Preliminary analysis of socioeconomic variable correlation with geospatial modeling in Costa Rica dengue epidemics

Análisis preliminar de la correlación de variables socioeconómicas con modelado geoespacial en la epidemia de dengue en Costa Rica

Cristina Soto-Rojas[1], Cesar Garita[2], Mariela Abdalah[3], Juan Gabriel Calvo[4], Fabio Sanchez[5], Esteban Meneses[6]

Soto-Rojas, C; Garita, C; Abdalah, M; Calvo, J. G; Sanchez, F; Meneses, E. Preliminary analysis of socioeconomic variable correlation with geospatial modeling in Costa Rica dengue epidemics. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on BioInspired Processing. Pag. 11-21.

https://doi.org/10.18845/tm.v37i7.7292

[1] National High Technology Center and Costa Rica Institute of Technology. Costa Rica
csoto@cenat.ac.cr
https://orcid.org/0000-0001-9180-1628

[2] Costa Rica Institute of Technology. Costa Rica
cesar@itcr.ac.cr
https://orcid.org/0000-0003-4592-3266

[3] National High Technology Center and Costa Rica Institute of Technology. Costa Rica
mabdalah@cenat.ac.cr
https://orcid.org/0000-0002-9790-2689

[4] University of Costa Rica. Costa Rica
juan.calvo@ucr.ac.cr
https://orcid.org/0000-0001-9948-9966

[5] University of Costa Rica. Costa Rica
fabio.sanchez@ucr.ac.cr
https://orcid.org/0000-0002-5552-3672

[6] National High Technology Center and Costa Rica Institute of Technology. Costa Rica
emeneses@cenat.ac.cr
https://orcid.org/0000-0002-4307-6000

Keywords

Geospatial modeling; data science; dengue epidemics.

Abstract

Dengue is a mosquito-transmitted disease that affects more than 5 million people worldwide. It is endemic in more than 100 countries and it has presence in 5 continents. Understanding the dynamics of dengue epidemics is crucial in reducing the massive public health impact this disease has. However, dengue is a complex phenomenon. There are many variables that contribute to the spread of the virus and the interconnection of those variables is not clear. We set out to explore the correlation of socioeconomic variables in dengue epidemics by using a geospatial model. Our study is centered in Costa Rica, a country with a repeated affectation by the virus. We found a possible relationship between number of dengue cases and some socioeconomic variables (dwellings with water pipes, location of work), which open the gates to consider including them in a more sophisticated epidemiological model.

Palabras clave

Modelado geoespacial; ciencia de datos; epidemia de dengue.

Resumen

El dengue es una enfermedad transmitida por mosquitos que afecta a más de 5 millones de personas en todo el mundo. Es endémica en más de 100 países y tiene presencia en los 5 continentes. Comprender la dinámica de las epidemias de dengue es crucial para reducir el impacto masivo que tiene esta enfermedad en la salud pública. Sin embargo, el dengue es un fenómeno complejo. Hay muchas variables que contribuyen a la propagación del virus y la interconexión de esas variables no está clara. Nos propusimos explorar la correlación de las variables socioeconómicas en las epidemias de dengue mediante el uso de un modelo geoespacial. Nuestro estudio se centra en Costa Rica, país con afectación reiterada por el virus. Encontramos una posible relación entre el número de casos de dengue y algunas variables socioeconómicas (viviendas con tubería de agua, ubicación del trabajo), que abren las puertas a considerar incluirlas en un modelo epidemiológico más sofisticado.

Introduction

According to the World Health Organization, the dengue disease is endemic in more than 100 countries [1]. The affected regions are Africa, the Americas, Eastern Mediterranean, South-East Asia, and Western Pacific; being the last two the ones hit the hardest. Dengue is a disease spread by the mosquito *Aedes Aegypti* and *Aedes Albopictus*. The transmission of the virus depends on two main elements: mosquitoes and humans. The dengue mosquitoes spread a virus that belongs to the *Flaviviridae* family, and four distinct serotypes can cause the disease with different variations of symptoms. That means there are four versions of the virus that causes dengue: DENV-1, DENV-2, DENV-3, and DENV-4. If a patient recovers from one of these serotypes, the patient may gain immunity to that specific serotype, but not to the others.

The dynamics of a dengue epidemic are complex. The available literature provides several approaches to model this disease. Most of those studies use climate variables, since they can influence the mosquito habitat conditions. As human activities also impact the spread of the virus, geospatial and socioeconomic variables have also been used in those models. To understand how dengue affects Costa Rica and its different regions, we embarked on a research project to explore available datasets from multiple sources and prevalent mathematical models. The final aim of the project is to shed light on variable correlation and model robustness for dengue epidemics in the country.

Background

Geographically weighted regression model

The simple linear regression is a classic model used to describe a variable of interest, known as the dependent variable, as a linear function of independent variables known as the predictor variables. This model has been used in many scenarios, delivering accurate results. However, if it considers the independence in the predictor variables, its effectiveness still has to be evaluated for geographical analysis.

When we consider the relation between the dependent variable and the predictor variables in multiple regions, it may happen that the relationship of variables changes from region to region. Moreover, in some regions the variable could have a different impact. If we consider a simple linear regression, differences from region to region could not be captured in the model that attempts to be fitted for all, and some valuable geographical analysis could be lost.

Geographically weighted regression (GWR) [16] considers each region and their neighborhoods as an individual local regression, and as an estimation of the coefficients. Then, the results of a regression for each region and the impact of the predictor variables for each case can be observed, providing a more segregated analysis that can bring more information than a simple linear regression. The formula for an GWR is:

$$y_i = a_{i0} + \sum_{k=1}^{N} a_{ik}\, x_{ik} + \epsilon_i$$

where y_i is the predicted variable for the region *i*, $\mathbf{a}_{i0}$ is the local intercept, $\mathbf{a}_{ik}$ is the coefficient of the variable **k** at the region **i**, x_{ik} is the value of the independent variable for the **k** variable at the region **i**, with **N** the number of regions, and ϵ_i the error term. One of these equations is calculated for each region and the coefficient calibration considers the neighborhoods of the region.

To determine the neighborhoods of each region there are two types of strategies that were used, the fixed one and the adaptive. The fixed one considers a fixed radius and includes the points inside this circle around the central region. In extreme cases, when the regions are small or large, the fixed approach may have problems with the estimations. The adaptive one considers a proportion of observations from the nearest neighbors and includes them; it adapts providing larger bandwidths to sparse data and smaller ones on the other case.

Related work

Hwa-Lung et al. [6] created a model that allows the generation of alerts within a week considering the spatio-temporal predictions of dengue fever cases. They used a stochastic Bayesian Maximum Entropy analysis and provide valuable spatial information of the dengue fever outbreaks. Åström et al. [7] proposes a model that considers the gross domestic product per capita (GPD), obtaining that both climatic variables and GDP influence the incidence of dengue. They predict the population at risk of dengue for 2050 under different combinations of the variables. One of their conclusions is that worsening global economic conditions would contribute to increase dengue incidence, especially on vulnerable urban populations.

Lowe et al. [8] showed that including climate information in a model for the case of Thailand improved the model for 79% of the provinces they modeled. They used a Bayesian framework considering spatial and temporal variables. They modeled the dengue relative risk for the next month of the data, and found that the climate variables of temperature and precipitation have a statistically significant contribution to the relative risk in the following month. Under this relationship, many models have aimed at predicting the behavior of the dengue disease considering these climatic variables. The variables of maximum temperature, humidity, and El Niño Southern Oscillation have been mainly considered, obtaining results that predict between 1 and 4 months of this phenomenon [9]-[11].

In the case of Costa Rica, we have an optimal climate for the mosquito. Vasquez et al. [12] proposed a predictive model using a generalized additive model and random trees that allow predicting the relative risk in 5 cantons of the country. Also, Sánchez and Calvo [13] proposed an epidemic model that allows exploring the transmission dynamics of this disease in the early life stage of mosquitoes and with an age structure in humans. This model allows a better analysis of the implications of this phenomenon by providing the age distribution of humans, and collaborates with prevention to learn more about the early life stages of the mosquito.

However, although multiple models have used climatic variables, Morin et al. [14] highlighted the complexity of the relationship between climatic variables and the factors that affect dengue transmission. That situation could explain some inconsistencies in associations between climatic variables and dengue, since ecological aspects, the development of the virus and host-species interactions are usually ignored.

Delmelle et al. [15] created a GWR model that considers socioeconomic and environmental variables and found that the main influencing variables are population density, socioeconomic status, proximity to tire shops and plant nurseries, and the presence of a sewage system. Naqvi et al. [17] used a GWR model and show that the temperature is the most significantly associated variable with the dengue fever in Pakistan.

Methods

Data Sources

Dengue Cases

Incidence dengue cases were obtained from the Ministry of Health of Costa Rica, and are found at the cantonal and regional level with the following characteristics:

- Data period: 2012-2013, 2015-2019.
- Data spatial segregation: socioeconomic region (6) and canton (84).
- Available variables: Number of cases for each year by epidemiological week.
- Some regions have 0 cases in some of the years. The data is complete.

Socio-economic

Information was extracted from the National Institute of Statistics and Censuses of Costa Rica, from which the following data was extracted from the National Census:

- Data period: 2011.
- Data spatial segregation: socioeconomic region and canton.
- Available variables: telephony, access to services, household distribution, education, housing characteristics, multidimensional poverty, Gini coefficient, household income, garbage disposal system, water system access, work and mobility.
- These files contain the value for each variable per house and per person. Some present the percentage, and some the total. The data is complete.

Spatial

Various geospatial layers were obtained from the National Territorial Information System of Costa Rica. They have the following characteristics:

- Data format: geospatial object from Web Feature Service.Data spatial type: polygons.
- Available layers: cantonal distribution.
- The data is complete.

Data Wrangling

Considering the irregular behavior of the incidence of dengue, we proceeded to group all the cases to have a greater representation in each canton of the incidence. On the other hand, the socioeconomic variables used are those of the 2011 census. These data were the last available information about the state of the cantons, since the centroids of the regions on the country are closer on the central area of the country and more distant on the external area, then the data were divided on being part of the Great Metropolitan Area (GMA) or not. The GMA delimits the central regions of the country and contains the closer centroids.

Model

For the model, the variable of accumulated cases by canton was used as a dependent variable. Multiple socioeconomic variables were taken into consideration as the independent variable, but those that presented a higher correlation with the data and better results in the model are shown. Results were obtained with the variables: access to water pipes, percentage of people who work in the same canton, percentage of people who work in another canton and percentage of people who are outside the labor force. Variables that were considered but did not show significant results were: deprived population, Gini index, population working in the primary, secondary and tertiary sectors, percentage of households in slum conditions.

Considering the GWR description, y_i is the dengue cases, and x_i is the different variables consider, finally the regions consider are the cantons. A GWR model was created for each variable. To evaluate the results of the model and determine if there was a relevant contribution of each variable, the results of R2 for each model were compared. The statistical significance was evaluated for each model:

$$\mathrm{t} = \mathrm{coefficient}/\mathrm{coefficient\_se}$$

A model was created for each variable for each case: GMA and non-GMA.

Experimental Setup

The computer used has an operating system Windows version. The code was implemented on R, version 4.2.0. The plotting library is tmap. The code used to run the model an generate the visualizations is available at the repository: https://gitlab.com/CNCA CeNAT/gwr-dengue.

Table 1. Software configuration.

Program	Version
Operating System (OS)	Windows 10 PRO
Processor	AMD Ryzen 7 2700 Eight- Core Processor 3.20 GHz
RAM	16gb

Results

Working in the same canton

This variable represents the percentage of the population in each canton that works in the same canton. The local R2 shows that the model has a modest adjustment, with values between 0.1 and 0.5 in the GMA case, and the coefficients associated show a positive correlation, especially in the upper left region (see Figure 1). In Figure 2 we observe that the model in the case of cantons outside the GMA does not show a good fit. Then it is possible to observe that the GMA seems to positively correlate people that work in the same canton with the cases.

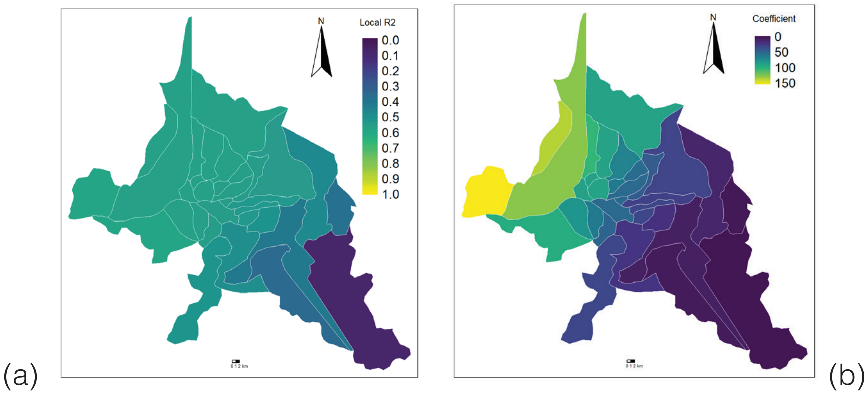


Figure 1. Results for population working on the same canton on the GMA (a) Local R2 (b) Coefficients.

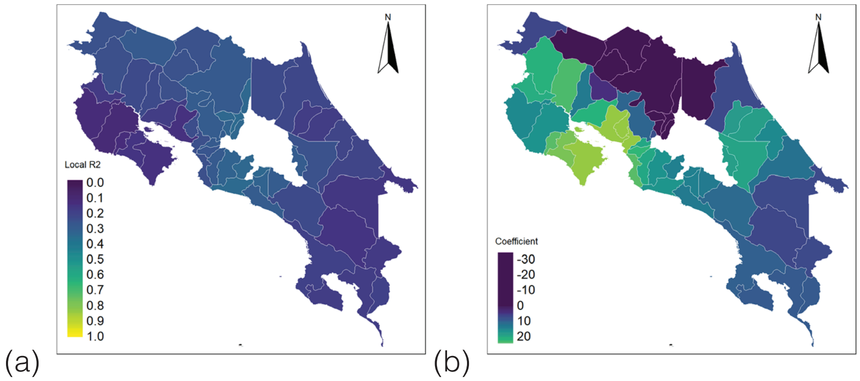


Figure 2. Results for population working on the same canton outside the GMA (a) Local R2 (b) Coefficients.

Working in another canton

This variable represents the percentage of the population that works in a different canton. The local R2 shows that the model has a modest adjustment, with values between 0.1 and 0.5 in the GMA case, and the coefficients associated show a negative correlation, especially in the upper left region (see Figure 3). It is important to note that in this case, the values of the coefficients are negative, so the model would be telling us that there may be a possible relationship between people leaving their canton to work and fewer cases of dengue.

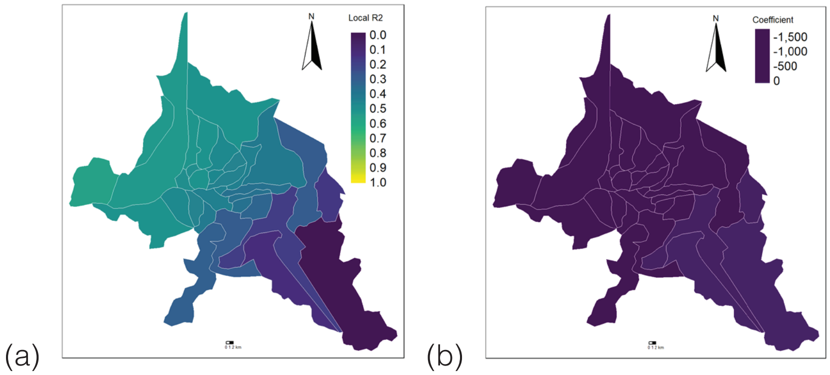


Figure 3. Results for population working in another canton on the GMA (a) Local R2 (b) Coefficients.

In the case of cantons outside the GMA, on Figure 4 we can see that, as in the case of the previous variable, robust results were not obtained.

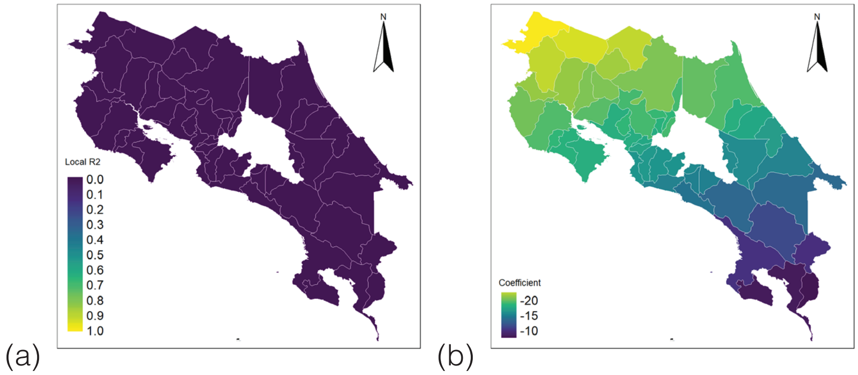
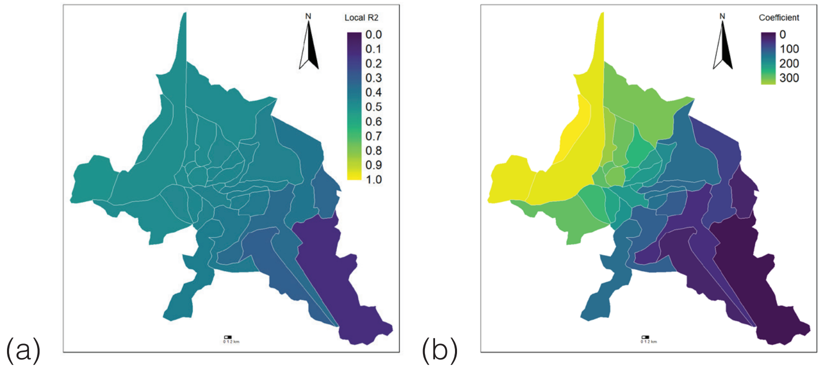


Figure 4. Results for population working in another canton outside the GMA (a) Local R2 (b) Coefficients.

Unemployed population

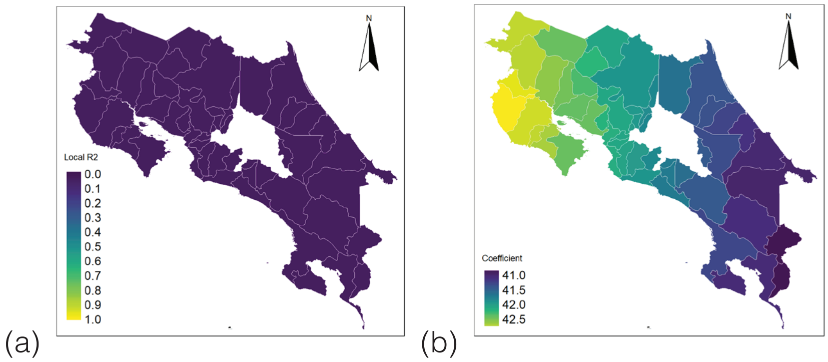
This variable represents the percentage of the population that is unemployed. In this case, we can see that the results follow a behavior similar to that of the two previous cases. The local R2 shows that the model has a modest adjustment, with values between 0.2 and 0.5 in the GMA case, and the coefficients associated show a positive correlation, especially in the upper left region (see Figure 5). In the GMA area, it allows us to analyze that, as with the variable of people who work in the same canton, we have positives coefficients. We could then observe a possible relationship between the incidence of dengue cases and the population that stays mainly in the same region, such as people outside the labor force together with those who work in the same canton.



(a) (b)

Figure 5. Results for unemployed population on the GMA (a) Local R2 (b) Coefficients.

In this case we can also observe in the Figure 6 that outside the GMA a good fit of the model was not obtained, similar to that with the other two variables.

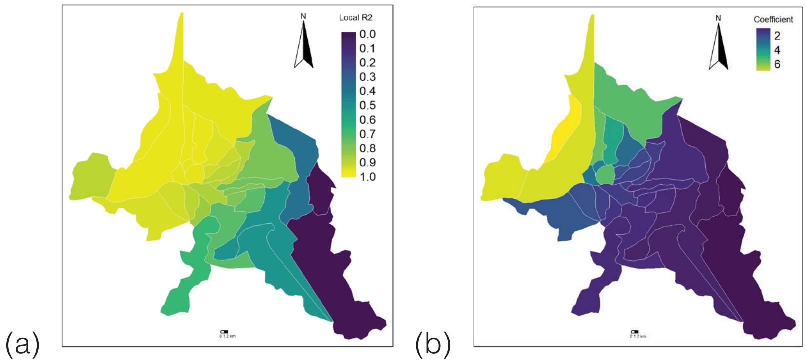


(a) (b)

Figure 6. Results for unemployed population outside the GMA (a) Local R2 (b) Coefficients.

Water pipes

This variable represents the dwellings that do not have access to pipes. The local R2 shows that the model has a strong adjustment, with values between 0.1 and 0.9 in the GMA case, and the coefficients associated show a positive correlation, especially in the upper left region (see Figure 7). For the case of the area outside the GMA, Figure 8 shows a good fit for the model. The local R2 shows that the model has a strong adjustment, with values between 0.4 and 0.9, with coefficients that are also positive. This can be interpreted as a possible relationship between access to pipes in the house and the incidence of dengue cases.



(a) (b)

Figure 7. Results for population without water pipes at households on the GMA (a) Local R2 (b) Coefficients.

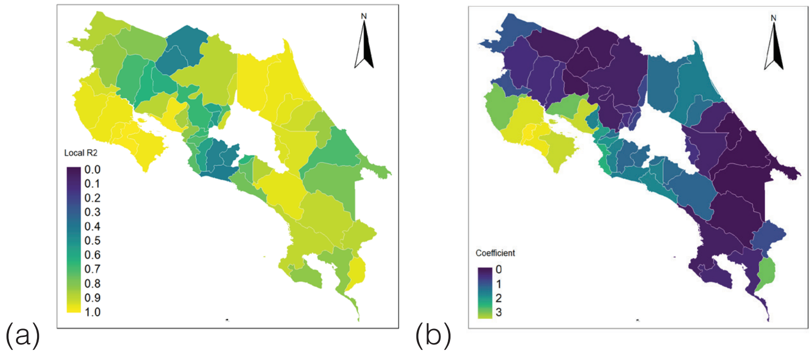


Figure 8. Results for population without water pipes at household outside the GMA (a) Local R2 (b) Coefficients.

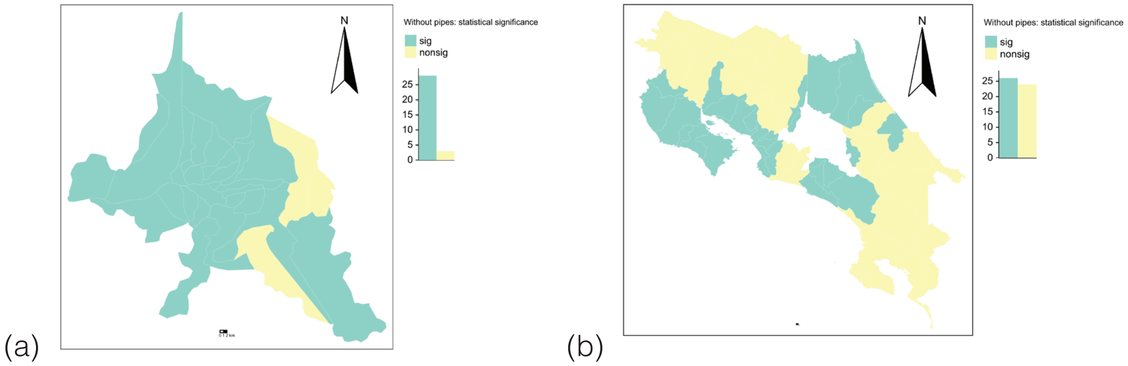

(a)
N
Without pipes: statistical significance
sig
nonsig
25
20
15
10
5
0
(b)
N
Without pipes: statistical significance
sig
nonsig
25
20
15
10
5
0

Figure 9. Statistical significance for population without water pipes at household (a) GAM (b) outside the GAM.

The statistical significance is evaluated for the results of the model. Figure 9 shows the case of the population without access to clean water, in the GAM region most of the regions show significance. Outside the GAM is almost the same significant and non-significant regions. The other variables can be found on the repository.

Summary

The results show that there are two possible relationships between socioeconomic variables and dengue cases that are worth exploring. The first is the employee mobility: there seems to be a positive correlation between minimum mobility (outside the labor force) and within the canton, while if they move outside the canton there seems to be a negative correlation. The behavior of the dengue mosquito is characterized by being a mosquito with little mobility, its life remains in the area near its breeding site. On the other hand, their feeding schedule is daytime. Considering these factors, it is relevant to analyze the mobility of people associated with their work, since work hours fit with the vector's feeding schedule.

The second is with the variable of dwellings with no access to water pipes, this variable presented a better fit in the model and shows a positive correlation with cases of dengue. We remark that there are more variables that influence this phenomenon of dengue, so these results cannot be interpreted as a direct relationship between these variables and cases of dengue, but they serve as a guide to identify the variables that can provide information that can be used in a more complex epidemiological model.

Conclusions and future work

The GWR model allows a geospatial analysis to evaluate if the variable that is being considered really has a possible relationship with the dependent variable, which helps to determine possible patterns in regions and a better understanding of the relationship between the variables related to dengue. It is an easy-to-implement model that allows a simple exploration of variables, in order to eventually determine whether or not to incorporate them into a more complex model.

The idea is to consider these variables to create an index that summarizes this socioeconomic information. This index is planned to be included as a parameter in an epidemiological model proposed by Sánchez and Calvo [13].

Access to data is usually challenging, not all variables are available for all the time ranges. Therefore, it would be interesting to have more detailed data for each canton updated by year, as well as data on the mobility of the different populations in the region, beyond the associated with the work.

In future iterations, the centroid of the regions could be better adjusted so that it is located in the most populated points of each canton, since it is currently only the centroid of the region, but the most populated area of the canton is not always located in the center.

Referencias

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] World Health Organization, UNICEF, et al. Operational guide using the web-based dashboard: Early warning and response system (ewars) fordengue outbreaks. 2020.

[3] Ministerio de Salud. Situación epidemiológica dengue, chikungunya y zika del ministerio de salud. data retrieved from: https://www.ministeriodesalud.go.cr/index. php/vigilancia-de-la-salud/analisis-de-situacion-de-salud. 2021

[4] Barrera R. Focks D. Dengue transmission dynamics: assessment and implications for control. In Report of the scientific working group meeting on dengue, 1-5, pages 92–108. WHO, October 2006.

[5] Rosen L. Rodhain F. Mosquito vectors and dengue virus-vector relation-ships. Dengue and dengue hemorrhagic fever, pages 45–60, 1997.

[6] Hwa-Lung Yu, Shang-Jen Yang, Hsin-Ju Yen, and George Christakos. A spatio-temporal climate-based model of early dengue fever warning in southern taiwan. Stochastic Environ- mental Research and Risk Assessment, 25(4):485–494, 2011.

[7] Christofer Åström, Joacim Rocklöv, Simon Hales, Andreas Béguin, Valerie Louis, and Rainer Sauerborn. Potential distribution of dengue fever under scenarios of climate change and economic development. Ecohealth, 9(4):448–454, 2012.

[8] Rachel Lowe, Bernard Cazelles, Richard Paul, and Xavier Rod ́o. Quantifying the added value of climate information in a spatio-temporal dengue model. Stochastic Environmental Research and Risk Assessment, 30(8):2067–2078, 2016.

[9] Elodie Descloux, Morgan Mangeas, Christophe Eugène Menkes, Matthieu Lengaigne, Anne Leroy, Temaui Tehei, Laurent Guillaumot, Magali Teurlai, Ann-Claire Gourinat, Justus Benzler, et al. Climate-based models for understanding and forecasting dengue epidemics. PLoS neglected tropical diseases, 6(2):e1470, 2012.

[10] Vivek Jason Jayaraj, Richard Avoi, Navindran Gopalakrishnan, Dhesi Baha Raja, and Yusri Umasa. Developing a dengue prediction model based on climate in tawau, malaysia. Acta tropica, 197:105055, 2019.

[11] M Hurtado-Díaz, H Riojas-Rodríguez, SJ Rothenberg, H Gomez-Dantés, and E Cifuentes. Impact of climate variability on the incidence of dengue in Mexico. Tropical medicine & international health, 12(11):1327–1337, 2007.

[12] Paola Vásquez, Antonio Loría, Fabio Sanchez, and Luis Alberto Barboza. Climate-driven statistical models as efective predictions of local dengue indicence in costa rica: a generalized additive model and random forest approach. Revista de Matemática: Teoría Y Aplicaciones, 27(1):1–21, 2020.

[13] Fabio Sanchez and Juan G Calvo. Dengue model with early-life stage of vectors and age- structure within host. Revista de Matemática: Teoría y Aplicaciones, 27(1):157–177, 2020.

[14] Cory W Morin, Andrew C Comrie, and Kacey Ernst. Climate and dengue transmission: evidence and implications. Environmental health perspectives, 121(11-12):1264–1272, 2013.

[15] Eric Delmelle, Michael Hagenlocher, Stefan Kienberger, and Irene Casas. A spatial model of socioeconomic and environmental determinants of dengue fever in cali, colombia. Acta tropica, 164:169–176, 2016.

[16] Chris Brunsdon, A. Stewart Fotheringham and Martin E. Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical analysis 28.4 (1996): 281-298.

[17] Naqvi, Syed Ali Asad, et al. ”Integrating Spatial Modelling and Space–Time Pattern Mining Analytics for Vector Disease-Related Health Perspectives: A Case of Dengue Fever in Pakistan.” International journal of environmental research and public health 18.22 (2021).