Geostatistical spatial modelling of dengue fever across São Paulo state, Brazil

Vanessa da Silva BRUM-BASTOS[1], Thiago Salomão de AZEVEDO[2], Maria Anice Mureb SALLUM[3] , Urška DEMŠAR[4]

[1] University of St Andrews, St Andrews, United Kingdom, vdsbb@st-andrews.ac.uk
[2] University of São Paulo, São Paulo, Brazil, azevedots@usp.br
[3] University of São Paulo, São Paulo, Brazil, masallum@usp.br
[4] University of St Andrews, St Andrews, United Kingdom, urska.demsar@st-andrews.ac.uk

Abstract: This paper explores geographic variability in relationships between the average dengue incidence at São Paulo state municipalities for 2000 to 2014. By linking dengue incidence to socioeconomic, environmental and climatological data, is possible to develop a more detailed picture of dengue incidence and creates a model for future prediction. Analysis is approached through spatial analysis using geographically weighted regression (GWR), which enables the investigation of local variations in incidence patterns. The results demonstrate that the variables that are traditionally assumed to affect dengue incidence, do not do so uniformly over space. Our findings present a starting point for a more detailed investigation as to why this heterogeneity exists, how each variable affects dengue incidence and a step further towards prediction models.

Keywords: dengue fever, geographic weighted regression, spatial modelling

1. Introduction

Dengue fever is an acute disease caused by four virus serotypes within the genus Flavivirus Forattini (2002). The World Health Organization (WHO) estimates that 390 million dengue infections occur worldwide each year and recently the disease has been spreading around Latin America and Caribbean; in these areas dengue epidemics have a 3-5 years cycle WHO (2009).

Brazil is between the thirty most endemic countries out of the one hundred that show the disease WHO (2014). There is a strong association between environmental conditions and mosquito-borne diseases Ebi et *al.* (2005); Small *et al.* (2003); Rogers *et al.* (2000). The unplanned urbanization, unappropriated sanitation, inefficient control of mosquitoes, population increase and the growing people flow underlie the spreading of dengue epidemic.

The increasing urbanization, started in 1950, promoted dramatic changes in São Paulo state, the unbridled growing of cities and the following environmental changes turned urban environments each time less and less sustainable Haughter *et al.* (1994). The peripherization of cities along with more than 75% of population living in urban areas exposes population to a high level of environmental insalubrity; which combined with the absence of planning has been increasing population exposure to epidemics in urban ecosystems.

Urban ecosystems are compounded by three intimate related ecological spheres: biotic, abiotic and anthropogenic. The abiotic sphere is related to the physical components of the urban site, for example, soil, topography, climate and hydrology. The biotic sphere is related to the "living mass" of the urban environment, for example, mosquitos, pigeons and humans. The anthropogenic sphere is related to all the man-made elements, for example, population density and sanitation conditions. These three spheres affect dengue epidemic distributions and are considered in our model through three groups of variables: environmental, climatological and socioeconomic.

Aedes aegypt life cycle, the dengue fever vector, is affected by the three different groups, but mainly by changes in temperature and precipitation Morin et al. (2013); Azevedo et al. (2012). Considering the

climate changes Earth may face in the upcoming decades, modelling the relationships between environmental, climatological and socioeconomic factors and dengue fever cases is an important topic in geographical health research. Such models can lead to better predictability of dengue fever occurrence for each climate future scenario; however the first step is to understand what these relationships currently are. For this purpose we used Geographically Weighted Poisson Regression (GWPR) Fotheringham et al. (2002) to investigate dengue fever incidence in São Paulo state – Brazil in the 2000 to 2014 period (Figure 1).

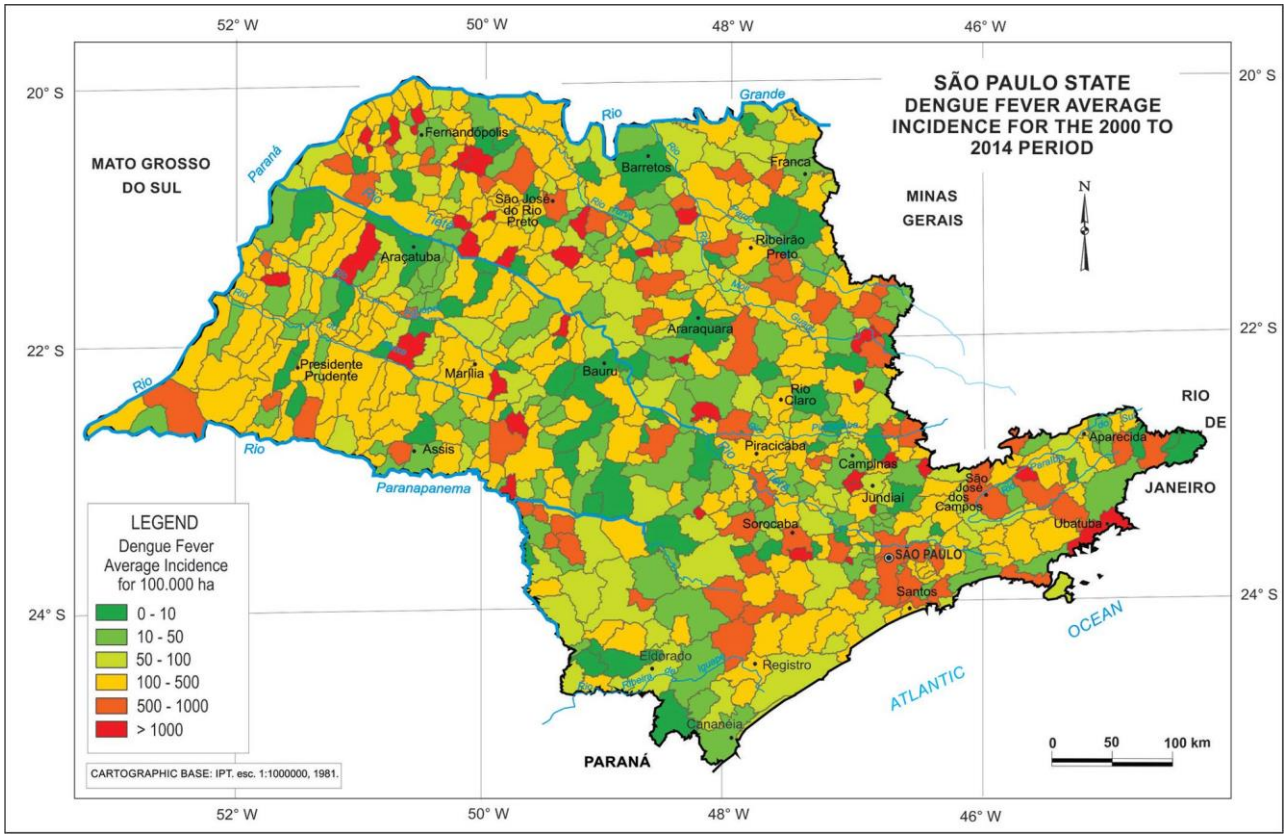

Figure 1 – Average dengue fever incidence for São Paulo state municipalities for the 2000 to 2014 period

2. Data

In this study we used four data sources: the report on autochthonous dengue cases between 2000 and 2014 in São Paulo state, socioeconomic data from SNIS (National Information System on Sanitation), boundary data from IBGE (Brazilian Institute of Geography and Statistics), satellite derived environmental and climatological data from AMBDATA (Environmental Variables for Species Distribution Model).

Dengue cases data were produced by CVE (Center of Epidemiological Surveillance from São Paulo State), the authority responsible for monitoring and controlling of diseases in São Paulo state. This data set contains information on the 645 municipalities and its respective number of autochthonous dengue cases for each year between 2000 and 2014.

We obtained a set of socioeconomic variables from the SNIS (2016) at municipality level for 330 municipalities, a set of rasters of environmental and climatological variables from the AMBDATA (2016) with 1 km of spatial resolution. The average for each environmental and climatological variable was calculated for each municipal boundary. We selected a subset of variables that were highlighted by previous studies as being associated with Aedes aegypt reproduction. We provide the list of selected variables and the reasoning behind their selection in the next section, where we discuss how we built our model.

All of the variables (dengue cases, socioeconomic, environmental and climatological information) were joined to a final spatial data set of 645 São Paulo municipalities, the boundaries and centroids of which we

obtained from the IBGE (Brazilian Institute of Geography and Statistics) through the FTP directory IBGE (2012). However, due to the socioeconomic data limitation we worked on 330 municipalities, the ones for which we had information on the three group of variables.

3. Methodology

To explain relationships between the dengue cases and the explanatory variables we adopted an established spatial statistical data analysis methodology Fotheringham *et al.* (2002), which consists of the following steps: 1) literature-based variable selection, 2) data-driven model optimisation and 3) a calibration and interpretation of the best possible global and local models. In step 1), we selected forty three potential explanatory variables based on literature and in step 2) reduced their number to twenty one using correlation analysis and model quality optimisation. These twenty three variables were then used as input into a global and a GWR model in step 3). This section describes the details of our modelling procedure.

3.1 Step 1: Literature-based variable selection

The initial selection of potential explanatory variables was performed based on literature review. We selected forty two variables that were deemed most likely to explain dengue fever incidence (Table 1). The forty two variables were categorised into 3 thematic groups: climate, environmental and socioeconomic related. The following list provides our reasoning behind inclusion of each of these variables per thematic group.

Table 1 – The forty three independent variables originally considered for inclusion in regression modelling. The eighteen variables found to be correlated with the highest number of other variables (shown on crossed fields in the table) were excluded from GWR modelling. Further, the variables marked with [b] were excluded during the model quality optimisation, leaving a set of twenty one(shown in bold) to be included in the final model.

Group	Variable	Name	Highly correlated with [a]
Environmental	Altitude	~~ALTIT~~	BIO1, BIO10, BIO5, BIO6, BIO8, BIO9, BIO11
	Slope	**SLP**	AVTME, HND100, HND50, PTCOV
	Height above the nearest drainage (100 m)	HND100 [b]	AVTMA, AVTME, SLP, HND50, HND500
	Height above the nearest drainage (50 m)	**HND50**	AVTMA, AVTME, SLP, HND100
	Height above the nearest drainage (500 m)	HND500 [b]	AVTMA, AVTME, HND100
	Percent tree coverage	**PTCOV**	BIO3, BIO12, SLP
	Drainage density	**DRADEN**	-
Climatological	Average precipitation	**AVPRE**	BIO4
	Annual average maximum temperature	AVTMA [b]	AVTME, HND50, HND100, HND500
	Annual average mean temperature	**AVTME**	AVTMA, SLP, HND50, HND100, HND500
	Annual average minimum temperature	**AVTMI**	BIO12, BIO18,
	Annual mean temperature	~~BIO1~~	ALTIT, BIO5, BIO6, BIO8, BIO9, BIO10, BIO11,
	Mean temperature of warmest quarter	~~BIO10~~	ALTIT, BIO1, BIO5, BIO6, BIO8, BIO9, BIO11
	Mean temperature of coldest quarter	~~BIO11~~	ALTIT, BIO1, BIO10, BIO5, BIO6, BIO8, BIO9
	Annual Precipitation	~~BIO12~~	AVTMI, BIO2, BIO7, BIO12, BIO13, BIO14, BIO16, BIO17, BIO18, PTCOV

	Precipitation of wettest month	BIO13	BIO7, BIO12, BIO16, BIO18
	Precipitation of driest month	~~BIO14~~	BIO2, BIO3, BIO12, BIO15, BIO17, BIO19
	Precipitation seasonality	**BIO15**	BIO3, BIO4, BIO14, BIO17, BIO19
	Precipitation of wettest quarter	BIO16 [b]	BIO7, BIO12, BIO13, BIO18
	Precipitation of driest quarter	**BIO17**	BIO3, BIO12, BIO14, BIO19
	Precipitation of warmest quarter	**BIO18**	AVTMI, BIO12, BIO13, BIO16,
	Precipitation of coldest quarter	BIO19 [b]	BIO3, BIO14, BIO15, BIO17
	Mean diurnal temperature range	**BIO2**	BIO3, BIO7, BIO12, BIO14,
	Isothermality	~~BIO3~~	BIO2, BIO14, BIO15, BIO17, BIO19, PTCOV
	Temperature seasonality	BIO4[b]	AVPRE, BIO15
	Maximum temperature of warmest month	~~BIO5~~	ALTIT, BIO1, BIO6, BIO8, BIO9, BIO10, BIO11,
	Minimum temperature of coldest month	~~BIO6~~	ALTIT, BIO1, BIO5, BIO6, BIO8, BIO9, BIO10, BIO11
	Temperature annual range	**BIO7**	BIO2, BIO12, BIO13, BIO16
	Mean temperature of wettest quarter	~~BIO8~~	ALTIT, BIO1, BIO5, BIO6, BIO8, BIO9, BIO10, BIO11
	Mean temperature of driest quarter	~~BIO9~~	ALTIT, BIO1, BIO5, BIO6, BIO8, BIO10, BIO11
Socio - economic	Urban population with direct selective waste collection (inhabitants)	~~CO165~~	AG005, AG026, CO165, CS050, ES004, ES026, G06B
	Urban population with direct selective waste collection provided by the local authority (inhabitants)	~~CS050~~	AG005, AG026, CO165, CS050, ES004, ES026, G06B
	Sewage network extension (km)	~~ES004~~	AG005, AG026, CO165, CS050, ES026, G06B
	Urban population with sanitary sewage (inhabitants)	~~ES026~~	AG005, AG026, CO165, CS050, ES004, G06B
	Resident urban population with sanitary sewage (inhabitants)	~~G06B~~	AG005, AG026, CO165, CS050, ES004, ES026
	Sewage treatment (%)	**IN016_AE**	-
	Waste collection per urban population (%)	**IN016_RS**	-
	Urban water supply services index (%)	**IN023_AE**	-
	Sewage treatment index (%)	**IN024_AE**	IN056_AE
	Total water supply services index (%)	**IN055_AE**	IN056_AE
	Total sewage treatment index (%)	**IN056_AE**	IN024_AE, IN055_AE
	Water network extension (km)	~~AG005~~	G06B, AG026, ES004, ES026, CO165, CS050
	Urban population with water supply services (inhabitants)	~~AG026~~	G06B, AG005, ES004, ES026, CO165, CS050

a High correlation means that the absolute value of the correlation coefficient between the two variables is more than 0.70

3.1.1 Group 1: climatological variables

Temperature and precipitation are key variables for dengue disease spread, highly frequent precipitation and a temperature range between 10 °C and 40°C, with 28°C being the optimum point, (Yang et al., 2009) compound the ideal environmental conditions for Aedes Aegypt reproduction and dispersion. Dengue epidemics usually happen on the warmest months of the year Watts *et al.* (1987), Consoli *et al.*(1994), Câmara *et al.* (2009), on which the temperature is favourable to the mosquito development and reproduction Johansson (2009). However it is on the rainy season that Aedes aegypt population reaches its peak and becomes a treat for human health Consoli *et al.* (1994).

3.1.2 Group 2: environmental variables

These variables are related to the land use, land cover and its environmental impacts. The inappropriate land use and land cover in urban areas produce rough surfaces, which are perfect habitat for Aedes aegypt reproduction and adaptation Azevedo et al (2012). Besides, urban centres are a synanthropic environment, which turn Aedes aegypt habits into much more domestic, accentuating its anthropophily Tauil (2001).

3.1.3 Group 3: socioeconomic variables

The socioeconomic variables are linked to the historical context in an emergent country like Brazil, mainly the demographic changes from the 1960's on. The migration to urban centres was responsible for an unexpected population growth, which resulted in insufficient habitation and poor sanitary conditions. Basic sanitation, mainly water supply services and garbage collection, is still inadequate in some urban centres; this lack of infrastructure increases the number of potential breeding places for Aedes aegypt mosquito (Tauil, 2001).

3.2 Step 2: Data-driven model optimisation

In this step, the set of forty three initial variables was reduced, first by removing correlated variables and second by optimising model quality. We first tested the forty three variables for collinearity by calculating a matrix of pairwise correlation coefficients. Eighteen variables that were highly correlated, i.e. pairwise coefficient > 0.07 Dormann et al. (2013), with the highest number of other variables (Table 1) were excluded from further consideration. This left us with twenty five variables to be used as input into model optimisation.

In the second part, we optimised the quality of the model by minimising the corrected Akaike Information Criterion (AICc), which is a measure of the quality of a statistical model Konishi and Kitagawa (2008). It can be used to compare two models calibrated on the same data set, the lower the AICc, the better the model quality Fotheringham et al. (2012). To find the best model, each of the twenty five variables that passed the collinearity test was, in turn, used to calibrate a one-variable GWPR model for the dengue fever (i.e. the dependent variable was average dengue incidence for the period and the only independent variable was one of the twenty five variables). For each of these twenty five one-variable models we calculated the AICc value. The variable with the smallest AICc value, indicating the best model quality, was kept. In the next step, the remaining twenty four variables not selected for the initial model were entered in turn, so that twenty four different two-variable models were produced. Again the models were calibrated and the variable resulting in the smallest AICc was retained as second variable in the final model. This operation was repeated until the AICc did not decrease anymore with the inclusion of an additional variable, indicating that the remaining variables did not improve the model. In our case, this occurred after twenty two iterations and excluded five variables marked with [b] in Table 1. Fig. 2 illustrates the model optimisation procedure.

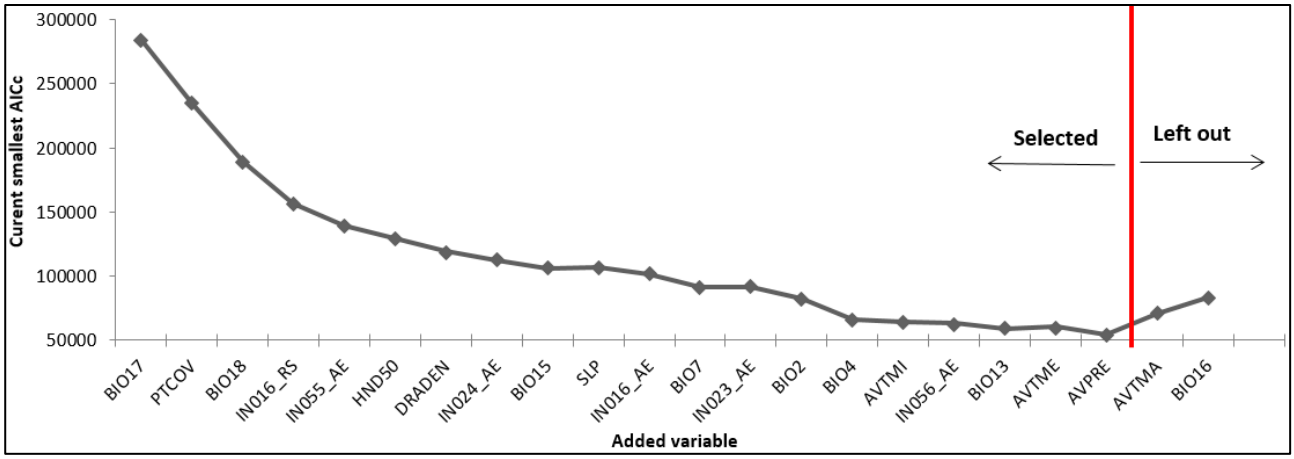

Figure 2. Model quality optimisation stops when there is no more decrease in AICc; it happened when testing 22-variable models. We therefore select the first 21 variables for the final model.

The final selection of the potential explanatory variables to be used for local modelling therefore includes the following (shown in bold in Table 1): BIO17, PTCOV, BIO18, IN016_RS, IN055_AE, HND50, DRADEN, IN024_AE, BIO15, SLP, IN016_AE, BIO7, IN023_AE, BIO2, BIO4, AVTMI, INO56_AE, BIO13, AVTME and AVPRE.

3.3. Step 3: local modelling

Multiple linear regression assumes that the relationships between variables are stationary in space, which contradicts one of the major assumptions in geography: spatial phenomena vary across space. GWR addresses this variability by producing a set of local models instead of with separate parameters estimated for each location in the data set (Fotheringham et al., 2002). The data points are weighted according to their distance from the regression point, so that points close to the regression point are more heavily weighted than points further away (Mansley and Demšar, 2015). GWR calibration results in a surface of local parameter estimates that can be mapped to investigate significant spatial variations in the relationships between the explanatory variables and dengue incidence (for more details on GWR see Mansley and Demšar, 2015 and Fotheringham et al., 2002).

4. Results

Results of the global regression model are presented in Table 2, showing global coefficients and their significance. The adjusted R^2 for the global model is 0.49. The results of global regression suggest a positive relationship between the dengue incidence and the following variables: IN016_RS, AVTMI, BIO4, IN056_AE, AVTME, BIO13, BIO18, IN055_AE, IN016_AE and BIO7. The rest of the relationships are negative in a global model.

The GWR model was calibrated for the twenty one variables as the global regression model. We ran a Poisson model with an adaptive bisquare kernel, which resulted in the optimal bandwidth of 99 nearest neighbours. This means that a local model for each of the 321 municipalities was generated using the weighted data from the nearest 99 municipalities. The local AICc value was 53419.91, an improvement of 14388.28 compared to the global model. Figure 3 shows that the local model replicates variations in turnout very well in SE and N, but performs least well in the NW. The average local adjusted R^2 for GWR model is 0.75 (averaging values from 0.51 to 0.98, Figure 3), also indicating improvement from the 0.49 adjusted R^2 of the global model.

Table 2 – Results of the global regression and summary statistics for GWR parameter estimates. All variables are statistically significant to 0.01 level in global regression

	Global regression		GWR parameter estimates			
Variable	Estimate	p-value	Mean	StDev	Min	Max
BIO15	-0.14	-104.35	-0.08	0.36	-0.75	1.21
HND50	-0.01	-102.16	0.01	0.05	-0.07	0.14
BIO2	-0.15	-83.24	-0.06	0.51	-2.13	1.05
PTCOV	-0.08	-79.29	-0.06	0.37	-0.80	0.69
DRADEN	-0.04	-60.83	-0.02	0.09	-0.24	0.17
IN024_AE	-0.03	-31.62	-0.09	0.14	-0.59	0.09
BIO17	-0.01	-27.28	-0.02	0.08	-0.23	0.26
AVPRE	-0.28	-26.09	-0.21	3.88	-7.08	8.61
IN023_AE	-0.03	-19.02	0.05	0.16	-0.39	0.57
DECLIVIDADE	-0.01	-5.25	-0.27	0.41	-1.77	0.60
IN016_RS	0.00	-3.47	0.05	0.12	-0.34	0.38
AVTMI	0.02	9.81	-0.09	0.62	-1.38	1.35
BIO4	0.01	16.29	0.01	0.30	-0.67	0.51
IN056_AE	0.03	25.96	0.08	0.08	-0.05	0.38
AVTME	0.24	26.28	0.15	3.14	-6.95	5.85
BIO13	0.02	45.71	0.02	0.07	-0.10	0.29
BIO18	0.01	63.94	0.02	0.04	-0.14	0.13
IN055_AE	0.09	80.78	0.01	0.06	-0.27	0.14
IN016_AE	0.01	112.24	0.00	0.01	-0.02	0.03
BIO7	0.29	157.98	0.22	0.56	-1.65	2.53

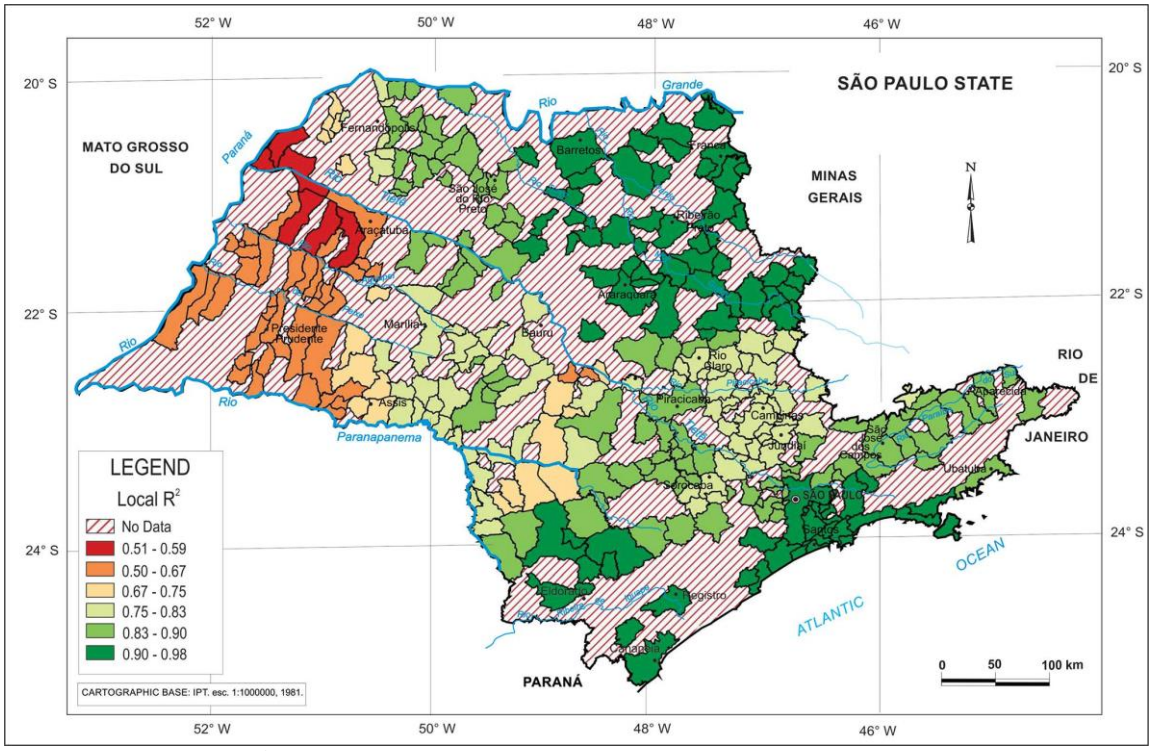

Figure 3 - Local R^2 for the GWR model. Model performs best in southeast and north areas of the São Paulo state and worst in the northwest central areas. No data values are linked to the states where socioeconomic information was not available

5. Conclusions

In this paper we conducted a spatial analysis of the relationships between environmental, socioeconomic and climatological variables and average dengue incidence in São Paulo state municipalities. Using a well-known local statistical methodology we built a local model that is in certain areas is able to explain up to 98% of the variation in the dengue incidence, a better performance than the 49% given by the global model. The model performs less well in the NW, indicating the difference in dengue incidence patterns between that particular region in the state and the others. The increase in the explained variance from the global to the local model suggests that there is substantial spatial variation in the characteristics that affect dengue incidence. This result provides evidence to the idea that dengue incidence is not only related to environmental and climatological conditions, but also to socioeconomic conditions. Our further research will detail the spatial variation on the relationships between the explanatory variables and dengue incidence, as well as use these results to predict dengue incidence for hypothetical socioeconomic and climate changes scenarios.

Acknowledgments

The authors acknowledge SWB (Science Without Borders) - CAPES (Coordination for the Improvement of Higher Education Personnel) for the financial support during the development of this study, through first author's PhD scholarship.

References

AMBDATA: Variáveis ambientais para Modelos de Distribuição de Espécies (SDMs). Eletronic database. São José dos Campos: INPE. Available at: http://www.dpi.inpe.br/Ambdata/index.php. Accessed on June 2016.

Azevedo, T. S.; Tavares, A. C.; Silva B. B. V.; Piovezan, R. ;Von Zuben, C. J.; André, N. I. (2012). Ilhas de calor e Aedes aegypti: um estudo preliminar para a cidade de Santa Bárbara d'Oeste, SP – Bra, utilizando sensoriamento remoto. I Congreso Latino americano de Ecología Urbana. Los Pòvorines. Buenos Aires Argentina.

Câmara, F. P.; Gomes, A. F.; Santos, G. T.; Câmara, D. C. P. (2009).Clima e epidemias de dengue no Estado do Rio de Janeiro. Revista da Sociedade Brasileira de Medicina Tropical 42(2):137-14.

Christofoletti, A. (1999). Modelagem de Sistemas Ambientais. São Paulo: Blucher. 256 p.

Consoli, R., Oliveira, R. L. (1994). Principais mosquitos de importância sanitária no Brasil. FIOCRUZ. 228p

Douglas, I. (1993). The Urban Environment. Baltmore: Edwart Arnold. 229 p.

Mansley, E. ; Demšar, U. (2015). Space matters: Geographic variability of electoral turnout determinants in the 2012 London mayoral election, Elect. Stud., vol. 40, pp. 322–334.

Forattini, O. P. (2002). Culicidologia médica. São Paulo: Edusp. v. 2. 860 p.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2002). Geographically weighted regression: the analysis of spatially varying relationships, Chichester: Wiley.

Yang, H. M. ; Macoris, M. L. G. ; Galvani, K. C. ; Andrighetti, M. T. M. ; Wanderley, D. M. V. (2009).Assessing the effects of temperature on the population of Aedes aegypti, the vector of dengue., Epidemiol. Infect., vol. 137, no. 8, pp. 1188–1202.

Haugther, G; Hunter, C. (1994). Sustainable Cities. Londres J. Kingsley Publishers / Regional Studies Association, Bristol.

IBGE. (2012). MALHA municipal digital do Brasil: situação em 2000 e 2010. Rio de Janeiro: IBGE. Available at: <ftp://geoftp.ibge.gov.br/malhas_digitais/> . Accessed on June 2016.

Johansson, M.A; Cummings, D.A.T; Glass, G.E.(2009). Multiyear Climate Variability and Dengue—El Niño Southern Oscillation, Weather, and Dengue Incidence in Puerto Rico, Mexico, and Thailand: A Longitudinal Data Analysis. PLoS Med 6(11): e1000168. doi:10.1371/journal.pmed.1000168, 2009.

Sistema Nacional de Informação sobre Saneamento – SNIS. Eletronic database. Brasília. Available at: http://www.snis.gov.br. Accessed on June 2016.

Tauil, P. L. (2001). Urbanização e ecologia do dengue Caderno de Saúde Pública, Rio de Janeiro, 17(Suplemento):99-102.

Watts, D.M.; Burke, D.S, Harrison BH.(1987). Effect of temperature on the vector efficiency of Aedes aegypti for dengue 2 virus. The American Journal of Tropical Medicine and Hygiene 36:143-152.

World Health Organization (2009). Dengue: Guidelines for diagnosis, treatment, prevention and control. World Health Organization (WHO) and the Special Programme for Research and Training in Tropical Diseases (TDR). 147p.

World Health Organization.(2014). Dengue: Guidelines for diagnosis, treatment, prevention and control. World Health Organization (WHO) and the Special Programme for Research and Training in Tropical Diseases (TDR).