

Carnegie Mellon

ML Model Uses Time,

Model for predicting transmission of COVID-19 can help policymakers monitor virus, inform health surveillance systems

Newswise — The COVID-19 pandemic presents unprecedented challenges to public health worldwide. Tracking the dynamics of the coronavirus permits governments, organizations, and individuals to make projections in an effort to curb the spread of the pandemic. But while a large amount of data about COVID is collected and publicly available, the information can be unreliable and subject to bias. In a new study, researchers analyzed data from Cali, Colombia, to develop a model that provides a template for tracking data, predicting transmission, and informing health surveillance systems.

The study was conducted by researchers at Carnegie Mellon University (CMU), Georgia Institute of Technology (Georgia Tech), Universitat Jaume I, and Universidad Nacional de Colombia. It is published in the *Journal of the Royal Statistical Society*.

“Most COVID-19 studies chronicle overall infection at a state or county level, reporting the aggregated number of cases in a particular region at a particular time,” explains Shixiang Zhu, assistant professor of data analytics at CMU’s Heinz College, who coauthored the study. “This tends to miss fine details of the virus’ propagation patterns.”

Zhu and his colleagues analyzed a high-resolution COVID-19 data set in Cali, the second-largest city in Colombia, with more than half the population living in neighborhoods of low socioeconomic status (SES), from March 15 to September 30, 2020. The data set, from the Municipal Public Health Secretary of Cali, documents the location and time of every confirmed case in the city, not just the combined number of cases or deaths in a geographic area.

The authors created a model based on a point process that changes over time and space, where previously infected individuals lead to new cases, and they used a neural network-based technique to account for the varying impact of location on this process. They also incorporated external influences imposed by city landmarks (e.g., churches, schools, town halls), and considered factors such as population density, since COVID-19 spreads through respiratory droplets, and aerosol transmission is higher in crowded and inadequately ventilated spaces.

The researchers also studied the real data, which revealed the unique dynamics of COVID-19 transmission and confirmed that several of the city's landmarks played an important role in the virus' spread. In particular, the model suggested an increased risk of contracting COVID-19 in the center, northeast, and northwest of Cali, which is where people of lower SES live, and a lower risk in the south of the city, which is where people of higher SES live.

Comparing the model with the real data, the study found that the model was successful at predicting the spread of COVID-19. As such, it can help policymakers monitor coronavirus dynamics and provide a template for tracking real-time data for future epidemics and informing health surveillance systems.

“High-resolution data sets like the one we used will be more widely available in the future, so the approach we used in Cali is not limited to that jurisdiction,” notes Zheng Dong, a Ph.D. student in machine learning at Georgia Tech's H. Milton Stewart School of Industrial and Systems Engineering, who led the study. “In fact, it can be used, extended, and adapted to several natural phenomena represented by locations in space and time.”

The study was funded by the National Science Foundation and the Universidad Nacional de Colombia.