



Satellite Imagery Analysis: What Can Hospital Parking Lots Tell Us about a Disease Outbreak?

Patrick Butler and Naren Ramakrishnan, *Virginia Tech*

Elaine O. Nsoesie and John S. Brownstein, *Harvard Medical School and Boston Children's Hospital*

Data mined from satellite imagery could serve as an early indicator of socially disruptive events like epidemics, especially in countries with limited surveillance resources.

Digital epidemiologists continually search for novel data sources that could be useful for surveying, mapping, and predicting infectious diseases. Examples include call data records, Web searches, social media text (for example, tweets), and online news reports.

A particularly interesting data source is satellite imagery. Researchers have used such historical images to assess population movement related to measles transmission in Niger¹ and to characterize environmental factors associated with hantavirus transmission.² Satellite imagery analysis is also regularly employed to measure as well as to predict company

growth and consumer demand, though it has not been used for prospective disease surveillance—until now.

Chicago-based Remote Sensing Metrics (www.rsmetrics.com) used satellite imagery of cars in Walmart parking lots to develop a regression model for making monthly predictions of the company's quarterly revenue. This application motivated us to collaborate with RS Metrics to examine hospital traffic as a possible indicator of an influenza epidemic. With funding provided by the Intelligence Advanced Research Projects Activity (IARPA) Open Source Indicators (OSI) program (www.iarpa.gov/Programs/ia/OSI/osi.html) to mine and assess public data sources for early indicators of

socially disruptive events such as disease outbreaks in Latin America, we focused on hospital parking lots in Mexico, Chile, and Argentina.

INITIAL DATASET

From RS Metrics, we obtained archival satellite imagery data of hospital parking lots in all three countries. Using Google Earth/Google Maps, Bing Maps, online hospital lists, and hospital ranking lists, we initially created a master list of 164 hospitals and then reduced this list to hospitals with parking lots having more than 40 spaces, yielding 74 unique locations. RS Metrics provided 2,575 satellite images of these hospitals taken at specific times each day from 1 November 2011 through 26 May 2013.

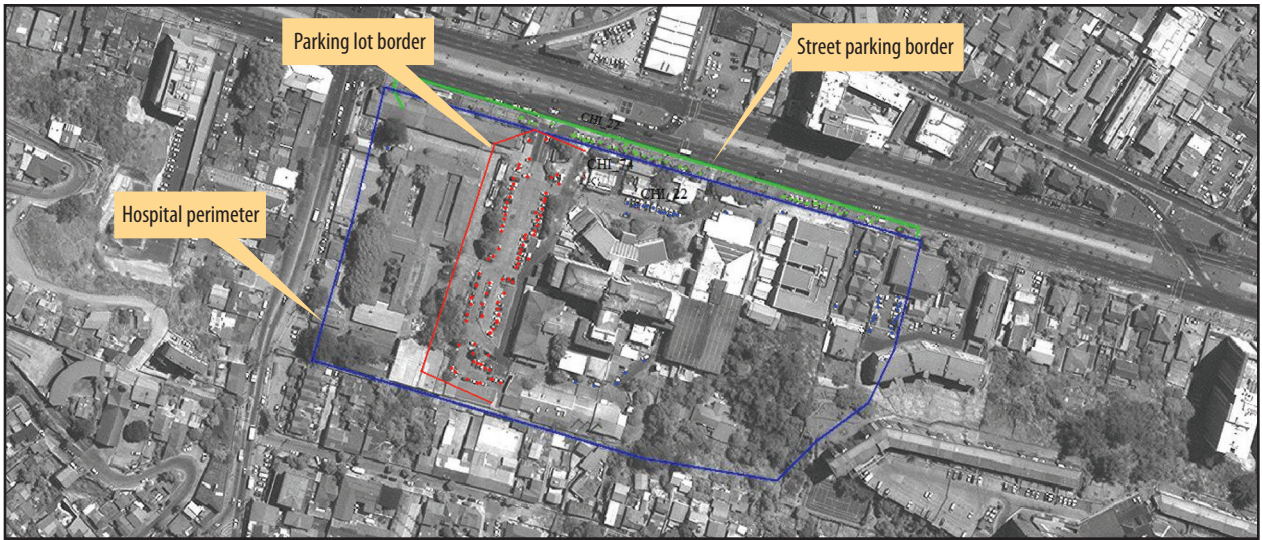


Figure 1. Satellite image of Hospital Dr. Gustavo Fricke in Valparaíso, Chile, with contours of key areas highlighted in different colors.

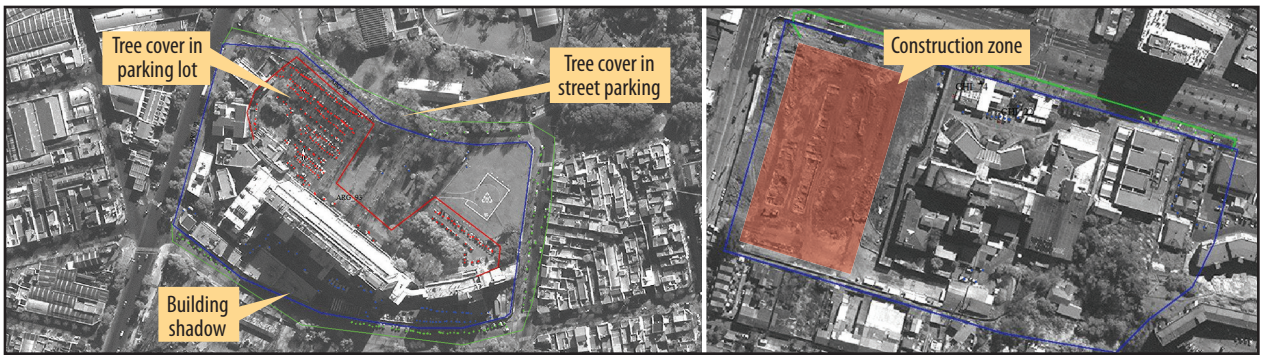


Figure 2. Examples of imperfections in satellite images of hospitals that can confound automatic algorithmic analysis of parking lots: tree cover and building shadows (left) and construction activity (right).

As Figure 1 shows, RS Metrics used virtual stencils to demarcate the parking lot border (red), hospital premises (blue), and street parking, if available (green), in each image for automated analysis. As Figure 2 shows, however, preprocessing revealed imperfections in many images—including tree cover, building shadows, construction activity, and difficulties precisely defining the contours—that could lead to over- or undercounting the number of vehicles. Furthermore, as archival images, they weren't regularly spaced in time, and representation was nonuniform across hospitals, making many images unsuitable. Consequently, of the original 2,575

images, we retained only 1,304 (50.6 percent).

For each of these images, RS Metrics used algorithms to automatically estimate the number of vehicles in the parking lot, along the hospital border, and on the street, the number of parking lot spaces, and the “fill” or occupancy rate. The dataset also included the date and time of each image as well as the hospital's geographic location (address, latitude and longitude) and name.

EXPANDING AND REFINING THE DATA

Using this dataset, we developed a least absolute shrinkage and

selection operator (LASSO) regression model³ to make weekly predictions of influenza-like illness (ILI) cases in Mexico, Chile, and Argentina. LASSO is a modified form of least squares regression that minimizes the sum of squared errors and also encourages sparsity in the number of terms utilized.

As ground truth, we obtained ILI data spanning the same time period as the hospital parking lot data from the Pan American Health Organization (PAHO). For each epidemiological week, PAHO provides the number of ILI cases per country in Latin America.

Based on an initial fit of the LASSO model to PAHO data, we

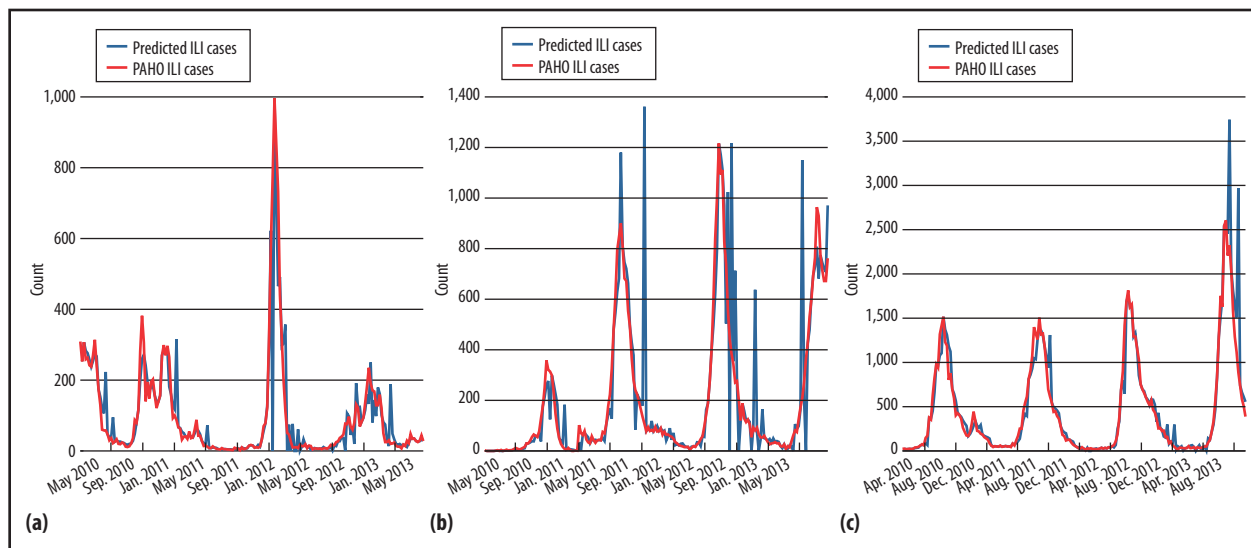


Figure 3. Comparison of influenza-like illness (ILI) case counts from the Pan American Health Organization (PAHO) and predicted ILI case counts using a LASSO (least absolute shrinkage and selection operator)-based parking lot occupancy rate model for (a) Mexico, (b) Chile, and (c) Argentina.

identified the most significant types of hospitals in our dataset. The majority (54 percent) were general care hospitals, such as Hospital Español Sociedad de Beneficencia Española and Hospital San José in Mexico. The model underemphasized specialized hospitals, such as those focusing on psychiatry.

RS Metrics expanded the list of hospitals in the initial dataset to include more general care facilities. This ultimately resulted in a new dataset that, after preprocessing, consisted of 2,564 satellite images of hospitals from January 2008 to September 2013. This expanded coverage was intended to capture trends before and after the H1N1 influenza pandemic of 2009.

DATA ANALYSIS

Using this expanded and refined dataset, we developed a LASSO model to forecast the weekly ILI case count based on PAHO data from the previous four weeks and parking lot occupancy rates. We used Spearman's rank correlation coefficient to assess the similarity in trends between parking lot occupancy rates and PAHO data. Although the correlation was not significant, peaks in parking lot volume appeared to precede peaks in influenza incidence, as Figure 3 shows. We obtained a normalized root mean square error of 0.074 for Mexico, 0.119 for Chile, and 0.58 for Argentina.


While our prediction method works surprisingly well, it also

creates many artificial peaks (false positives). This likely stems from two reasons. First, as a flu outbreak progresses the intensity of cases in different regions of the country vary over time, and thus our model requires extra data to better infer the locations of such outbreaks. Second, because the data for each hospital is collected at irregular intervals, it can quickly become out of date; to some extent, this problem can be mitigated by training on only the most recent data.

Monitoring hospital traffic as an early indicator of disease outbreak is a promising concept, especially for countries with limited public health surveillance resources. As our study indicates, however, satellite imagery data needs to be carefully defined, extracted, and refined. Moreover, our current model doesn't include other factors that impact parking lot occupancy trends including natural disasters, social unrest, seasonality, and the hospital's distance from a metropolitan region. In future work, we plan to

DISCLAIMER

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

incorporate such factors as well as explore the use of other novel data sources for ILI prediction. 

References

1. N. Bharti et al., "Explaining Seasonal Fluctuations of Measles in Niger Using Nighttime Lights Imagery," *Science*, 9 Dec. 2011, pp. 1424–1427.
2. G.E. Glass et al., "Satellite Imagery Characterizes Local Animal Reservoir Populations of Sin Nombre Virus in the Southwestern United States," *Proc. Nat'l Academy of Sciences*, 24 Dec. 2002, pp. 16817–16822.

3. R. Tibshirani, "Regression Shrinkage and Selection via the LASSO," *J. Royal Statistical Soc., Series B*, vol. 58, 1996, pp. 267–288, 1996.


Patrick Butler is a PhD candidate in computer science at Virginia Tech. Contact him at pabutler@cs.vt.edu.

Naren Ramakrishnan, *Discovery Analytics* column editor, is the Thomas L. Phillips Professor of Engineering at Virginia Tech and director of the university's *Discovery Analytics* Center. Contact him at naren@cs.vt.edu.

Elaine O. Nsoesie is a postdoctoral research fellow at Harvard Medical School and Boston Children's Hospital. Contact her at onelaine@vt.edu.

John S. Brownstein is an associate professor at Harvard Medical School and directs the Computational Epidemiology Group at the Children's Hospital Informatics Program, Boston Children's Hospital. Contact him at john.brownstein@childrens.harvard.edu.



 Selected CS articles and columns are available for free at <http://ComputingNow.computer.org>.



stay connected.

Keep up with the latest IEEE Computer Society publications and activities wherever you are.

IEEE  computer society

| | |
|--------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
|  @ComputerSociety @ComputingNow |  facebook.com/IEEEComputerSociety facebook.com/ComputingNow |
|  IEEE Computer Society Computing Now |  youtube.com/ieeecompulersociety |