incidence of COVID-19 confirmed cases. (a) April-May data, (b) April-July data, (c) Exceedance probabilities $P(\exp(Y) > k \mid X)$ associated with the covariate-adjusted latent field $Y$ (April-July data).

Fig. 9. Cumulative number of risk cases per week in risk cluster A from April to July (solid red line), and epidemiological curve of confirmed cases from March 12 to August 3 (solid black line). Both curves indicate risk cases with symptoms (April-May) preceded confirmed COVID-19 cases occurring during the exponential phase of the pandemic. Cluster A intersects 10 relevant neighborhoods: Bolivar, Torres del Santuario, Barrio de San Sebastián, Barrio de San Miguelito, Himno Nacional, Julián Carrillo, San Luis Rey, San Juan de Guadalupe, Independencia y Barrio de Guadalupe. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).
infection hotspots confirmed later, and that early risk-clusters predict between March 12 and August 31, 2020.

Persistence and growth of clusters of confirmed COVID-19 cases in the SLPMA between March 12 and August 31, 2020.

| Neighborhoods with five top clusters | Cases accumulated | Cluster growth across time | Persistence (months) |
|--------------------------------------|-------------------|---------------------------|----------------------|
| Centro                               | 161               | 82.0% (M4-M6)             | 3                    |
| Tequisquian area                     | 138               | 81.1% (M3-M6)             | 4                    |
| Progreso                            | 107               | 0.0% (M6)                 | 1                    |
| Simón Díaz                          | 97                | 25.8% (M5-M6)             | 2                    |
| Capricornio                          | 97                | 0.0% (M6)                 | 1                    |
| Las Mercedes                         | 69                | 68.1% (M4-M5)             | 2                    |
| Satellite                            | 66                | 0.0% (M5)                 | 1                    |
| San Ángel                            | 26                | 57.7% (M3-M4)             | 2                    |
| Jardines del Sur                     | 24                | 50.0% (M3-M4)             | 2                    |
| 21 de Marzo                          | 12                | 50.0% (M2-M3)             | 2                    |
| San Ángel Inn                        | 9                 | 0.0% (M3)                 | 1                    |
| Industrial Aviación                  | 7                 | 0.0% (M2)                 | 1                    |
| Maya Mil                             | 6                 | 0.0% (M2)                 | 1                    |
| Forestal                             | 5                 | 0.0% (M1-M2)              | 2                    |
| La Loma                              | 5                 | 0.0% (M1-M2)              | 2                    |
| Tangamanga                           | 2                 | 0.0% (M1)                 | 1                    |
| San Pedro                            | 2                 | 0.0% (M1)                 | 1                    |

a M1, month 1; M2, month 2; M3, month 3; M4, month 4; M5, month 5; M6, month 6.

Key findings on RQ4.

Most CDS contributions corresponded to non-risk cases (Fig. 4b); risk cases amounted to 15.77%, and their maximum intensity was located in a densely populated residential area of high mobility and social interaction (Fig. 4c).

Risk cases from Period II (April-July) had statistically significant spatial clustering independent of population density across geographical scales from 5 to 730 m (Fig. 6 d-f) and risk clusters were computed at a scale of 730 m. Similar results were obtained for clusters of COVID-19 cases confirmed in Period I (April-May) at a scale of 600 m (Fig. 6 a-c). Many Period II clusters have significant overlap, are more prominent and expand several Period I clusters (Fig. 8).

Period II clusters also overlap city areas where the exceedance probabilities are close to one (Fig. 7). Risk case clusters in areas with a high incidence of risk of symptoms in Period I evolved to become the larger confirmed COVID-19 clusters of Period II (Fig. 8).

Risk case clusters A-E (Fig. 8) were compared with the quantitative cluster analysis of COVID-19 confirmed cases. Cluster A (294 cases) had several cases comparable to the top-five confirmed neighborhoods. Cluster B (138 cases) and Cluster C (37 cases) intersect an area ranking second among confirmed SARS-CoV-2 hotspots. Cluster D (56 cases) includes the largest city park surrounded by a trail for strolling/jogging/biking. Cluster B (68 cases) on the northwest side of the park intersects five neighborhoods. Cluster F (77 cases) intersects the Graciano Sanchez municipality, which includes seven neighborhoods.

Comparing the curve of cluster A showing the cumulative number of risk cases per week surveyed in Period II with that of confirmed cases in March 12 and August 3 (Fig. 9) we observed that risk cases started in Period I, before the transition to the exponential phase of confirmed COVID-19 cases.

| Source | Features |
|--------|----------|
| The highest proportion of CDS users were those of 20–29 years (44.02%) and 30–39 years (26.96%). The predominant age groups were, in descending order, those of 10–19 years (18.65%), 20–29 years (17.54%), 0–9 years (16.65%), and 40–49 years (13.29%). Children under 10 years were not considered in this study. The predominant CDS groups are adults with a smartphone and internet connection. The elderly population is less likely to have access to and master the mobile technology deployed in the field; the analysis was directed to risk cases younger than 65 years. The CDS app was available for Android OS only: 68.3% of Mexico’s population has access to a smartphone, with 82.5% Android OS users, and 17.1% iOS users; the rest use Samsung, Windows, and PlayStation devices. Asymptomatic individuals infected with SARS-CoV-2 are not eligible for COVID-19 testing due to the limited number of tests available and therefore are not considered in the SSSLP data. |

An adequate response to the COVID-19 pandemic demands a fine-grained understanding of the population under study, since interventions have disparate effects on target populations, depending on multiple factors [28]. Since variation in the outcomes of pandemic interventions involves allocating extremely limited resources to different population sectors elicits different outcomes, the CDS platform may be used as a valuable tool to optimize resource allocation.

Having a limited supply of COVID-19 tests is not a constraint unique to San Luis Potosí. Whatever few tests available used to diagnose those already with symptoms is not as informative when 40%–45% of those infected with SARS-CoV-2 will remain asymptomatic and the potential of the virus to spread silently may be significant [29]. Whereas extensive contact tracing via testing has been successful in South Korea and other countries, such a scale of testing is infeasible in Mexico. One possible solution involves using heterogeneous population factors to decide how to allocate tests differentially [30]. This approach may profit from using a tool as CDS to quantify risk and different citizen profiles in diverse geographies. Information on the spatial distribution of cases during the spread of transmissible diseases is crucial to design and execute effective interventions. The CDS platform allows documentation and exploration of COVID-19 risk self-assessment data to enable epidemiological intelligence tasks, and may be extended to address epidemiological events concurrent with COVID-19, such as the next expected influenza A and B outbreaks. Survey data could also provide useful information to allocate vaccination campaigns across the state of San Luis Potosí and elsewhere.

Despite the biases of this study (Table 9), our results support the hypothesis that the risk clusters identified early coincide with the infection hotspots confirmed later, and that early-risk-clusters predict the areas of further SARS-CoV-2 spreading.

The key points for future work are listed in Table 11.

Authors’ contributions

Conception and design of the study: ACG, ARML, FDB, MLRM, MOR, RLR, SRC, VCRV. Data acquisition: CAA, FHM, MALIS, MCI, EEHH. Analysis and interpretation of data: ACG, AEMN, FDB, MSGR, RLR, SRC, VCRV, VUMD. Drafting the manuscript: FMC, LAR, SRC, VCRV, RLR. All authors revised the manuscript and read and approved the version submitted.

Future work key points.

An adequate response to the COVID-19 pandemic demands a fine-grained understanding of the population under study, since interventions have disparate effects on target populations, depending on multiple factors [28]. Since variation in the outcomes of pandemic interventions involves allocating extremely limited resources to different population sectors elicits different outcomes, the CDS platform may be used as a valuable tool to optimize resource allocation.

Having a limited supply of COVID-19 tests is not a constraint unique to San Luis Potosí. Whatever few tests available used to diagnose those already with symptoms is not as informative when 40%–45% of those infected with SARS-CoV-2 will remain asymptomatic and the potential of the virus to spread silently may be significant [29]. Whereas extensive contact tracing via testing has been successful in South Korea and other countries, such a scale of testing is infeasible in Mexico. One possible solution involves using heterogeneous population factors to decide how to allocate tests differentially [30]. This approach may profit from using a tool as CDS to quantify risk and different citizen profiles in diverse geographies. Information on the spatial distribution of cases during the spread of transmissible diseases is crucial to design and execute effective interventions. The CDS platform allows documentation and exploration of COVID-19 risk self-assessment data to enable epidemiological intelligence tasks, and may be extended to address epidemiological events concurrent with COVID-19, such as the next expected influenza A and B outbreaks. Survey data could also provide useful information to allocate vaccination campaigns across the state of San Luis Potosí and elsewhere.

Summary table

- Mobile phone data providing services such as contact tracing, mobility, and proximity assessment have proven to be of use for COVID-19 epidemiologic surveillance and support of appropriate public health interventions in several countries.
- We developed and successfully deployed CDS, a mobile crowdsourcing platform for COVID-19 surveillance that includes the CDS app, to survey self-reported signs, symptoms, and vulnerabilities in a city located in central Mexico.
- Real-time assessment of the users’ self-reported risk factors by the CDS platform provides early detection, prediction, and evolution of actual COVID-19 case clusters.
Declarations of Competing Interest

The authors report no declarations of interest.

Acknowledgments

The study was partially funded by Consejo Estatal Electoral y de Participación Ciudadana (CEEPAC) de San Luis Potosí, Consejo Potosino de Ciencia y Tecnología (COPOCYT), IPICYT’s Youth Innovation Laboratory, and Servicios de Salud del Estado de San Luis Potosí (SSSLP). We are also grateful to the ACM special interest group on economics and Facebook.

References

[1] T.P. Velavan, C.G. Meyer, The COVID-19 epidemic, Trop. Med. Int. Health 25 (2020) 278–280, https://doi.org/10.1111/tmi.13383.

[2] V. Suárez, M. Suárez Quezada, S. Oroz Ruiz, E. Ronquillo De Jesús, Epidemiología de COVID-19 en México: del 27 de febrero al 30 de abril de 2020, Rev. Clínica Española. 220 (8) (2020) 463–471, https://doi.org/10.1016/j.rcle.2020.05.007.

[3] Secretaría de Salud, Gobierno de México, Inicia la fase 3 por COVID-19: Coronavirus, Inicia La Fase 3 Por COVID-19, 2020 (Accessed 24 June 2020), https://coronavirus.gob.mx/2020/04/23/inicia-la-fase-3-por-covid-19-2/.

[4] Coronavirus Resource Center, Mortality Analyses - Johns Hopkins Coronavirus Resource Center, 2020 (Accessed 24 June 2020), https://coronavirus.jhu.edu/data/mortality.

[5] E.B. Hodcroft, Preliminary case report on the SARS-CoV-2 cluster in the UK, France, and Spain, Swiss Med. 150 (2020), https://doi.org/10.4414/smc.2020.20212.

[6] J. Cai, W. Sun, J. Huang, M. Gamber, J. Wu, G. He, Indirect Virus Transmission in Wuhan based on social media data, ISPRS Int. J. Geo-Inf. 9 (2020) 1343–1345, https://doi.org/10.3390/ijgi9060312.

[7] J.R. Koo, A.R. Cook, M. Park, Y. Sun, J.T. Lim, C. Tam, B.L. Dickens, Interventions to mitigate early spread of SARS-CoV-2 in Singapore: a modelling study, Lancet Infect. Dis. 20 (2020) 678–688, https://doi.org/10.1016/S1473-3099(20)30162-6.

[8] H. Rosman, A. Kenhet, S. Shilo, A. Gavioli, T. Bauman, O. Cohen, E. Shelly, R. Balicer, B. Geiger, Y. Dor, E. Segal, A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys, Nat. Med. (2020) 1–4, https://doi.org/10.1038/s41591-020-0857-9.

[9] Z. Peng, R. Wang, L. Liu, H. Wu, Exploring urban spatial features of COVID-19 transmission in Wuhan based on social media data, ISPRS Int. J. Geo-Inf. 9 (2020) 402, https://doi.org/10.3390/ijgi9050402.

[10] Secretaría de Salud, Comunicado emergente sobre el uso de pruebas para la detección de antígenos SARS-CoV-2 en México, 2020. https://coronavirus.gob.mx/wp-content/uploads/2020/11/Prueba_antigenica_COVID_11Nov2020.pdf.

[11] M. Azevedo Silva, COVID-19 Apps, 2020. https://viena.org/document/covi d-19-Apps.

[12] Human Security Unit of the United Nations, Human Security Handbook. An Integrated Approach for the Realization of the SDG’s, 2016. https://www.un.org/humansecurity/wp-content/uploads/2017/10/hs2.pdf.

[13] J. Phutharak, S.W. Loke, A review of mobile crowdsourcing architectures and challenges: toward crowd-empowered internet-of-things, IEEE Access 7 (2019) 204–324, https://doi.org/10.1109/ACCESS.2018.2885353.

[14] INEGI, Catálogo Único de Claves de Áreas Geostatísticas Estatales, Municipales y Localidades, 2020. https://www.inegi.org.mx/app/ageemi/.

[15] WHO, Coronavirus Disease (COVID-19) Situation Reports, 2020. https://www. who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/.

[16] World Health Organization, Digital technology for COVID-19 response, Newsletter (2020) 1 (Accessed 13 October 2020), https://www.who.int/news/item/03 -04-2020-digital-technology-for-covid-19-response.

[17] Sistema Nacional de Vigilancia Epidemiológica, Síntico oficial COVID-19 México, 2020 (Accessed 13 October 2020), https://covid19.sanea.gob.mx/.

[18] Y. Zhou, Q. Yang, J. Chi, B. Dong, W. Lv, L. Shen, Y. Wang, Comorbidities and the risk of severe or fatal outcomes associated with coronavirus disease 2019: a systematic review and meta-analysis, Int. J. Infect. Dis. 99 (2020) 47–56, https://doi.org/10.1016/j.ijid.2020.07.029.

[19] International Labor Organization, Informal Employment in Mexico: Current Situation, Policies and Challenges, 2014. https://www.iilo.org/wcmsp5/groups/public/---americas/---ro-lima/documents/publication/wcms_245889.pdf.

[20] Secretaría de Salud, Flujoograma diagnóstico de la enfermedad respiratoria aguda en el contexto de la sedinemia COVID-19, 2020. https://coronavirus.gob. mx/wp-content/uploads/2020/10/Algoritmo_Dx_COVID_Flu_05oct2020.pdf.

[21] S.M. Kisliger, Geographic and Demographic Transmission Patterns of the 2009 A/H1N1 Influenza Pandemic in the United States, University of Cambridge, 2017. http://www.repository.cam.ac.uk/handle/1810/273772.

[22] C. Wang, L. Han, G. Stein, S. Day, C. Bien-Gund, A. Mathews, J.J. Ong, P.Z. Zhao, S. Menn, A.M. Valdes, M.B. Freidin, C.H. Sudre, L.H. Nguyen, D.A. Drew, T.D. Spector, Real-time tracking of self-reported symptoms to predict potential COVID-19, Nat. Med. 26 (2020) 1037–1040, https://doi.org/10.1038/s41591-020-0916-3.

[23] C. Wang, L. Han, G. Stein, S. Day, C. Bien-Gund, A. Mathews, J.J. Ong, P.Z. Zhao, S. F. Wei, J. Walker, R. Chou, A. Lee, A. Chen, B. Bayus, J.D. Tucker, Crowdsourcing in health and medical research: a systematic review, Infect. Dis. Poverty 9 (2020) 1–48, https://doi.org/10.18637/jss.v063.i07.

[24] M. Akbarpour, C. Cook, A. Marzouli, A. Nagaraj, M. Saccorla, P. Tebaldi, S. Vasserman, H. Yang, Socioeconomic network heterogeneity and epidemic transmission, Proc. Nat. Acad. Sci. U. S. A. 117 (2020) 15835–15842, https://doi.org/10.1073/pnas.2004186117.

[25] M. Kissler, Human Security Handbook: An Integrated Approach for the Realization of the SDG’s, 2016. https://www.wiley.com/en-mx/Applied+Social+Statistics+ for+Public+Health+Data-p984071397718.

[26] M. Kissler, Geographic and Demographic Transmission Patterns of the 2009 A/H1N1 Influenza Pandemic in the United States, University of Cambridge, 2017. https://doi.org/10.17863/CAM.20834.

[27] L. Cuaca, Intensity estimation for spatial point processes observed with noise, Scand. Stat. Theory Appl. 35 (2008) 322–334, https://doi.org/10.1111/j.1467- 9469.2007.00583.x.

[28] B.M. Taylor, T.M. Davies, B.S. Rowclinson, P.J. Diggle, Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-gaussian cox processes in R, J. Stat. Softw. 63 (2015) 1–48, https://doi.org/10.18637/jss.v063.i07.

[29] L.A. Waller, C.A. Gotway, Applied Spatial Statistics for Public Health Data, John Wiley & Sons, 2004. https://www.wiley.com/en-mx/Applied+Social+Statistics+ for+Public+Health+Data-p984071397718.

[30] J. Phuttharak, S.W. Loke, A review of mobile crowdsourcing architectures and challenges: toward crowd-empowered internet-of-things, IEEE Access 7 (2019) 204–324, https://doi.org/10.1109/ACCESS.2018.2885353.

[31] D. Pulido, B. Basurto, M. Cándido, J. Salas, Geospatial Spread of the COVID-19 Pandemic in Mexico, ArXiv Prepr, 2020 (Accessed 29 June 2020), https://arxiv. org/abs/2006.07784.

[32] E. Hernández-Garduño, Obesity is the comorbidity more strongly associated for Covid-19 in Mexico. A case-control study, Obes. Res. Clin. Pract. 14 (2020) 375–379, https://doi.org/10.1016/j.orcp.2020.06.001.