RBPD
Revista Brasileira de Planejamento e Desenvolvimento

Revista Brasileira de Planejamento e Desenvolvimento
ISSN: 2317-2363

https://periodicos.utfpr.edu.br/rbpd

Integration of Big Data and public health: an innovative methodology for predicting dengue cases in Espírito Santo and Distrito Federal

Rodrigo Straessli Pinto Franklin
rodrigo.franklin@ufes.br
Universidade Federal do Espírito Santo. Vitória. Espírito Santo. Brasil.

Rodrigo Emmanuel Santana Borges
rodrigoesborges@gmail.com
Instituto de Pesquisas e Estatísticas do Distrito Federal. Distrito Federal. Brasil.

Everlam Elias Montibeler
everlamelias@gmail.com
Universidade Federal do Espírito Santo. Vitória. Espírito Santo. Brasil.

Daniel Rodrigues Cordeiro
danielrodriguesco@gmail.com
Universidade Iguaçu. Nova Iguaçu. Rio de Janeiro. Brasil.

Edina Pereira dos Santos
edinapereirasantos@gmail.com
Universidade Federal do Espírito Santo. Vitória. Espírito Santo. Brasil.

ABSTRACT

O artigo em questão, desenvolvido com base em uma metodologia robusta, teve como objetivo estabelecer uma correlação entre o interesse público sobre a dengue, conforme indicado pelos dados do Google Trends, e os casos notificados da doença pela Secretaria Estadual de Saúde do Espírito Santo, Brasil, por um lado, assim como a correlação entre dados similares de tendência e a incidência de dengue no Distrito Federal. Esta pesquisa inovadora reconstruiu uma série temporal do interesse pela dengue em 49 dos 78 municípios do Espírito Santo, visando compreender melhor a dinâmica desta arbovirose em contextos específicos. A dengue representa um desafio significativo para a saúde pública global, com uma distribuição geográfica que se expande rapidamente, principalmente em regiões de clima tropical e subtropical. A metodologia aplicada neste estudo foi de caráter exploratório, mas mostrou grande potencial, pela forte associação verificada mesmo em um modelo tão simples. Assim, a pesquisa propôs uma abordagem empiricamente fundamentada e inovadora, visando contribuir significativamente para o combate à dengue no Espírito Santo e potencialmente em outras regiões com desafios semelhantes.

KEYWORDS: Dengue, Big Data, Epidemiologia, Espírito Santo, Distrito Federal.

1 INTRODUCTION

Dengue fever is a persistent and worrying public health problem in Brazil. The disease, transmitted by the Aedes aegypti mosquito, is highly endemic in several regions of the country, with epidemic peaks that cause a significant impact on the population and health systems (MINISTRY OF HEALTH, 2024a).

In 2024, Brazil faces one of its worst dengue outbreaks in decades, with more than 3.5 million probable cases and more than 1,600 confirmed deaths as of April 16, according to data from the Ministry of Health (2024b). This number represents an increase of 114% compared to cases registered in 2023.

Several factors contribute to the persistence of dengue fever in Brazil, such as precarious sanitary conditions, disorderly urbanization, climate change and the population's lack of awareness about the importance of measures to control the transmitting mosquito (MENDONÇA; SOUZA; DUTRA, 2009; MARQUES et al. 2024).

On the other hand, the scope and availability of practically real-time information about search trends and opinions of the population, offered publicly, albeit with some limitations, by large companies such as Google or Twitter, opens the possibility of using it in managing outbreaks. and dengue epidemics.

In fact, the use of posts on the X network – formerly Twitter – is officially one of the bases of the Infodengue tool, developed by Fiocruz in partnership with the School of Applied Mathematics of the Getúlio Vargas Foundation (CODEÇO et al., 2023).

The present work contributed to this area by seeking to show the potential for using the Google Trends public information source, for more than 50 municipalities in Espírito Santo, by verifying its capacity as a source for nowcasting the number of cases.

To this end, it is structured into four sections in addition to this one. The next section presents a literature review on the analysed object. Next, we present the sources and the method used to verify the ability to predict search trends identified by Google Trends in relation to practically present cases of dengue. Section number 4 briefly presents and discusses the results. Finally, final considerations are made on scope and suggestions for future advances.

2 PUBLIC POLICY FOR DENGUE AND BIG DATA

The "Dengue Interest Index Monitoring Panel" project is an initiative of the Cities Laboratory (LabCidades), which is part of the Data Office nucleus of the Federal University of Espírito Santo (UFES). It was developed in collaboration with the Government of the State of Espírito Santo and the Secretariat of Science, Technology, Innovation and Professional Education (SECTI), aiming to use data analysis to anticipate trends in dengue cases through the interest shown by the population in the search for internet on the subject.

Bitar (2022), in her dissertation on predictive models of dengue transmission scenarios, was based on the use of disease incidence data and factors associated with its transmission in humans, excluding analyses that limited to entomological or climatic data alone. By examining 75 types of predictor variables, categorized into 7 distinct groups, the research revealed a practical and comprehensive

application of the models in the territory, standing out at the municipal, state and provincial levels. The results demonstrated the ability of modelling techniques to anticipate outbreaks up to three months in advance and provide impressive accuracy for up to four weeks, reiterating the complexity of dengue transmission dynamics and the need for continuous adaptations in predictive analyses. This study emphasized the importance of integrating socioeconomic and demographic factors, in addition to territorial and health contexts, into predictive models for a more holistic and effective understanding of dengue prevention and control.

The integration of emerging technologies, such as Big Data, in the formulation and implementation of public policies offers a significant opportunity to improve epidemiological surveillance, predict outbreaks and optimize resource allocation. Infodemiology, which involves the analysis of large volumes of data generated by internet searches, has emerged as a promising tool to complement traditional epidemiological methods (BRIGO et al., 2014; SOUSA-PINTO et al., 2020; PAGUIO et al., 2020).

Additionally, analysis of search trends can reveal geographic and temporal patterns of public interest that are critical to health policy planning and implementation. A study on Systemic Lupus Erythematosus (SLE) showed that searches related to the disease were significantly influenced by media events and the introduction of new treatments, highlighting the importance of effective communication and the impact of celebrities in promoting health awareness (SCIASCIA; RADIN, 2017). In a similar way, the promotion of public policies for dengue can benefit from the use of Google Trends data to identify periods of greater public interest and adjust education and prevention campaigns (PHILLIPS et al., 2018).

2.1 Google Trends as source for nowcasting

Google Trends is a powerful tool that analyses the popularity of search terms on Google over time, providing insights into people's behavior and interests. This tool has proven particularly useful in fields such as epidemiology, where search patterns can indicate the emergence or development of epidemics. For example, a sudden increase in searches for symptoms like "fever" and "cough" in a specific region could signal a possible flu outbreak even before cases are officially reported to health authorities.

The concept of nowcasting, or near-real-time forecasting, applies well to using Google Trends to monitor epidemics. The ability to obtain up-to-date data quickly allows public health professionals to respond more promptly to emerging health threats. During the H1N1 pandemic in 2009, for example, researchers used Google Trends to track the spread of the virus, correlating spikes in searches for information related to the virus with areas where new cases were being diagnosed, enabling a more agile response (COOK et al., 2011).

Furthermore, studies have shown that searches for terms related to infectious diseases, such as "epilepsy" and "seizures", showed significant increases in times of high media coverage, indicating the influence of the media on online search behavior (BRIGO et al., 2014; PHILLIPS et al., 2018).

Despite its usefulness, using Google Trends for epidemic forecasting faces significant challenges. The main limitation is the representation of the data, as not

all people use Google or search for their symptoms online. Furthermore, as previously mentioned, searches can be influenced by factors such as media coverage, which can lead to search spikes not directly related to the actual number of cases. Phillips et al. (2018) identified that search spikes on Google Trends were associated with public awareness events and media coverage and found significant increases in searches for breast cancer during breast cancer awareness month and after celebrity announcements.

Google Trends (GT) can also be used to monitor the effectiveness of public health campaigns and the population's response to these initiatives. For example, during vaccination campaigns, an increase in searches for "vaccination points" or "vaccine side effects" may indicate a high level of public engagement with the campaign. This not only helps assess the impact of health communication strategies but can also indicate areas where additional information or resources may be needed.

Another limitation is related to the lack of absolute data, the dependence on the Google algorithm, which is not published, and the inability to control for confounding factors such as race, socioeconomic status and educational level. Furthermore, GT data is more suitable for urban areas and common topics, while being less effective in rural areas and rare topics (PHILLIPS et al., 2018).

Therefore, although Google Trends data offers valuable insight, it should be used in conjunction with other epidemiological data sources to distinguish between actual biological epidemics and the interest/apprehension generated by the media and public in order to generate a more accurate analysis. it needs (SOUSA-PINTO et al., 2020).

Finally, while Google Trends offers near real-time insight, correctly interpreting the data requires specific expertise. Epidemiologists and data analysts must work together to develop models that fit Google search data to known epidemiological patterns. Interdisciplinary collaboration is crucial to maximize the potential of this tool, not only for nowcasting, but also to better understand disease transmission dynamics and refine public health intervention strategies. In summary, Google Trends is a promising tool that, if used cautiously and in combination with other methodologies, can significantly improve epidemic forecasting and management.

3 METHOD

Code was structured to obtain the daily Google Trends series for cities in Espírito Santo and Brasília, Distrito Federal, for the years 2023 and 2024. Such data was weekly grouped by average.

In the case of Espírito Santo, it was possible to obtain information for 52 of the 78 cities in the federative unit, namely: Afonso Cláudio, Alegre, Alfredo Chaves, Anchieta, Aracruz, Atílio Vivácqua, Baixo Guandu, Barra de São Francisco, Boa Esperança, Cachoeiro de Itapemirim, Castelo, Colatina, Conceição da Barra, Conceição do Castelo, Domingos Martins, Ecoporanga, Fundão, Guaçuí, Guarapari, Ibatiba, Ibiraçu, Iconha, Irupi, Itaguaçu, Itapemirim, Iúna, Jaguaré, Jerônimo Monteiro, Linhares, Mantenópolis, Marataízes, Mimoso do Sul, Montanha, Muniz Freire, Muqui, Nova Venécia, Pinheiros, Piúma, Ponto Belo, Presidente Kennedy, Santa Leopoldina, Santa Maria de Jetibá, Santa Teresa, São Gabriel da Palha, São

Mateus, Serra, Sooretama, Venda Nova do Imigrante, Viana, Vila Valério, Vila Velha and Vitória.

For dengue incidence data, all dengue bulletins published by the Espírito Santo Health Department were compiled via code, which contained, for practically every week, the incidence per 100,000 inhabitants in each city.

To obtain the incidence in absolute terms, population estimates were obtained for each city analysed via the API of the Infodengue Project, from Fiocruz. For the Federal District, the absolute incidence was obtained directly from this tool. The need to process bulletin data and use population data to reach absolute incidences in Espírito Santo was necessary once the State stopped reporting through the national health notification system (SINAN – Ministry of Heath, 2007) occurrences related to dengue and, at the same time, for not making the illnesses recorded publicly and in real time or updated on its own platform that it developed.

Once the series were compiled, it was decided to use panel data analysis with a stationary effects model for Espírito Santo, and direct correlation for the Distrito Federal. The analyzes were separated due to different sources and nature (more indirect or more direct) of the data.

3.1 Tested variables and the model

Regression analysis aims to measure the dependence of a variable (variable to be explained) in relation to one or more independent variables (explanatory variables), whose objective is to estimate and/or predict the behavior of the object under study (MARAVALHAS; SILVA JR., 2019; SILVA, SOUZA; CYSNEIROS, 2019).

In Hair Jr. et al. (2005), it appears that a multiple linear regression model uses more than one independent variable, and these models can be classified into time series, cross-sectional or panel data regressions. The latter has two subclassifications, those of time and individuals, as demonstrated in Equation (1):

$$Y_{i,t} = \beta_0 + \beta_1 x_{1i,t} + \cdots + \beta_n x_{ni,t} + \varepsilon_{i,t} \quad (1)$$

Where:

$Y_{i,t}$ is the dependent variable;

$x_{ni,t}$ existing independent variables of i individuals, representing the cross-section data in which $i \in (1, 2, 3, \dots, N)$ and t the number of periods indicating the series and time, in which $t \in (1, 2, 3, \dots, N)$;

β_0, β_1e β_n are the regression parameter; and

$\varepsilon_{i,t}$ is the representing term of the residual or regression error

For Hsiao (2014), one of the main advantages of panel data analysis is the use of a greater amount of information, the reduction of collinearity problems and the increase in estimation efficiency. The union of time series and cross section increases the degree of freedom of the sample, that is, it increases the number of observations in relation to the number of parameters that we seek to estimate. The larger the sample size, the greater the degree of freedom and, consequently, the better the result of the parameters.

Panel data are essential for analysing the application of theory in empirical research and understanding the behavior of certain variables over time. They capture an individual's change over time, allowing you to identify cycles of change and understand their impacts (MENEGHINI; LANA, 2024). A panel is said to be balanced when information is available for all individuals during all periods considered in the analysis and unbalanced when not all observations are available for all individuals and all periods (OLIVEIRA; BRUNEO, 2019).

The first variable in the model was Absolute Incidence of Dengue (AID): The dependent variable of the model, which represents the total number of dengue cases registered in each municipality each epidemiological week.

The second variable in the model was Searches on Google Trends (TRENDS): The independent variable of the model, which represents the volume of searches related to dengue on Google Trends for each municipality each year. Google Trends provides data on how often users search for certain terms, which can be an indicator of the population's interest and concern about the disease.

Therefore, the equation of this work can be represented as Equation (2):

$$IAD_{i,t} = \beta_0 + \beta_1 TRENDS_{1i,t} + \varepsilon_{i,t} \quad (2)$$

Therefore, the variables in this research were divided as follows: $y_{i,t} = IAD_{i,t}$ (dependent variable – VD), represents the total number of dengue cases registered in each municipality at each epidemiological week; $x_{1i,t} = TRENDS_{1i,t}$ (independent variables – VI), represents the volume of searches related to dengue on Google Trends for each municipality each year. The regression coefficients or parameters $\beta_0$and $\beta_1$are described by Montgomery, Peck and Vining (2012), as: $\beta_0$being the intercept or linear coefficient, which corresponds to the mean of Y when all control variables are equal to zero, that is, it represents the value of Y when X is equal to zero; the coefficient β_1, such as partial or angular regression coefficients.

For this study, i indicates the municipalities studied, $\beta_0$is the intercept that is represented by a scalar variable and is fixed in time (predictable, non-random value), β_1, the angular coefficient (parameter) that will be tested and $x_{1i,t}$ is the independent variable, which varies depending on time t of the periods studied and, in each municipality, i of ES and DF.

Y$Y_{i,t}$is the dependent variable that measures the total number of dengue cases registered in each municipality and $\varepsilon_{i,t}$is the unobservable value of the specified individual effect, that is, the error/disturbance variable which means that the impacts of the independent variable (trends) on the absolute dengue rate (dependent variable) do not reach the expected value.

Regarding the best method, Baltagi (2013) states that there is no consensus on which would be the best technique, whether stationary effects or variable effects. The main difference between the two models is fundamentally how each treats unobserved effects. The usual strategy for specifying the stationary or random nature of effects is done by applying the Hausman test (1978) under the null hypothesis that the GLS estimates (random effects) are consistent. If the null hypothesis is rejected, the effects are considered stationary, and the model is estimated using ordinary least squares (OLS). If the null hypothesis is accepted, we would have the case of random effects, and the model would then be estimated by generalized least squares (GLS).

The stationary effects model is a panel data analysis technique widely used to study longitudinal or temporal data where multiple observations are made on the same subjects (in this case, municipalities) over time. This model is particularly useful for controlling the influence of unobserved variables that may vary across subjects but not over time. When applying the stationary effects model, each municipality would have its own intercept in the model, allowing the analyst to focus only on variations within each municipality over time, such as fluctuations in dengue incidences and search volumes on Google Trends.

To analyse the incidence of dengue in the municipalities of Espírito Santo and its relationship with the volume of searches on Google Trends, the stationary effects model can be a simple and direct initial approach. By analysing data weekly for two distinct years, this model can help identify significant patterns or trends in disease behavior that are consistent within each county over time. For example, an increase in searches for dengue-related terms on Google Trends may precede or coincide with outbreaks of the disease in municipalities, indicating a possible early warning that can be exploited by public health authorities to intervene more quickly.

Furthermore, the use of the stationary effects model facilitates the comparison between different municipalities, controlling the intrinsic characteristics of each one that do not change over time, such as geographic and socioeconomic aspects that can influence the prevalence of dengue. The correlation between search volume and dengue cases, adjusted for these stationary effects, provides a clearer and more isolated analysis of the impact of the variables of interest. Thus, this model not only simplifies initial data exploration but also highlights specific areas for more detailed investigations or focused interventions, making it a valuable tool for public health monitoring and management.

Furthermore, the use of the stationary effects model facilitates the comparison between different municipalities, controlling the intrinsic characteristics of each one that do not change over time, such as geographic and socioeconomic aspects that can influence the prevalence of dengue. The correlation between search volume and dengue cases, adjusted for these stationary effects, provides a clearer and more isolated analysis of the impact of the variables of interest. Thus, this model not only simplifies initial data exploration but also highlights specific areas for more detailed investigations or focused interventions, making it a valuable tool for public health monitoring and management.

4 RESULTS AND DISCUSSION

In the first regression analysis, we sought to verify the effects of the Google Trends variable (TRENDS - independent variable), on the Absolute Incidence of Dengue (AID - dependent variable), in the municipalities of Espírito Santo, in the period from 2023 to 2024, counting with the 52 individuals in the equation (in this case, the municipalities of ES).

In the second regression analysis, we sought to verify the effects of the Google Trends variable (TRENDS - independent variable), on the Absolute Incidence of Dengue (AID - dependent variable), in the municipalities of the Distrito Federal, in

the period from 2023 to 2024, counting with the 16 individuals in the equation (in this case, the municipalities of DF).

Table 1 - Results of panel data analysis - stationary effects model - cities in Espírito Santo - 2023 and 2024

Model 1: MQO, using 2806 observations Included 52 cross section unities Dependent Variable: AID **Regression data:**					
Variable	***Coefficiente***	***Standart Error***	***Reason-t***	***P-value***	
Trends	8,23064	0,13366	61,577	<0,0001	***
Data based statistics:					
Square Resíd. sum	30723000	Total square sum		72284000	
Square-R	0,57497	Ajusted square - R		0,57466	
F (1, 2803)	3791,77	P-value(F)		2,22e-16	

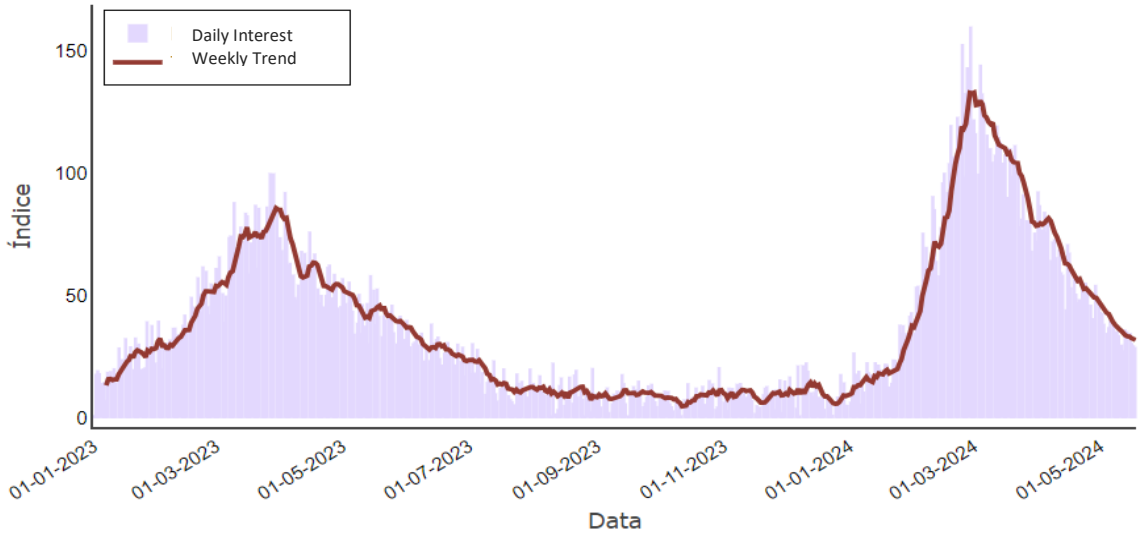
Source: The results presented above were prepared by the authors, based on research data.

Notes: *90% significance; **95% significance; and ***99% significance.

An important item to analyse in the regression model is the statistical significance of the variables, which is represented by the “p-value” column, and as noted, the variables have a statistical significance of 99%. In this way, it was possible to analyse the relationship between the independent variable and the dependent variable (Trends x AID). Another important item to analyse is the adjusted R2 value, which for this first model was 0.57466, that is, it means that the model explains 57.47% of the variation in the absolute incidence of dengue. This means that the model fits the data well and that Google Trends searches are a good predictor of dengue incidence.

For Table 1, we thus obtained an estimated regression coefficient for the "trends" variable of 8.23064, with a standard error of 0.13366. This positive and statistically significant value (p-value < 0.0001***) indicates that there is a strong positive association between the Absolute Incidence of Dengue (AID, dependent variable) and searches on Google Trends (Trends, independent variable), for the municipalities included in the analysis of Espírito Santo. Finally, the highly significant F test (p-value < 2.22e-16) confirms that the model as a whole is statistically significant. This means that there is a general relationship between the incidence of dengue and searches on Google Trends. This correlation can be better visualized inFigure 1, through the Dengue Interest Index Monitoring Panel, developed by LabCidades – Laboratory of Data Science Projects Applied to Cities.

Figure 1 – Trend in interest in information about dengue in ES - 2023 and 2024



Source: Painel da Dengue - LabCidades (2024).

As observed in the panel above, data from searches for interest in information about dengue in the cities of Espírito Santo are in line with weekly trends in dengue cases. For LabCidades (2024), this allows public authorities to plan effective actions against dengue cases up to two weeks in advance.

Table 2 - Results of panel data analysis - stationary effects model - cities in the Federal District - 2023 and 2024

Model 2: MQO, using 68 observations
Included 16 cross section unities included
Dependent variable: AID

Regression data:

Variable	*Coefficiente*	*Standart Error*	*Reason-t*	*P-value*	
Trends	230,290	11,013	20,91	<0,0001	***
Data based statistics:					
Square resíd. Sum	81912000	Total square sum		632910000	
R-square	0,87058	Ajusted square – R		0,8666	
F (1, 65)	437,233	P-value(F)		2,22e-16	

Source: The results presented above were prepared by the authors, based on research data.

Notes: *90% significance; **95% significance; and ***99% significance.

As mentioned previously, an important item to analyse in the regression model is the statistical significance of the variables, which is represented by the “p-value” column, and as noted, the variables have a statistical significance of 99%. In this way, it was possible to analyse the relationship between the independent variable and the dependent variable (Trends x AID). Another important item to analyse is the adjusted R2 value, which for this second model was 0.8666, that is, it means that the model explains 86.70% of the variation in the absolute incidence of dengue. This means that the model fits the data well and that Google Trends searches are a good predictor of dengue incidence.

For Table 2, we obtained an estimated regression coefficient for the variable "trends" that was even higher at 230.290, with a standard error of 11.013. As

explained, we attribute this high association to the way the data was compiled, subject to review and harmonization. This positive and statistically significant value (p-value < 0.0001***) indicates that there is a strong positive association between the Absolute Incidence of Dengue (AID, dependent variable) and searches on Google Trends (Trends, independent variable), for the municipalities included in the Distrito Federal analysis. Finally, the highly significant F test (p-value < 2.22e-16) confirms that the model as a whole is statistically significant. This means that there is a general relationship between the incidence of dengue and searches on Google Trends.

The results in Tables 1 and 2 are in line with the findings of Monnaka and Oliveira (2021), who found that the incidence of dengue and yellow fever in the State of São Paulo showed a strong correlation with the popularity of their terms measured by Google Trends in weekly periods. The Google Trends tool provided early warning, with high sensitivity, to detect outbreaks of these diseases.

Other results along the same lines were found by Liu et al. (2016), who used another internet search tool (Baidu Search Index) to develop a predictive model for dengue outbreaks in Guangzhou and Zhongshan in China. The study indicated that internet-based surveillance systems are lower cost and can be a valuable complement to traditional surveillance.

Ho et al. (2018), also state the importance of using search tools such as Google Trends to complement traditional disease surveillance methods combined with other factors that could potentially identify dengue outbreaks and help with health decisions. The authors used Google Dengue Trends (GDT) in the study to study the metropolitan region of Manila, in the Philippines.

Husnayain, Faud and Lazuardi (2019) highlight that Google Trends data has a linear time series pattern and is statistically correlated with official annual reports on dengue. Identification of information-seeking behavior is necessary to support the use of Google Trends for disease surveillance in Indonesia.

5 FINAL CONSIDERATIONS

The results of the analysis demonstrate a strong positive association between the absolute incidence of dengue and searches on Google Trends in the 52 municipalities of Espírito Santo, as well as in the 16 municipalities of the Federal District. Robust evidence was obtained for the association between the volume of searches as reported by Google Trends and Dengue notifications in the studied municipalities. This suggests that searches on Google Trends can be a useful indicator for monitoring dengue activity and directing actions to prevent and control the disease.

From this information, it will be possible to create models with highly adjusted prediction capacity, taking advantage of studies such as those by Bitar (2022) and Infodengue (Codeço et al., 2022), in addition to greater coverage than that available via other social networks such as (formerly Twitter).

The analysis presented provides evidence of a strong association between the absolute incidence of dengue and searches on Google Trends in Espírito Santo and the Federal District. This information can be used by public authorities and health professionals to monitor dengue activity, direct actions to prevent and control the

disease and evaluate the impact of interventions. However, it is important to consider the limitations of the analysis and conduct additional research to better understand the relationship between these indicators.

Regarding the limitations of the research, the following points were highlighted: a) study lasting only two years; b) sample size; and c) reduced number of independent variables in the model. To indicate future research, it is suggested: a) replication of the model in other municipalities in other states; and b) inclusion of new variables in the model.

Integração de Big Data e saúde pública: Uma metodologia inovadora para a previsão de casos de dengue no Espírito Santo e Distrito Federal

RESUMO

O artigo em questão, desenvolvido com base em uma metodologia robusta, teve como objetivo estabelecer uma correlação entre o interesse público sobre a dengue, conforme indicado pelos dados do Google Trends, e os casos notificados da doença pela Secretaria Estadual de Saúde do Espírito Santo, Brasil, por um lado, assim como a correlação entre dados similares de tendência e a incidência de dengue no Distrito Federal. Esta pesquisa inovadora reconstruiu uma série temporal do interesse pela dengue em 49 dos 78 municípios do Espírito Santo, visando compreender melhor a dinâmica desta arbovirose em contextos específicos. A dengue representa um desafio significativo para a saúde pública global, com uma distribuição geográfica que se expande rapidamente, principalmente em regiões de clima tropical e subtropical. A metodologia aplicada neste estudo foi de caráter exploratório, mas mostrou grande potencial, pela forte associação verificada mesmo em um modelo tão simples. Assim, a pesquisa propôs uma abordagem empiricamente fundamentada e inovadora, visando contribuir significativamente para o combate à dengue no Espírito Santo e potencialmente em outras regiões com desafios semelhantes.

PALAVRAS-CHAVE: Dengue, Big Data, Epidemiologia, Espírito Santo, Distrito Federal.

REFERENCES

BALTAGI, B. H. Econometric analysis of panel data. United States: John Wiley Professio, 2013.

BITAR, Rachel Helen Borges da Silva. Predictive models of dengue transmission scenarios: A scoping review. 2022. 125f. Master's Thesis (Public Health Policy) - Oswaldo Cruz Foundation – Fiocruz, Brasília, Federal District, 2022.

BRIGO, F.; IGWE, S. C.; AUSSERER, H.; NARDONE, R.; TEZZON, F.; BONGIOVANNI, L. G.; TRINKA, E. Why do people Google epilepsy? An infodemiological study of online behavior for epilepsy-related search terms. Epilepsy & Behavior, v. 31, n. 1, p. 67-70, jan. 2014.

CODEÇO, C. T.; BASTOS, L. S.; ARAÚJO, E. C.; et al. Infodengue Group Report 02/23, PROCC/Fiocruz e EMap/FGV, revised version on October 26 de 2023.

CODEÇO, C. T.; OLIVEIRA, S. S.; FERREIRA, D. A. C.; RIBACK, T. I. S.; BASTOS, L. S.; LANA, R. M.; ALMEIDA, I. F.; GODINHO, V. B.; CRUZ, O. G.; COELHO, F. C. Fast expansion of dengue in Brazil. The Lancet Regional Health – Americas, v. 12, n. 1, p. 1-3, 2022.

COOK, S.; CONRAD, C.; FOWLKES, A. L.; MOHEBBI, M. H. Assessing Google Flu Trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. PloS One, v. 6, n. 8, p. 1-8, ago. 2011.

HAIR Jr., J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. Multivariate data analysis. Porto Alegre, RS: Bookman, 2005.

HAUSMAN, J. A. Specification tests in econometrics. Econometrica, v. 46, n. 6, p. 1251-1271, 1978.

HO, H. T.; CARVAJAL, T. M.; BAUTISTA, J. R.; CAPISTRANO, J. D. R.; VIACRUSIS, K. M.; HERNANDEZ, L. F. T.; WATANABE, K. Using Google Trends to examine the spatio-temporal incidence and behavioral patterns of dengue disease: a case study in Metropolitan Manila, Philippines. Tropical Medicine and Infectious Disease, v. 3, n. 4, p. 1-16, 2018.

HSIAO, Ch. Analysis of Panel Data. Cambridge, UK: Cambridge University Press, 2014.

HUSNAYAIN, A.; FUAD, A.; LAZUARDI, L. Correlation between Google Trends on dengue fever and national surveillance report in Indonesia. Global Health Action, v. 12, n. 1, p. 155-165, 2019.

LABCIDADES (Org.). Monitoring panel of the Interest Index for dengue in Espírito Santo – published in 2024. Available in: https://labcidades-ufes.shinyapps.io/dengue/. Access at: 11 jun. 2024.

LIU, K.; WANG, T.; YANG, Z.; HUANG, X.; MILINOVICH, G. J.; LU, Y.; HU, W. Using Baidu Search Index to Predict Dengue Outbreak in China. Scientific Reports, v. 6, n. 1, p. 1-9, 2016.

MARAVALHAS, F. B.; SILVA Jr., L. H. Socioeconomic and demographic factors associated with the existence of museums in Brazilian municipalities: a study applying a linear regression model. Social Studies Notebooks, v. 34, n. 1, p. 1-23, 2019.

MARQUES, A. B.; OLIVEIRA, A. G. M. G. de; RODRIGUES, E. C.; SANTOS, G. F. S.; COSTA, K. O. Dengue: current perspectives and future challenges. Brazilian Journal of Health Review, v. 7, n. 1, p. 6765-6773, 2024.

MENDONÇA, F. de A.; SOUZA, A. V. E.; DUTRA, D. de A. Public health, urbanization and dengue in Brazil. Society & Nature, v. 21, n. 3, p. 257-269, 2009.

MENEGHINI, E. M. P.; LANA, J. Pensata: choosing between fixed and random effects in panel data analysis. Internext, v. 19, n. 1, p. 16-23, 2024.

MINISTRY OF HEALTH (Org.). Department of Health Surveillance. Department of Epidemiological Surveillance. Notifiable Diseases Information System – SINAN: standards and routines. 2nd ed. Brasília: Ministry of Health, 2007. (Series A. Standards and Technical Manuals).

MINISTRY OF HEALTH (Org.). Ministry of Health announces national D-Day to combat dengue – published in 27 de fev. de 2024a. Available in: <https://www.gov.br/saude/pt-br/assuntos/noticias/2024/fevereiro/ministerio-da-saude-anuncia-dia-d-nacional-para-combater-a-dengue>. Access at: 24 jun. 2024.

MINISTRY OF HEALTH (Org.). Ministry of Health updates epidemiological scenario on dengue in Brazil – published in 06 de mar. de 2024b. Available in: <https://www.gov.br/saude/pt-br/assuntos/noticias/2024/marco/ministerio-da-saude-atualiza-cenario-epidemiologico-sobre-a-dengue-no-brasil>. Access at: 24 jun. 2024.

MONNAKA, V. U.; OLIVEIRA, C. A. C. D. Correlation and sensitivity of Google Trends for dengue and yellow fever outbreaks in the state of São Paulo. Einstein, v. 19, n. 1, p. 1-6, 2021.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. Introduction to linear regression analysis. New Jersey, USA: Wiley, 2012.

OLIVEIRA, R. A.; BUENO, L. R. The impact of PRONAF financing on agricultural indicators in crops in the State of Paraná: an analysis of data in a Redes panel. Regional Development Magazine, v. 24, n. 1, p. 292-309, 2019.

PAGUIO, J. A.; YAO, J. S.; DEE, E. C. Silver lining of COVID-19: Heightened global interest in pneumococcal and influenza vaccines, an infodemiology study. Vaccine, v. 38, n. 34, p. 5430-5435, 2020.

PHILLIPS, C. A.; LEAHY, A. B.; LI, Y.; et al. Relationship between state-level Google online search volume and cancer incidence in the United States: Retrospective study. Journal of Medical Internet Research, v. 20, n. 1, p. 1-9, 2018.

SCIASCIA, S.; RADIN, M. What can Google and Wikipedia tell us about a disease? Big data trends analysis in systemic lupus erythematosus. International Journal of Medical Informatics, v. 107, n. 1, p. 65-69, 2017.

SILVA, W. J. F.; SOUZA, R. M. C. R.; CYSNEIROS, F. J. A. Polygonal data analysis: A new framework in symbolic data analysis. Knowledge-Based Systems, v. 1, n. 163, p. 26-35, 2019.

SOUSA-PINTO, B.; ANTO, A.; CZARLEWSKI, W.; et al. Assessment of the impact of media coverage on COVID-19–related Google Trends data: Infodemiology study. Journal of Medical Internet Research, v. 22, n. 8, p. 1-12, 2020.

Recebido: 17 jul. 2025.
Aprovado: 22 ago. 2025.
DOI: 10.3895/rbpd.v14n2.19159
Como citar: FRANKLIN, R. S. P.; BORGES, R. E. S.; MONTIBELER, E. E.; CORDEIRO, D. R.; SANTOS, E. P. Integration of Big Data and public health: an innovative methodology for predicting dengue cases in Espírito Santo and Distrito Federal. **R. Bras. Planej. Desenv.** Curitiba, v. 14, n. 03, p. 613-627, set./dez. 2025. Disponível em: <https://periodicos.utfpr.edu.br/rbpd>. Acesso em: XXX.
Correspondência:
Daniel Rodrigues Cordeiro
RJ-105, 2134 - Luz, Nova Iguaçu - RJ