RISK ASSESSMENT OF HEPATITIS-A TRANSMISSION THROUGH THE USE OF SOCIODEMOGRAPHIC AND REMOTE SENSING DATA: A CASE STUDY OF THE STATE OF PARÁ, BRAZIL

Philipe Riskalla Leal[1], Ricardo José de Paula Souza e Guimarães[2], and Milton Kampel[1]

[1]Instituto nacional de Pesquisas Espaciais (INPE)
[2]Instituto Evandro Chagas

November 23, 2022

Abstract

Hepatitis-A is a waterborne infectious disease transmitted by the eponymous hepatitis-A virus (HAV). Due to the disease's sociodemographic and environmental characteristics, this study applied public census and remote sensing data to assess risk factors for hepatitis-A transmission. Municipality-level data were obtained for the state of Pará, Brazil. Generalized linear and non-linear models were evaluated as alternative predictors for hepatitis-A transmission in Pará. The Histogram Gradient Boost (HGB) regression model was deemed the best choice (RMSE= 2.36, and higher Rˆ2 = 0.95) among the tested models. Partial dependence analysis (PDA) and permutation feature importance analysis (PFI) were used to investigate the partial dependences and the relative importance values of the independent variables in the disease transmission prediction model. Results indicated a complex relationship between the disease transmission and the sociodemographic and environmental characteristics of the study area. Population size, lack of sanitation, urban clustering, year of notification, insufficient public vaccination programs, household proximity to open-air dumpsites and storm-drains, and lack of access to healthcare facilities and hospitals are sociodemographic parameters related to HAV transmission. Turbidity and precipitation are the environmental parameters closest related to disease transmission. This study reinforces the need to incorporate remote sensing data in epidemiological modelling and surveillance plans for the development of early prevention strategies for hepatitis-A.

RISK ASSESSMENT OF HEPATITIS-A TRANSMISSION THROUGH THE USE OF SOCIODEMOGRAPHIC AND REMOTE SENSING DATA: A CASE STUDY OF THE STATE OF PARÁ, BRAZIL

Philipe Riskalla Leal[1], Ricardo José de Paula Souza e Guimarães[2], Milton Kampel[1]

[1]National Institute for Space Research (INPE, Instituto Nacional de Pesquisas Espaciais)
[2]Evandro Chagas Institute

Corresponding author: Philipe Riskalla Leal (leal.philipe@gmail.com)

Key Points:

- Hepatitis-A is a waterborne infectious disease responsible for approximately 70,000 deaths per year around the world.
- In this work, sociodemographic and environmental factors are related to hepatitis-A transmission by applying census and remote sensing data.
- This research stresses the need to incorporate remote sensing data to epidemiological modelling for prevention and surveillance plans.

Abstract

Hepatitis-A is a waterborne infectious disease transmitted by the eponymous hepatitis-A virus (HAV). Due to the disease's sociodemographic and environmental characteristics, this study applied public census and remote sensing data to assess risk factors for hepatitis-A transmission. Municipality-level data were obtained for the state of Pará, Brazil. Generalized linear and non-linear models were evaluated as alternative predictors for hepatitis-A transmission in Pará. The Histogram Gradient Boost (HGB) regression model was deemed the best choice ($RMSE$= 2.36, and higher R^2 = 0.95) among the tested models. Partial dependence analysis (PDA) and permutation feature importance analysis (PFI) were used to investigate the partial dependences and the relative importance values of the independent variables in the disease transmission prediction model. Results indicated a complex relationship between the disease transmission and the sociodemographic and environmental characteristics of the study area. Population size, lack of sanitation, urban clustering, year of notification, insufficient public vaccination programs, household proximity to open-air dumpsites and storm-drains, and lack of access to healthcare facilities and hospitals are sociodemographic parameters related to HAV transmission. Turbidity and precipitation are the environmental parameters closest related to disease transmission. This study reinforces the need to incorporate remote sensing data in epidemiological modelling and surveillance plans for the development of early prevention strategies for hepatitis-A.

1. Introduction

Risk assessment and vulnerability analyses are common practices in epidemiology (AVANZI et al., 2018; GULLÓN et al., 2017; WHO, 2014). Evidence from around the world confirms that climate change can affect distribution and occurrence of diseases, a major concern for policy making and healthcare facilities (UN, 2007). The health of human populations is sensitive to shifts in weather patterns and other aspects of climate change (SMITH et al., 2015). Weather events and climate change are important drivers of the transmission of waterborne diseases – for instance, cholera, dysentery and waterborne hepatitis are expected to have higher incidence, or even spread to new areas (AHERN et al., 2005; DAVIES et al., 2015).

Most of the burden of climate change will be borne by developing countries, where the incidence of viral hepatitis and other communicable diseases has traditionally been high, and where healthcare systems still lack proper coverage for health-related products and services (CARBALLO et al., 2013). Previous reports have indicated that the main causes of waterborne diseases are related to contamination of water supply systems, usually through increased run-off from surrounding areas or by inundation (CANN et al., 2013). Nevertheless, other factors, e.g., climate variability, also influence waterborne disease transmission (WHO, 2009).

Hepatitis-A is an infectious disease transmitted by the eponymous hepatitis-A virus (HAV) and accounts for approximately 70,000 deaths per year around the world (WHO, 2016). Hepatitis-A may cause debilitating symptoms and lead to acute liver failure, which is associated with high mortality (WHO, 2019). HAV transmission occurs in different ways, though the fecal-oral route is the most common worldwide (FIORE; WASLEY; BELL, 2006). Fecal-oral transmission occurs when a susceptible person has direct contact with an infectious person or ingests contaminated food or water (WHO, 2011). The latter transmission route is intimately dependent on sanitary, social, cultural and environmental conditions (CLEMENS et al., 2000; FIORE; WASLEY; BELL, 2006; JACOBSEN; KOOPMAN, 2005; MS, 2005; NUNES et al., 2016; PEREIRA; GONÇALVES, 2003).

Previous studies have indicated that hepatitis-A transmission may be related to extreme precipitation and flooding events (GULLÓN et al., 2017; MARCHEGGIANI et al., 2010). In Brazil, extreme precipitation events have been positively related to HAV outbreaks (SANTOS et al., 2019). In Spain, intense rainfall has also been associated with greater incidence of hepatitis-A (GULLÓN et al., 2017). From a climate change perspective, one may expect more intense and more frequent precipitation events in the future (CAMUFFO; DELLA VALLE; BECHERINI, 2018; UN, 2007). This fact and how it bears upon epidemiological outbreaks pose a great challenge for policy making, public health agencies and management planning (MARCHEGGIANI et al., 2010).

Hepatitis-A is endemic in Brazil and affects mostly children, adolescents and young adults (CLEMENS et al., 2000; MS, 2002). Reported cases are heterogeneously spread throughout the country. Specifically, in the northern region, the HAV-related mortality rate has been increasing since 2013. Between 2012 and 2016, the HAV mortality coefficient doubled, reaching 35 cases per million inhabitants (MS, 2018). An anti-HAV vaccine only became available in 2014 and is now included in the National Vaccination Calendar of the Unified Health System (SUS), Brazil's public health system (MS, 2014).

Given the importance of effective prevention and control of hepatitis-A in Brazil and similar places, assessment of the main factors associated with disease transmission is paramount. In order to determine where disease-favoring conditions are present in the environment, remote sensing can be of great importance to assess disease-related environment factors (PATEL, 2020). This assessment can provide meaningful insights for controlling disease transmission.

In light of the topics above, this study assessed how hepatitis-A transmission relates to environmental data detectable by remote sensing and to sociodemographic data derived from the national census and from vaccination programs of the state of Pará (in the Amazon region), Brazil. Various models were tested to best identify and characterize the main variables associated with the hepatitis-A transmission. A municipality grid was applied to perform the spatial aggregation among the datasets.

2. Material and methods

2.1. Study area

Epidemiological, sociodemographic and remote sensing data were obtained for the northern state of Pará, Brazil. In this region, floods are gradual and natural to the ecosystem dynamics (IBGE, 2019). The state of Pará comprises 144 municipalities and six mesoregions (Figure 1). The geographical limits of the municipalities were obtained from the Brazilian Institute of Geography and Statistics (IBGE) (IBGE, 2019) and their grid was applied to spatially integrate the different datasets of this study. The municipality was the political unit of choice, being the smallest political-administrative unit of the Brazilian federative republic (RAMALHO, 2020).

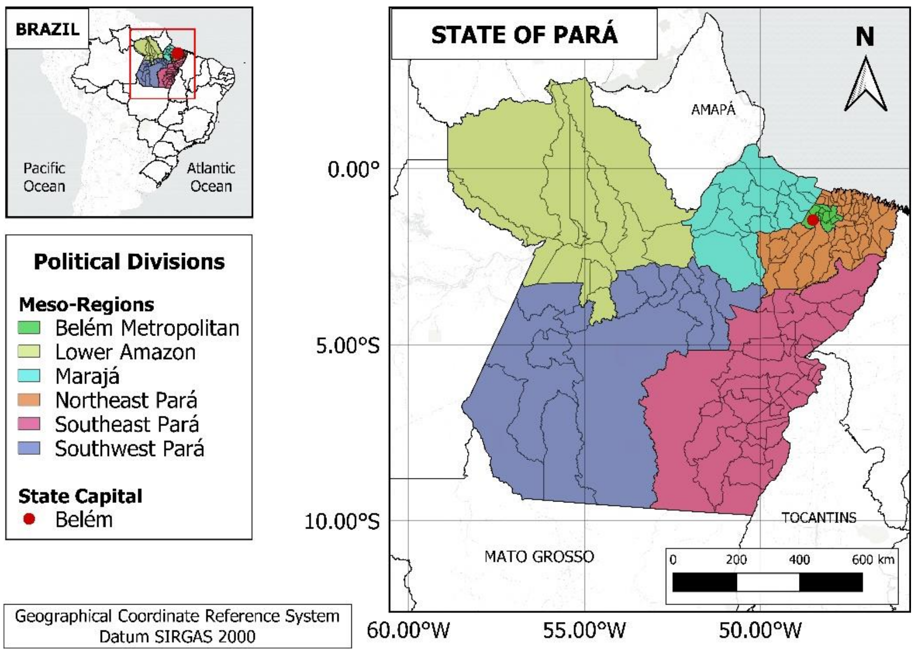

Figure 1: Study area: state of Pará (Brazil). Geographical definitions by the Brazilian Institute of Geography and Statistics (IBGE) (IBGE, 2019).

2.2. Epidemiological data

Information on hepatitis-A cases was obtained from the Notifiable Diseases Information System (SINAN) of Brazil's Ministry of Health (MS, 2007). Data included individual names and addresses, all of which were omitted to ensure and preserve confidentiality, and comprised Pará's residents confirmed to be infected with the hepatitis-A virus between January 2008 and December 2017. The data was aggregated by municipality and month. The epidemiological dataset was geocoded by municipality and consists of 5,500 reported positive new cases (RPC), representing 4.26% of all RPCs in Brazil.

2.3. Sociodemographic data

Annual data on the coverage of the anti-HAV vaccination program and on the number of live births in each municipality were obtained from the SUS's Information Technology Department platform (DATASUS) (MS, 2019), encompassing annual vaccination rates per municipality for the 2014-2017 period. Population coverage of anti-HAV vaccination is the ratio between vaccinated individuals (infants and children under 2) and the total population of a given municipality.

A total of eight variables were obtained from the IBGE's 2010 census data (IBGE, 2010): households with/without sanitation; households near storm drains; households near open-air sewage discharge; households near open-air dumpsites; households with running water; households with water-wheel; and households with a self-supplied water. These census data indicate the number of households in each condition. Therefore, each variable was transformed into relative percentages by dividing the number of households by the total number of households in each municipality. The annual population estimate per municipality was also obtained from the IBGE (IBGE, 2017). The demographic data was applied to evaluate the incidence of the disease in each municipality. A temporal dependence was also incorporated to the model by adding the covariate "year". The variable reflects the year of notification of each reported case of hepatitis-A in the epidemiological data of the SINAN.

All geographical and political boundaries and shapes (municipalities and mesoregions) were obtained from the IBGE (IBGE, 2019). The municipalities' centroid coordinates (longitude and latitude) were taken as covariates during the modelling, enabling the integration of spatial dependence into the models. The municipalities' centroid coordinates were previously reprojected for the SIRGAS 2000 polyconic projection.

2.4. Environmental data

The Google Earth Engine (GEE) platform allows easy access to several global remote sensing datasets thanks to the computational processing power of Google servers (GORELICK et al., 2017). The platform was used to retrieve environmental variables detectable by remote sensing pertaining to hepatitis-A modelling. For the present study, eight variables were selected: surface daytime temperature (SDT), surface nighttime temperature (SNT), turbidity, total suspended matter (TSM), enhanced vegetation index (EVI), normalized difference index ($NDVI$), precipitation, and hydrological mobility index (HMI) (see Table 1). All remote sensing variables were aggregated monthly over the study period (2008-2017).

Table 1: General characteristics of the remote sensing variables used in this study (Data access: Google Engine Platform).

Data	Source	Sensor	Spatial resolution	Spatial aggregation	Temporal resolution
Daytime and nighttime surface temperature	NASA[a]/USGS[b]	MODIS[c]	1x1 km	Average per municipality	8 days
Surface spectral reflectance[d]	NASA[a]/USGS[b]	Landsat series	30x30 m	Average per municipality	16 days
EVI/ $NDVI$	NASA[a]/USGS[b]	MODIS[c]	250x250 m	Average per municipality	16 days
Altimetry	SRTM[e]	Radar	30x30 m		
Precipitation	Climate Hazards Group	Multi-plataform[f]	4 x 4 km	Average per municipality	Daily

(a) NASA: National Aeronautics and Space Administration. (b) USGS: United States Geological Survey. (c) MODIS: Moderate Resolution Imaging Spectroradiometer. (d) Landsat surface spectral reflectance atmospherically corrected by the LASRC algorithm (U.S. GEOLOGICAL SURVEY, 2019).[33] (e) SRTM: - Shuttle Radar Topographic Mission (SAINT-EXUPÉRY et al., 2007).[34] (f) Precipitation from the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) dataset. (FUNK et al., 2015) The dataset comprises different platforms, orbiting sensors and in situ meteorological station data.

Data on surface daytime and nighttime temperatures (SDT and SNT, respectively) were derived from the Moderate Resolution Imaging Spectroradiometer (MODIS), product MOD11A2, with 1 km^2 spatial resolution (WAN, Z., HOOK, S., HULLEY, 2015). SDT and SNT are important factors that induce human behavior, as fluctuations in their values indirectly influence human activities such as bathing, hydration and water recreation (PARSONS, 2003). Thus, one should expect oscillations in daily temperatures to influence HAV transmission.

The Landsat surface reflectance dataset was used to estimate the TSM and the turbidity of the waterbodies of each municipality in Pará. These water-related parameters include water transparency (ALCÂNTARA; CURTARELLI; STECH, 2016; ODY et al., 2016; RODRIGUES et al., 2017), transmittance (LEE et al., 2015), and, consequently, the amount of solar irradiance available in the system. Since solar irradiance directly influences the survival of HAV in aquatic systems through photodegradation (BALES; LI, 1993; HU et al., 2015; MAVIGNIER; FRISCHKORN, 1992), TSM and turbidity are expected to be indirectly related to HAV survival, and therefore, to viral transmissivity.

Turbidity was estimated using a semi-empirical algorithm previously validated for both estuarine and coastal waters (DOGLIOTTI et al., 2015). The algorithm relates turbidity to remote sensing reflectance at wavelength (λ), with $\rho_{w(\lambda)}$. $\rho_{w(\lambda)}$ is defined as the ratio of water-leaving radiance ($L_{w_{(\lambda)}}$) and the above-water downwelling irradiance ($E_{0+_{(\lambda)}}$). The resulting turbidity is expressed in Formazin Nephelometric Units (FNU). The algorithm was validated for independent environments, with stable performance and relative mean error below 13.7%. The algorithm is described in Equation 1,Equation 2 and Equation 3. $A_{(\lambda)}$ and $C_{(\lambda)}$ are spectral conditional constants that follow the conditional rules from Equation 3. w is a linear mixture factor for cases in which $\rho_{w(\lambda)}$ is between 0.05 and 0.07 (sr^{-1}).

$$T\left(\rho_{w(\lambda)}\right) = \frac{A_{(\lambda)} * \rho_{w(\lambda)}}{\frac{\left(1 - \rho_{w(\lambda)}\right)}{C_{(\lambda)}}} \qquad \text{Equation 1}$$

$$w = \left[\frac{\rho_{w_{(\lambda=645)}} - 0.05}{0.02}\right] \qquad \text{Equation 2}$$

$$T = \begin{cases} T\left(\rho_{w(\lambda=645)}\right), & \rho_{w(\lambda=645)} < 0.05 \therefore A_{(\lambda)} = 228.1, C_{(\lambda)} = 0.1641 \\ T\left(\rho_{w(\lambda=859)}\right), & \rho_{w(\lambda=859)} \geq 0.07 \therefore A_{(\lambda)} = 3078.9, C_{(\lambda)} = 0.2112 \\ (1 - w) \cdot T\left(\rho_{w(\lambda=645)}\right) + w \cdot T\left(\rho_{w(\lambda=859)}\right), & 0.05 \leq \rho_{w(\lambda=645)} < 0.07 \therefore \end{cases} \qquad \text{Equation 3}$$

TSM was estimated using a generalized algorithm validated for continental waters (ALCÂNTARA et al., 2016). The algorithm has been previously validated with performance values for root mean square error ($RMSE$) equal to 24.62 (ALCÂNTARA et al., 2016). The algorithm defines TSM as a second degree polynomial function of the ratio of two remote sensing reflectances. For this study, the algorithm was applied to the MODIS dataset, given its higher temporal resolution (daily) vis-à-vis Landsat's (~16 days). Therefore, the spectral bands were corrected to the nearest available band from MODIS (see Equation 4 and Equation 5).

$$TSM = 0.03 * index^2 - 0.08 * index + 0.9 \qquad \text{Equation 4}$$

$$index = \rho_{w_{(\lambda=555)}} / \rho_{w_{(\lambda=469)}} \qquad \text{Equation 5}$$

Since water body dynamics is mainly influenced by climatic and inter-annual variability (i.e., tides, rain cycles, temperature oscillations) (SIMONS; SENTÜRK, 1976), as well as by land use and land coverage changes that directly impact transport of sediments, deposition of materials and biochemistry fluxes (SIMONS; SENTÜRK, 1976), EVI and $NDVI$ were also integrated into the model. Both indexes can be related to surface vegetation coverage (DA SILVA et al., 2019) and both were derived from MODIS product MOD13Q1, with 1×1 km spatial resolution (JUSTICE et al., 1998).

Data from the Climate Hazards Group Infrared Precipitation with Station (CHIRPS) (FUNK et al., 2014) were applied to assess monthly accumulated precipitation in the municipalities of Pará. CHIRPS data have a spatial resolution of ~5.6 × 5.6 km^2 and encompass nearly 30 years of quasi-global rainfall data (50°S-50°N). CHIRPS provides gauge-precipitation satellite estimates with low latency, high resolution, low bias, and long record period (FUNK et al., 2015).

The digital elevation dataset from the Shuttle Radar Topography Mission (SRTM) (SRTM, 2015) and the CHIRPS precipitation dataset were used to estimate the Hydrological Mobility Index (HMI). Both datasets were spatially resampled to the same spatial resolution of the CHIRPS dataset (which has coarser spatial resolution). The index describes the hydrological flushing potential of a given surface (FONSECA et al., 2007) and, thus, can be associated with pathogen dispersal in the environment, serving both as a flusher and a retainer of the virus, influencing disease transmission (BARBOSAI et al., 2017; FONSECA et al., 2007).

Another five environmental variables were also later derived from the CHIRPS dataset to be incorporated in the hepatitis-A modelling: PPF 1.0%, PPF 5.0%, PPF 90.0%, PPF 99.0% and PPF 99.9%, where PPF stands for point-probability function. Each PPF represents the cumulative number of monthly precipitation occurrences given an intensity threshold that might be expected from the PPF of a predefined family of probability distribution functions (PDF). The PPF approach was applied to evaluate the potential relationship between disease transmission and extreme precipitation events (DIAZ; MURNANE, 2008; GULLÓN et al., 2017; MARCHEGGIANI et al., 2010). Since there is still much to be considered with respect to extreme precipitation events, this statistical approach was based on prior similar epidemiological studies (CURRIERO et al., 2001; GULLÓN et al., 2017). In brief, the algorithm for the derivation of these secondary precipitation variables can be described in three steps, as follows:

First, the precipitation time-series is linearly decomposed into three time components: the trend ($T_{(t)}$), the seasonal ($S_{(t)}$) and the residue ($R_{(t)}$). This approach assumes that the trend changes linearly over time, implying a linear additive structure (Equation 6). In addition, the decomposition assumes that seasonality presents constant frequency (width of cycles) and amplitude (height of cycles) over time.

$$Y_{(t)} = T_{(t)} + S_{(t)} + R_{(t)} \qquad \text{Equation 6}$$

Second, a Pearson Type III probability distribution family is fit into $R_{(t)}$ by means of a Maximum Likelihood Estimation (MLE) (VIRTANEN et al., 2020). This PDF family is defined in terms of the mean (μ), the standard deviation (σ) and the skewness ($skew$) of the distribution (VOGEL; MCMARTIN, 1991) (Equation 7). This produces a large number of different distributions, both skewed and symmetrical, and is reduced to a standard frequency function when skewness is zero. This type of distribution is largely used by the U.S. Army Corps of Engineers in flood frequency analysis, by the National Oceanic and Atmospheric Administration in precipitation data analysis, and by the U.S. Navy (FEDERAL AVIATION ADMINISTRATION (FAA), 2003).

$$\mathrm{PDF}(x|\, skew, \sigma, \mu) = \frac{|\beta|}{\Gamma_{(a)}} * [\beta * (x - \zeta)]^{(a-1)} * \exp[-\beta * (x - \zeta)] \qquad \text{Equation 7}$$

where

$$\beta = \frac{2}{skew * \sigma}$$ Equation 8

$$a = (\sigma * \beta)^2$$ Equation 9

$$\zeta = \mu - (\frac{\alpha}{\beta})$$ Equation 10

$$\Gamma_{(x)} = \int_0^\infty t^{(x-1)} * e^{(-t)} dt$$ Equation 11

Finally, $skew$ and σ are the skewness and the standard deviation of the time-series, respectively.

Once the PDF is fitted for $R_{(t)}$, its hyper-parameters as well as the selected percentiles (1.0%, 5.0%, 90.0%, 99.0%, and 99.9%) are used to retrieve thresholds for later classification of $R_{(t)}$. The thresholds are then assessed by means of the point probability function ($PPF_{(x)}$) of the given PDF. PPF is defined as the inverse of a cumulative distribution function (CDF). PPF is also called probability quantile function in statistics literature (WASSERMAN, 2009), but the PPF nomenclature is used here. The Pearson Type III PPF is defined in Equation 12.

$$PPF_{(q|skew,\sigma,\mu)} = CDF^{-1}_{(q|\alpha,\beta,\zeta)} = \frac{1}{\Gamma_{(a)}} * \frac{\left[\int_0^q t^{(a-1)} * e^{(-t)} dt\right]}{\beta} + \zeta$$ Equation 12

For the third and final step of the algorithm, the thresholds derived from Equation 12 are then used to classify $R_{(t)}$. The classified $R_{(t)}$ is then aggregated monthly for each threshold. These parameters are used as proxies for the evaluation of precipitation disaster events, since they can be highly significant for waterborne diseases such as hepatitis-A (FREITAS et al., 2015).

2.5. Data pre-processing

Prior to analyzing the data, all variables and all hepatitis-A cases were aggregated per municipality and per month. Remote sensing variables were averaged per month and per municipality. Precipitation data were summed monthly and averaged spatially for each municipality. Elevation and declivity data were averaged spatially for each municipality.

2.6. Statistical analyses

Multivariate regression analyses were used to evaluate the best model for assessing the main factors that impact hepatitis-A transmission. The evaluated regression models used here were: a) the Generalized Linear Model (GLM); b) the Multilayer Perceptron (MPL) deep-learning algorithm; c) the Gradient Boost (GB); d) the Decision Tree (DT); e) the Histogram

Gradient Boost (HGB). All algorithms are implemented in the Python's Statsmodels package (PEDREGOSA et al., 2011).

In the GLM model, the Poisson and Negative Binomial (NB) probability distribution families were used. In the Poisson distribution, each $Y_{(i)}$ is a random variable in which the Poisson distribution has an expected value ($\mu_{(\mathrm{i})}$) (Equation 13) that represents the number of observed events in a given $\mathrm{municipality}_{(i)}$.

$$Yi \sim Poisson\ (\mu_{(i)}) \qquad \text{Equation 13}$$

The expected value ($\mu_{(i)}$) was assumed to be the linear sum of each relative risk coefficient ($\theta_{(i)}$) and the respective linear expected value ($\mathrm{E}_{(i)}$) (Equation 14). In this study, the relative risk coefficient represents the relative increase in hepatitis-A transmission in $municipality_{(\mathrm{i})}$, while $\mathrm{E}_{(i)}$ is the expected hepatitis-A transmission in $municipality_{(\mathrm{i})}$ under the null hypothesis. Under this hypothesis, the transmission risk of the disease is constant over the entire area of study. The relative risk can take on real values between zero and +∞. If the relative risk is 1, this would mean that all verified municipalities have the same average risk of infection in the area of study; if less than one, it would mean that the municipality's transmission risk is lower. If higher than one, it would mean that the municipality's transmission risks is higher.

$$\mu_{(i)} = E_{(i)} * \theta_{(i)} \qquad \text{Equation 14}$$

Alternatively to the Poisson family distribution, the negative binomial (NB) family is also commonly used to model counting processes, the main difference being that it allows for over-dispersion of the data. Under this assumption, the data follow an expected value $E\big(\mathrm{Y}_{(i)}\big) = \mu_{(i)}$ and variance $V\big(\mathrm{Y}_{(i)}\big) = \mu_{(i)} + (\mu_{(i)})^2/\kappa$ (FOX, 2008). Unless the parameter κ is large, the variance of Y increases more rapidly than for a Poisson distributed variable. By defining the expected value of $\mathrm{Y}_{(i)}$ as a random variable, it is possible to incorporate additional variability among observed counts. The PDF of a NB variable is described in Equation 15.

$$p_{Y_{(y|\mu,\kappa)}} = \exp\{[\mathrm{y}\log\left(\frac{\mu}{(\mu + \kappa)}\right) - \kappa * \log(\mu + \kappa)] + \kappa * \log(\kappa) + \log\big(\Gamma_{(\kappa+y)}\big) - \log\big(\Gamma_{(\kappa)}\big) - \log(y!) \qquad \text{Equation 15}$$

The MPL algorithm is a non-linear model. It assumes that the relationship between the covariates and the dependent variable can be defined by an association of neurons structured in sequential layers (WILDE, 2013). The MPL algorithm accepts several types of activation functions ($log - loss$, $identity$, $tanh$, $relu$, etc.). In this study, the $relu$ activation function (Equation 16) together with stochastic gradient descent $adam$ solver (KINGMA; BA, 2015) were applied to evaluate the weights of the neuron matrix.

$$\mathrm{f(x)} = \max(0, \mathrm{x}) \qquad \text{Equation 16}$$

Gradient Boost (GB), Decision Tree (DT) and Histogram Gradient Boost (HGB) are machine learning (ML) algorithms that can perform both classification and regression tasks.

They are capable of fitting complex datasets in an additive model approach (BOEHMKE; GREENWELL, 2019). These ML algorithms can capture non-linear relationships between the covariates and the dependent variable in forward stage wise fashion (PETRERE-JR; FRIEDMAN, 2000) by minimizing the negative gradient of a given loss function (PEDREGOSA et al., 2011). Machine learning is greatly influenced by its hyper-parameters setting. Therefore, tuning these hyper-parameters is an essential step in analysis. For each model, a grid-search technique (UNPINGCO, 2016) was applied to retrieve the respective best fitting hyper-parameters of each model configuration. The $RMSE$ loss function (Equation 17) was applied to fit each model and, respectively, select the best hyper-parameters. The coefficient of determination R^2 (Equation 19) was also applied for later inter-comparison of models.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}[(\hat{y}_i - y_i)^2]}{n}} \quad \text{Equation 17}$$

$$\text{MSE} = \frac{\sum_{i=1}^{n}[(\hat{y}_i - y_i)^2]}{n} \quad \text{Equation 18}$$

$$R^2 = \left(1 - \frac{MSE}{\sum_{i=1}^{n}[(y_i - \text{E(y)})]^2}\right) = 1 - \frac{MSE}{TSS} \quad \text{Equation 19}$$

After selecting the best regression model for the number of cases of hepatitis-A (the one with the lowest $RMSE$), a partial dependence analysis (PDA) and the permutation feature importance (PFI) were verified. The PDA can depict the relationship between the dependent and the independent variables of the model (MOLNAR, 2019). It graphically structures the variables' marginal effects (whether linear, monotonic or more complex) (PETRERE-JR; FRIEDMAN, 2000). PFI is a model inspection technique especially useful for non-linear/complex estimators (PEDREGOSA FABIAN et al., 2011) and is defined as the decrease in a model score (e.g., $RMSE$) when a single covariate is randomly shuffled (PAVLOV, 2019). A shuffling effort of 99 shuffles was applied for the PFI analysis.

3. Results

A set of six different techniques was applied to model hepatitis-A transmission. Of all models tested (Table 2), HGB Regression proved to be the best in terms of $RMSE$ and R^2 criteria. GB obtained the lowest $RMSE$ of all models, despite its low non-biased R^2. GLM-Poisson, MPL, and DT returned negative R^2 scores, indicating biased estimates. Results from the grid-search analyses can be found in the supporting information.

Table 2: Relationship of the best-fitted models with respective residual fitness.

Models	$RMSE$	R^2	R^2 adjusted	Fitting time (s)	Log-likelihood	Deviance	χ^2
GLM - Poisson	11.311	-3.399	0.331	0.112	$-7.27 * 10^4$	$1.27 * 10^5$	$2.78 * 10^5$
GLM – NB	168.477	0.010	0.323	1.050	$-2.87 * 10^4$	$2.862 * 10^4$	$6.39 * 10^4$

MPL	0.100	-249.366	N/A	3.288	N/A	N/A	N/A
GB	0.094	0.126	N/A	3.543	N/A	N/A	N/A
DT	0.000	-6.061	N/A	0.145	N/A	N/A	N/A
HGB	2.358	0.953	N/A	2.843	N/A	N/A	N/A

After selecting the HGB model, a partial dependence analysis (PDA) was applied to indicate the relative dependence of each variable. The results of the PDA reflected how each variable related to hepatitis-A transmission. PDA values varied between -2.4 and zero (Figure 2). Positive relations were observed for population size, households near open-air sewage discharge, households near open-air dumpsites, and latitude. Negative relations were observed for vaccination coverage, households with public water supply, households with waterwheels, and the municipalities' centroid longitude. A constant relation was observed for the variable households with sanitation. More complex (non-linear) relations were observed for the variables households near storm-drains, households with local water supply and year of notification. The dependences of households near storm-drains, households with local water supply and year of notification presented a bell-shaped pattern, indicating that they varied depending on the municipality and/or period studied.

The environmental variables with positive relations were turbidity, precipitation and $NDVI$ (the latter one in lesser degree) (Figure 2). IMH and EVI were negatively related. A constant relation was observed for SDT, SNT, TSM, and all PPF derived variables. For normalized turbidity values below 1.8, a partial dependence plot of turbidity indicated no clear relationship with disease transmission, but for higher values partial dependence was positively related to disease transmission. Precipitation and $NDVI$ relative dependences were non-linearly associated with disease transmission, although they denoted an average positive trend. With respect to precipitation, there was nearly constant partial dependence for below-average values; for near average values, precipitation had a negative dependence effect; for above average values, precipitation had positive dependence. In respect to $NDVI$, for below zero normalized values, $NDVI$ denoted constant dependence with disease transmission; for higher values, $NDVI$ dependence was positive. EVI denoted an inverse pattern with respect to $NDVI$.

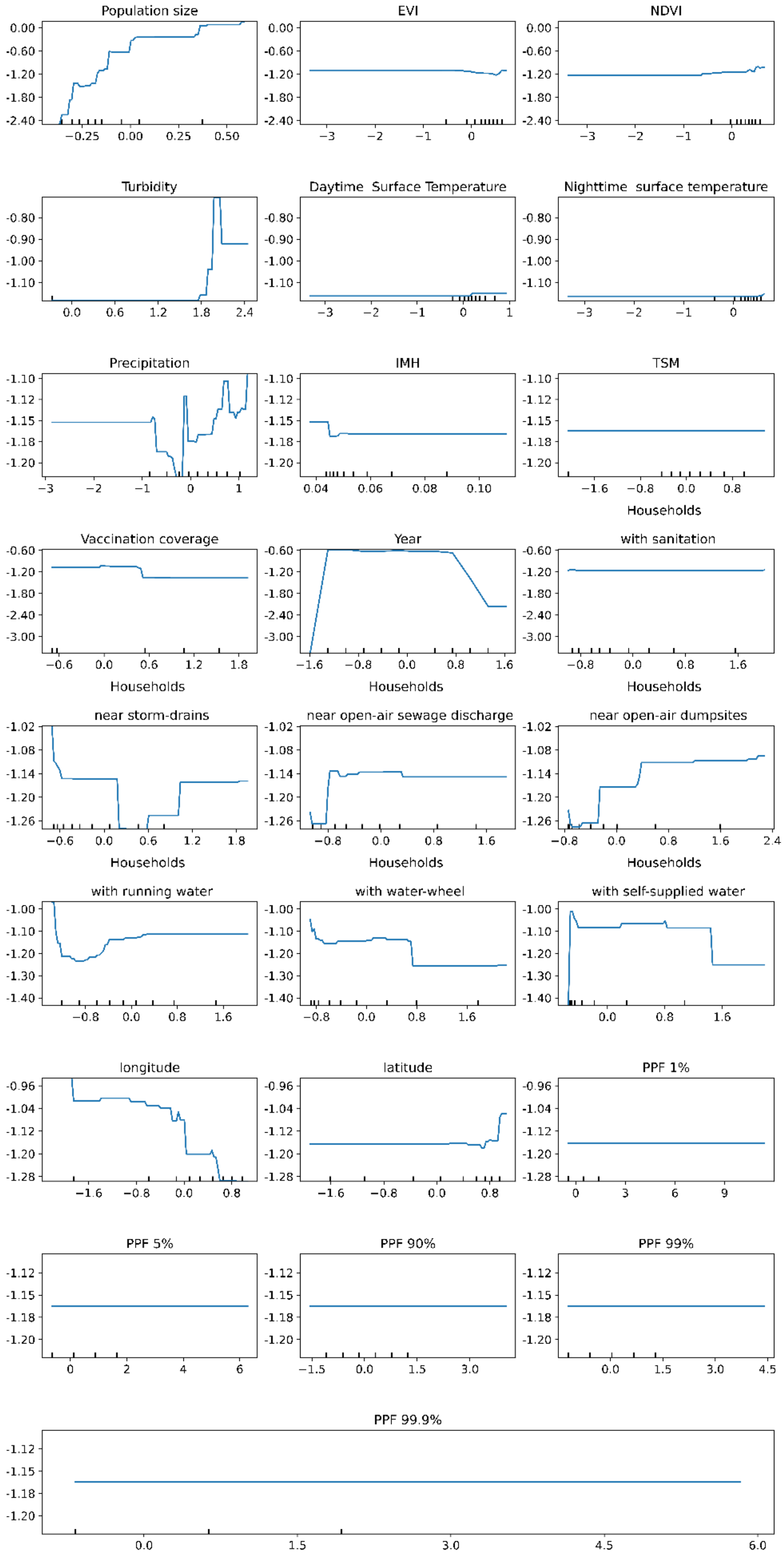

Figure 2: Results of the partial dependence analysis of the explanatory variables for hepatitis-A from the $HGB\ regression$ model. Marks on the x-axis indicate the data distribution.

PFI analysis depicted the relative importance of each environmental and sociodemographic parameter in the HGB model (Figure 3). In decreasing order of importance, population size, $NDVI$, latitude, year of notification and households near open-air dumpsites were the five most significant variables in the model. TSM and all PPF variables were the least significant variables in the model. Uncertainty with regard to PFI values were similar; in decreasing order of uncertainty, the variables were population size, year of notification, vaccination coverage and households near open-air dumpsites.

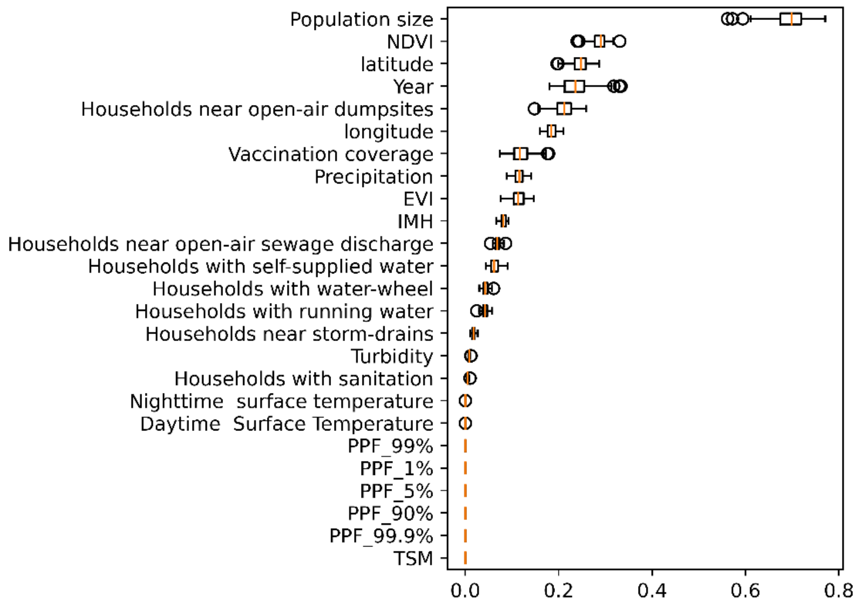

Figure 3: Permutation feature importance analysis depicting the relative importance of each covariate in the HGB model. Shuffling effort: 99 times.

4. Discussion

This study evaluated hepatitis-A transmission by means of sociodemographic and environmental parameters from the state of Pará, Brazil. The observed relations were mostly complex, indicating that multiple interaction effects control disease transmission.

The sociodemographic variables closest related to the disease were the national public vaccination coverage, longitude of the municipality centroids, year of notification and location of households near open-air dumpsites and near storm-drains.

Given the importance of public vaccination in mitigating hepatitis-A transmission (FIORE; WASLEY; BELL, 2006; WHO, 2011), the vaccination relative dependence values were expected to be higher, if not the highest of all variables of the model. The observed low relative dependence was associated with the insufficient coverage rate of the public vaccination program (MS, 2014), as well as with the problems arising from lack of sanitation, sewage disposal and drinking water in the study area (FREITAS et al., 2015; IBGE, 2010; UN, 2007).

The longitude of the municipality centroids showed that disease transmission is space-dependent. Westerly municipalities (longitudes < 50°W) had higher risk of hepatitis-A transmission than easterly municipalities (longitudes > 50°W), a spatial pattern that reflects the sociodemographic characteristics of the study area, where relatively richer and more developed municipalities tend to be located on the eastern part of the state (DO; HUMANO; MUNICÍPIOS, 2010). These findings reinforce the importance of clean drinking water and proper sociodemographic conditions for controlling hepatitis-A transmission (JACOBSEN; KOOPMAN, 2005).

Households near storm-drains were both negatively and positively related to hepatitis-A transmission. For municipalities with a low percentage of households near storm-drains, the relationship was negative, whereas positive dependence was observed for municipalities with high percentage of households near storm-drains. This pattern is associated with population density, storm-drain clogging and the rate of contact of the population with contaminated water-bodies. A similar dual pattern was observed in a previous study, in which the authors suggested that a variable’s dependence duality is a reflection of internal spatial variations of disease transmission in the study area (ROGERS, 2000). This reinforces the notion that epidemiological programs, policy-making and strategy planning must be specific to each area. Only then it is possible to properly consider the unique epidemiological factors associated with a disease’s transmission.

The environmental variables most related to hepatitis-A were turbidity and precipitation. Turbidity has a complex relationship with disease transmission and its dependence pattern is suggested by a peaked Gaussian distribution shape. Lower values of turbidity did not influence disease transmission; for average values, the turbidity was positively associated; and for higher values turbidity was negatively related to disease transmission. The peaked Gaussian distribution shape dependence was attributed to different characteristics of the limnological environment, e.g., increased untreated sewage discharge into the environment (GUIMARAENS; CODEÇO, 2005), contamination of waterbodies nearby, and particle sedimentation (JAMES; LECCE, 2013; UNESCO, 1982). Aside from the fact that untreated sewage is directly linked to virus dispersion and propagation of the disease (GUIMARAENS; CODEÇO, 2005), wastewater also influences the attenuation of light in the water column, increasing the turbidity of the water body (OLIVEIRA et al., 2018). The higher the turbidity, the more suspended particles there are in the water column. And more suspended particles mean a greater adherence rate of other materials (organic and inorganic), leading to an increase in sedimentation rates (GALVEZ; NIELL, 1993; THORNTON, 1990; WILDE, 2013). As a consequence, the suspended particles may act as binding agents in the limnological environment; in sufficiently large number, these particles can more efficiently bind particles like the hepatitis-A virus (KENDALL; KENDALL, 2012), increasing its deposition rate. If there are less HAV available in the system, the chances of infection are reduced, directly diminishing disease transmission. In some cases, increases in turbidity can also be related to increases in water turbulence (KNOBLAUCH, 1999). As turbulence increases, higher dispersion forces act on the HAV present in the water column (SIMONS; SENTÜRK, 1976). As a consequence, turbulence acts as a cleaning agent that diminishes the virus pool available for potential infection (GURJÃO, 2015; SIMONS; SENTÜRK, 1976).

Precipitation also denoted a non-linear association with disease transmission. For below average precipitation, the effect was nearly constant; for near average values, precipitation had a negative effect on disease transmission, while above average precipitation had a positive effect. These findings are associated with the same factors as turbidity. Lower precipitation events induce less turbulent behavior in water bodies, and consequently a higher deposition rate. Under this scenario, HAV is expected to be less present in water systems. The opposite is also true. Under higher precipitation events, the deposition rate is reduced with the increase in water turbulence. Furthermore, under intense precipitation events, contamination of public water supply systems due to increased run-off from surrounding areas, to inundation processes (CANN et al., 2013) and/or to flushing of streets, ponds and other potential water sources (CANN et al., 2013) is an important factor positively related to hepatitis-A transmission.

Previous studies report that hepatitis-A transmission is related to extreme precipitation and flooding events (GULLÓN et al., 2017; MARCHEGGIANI et al., 2010). This study, however, by applying the PPF methodology to the study area, found no statistical evidence supporting such a statement. Despite the different tested PPFs, their respective relative dependencies were constant for disease transmission. As disaster events may impact public health in different time frames – from short-lasting impacts (hours) to long-lasting ones (years) (FREITAS et al., 2015), a time lag effect can impinge a direct assessment of the disease transmission. Thus, future studies are required to investigate this temporal dependence. Auto Regressive Integrated Moving Average (ARIMA) models might be a possible alternative. The technique is widely used in epidemiology (LUZ et al., 2008) and has been applied to other waterborne diseases (CHADSUTHI et al., 2012). Furthermore, given the variability and lack of consensus on how to measure and depict extreme precipitation events (GULLÓN et al., 2017), other methodological approaches to extreme events are required to properly assess a potential relationship with disease transmission.

This study reiterates how important it is for public health practitioners and water companies to be aware of the risks related to waterborne disease outbreaks. The methods applied here can be extended to other waterborne diseases, reinforcing the applicability of this work. Given the impacts of extreme weather events on waterborne diseases, especially under a scenario of climate change, health disparities are likely to occur in the near future. A population's ability to adapt to and limit the effects of such events is likely dependent on socioeconomic and environmental circumstances, as well as on the information and technology available (GULLÓN et al., 2017). By considering the expected increase in hepatitis-A transmission due to climate change, and given the increase in population density and the lack of proper sanitation and vaccination in the area of study, this essay may be of interest for early warning planning in the public health sector (FORD et al., 2009).

5. Conclusions

This study assessed the relationship between hepatitis-A transmission and environmental and sociodemographic variables in the state of Pará, Brazil. Generalized linear and non-linear models were examined as alternative predictors for hepatitis-A. The best-suited model was the Histogram Gradient Boost (HGB). Population size, lack of sanitation and of proper public vaccination, households' proximity to open-air dumpsites and storm-drains, and insufficient access to healthcare facilities and hospitals were the sociodemographic parameters related to HAV transmission. Turbidity and precipitation were the environmental parameters more closely related to disease transmission, and it was found that hepatitis-A transmission was positively associated with periods of average turbidity and more intense precipitation. This study stresses the need to incorporate remote sensing data to epidemiological modelling and surveillance plans in order to develop early prevention strategies for waterborne diseases.

Future studies are required to investigate potential time-dependent relationships of the disease with the environment and society. In addition, properly mitigating the spread of hepatitis-A will only be possible if we invest in alternative strategies for sustained disease control and relief, which are essential for public health policymakers, vaccine developers and disease control specialists to make robust estimates of current and future distribution of disease transmission around the world.

Acknowledgements

We thank the State Health Department of Pará (SESPA) for providing the epidemiological data used in this study.

Ethics approval and consent to participate

This study used secondary, publicly available and unrestricted data that do not identify individuals. The results were aggregated by municipality and, therefore, it was not necessary to go through the Research Ethics Committee (CEP) in accordance with Law no. 12,527/2011 that ensures public access to information (http://www.planalto.gov.br/ccivil_03/_ato2011-2014 /2011/lei/l12527.htm).

Consent to publish

The Geoprocessing Laboratory of the Evandro Chagas Institute of the Ministry of Health is authorized by State Health Department of Pará (SESPA) to use and publish data from the Notifiable Diseases Information System (SINAN) and Epidemiological Surveillance Information System (SIVEP).

Availability of data and materials

The Geoprocessing Laboratory of the Evandro Chagas Institute of the Ministry of Health is authorized to use and publish analyses of data made available by their respective sources. The present applied data and respective processing scripts can be provided on demand to eventual interested parties.

Conflict of Interest

The authors declare that they have no actual or potential competing financial interests.

Funding

PRL received support from the Coordination for the Improvement of Higher Education Personnel, CAPES (finance Code 001), and was partially supported by the National Research

Council, CNPq (grant #313588/2019-8) under program 2019-2023 (no. 4444327/2019-5) of the National Institute for Space Research, INPE; RJPSG was partially supported by Brazil´s National Research Council (CNPq, grant #313588/2019-8).

Authors' contributions

Concept: PRL, MK, RJPSG. Methodology: PLR, RJPSG, MK. Formal analysis: PRL. Resources: MK, RJPSG. Writing (original draft preparation): PRL. Writing (revision and editing): PRL, MK, RJPSG. Supervision: MK, RJPSG.

All the authors have read and approved the final version of the manuscript.

Authors' Information

Philipe Riskalla Leal (PRL) – National Institute for Space Research (INPE), Av. dos Astronautas 1758, São José dos Campos, Brazil, 12227-010 {philipe.leal@inpe.br}.

BSc in Biological Sciences from the Federal University of São Paulo (UNIFESP), MSc in Sciences from the University of São Paulo (USP). PRL is currently a PhD student in the Remote Sensing Postgraduate Program at the National Institute for Space Research (INPE) and a member of the Monitoring Oceans from Space Multi-User Laboratory (MOceans-INPE). PRL has experience in the areas of Geosciences, Limnology, Water Resources and Environment, with emphasis on the use of geoprocessing, remote sensing and geographic information systems techniques.

Milton Kampel (MK) – National Institute for Space Research (INPE), Av. dos Astronautas 1758, São José dos Campos, Brazil, 12227-010 {milton.kampel@inpe.br}.

BSc in Oceanography from the State University of Rio de Janeiro, MSc in Remote Sensing from the National Institute for Space Research (INPE), PhD in Oceanography from the University of São Paulo, with postdoctoral degree from the Bedford Institute of Oceanography, Canada. MK is currently Senior Researcher and Leader of the Monitoring Oceans from Space Multi-User Laboratory (MOceans-INPE) and Tenured Professor of the Postgraduate Program in Remote Sensing (INPE). MK has experience in the areas of Geosciences, Oceanography, Climate and Environment, with emphasis on the use of geotechnologies, remote sensing and geographic information systems.

Ricardo José de Paula Souza e Guimarães (RJPSG) – Instituto Evandro Chagas, Rodovia BR-316 km 7 s/n, Ananindeua, Brazil, 67030-000 {ricardojpsg@gmail.com}.

BSc in Biological Sciences from the University of Taubaté, MSc in Remote Sensing from the National Institute for Space Research, PhD in Biomedicine from the Institute of Teaching and Research at Santa Casa de Belo Horizonte. RJPSG is currently a Full Technologist 2 (Biomedical Research and Investigation in Public Health) and responsible for the Geoprocessing Laboratory of the Evandro Chagas Institute of the Ministry of Health; Tenured Professor of the Postgraduate Program in Epidemiology and Health Surveillance (PPGEVS) at the Evandro Chagas Institute. RJPSG has experience in Geosciences, with emphasis on

Geotechnologies, Remote Sensing, Geographic Information Systems, Spatial Epidemiology, Public Health. RJPSG works on the following areas: geoprocessing in health and the environment, mobile geotechnology, spatial and geostatistical analysis.

References

AHERN, M. et al. Global Health Impacts of Floods: Epidemiologic Evidence. **Epidemiologic Reviews**, v. 27, n. 1, p. 36–46, 2005.

ALCÂNTARA, E.; CURTARELLI, M.; STECH, J. Estimating total suspended matter using the particle backscattering coefficient: results from the Itumbiara hydroelectric reservoir (Goiás State, Brazil). **Remote Sensing Letters**, v. 7, n. 4, p. 397–406, 2016.

ALCÂNTARA, E. H. et al. Spatiotemporal total suspended matter estimation in Itumbiara reservoir with Landsat-8/OLI images. **International Journal of Cartography**, v. 2, n. 2, p. 148–165, 2016.

AVANZI, V. M. et al. Risk areas for hepatitis A, B and C in the municipality of Maringá, Paraná state, Brazil 2007-2010. **Geospatial Health**, v. 13, n. 607, p. 188–194, 2018.

BALES, R. C.; LI, S. MS-2 and poliovirus transport in porous media: hydrophobic effects and chemical perturbation. **Water Resources Research**, v. 29, n. 4, p. 957–963, 1993.

BARBOSAI, V. S. et al. Os Sistemas de Informação Geográfica em estudo sobre a esquistossomose em Pernambuco. **Revista de Saúde Pública**, v. 51, p. 1–10, 2017.

BOEHMKE, B.; GREENWELL, B. **Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow.** . [s.l: s.n.], 2019.

CAMUFFO, D.; DELLA VALLE, A.; BECHERINI, F. A critical analysis of the definitions of climate and hydrological extreme events. **Quaternary International**, n. October 2017, p. 0–1, 2018.

CANN, K. F. et al. Extreme water-related weather events and waterborne disease. **Epidemiology and Infection**, v. 141, n. 4, p. 671–686, 2013.

CARBALLO, M. et al. Migration, hepatitis B , and hepatitis C. **Viral Hepatitis: Fourth Edition**, p. 506–514, 2013.

CHADSUTHI, S. et al. Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time-series and ARIMAX analyses. **Asian Pacific Journal of Tropical Medicine**, v. 5, n. 7, p. 539–546, 2012.

CLEMENS, S. A. et al. Soroprevalência para hepatite A e hepatite B em quatro centros no Brasil. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 33, n. 1, p. 1–10, 2000.

CURRIERO, F. C. et al. The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948-1994. **American Journal of Public Health**, v. 91, n. 8, p. 1194–1199, 2001.

DA SILVA, V. S. et al. Methodological evaluation of vegetation indexes in land use and land cover (LULC) classification. **Geology, Ecology, and Landscapes**, v. 00, n. 00, p. 1–11, 2019.

DAVIES, G. I. et al. Water-borne diseases and extreme weather events in Cambodia: Review of impacts and implications of climate change. **International Journal of Environmental Research and Public Health**, v. 12, n. 1, p. 191–213, 2015.

DIAZ, H. F.; MURNANE, R. J. Preface: The significance of weather and climate extremes to society: An introduction. **Climate Extremes and Society**, v. 9780521870, p. xiii–xvi, 2008.

DO, E.; HUMANO, D.; MUNICÍPIOS, N. O. S. SÍNTESE DO ÍNDICE DE DESENVOLVIMENTO HUMANO MUNICIPAL – IDHM PARA O ESTADO DO PARÁ. 2010.

DOGLIOTTI, A. I. et al. A single algorithm to retrieve turbidity from remotely-sensed data in all coastal and estuarine waters. **Remote Sensing of Environment**, v. 156, n. October 2014, p. 157–168, 2015.

FEDERAL AVIATION ADMINISTRATION (FAA). Using Modern Computing Tools to Fit the Pearson Type III Distribution to Aviation Loads Data. **Dot/Faa/Ar-03/62**, n. September, 2003.

FIORE, A. E.; WASLEY, A.; BELL, B. P. Prevention of hepatitis A through active or passive immunization: recommendations of the Advisory Committee on Immunization Practices (ACIP).Atlanta, U.S.: Coordinating Center for Health Information and Service, Centers for Disease Control and Prevention (CDC), U.S. Department of Health and Human Services, 2006Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16708058>. Acesso em: 12 sep. 2018.

FONSECA, F. R. et al. Desenvolvimento de um índice hidrológico para aplicação em estudos de distribuição da prevalência de esquistossomose em Minas Gerias. **Anais XIII Simpósio Brasileiro de Sensoriamento Remoto**, n. January, p. 2589–2595, 2007.

FORD, T. E. et al. Using satellite images of environmental changes to predict infectious disease outbreaks. **Emerging Infectious Diseases**, v. 15, n. 9, p. 1341–1346, 2009.

FOX, J. **Applied regression and generalized linear models**. . In: **Applied regression analysis and generalized linear models**. 2. ed. Thousand Oaks, CA, US: Sage Publications, Inc, 2008. p. 379–424.

FREITAS, C. M. DE et al. Desastres naturais e saúde : uma análise da situação do Brasil. **Ciência & Saúde Coletiva**, v. 19, n. 9, p. 3645–3656, 2015.

FUNK, C. et al. The climate hazards infrared precipitation with stations - a new environmental record for monitoring extremes. **Scientific Data 2**, v. 2, 8. 2015.

FUNK, C. C. et al. A quasi-global precipitation time series for drought monitoring: U.S. Geological Survey Data Series 832. **Usgs**, n. January, p. 4, 2014.

GALVEZ, J. A.; NIELL, F. X. **Sedimentation and Mineralization of Seston in a eutrophic reservoir, with a tentative sedimentation model**. . In: STRASKRABA, M.; TUNDISI, J. G.; DUNCAN, A. (Eds.). . **Developments in Hydrology: Comparative Reservoir Limnology and Water Quality Management**. Málaga: KLUWER ACADEMIC PUBL, SPUIBOULEVARD 50, PO BOX 17, 3300 AA DORDRECHT, NETHERLANDS, 1993. p. 119–126.

GORELICK, N. et al. Google Earth Engine: Planetary-scale geospatial analysis for everyone. **Remote Sensing of Environment**, v. 202, p. 18–27, 2017.

GUIMARAENS, M. A. DE; CODEÇO, C. T. Experiments with mathematical models to simulate hepatitis A population dynamics under different levels of endemicity. **Caderno de Saúde Pública**, v. 21, n. 5, p. 1531–1539, 2005.

GULLÓN, P. et al. Association between meteorological factors and hepatitis A in Spain 2010–2014. **Environment International**, v. 102, p. 230–235, 2017.

GURJÃO, T. C. M. **GENÓTIPOS DO VÍRUS DA HEPATITE A (VHA) DETECTADOS EM DIFERENTES ECOSSISTEMAS AQUÁTICOS E A RELAÇÃO DO VHA COM OS INDICADORES DE QUALIDADE DA ÁGUA, BELÉM, PARÁ, BRASIL**. . Dissertação (Mestrado em Biologia) — Brasil: Universidade

Federal do Pará, 2015.

HU, Z. et al. Temporal dynamics and drivers of ecosystem metabolism in a large subtropical Shallow Lake (Lake Taihu). **International Journal of Environmental Research and Public Health**, v. 12, n. 4, p. 3691–3706, 2015.

IBGE. CENSO DEMOGRÁFICO 2010: características da população e dos domicílios: resultados do universo. **Sidra: sistema IBGE de recuperação automática**, 2010.

IBGE. Estimativas da população residente no Brasil e Unidades da Federação em 1º de julho de 2017Rio de Janeiro: [s.n.], 2017. Disponível em: <ftp://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2017/estimativa_dou_2017.pdf >.

IBGE. Divisões politico-administrativas do Brasil. 2019.

JACOBSEN, K. H.; KOOPMAN, J. S. The effects of socioeconomic development on worldwide hepatitis A virus seroprevalence patterns. **International Journal of Epidemiology**, v. 34, n. 3, p. 600–609, 2005.

JAMES, L. A.; LECCE, S. A. **Impacts of Land-Use and Land-Cover Change on River Systems**. . In: JOHN F. SHRODER (Ed.). . **Treatise on Geomorphology**. [s.l.], Academic Press, 2013. v. 9p. 768–793.

JUSTICE, C. O. et al. The Moderate Resolution Imaging Spectroradiometer (MODIS): land remote sensing for global change research. **IEEE Transactions on Geoscience and Remote Sensing**, v. 36, n. 4, p. 1228–1249, 1998.

KENDALL, K.; KENDALL, M. **Adhesion of Cells , Viruses and Nanoparticles.** . [s.l.], SPRINGER, 2012.

KINGMA, D. P.; BA, J. L. Adam: A method for stochastic optimization. **3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings**, p. 1–15, 2015.

KNOBLAUCH, H. Overview of density flows and turbidity currents. **Water Resources Research Laboratory**, n. June, p. 27, 1999.

LEE, Z. P. et al. Secchi disk depth: A new theory and mechanistic model for underwater visibility. **Remote Sensing of Environment**, v. 169, n. November, p. 139–149, 2015.

LUZ, P. M. et al. Time series analysis of dengue incidence in Rio de Janeiro, Brazil. **American Journal of Tropical Medicine and Hygiene**, v. 79, n. 6, p. 933–939, 2008.

MARCHEGGIANI, S. et al. Risks of water-borne disease outbreaks after extreme events. **Toxicological and Environmental Chemistry**, v. 92, n. 3, p. 593–599, 2010.

MAVIGNIER, A. L.; FRISCHKORN, H. Physical, chemical and bacteriological study of Cocó River, Fortaleza - Ceará. Anais do 1 simpósio de Recursos hídricos do nordeste, Recife, 25-27 nov. **Anais**...Fortaleza: 1992

MOLNAR, C. **Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.** . 1. ed. [s.l.], Lulu, 2019.

MS. **Programa nacional de hepatites virais: avaliação da assistência as hepatites virais no Brasil 2002.** . 1° edição ed. Brasília - DF:, Brasil: MINISTÉRIO DA SAÚDE, 2002.

MS. **Programa nacional para a prevenção e o controle das hepatites virais: manual de aconselhamento em hepatites virais.** . [s.l.], MINISTÉRIO DA SAÚDE. SECRETARIA DE

VIGILÂNCIA EM SAÚDE. DEPARTAMENTO DE VIGILÂNCIA EPIDEMIOLÓGICA, 2005. v. Série D

MS. **Sistema de informação de agravos de notificação (SINAN): normas e rotinas.** . 2. ed. Brasília: Ministério da Saúde, 2007.

MS. Informe técnico da introdução da vacina adsorvida Hepatite-A (inativada)Brasília, Brasil: MINISTÉRIO DA SAÚDE, 2014Disponível em: <http://portalarquivos2.saude.gov.br/images/pdf/2015/junho/26/Informe-t--cnico-vacina-hepatite-A-junho-2014.pdf>. Acesso em: 30 jul. 2018.

MS. HEPATITES virais 2018 (Secretaria de Vigilância em Saúde − Ministério da Saúde, Ed.)**Boletim Epidemiológico**, [s.l.], MINISTÉRIO DA SAÚDE; SECRETARIA DE VIGILÂNCIA EM SAÚDE, 2018Disponível em: <http://portalarquivos2.saude.gov.br/images/pdf/2018/julho/05/Boletim-Hepatites-2018.pdf>. Acesso em: 7 sep. 2018.

MS. Base de Dados - DATASUS. 2019.

NUNES, H. M. et al. Soroprevalência da infecção pelos vírus das hepatites A, B, C, D e E em município da região oeste do Estado do Pará, Brasil. **Revista Pan-Amazônica de Saúde**, v. 7, n. 1, p. 55–62, 2016.

ODY, A. et al. Potential of high spatial and temporal ocean color satellite data to study the dynamics of suspended particles in a micro-tidal river plume. **Remote Sensing**, v. 8, n. 3, 2016.

OLIVEIRA, A. R. M. DE et al. ESTIMATION ON THE CONCENTRATION OF SUSPENDED SOLIDS FROM TURBIDITY IN THE WATER OF TWO SUB-BASINS IN THE DOCE RIVER BASIN monitoring , variables Knowing the relationship between the total suspended solids concentration (TSS), turbidity in the waters , a. **Engenharia Agrícola**, v. 38, n. 5, p. 751–759, 2018.

PARSONS, K. **Human Thermal Enviroments.** . 2. ed. London, England: Taylor & Francis, 2003.

PATEL, K. Of Mosquitoes and Models : Tracking Disease by Satellite. Disponível em: <https://earthobservatory.nasa.gov/features/disease-vector?src=eoa-features>. Acesso em: 22 jul. 2020.

PAVLOV, Y. L. Random forests. **Random Forests**, p. 1–122, 2019.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PEDREGOSA FABIAN et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PEREIRA, F. E. L.; GONÇALVES, C. S. Hepatite A. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 36, n. 3, p. 387–400, 2003.

PETRERE-JR, M.; FRIEDMAN, J. Greedy Function Approximation: A Gradient Boosting Machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2000.

RAMALHO, H. D. **A caracterização do município como entidade federativa**. . . 2020, p. 1–20.

RODRIGUES, T. et al. Retrieving Total Suspended Matter in Tropical Reservoirs Within a Cascade System with Widely Differing Optical Properties. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 10, n. 12, p. 5495–5512, 2017.

ROGERS, D. J. Satellites, Space, Time and the African Trypanosomiases. **Advances in Parasitology**, v. 47, 2000.

SAINT-EXUPÉRY, A. DE et al. The Shuttle Radar Topography Mission. **Reviews of Geophysics**, v. 45, n. 2, p. 248, 2007.

SANTOS, K. D. S. et al. Perfil da hepatite A no município de Belém, Pará, Brasil. **REvista visa em debate**, v. 7, n. 2, p. 18–27, 2019.

SIMONS, D. B.; SENTÜRK, F. **Sediment transport technology.** . Fort Collins, USA, USA: Water Resources Publications, 1976. v. 1

SMITH, K. R. et al. Human health: Impacts, adaptation, and co-benefits. **Climate Change 2014 Impacts, Adaptation and Vulnerability: Part A: Global and Sectoral Aspects**, p. 709–754, 2015.

SRTM. The Shuttle Radar Topography Mission (SRTM) Collection User Guide. p. 1–17, 2015.

THORNTON, K. W. **Sedimentary Processes**. . In: **Reservoir Limnology: Ecological Perspectives**. [s.l.], John Wiley & Sons, 1990. p. 43–69.

U.S. GEOLOGICAL SURVEY. Landsat 8 Surface Reflectance Code (LASRC) Poduct Guide. (No. LSDS-1368 Version 2.0). n. May, p. 40, 2019.

UN. Climate Change: Impacts, Vulnerabilities and Adaptation in Developing Countries.**United Nations Framework Convention on Climate Change**, [s.l: s.n.], 2007Disponível em: <http://unfccc.int/resource/docs/publications/impacts.pdf>.

UNESCO. **Sedimentation problems in river basins.** . In: **Studies and reports in hydrology**. Paris, France: [s.n.], 1982. . p. 152.

UNPINGCO, J. **Python for probability, statistics, and machine learning.** . [s.l: s.n.], 2016.

VIRTANEN, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. **Nature Methods**, 2020.

VOGEL, R. W.; MCMARTIN, D. E. Probability Plot Goodness-of-Fit and Skewness Estimation Procedures for the Pearson Type 3 Distribution. **Water Resources Research**, v. 27, n. 12, p. 3149–3158, 1991.

WAN, Z., HOOK, S., HULLEY, G. MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006 [Data set]. 2015.

WASSERMAN, L. **All of Statistics: A Concise Course in Statistical Inference.** . [s.l.], Springer Berlin Heidelberg, 2009. v. 46

WHO. Protecting health from climate change: connecting science, policy and people**WHO Library Cataloguing-in-Publication Data**, Denmark: World Health Organization, 2009Disponível em: <https://apps.who.int/iris/bitstream/handle/10665/44246/9789241598880_eng.pdf;jsessionid=76ECF990F9BB0FB66A05CEF32C24613C?sequence=1>. Acesso em: 10 feb. 2020.

WHO. Evidence based recommendations for use of hepatitis A vaccines in immunization services : Background paper for SAGE discussions. n. October, 2011.

WHO. Gender, climate change and health**WHO Library Cataloguing-in-Publication Data**, [s.l.], WHO Press, 2014Disponível em: <https://apps.who.int/iris/bitstream/handle/10665/144781/9789241508186_eng.pdf;jsessionid=FD60C4C0643A7E9306E66D67944C458B?sequence=1>. Acesso em: 10 feb. 2020.

WHO. WHO: Viral Hepatitis 2016-2021[s.l: s.n.], 2016

WHO. Hepatitis A. Disponível em: <https://www.who.int/immunization/diseases/hepatitisA/en/>. Acesso em: 8 may. 2019.

WILDE, P. DE. **Neural Network Models: theory and projects.** . 2. ed. [s.l.], Springer, 2013. v. 369

Supporting Information

The best hyper-parameters of each evaluated regression model are presented in Table 3.

Table 3: Best hyper-parameter settings of the grid search analyses of each tested model.

Models	HL	Learning rate	Leaf size	Min samples per leaf	Nearest neighbors	Max depth	N estimators
GLM	N/A	N/A	N/A	N/A	N/A	N/A	N/A
MPL	(3, 4)	0.001	N/A	N/A	N/A	N/A	N/A
GB	N/A	0.1	N/A	N/A	N/A	17	188
DT	N/A	N/A	N/A	N/A	N/A	N/A	N/A
HGB	N/A	0.05	150	13	N/A	20	N/A

HL: hidden layers - (N° of neurons per layer)

“N/A” indicates a hyper-parameter that is not applicable to a given model.