UNIVERSITY OF PAVIA

DEPARTMENT OF ELECTRICAL, COMPUTER AND BIOMEDICAL ENGINEERING

DOCTOR OF PHILOSOPHY IN ELECTRONIC ENGINEERING

MODELING URBAN AREAS EPIDEMIOLOGICAL RISK EXPOSURE USING MULTISPECTRAL SPACEBORNE DATA

Supervisor: Prof. Paolo Ettore GAMBA

Dissertation of

Oladimeji Ezekiel MUDELE

A.Y. 2019/20

Declaration of Authorship

I, Mudele Oladimeji Ezekiel, declare that this thesis titled, 'MODELING URBAN AREAS EPIDEMIOLOGICAL RISK EXPOSURE USING MULTISPECTRAL SPACEBORNE DATA' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Abstract

In recent decades, the world has been fast urbanizing. More than half of the world's human population now live in urban areas. Such high density of urban population is resulting in air and water pollution, land degradation, and infectious diseases spread risks prominence. However, the increasing quality (in terms of finer spatial and temporal resolution)and quantity of Earth Observation (EO) satellite data provide new perspectives for analysing these phenomena.

Within the specific domain of epidemiological risks dynamics in urban areas which is the focus of this work, the use of multispectral optical EO sensor data has created new opportunities. These data through their visible, near, mid, far and thermal infrared bands provide planetary-scale access to environmental variables such as temperature, humidity, and vegetation types,location and conditions. Since these environmental variables affect the development of vectors causing infectious disease (e.g., mosquitoes), there is the possibility to use EO data to estimate them, and obtain disease risk models.

The *Ae. aegypti* mosquito species transmits Zika, Dengue, and Chikungunya, diseases widespread in more than 100 world countries, and is concentrated in urban areas. The development of this vector depends significantly on local environmental temperature, humidity, precipitation and vegetation. In this regard, multispectral EO data can provide globally consistent and scalable sources to obtain the required environmental variable inputs, and extract significant and consistent monitoring and forecasting models for vector population.

In the domain of modeling mosquito vector population based on EO data features, spatial models require detailed vegetation maps as input. In this regard, there is the need for robust methods to extract this information from freely available high resolution EO data. Temporal models, on the other hand, suffer quality and explainability, which make them limited in application potential. This thesis is targeted towards mitigating these challenges.

Specifically, this thesis report the following contributions:

1. A method to map vegetation types in urban areas at high spatial resolution using Sentinel-2 multispectral EO data. The results show an improvement in the quality of the resulting vegetation maps with respect to what is available by means of state-of-the-art techniques.

2. A method that combines EO-based spectral indices, temperature layers, and precipitation measurement to model the temporal evolution of local mean *Ae. aegypti* population. The approach leverages the random forest (RF) machine learning (ML) technique and its embedded nonlinear features importance ranking (mean decrease impurity, MDI) to rank the effects of environmental variables and explain the resulting model. The results here show that by ranking the environmental variables with MDI, it is possible to isolate variables

that make highest contributions to the vector population development and explain their individual effects.

3. A weighted generalized linear modeling (GLM) technique to predict *Ae. aegypti* population using multispectral EO data covariate inputs. GLMs are generally simple to implement and explain, but do not provide the same level of prediction quality as ML methods. The proposed weighted GLM compares well with ML techniques in quality, and provides capability for more explicitly interpretation of the results.

4. A recurrent neural network (RNN) technique for spatio-temporal modeling of *Ae. aegypti* population at the urban block level using multispectral EO data as inputs. This study is needed because spatial models obscure seasonality effects while temporal model are blind to spatial changes in micro-climates. The proposed technique shows great promise with respect to the use of free multispectral EO data for spatio-temporal epidemiological modeling. Precisely, the model obtained with RNN showed better quality when compared to other traditional machine learning models.

All the proposed techniques have been applied in the Latin American region where the risk of *Ae. aegypti* vector transmitted diseases are the highest in the world. They were validated thanks to the long term partnership with the University of Alagoas in Maceiò (Brazil) and the Brazilian company: ECOVEC.

Acknowledgements

Firstly, I acknowledge the financial support of the European Commission through Horizon 2020 research and innovation programme, grant agreement No 734541 - Project "EOXPOSURE".

In addition, I acknowledge the support of the Telecommunication and Remote Sensing Laboratory of the University of Pavia Italy where I have been engaged in my doctoral research. My utmost appreciation goes to my advisor, Prof. Paolo Ettore Gamba, and also to Prof. Fabio Dell'Acqua who is the project coordinator for "EOXPOSURE".

I acknowledge the contributions of Universidad Nacional de Córdoba, Argentina, Comisión Nacional de Actividades Espaciales (CONAE), Córdoba, Argentina, and Universidade Federal de Alagoas, Maceió, Brazil where I have passed some time during my doctoral study. Particularly, I express my appreciation to Prof. Marcelo Scavuzzo and Prof. Alejandro Frery who provided direct support for me during my time at these named institutions.

I am grateful to my parents, Mr Abiodun Mudele and Mrs Oluwatoyin Mudele, and my brothers, Abiodun, Abiola and Oluwatosin for the never ending support and love. I also acknowledge the emotional and experiential support from my friends namely: Moses Koledoye, Yilkal Belay, Olawale Ajani, Ayodeji Babalola.

To God be the Glory.

Contents

[1] This chapter has been published as a standalone paper as O. Mudele, P. Gamba "Mapping vegetation in urban areas using Sentinel-2", in the Proc. of the 2019 Joint Urban Remote Sensing Event (JURSE2019), Vannes (France), 2019, unformatted CD-ROM, doi: 10.1109/JURSE.2019.8809019.

[2] This chapter has been published as a standalone paper as O. Mudele, F. Bayer, L. Zanandrez, A. Eiras, P. Gamba, "Modeling the Temporal Population Distribution of Ae. aegypti Mosquito using Big Earth Observation Data," IEEE Access, doi: 10.1109/ACCESS.2020.2966080, vol. 8, no. 1, pp. 14182-14194, Jan. 2020.

[3] This chapter has been published as a standalone paper as O. Mudele, A. C. Frery, L. Zanandrez, A. Eiras, P. Gamba, "Modeling dengue vector population with earth observation data and a generalized linear model." Acta Tropica, doi: https://doi.org/10.1016/j.actatropica.2020.105809, vol. 215, mar. 2021.

List of Figures

List of Tables

Chapter 1

Monitoring health risk in urban areas using remote sensing

This chapter gives a broad introduction on remote sensing as a tool for monitoring urban areas, with main focus on epidemiological applications of satellite data. It introduces the main motivations and objectives of this thesis and presents its structure and organization.

1.1 Why Urban Remote Sensing?

Urbanisation is one of the most prominent phenomena that characterize the 21^{st} century. It is a complex socio-economic process that results in the expansion of built-up extents due to the conversion of formerly rural into urban settlements and movement of rural populations to cities. According to the "World Urbanization Prospects 2018" report of the UN [1], in 2007, for the first time in recorded history, the population of the world became more urban than rural. This process is expected to continue for many decades in the future. Globally, at least 55% of world population now live in urban areas, and this percentage is set to increase to 68% by 2050.

The effects of urbanisation are two-edged. On one hand, it fosters benefits including better access to healthcare, sanitation, quality education, better employment, entrepreneurship, technological innovation, etc. Its downsides, however, include environmental degradation, pollution, diseases, and infections, among others. From an epidemiological point of view, the negative effects of urbanisation are more prominent in resource-poor countries due to lack of planning which has led to haphazardly expanding cities within which shanties and slums have also developed. While more advanced countries have the regulatory framework to ensure that minimum

standards of living conditions are met in most parts of the urban areas, resource-poor countries do not have such frameworks, and thus are at the risk of bearing the brunt of the negative spillover effects from global urbanisation [2].

In general, the rise of urban areas and growing population in many parts of the world calls for interventions from public and private actors. Such interventions require data to provide knowledge of how cities and their environmental processes are changing. In addition, for the emerging urban areas to support healthy living, there is the need to develop tools that can help monitor the health risk exposure by humans in these increasingly congested areas.

There are different types and sources of data that can be used for urban areas monitoring purposes. They include population census, household survey based socio-economic data, mobile phone records [3], aerial photographs and remote sensing, among others. However, with the recent strides in space missions and increasing access to petabytes of moderate and high spatial resolution satellite images of the earth surface, directions in urban monitoring have largely shifted towards the use of remote sensing data [4]. Remote sensing data, also referred to as Earth Observation (EO) data in some literature[1], are applicable and have already been successfully applied for many urban monitoring studies. The advantages of remote sensing data for this kind of applications include: free availability in many cases; global coverage; high spatial and temporal resolutions compared to other techniques, and specifically ground data collection.

Remote sensors leverage different physical principles to provide information about the electromagnetic properties of observed land surface by measuring the energy reflected (passive optical sensors), emitted (thermal infrared) or scattered (active radar sensors) by the scene. Consequently, most remote sensing data can be categorised into either optical or radar (mostly synthetic aperture radar, SAR) data. These two data types provide a variety of information on urban areas surface properties. For instance, the optical energy reflected by vegetation depends on properties like leaf structure, moisture and pigmentation, while the scattered radar energy by the same vegetation depends on the size, shape, orientation and dielectric properties with respect to the wavelength of incident microwave energy. Optical data used in urban studies are commonly available in multispectral format: containing a few bands ranging from visible to infrared wavelengths in the electromagnetic spectrum to provide spectral signatures of materials on the surface of the Earth. Radar data are typical only generated at a single wavelength band for each sensor [5]. Both data types have been applied in urban studies (differently or in combination) for the purposes of urban extent mapping [6–8], land use and land cover mapping [5, 9], urban meteorology and heat island effects [10], building footprints detection [11], disaster management [12, 13], change detection and urban sprawl [14–16], among others.

[1] For this reason, in this work the term ”remote sensing data” is used interchangeably with ”Earth Observation (EO) data”.

1.2 Remote Sensing Data Application for the 2030 Sustainable Development Goals

Uncontrolled development along with recent economic events pose great risks to stable functioning of the Earth's systems; atmosphere, forests, water, and biodiversity. As a result, in year 2015, The 2030 Agenda for Sustainable Development was adopted by all United Nations to chart a course for peace and prosperity for the majority on the planet. The 17 sustainable development goals (SDGs) were adopted at the heart of this agenda. These goals are meant to help countries to measure, manage and monitor progress on economic, social and general environmental sustainability [17].

Remote sensing data can support various activities leading towards the actualisation of the SDGs. In this work the focus is on goal number 11, aiming at making cities and human settlements more inclusive, safer, more resilient and sustainable. In support of this goal, alongside vector-borne diseases risks modeling application which is the main focus of this thesis, remote sensing data have been used for other activities related to urban areas analysis including: housing conditions, sustainable transportation [18], air pollution analysis, and human health indicators modeling [19].

With regards to housing conditions mapping, remote sensing data have been applied to measure intra-urban poverty in the socioeconomically divergent city of Medellin, Colombia as presented in [20]. In that study, a land cover classification map was combined with texture and structure features obtained from very high resolution Quickbird RS data to estimate the slum index of the city. The slum index is a value ranging from from 0 (no slum-like households present) to 5 (all households in the area lack all five of the features defined by UN-Habitat [21]). Results from that study show that the RS variables explain up to 59% of the intra-urban variability of the slum index. This kind of approach can serve to lower the cost of socioeconomic surveys.

In relation to air pollution in urban areas, particulate matter (e.g., PM2.5) is one major urban air pollutant which causes respiratory and lungs diseases. In [22], aerosol optical thickness retrievals from multispectral remote sensing data has been combined with ground measurement of PM2.5 mass concentration to assess air quality in different cities across the world. The results show an excellent linear correlation of 0.96 between RS data and ground-based values.

On the use of RS data for health indicators modeling, the authors of [23] produced land use and land cover maps from multispectral RS data spanning a period of 25 years in the city of Atlanta, Georgia, USA. They observed dramatic changes in land use and land cover due to loss of forests and croplands to urbanisation. These changes have led to rising surface temperatures and ground level ozone production. Correlation analysis showed that the RS data features strongly correlate

with volatile organic compounds and nitrogen oxide emissions which are weakly linked to the rates of cardiovascular and chronic lower respiratory disease development.

Among the possible challenges quickly recapped in the previous paragraphs, the focus of this thesis is on the exploitation of urban remote sensing for health risk exposure evaluation, with particular focus on vector-borne diseases. In order to introduce the state of the art in this topic, therefore, the next section presents an overview of how EO data have been exploited for this task.

1.3 Remote sensing for vector-borne disease risks mapping: overview and state of the art

According to the World Health Organization, more than half the world's population is at risk for vector-borne diseases. A major challenge posed by urbanisation is the increase in infectious disease spread due to more people being potentially exposed. Even though the quality of life can be relatively better in urban areas compared to their neighbouring rural counterparts, a deep dive into the intricacies of cities often reveals more underlying information. Cities around the world are composed of heterogeneous groups with different living conditions, microclimates, and economic zones. The overcrowding effects of uncontrolled migration have led to the creation of slums and shanty towns in many urban cities. These slums are characterised in many cases by poor housing, lack of fresh water, bad sanitation facilities, and are exposed to more diseases. These effects have been shown to facilitate the transmission of diseases (e.g., typhoid, cholera, tuberculosis) by supporting breeding grounds for vectors like mosquitoes, rodents, ticks, and more [24].

Many studies have been performed to assess the relationships between environmental variables and the geographical distribution/number of disease transmission vectors in urban areas. In fact, such explorations date back to years well before the 20^{th} century when Hippocrates had recognised the relationships between landscape and human health. The term "landscape epidemiology" was later coined by the Russian parasitologist Evgeny Nikanorovich Pavlovsky to describe the theory that the prevalence of certain carrier vectors and the diseases they cause in certain locations are influenced by the variables of their host environment. The environment provides the necessary conditions for survival, transmission and reproduction of the diseases or their causal vectors. Landscape epidemiology has since evolved into a thriving domain which includes the exploitation of remote sensing data and other geographical information systems to explore, model, and understand the prevalence of diseases/vectors as a function of underlying driving environmental influences. Examples of infectious disease carrying vectors which have

been studied and have shown to depend on environmental conditions for their survival and reproduction include mosquitoes, rodents, and ticks. Diseases caused by these named vectors have dominated landscape epidemiology research studies till date [25]. Some of the environmental parameters that can be measured from remote sensing data and have been used to model the prevalence of these diseases include land surface temperature, vegetation types and conditions, humidity, and precipitation, among others [26]. Even though meteorological collecting station on the ground provide information on temperature, precipitation and humidity, the scarcity of such stations means their data can only provide meaning information for country-level or global studies. Instead, urban-level (regional or municipal) studies require more details. Also, for cases where vegetation condition and types are of effect in vector prevalence, meteorological sources may only be partially useful.

In this work, epidemiology studies involving multispectral sensor data is of particular interest. As we know, the surface of the earth interacts with the incident electromagnetic energy emitted by the sun. This incident energy is either absorbed, reflected, or transmitted by the objects on the Earth surface. The reflected energy is usually sensed in the visible and infrared wavelength bands. This is the basics of optical remote sensing. Multispectral remote sensors are optical sensors with a few spectral bands (typically < 10) [27]. Figure 1.1 presents an illustration of a typical optical remote sensing scenario.

One simple way to obtain information about the Earth surface from multispectral remote sensing data is through spectral indices. Spectral indices are obtained by combining the spectral reflectance of two or more bands of a multispectral data to highlight the relative abundance of

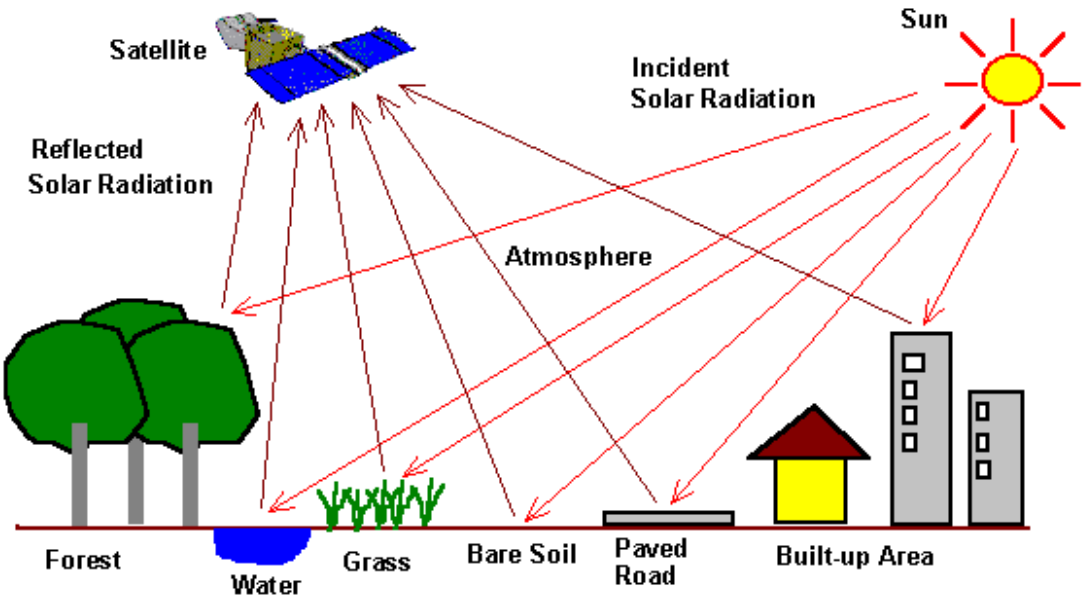


FIGURE 1.1: A typical optical remote sensing scenario (from [27])

certain features of interest. The common spectral indices which find application in landscape epidemiology and their formulas are presented in Table 1.1.

TABLE 1.1: Spectral indices based on multispectral remote sensing data commonly used in landscape epidemiology.

Index	Formula
Enhanced Vegetation Index (EVI) [28]	$\frac{\text{NIR}-\text{Red}}{\text{NIR}+\text{C1}\times\text{Red}-\text{C2}\times\text{Blue}\times\text{L}}$
Normalized Difference Vegetation Index (NDVI) [28]	$\frac{\text{NIR}-\text{Red}}{\text{NIR}+\text{Red}}$
Normalized Difference Water Index - (NDWI) [29]	$\frac{\text{NIR}-\text{SWIR}}{\text{NIR}+\text{SWIR}}$
Normalized Burn Ratio (NBR) [30]	$\frac{\text{SWIR}_1-\text{SWIR}_2}{\text{SWIR}_1+\text{SWIR}_2}$

Blue: Blue band ($\approx 490\,\text{nm}$)
Red: Red band ($\approx 700\,\text{nm}$)
NIR: Near infrared band ($\approx 850\,\text{nm}$)
SWIR: Shortwave infrared band ($\approx 1500\,\text{nm}$—$2200\,\text{nm}$)
C1 and C2 are coefficients of aerosol resistance, L is canopy background adjustment, and G is gain factor. For MODIS EVI product, C1 = 6, C2 = 7.5, L = 1 and G = 2.5 [28].

Other information that can be obtained from multispectral remote sensing data and are useful for epidemiological studies are land cover maps and land surface temperature values [31]. However, temperature information can only be retrieved from data with thermal infrared band(s), i.e. collecting not only the energy emitted by the sun, but the one emitted by the Earth as well.

To study epidemic spread using remote sensing data, the considered disease or carrier vector's development and prevalence must vary spatially and/or temporally based on certain environmental conditions that can be estimated using remote sensing data. These environmental variables correspond to the disease risk factors. In practice, these risk factors are used to map the spatial, temporal or spatio-temporal disease risk prevalence. To achieve the aim of obtaining these models, ideally, there is the need for: (i) quality input from remote sensing data to represent the various hypothesized environmental risk factors, (ii) in-situ data of disease or vector prevalence, ideally collected on the field by trained professionals, (iii) automatic learning methods to find the underlying relationships between the disease dynamics and the risk factors, and (iv) domain expertise to understand the results of the modeling procedure and assess the quality of such models from an epidemiological standpoint.

With regard to automatic learning methods that are used to model the relationship between the disease/vector spread and the hypothesized environmental covariates, theoretically any technique that applies to multivariate prediction in other domains of remote sensing can be used. Most of these techniques belong to the classes of statistical models, machine learning and neural networks.

Since the early days of applying remote sensing data for landscape epidemiology, optical multispectral remote sensing data has dominated this area [32]. The Landsat programme of the National Aeronautics and Space Administration of the US government (NASA) is the longest-running EO mission for multispectral satellite image acquisition. The program was launched on July 23, 1972 [33]. Since the commencement of the Landsat programme, there has been many additional missions. This increased interest, financing, and applications for EO missions has led to advances resulting in higher resolution sensors and more freely accessible data [34]. In addition, and to complement other developments in space technology, rapid development in available electronics, computing power, data storage capabilities, and learning algorithms have enhanced the possibilities that can be reached with EO space mission investments. These development have rippled into and shaped the epidemiology use cases of EO data too, and still continues to do so.

As with other applications that have and continue to leverage EO data over the years, landscape epidemiology has evolved through the different times defined by the evolution of available data. Section 1.3.1 provides an overview of early applications of remote sensing data for epidemiological applications, while the following Section highlights the new trends and their promised improvements, as well as existing and known limitations.

1.3.1 Early applications of remote sensing data for Landscape Epidemiology

From the reviews of the use of EO data for epidemiological studies as presented in [32] and [19], the majority of the earliest studies in this domain utilized data from Landsat's Multispectral Scanner (MSS) and Thematic Mapper (TM) along with National Oceanic and Atmospheric Administration (NOAA)'s Advanced Very High Resolution Radiometer (AVHRR) optical multispectral sensors. There are also studies, though very few, that utilized data from France's Système Pour l'Observation de la Terre (SPOT) [35] sensors, which is a commercial mission with multispctral data available at up to $10\,\mathrm{m}$ spatial resolution. Some of the diseases that were considered in the most prominent earlier studies are Malaria, Eastern Equine Encephalomyelitis, Filariasis, Rift Valley Fever, Schistosomiasis, Trypanosomiasis, Lyme disease, Dracunculiasis, and Leishmaniasis [19, 32]. It is worth noting that the first four out of these nine diseases named here are transmitted by mosquito species. This goes to emphasize how mosquito-borne diseases have dominated this study from its early days.

AVHRR is an operational multispectral sensor on board the NOAA polar orbiting satellites. At any given time AVHRR is active on two satellites orbiting the Earth in opposite directions, allowing for total global coverage twice daily. It produces data at $1.1\,\mathrm{km}$ spatial resolution comprising five bands. Most early applications of EO data for epidemiology relied on AVHRR data information to model spatial and temporal variability of the required environmental variables.

At $1.1\,\mathrm{km}$ resolution with global coverage, AVHRR data lends itself to observing surface conditions at regional or continental scale. Its low spectral resolution however is its major drawback

The authors of Ref. [36] used the single day mean temperature of the AVHRR sensor data to model the village-level spatial prevalence of Bancroftian filariasis in the Southern Nile Delta area of Egypt at $10\,\mathrm{km}^2$ spatial resolution. The results show significant correlation between the AVHRR measured day temperature and filariasis prevalence.

Malaria, which is transmitted by anopheline mosquitoes, is also one of the earliest applications of EO data for infectious diseases spread monitoring. The authors of Ref. [37] extracted vegetation and temperature information obtained from AVHRR data over a period of 5 years (1990—1994) to predict maps of malaria seasonality in Kenya. That study showed significant correlation between vegetation condition (lagged by one month) and malaria admissions.

Landsat TM data has a temporal resolution of 16 days, spatial resolution of up to $30\,\mathrm{m}$ in about eight bands, making it suited for spatial modeling in landscape epidemiology. The study in [38] focuses on spatial modeling of mosquito larval ecology based on theoretical knowledge of the direct links that exist between the environment, mosquito larvae, and production of adult mosquitoes. To obtain environmental risk factors information, analysis of selected Landsat TM images of dry and wet season land surface information in April 1986 and October 1987 were analysed to map 16 land cover units. Landsat TM data provides major advantages for this study due to its high spatial resolution ($30\,\mathrm{m}$) in non-thermal bands, and the middle infrared bands (bands 5 or 7) which are sensitive to differences in plant and soil moisture.

1.3.2 New Earth observation sensors and missions: how much progress has been made with regards to urban epidemics monitoring?

While spatial analyses help to understand hotspot locations in terms of prevalence of disease drivers, the temporal dimension of the disease spread mechanisms and how they interplay with the environment has also been shown to be useful in this study domain, especially for local and regional studies. As revealed in [32], until the mid-2000s most attempts in landscape epidemiology did not consider the temporal dimension, and thus lacked critical components of seasonality in their models and results. Monitoring environmental variables in the temporal dimension requires time-series of overlapping EO images over the same scene. Quality images of this sort are either scarce or expensive during the periods of early exploration. While NOAA-AVHRR at that time presented the best possibilities for such analyses due to its 1.1km resolution (for vegetation monitoring indices) and high temporal resolution (twice daily), its limited coverage was a critical demerit.

A significant turning point for temporal analysis and possibility to map more diseases based on EO data came with the launch of NASA's Earth Observing System (EOS) mission. The Moderate Resolution Imaging Spectroradiometer (MODIS) is the key instrument onboard the EOS Terra satellite which was launched in December 1999. Another MODIS instrument was launched on the EOS Aqua satellite in 2002. MODIS provides substantial improvements in spatial resolution (up to 250 m), 36 spectral bands, 12-bit radiometric resolution in the visible, near, mid, and far infrared bands, and an enhanced set of preprocessed and freely available products at a global scale [39]. The MODIS vegetation and landcover suites have found extensive applications in landscape epidemiology especially because of its global coverage and temporal resolution. Some of the MODIS land products which have been used in landscape epidemiology studies are presented in Table 1.2.

TABLE 1.2: Details of MODIS data products (and derivable layers) used in temporal studies *Ae. aegypti* vector and diseases spread risks, and the proxy environmental variable they represent.

Data product	Band(s) used	Feature	Spatial resolution (meters)	Temporal resolution (days)	Proxy to:
MODIS MOD11A2	`LST_Day_1km`	Daytime LST [a]	1000	8	Maximum temperature
MODIS MOD11A2	`LST_Night_1km`	Night-time LST	1000	8	Minimum temperature
MODIS MOD13Q1	`NDVI` or `EVI`	NDVI [b] or EVI [c]	250	16	Vegetation condition
MODIS MOD13A3	`NDVI` or `EVI`	NDVI [b] or EVI [c]	1000	16	Vegetation condition
MODIS MOD13Q1	`sur_refl_b02` and `sur_refl_b07`	NDWI [d]	250	16	Surface moisture and humidity

[a] LST: Land surface temperature [40]
[b] NDVI: Normalized Difference Vegetation Index (NDVI)
[c] EVI: Enhanced Vegetation Index (EVI)
[d] NDWI: Normalized Difference Water Index (NDWI)

1.4 Modeling the distribution of *Ae. aegypti* mosquitoes using remote sensing data

Among all possible vectors for diseases, in this work, we are going to focus on *Ae. aegypti* mosquitoes. At the beginning of the doctoral study that has led to this dissertation, a systematic literature review was carried out to aggregate, process and analyse existing body of work in the domain of vector population and diseases risks modeling using EO data. The process involved the use of Google Scholar platform [41] to search for method-based studies in this domain. The main goal of the review was to analyse gaps in methods and input data quality that are potential high-impact contributions. After pre-selecting studies that qualify through the search criteria, they were summarized to extract information including EO data products used and their quality (spatial and temporal resolution), modeling methods applied, potential gaps in methods applied, and future direction proposed by each authors. The remaining paragraphs of this section communicate the information gathered from this review process, while Section 1.5 presents the challenges which were discovered and have been made the focus on this dissertation.

Female mosquitoes of this species are the mainly responsible for the spread of Dengue and other arboviruses, including Zika, Chikungunya. They breed in artificial containers in urban environments and poses significant public health threats in urban areas in more than 100 countries, which account for half of the world's population [42]. The disease transmission by *Ae. aegypti* depends on many factors such as: the susceptibility of the exposed population [43], socio-economic conditions [44], viral life cycle and, notably, the local vector density [45].

Entomological studies have shown that the oviposition, life cycle, and population dynamics of this mosquito species are significantly affected by abiotic and biotic environmental conditions including precipitation, temperature, vegetation condition, and humidity [46, 47]. Precipitation affects the volume of water in hatching containers, which, in turn, determines the population of hatched mosquitoes [42]. There is evidence in the literature that higher precipitation generally should result in a higher dengue risk [48]. Temperature affects the reproduction rates of *Ae. aegypti*, as well as the incubation period of the carried viruses. Higher temperatures shorten the extrinsic incubation period, accelerate vector development and, consequently, increase adult population and virus risk exposure for humans [49, 50]. Furthermore, vegetation canopy acts as a shield to protect hatching water containers from sunlight, reducing evaporation, decreasing sub-canopy wind speed and improving *Ae. aegypti* mosquito development [51]. Besides, higher humidity is associated with high Dengue virus propagation by *Ae. aegypti*, with effects also shown on the vector population [42, 52]. By obtaining features related to these environmental conditions from EO data and collecting field data on the vector population dynamics, it is possible to model the temporal and spatial distribution of *Ae. aegypti*, and the consequent disease risks exposure [42, 53, 54].

Major contributions focusing on modeling *Ae. aegypti* vector prevalence and/or the consequent risks of Zika, Dengue or Chikungunya are mostly analyzed with respect to either their spatial or temporal patterns. Spatial modeling is applied to map the geographical variability of the diseases risks in terms of breeding site suitability and/or disease prevalence hotspots over a geographical area of interest. Temporal modeling focuses on capturing the relationship between the vector/disease risk and the environmental risk factors across the difference climatic periods of the year, and using the modelled relationship to predict possible future events (forecasting). In both cases, the goal of the study determines the type of data that is used; local spatial modeling prioritizes spatial resolution and information with high spatial variability, while temporal modeling prioritizes temporal resolution and variability. A spatio-temporal model considers both dimensions.

One typical spatial modeling focused study is presented in [55]. That study uses environmental, entomological and demographic factors to spatially map the possible niches that contain breeding sites where *Ae. aegypti* lay their eggs. To represent the environmental content of the study area (Targatal, Argentina), an unsupervised land cover classification procedure (using the

k-means algorithm) was performed to identify seven land cover classes (bare soil, low vegetation (grass), high vegetation (trees), urban buildings, superficial water, shadows, pasture, and crops) at $10\,\mathrm{m}$ spatial resolution from SPOT 5 multispectral satellite image products. A Maximum Entropy prediction technique [56] was then used to map the ecological niche distribution by detecting non-random relationships between the georeferenced records of the presence of the species and a set of buffer images derived from the initially obtained land cover classes to characterize the environment under study. The resulting model showed that the environmental variables that best explain 75% of the distribution of *Ae. aegypti* breeding sites were: the percentage of bare soil (44.9%), the percentage of urbanization (13.5%), and the water distribution (11.6%). While the study provides a good baseline with regards to spatial modeling of *Ae. aegypti* distribution using EO data, its results are expensive to reproduce because, as earlier mentioned, SPOT satellite data are not freely accessible.

Another study published in [57] explores the spatial modeling of dengue incidence for the same city of Targatal, Argentina. The study is based on the hypothesis that the spatial pattern of dengue outbreak is a cooperative result of multiple factors including environmental, demographic, entomological and epidemiological factors. As outlined by the study authors, these hypothesized factors can also be grouped into micro-scale (e.g. mosquito breeding sites), medium-scale (e.g. housing conditions) and macro-scale effects (e.g. vegetation, temperature, rivers, etc.). For the study experiments, Landsat 5 TM data was used to obtain macro-scale habitat environmental effects. The environmental descriptors which were extracted from the data are: distance to main streets and roads, distances to river, distance to vegetation, tasselled cap brightness, tasselled cap greenness, tasselled cap wetness, and Landsat 5 bands 1—-7. Tasselled cap greenness, brightness and wetness are feature layers obtained by performing a Tasselled cap transformation [58] on the the input multi-spectral image data. These layers provide proxy information to soil brightness, vegetation and soil status (including humidity), respectively. The baseline information classes which were used for the "distance to" environmental descriptors (i.e. roads, rivers and vegetation classes) were derived by visual interpretation and maximum likelihood classification. Visual inspection and threshold was then used to construct a decision tree using a total of 487 suspected geo-referenced dengue cases. The resulting model was used to obtain the spatial risk map of the disease spread. The obtained map provides a platform to discuss some of the potentials and drawbacks of remote sensing data for epidemiological modeling. In general, the authors concluded that the local spatial variation of dengue risks might not be fully explainable by macro-scale effects which are obtainable from remote sensing data layers. The authors also acknowledge the need for model explanability (or "interpretability") in such a way that the relationship between the dengue prevalence and each considered environmental factor can be better understood and the most influential predictor variables can be known. Another drawback of that is the manual modeling of the decision tree based prediction

modeling, hence its lack of scalability. To mitigate this particular drawback, machine learning techniques may be applied.

Another relevant study which is focused on $5\,\mathrm{km}$ resolution global mapping of environmental suitability Zika virus is presented in [42]. That study uses a boosted regression tree algorithm [59] to establish a multivariate empirical relationship between probability of Zika virus and environmental conditions in locations where confirmed cases of the disease has been reported. Due to the global scale of the study, socio-economical information was considered alongside other hypothesised environmental variables of effect. While gridded meteorological data from WorldClim data [60] were sufficient to obtain temperature, humidity, and precipitation records for the global scale modeling task, vegetation information was obtained from the MODIS EVI (MODIS 13A3) [61] data product. Also, the MODIS collection 5 (C5) [62] land cover product was used to augment the urban growth rates data provided by the United Nations Population Division [63]. The resulting model of that study predicted high levels of risk in many areas within the tropical and sub-tropical zones. Also, large portions of the Americas were predicted to be suitable for transmission. The highest risk areas were predicted to be concentrated in Brazil, followed by Venezuela and Colombia. A major takeaway from this study from a remote sensing standpoint is that in global scale studies which are at very coarse resolution, though meteorological data can be informative, remote sensing data (in this case MODIS) still find useful application.

For a global scale disease or vector niche spatial mapping, it might be difficult to quantify and rank the influence of each considered environmental variable on the prediction output and rank the most important variables, because their importance might change from one location to another depending on medium-scale effects and local climate variability pattern. While predictions at global scale help to map the variability of risk across different climate zones, they do not provide information for local authorities to perform further studies and understand which variables affect more the results.

In general, to a large extent, apart from data constraints, the scale of the analysis (municipal, national or global) and the domain expertise in terms of personnel leveraged by each study determines the kinds of variables that are used in spatial models of *Ae. aegypti* and, thus, its resultant disease distribution. This is why the explanation of the effect of the different variables is important, since it can help to provide hints on which environmental variables are actually relevant. While global or national spatial models may well depend on macro-scale effects [42], local (municipality-level) spatial risk mapping requires information that varies at smaller scales, i.e. that follows or is affected by micro-scale effects [55]. As a result, for local scale tasks there is the need of new methods with higher quality inputs from new freely available sensor data with fine spatial resolution, such as the European Sentinel constellation multispectral and radar data, instead of the previously used proprietary and expensive SPOT data [55].

Moreover, as mentioned above, solely spatial models tend to obscure seasonal information which is essential to capture the temporal dimension of the disease or vector dynamics. Hence, temporal models require more attention as well. A study that combines spatial and temporal modeling (spatio-temporal) is presented in. [64]. The authors developed a national scale modeling of dengue in Brazil from 2010 to 2017. NDVI, NDWI, and NBR spectral indices were obtained from Landsat 5, 7, 8 and Sentinel-2 images covering Brazil weekly over the period of 7 years. For each observation week, EO data pixel values and in-situ dengue incidence data were aggregated up to the municipality level. This aggregation helps to deal with missing pixels data due to cloud cover and other distortions. Distributed logarithmic non-linear models were used to fit the relationships between the municipality level dengue incidence values and the hypothesized explanatory covariates (NDVI, NDWI and NBR). The distributed lag non-linear model [65] allows for model explanation in terms of how a particular environmental variable contributes to the dengue prevalence as either a leading or lagging variable. The results of the study show that if there is a high NDVI value, then there will be an uptick in dengue incidence in 5 weeks, i.e NDVI has a 5-week lagging effect on dengue prevalence in Brazil. This kind of lagging effect and variable power quantification is useful in temporal studies as it helps public health players to better plan vector/disease control activities at urban and national levels.

Temporal prediction models of *Ae. aegypti* population or dengue risks use past and present prevalence information to predict the future. To achieve this, they use environmental information of temporal variability that affect the *Ae. aegypti* dynamics. State-of-the-art studies that focus solely on temporal modeling of *Ae. aegypti* or the consequent disease distribution at local scale (municipal/regional) use MODIS data products. This is because of the daily temporal resolution of the data product. In practice, since most temporal modeling studies are conducted on weekly cycles and optical data are subjected to atmospheric distortions such as clouds, weekly composite MODIS data products that combine multiple acquisitions to remove clouds and other noise are usually sufficient. Advanced MODIS specifications ensure that the environmental variables of interest (temperature, vegetation condition, humidity, etc) can be obtained at finer spatial resolution than with AVHRR data, and made available freely as ready-made composite products that are atmospherically corrected and georegistered. The major MODIS data products and derivable layers which are are used in temporal studies of the *Ae. aegypti* are presented in Table 1.2 [53, 54, 66, 67].

The prediction algorithms (either machine learning or statistical models), model selection, and variable importance ranking (model explanation) methods form major parts of *Ae. aegypti* studies that propose methodologies for temporal modeling [53, 54, 66, 67]. To remove redundancy among candidate covariates, drop irrelevant variables, and improve model quality, model selection is often performed. We can find a framework with all these parts in [66]. That study introduced the use of the MODIS product suite for temporal forecasting of *Ae. aegypti* oviposition activity. Specifically, temporal series of NDVI and diurnal (day and night) land surface

temperature (LST) were used to forecast *Ae. aegypti* oviposition using data from October 2005 to September 2007 (2 years) in San Ramón de la Nueva Oran city of Argentina. As the authors of the study explained, NDVI was used because it provides proxy information to humidity and precipitation, while LST gives an approximation of the environmental temperature. The specific MODIS data products used for the study are MOD13Q1 NDVI and MOD11A2 LST products. Details of these data products can be found in Table 1.2. Two linear models (LM) were fitted using: (i) environmental variables with time lag up to 24 weeks, and (ii) environmental variables without time lagged variables. The linear correlation coefficient was used for model selection, to remove multicollinearity among input prediction variables, and to improve the quality of the models. The best model of that study was obtained without including lagged effects. Also, the resulting best model was tested using recent MODIS obtained environmental variables from July 2014 to January 2016. The results show that the model developed using data from 2005 to 2007 was able to predict a potential outbreak around January 2015. However, while pairwise linear correlations are easy to compute for model selection purposes, it only considers linear relationships among variables, and only in a pairwise mode, ignoring complementary and nonlinear relationships among covariates. Also, linear regression which was applied as the prediction algorithm in that study has the major advantage of being explicitly explainable with meaningful coefficients that signal variable importance and direction of effect (positive or negative) on the modelled oviposition activity. However, its assumption of linearity in the relationship between *Ae. aegypti* oviposition activity and the surrounding environment is a strong bias that makes it prone to underfitting.

Another study presented in [53] is based on the method presented based on the sole use of MODIS data product in [66]. The study presents the temporal modeling of oviposition activity of *Ae. aegypti* vector in Targatal city (Argentina) for four years (August 2011 to July 2015). The response variable was obtained from data collected with 50 or 100 ovitraps, randomly placed around the city during the observation period. In addition to NDVI and LST variables which were estimated also from the MOD13Q1 and MOD11A2 MODIS products, other covariates which were considered in the study are NDWI and precipitation. The NDWI variable layers were included as proxy to humidity. Local precipitation information was obtained from Tropical Rain Measurement Mission (TRMM) data [68] and the Global Precipitation Mission (GPM) [69]. TRMM is a joint mission of NASA and the Japan Aerospace Exploration Agency which uses several instruments including radar, microwave imaging, and lightning sensors to detect rainfall. Since TRMM only provided data until June 2015, GPM products were used to ensure continuity of precipitation information availability. Each resulting environmental variable feature was considered with up to three weeks of lag to represent asynchronous effects. LM was used to fit the relationship between the vector oviposition activities and the chosen environmental variables. Model selection was performed using manual forward analysis by testing the effect of each variable and the significance of including it into the model. From the resulting

LM equation, it was seen that low temperatures inhibit oviposition activity in the location of study. NDWI, on the other hand, showed positive effect on *Ae. aegypti* oviposition activity. As revealed by that study, beyond good prediction quality, the ability to explain a disease spread or vector population model (i.e being able to rank the effects of each predictor variable) provides valuable information for public health authorities. One drawback of this study is that the model selection approach, being manual, does not scale well.

To address the bottleneck in prediction quality by LM, the authors of [54] compared different machine learning models and statistical models for the task of temporal modeling of *Ae. aegypti* oviposition. That study juxtaposed the results obtained with LM in [53] (as discussed above) with nonlinear models (e.g., the generalised linear model - GLM), and machine learning techniques (Support Vector Regression - SVR, multilayer perceptron - MLP, k-nearest neighbors - KNN, and decision trees regression - DTR). Exactly the same data (MODIS and TRMM/GPM) for the same study area (Targatal) as described in [53] was considered, including the use of lagged variables up to 3 weeks. Since a manual forward selection would have been computationally burdensome in this case involving multiple complex models, the linear correlation coefficient was used for model selection. The results show that nonlinear models perform better than the linear regression for vector oviposition activity prediction, thus making a case against LM which had dominated previous studies [53, 66]. Specifically, MLP, KNN and SVM improve the resulting temporal models of vector oviposition activity. Unlike LM and GLM however, MLP, KNN and SVM work as black box models and thus the contribution of each covariate included in the prediction cannot be easily inferred.

Table 1.3 presents an overview (RS data products usage) for *Ae. aegypti* population modeling. From this table and the summary presented in this section, it seems that the spatial or temporal dimension and the geographical scale of the study area influence which input environmental variables and RS data sources are best used as model covariates.

1.5 Challenges and objectives

From the overview of the previous sections, the following challenges emerge and have been selected as the main scientific objectives of this thesis.

- Spatial mapping models largely include buffer layers (e.g. “distance to” layers) which are obtained from landcover classification. In this sense, SPOT dataset provided the best spatial resolution among the studies which has been reviewed. This dataset however are not freely available for scientific investigation purposes. Also, none of the reviewed studies considered implementation of a fully optimised landcover classification — defining more classes and using more robust classification techniques to ensure that the derived

TABLE 1.3: Featured studies that apply RS data sources for spatial and/or temporal *Ae. aegypti* and related disease risks modeling. Similar studies using the same sensor data have been combined together in this table.

Study	Sensor (data product)	Environmental variable	Feature layer	Spatial resolution	Temporal resolution
[42]	MODIS (MOD13Q1)	Vegetation	EVI	250 m	16 days
	MODIS (MOD44W)	Human presence	Urban extent	250 m	Annual
[54, 66, 67]	MODIS (MOD13Q1)	Vegetation	NDVI	250 m	16 days
	MODIS (MOD13Q1)	Humidity	NDWI	250 m	16 days
	MODIS (MOD11A2)	Temperature	LST	1000 m	8 days
	TRMM	Precipitation	Precipitation	0.1 deg	Daily
[57, 70]	Landsat TM 5	Soil moisture	Distance to water	30 m	16 days
		Vegetation	Distance to vegetation		
		Soil moisture	Tasseled cap brightness		
		Humidity	Tasseled cap wetness		
		Vegetation	Tasseled cap greenness		
		Vegetation	NDVI		
		Land surface	Bands 1—7		
[55]	SPOT 5	Land cover	Classification map	10 m	5 days
		Land surface	Spectral feature		
		Temperature	NBRT		
[71]	MODIS (MOD13Q1)	Vegetation	NDWI	250 m	16 days
	MODIS (MOD13Q1)	Humidity	NDWI	250 m	16 days
	DMSP-OLS	Human presence	Night-time light	1000 m	Annual
[72]	Landsat 7 ETM+	Humidity	Tasseled cap brightness)	30 m	16 days
		Vegetation	Tasseled cap greenness		
		Humidity	Tasseled cap wetness		

layers are of high quality with respect to the ground truth. Ref. [55] used an unsupervised technique (k-means) for land cover classification, while [57] used a maximum likelihood approach. From a remote sensing standpoint, more novel approaches can be used to improve the quality of the models. Sentinel-2 fits the bill for such case since it brings a 10 m spatial resolution and a global coverage, enabling the possibility to obtain in wide geographical areas the same level of details provided by the use of SPOT data in local scale risk mapping projects. Additionally, Sentinel-2 data provides even better spectral and temporal resolutions which can be used to recognize more land covers, and achieve, for instance, a more accurate map of urban vegetation. This thesis makes a contribution in this direction.

- With regards to temporal modeling, the reviewed literature shows that there is a trade-off between model quality and explainability. Explanability in this domain is important because public health management agencies may want to use such models to improve their understanding of local disease spread dynamics. Linear models and generalized linear models introduce intuitive equations with which the effects of the environment variables can be explicitly ranked and explained. However, these models do not provide the best quality of prediction. On the contrary, non-linear machine learning methods such as neural (deep) networks and Support Vector Machines provide better risk prediction quality,

but work as black boxes that cannot be easily explained [54]. As a result, there is the need to develop Ae. aegypti vector dynamics prediction frameworks – based on machine learning or statistical models — that offer a combination of high prediction quality and good explainability. Also, to improve the use of the linear correlation coefficient or the manual forward selection which have been used so far for model selection in state-of-the-art studies, any new proposed method should be able to handle model selection in an automatic (or quasi-automatic) way, and also to consider nonlinear relationships among covariate features. To both these points this thesis is meant to provide a significant contribution.

- In the reviewed state-of-the-art studies in *Ae. aegypti* modeling, we have mentioned spatial mapping at municipality and global scale using macro-scale effects, temporal modeling at municipality level, and spatio-temporal modeling at national scale. In the hypothetical case of local authorities needing to optimise vector control activities within at sub-municipal (i.e., block) level, current approaches in literature only provide spatial models. Consequently, there is the need for sub-municipality-level spatio-temporal models that capture the variability of vector population or disease risks across neighbourhoods and different micro-climatic zones. This is also a point that is going to be addressed in one of the chapters of this thesis.

- Finally, on the issue of temporal modeling *Ae. aegypti* population and the consequent disease risk based on environmental variables estimated from remote sensing data, there are other areas with possibilities for incremental improvements. One of them is the modeling technique used. As a first attempt in this direction, in this thesis a deep learning approach will be introduced.

1.6 Specific contributions

Towards addressing the problems and objectives already stated, this dissertation makes the following specific contributions:

- A high resolution urban vegetation mapping procedure with Sentinel-2 data. This is targeted towards providing a methodology to derive qualitative inputs for spatial vector and diseases modeling.

- An accurate and explainable machine learning approach for EO-based temporal population modeling of *Ae aegypti* population at municipality level. This particular contribution features a Random forest (RF) regression along with its built-in quantitative measure of the variable importance (MDI) which was used to extract and rank the most informative

environmental features which impact the mosquito population. Operationally, such ranking information can help investigators to understand major risk driving mechanisms in different locations and seasons.

- A robust statistical regression approach for ***Ae aegypti*** vector population modeling using EO data. This thesis has proposed a Weighted Poisson GLM approach, which is able to achieve machine learning (ML) quality results, while also providing the capability to explicitly interpret the causality in the model. The contribution here is that principal investigators, using this method, can have both a qualitative model and an intuitive equation with which they can explain the environmental effects without the need of a relations curves which are less easy to interpret.

- A spatio-temporal forecast epidemiological modeling methodology based on Recurrent Neural Network (RNN). This contribution focuses on sub-municipality-level vector population forecasting so as to capture different behaviours of the vector distribution across different neighborhoods.

1.7 Dissertation organisation

As already briefly motivated above, multispectral EO data which are becoming increasingly freely available provide the possibility to model the prevalence of epidemiological spread and vector development in space and time. However, the resulting models need to be explainable as much as possible in order to serve as empirical guide for vector control actions. This thesis explores both spatial and temporal epidemiological modeling from the standpoints of EO data input quality, modelling techniques, and empirical explainability of the resulting models. Accordingly, the contributions are organised into six chapters.

This chapter has introduced the objectives of the thesis, as well as provided a short review of previous works that has been performed with regards to using multispectral EO data for epidemiological modeling and monitoring purposes.

Chapter 2 targets the possibility to improve the quality of inputs for urban level spatial modeling of *Ae. aegypti* population and its causal disease prevalence in urban areas around the world. Accordingly, the chapter presents a novel method for mapping different vegetation types in urban areas at $10\,\mathrm{m}$ spatial resolution using Sentinel-2 data.

Chapter 3 introduces a framework for modeling the temporal distribution of *Ae. aegypti* at municipality level in an urban area. This framework combines EO-based environmental features extracted from NDVI, NDWI, LST, and precipitation measurements as inputs to a random forest ML algorithm. RF's embedded nonlinear features importance ranking, based on mean decrease

impurity (MDI), is applied to rank the environmental variables and select the most informative environmental features, thus providing explainability. This explanability is explored in practice through relations curves which are introduced in the chapter.

To reduce the computational burden and the domain expertise required to explain EO-based *Ae. aegypti* temporal models by the use of MDI , Chapter 4 introduces a weighted GLM technique for the same temporal modeling task as in Chapter 3. Indeed, GLMs provide model equations which are intuitive to understand and provide hints on the effects of each variable. The weights of the GLM serve here to improve the model quality towards something comparable to the RF model, while retaining the model nas usefulness for application to disease control actions in urban areas.

Chapter 5 introduces a time series *Ae. aegypti* population forecast (one-week-ahead) which is spatially disaggregated at the neighborhood level. This proposed technique leverages the prediction quality of Recurrent Neural Networks (RNNs) which take as inputs MODIS and GPM datasets to represent the spatio-temporal vegetation, humidity, temperature and precipitation conditions.

Finally, Chapter 6 provides a comprehensive review of the contributions of the previous chapters and concludes the thesis by summarizing its achievements and highlighting possible future research paths.

Chapter 2

Mapping urban vegetation with Sentinel-2 data

2.1 Introduction[1]

Urban green spaces impact the urban ecosystem from the standpoints of health and quality of life. Vegetation may vary in height, sizes, canopy, and species, with each of these variations resulting in a different environmental impact. Major parts of urban areas where vegetations are often found include parks, Government reserved areas, river banks, domestic gardens, and street trees. Their impacts include pollution removal, noise attenuation, wind storm control, temperature reduction, ground water replenishment, recreation for citizens, and epidemiological effects. It is therefore important to study the quality and presence of vegetation in urban areas. Urban vegetation maps are needed in various applications and studies to improve the lives of urban dwellers [73].

With regards to landscape epidemiology, as introduced in Chapter 1, vegetation types and structures affect the development and consequent density of *Ae. aegypti* mosquito species. As a result, studies on the geographical variability of this mosquito vector population in terms of breeding site suitability and/or disease prevalence hot spots use vegetation maps to obtain vital covariate input layers. As stated, previous studies in this domain have used either Landsat or SPOT data. While Landsat does not provide enough spatial and temporal resolution for high-quality urban vegetation mapping, SPOT data are not available for free.

Sub-pixel vegetation classification approaches have been developed to address the issue of insufficient spatial resolution in Landsat data. The authors of Ref. [74] explored different approaches

[1]This chapter has been published as a standalone paper as O. Mudele, P. Gamba "Mapping vegetation in urban areas using Sentinel-2", in the Proc. of the 2019 Joint Urban Remote Sensing Event (JURSE2019), Vannes (France), 2019, unformatted CD-ROM, doi: 10.1109/JURSE.2019.8809019.

to extract information on urban vegetation abundance using single time point Landsat ETM+ data. In general, medium resolution RS data may produce low mapping accuracy values in urban areas if subjected to pixel-level classification methodologies. In highly heterogeneous urban surfaces, the medium resolution data like Landsat ETM+ produce mixed pixels which are comprised of multiple land cover types. As a result, the authors of that study considered the use of sub-pixel level methods. Such methods apply spectral unmixing techniques [75] to address the problem of mixed pixels. In [74], three spectral unmixing approaches: linear regression analysis, linear spectral unmixing and multi-layer perceptrons, were compared to delineate vegetations from Landsat ETM+ data in the European city of Brussels. The results show that the approaches considered work better when the predictions are spatially aggregated to neighbourhood levels. While unmixing approaches do provide means to work with medium resolution data for urban areas vegetation analysis, the weakness of this class of methods is that the unmixing algorithms make assumptions about how the different land cover classes mix into pixels. For example, linear spectral unmixing, which is the most commonly used approach, assumes that the spectral signature of a mixed pixel is the weighted linear combination of all its component land cover classes (endmembers). Also, other authors found that vegetation estimates derived from spectral mixture modelling appear less sensitive to background soil reflectance [76].

To avoid the bottleneck of spectral unmixing approaches, other studies have considered the use of very high resolution (sub-meter) multispectral data for urban vegetation mapping. Some of these studies also integrate vegetation phenology by considering multi-temporal combinations of such RS data. In the study presented in [77], multi-temporal WorldView-2 images were applied for the purpose of mapping vegetation functional types in urban areas. WorldView-2 data has eight multispectral bands with a spatial resolution of $1.84\,\mathrm{m}$. Obviously, at this resolution, it is possible to delineate heterogeneous properties in urban surfaces. Moreover, pixel and sub-pixel-based classification methods can produce spurious pixels (classification noise) when applied to high resolution imagery like WorldView-2. As a result, an object-based classification approach which considers spatial neighbourhoods was applied in that study. The results showed that by considering vegetation phenology through the incorporation of overlapping multi-temporal scenes, the classification overall accuracy improved from 82.3% (without phenology) to 91.1%. The bottleneck with reproducing these results however is that WorldView-2 imagery are not available for free.

To counter the need for unmixing-based approaches as presented in [74] on freely available Landsat imagery and the need to acquire commercial WorldView-2 images for the task of urban vegetation mapping, this study presents a procedure to extract urban vegetation types from Sentinel-2 (S-2) multispectral data by combining multispectral and multi-temporal information. The presented method is based on the use of multi-temporal (seasonally aggregated) Normalized Difference Spectral Vector (NDSV) features to improve vegetation classes separability. For this

purpose, a robust classifier - Random Forest (RF) - was applied and compared with Classification and Regression Trees (CART) and Support Vector Machines (SVM).

Quantitatively, the study results show that RF performs 6% better than SVM which is the second best performing classifier, and seasonal aggregation quantitatively reduces the confusion between the vegetation classes of interest. In addition, qualitative analyses of the resulting vegetation maps show that the NDSV feature provides better separability of vegetation classes.

From a vector population modeling point of view, the method proposed in this chapter can be applied to generate vegetation maps which can be used as input to achieve modeling objectives, and possibly to explain spatial variations in vector population concentration based on types and diversity of vegetation in urban areas.

Section 2.2 presents an introduction to S-2 data. The methodology is presented in Section 2.3. The experimental procedure and results are then presented in Sections 2.4.2 and 2.4.3.

2.2 Sentinel-2 data

The Sentinel program was designed and initiated to meet specific needs of the European Space Agency's (ESA) Copernicus program. The goal of this program is to replace older EO missions and ensure continuity of data availability while also enhancing the quality of available data. Among the five missions contained in the Sentinel program, Sentinel-1 (C-band radar instrument) and S-2 optical imaging provide land surface information and can be used for vegetation monitoring purposes. Unlike the Landsat program which is composed only of multi-spectral optical imaging instruments, the Sentinel program leverages both multi-spectral optical imaging and radar instrument to provide information on land, ocean and the atmosphere. This ensures that the program can deliver the advantages across both technologies, optical and radar.

Sentinel-2 (S-2) is a high-resolution multi-spectral imaging mission with two satellites flying in the same orbit but with a phase difference of $180°$ to provide a high temporal resolution of about 5 days. Each satellite carry an optical instrument payload that samples 13 spectral bands: four at $10\,\mathrm{m}$, six at $20\,\mathrm{m}$, and three bands at $60\,\mathrm{m}$. The S-2 satellite weighs approximately 1.2 tonnes and has lifespan of about 7 years [78]. A poster of Sentinel-2 satellite is presented in Figure 2.1 while Table 2.1 presents a summary of the bands present in Sentinel-2 image.

2.3 Methodology

To mitigate errors in the data due to atmosphere and sensor acquisition parameters, there is the need for data calibration, co-registration of overlapping scenes, and cloud cover filtering to

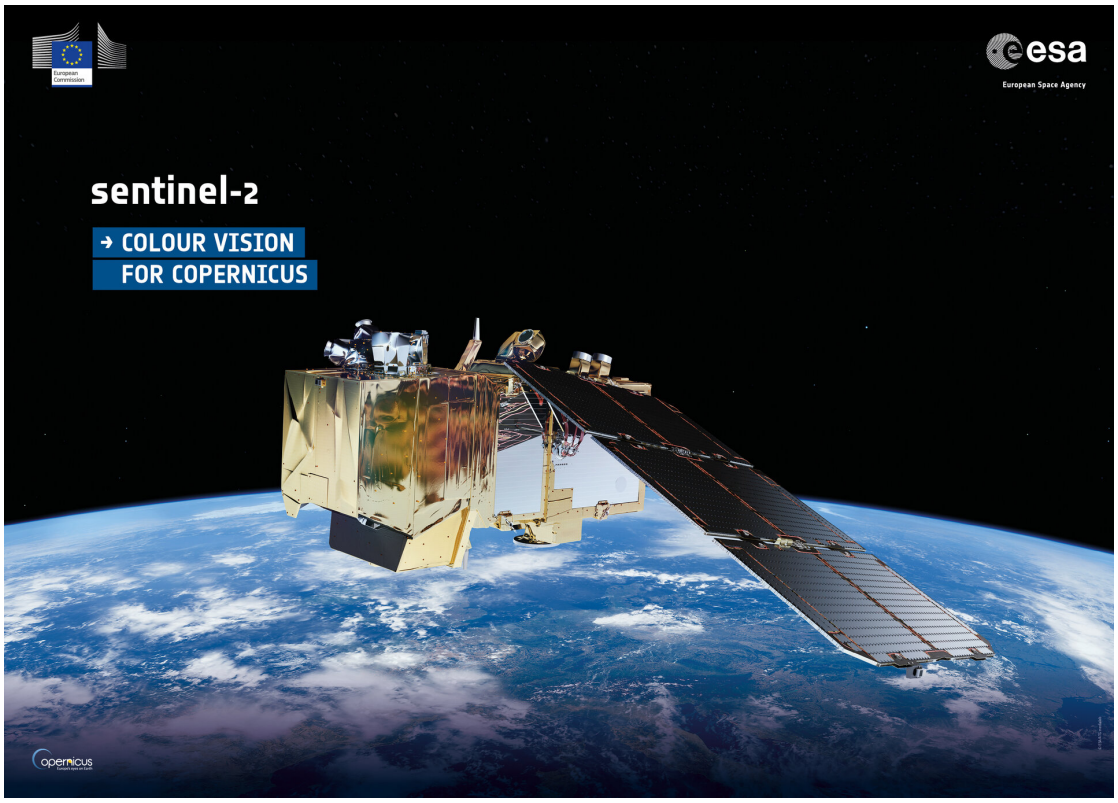


FIGURE 2.1: Sentinel-2 satellite poster (From [79])

TABLE 2.1: Spectral bands description of Sentinel 2 image

Spectral bands	Wavelength (μm)	Resolution (m)
Band 1 – Aerosols	0.443	60
Band 2 – Blue	0.490	10
Band 3 – Green	0.580	10
Band 4 – Red	0.665	10
Band 5 – Red Edge 1	0.705	20
Band 6 – Red Edge 2	0.740	20
Band 7 – Red Edge 3	0.783	20
Band 8 – Near Infrared (NIR)	0.842	10
Band 8A – Red Edge 4	0.865	20
Band 9 – Water vapor	0.940	60
Band 10 – Cirrus 2	1.375	60
Band 11 – Short Wave Infrared (SWIR) 1	1.610	20
Band 12 – Short Wave Infrared (SWIR) 2	2.190	20

select images suitable for the procedure. These steps were already implemented in the cloud computing platform used in this work (Google Earth Engine [80]). To mitigate other forms of error due to residual cloud covers and other errors in single acquisitions, multi-temporal overlapping scene combination (in case of using multiple overlapping data) and/or selection (in case of using single image input) was performed. For scene combination, the greenest pixel compositing method was applied [6]. The output of this method is a mosaic layer constituting pixels

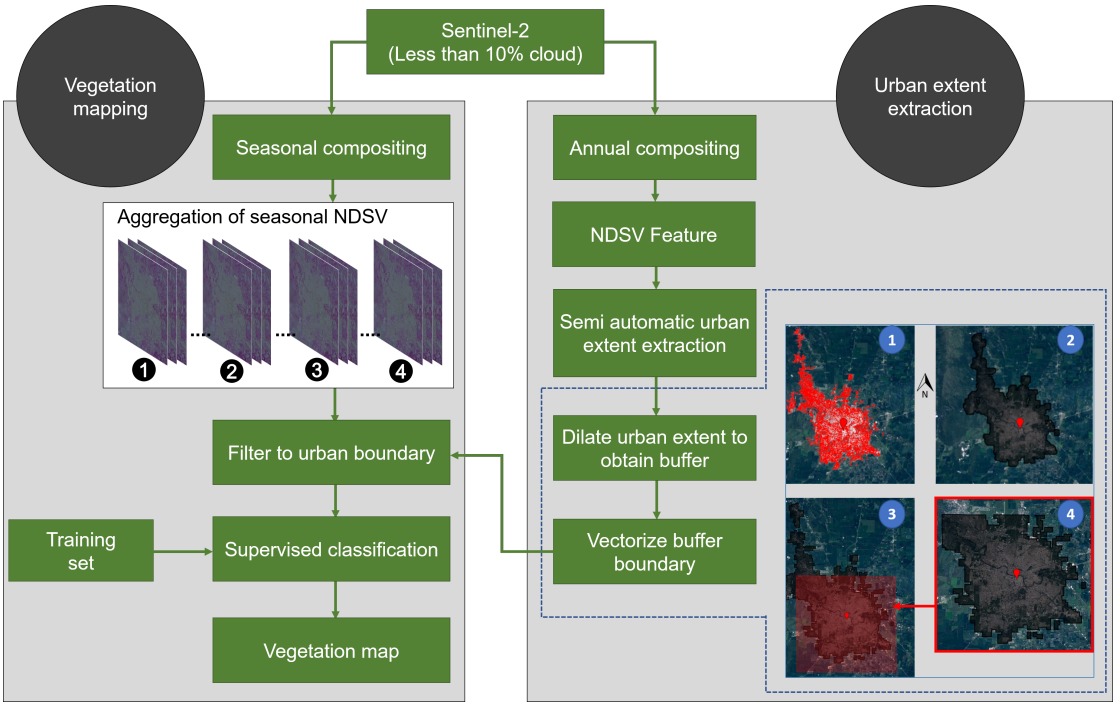


FIGURE 2.2: Proposed methodology for vegetation mapping with Sentinel-2 data. The framework is divided into two main blocks: urban extent extraction (explained in Section 2.3.3); and vegetation mapping (explained in Section 2.3.4)

.

with the highest NDVI value among all images in the considered multi-temporal stack. The resulting greenest pixel layer is useful to discriminate between artificial surfaces and vegetation types, thus making it effective for the specific goals of this study.

To map vegetation types within an urban area, it is important to define the urban boundary. The procedure implemented for this work includes an initial semi-automatic urban area extraction in the area of interest. Then, a buffer zone around the extracted urban extent is obtained and used as geographical bound for the urban vegetation mapping procedure. A full description of the processing chain is shown in Figure 2.2. The urban extent delineation procedure is described in Section 2.3.3 and the subsequent vegetation mapping using the obtained urban areas bound is described in Section 2.3.4. Both sub-methods (urban extent extraction and vegetation mapping) are based on the normalized difference spectral vector (NDSV) feature space. Hence, the NDSV feature is introduced in Section 2.3.1.

2.3.1 Normalized Difference Spectral Vector (NDSV)

NDSV was introduced in [81] as a feature space for effectively mapping urban areas from satellite images by leveraging all information contained in the bands of the multi-spectral data, thereby reducing errors and ambiguities occurring as a result of differences in acquisition mode, space, time, etc.

As already introduced in Chapter 1, there are a wide range of indices in the remote sensing literature which have been developed for different land cover discrimination purpose. Some of such indices (NDVI and NDWI) have been introduced in Table 1.1. Generally, a normalised difference index is defines as presented in Equation 2.1.

$$f(B_i, B_j) = \frac{B_i - B_j}{B_i + B_j}, \tag{2.1}$$

where B_i and B_j are two generic bands. Differently from a single normalized difference index where a single value is computed for every pixel, NDSV creates a vector of values on every pixel. By applying the NDSV transform to satellite images, we produce data that are inherently normalized and globally consistent. Also, the correlation across all individual indices that make-up the vector will provide all the dimensions needed to analyze the individual contribution of different features in urban areas (e.g. Low density human settlements mixed with green plants, asphalt surfaces that are not built-up, etc.). The vector space also gives us more reinforcement points in class probability estimation, compensating for the drawbacks of every single index with the other ones.

The NDSV for an n-band multi-spectral image can be computed for each pixel as shown in Equation 2.2:

$$NDSV = \begin{bmatrix} f(B_1, B_2) \\ f(B_1, B_3) \\ f(B_1, B_4) \\ \vdots \\ f(B_2, B_4) \\ f(B_2, B_6) \\ \vdots \\ f(B_{n-1}, B_n) \end{bmatrix}, \tag{2.2}$$

2.3.2 Brief introduction of machine learning algorithms used

The machine learning methods used in [82] and [6] for NDSV-based urban mapping have been chosen for this study. They are:

- Classification and regression trees (CART): a decision tree obtained via recursive partitioning of the data space and fitting a simple prediction model within each obtained partition [83].

- Support vector machines (SVM): a learning technique that classifies all the data points by defining a hyper-plane to separate all classes. Hyper-plane selection is done by maximizing distances between nearest data points in each classes and hyper-plane. This distance is referred to as "Margin". Compared to decision trees based methods (Random forest and CART), it is more computationally intensive and takes more processing time [84].

- Random Forest (RF): an ensemble decision tree learning and classification model. Ensemble learning is a divide-and-conquer approach to improve classification performance. In this sense, RF combines decision trees to produce a more robust classification result [85]. RF has an advantage over CART in that it combines multiple versions of CART-like decision trees to create a more robust model. Also, it shows robustness with respect to the quality of training samples, as well as against sample imbalance and overfitting [86, 87].

In [82], SVM and CART have been applied to produce high quality urban maps based on NDSV input obtained from Landsat 5 and 7 data. SVM, CART and RF classifiers also showed quality results in the study presented in [6] which validates NDSV-based urban mapping across three different locations: Brazil; South East China; and Indonesia.

The robustness of RF in sample imbalance scenarios is of particular interest for this study since the training samples which have been collected are unbalanced across classes. As a result, RF was selected as the main classifier for this work. To justify this choice, its results were benchmarked against that of CART and SVM.

2.3.3 Urban extent extraction

A semi-supervised urban extent extraction procedure proposed in [6] which is based on the normalized difference spectral vector was used in this work. Our implementation of this procedure takes a single Sentinel-2 data as input (annual greenest pixel composite), applies the NDSV transformation and selects training points automatically from the urban class of the GlobCover product [88]; a coarser (300m spatial resolution) global land cover map. A simple CART classifier is trained using the automatically selected training points. The output map is then spatially regularized using a morphological filter to remove pixel-level classification noise. Specifically, here, a "close" operator with square kernel of $1\,\mathrm{km}$ width is applied.

Since the study interest is the boundary of the urban area plus its fringes, a buffer around the resulting map is obtained using a dilate morphological operation. The boundary of this buffer

is vectorized to produce the final urban area plus fringes boundary. An example for the city of Cordoba (Argentina) is shown in Figure 2.3.

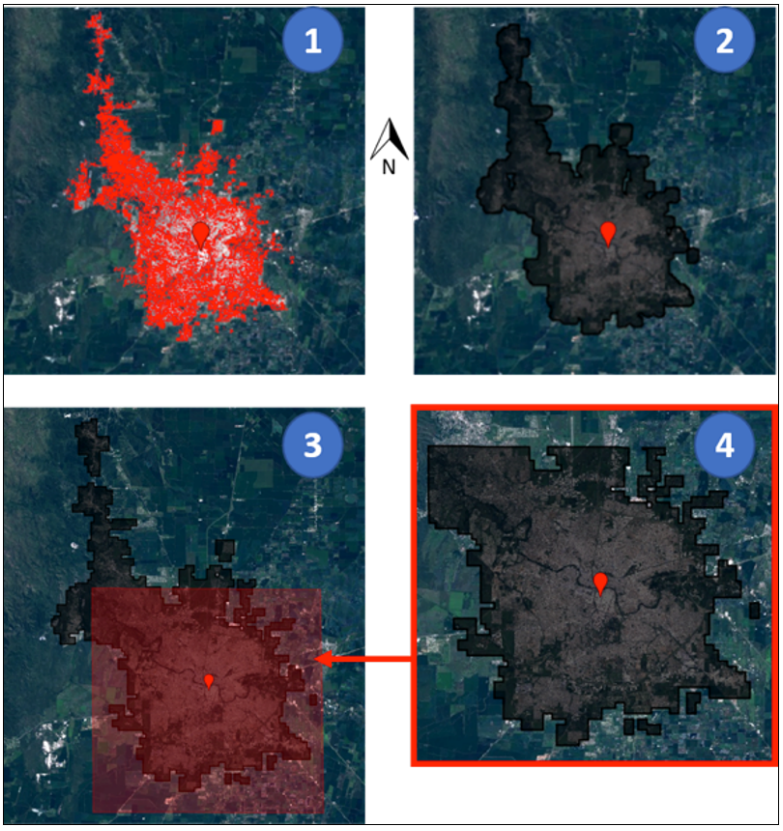


FIGURE 2.3: Semi-automatic urban mapping and buffer mask extraction procedure. Steps are as follow: 1- extraction of urban extents; 2- buffer application using a morphological dilation; 3- sub-area extraction; 4- final result.

2.3.4 Vegetation mapping

Vegetation mapping is performed inside the urban and peri-urban areas within the bound geometry obtained for the urban extent extraction stage. This task is performed by means of supervised classification. As with every supervised procedure, the final map quality depends on the selection of the representative samples used to train the classifier. These points have been mostly collected by field visit, but augmented by looking at high-resolution aerial images. Constraints considered in selecting the classification legend include spatial scale of data, season of the year when points are collected and, more importantly, the goals of the work. Based on these constraints, the following classes were defined: Trees, Bushes, Grass, Water, Artificial surfaces, and Bare soil.

To obtain a single image for classification input, scene combination was performed annually, seasonally and bi-monthly. There are two major motivations for considering seasonal/bi-monthly aggregation. First, vegetation behave dynamically across different times of the year. Seasonal

and bi-monthly aggregation incorporate these phenological behaviours/changes across all seasons in the spectral description of the vegetation. Second, annual greenest pixels could still contain cloudy pixels. However, it is most unlikely for the same pixel to be cloudy across all seasons of the year.

In the seasonal and bi-monthly cases, the output seasonal greenest pixel mosaics are stacked to produce a single aggregated dataset. Then, to remove the errors due to differences in the acquisitions across the different images in the input collection, the composited data is converted to the above mentioned NDSV feature. In the cases of seasonal and bimonthly aggregation, the NDSV is computed for each of the output (seasonal or bi-monthly) composite before aggregation. Results obtained with NDSV input are benchmarked against using only the original spectral features.

2.4 Data, Experimental procedure, and Results

2.4.1 Study area and data

The proposed methodology was tested with data covering the city of Cordoba, Argentina, located at 31°24'30" S, 64°11'02" W. Cordoba covers an area of $576\,\mathrm{km}^2$ with a population of nearly 1.5 million [89].

S-2 dataset within one-year (Sept. 2017 to Aug. 2018) were considered. After selecting images with less than 10% cloud cover, 51 images were retained and combined. 10 of the 13 bands in S-2 images, i.e. those giving information about land covers were selected for this work, discarding specifically the Aerosol, Cirrus and Water vapor bands. This resulted in a 10 spectral band single input image for the annual greenest pixel composite —- 40 bands and 60 bands in the seasonal and bi-monthly aggregates respectively. For the NDSV feature, a 45-band feature was obtained from the annual greenest, 180 bands for seasonal aggregate and 270 bands for bi-monthly aggregate.

Field visits were made to select points representative of the Trees, Bushes, Grass, Water and Bare soil classes. Points for the artificial surfaces class were selected instead by visual inspection of Google Earth™ images. A total of 1,428 points covering all classes were collected. A major limitation of the points collected is their geographical distribution (see Figure 2.4). This occurred due to constraints including inaccessibility of many areas. Thus, the training of the classifier and validation of the resulting model was only done on points taken within the areas that could be accessed.

It is important to stress the class imbalance in terms of ground truth sample points collected among the considered vegetation classes. This occurred as a result of more availability of certain

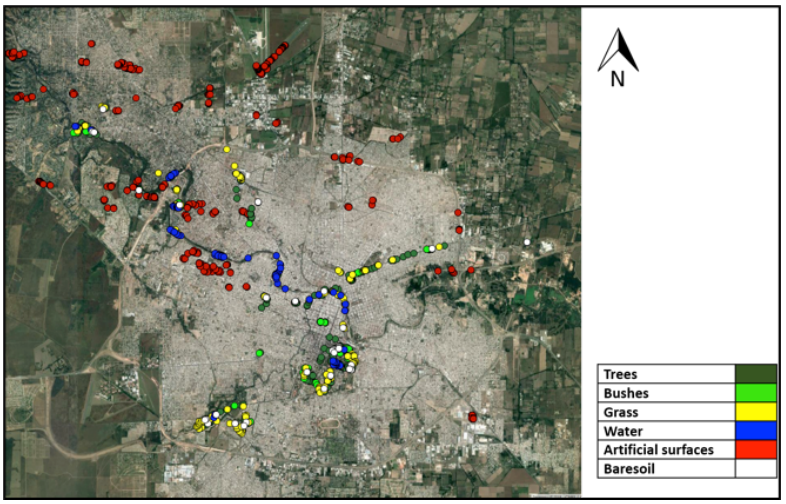


FIGURE 2.4: Landsat view of the city of Cordoba with the selected training and validation points overlaid.

classes in the accessible areas during points collection. Figure 2.5 shows a bar plot of the number of points collected for each vegetation class of interest. It can be seen from this figure that there are more than double Trees points selected than Bushes points.

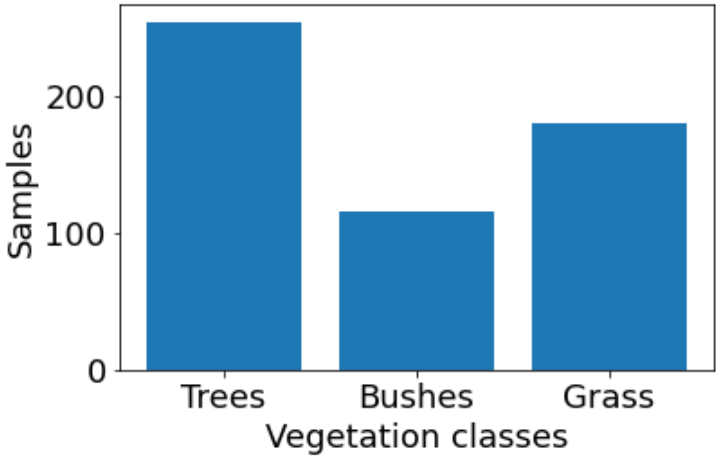


FIGURE 2.5: Number of ground truth sample points selected for each considered vegetation class.

2.4.2 Experimental procedure

All obtained maps were evaluated with respect to their overall accuracy (O.A.), Kappa coefficient, Producer accuracy and User accuracy [90]. 80% of the field collected data points per class were used as training set, while the rest were used as test set. 20-fold cross-validation was used to ensure unbiased accuracy values.

For RF and SVM, grid search experiments were performed to find the optimal model parameters. In the case of SVM, the searched parameters are the "decision procedure" and "type of kernel function". Specifically, the "decision procedure" is searched among "voting" and "margin" options. Then, using the best "decision procedure", the search for the optimal "type of kernel function" was performed. In the case of RF, the "Number of trees" parameter was searched. All these initial experiments were conducted with seasonally aggregated spectral feature input, and the best model obtained at this stage was chosen for further experiments.

To make the case for the need for seasonal aggregation, in light of its increased computational complexity through increasing dimensionality, an experiment was conducted to compare seasonally aggregated and yearly composited spectral feature inputs. Also, a model with bi-monthly aggregated spectral feature input was also tested to see validate that seasonal phenological information is sufficient for the task at hand.

Finally, to make a case for the NDSV feature space, models with seasonally aggregated spectral feature and NDSV feature inputs were compared. All results are presented in Section 2.4.3.

2.4.3 Results

2.4.3.1 Quantitative results

Figure 2.6 shows the overall accuracy as a function of the "type of kernel functions" and "decision procedure" in SVM. This figure reveals that the desired classes are better separable by the combination of "margin" decision procedure with polynomial kernel functions. Specifically, the 5th order polynomial function produces the best result among all tests conducted. These optimal parameters are used in all further experiments conducted with SVM.

Figure 2.7 presents the O.A of the obtained maps results as a function of the "Number of trees" in RF. This plot that better performance is gained until a peak is reached at 100 trees. Beyond that, the model quality degrades. Hence, all further RF in this study are fitted with 100 trees.

Table 2.2 compares the results obtained with the optimal versions of all tested classifiers fitted using seasonally aggregated spectral feature S-2 data input. This table shows that RF provides the best result with regards to both O.A and Kappa. Specifically, RF performs 6% better than SVM which produces the second best performance. This better performance by RF can be attributed to a much improved User accuracy for the Bushes class with respect to other classes. As can be seen from Figure 2.5, the Bushes class has less ground truth examples selected. Hence, RF, here, as shown to be robust with respect to handling sample imbalance better than SVM and CART. SVM provides the worse performance for the Bushes class in both Producer and User accuracy. The performance of CART as the least performing model in general shows

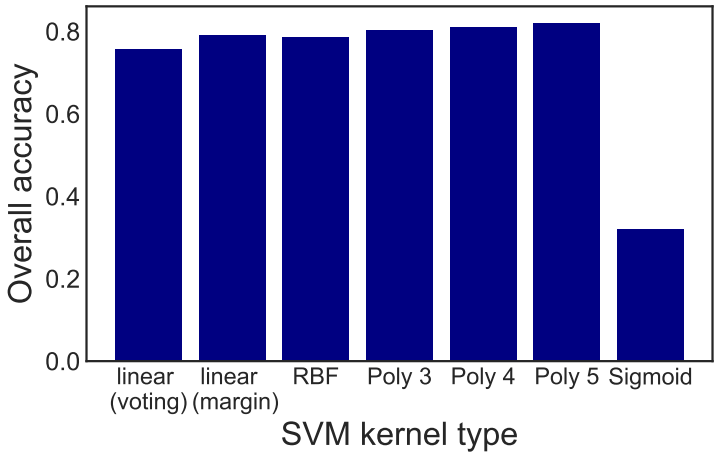


FIGURE 2.6: Classification O.A results using different voting decision procedures (voting or margin) and kernel functions (linear, sigmoid, radial basis function (RBF) and polynomial) in the fitted SVM classifier. Comparison of different decision procedures was done using the linear kernel function. All other kernel function results shown are obtained with margin decision procedure. ("Poly #" = Polynomial function of order #).

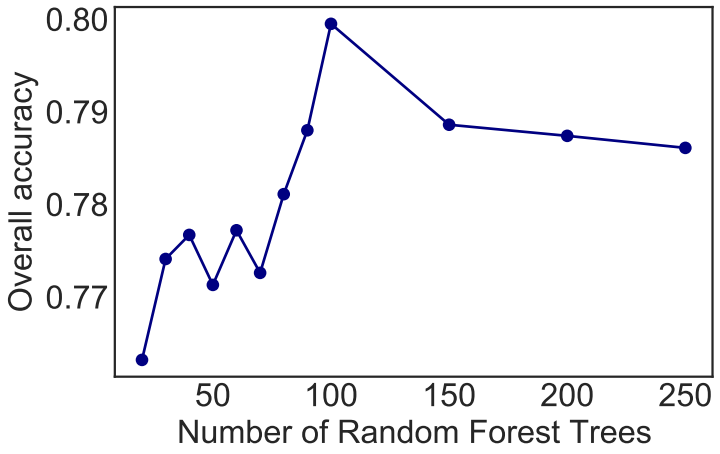


FIGURE 2.7: Plot of overall accuracy of classification as a function of number of Random Forest trees.

the advantage of combining multiple trees as obtainable with RF which has produced the best performance. All further experiments are conducted using RF classifier (100 trees).

Table 2.3 compares results obtained with aggregated (seasonal and bi-monthly) and annual composite spectral feature inputs. This table shows that phenological information obtained by aggregations along the time dimension generally improve the accuracy of the classification map. This is specifically seen in the lower confusion between Bushes and Grass. Since it is difficult to find pure bush pixels at $10\,\mathrm{m}$ spatial resolution, the phenological information improve the separability between the classes. Bi-monthly aggregation, however, do not provide additional

TABLE 2.2: Classification results with seasonally aggregated spectral feature input.

Classifier	Classes ⇒	Trees	Bushes	Grass	Water	Artificial surfaces	Bare soil
RF	Prod	0.841	0.577	0.756	0.814	0.983	0.453
	User	0.746	0.878	0.759	0.862	0.935	0.821
	O.A	0.861 ± 0.020					
	Kappa	0.800 ± 0.026					
SVM	Prod	0.781	0.360	0.641	0.880	0.956	0.407
	User	0.671	0.534	0.734	0.708	0.957	0.606
	O.A	0.812 ± 0.020					
	Kappa	0.730 ± 0.027					
CART	Prod	0.639	0.652	0.596	0.745	0.933	0.395
	User	0.665	0.564	0.618	0.720	0.946	0.348
	O.A	0.778 ± 0.016					
	Kappa	0.685 ± 0.023					

improvement with respect to seasonal to justify the resulting increase in dimensionality.

Table 2.4 compares quantitative results obtained with the seasonally aggregated NSDV and Spectral features inputs to RF classifier. From this table, there are no statistically significant differences between results obtained using either of the two input features. However, there are qualitative differences in the maps produced by both inputs. These differences are explored in Section 2.4.3.2.

TABLE 2.3: Comparing results with seasonally aggregated, bi-monthly aggregated, and annual composite spectral feature inputs to RF classifier (Number of trees = 100)

Classifier	Classes ⇒	Trees	Bushes	Grass	Water	Artificial surfaces	Bare soil
Seasonally aggregated	Prod	0.841	0.577	0.756	0.814	0.983	0.453
	User	0.746	0.878	0.759	0.862	0.935	0.821
	O.A	0.861 ± 0.020					
	Kappa	0.800 ± 0.026					
Bi-monthly aggregated	Prod	0.839	0.602	0.744	0.810	0.987	0.335
	User	0.764	0.864	0.785	0.815	0.928	0.913
	O.A	0.862 ± 0.020					
	Kappa	0.800 ± 0.026					
Annual composite	Prod	0.781	0.523	0.727	0.734	0.973	0.193
	User	0.690	0.805	0.681	0.783	0.931	0.708
	O.A	0.826 ± 0.019					
	Kappa	0.746 ± 0.026					

TABLE 2.4: Comparing results with seasonally aggregated and annual composited NDSV feature inputs to RF classifier (Number of trees = 100)

Classifier	Classes ⇒	Trees	Bushes	Grass	Water	Artificial surfaces	Bare soil
Spectral feature	Prod	0.841	0.577	0.756	0.814	0.983	0.453
	User	0.746	0.878	0.759	0.862	0.935	0.821
	O.A	0.861 ± 0.020					
	Kappa	0.800 ± 0.026					
NDSV feature	Prod	0.849	0.535	0.771	0.806	0.979	0.240
	User	0.732	0.870	0.755	0.856	0.929	0.779
	O.A	0.853 ± 0.019					
	Kappa	0.786 ± 0.018					

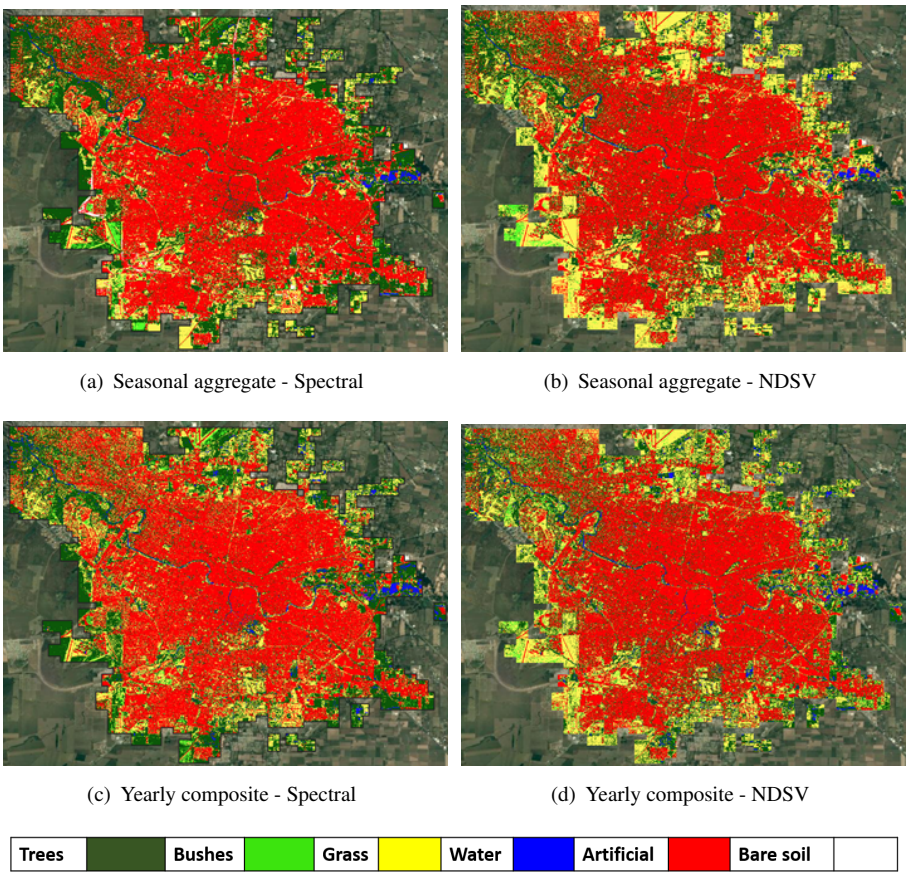
The quantitative results discussion which has been presented in this section focused on the general results, as well as the specific results for the vegetation classes. However, it is important to mention that the generally low accuracy of Bare soil class in all the experiments is due to small amounts of ground truth samples which have been selected. Similarly, there were more Artificial surfaces class samples since the samples were selected by visual inspection of Google Earth™. This is responsible for the high accuracy in the Artificial surfaces class.

2.4.3.2 Qualitative results

To expand on all quantitative results presented in Section 2.4.3.1, the corresponding maps obtained from each feature input considered (excluding bi-monthly aggregated feature) are presented in Figure 2.8. As shown in this figure, there are discrepancies between all the maps presented, but there is the need to zoom into specific areas in other to clearly observe the discrepancies and assess the quality of each map. As will be shown in further expositions in this section, the difference between the presented map are more prominent in the the peri-urban edges of the city. Some of these discrepancies are not captured by the quantitative results due to limitation in the geographical distribution of the ground truth data.

To further analyse the qualities of the resulting map, Figure 2.9 presents a comparison of the Trees class maps obtained with Spectral and NDSV (both seasonally aggregated) inputs. In this figure, the three zones of major disagreements among these maps are highlighted with red bounding boxes.

Figure 2.10 presents a ground truth visualization of three zones highlighted in Figure 2.9. This visualization offers the possibility for qualitative assessments of the maps. It shows that even though both Spectral and NDSV seasonally aggregated inputs produce similar quantitative results as shown in Table 2.4, NDSV input actually produces better results, especially with respect



(a) Seasonal aggregate - Spectral

(b) Seasonal aggregate - NDSV

(c) Yearly composite - Spectral

(d) Yearly composite - NDSV

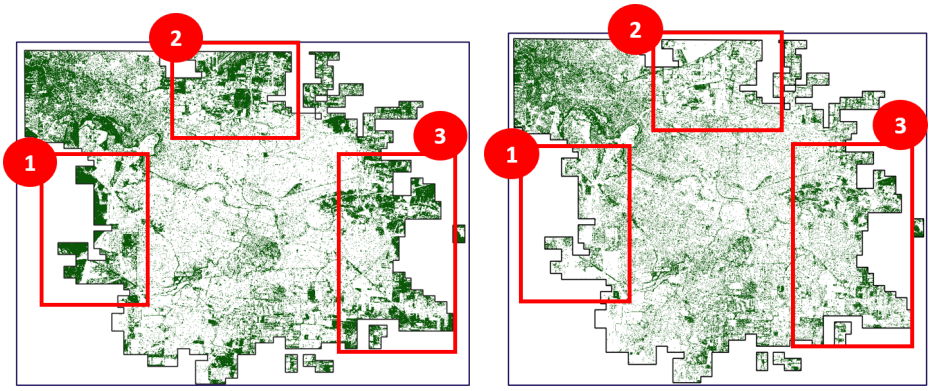
(e) Maps legend

FIGURE 2.8: Classification maps. Figure 2.9 highlights major disagreement areas in the Trees class for maps obtained with seasonally aggregated NDSV feature inputs.

to mapping the vegetation classes. Specifically, we can see that the Spectral input leads to more confusion of Trees and Grass classes than the NDSV input.

2.5 Chapter conclusions

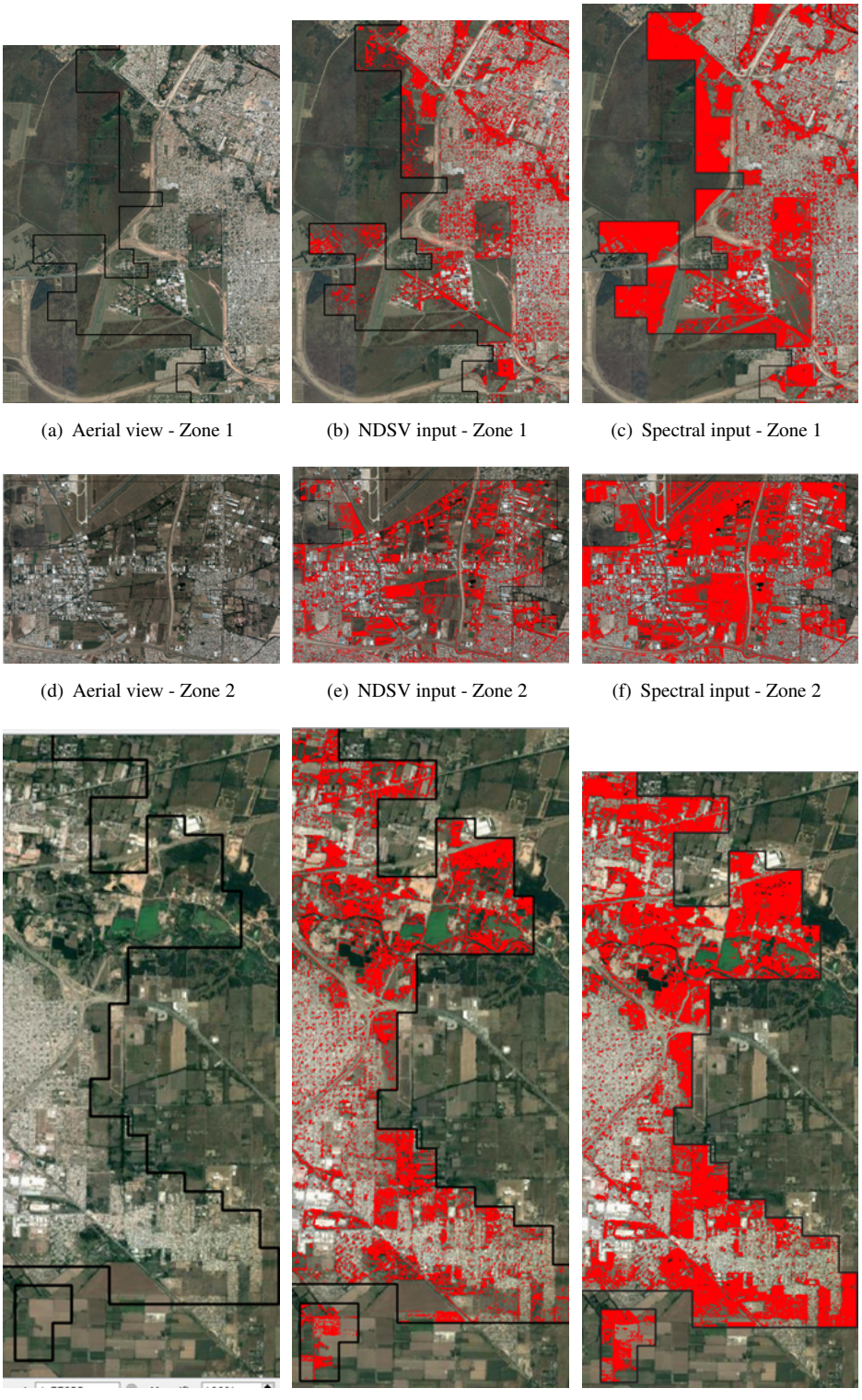
The study presented in this chapter proposes a methodology for mapping urban vegetation from Sentinel-2 data, exploiting their high temporal resolution to create seasonally aggregated inputs. Also, NDSV features were compared with spectral features, and the RF classifier with SVM and CART classification models. It was found that seasonally aggregated inputs show better performance than annual aggregates for the task at hand, while NDSV improves the separation between Trees and Grass. In particular, in cases where spatial and spectral resolutions are insufficient to differentiate among classes, phenological information helped to improve classes separability.



(a) Trees map - Spectral input (b) Trees - NDSV input

FIGURE 2.9: Trees classification maps obtained with seasonally aggregated Spectral and NDSV feature inputs. The highlighted zones are areas with significant qualitative disagreements among the maps. These zones are overlaid on aerial images showing ground truth classes in Figure 2.10

A major limitation of the study is in the geographical distribution of the training and validation point samples. In line with this, future studies can consider developing an automatic way to augment field collected ground truth data. Also, more vegetation classes could have been defined. Further studies could be conducted to mitigate these limitations.



(a) Aerial view - Zone 1 (b) NDSV input - Zone 1 (c) Spectral input - Zone 1

(d) Aerial view - Zone 2 (e) NDSV input - Zone 2 (f) Spectral input - Zone 2

(g) Aerial view - Zone 3 (h) NDSV input - Zone 3 (i) Spectral input - Zone 3

FIGURE 2.10: Trees class maps from selected zones in Figure 2.9. By comparing the class maps with ground truth through the aerial images in the background, it is seen that the Spectral input, unlike NDSV, creates more confusion among Trees and Grass classes.

Chapter 3

Modeling the temporal population distribution of Ae. aegypti mosquitoes using big Earth observation data

3.1 Introduction [1]

As mentioned in Section 1.4, one major area currently experiencing rapid innovation due to EO data is landscape epidemiology [91]. In this field, environmental variables that are proxies to favorable conditions for the spread of disease causing vectors are extracted from EO data and used to model the distribution of these vectors [25].

Widespread and clinically important viruses including Zika, Chikungunya, Dengue and Yellow fever are transmitted mainly by the female *Ae. aegypti* mosquito species [89, 92]. The *Ae. aegypti* species is known to be adaptable to urban environments due to its inclination to being bred in artificial containers [42, 53, 54]. Spread of the causal diseases by this vector is known to correlate with the local adult vector population [93]. Environmental conditions including precipitation, vegetation conditions, temperature, and humidity have been shown in previous works to significantly influence *Ae. aegypti* mosquito development [42, 54, 94]. After obtaining these environmental variables from EO data, statistical and machine learning (ML) models of disease outbreak can be used for vector population prediction [95].

In spite of the works in this study domain, there are still some gaps. With regards to temporal modelling of diseases/vector prevalence, one of the major gaps is in the selection of the most informative environmental features subset to obtain the model with best fit and least redundancy.

[1] This chapter has been published as a standalone paper as O. Mudele, F. Bayer, L. Zanandrez, A. Eiras, P. Gamba, "Modeling the Temporal Population Distribution of Ae. aegypti Mosquito using Big Earth Observation Data," IEEE Access, doi: 10.1109/ACCESS.2020.2966080, vol. 8, no. 1, pp. 14182-14194, Jan. 2020.

To address the tradeoff between quality and explanability, this chapter presents a methodology to model the temporal population distribution of female *Ae. aegypti* mosquito based on time series environmental variables obtained from freely accessible EO data products and field collected mosquito population data. To achieve this aim, a random forest (RF) [96, 97] model is considered due to: (i) its robustness against unbalanced data with good performance in complex problems and (ii) its embedded mean decrease impurity (MDI) variable importance measure. MDI is used to rank and extract the most relevant environmental feature subset for modeling the vector population. It considers also nonlinear associations between candidate features and provides an avenue to explain the resulting model. For benchmark purposes, RF is compared with other ML models already used in similar works namely: k-Nearest Neighbors (KNN), Support Vector Regression (SVR), Decision Tree Regression (DTR), and Multi Layer Perceptron (MLP), and statistical regression models such as GLM [98] and linear regression model (LM). Finally, considering the best model, fitted with the most informative features as ranked by MDI, the effects of the selected features on the vector population are explained.

3.2 Study Area and EO Data Variables

In this study, the average population counts of female *Ae. aegypti* mosquitoes is modeled as a function of selected RS environmental variables. The methodology of this study is applied to data for the municipality of Vila Velha, located between latitudes $20°19'$ and $20°32'$ South, and longitudes $40°16'$ and $40°28'$ West, on the coast of the Espírito Santo State of Brazil. This municipality covers a total area of $209.965\,\mathrm{km}^2$, with an estimated population of $486,208$ inhabitants [99], and it is the main city in the metropolitan region of the capital Vitoria. Vila Velha has been chosen as the study area due to availability of ground collected data. Also, the Espírito Santo state has verified $63,847$ dengue cases, with an incidence of 1588.8 per $100,000$ inhabitants in year 2019, with Vila Velha having $6,557$ of those cases ($1,348.6$ per $100,000$ inhabitants) , which is far higher than the limit to be considered an epidemic [100]. Figure 3.1 shows the location of the studied area.

Brazil has implemented several vector control strategies since 2002 as part of its National Dengue Control Program (PNCD) initiative [101]. This program addresses important components including: epidemiological surveillance, vector control, environmental sanitation, among others [102]. PCND relies exclusively on the larval collection, which is not as effective as adult mosquito sampling for *Ae. aegypti* surveillance and control [103]. Since 2005, Ecovec —- an expert real-time mosquito surveillance company based in Belo Horizonte, Brazil —- has been implementing the Integrated Aedes Monitoring System (*Monitoramento Integrado do Aedes: MI-Aedes*®. This program was initiated as an operational improvement to PNCD with regards to operational adult mosquito surveillance, and it is still in operation. The MI-Aedes® platform

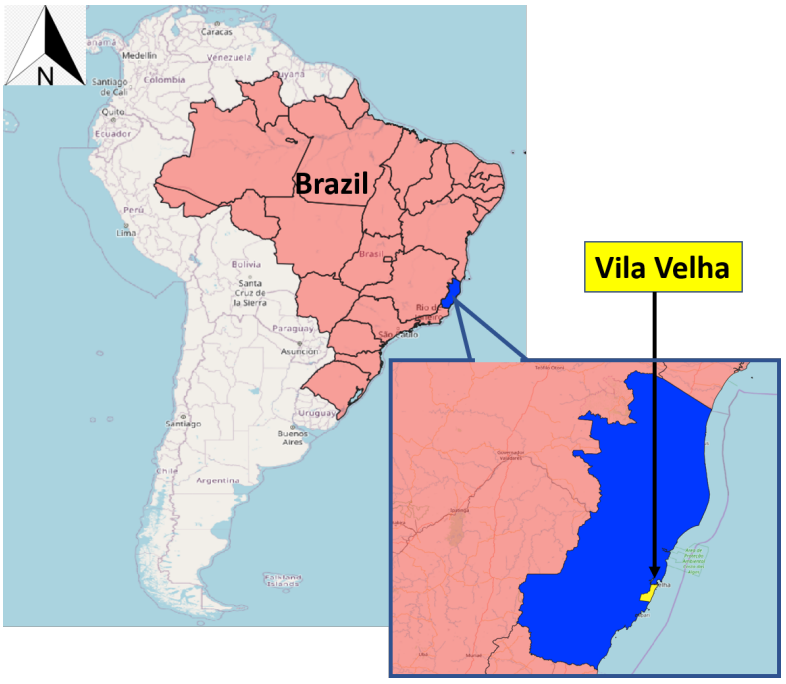


FIGURE 3.1: Location of Vila Velha in Brazil.

consists of (i) adult sticky traps (commercially named MosquiTRAP®) to monitor adult Ae. aegypti population; (ii) data entry mobile app; (iii) georeferenced hotspot areas map for targeting vector control [104]. It is adequate for large scale monitoring because it does not need electricity, presents low cost per unit, and showed to be a reliable tool for early detection of dengue transmission risk [103]. MI-Aedes® was extended to Vila Velha in 2017, and its data was used for vector control improvement in 2018.

Since the commencement of MI-Aedes® in Vila Velha, vector control actions in this area have been subjected to revisions and improvements at intervals based on previous mosquito population data collected. For example, data collected in 2017 was used to improve control actions from the 1st week of 2018 by directing more control actions to places with higher adult mosquito population. One major advantage of the MI-Aedes® program is that it collects weekly data of adult vector population, not immature forms. This ensures that the data highly correlates with human diseases infection cases since only adult female mosquitoes are responsible for the transmission [103].

This study used MI-Aedes® data in Vila Velha from week 15 in 2017 to week 34 in 2019 — a total of 124 weeks. During this period, the program utilized homogeneously positioned 791 adult MosquiTRAPs® to obtain weekly Ae. aegypti population data across the whole municipality. To reduce the effects of spatial autocorrelation of collected data, each trap is placed within at least $250\,\mathrm{m}$ from all surrounding ones. A team of trained field workers acquires mosquito population data weekly by inspecting sticky cards set inside each trap. Ae. aegypti mosquitos are identified, counted, and their presence registered in a mobile app that, in turn, sends data to

an online database. To maintain operational standards on the field during the collection tenure, an Ecovec specialist visited the field every six months to check the conditions of the traps and to update staff training. Figure 3.2 shows the location of the MosquiTraps® with in the study area.

Operative conditions changes driven by vector control actions which started from the 1st week of 2018 and an update in the MI-Aedes® platform starting from week 42 in 2018 which led to changes in identification number of the mosquito traps and slight changes in their geopositions are accounted for in the study analyses. In the cases of trap location changes, the traps were moved to new adjacent residences due to monitoring difficulties or because residents specifically requested the changes. All in all, trap relocations beyond a 40 meters radius were rare. As a result, the consequent analyses were split in three time batches, as presented in Table 3.1. Since previous works [53, 54] have shown that it is sufficient to work with weekly observation samples, the *Ae. aegypti* population data for each batch were cleaned to obtain female mosquito population per mosquito trap on a weekly basis. It is important to note two points here: firstly, lower vector population and correlation with environment condition are expected starting from the Batch 2 in the study analyses. This is due to improvements in operative mechanisms in control actions. Secondly, the three data Batches do not exactly span the same time and seasons of the year.

TABLE 3.1: Mosquito population data Batches used in this study

Batch	Date range	Total weeks	Differentiating condition
1	10/04/2017 - 31/12/2017	38	-
2	02/01/2018 - 05/10/2018	40	Vector control improvement
3	08/10/2018 - 23/08/2019	46	Update to the MI-Aedes® platform

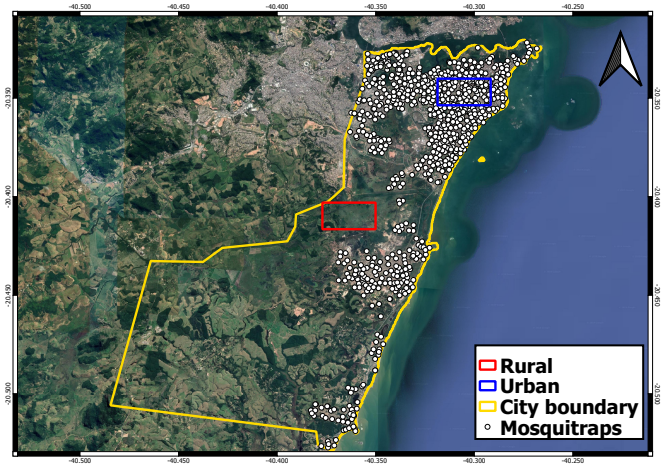


FIGURE 3.2: Aerial view of Vila Velha municipality, data collection mosquito trap locations, and urban and rural surface zones selected to extract the environmental variables.

As already mentioned in Section 1.3.2, in studies that focus on modeling the temporal distribution of *Ae. aegypti* based on EO data derived environmental factors, NASA's MODIS EO data products have been recently used [53, 54]. This is due to its free access, global availability and non-requirement of much further processing to obtain the needed environmental features [39].

For this work, data products providing information corresponding to surface temperature, precipitation, humidity, and vegetation conditions were collected for the weeks matching the vector population data per batch. NWVI layers obtained from MODIS MOD13Q1 product were used to obtain information about vegetation condition. The near and mid-infrared bands of the same data product was used to compute the NDWI using Gao's definition [29]. This layer provides information about vegetation water content which is a proxy to humidity and surface moisture [54]. In addition, estimates of the minimum and maximum surface temperatures have been obtained from the day and night-time land surface temperature (LST) bands of the 8-day composite MODIS MOD11A2 product at $1\,\mathrm{km}$ spatial resolution [40]. Finally, the calibrated precipitation band of the Global Precipitation Measurement (GPM) mission data product was used to obtain precipitation information [69], although in a very coarse way, because this data is available daily at $0.1\,\mathrm{deg}$ spatial resolution ($\approx 11\,\mathrm{km}$). The mathematical definitions of both NDVI and NDWI have been provided in Table 1.1, and the details of all utilized EO data products can be found in Table 1.2.

3.3 Data preparation for modeling

A high-level schematic of the whole methodology of the chapter study as described in this section is presented in Figure 3.3. All the blocks in this schematic, except for modeling and model selection, are captured under data preparation. The data preparation blocks are detailed in this section.

All data were downloaded using the `export` object of the `JavaScript` application programming interface of google earth engine (GEE). This object has resampling and reprojection methods abstracted into it for easy co-registration of data with different properties. For each download task, we obtained the output data at a resampled spatial resolution of $250\,\mathrm{m}$ because this is the minimum distance separating neighboring mosquito traps. We used the `scale` parameter of the `export` object to set this spatial resolution value. We set the common coordinate reference system projection for all downloaded data as `EPSG:4326` using the `crs` parameter.

As previously implemented in [53, 54] and [105], to find the relationship between the measured number of mosquitoes and the model output exploiting EO data, two buffer areas of size $18\,\mathrm{km}^2$ were defined, each from which distinct temporal characterization of the considered EO variables are obtained within our study area. One of the zones is in the densely urbanized part of the city

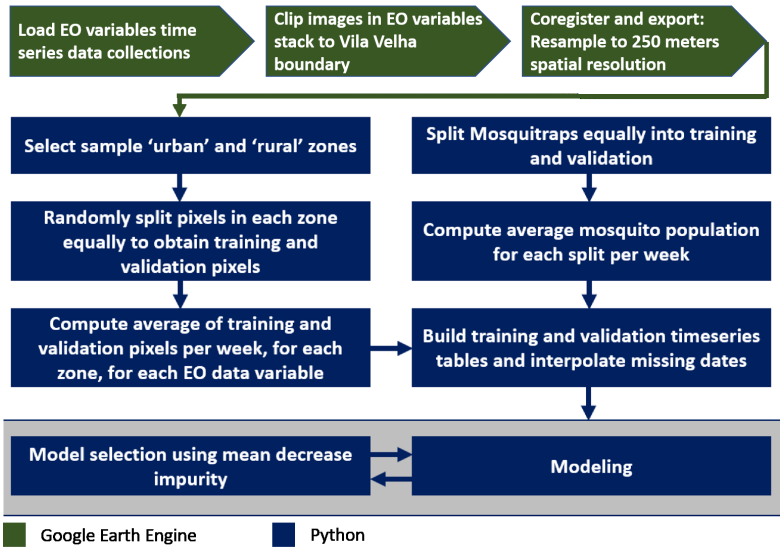


FIGURE 3.3: Schematic of the study methodology.

(labeled as "urban"), and the other in the peri-urban zone (labeled as "rural") areas. These two areas are shown in Figure 3.2. The hypothesis for sampling both surface types is that the dynamics of the EO variables are temporally different across urban and rural surfaces, and so are their effects on the vector population. All obtained models are fitted using a combination of EO variable features extracted from these two surface samples.

To obtain training and validation climatic covariates data for our model, the pixels in each selected zone are randomly and equally split, and the mean for each split per image in the temporal stack is computed. This process was done for all image stacks representing the different climatic variables. We used a fifth-order spline interpolation to obtain values for missing dates. We used time-delayed observations with weekly steps up to two weeks to account for non-synchronous environmental effects in the *Ae. aegypti* vector development life cycle [106]. As in [54], all the variables were standardized (z-score). This was done because variable rescaling is a good practice, especially for training a neural network [107].

3.4 Modeling Procedure

To model the weekly mean number of mosquitoes per mosquito trap (Y), we consider fitting a RF model. For prediction benchmark purpose, ANN, SVR, KNN, DTR, LM, and GLM fitting models were also considered. Finally, a more parsimonious RF model, labeled RF*, considering only the most informative climate variables obtained by using the MDI to rank all predictor covariates was also implemented and compared with the other ones. For prediction performance comparison, the usual quantitative measures between the observed data (y) and the predicted values ($\hat{y}$) were considered, namely: the linear correlation coefficient (R), the root mean square

error (RMSE), and the mean absolute percentage error (MAPE) [108]. We use R and MAPE to measure relative qualities, and RMSE to measure absolute fit of our models. Background details RF implementation (function and model parameters) for this study are presented in Section 3.4.1. Summary of the other models fitted for comparison are presented in Section 3.4.2. All modeling computation are implemented with with R programming [109].

3.4.1 Random Forest Regression

Random forests (RF) [96] are one of the most effective computationally intensive procedures to improve on unstable estimates, especially when it is difficult to find a good model due to problem complexity [97]. It is a predictor consisting of a collection of several randomized regression trees [97]. Let $\mathbf{X} \in \chi \subset \mathbb{R}^p$ an input vector related to p features and $Y \in \mathbb{R}$ the response random variable, the objective is to estimate the regression function $m(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$. Given a training sample $\mathbb{D}_n = ((\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n))$ of independent random variables distributed as the independent prototype pair $(\mathbf{X}, Y)$, the goal is to use the data set $\mathbb{D}_n$ to construct an estimate $m_n : \chi \to \mathbb{R}$ of the function m. To this aim, the random forest consists of a collection of M randomized regression trees. For the jth tree in the family, with $j = 1, \ldots, M$, the predicted value at each j is denoted by $m_n(\mathbf{x}; \Omega_j, \mathbb{D}_n)$, where $\Omega_1, \ldots, \Omega_M$ are random variables independent of $\mathbb{D}_n$. The variables Ω_j are used to resample the training set prior to the growing of individual trees and to select the directions for splitting. Then, the trees are combined to form the forest estimate given by:

$$\hat{y} = m_{M,n}(\mathbf{x}; \Omega_1, \ldots, \Omega_M, \mathbb{D}_n) = \frac{1}{M} \sum_{j=1}^{M} m_n(\mathbf{x}; \Omega_j, \mathbb{D}_n). \tag{3.1}$$

In this procedure, we have also to set other tuning parameters, i.e., the number of possible directions for splitting (m_s) at each node of each tree, and the lowest number of examples (m_e) in each cell to cause a split. In this work we set initially $M = 500$, $m_s = \lceil p/3 \rceil$ and $m_e = 5$.

One important characteristic of the random forest is that it can be used to rank the importance of the input variables (i.e., the features we extract from EO data) over the pattern variability of the target variable Y. In this work, we considered the MDI calculated based on the reduction in sum of prediction squared errors averaged over all trees whenever a variable is chosen to split [110].

3.4.2 Other Predictive Models

As mentioned above, for comparison purposes a number of other regression models have been implemented and their results compared with the ones by the proposed RF procedure. These

models have been selected due to their application in a similar study presented in [54]. Specifically, the selected models are:

- the Linear Regression Model (LM): the most used predictive model in different fields. With respect to the topic of this paper, the LM has been used in similar works in [53] and [54], to model the *Ae. aegypti* oviposition activity in a northern Argentine city. In this work, LM is implemented using the `lm()` function in `R`, and an ordinary least square estimator is considered to estimate the regression parameters.

- The Generalized Linear Model (GLM): a class of regression models able to model response variables in the exponential family of distributions [98]. Since the weekly mean value of mosquitoes population (y) is continuous and always in the space of positive real numbers ($y \in \mathbb{R}^+$), the Gaussian assumption considered for inference in the LM can lead to poor results and predictions. In GLM, instead, the gamma distribution is considered to model y, together with the log link function in the GLM regression structure.

- Artificial Neural Networks (ANN): one of the most used ML methods for geoscience problems. More recently, deep learning methods – corresponding to ANN architectures with several hidden layers [111] – became widely applied in many EO data processing problems [112, 113]. The ANN is a nonlinear data modeling technique used to model complex relationships between sets of input and output variables [114]. In this work, a multilayer perceptron with three layers, each with three neurons, is used. Each of the neuron is activated with a logistic activation function, and the resilient backpropagation with weight backtracking algorithm [115] was considered for training the neural network.

- Support Vector Regression (SVR): an approach based on support vectors with radial basis function kernel with tuning parameter γ set as 1/(number of features), tuned via the the epsilon-regression method.

- k-Nearest Neighbor Regression (KNN): a regression based on the 4-nearest neighbors according to Euclidean distance and uniform weights for local interpolation.

- Decision Trees Regression (DTR): a regression model in the form of a tree structure [116]. It is a simple but powerful prediction method [117]. This model was fitted with default parameters.

3.5 Results and Discussion

All the EO data variables used in this work were considered both in an urban surface zone (U) and in a peri-urban, or rural zone (R). We thus have the following variables: NDVI-U, NDWI-U, TempD-U, TempN-U, and Prec-U, as well as NDVI-R, NDWI-R, TempD-R, TempN-R, and

Prec-R, where TempD is the daytime temperature and TempN is the night-time temperature. Table 3.2 reports the mean, median, standard deviation (SD), minimum (Min) and maximum (Max) of the EO data variables together with the female mosquito population (y). These measures were computed for the three Batches of time considered in our study (cf. Table 3.1) in a separate way, to evaluate the influence of the vector control program improvement implemented in 2018 and 2019, affecting Batches 2 and 3. We observed that the average value of the number of female mosquitoes in Batch 1 is 0.1910, while the values in Batches 2 and 3 are 0.1267 and 0.1134, respectively. This represents about 40% decrease in the mean number of mosquitoes from Batch 1 compared to subsequent Batches considered and highlights the efficacy of the data-driven vector control program improvement which has been implemented. Furthermore, the nonparametric Kruskal-Wallis test [118] rejects the null hypothesis of three populations being equally distributed (p-value = 0.0092), confirming that the decrease in the mean number of mosquitoes from Batches 1 to 2 and 3 is statistically significant. In addition to the Kruskal-Wallis test, the Nemenyi test shows that the population from Batch 1 differs from Batches 2 and 3, but population distributions of mosquitoes of Batches 2 and 3 do not differ statistically. Also, we can note that the standard deviation of the number of mosquitoes decreases from 0.1446 in Batch 1 to 0.0956 and 0.0315 in Batches 2 and 3, respectively. The boxplot in Figure 3.4 shows the differences among the population of female mosquitoes in the different Batches. Particularly, we can see that the maximum value in Batch 3 is four times less than the maximum number in Batch 1.

The environmental characteristics, as presented in Table 3.2, slightly differ across the three considered batch periods due to differences in observation time and seasons. Also, it can be seen that there was more precipitation in 2017 (Batch 1) than 2018 and 2019 (Batches 2 and 3). Regarding the urban and rural zones, the average land surface temperature in urban zone, TempD-U and TempN-U, respectively, is, as expected, higher and more variable than in the rural zone. Moreover, as equally expected, the opposite is the case for both NDVI and NDWI.

The linear correlation between each environmental variable, including lagged ones, and the female mosquito population, y, is presented in Table 3.3. This table shows, first of all, that none of the considered covariates show very strong linear correlation (above 0.7) with y. However, in all Batches, the non-lagged daytime temperature variables show the highest correlation with y: TempD-R and TempD-U — in that order — in both Batches 1 and 2, and only the former in Batch 3. In general, more EO variables show higher correlation with the vector population in Batch 2, with five non-lagged variables showing correlation above 0.4, compared to the other Batches.

In addition, we can note that the lagged NDWI-R and NDVI-R and the precipitation variables show greater correlation with y than with non-lagged ones. This observation can be explained by the aquatic cycle of *Ae. aegypti* from egg to adult being approximately 7-9 days [106], thus

TABLE 3.2: Descriptive measures of the EO-based environmental variables of interest, computed separately for each Batch

Variable	Mean	Median	SD	Min	Max
Batch 1 (2017)					
y	0.1910	0.1587	0.1446	0.0319	0.7873
NDVI-U	0.2269	0.2260	0.0261	0.1702	0.2841
NDVI-R	0.7280	0.7395	0.0448	0.5826	0.7983
NDWI-U	0.0024	0.0035	0.0350	-0.0994	0.0748
NDWI-R	0.5407	0.5492	0.0473	0.3952	0.6640
TempD-U (°C)	31.4166	31.0544	4.1390	22.5969	39.3821
TempD-R (°C)	27.7208	27.7761	2.9338	20.5000	33.8600
TempN-U (°C)	20.0180	20.6767	2.4379	13.4942	23.5909
TempN-R (°C)	19.0909	19.3264	1.8757	13.6278	23.1461
Prec-U ($\mathrm{mm\,h^{-1}}$)	2.9118	0.2000	7.5662	0.0000	43.2500
Prec-R ($\mathrm{mm\,h^{-1}}$)	2.9039	0.5375	6.7846	0.0000	37.1750
Batch 2 (2018)					
y	0.1267	0.0923	0.0956	0.0531	0.5057
NDVI-U	0.2500	0.2474	0.0179	0.2086	0.2996
NDVI-R	0.7475	0.7554	0.0261	0.6802	0.7922
NDWI-U	0.0030	-0.0021	0.0365	-0.0591	0.1065
NDWI-R	0.5923	0.5965	0.0518	0.4762	0.6925
TempD-U (°C)	31.8615	31.5774	3.4781	26.4761	38.8394
TempD-R (°C)	27.5138	26.9434	2.6868	23.7850	33.8161
TempN-U (°C)	21.2936	21.0190	2.2088	16.3886	25.3319
TempN-R (°C)	20.1959	19.6536	2.0855	16.5067	23.7806
Prec-U ($\mathrm{mm\,h^{-1}}$)	2.4707	0.0654	4.9280	0.0000	23.0803
Prec-R ($\mathrm{mm\,h^{-1}}$)	1.7527	0.0990	3.4678	0.0000	16.0950
Batch 3 (2018-2019)					
y	0.1134	0.1075	0.0315	0.0620	0.1922
NDVI-U	0.2557	0.2551	0.0188	0.2117	0.3075
NDVI-R	0.7210	0.7265	0.0331	0.5843	0.7829
NDWI-U	-0.0137	-0.0200	0.0324	-0.0762	0.0661
NDWI-R	0.5350	0.5450	0.0538	0.4031	0.6272
TempD-U (°C)	33.6604	33.8508	4.3575	18.6561	39.0947
TempD-R (°C)	28.9434	29.4046	2.3435	24.7489	33.1411
TempN-U (°C)	22.0600	22.4421	2.1145	17.3431	25.2092
TempN-R (°C)	20.6418	21.1400	2.6758	13.2544	24.1517
Prec-U ($\mathrm{mm\,h^{-1}}$)	1.6132	0.1596	2.8811	0.0000	11.3604
Prec-R ($\mathrm{mm\,h^{-1}}$)	1.3545	0.0419	2.4574	0.0000	9.5575

Units: °C: degrees Celcius, $\mathrm{mm\,h^{-1}}$: millimeters per hour

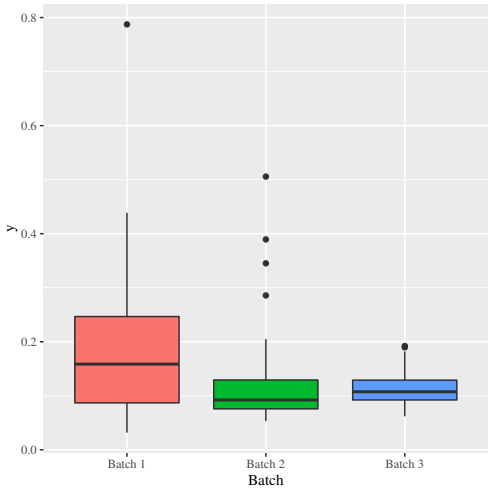


FIGURE 3.4: Boxplot comparing the distribution of the number of female mosquitoes in the three considered periods.

making the transition dependent on non-synchronous environmental effects, as also highlighted in [119]. We observe that the linear relationships between the targets and the predictors could be insufficient to describe the temporal variance in the mosquito population, justifying the need for non-linear approaches such as the considered ML methods.

Considering the results for the validation dataset, the quantitative measures of the prediction obtained from each considered model are shown in Table 3.4 for all Batches. In Batch 1, the measures show that GLM produces the worst results in terms of R and RMSE, while LM the worst in terms of MAPE. The best performances are reached in R by RF, in RMSE by RF*, and in MAPE by ANN. The outstanding performance of RF, and its more parsimonious variant, RF*, is highlighted by the correlation measure of 0.90 and 0.86, and RMSE of 0.0369 and 0.0396, respectively. Also, in Batch 2, we obtain the best result in terms of RMSE with RF*, followed by the fully fitted RF. In this Batch, however, best results considering R and MAPE are obtained using KNN. In Batch 3, RF* produced the best model as measured by R, RMSE and MAPE. Considering RMSE, RF* produces equal or better quality in all Batches. In general, from the

TABLE 3.3: Correlation matrix between the average number of mosquitoes (y) and each covariate among the pool of EO-based variables in Table 3.2

Covariate	Lag 0	Lag 1	Lag 2	Lag 0	Lag 1	Lag 2	Lag 0	Lag 1	Lag 2
NDVI-U	-0.3097	-0.0926	-0.1127	-0.1401	-0.2117	-0.3098	-0.1527	-0.0546	0.0913
NDVI-R	0.0030	0.0958	-0.0142	-0.5025	-0.5852	-0.4079	0.0439	0.0807	0.3296
NDWI-U	-0.3989	-0.2592	-0.1480	-0.2454	-0.2772	-0.2118	0.1716	0.0320	-0.0873
NDWI-R	0.0624	0.1566	0.1342	-0.2702	-0.3954	-0.3392	0.0380	0.1005	0.1480
TempD-U	0.4046	0.3086	0.2236	0.5810	0.4700	0.3565	-0.1082	-0.0414	-0.0750
TempD-R	0.4157	0.3327	0.1532	0.6418	0.4962	0.4988	-0.2703	-0.1654	-0.2530
TempN-U	-0.2959	0.0895	0.1403	0.4835	0.4150	0.3578	-0.1004	-0.0566	-0.0871
TempN-R	0.1105	0.1701	0.0328	0.4323	0.3584	0.4254	-0.1352	-0.1698	-0.0909
Prec-U	0.0084	-0.0349	-0.0553	-0.0185	-0.1644	0.0360	-0.0969	-0.1376	0.0578
Prec-R	0.0261	-0.0368	-0.0335	0.0053	-0.1595	-0.0232	-0.0384	-0.1142	0.0589

whole results, though the statistical models in certain cases produced sufficient or comparable measured qualities, there are much more cases of better performances by machine learning models. For example, LM produces comparable quality with respect to RF in all Batches of our analyses if we look only at the explained variance which is measured by R. When we consider the RMSE, however, we see that that RF performs better than LM in terms of absolute fit. LM's high correlation measure of 0.8062 and 0.8535 in Batches 1 and 2 respectively are as a result of better correlation between more of their EO covariates and y compared to what is obtained in Batch 3. The much lower R measure for the all models on Batch 3 data could be as a result of time cumulative effect of data driven improvement in the vector control actions which led to less variation in the mosquito population due to environmental changes. This same cause may be used to explain the lower correlation among more EO covariates and y in the Batch 3 with respect to Batches 1 and 2, as earlier presented in Table 3.3.

Juxtaposing these results in light of time sample sizes across the batches, as shown in Table 3.1, it is seen that even though there are more weeks (i.e. more data samples) considered in Batch 3 (46 weeks) compared to batch 1 (36 weeks), we still obtain much better R values in Batch 1 for all the considered methods. This shows that the control action effects take precedence over sample size in the modeling task explored in this study.

Figure 3.5 shows a scatterplot of the actual and predicted female mosquito population values. In this plot, GLM, LM, and DTR are not considered due to their low RMSE figures. In all the Batches, the EO-based environmental features show significant effects on the vector population, less so in Batch 3. The environmental effects in Batches 2 and 3 show pointers with which the vector control program can be further improved. A plausible approach to consider is weighting the intensity of the program implementation by the most relevant environmental conditions across the year. This can ensure a higher control intensity during periods of favorable climate for vector development, thus reducing risks exposure during these periods and helping to optimize vector control resources allocation. As opposed to current common practice which favors the application of control actions mostly in rainy season, these actions should also be performed in dry season and periods of low vector population. Implementing control actions during the dry

TABLE 3.4: Quality measures of predictions in the validation dataset

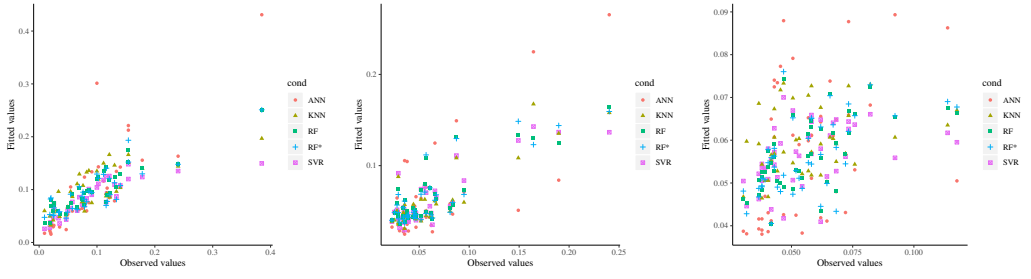
	Batch 1			Batch 2			Batch 3		
Model	R	RMSE	MAPE	R	RMSE	MAPE	R	RMSE	MAPE
GLM	0.6274	0.0722	60.6487	0.6366	0.0490	43.0603	0.5707	0.0180	23.8297
LM	0.8062	0.0480	75.0815	0.8535	0.0309	44.3038	0.5697	0.0168	23.3971
ANN	0.8420	0.0447	34.1122	0.7202	0.0360	44.3968	0.4048	0.0192	24.3962
SVR	0.8270	0.0471	36.1444	0.8651	0.0252	30.9242	0.4096	0.0174	23.8113
KNN	0.7456	0.0478	64.1194	0.8839	0.0223	30.7137	0.3662	0.0180	27.1013
DTR	0.6334	0.0547	67.9013	0.7925	0.0281	32.5429	0.3466	0.0179	25.7221
RF	0.9066	0.0369	56.9834	0.8530	0.0255	35.4397	0.5981	0.0156	21.4197
RF* (8 features)	0.8618	0.0396	61.8109	0.8594	0.0247	33.4977	0.5885	0.0156	21.5879

season has been shown in [120] to suppress the population of eggs laid during this season and consequently reduce the hatched population in the following rainy season.

In Table 3.5, we present a summary of the observed and fitted data: mean, median, minimum (Min), maximum (Max), first (Q1) and third (Q3) quartiles. In Batch 1, LM exaggerates the minimum value, producing a negative minimum value which has no physical meaning in relation to mosquito population. Considering that this same model produced a good quality measure of R, we deduce that LM, regardless of the quality score, is not a good model for predicting mosquito population because its distribution assumption of the response variable is inappropriate for this study use case. This problem, however, does not occur with GLM, due to its gamma distribution assumption which assumes positive values for y. GLM produces a good estimated minimum value in all Batches, but still exhibits low prediction performance, which can be seen in its predicted mean and Q3 which are above the observed values for all Batches. Among the ML methods, ANN produces better prediction in relation to minimum and maximum values. RF and RF*, though both overestimate the minimum in all Batches produced good estimated mean values and showed good balance. The results in Figure 3.5, Table 3.4 and Table 3.5 show that the best models for predicting the vector population are RF* and RF, followed by SVR.

Figure 3.6 shows the mosquito population data (validation set) along with the predicted values by the best models. It is seen that the SVR and KNN models underestimate the seasonal spikes in Batches 1 and 2. The model performances are reduced in Batches 2 and 3 to their overestimation of the very low mosquito population value points resulting from the improved control activities in the Batch periods.

In further analysis of RF*, here, we discuss the selected most informative features subset across the considered observation time periods. Figure 3.7 shows the average value of the MDI for each variable considering 50 RF replicas. For Batch 1, the eight selected most informative EO variables are TempN-R, TempD-U, TempN-R1, TempN-U2, TempD-R, TempN-U, NDVI-U and NDWI-U. For Batch 2, the selected variables features are: TempD-R1, NDVI-R1, TempD-R, NDWI-R2, NDVI-R, NDWI-R1, TempD-R2 and NDWI-R2. For Batch 3, they are: TempD-R,

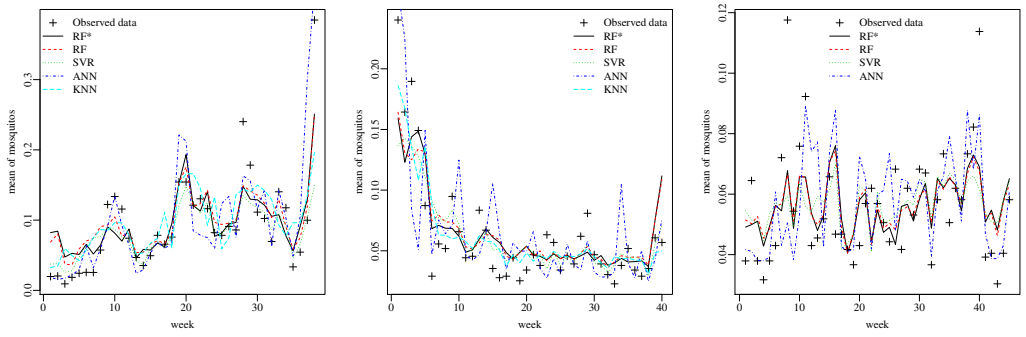


(a) Batch 1 (b) Batch 2 (c) Batch 3

FIGURE 3.5: Scatterplot of observed and predicted values by means of different predictors: ANN, KNN, RF, RF* and SVR (see the text for more explanation).

TABLE 3.5: Summary of the observed and fitted data

	Min	Q1	Median	Mean	Q3	Max
			Batch 1			
y	0.0093	0.0474	0.0803	0.0933	0.1204	0.3848
GLM	0.0063	0.0397	0.0745	0.1043	0.1611	0.4340
LM	-0.0724	0.0413	0.0793	0.0946	0.1330	0.3565
ANN	0.0157	0.0546	0.0752	0.0995	0.1312	0.4309
SVR	0.0257	0.0610	0.0822	0.0838	0.1090	0.1496
KNN	0.0325	0.0621	0.0928	0.0984	0.1357	0.1967
DTR	0.0392	0.0392	0.0741	0.0986	0.1411	0.1715
RF	0.0362	0.0703	0.0918	0.0985	0.1224	0.2512
RF*	0.0461	0.0646	0.0858	0.0956	0.1187	0.2510
			Batch 2			
y	0.0228	0.0351	0.0461	0.0619	0.0638	0.2402
GLM	0.0208	0.0384	0.0695	0.0890	0.0976	0.5542
LM	0.0140	0.0376	0.0716	0.0795	0.0977	0.3030
ANN	0.0247	0.0370	0.0488	0.0694	0.0857	0.2653
SVR	0.0309	0.0456	0.0498	0.0614	0.0703	0.1351
KNN	0.0310	0.0450	0.0487	0.0595	0.0566	0.1688
DTR	0.0438	0.0438	0.0438	0.0618	0.0641	0.1439
RF	0.0370	0.0454	0.0500	0.0642	0.0701	0.1769
RF*	0.0367	0.0436	0.0507	0.0639	0.0678	0.1856
			Batch 3			
y	0.0303	0.0417	0.0531	0.0563	0.0657	0.1176
GLM	0.0303	0.0440	0.0576	0.0606	0.0702	0.1020
LM	0.0251	0.0453	0.0598	0.0596	0.0693	0.0888
ANN	0.0380	0.0419	0.0599	0.0574	0.0659	0.0893
SVR	0.0409	0.0515	0.0568	0.0566	0.0626	0.0700
KNN	0.0389	0.0544	0.0591	0.0592	0.0657	0.0777
DTR	0.0478	0.0478	0.0584	0.0576	0.0673	0.0673
RF	0.0405	0.0504	0.0558	0.0565	0.0626	0.0743
RF*	0.0405	0.0491	0.0547	0.0562	0.0636	0.0760

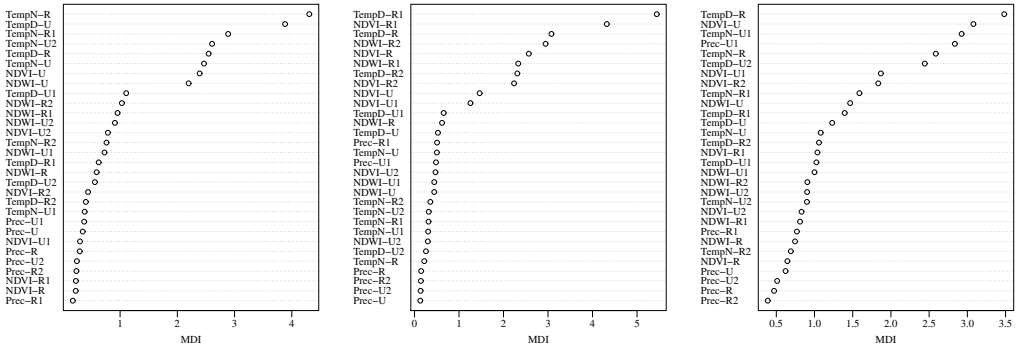


(a) Batch 1 (b) Batch 2 (c) Batch 3

FIGURE 3.6: Time series of the actual (points) versus predicted (lines) values on validation data, using only a subset of the prediction algorithms in Fig. 4: ANN, SVR, RF and RF*.

NDVI-U, TempN-U1, Prec-U1, TempN-R, TempD-U2, NDVI-U1, NDVI-R2. The improved control actions in Batches 2 and 3 are mostly responsible for the differences in important variables across all Batches. Only TempD-R is commonly selected in all the Batches. In addition, we can see that temperature features extracted from the rural surface zone provide the highest quality of information to our models in all Batches: TempN-R, TempD-R1 and TempD-R for Batches 1, 2 and 3, respectively. This is consistent with previous studies which show that non-artificial surface characterizations of the environment are more informative for predicting *Ae. aegypti* vector population [53, 54], and that the weekly (or daily) temperature is the most important environmental condition affecting the development of *Ae. aegypti* [106]. The non-synchronous effects of temperature is also seen in the selection of both TempN-R and TempN-R1 in Batch 1, and TempD-R, TempD-R1 and TempD-R2 in Batch 2. Of all eight selected subset variables, six of them lagged in both Batches 2 and 3, compared to only two in Batch 1. This shows that most of the environmental effects on vector population during the improved vector control regimes come from non-synchronous compounding effects. Six temperature variables are selected in Batch 1. This number reduced to three in both Batches 2 and 3. We deduce from this that the effect of non-temperature environmental variables increases during the improved control actions regime.

Figure 3.8 presents the relationship between the mosquito population y and the standardized values of environmental features selected for RF*. It is seen that the relationships vary across the different features (x). For example, Figure 3.8(a) shows that in Batch 1, higher values of all six selected temperature variables, between their mean and two standard deviations above their mean, result in larger numbers of mosquitoes. Also, for Batches 2 and 3, as shown in Figures 3.8(b) and 3.8(c), even with the improved anthropological interference due to better control actions, when the three selected most informative temperature variables in each case synchronously rise above one standard deviation from their mean, there is a rapid increase in the vector population. These results are in accordance with laboratory studies presented in [121, 122]. These works show that at higher temperatures below $40\,^{\circ}\mathrm{C}$, the development life



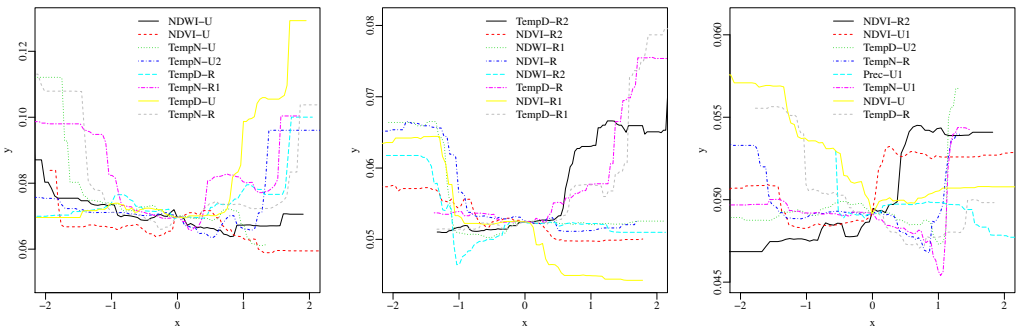
(a) Batch 1 (b) Batch 2 (c) Batch 3

FIGURE 3.7: Average values of MDI considering 50 RF replicas.

cycle of *Ae. aegypti* is accelerated, thus increasing the vector population. In addition, the extrinsic incubation period (EIP) of *Ae. aegypti* – the time interval between virus agent acquisition by the mosquito vector and the moment it is able to transmit it to humans – reduces at high temperature [94, 123]. The results in Figure 3.8(b) also show that the improved control program in starting from Batch 2 is effective except in cases of synchronous extreme values of important temperature variables. This observation shows pointers that can be used to further improve the efficacy of the control actions in subsequent years.

Regarding non-temperature environmental effects, the relation curves for Batch 3, Figure 3.8(c), also show that at about -0.5 standard deviation below the mean of Prec-U1, there is a spike in the vector population. As reported in [54, 121], in regions where primary larval sites are in rain-filled containers, rainfall has been shown to positively correlate with larval and adult mosquito abundance. Further studies have corroborated the importance of humidity and vegetation conditions in abundance and reproduction of mosquito species, with longer dry season and lower relative humidity resulting in higher egg mortality [124]. Fully developed tree canopies by providing shade can reduce evaporation of hatching water, and can also increase near ground humidity, thus increasing density of *Ae. aegypti* mosquito larvae [42, 51]. Stronger positive effects of vegetation conditions can be observed through the effects of NDVI-R2 and NDVI-U1 in Batch 3.

Finally, and in general, this research shows, based on our study data, that the random forest model performs better than all the other models for predicting the mosquito population from EO-based variables. By selecting the most informative features using the MDI measure, it is possible to obtain a much less complex model with comparable (or even better) results, labeled in this work as RF*. Additionally, with the MDI ranking it is possible to understand and explain the rationale for the model variables and analyse the effects of each EO variable on the vector population. This better understanding is gained by considering the relation graphs based on RF*, which provide a visualization of the the nonlinear relationships among the variables.



(a) Batch 1 (b) Batch 2 (c) Batch 3

FIGURE 3.8: The relation between the mean of mosquitoes (y) and each covariate (x).

3.6 Chapter Conclusions

In this Chapter, a procedure for modeling the temporal population distribution of female *Ae. aegypti* mosquitoes starting from freely available EO data has been presented. Specifically, MODIS and GPM EO data products were used to obtain estimates of temperature, precipitation, moisture and vegetation conditions.

The main outcome is a procedure for RF-based explainable modeling of the target quantity using EO data. A quantitative measure of the variable importance (MDI)— wrapped in RF — was also used to extract the most informative environmental features, and obtain a less complex but still accurate predictive RF model, labeled RF*. To prove their robustness, the resulting models were compared to other machine learning models including SVR, ANN, KNN and DTR, as well as statistical models, such as LM and GLM.

The proposed RF-based approach is capable of mapping the complex relationship among the EO variables and the vector population. Furthermore, the features selected thanks to the MDI ranking may be empirically interpreted, and provide hints about the relationship among vector population and these environmental conditions from an operational point of view.

Further studies can consider applying the proposed method to longer time scales for monthly or annual modeling. Such studies can incorporate other considerations including seasonal impact of the different EO variables features.

Chapter 4

Modeling Dengue Vector Population with Earth Observation Data and a Generalized Linear Model

4.1 Introduction[1]

Chapter 3 presented a methodology for municipality level temporal modeling of *Ae. aegypti* vector population based on environmental conditions estimated for EO data features. As part of that methodology, random forest (RF) and its embedded features ranking metric (the mean decrease impurity - MDI) have been applied for both modeling and model selection. While that study contributes progress to the problem which is being addressed, it still leaves some gaps. One of them is that the use of the relation curves for the explanation of the variables effects is not very intuitive and requires some domain expertise. In addition, while MDI provides variable importance ranking it doesn't spècify the directionality (positive or negative) of the effects of the variables.

Instead of RFs, generalised linear models (GLM) may be used. They provide model equations that are understandable and intuitive to explain, usually without the same prediction power as RFs. The leading question in this chapter is, therefore, the following one.

> "How can we improve a GLM towards obtaining machine learning (ML) quality results, while also having the capability to explicitly interpret the causality in the model?"

[1]This chapter has been published as a standalone paper as O. Mudele, A. C. Frery, L. Zanandrez, A. Eiras, P. Gamba, "Modeling dengue vector population with earth observation data and a generalized linear model." Acta Tropica, doi: https://doi.org/10.1016/j.actatropica.2020.105809, vol. 215, mar. 2021.

To this aim, in the following sections a weighted Poisson GLM is proposed [125]. The results show that this technique produces comparable prediction power than state-of-the-art ML methods, thus answering our scientific question.

4.2 Material

4.2.1 Study and field data

This study is applied in the same location (Vila Velha) and with the the same data (MI-Aedes®) which has been introduced in Section 3.2 in the previous chapter. Refer to Table 3.1 for more details.

In particular, for this study, MI-Aedes data in Vila Velha in 2017 (epidemiological weeks 17–52 = 36 weeks) and 2018 (epidemiological weeks 1–40 = 40 weeks) are used. These data correspond to Batches 1 and 2 in the study presented in Chapter 3, i.e Batch 3 data are not considered in this chapter.

A noteworthy difference in the field data processing in this chapter is that for each mosquito traps split (training/validation), the sum of mosquito population (rather than average (as in Chapter 3) are considered. The reason is to ensure that the response variable is a natural number, in accordance with the requirements of a Poisson GLM. Moreoevr, spatial data pooling (a sum, in this case) helps to mitigate the problem of site-level sparsity, i.e. the fact that at a given time point many individual traps report zero values. Consequently, the response variable used throughout this chapter is the sum of the mosquitoes per time point at municipality level. This response variable is then modelled based on EO features representing environmental conditions as derived in Section 4.2.2.

4.2.2 Environmental variables

As in Section 3.2, EO data sets obtained fron (MODIS data are used as proxies to humidity, temperature and vegetation conditions. To better represent environmental conditions covariates, these EO-derived features are split between the same “urban” and “rural” zones defined in the previous chapter and visualized in Figure 3.2.

All data preparation and features extraction processes are performed using the `python` [126] programming environment, with the `gdal`, `numpy` and `pandas` libraries.

4.3 Modeling

We model the mosquito population (Y) as a function of climatic covariates (X) subject to errors. For this, we consider a Poisson Generalized Linear Model (GLM), and we compare the logarithm, identity, and square root link functions for fitting the model to our data. We then applied stepwise regressions based on the Akaike Information Criterion (AIC) to select the model of better fit and best quality input features. We derived the residuals $e_i = y_i - \widehat{y}_i$ of our initial fully fitted GLM for the best performing link function, and obtained weights $w_i = |e_i|^{-1}$ to fit a weighted GLM (GLM-W).

We further considered a more parsimonious GLM-W model: GLM-W*, also selected using AIC based stepwise regression. We implemented the GLM in code using the `glm` function of `MASS` package in the `R` programming environment. The weighting is implemented using the `weights` parameter of this function. For prediction benchmarking, we considered ML algorithms: RF and SVM, and a more parsimonious RF, labeled RF*, fitted with the most informative climatic variables as ranked by the Mean Decrease Impurity (MDI) — a feature quality importance ranking technique for RF [110]. We chose as informative those variables with the highest MDI. We used `e1071` and `randomForest` packages to implement SVM and RF, respectively, in `R`.

We assessed the prediction quality with AIC and Mean Absolute Error (MAE) [127] for all GLMs, and only the latter for the ML methods. Smaller values of AIC and MAE are desirable for better model quality.

4.3.1 Problem Formulation

Let $X \in \chi \subset \mathbb{R}^p$ be an input vector related to p climatic covariates and $Y \in \mathbb{N}$ our mosquito population count variable. The goal is to estimate a function, m, such that $m(x) = \mathbb{E}(Y \mid X = x)$. This goal is formulated in the form of the following generalized linear model:

$$\widehat{y}_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{4.1}$$

where superscript T denotes transposition; $\widehat{y}_i$ is the predicted mosquito population for the ith epidemiological week; k is the number of climatic covariates used to fit the model; x_{ij} is the value of the climatic variable j in the ith week; β_0 is the value of $\widehat{y}$ when all the climatic variables equal zero; and β_j is the gradient of $\widehat{y}$ with respect to the jth climatic variable. The observed mosquito population y_i is an instance of the random variable Y_i with a mean estimated as $\widehat{y}$ by our model.

4.3.2 Poisson Regression Modeling

This is one of the models commonly used in epidemiological studies. It is one of the families of GLMs, suitable for cases of discrete non-negative dependent variables with non-Normal distribution [128, 129].

A Poisson regression model assumes that the response Y_i follows a Poisson distribution:

$$\Pr(Y_i = y \mid \mu_i) = \frac{e^{-\mu_i}\mu_i^y}{y!}, \tag{4.2}$$

where $\mu_i = \mathbb{E}(Y_i \mid \mathbf{x}_i) \geq 0$ is the mean and variance of Y_i. The relationship between μ_i and the linear predictor is given by a link function, g, such that $g(\mu_i) = \mathbf{x}_i^T\beta$.

In this work, models derived with the identity, logarithm and square root link functions as shown in eqs. (4.3a) to (4.3c), respectively, are applied and benchmarked.

$$\mu_i = \mathbf{x}_i^T\beta, \tag{4.3a}$$

$$\mu_i = \exp\left(\mathbf{x}_i^T\beta\right), \tag{4.3b}$$

$$\mu_i = \left(\mathbf{x}_i^T\beta\right)^2. \tag{4.3c}$$

4.3.3 Model selection with Akaike Information Criterion

Model selection refers to a group of techniques applied to choose a minimal-sized subset of high-quality features to shorten training time, improve model performance, and reduce the cost of collecting and processing data. For this work, we used AIC as a selection criterion because of its simplicity and versatility [130].

AIC is an information-theory-based model selection criterion and quality metric which takes the form of a negative log likelihood plus a penalty term:

$$\text{AIC} = -2\ln L(\widehat{\theta}) + 2k, \tag{4.4}$$

where $L(\widehat{\theta})$ is the maximized likelihood function, $-2\ln L(\widehat{\theta})$ is the lack of fit component, and k is the number of independent covariate features in the model, i.e., its complexity. The model with the lowest AIC is chosen as the best [130, 131].

The best quality features subset $S \in X \subset \mathbb{R}^k$ with AIC to obtain the model $m(S)$ are selected by the condition:

$$m(S) = \arg\min_k \text{AIC}(m(X = x^{(k)})), \tag{4.5}$$

where $x^{(k)}$ is a combination of k features in X. Exhaustively searching the space of our model $m(X)$ for all possible subset models $m(S)$ in order to select the best has exponential order of magnitude: $\mathcal{O}(n) = 2^n + 1$. Consequently, there is a need for greedy solutions that make a locally optimal choice at each stage of the search. Two commonly used methods are forward selection and backward elimination.

For this study, stepwise selection [132] is applied, This method combines forward and backward selection by finding and selecting the model with the smallest AIC by removing or adding variables at every step from all the explanatory variables. Being a greedy approach, this technique is susceptible to selection errors, but is fast and easy to implement.

4.3.4 Machine Learning models

To compare with the approach presented in Chapter 3, a Random Forest (RF) model is fitted. MDI ranking and parsimonious RF selection and fitting are also performed and compared with the proposed GLM. In addition, to provide an extra baseline, SVM is also implemented: using a radial basis function kernel with tuning parameter set as $1/(\text{number of features})$, tuned via the ε-regression method.

4.4 Experimental Results

The resulting EO covariates feature labels used in Chapter 3 (Section 3.5) are retained, i.e. "urban" variables: NDVI-U, NDWI-U, TempD-U, TempN-U, and Prec-U, and "rural" variables: NDVI-R, NDWI-R, TempD-R, TempN-R, and Prec-R, where TempD is the daytime temperature, TempN is the night-time temperature, and Prec is precipitation. Considering the lagged variables up to two weeks, our fully fitted model comprises 30 EO derived covariates. The suffixes "#1" and "#2" indicate one and two weeks lagged variables, respectively. For example, "NDVI-U2" denotes the two-week lagged urban NVDI variable.

Table 4.1 reports the mean, median, standard deviation (SD), minimum (Min) and maximum (Max) of the EO data variables together with the mosquito population (y). These measures, as with all further results, are reported differently for both years (2017 and 2018) to account for differences due to an increase in control actions in 2018 (See Table 3.1). As shown in the table, the median total mosquito population is 96.5 in 2017 and 73.0 in 2018. This represents a $24.35\,\%$ decrease in 2018, most likely due to improvements in control actions in 2018 as reported by the data providers and also validated in Chapter 3. This fact justifies our decision to analyze the data for each year separately.

TABLE 4.1: Descriptive measures of the EO-based environmental variables of interest (#U and #R denote "urban" and "rural", respectively. TempD: daytime temperature, TempN: night-time temperature, Prec: Precipitation).

Variable	Mean	Median	SD	Min	Max
			Year 2017		
y	111.984	96.500	84.854	17.000	454.000
NDVI-U	0.229	0.229	0.015	0.198	0.263
NDVI-R	0.728	0.730	0.039	0.583	0.798
NDWI-U	0.005	0.004	0.022	-0.050	0.075
NDWI-R	0.545	0.544	0.038	0.476	0.664
TempD-U	31.284	30.980	4.100	22.597	37.065
TempD-R	27.753	27.711	3.016	20.500	33.860
TempN-U	20.077	21.069	2.305	13.494	22.498
TempN-R	18.924	18.955	1.865	13.628	23.146
Prec-U	3.272	0.463	7.974	0.000	43.250
Prec-R	3.197	0.938	7.087	0.000	37.175
			Year 2018		
y	100.250	73.000	75.622	42.000	400.000
NDVI-U	0.250	0.247	0.018	0.209	0.310
NDVI-R	0.748	0.755	0.026	0.680	0.792
NDWI-U	0.003	-0.002	0.036	-0.059	0.107
NDWI-R	0.592	0.597	0.052	0.476	0.693
TempD-U	31.862	31.577	3.478	26.476	38.839
TempD-R	27.514	26.943	2.687	23.785	33.816
TempN-U	21.294	21.019	2.209	16.389	25.332
TempN-R	20.196	19.654	2.086	16.507	23.781
Prec-U	2.471	0.065	4.928	0.000	23.080
Prec-R	1.753	0.099	3.468	0.000	16.095

Units: °C: degrees Celcius, $\mathrm{mm\,h^{-1}}$: millimeters per hour

Table 4.1 is similar to what has been presented in Table 3.2 except for the changes in the response variables statistics since the sum of the mosquito population (rather than average) is considered in this study. Hence the same observations which have been drawn in Section 3.5 about the statistics of the EO covariates remain valid. For example, the mean and median land surface temperature in the urban area are higher than in the rural zone. These is because vegetated surfaces absorb more heat energy and contain more moisture.

4.4.1 Statistical regression approach

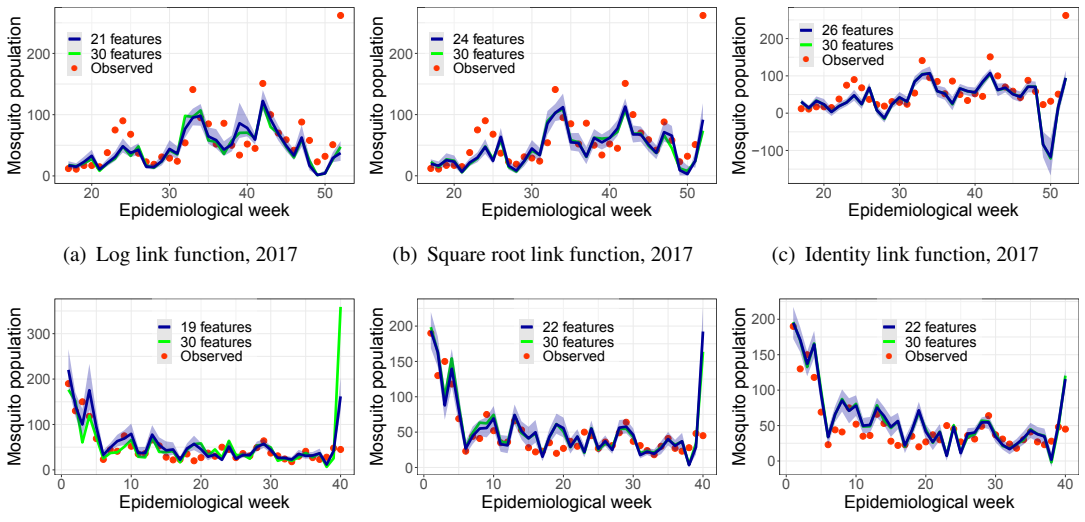
Figure 4.1 presents observed and fitted values for fully fitted and stepwise selected GLMs across the three link functions. The link functions which have been applied are hyperparameters of the GLM and do not have any specific significance with respect to the mosquito population which is being modeled. However, it is important to know that while both the logarithm and square root

link functions introduce non-linearity into the model, the identity link function assumes linearity between EO covariates and the target vector population.

In Figure 4.1(c), it may be noted that the identity link predicted negative mosquito population values in 2017. This result shows a potential weakness in studies that model the population or oviposition activities of mosquito vectors using linear models with no non-linear link function. One of such studies can be found in [53]. The identity link function does not constrain the model response to positive values, resulting in the possibility to predict negative values with no physical meaning. The issue however, as seen in Figures 4.1(a), 4.1(b), 4.1(d), and 4.1(e), doesn't affect models with non-linear link functions. This goes to show that the relationship between the mosquito vector population and the predictor EO variables is better modeled non-linearly.

Table 4.2 presents the results for fully fitted and stepwise selected GLM with the three link functions. In 2017, using the log link function, 9 variables were dropped to produce the most parsimonious model compared to other link functions. More parsimony means a less complex model with less redundancy. The less parsimony shown here by the log link function actually results in better MAE and AIC values, meaning a better model of mosquito vector population. The implication of this is that model selection is a useful component towards finding good model fit for the task of this study. In addition, a log link function provides better non-linearity between the used EO data features and the mosquito population.

In 2018, even though the validation MAE does not improve by applying stepwise regression in the case of square root linked GLM, there is still a significant improvement in terms of the AIC



(a) Log link function, 2017 (b) Square root link function, 2017 (c) Identity link function, 2017

(d) Log link function, 2018 (e) Square root link function, 2018 (f) Identity link function, 2018

FIGURE 4.1: Observed (points) and predicted (lines) values with the three link functions and 95 % confidence intervals. The models with 30 features are the full models, while those with less are stepwise selected.

and parsimony of the model. In addition, again, by applying modeling selection to the GLM fitted with the log link function, 11 variables were dropped and the validation MAE is reduced by about 25%. This shows the robustness of the log link function across different control action regimes. Taking these results into consideration, all further models are fitted with the log link function.

TABLE 4.2: Quality measures obtained with three link functions when the model is fit with the full independent feature set

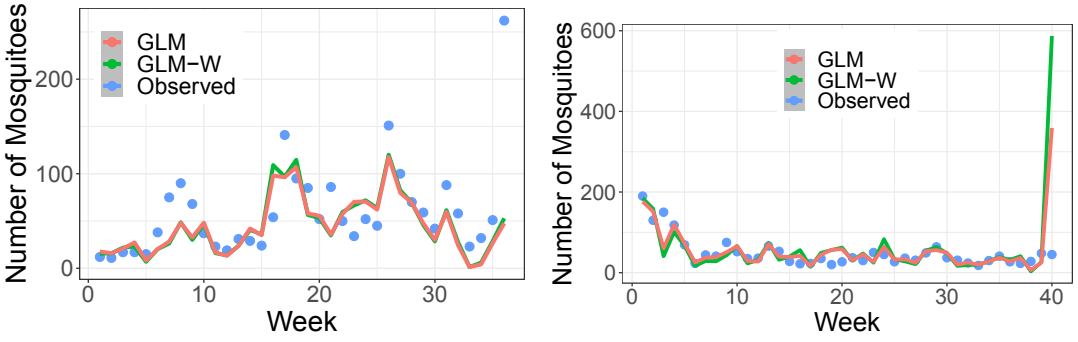
(a) Year 2017

Link	# features	AIC		MAE
log	30	272.264	Training	1.852
			Validation	25.467
	21	259.679	Training	2.905
			Validation	24.299
Identity	30	290.668	Training	4.682
			Validation	32.051
	26	283.077	Training	4.605
			Validation	31.794
Sqrt	30	278.748	Training	3.275
			Validation	24.539
	24	270.609	Training	3.711
			Validation	23.824

(b) Year 2018

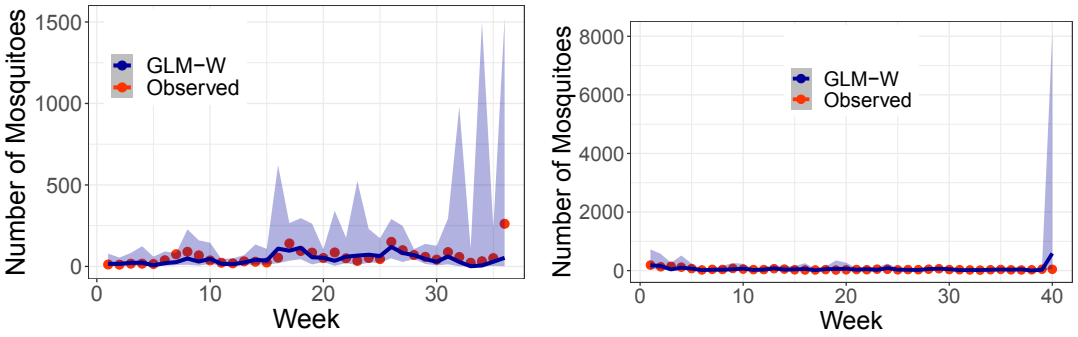
Link	# features	AIC		MAE
log	30	311.426	Training	4.386
			Validation	20.503
	19	298.087	Training	5.085
			Validation	15.557
Identity	30	307.098	Training	4.358
			Validation	17.307
	22	293.710	Training	4.594
			Validation	16.953
Sqrt	30	309.745	Training	4.590
			Validation	15.738
	22	296.475	Training	4.613
			Validation	16.296

The performance of GLM-W is compared with that of the the fully fitted log link GLM in the plots shown in Figure 4.2. In 2017, there are only small noticeable differences between the two models. In 2018, the only significant difference was in the 40th epidemiological week which is an outlier week. Ideally, the model will require more data to be able to capture the pattern in this particular week. One drawback of the model weighting is the increment in the uncertainty of the predictions as shown in Figure 4.3. Hence, a stepwise regression with AIC to obtain GLM-W* is useful to filter out features with the highest uncertainty and retain only the useful ones as obtained and shown in Figure 4.4.



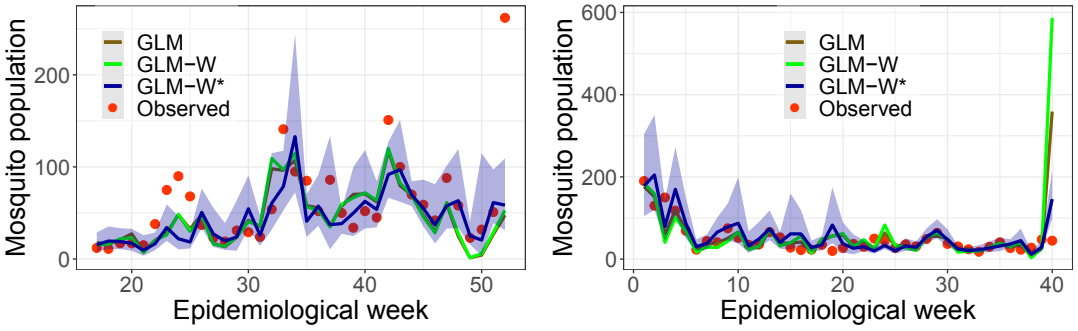
(a) 2017 (b) 2018

FIGURE 4.2: Time series of the observed values (points) and predicted values (lines) for GLM and GLM-W.



(a) 2017 (b) 2018

FIGURE 4.3: Time series of the observed values (points) and predicted values (lines) for GLM-W — with 95% confidence intervals.



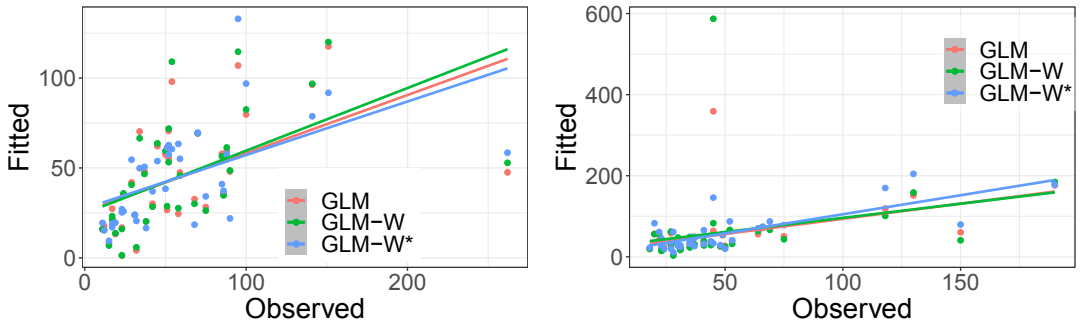
(a) 2017 (b) 2018

FIGURE 4.4: Observed and predicted values for GLM, GLM-W and GLM-W* — with 95 % confidence intervals for GLM-W*.

Considering GLM-W*, it can be seen that in 2017 (see Figure 4.4(a)), it has only six observed points not captured by the line of best fit and 95 % confidence interval. Alternatively, the stepwise selected GLM — log link without weights — has 18 fitted points outside its confidence interval in that same year (cf. Figure 4.1(a)). In 2018, as shown in Figure 4.4(b), only four observed points are outside the confidence interval, against eleven in the unweighted stepwise selected GLM (cf. Figure 4.1(d)). These results show that the proposed weighting approach along with model selection improves the quality of resulting vector population model fit, especially when we consider the confidence interval.

Figure 4.5 shows the yearly scatterplots of fitted versus observed values. From that figure, one can conclude that there is stronger correlation between observed and fitted values in GLM-W*, and, thus the weighted regression model performs better.

Table 4.3 presents measures of the quality of the results of the GLM models fitted so far. This table reveals that for both years, GLM-W* produces a good combination of best parsimony and lowest AIC and MAE among all fitted models. Starting from a full GLM, through GLM-W to GLM-W*, the AIC of the resulting model can be reduced by more than a factor of five in



(a) 2017 (b) 2018

FIGURE 4.5: Scatterplots and regression lines of observed and predicted values for GLM, GLM-W and GLM-W*.

2017 and a factor of three in 2018. The number of input environmental covariates selected are also reduced by a factor of three in 2017, and more than a factor of two in 2018. Besides, the gap between training and validation quality (MAE) is reduced with GLM-W*, showing that GLM-W* is less prone to overfitting.

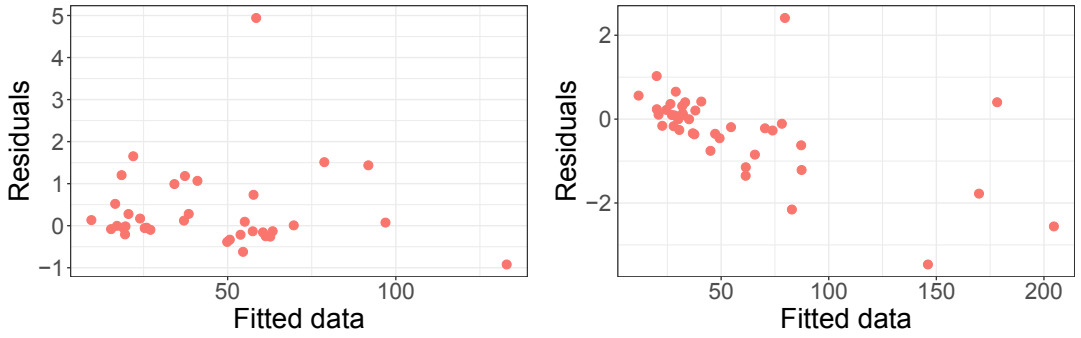
TABLE 4.3: Quality measures for GLM, GLM-W and stepwise selected GLM-W (GLM-W*)

Year	Model	# features	AIC		MAE
2017	GLM	30	272.264	Training	1.852
				Validation	25.467
	GLM-W	30	86.790	Training	2.321
				Validation	25.510
	GLM-W*	10	53.658	Training	13.916
				Validation	23.297
2018	GLM	30	311.426	Training	4.386
				Validation	20.503
	GLM-W	30	114.228	Training	5.368
				Validation	29.748
	GLM-W*	14	88.997	Training	8.666
				Validation	19.334

The quality of GLM-W* is further assessed by analysing the residuals. The expectation for a good model is that the residual plot should not follow a known distribution. Our models display this property as presented in Figure 4.6, thus showing that the resulting model is good in both years.

4.4.2 Comparison with Machine Learning

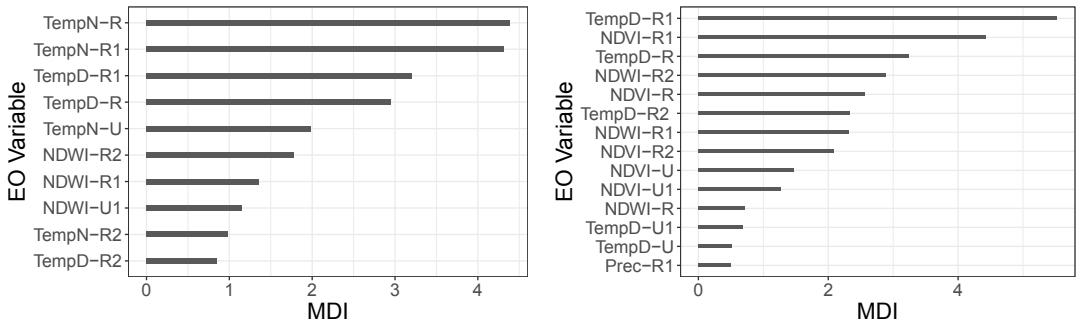
The stepwise selected models with the lowest AIC values, GLM-W*, in 2017 and 2018, have 10 and 14 covariates. For comparability purposes, RF* is fitted in 2017 with the top 10 MDI ranked variables, and in 2018 with the top 14 MDI ranked variables for the year. In this way, RF* for each year with the same number of covariates features as the corresponding GLM-W* is obtained. Figure 4.7 shows the average MDI for each top-ranked selected variable considering



(a) 2017 (b) 2018

FIGURE 4.6: Adjusted residuals plot for GLM-W*.

50 RF replicas per year. Since MDI is embedded in the random forest, it is a model-driven variables ranking metric and thus does not necessarily select the same variables as with AIC for GLM-W*.



(a) 2017 (b) 2018

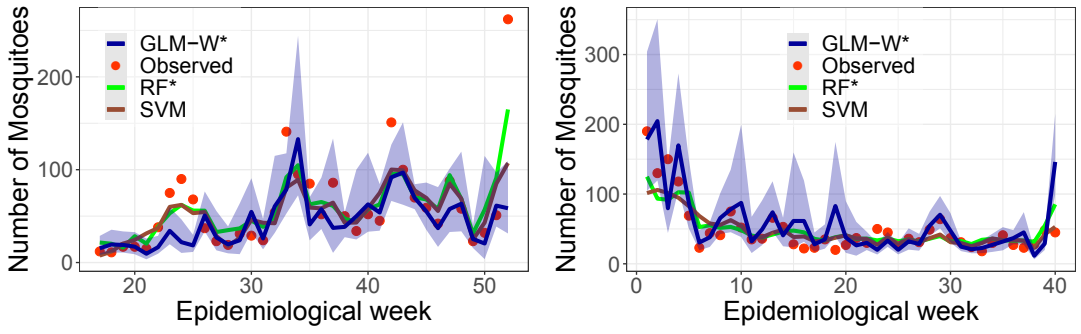
FIGURE 4.7: Average Mean Decrease Impurity (MDI) for selected variables considering 50 replicas of RF. Higher values of MDI signify higher relevance for the model.

Table 4.4 presents quality measures for these models. In both years, RF* has the best prediction compared to RF and SVM, considering parsimony and MAE. Also, it performs better than GLM-W* considering the absolute fit measured by MAE.

TABLE 4.4: Quality measures for SVM, RF, RF*

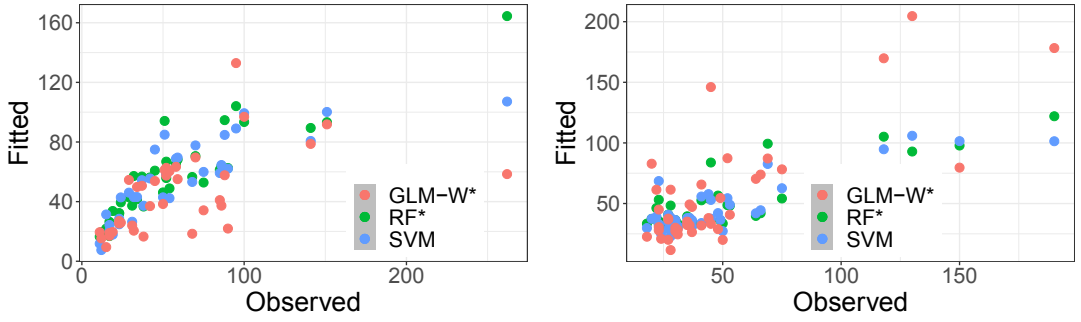
Year	Model	# features		MAE
2017	SVM	30	Training	10.271
			Validation	17.730
	RF	30	Training	8.965
			Validation	17.919
	RF*	10	Training	8.318
			Validation	**17.060**
2018	SVM	30	Training	8.846
			Validation	13.282
	RF	30	Training	8.345
			Validation	14.505
	RF*	14	Training	8.127
			Validation	14.585

Figure 4.8 presents a comparison between SVM, RF* and GLM-W* in line plots, with 95 % confidence intervals shown for GLM-W*. Here, it can be seen that the predictions by all ML algorithms tested, and specifically RF*, is better than that of GLM-W*. As already mentioned and validated in Chapter 3, ML methods generally show better predictive capabilities than statistical models for most applications. However, taking a deeper look, it can also be seen that the 95 % confidence intervals of GLM-W* for both years include most of the predicted points by the ML methods. For example, in 2017, the line of best fit and confidence band of GLM-W* only fails to capture the predictions by RF* in epidemiological weeks 24, 25 and 52. The ML methods' predictions for these few weeks, especially in the 52nd week, are mostly responsible for the much improved MAE, thus making the metric unfairly biased. Also, even though RF* performs better in the 52nd week, the observed value for this week is not reached by any of the models because it is an outlier. A way to improve prediction for this week is to include more data samples with such very high observed values to derive a suitable relationship based on more examples. In other weeks in 2017, however, GLM-W* produces better or comparable predictions. Specifically, in epidemiological weeks 19 to 21, 27 to 29, 44 to 46 and 48 to 51, GLM-W* fits better.



(a) 2017 (b) 2018

FIGURE 4.8: Observed and predicted values for SVM, RF* and GLM-W* — with 95 % confidence intervals for GLM-W*.



(a) 2017 (b) 2018

FIGURE 4.9: Scatterplot of observed and predicted values for SVM, RF* and GLM-W* — with 95 % confidence intervals for GLM-W*.

Figure 4.8(b) shows that in 2018, GLM-W* produced better fit in epidemiological weeks 1, 5 to 7, 9, 13, 28, and 29, and its predictions in the remaining other weeks compare very well to ML. The overestimated prediction can explain the better MAE measure obtained by ML models in this year by GLM-W* in the last observation week. Thus, as demonstrated, GLM-W*, while being less complex and more explainable, follows the observed value for, at least, as many prediction points as the ML algorithms. Its confidence intervals include predictions by the ML methods for most weeks — in 2017, 31 out of 36 weeks, and in 2018, 36 out of 40 weeks. This makes it sufficient to provide an excellent statistical explanation of the mosquito population count.

Furthermore, from the scatterplots presented in Figure 4.8, it can be seen that for both years that, indeed, with the exemption of a few points, GLM-W* does produce a comparable model. It is important to note that the explanability and intuitiveness of working with GLM makes it useful in spite of the possibility quality compromise. This is specifically true in the case of epidemiological application where variable importance and effects directionality are informative for planning control activities. From the results which have been presented so far, GLM-W* offsets, or perhaps removes quality compromise one has to make for applying a GLM for the specific problem case addressed in this study.

Table 4.5 presents a summary of the observed and fitted data: mean, median (Q2), minimum (Min), maximum (Max), first (Q1) and third (Q3) quartiles. The mean and median values report the balance of the model predictions while the Min and Max reveal how well the model captures spikes and valleys of the observed data. In 2017, the predicted Max values by all GLMs are better than what is obtained with SVM, and in moving from GLM (unweighted with log link function) to GLM-W and finally to GLM-W* predicted Min is always improved. RF*, however, produces the best Max value prediction, as discussed above for the 38th week in Figure 4.8(a). Another significant improvement that is obtained in 2017 from weighting our GLM before stepwise selection with AIC is in the predicted minimum value. As a result, GLM-W* produces better minimum predicted value than ML models and improves concerning the unweighted GLM. RF and RF*, on the other hand, overestimate the minimum value. In 2018, GLM-W* produced better Min, Q1, median, and Max than RF, RF* and SVM. While the ML models produce the best mean predictions, they significantly underperform in predicting the maximum observed response variable, which corresponds to the first week of this year (cf. Figure 4.8(b)). GLM-W* is the only model that makes an adequate prediction in this week.

4.5 Discussion

Environmental variables affect *Ae. aegypti* population directly and indirectly. Temperature, for example, has been shown to foster faster development of immature forms into adult mosquitoes [45],

TABLE 4.5: Summary of the observed and fitted data (all GLMs fitted with log link function).

		Min	Q1	Q2	Mean	Q3	Max
2017	y	11.00	27.75	50.50	59.33	77.50	262.00
	GLM	1.23	23.04	38.92	44.94	60.21	117.61
	GLM-W*	9.48	21.58	39.72	45.09	58.98	132.95
	SVM	7.51	35.79	54.96	54.27	70.92	107.15
	RF	21.27	41.58	56.46	58.60	68.63	159.06
	RF*	17.64	36.40	54.66	58.07	68.68	163.08
2018	y	18.00	27.75	36.50	48.95	50.50	190.00
	GLM	6.51	27.68	38.74	55.57	57.16	358.96
	GLM-W	3.24	27.68	39.48	60.50	57.53	587.05
	GLM-W*	11.63	28.79	37.79	57.00	71.25	204.63
	SVM	22.97	33.46	37.78	46.47	53.35	105.92
	RF	28.54	34.85	38.72	48.69	53.07	120.74
	RF*	28.89	33.57	38.21	48.51	52.71	126.70

whereas precipitation influence the availability of breeding sites for the reproduction of the species, which could lead to a delayed increase in the population after rainy periods [133]. At the same time, higher temperatures increase evaporation, thereby enhancing humidity, which has been associated with increased feeding activity, survival, and egg development [134]. Land cover and vegetation can increase humidity, whereas decreasing land surface temperature, having indirect effects on vector dynamics [135].

[136] used a generalized additive model applied to adult mosquito data from a subtropical city and found that despite being associated to an increased development of the vector, humidity levels above 79 % presented negative correlation with vector population. This was explained as accounting for specific conditions presented in the area where data were collected, which had higher humidity when the temperature was lower, suggesting that temperature had a more substantial effect on vector survival. Thus, the relationship between *Ae. aegypti* and the environment may depend on complex interactions and correlations among the variables, which can affect the vector ecology differently, depending on their range. Here, we conduct sensitivity analyses to understand these complex relationships for the city of Vila Velha in the same vein.

Our previous analysis shows that GLM-W* is a good model for predicting the population of *Ae. aegypti* in Vila Velha. The advantage of using a GLM against ML (RF or SVM) is that it provides equations with which the model can be understood by public health managers and integrated into real-life public health systems. Table 4.6 shows the fitted models for the two years under analysis.

In 2017, NDVI-U made the highest positive contribution to the *Ae. aegypti* population. Also, NDWI-R1, NDVI-R2, and NDWI-U2 showed positive influence. The greatest negative influence on the vector population is by TempN-U. In 2018, NDWI-R1 and NDWI-U1, respectively, made the highest and second-highest positive contributions, while NDWI-R2, NDWI-U2 and NDVI-R1, in that order, made the highest negative contributions. We see in both years that all selected NDWI (humidity proxy) variables have a relatively strong influence on the predicted responses. For instance, NDWI-R1 maintains a very strong positive influence in both years. In general, the

TABLE 4.6: Fitted models for $\widehat{\mu} = \exp\left\{\beta_0 + \sum_{j=1}^{k} \beta_j x_j\right\}$

(a) Parameters for 2017

β_j	x_j
-15.828	
0.188	Prec-R
0.213	TempD-R1
0.286	TempD-U
8.924	NDVI-R2
9.384	NDWI-R1
11.019	NDWI-U2
14.379	NDVI-U
-0.180	TempD-U2
-0.180	Prec-U
-0.233	TempN-U

(b) Parameters for 2018

β_j	x_j
-14.491	
0.169	TempD-R1
0.598	TempN-R2
6.649	NDWI-U1
8.200	NDWI-R1
-0.063	Prec-R1
-0.171	Prec-R
-0.233	TempN-R
-0.251	TempN-U1
-9.350	NDWI-U2
-11.379	NDVI-R1
-0.267	TempD-R2
-13.924	NDWI-R2

equation coefficients, β_j, show that the humidity variables in Vila Velha are the most influential on *Ae. aegypti* population in both years.

The literature corroborates these results. For instance, [137] compared adult *Aedes* mosquito data collected over 12 weeks to oviposition data collected in the same period and around the same points. This study also utilized the sticky cards MosquiTraps® for adult *Aedes* population surveillance. The data were correlated with temperature and precipitation obtained from meteorological weather sources, showing positive contributions from temperature and adverse effects from precipitation on the adult vector population. Our study shows the same effect in Vila Velha, as revealed by Table 4.6. For example, we see that Prec-U and Prec-R — the non-lagged "urban" and "rural" surface precipitation variables — are both selected as informative variables. This highlights the effect of instantaneous precipitation on the mosquito population, even across different control actions regimes. Cumulatively, there are more negative influences among precipitation variables selected in both years. This agrees with the results from Ref. [137], in which the authors explained that MosquiTraps® capture more mosquitoes in the dry season probably because they compete with other potential breeding sites that were filled with water in the rainy period.

In consonance with these effects, our results show that TempD-R1 and TempD-U both have a positive influence on the vector population, regardless of differing vector control conditions. The results in also shows the positive effects of temperature on the efficiency of adult vector population growth.

A model-based fully on EO data makes it possible to account for microclimatic ("urban" vs "rural") effects. Our results show that a significant number of rural variables — six out of ten in 2017 and nine out of fourteen in 2018 — are selected as part of the most informative groups in both years of our study. Perhaps the reason for this is that rural variables do not suffer

too much variation during the day, and are less susceptible to the volatility of urban surface conditions. Using this knowledge, operational surveillance and control action systems, including MI-Aedes®, can scale their impact by relying on freely available and globally consistent EO data sources, and, by this, can also use rural surface variables to improve the information available to their models.

Public health managers will have the GLM equation(s) at hand, so they will not need weekly complete mosquito data to make forecasts. Hence, the models can be used to plan control activities without full mosquito data. This is quite a common practice. For example, in many Brazilian municipalities, including Vila Velha (as reported by Ecovec), surveillance activities are performed by larvae surveys. which are very time and cost intensive. The costs of operational surveillance programs can be roughly cut by a half by alternating weekly planning control activities and trap inspections. The model(s) presented here may be used in such situations.

Another limitation is the generalization of the models. The variable coefficients, β_j, we obtained for Vila Velha are not generalizable into all areas since they depend on environmental conditions that change in different climate zones. Future studies can, however, investigate how well these coefficients, β_j, generalize to other municipalities with similar meteorological and climatic profile as those in Vila Velha.

4.6 Chapter conclusions

This chapter introduces a Poisson GLM to model the temporal relationship between adult *Ae. aegypti* population and relevant biotic and abiotic environmental variables, such as precipitation, temperature, humidity, and vegetation condition. The quality of the model improves its ability to fit the data, its robustness, and its interpretability, With respect to the RF-based approach presented in chapter 3.

This model has been applied to explain the effects of the environment of vector population across two different control regimes. This type of analysis may improve the control actions by local authorities. Moreover, the meaning of the model equation coefficients is intuitive and can be used to improve planning and resource allocation with the aim of a more efficient surveillance.

Chapter 5

Dengue Vector Population Forecasting using Multi-source Earth Observation Products and Recurrent Neural Networks

5.1 Introduction

Chapters 3 and 4 have explored ways to model and understand the interaction among environmental variables and *Ae. aegypti* mosquito counts in the temporal domain. Other studies, including [42, 54] and [53], have performed similar studies in either the temporal or the spatial domain. However, none of these studies combine these two dimensions into a single pipeline, i.e, nine of the performs a real spatio-temporal modeling. Finally, these studies are all devoted to "nowcasting", as opposed to forecasting, and only at the municipality level.

In this chapter, a first attempt is proposed at a time series *Ae. aegypti* forecast (one-week-ahead), which is spatially disaggregated at the neighborhood (urban block) level. To this aim, the same freely available EO satellite image products as in Chapters 3 and 4 are used for the estimation of the environmental features of interest. Forecast models, as opposed to nowcasting models, serve to enable operational disease outbreak surveillance systems to anticipate and better plan for future disease spread events.

Specifically, the above mentioned model is obtained starting from state of the art models, routinely applied to *e.g.*, economics [138], weather, and environmental state predictions [139]. Some of the most frequently used algorithms include autoregressive moving average models (ARMA), autoregressive integrated moving average models (ARIMA) [140], random forest, and

more recently neural and deep learning networks [141]. Among the last class of approaches, Recurrent Neural Networks (RNNs) [142, 143] has been used for sequential data modeling, and show great capability to capture multivariate non-linear interactions in data sequences [144]. However, RNNs suffer from the problem of vanishing and exploding gradients over long sequences [140]. As a result, long short-term memory (LSTM) [145] and gated recurrent unit (GRU) [146], which are variants of RNN designed to mitigate these setbacks, were designed. They have found successful application in many fields, *e.g.*, machine translation [147], speech recognition [148], and other time series forecasting tasks [141]. Recent state-of-the-art applications consider the use of LSTM and GRU in an encoder-decoder fashion [141]. All these works suggest that using RNNs (with LSTM and GRU) could be an efficient way to tackle the problem at hand.

Accordingly, the research question in this chapter is whether an accurate geographically distributed time series prediction for *Ae. aegypti* numbers at the neighbourhood level is possible using EO data as inputs and using RNNs. To try and find an answer, we start from the analysis in [149], which shows that in a group of concurrent mosquito population time series data in a specific area, and over a sufficient amount of time, there exist multiple subgroups (or clusters) of temporally homogeneous time series in different spatial points. The results from that study show that the temporal data distribution within the same cluster can be approximated as a single signal, the centroid (or mean) of this cluster. Leveraging this technique, the problem of neighborhood-level vector population modeling has been split into two steps: (i) finding vector population time series clusters along the spatial axis obtaining their mean time series, and (ii) deriving a model of the obtained means using environmental information at the neighborhood-level from free EO products. Point (i) is achieved using the K-means clustering algorithm, and point (ii) using RNNs.

5.2 Background on RNNs

Unlike feed forward neural networks, RNNs are a kind of neural networks with loops, which allow them to learn sequential dependency in data. Given $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ with $\mathbf{x}_t \in \mathbb{R}^u$ as input independent covariate features, a simple RNN can be expressed as follows:

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t), \tag{5.1}$$

where $\mathbf{h}_t \in \mathbb{R}^v$ is the hidden state at time t, and v is the number of hidden units which is an hyperparameter to set.

Due to the problem of vanishing gradients RNNs, the function f is estimated using LSTMs [145]. An LSTM maintains a hidden state, $\mathbf{h}_t$, and a memory cell state, $\mathbf{s}_t$, that are updated at every

time step, and used to determine the output at that same time. At each time step, the access to s_t is controlled by three sigmoid gates: forget gate $\mathbf{f}_t$, input gate $\mathbf{i}_t$, and output gate $\mathbf{o}_t$. The mathematical formulations of these gates and the resulting $\mathbf{h}_t$ and $\mathbf{s}_t$ are summarized as follows:

$$\begin{aligned}
\mathbf{f}_t &= \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}; \mathbf{x}_t] + \mathbf{b}_f), \\
\mathbf{i}_t &= \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}; \mathbf{x}_t] + \mathbf{b}_i), \\
\mathbf{o}_t &= \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}; \mathbf{x}_t] + \mathbf{b}_o), \\
\mathbf{s}_t &= \mathbf{f}_t \odot \mathbf{s}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_s[\mathbf{h}_{t-1}; \mathbf{x}_t] + \mathbf{b}_s), \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{s}_t),
\end{aligned} \tag{5.2}$$

where $[\mathbf{h}_{t-1}; \mathbf{x}_t] \in \mathbb{R}^{v+u}$ is a concatenation of the previous hidden state $\mathbf{h}_{t-1}$ and the current input $\mathbf{x}_t$; $\mathbf{W}_f$, $\mathbf{W}_i$, $\mathbf{W}_o$, $\mathbf{W}_s \in \mathbb{R}^{v \times (v+u)}$, and $\mathbf{b}_f$, $\mathbf{b}_i$, $\mathbf{b}_o$, $\mathbf{b}_s \in \mathbb{R}^m$ are learnable weight and bias parameters, respectively; σ, $\tanh$ and $\odot$ are the logistic sigmoid activation function, the hyperbolic tangent function, and the Hadamard product, respectively.

5.3 Methodology

5.3.1 Notation and Problem statement

Let's consider a database of concurrent time series of *Ae. aegypti* mosquito population collected using M mosquito traps: $Y = \{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(M)}\}$, with $\mathbf{y}^{(m)} \in \mathbb{R}^P$, where P is the observation period (e.g., the number of weeks in case of weekly monitoring), is the vector of data collected at the m-th mosquito trap, i.e. $\mathbf{y}^{(m)} = (y_1^{(m)}, y_2^{(m)}, \ldots, y_P^{(m)})$. Instead, let's denote $\mathbf{y}_t = (y_t^{(1)}, y_t^{(2)}, \ldots, y_t^{(M)})$ the vector of mosquito numbers collected at all M traps in a particular t-th week.

Now, let's partition Y into K clusters of time series $C = \{C^{(1)}, C^{(2)}, \ldots, C^{(K)}\}$ with means (i.e., cluster centroids) $\{\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \ldots, \mathbf{c}^{(K)}\}, \mathbf{c}^{(k)} \in \mathbb{R}^P$. The clusters form a partition of Y, i.e., $C^{(k)} \subseteq Y, (k = 1, 2, 3, \ldots, K), \cap_{k=1}^{K} C^{(k)} = \emptyset$ and $\cup_{k=1}^{K} C^{(k)} = Y$. We present details of the clustering algorithm in Section 5.3.3.

Finally, let's consider N environmental variables (or proxies to them extracted from EO data), whose measure are available for the same P time instants in the M locations of the mosquito traps. Let's denote the whole set of values of these variables as $\mathbf{V} \in \mathbb{R}^{N \times M \times P}$. By clustering $\mathbf{V}$ according to the partition of Y, $\mathbf{V}$ is reduced to $\mathbf{X} \in \mathbb{R}^{N \cdot K \times P}$, where $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N \cdot K)}\}$, and $\mathbf{x}^{(i)} \in \mathbb{R}^P$. Eventually, let $\mathbf{x}_t \in \mathbb{R}^{N \cdot K}$ be the set of mean values of the environmental variables for each cluster at the time instant t, i.e. $\mathbf{x}_t = \{x^{(1)}(t), x^{(2)}(t), \ldots, x^{(N \cdot K)}(t)\}$.

Using a temporal window of size $T \ll P$, we formulate our forecast model as a nonlinear autoregressive exogenous model (NARX):

$$\widehat{\mathbf{c}}_t = F([\mathbf{c}_{t-T}, \ldots, \mathbf{c}_{t-1}]; [\mathbf{x}_{t-T}, \ldots, \mathbf{x}_{t-1}]), \tag{5.3}$$

where ";" denotes the time point concatenation, and, as before, $k = 1, 2, \ldots, K$. $F(.)$ is selected to be an LSTM model (see Section 5.3.2). The model output $\widehat{\mathbf{c}}_t \in \mathbb{R}^k$ is a vector of the forecast values of mean mosquito population for the k clusters in the t-th week based on T trailing autoregressive (vector population) and exogenous (environmental conditions) components.

5.3.2 Adaptation of RNNs for this work

For this study, an encoder-decoder LSTM [150] architecture is applied due to its recorded success in many applications including time series forecasting. The encoder is an LSTM which encodes the input sequence, $\mathbf{c}_{t-T}, \ldots, \mathbf{c}_{t-1}$ and $\mathbf{x}_{t-T}, \ldots, \mathbf{x}_{t-1}$ within a time window of length T, into a learned representation $\mathbf{h}_t \in \mathbb{R}^v$ and memory cell state $\mathbf{s}_t \in \mathbb{R}^v$ is the encoder output size. For time series prediction tasks, the decoder is usually a stack of LSTM and a fully connected (dense) neural network with a non-linear activation. The decoder LSTM takes $\mathbf{h}_t$ as input, copies it over the length of T, and generates the decoder hidden state $\mathbf{d}_t$. The fully connected layer takes $\mathbf{d}_t$ as input and produces $\widehat{\mathbf{c}}_t$. For this study, a fully connected layer (dense) with a rectified linear activation function (ReLU) [151] is added on top of the decoder LSTM to map the output of the LSTM to a vector of forecast mosquito populations. Figure 5.1 illustrates this adapted encoder-decoder LSTM.

Considering a window of size T and subwindows of size P such that $T \ll P$ and $T \mod P = 0$, the model is:

$$\mathbf{h}_t = f_1([\mathbf{c}_{t-T}, \ldots, \mathbf{c}_{t-1}]; [\mathbf{x}_{t-T}, \ldots, \mathbf{x}_{t-1}]), \tag{5.4}$$

$$\mathbf{d}_t = f_2(\mathbf{h}_t), \tag{5.5}$$

$$\widehat{\mathbf{c}}_t = \vartheta(\mathbf{W}_d \mathbf{d}_t + \mathbf{b}_d), \tag{5.6}$$

where $\mathbf{W}_d$ and $\mathbf{b}_d$ are learnable parameters of the decoder fully connected layer, and $\mathbf{d}_t$ and $\mathbf{y}_t$ are the decoder hidden state and model output for time t prediction, respectively; $f_1(.)$ and $f_2(.)$ are the encoder and decoder LSTMs, respectively; ϑ is the ReLU activation function, which is defined for an arbitrary input $x \in \mathbb{R}$ as:

$$\vartheta(x) = \max\{0, x\}. \tag{5.7}$$

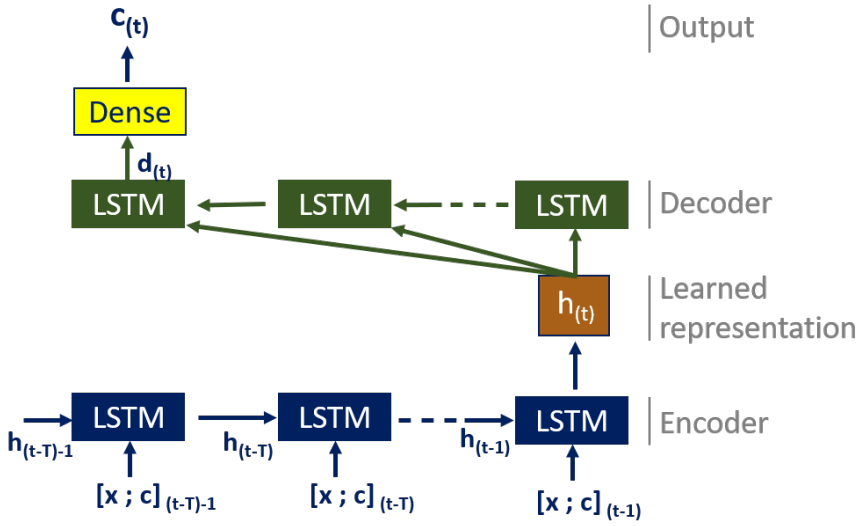


FIGURE 5.1: Architecture of our adapted encoder-decoder LSTM. The encoder output $\mathbf{h}_t$ is replicated into T copies to feed each time point of the decoder. The dense layer maps the decoder output to the desired prediction output. ";" signifies concatenation; T is the size of a temporal window; $\mathbf{c}_t$ is the predicted output vector at time t; $\mathbf{x}_{t-1}$ is a vector of the EO covariate features at time $t-1$.

The choice of the ReLU activation for the model output layer is justified by the need to produce positive real number output predictions i.e $\widehat{c}_t^{(i)} \in \mathbb{R}^+ \; \forall \; \widehat{c}_t^{(i)} \subseteq \widehat{\mathbf{c}}_t$, since it is impossible to have negative mosquito vector population values.

5.3.3 Time series clustering

The clustering applied to the set of mosquito trap records Y is implemented by means of the standard K-means algorithm with Euclidean distance. This algorithm is simple to implement and converges fast [152, 153].

Still, due to the unsupervised nature of clustering, there is the need to determine the optimal number of clusters K. The goal is selecting K to minimize the total intra-cluster variation, also known as total within-cluster sum of square variation or distortion. To this aim, the elbow method [154] is applied. The resulting set of distortions is then plotted, the "optimal" K is selected as the *"sweet spot"* where there is a bend ("elbow") in the curve indicating a significant reduction in the gradient of the distortion with respect to K. The distortion is computed as as:

$$J = \sum_{k=1}^{K} \sum_{m=1}^{M} \left\| \mathbf{y}^{(m)} - \mathbf{c}^{(k)} \right\|^2, \tag{5.8}$$

5.3.4 Random forest model for benchmarking

To prove their significance, the results of this work will be compared to those by using multi-output Random Forests (RF) is fitted as a baseline NARX (following eq. (5.3)). However, due to the non-recurrent nature of RF, it is not possible to consider the sequential ordering of lagged environmental effects within the considered time window T. Hence, they are concatenated them into a single vector, ignoring their temporal ordering. The number of trees is set to the commonly selected value of 500 (e.g., as in Sections 3.4.1 and 4.3.4).

5.4 Materials

5.4.1 Study area and field data

This research is based on MI-Aedes® collected adult *Ae. ae gypti* mosquito counts collected in the towns of Vila Velha and Serra in Espírito Santo State (region), Brazil. Vila Velha is between latitudes $20°19'$ and $20°32'$ South, and longitudes $40°16'$ and $40°28'$ West. It covers a total area of $209.965\,\mathrm{km}^2$, and has an estimated population of $486,208$ people. Serra is between latitudes $20°7'$ and $20°12'$ South, and longitudes $40°18'$ and $40°30'$ West. It covers a total area of $553\,\mathrm{km}^2$, and has an estimated population of $507,598$ people. Both towns are about $40\,\mathrm{km}$ apart. Figure 5.2 presents their geographical locations.

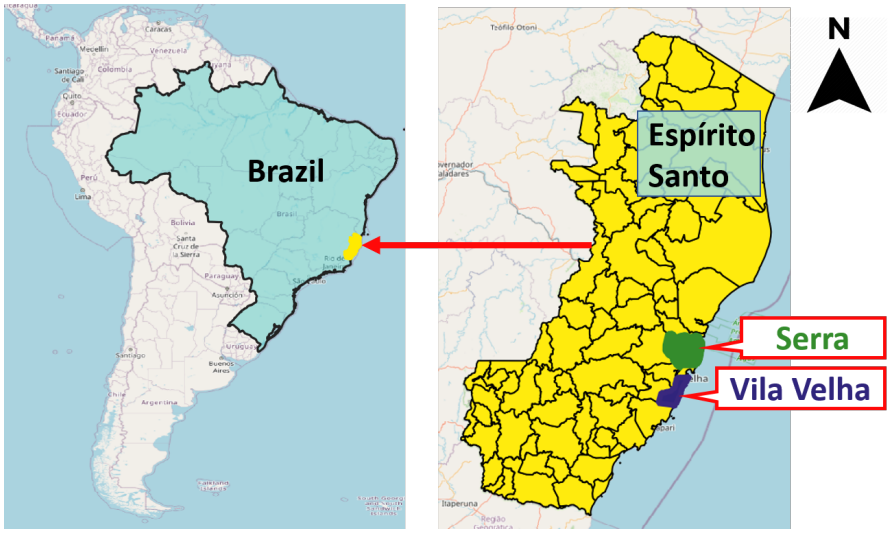


FIGURE 5.2: Geographical location of the considered study areas: Vila Velha and Serra

These locations are very relevant to dengue risk mapping. Indeed, in 2019 the Espírito Santo state verified $63,847$ dengue cases, with an incidence of 1588.8 per $100,000$ inhabitants. Vila Velha had $6,557$ of those cases ($1,348.6$ per $100,000$ inhabitants), which is far greater than the epidemic threshold.

For this study, in Vila Velha, data for weeks corresponding to years 2017 and 2018 (Batches 1 and 2 in Table 3.1) are applied. As already mentioned in Section 3.2, these data contain mosquito counts collected weekly with 791 MosquiTRAP® devices.

In Serra, corresponding data also for years 2017 and 2018, spanning 27/04/2017 to 06/12/2018, and which are collected with 1127 devices. All the mosquito traps in both locations are placed at least at $250\,\mathrm{m}$ apart. Data were acquired on site weekly by a team of trained field workers by inspecting the sticky cards set inside each trap. *Ae. aegypti* mosquitos were identified, counted, and their presence and number registered using the MI-Aedes® framework which has been detailed in Section 3.2.

Also, as a result of the control activities optimization performed at the start of year 2018 in both locations, the data collected are divided into two temporal regimes: 2017 and 2018. In Vila Velha, the data for 2017 spans from 10/04/2017 to 31/12/2017 (epidemiological weeks 15 to 52, 36 weeks), while the data for 2018 spans from 02/01/2018 to 05/10/2018 (epidemiological weeks 1 to 40, 40 weeks). In Serra, the data for 2017 spans from 27/04/2017 — 30/12/2017 (epidemiological weeks 17 to 52, 36 weeks), while the data for 2018 spans 05/01/2018 to 05/10/2018 (epidemiological weeks 1 to 49, 40 weeks). Note that weeks 7 and 8 for year 2018 were not provided.

To pre-process the data, traps missing data even for just one of the weeks, or with zero mosquito reported in all weeks, were filtered out. This resulted in a final set of 193 and 325 trap records in 2017 and 2018, respectively, out of the initial 791 points in Vila Velha. Similarly, in Serra, the final set includes 567 trap records in 2017 and 95 in 2018. To reduce"data noise" in the obtained series from each retained trap, an exponential moving average filter with a span of five weeks was applied to the records.

5.4.2 Environmental variables from EO data

This study leverages the same freely available EO products as with other studies focusing on the same theme which have been presented in Chapters 3 and 4. Specifically, for the purpose of recall, MODIS dataset are used to obtain humidity, vegetation, and temperature while the Global Precipitation Mission data (GPM) is used to obtain precipitation information. MODIS MOD13Q1 data product is used for vegetation and humidity information while MODIIA2 is used for temperature information (day and night-time temperatures). The details of these MODIS products are already presented in Table 1.2, while the specifications of GPM data product has been introduced in Section 3.2. In line with the requirement for a spatio-temporal study, all data are resampled (by nearest neighbor) to a common resolution of $250\,\mathrm{m}$ which is the minimum distance between neighbouring mosquito trap locations and also the native resolution of the highest spatial resolution data product used (MOD11A2).

5.4.3 Data extraction and transformation

Since the approach is supervised, there is the need to select training and validation samples. In this research, this step was performed after the clustering, because the forecast model is applied (see eq. (5.3)) to the cluster representative values.

Therefore, for each cluster the average values of the environmental covariate features at each point in time ($\mathbf{x}_t$) were computed by averaging the EO proxy values in the locations of the traps assigned to that cluster. Moreover, since the EO data have a different temporal sampling than the mosquito counts, their temporal records were interpolated using a cubic spline interpolation [155]. Specifically, both NDVI and NDWI data which are available every two weeks were interpolated to obtain weekly values.

The mosquito count records for each cluster, per time point, were randomly spatially subdivided into two sets: one for training, and the other one for model testing. Accordingly, the $\mathbf{c}_{t-T}, \ldots, \mathbf{c}_{t-1}$ vectors in eq. (5.3), were estimated using only the training set in the training phase, and the test set in the model testing phase.

Finally, the resulting training data (target and predictor variables combined) is then randomly subdivided along time to extract $20\,\%$ of the time-points to be used for validation of the model during training.

5.5 Experimental Results

5.5.1 Training procedure

The model was trained for one-week-ahead *Ae. aegypti* population prediction starting from T training populations (autoregressive component) and environmental condition features (exogenous component). As a result, we obtained predictions starting at $t = T + 1$.

The adaptive learning rate optimization algorithm (Adam) [156] was selected to train the neural network with a learning rate of 0.001. The objective function for parameters learning through backpropagation was set to the mean absolute error (MAE) loss [141]. A dropout rate of 0.2 was used in the decoder to avoid overfitting, and a batch size of 1 because the data set is not too large. All the models were trained in 100 epochs, and the model with the best validation accuracy was selected and saved.

5.5.2 Parameter Settings

The three key parameters required in an encoder-decoder LSTM are: (i) the temporal window size T, (ii) the encoder **h** output vector size, and (iii) the decoder **d** output vector size. For simplicity, as in Ref. [141], both the encoder and the decoder ($\mathbf{h} \in \mathbb{R}^v \ni \mathbf{d}$) use a single layer each, i.e., their output size is v. We considered the window size T as one of the three possible values $\{3, 6, 9\}$, while the value of v is selected from the set $\{16, 32, 64, 128\}$.

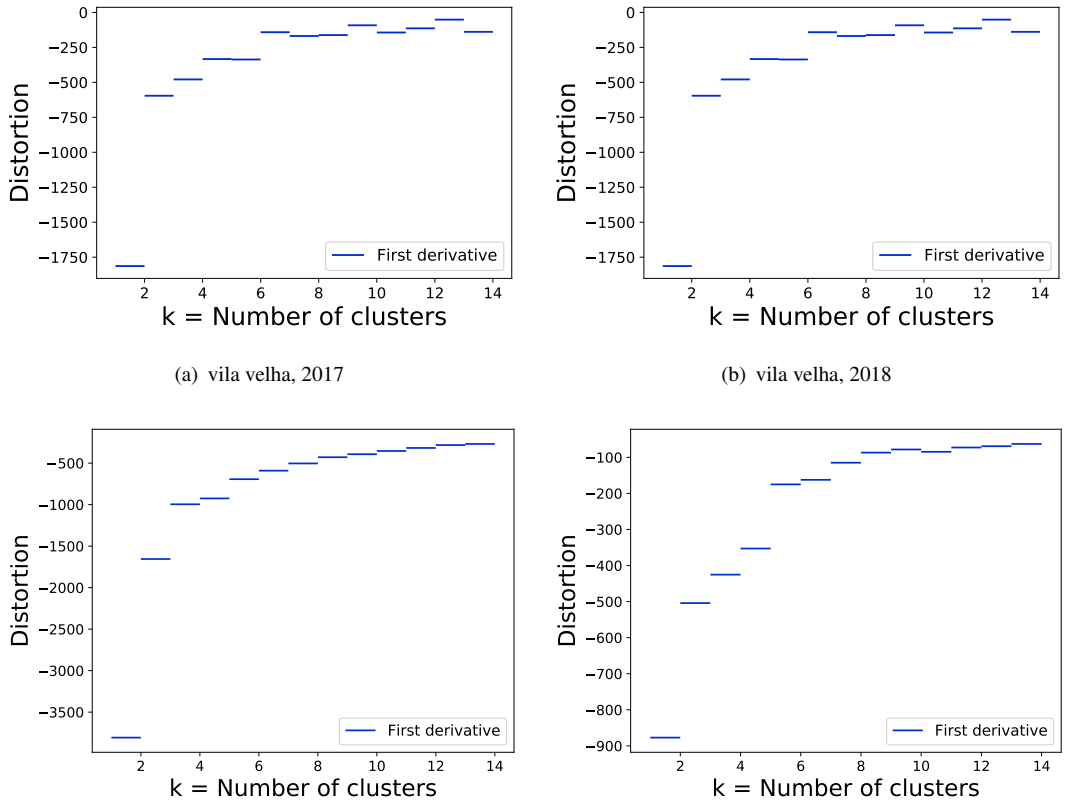
5.5.3 Clustering results

To address the stability problem of the K-means clustering method, the elbow plot is obtained as the average of 20 repetitions over the sequence $1 \leq K \leq 14$. Figure 5.3 shows the piecewise approximated first derivatives of the resulting K-means elbow plots in Vila Velha (Figures 5.3(a) and 5.3(b)) and Serra (Figures 5.3(c) and 5.3(d)) in both years. In Vila Velha, beyond $K = 6$, there is no significant jump in the derivative, hence we chose $K = 6$ as optimal elbow point in that location in both years. For Serra, on the other hand, $K = 5$ is the optimal elbow point which is chosen for both years. The clusters are derived and labeled such that the 6 clusters in Vila Velha are labelled Cluster 1A–6A in 2017 and Cluster 1B–6B in 2018. The A and B suffixes of the cluster labels are codes for years 2017 and 2018, respectively. Following the same convention, Cluster 1A–5A and 1B–5B representing the 5 clusters for both years in Serra are also derived.

The line plots of Figure 5.4 show the mosquito population temporal patterns for each resulting cluster for each year and location. Different clusters are differentiated by their temporal distribution and range (see the y-axes of plots). In an epidemiological sense, periods of maximum spikes are indicative of highest possible disease outbreak risks. In each case presented in Figure 5.4, the K-means clustering has helped to identify underlying common patterns that describe the vector development activity at the different neighborhoods where trap observations have been carried out.

Furthermore, it is seen from Figure 5.4 that in spite of the inter-cluster temporal pattern differences, there are similar patterns in sub-sequences across multiple clusters and locations. Such similarities correspond to weeks of typical macro-climatic effects at municipal and regional levels. By the hypothesis of vector population dependency on abiotic and biotic environmental effects, differences in temporal patterns of the cluster centers correspond to differences in micro-climatic effects which differ across clusters, and are shared within the same cluster.

A micro-climate is a local set of atmospheric conditions that differ from those in the surrounding areas. In Vila Velha, in 2017, local peaks can be observed in the neighborhood of observation



(a) vila velha, 2017

(b) vila velha, 2018

(c) Serra, 2017

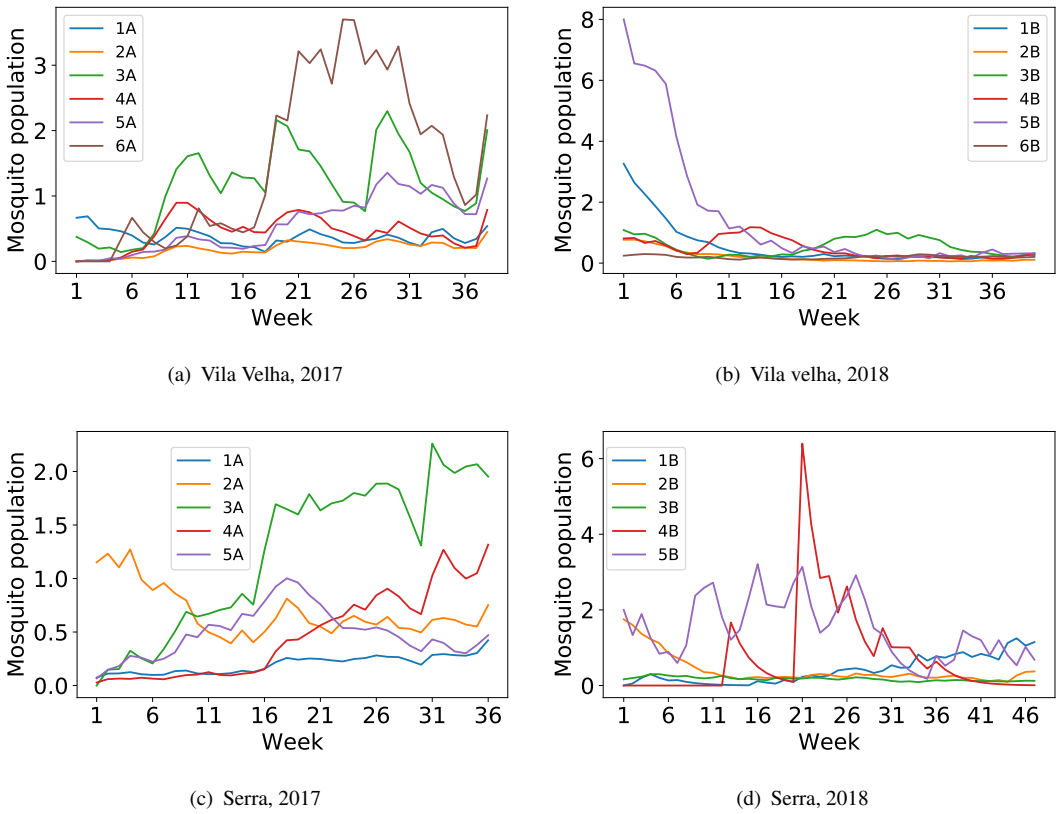
(d) Serra, 2018

FIGURE 5.3: Numerical first derivatives of the elbow plots for selecting the optimal number of k-means clusters in 2017 and 2018. The plots show that $k = 6$ is the elbow point in both years in Vila Velha, while $k = 5$ is the elbow point in both years in Serra.

weeks 13, 21, 29, and 38 (corresponding to epidemiological weeks 27, 35, 43 and 52, respectively). For each cluster, however, the range and duration of the peaks differ. In 2018, the population of the mosquitoes is always decreasing between observation (and epidemiological weeks) 1–9 for all clusters. There are spikes with local peaks in the neighborhood of week 25 in clusters 3 and 6. For Serra, in 2017, we have local peaks in the neighborhood of observation weeks 18, 28 and 32 (epidemiological weeks 34, 44 and 48, respectively) for all clusters.

Common patterns are also observed among some subsets of clusters. For example, all clusters except 2A exhibit increasing vector population between observation weeks 6–16. Still in Serra, in 2018, Clusters 4B and 5B exhibit different patterns all through the year with respect to the other clusters that are always close to zero. Inter-cluster similarities per location can be attributed to municipality-level macro-climatic effects.

We also see patterns that are common to both Vila Velha and Serra. For example, in 2017 there are local peaks in the neighborhood of epidemiological weeks 34–36 and 43–44 in both test locations. These similarities can be attributed to regional macro-climatic effects.



(a) Vila Velha, 2017

(b) Vila velha, 2018

(c) Serra, 2017

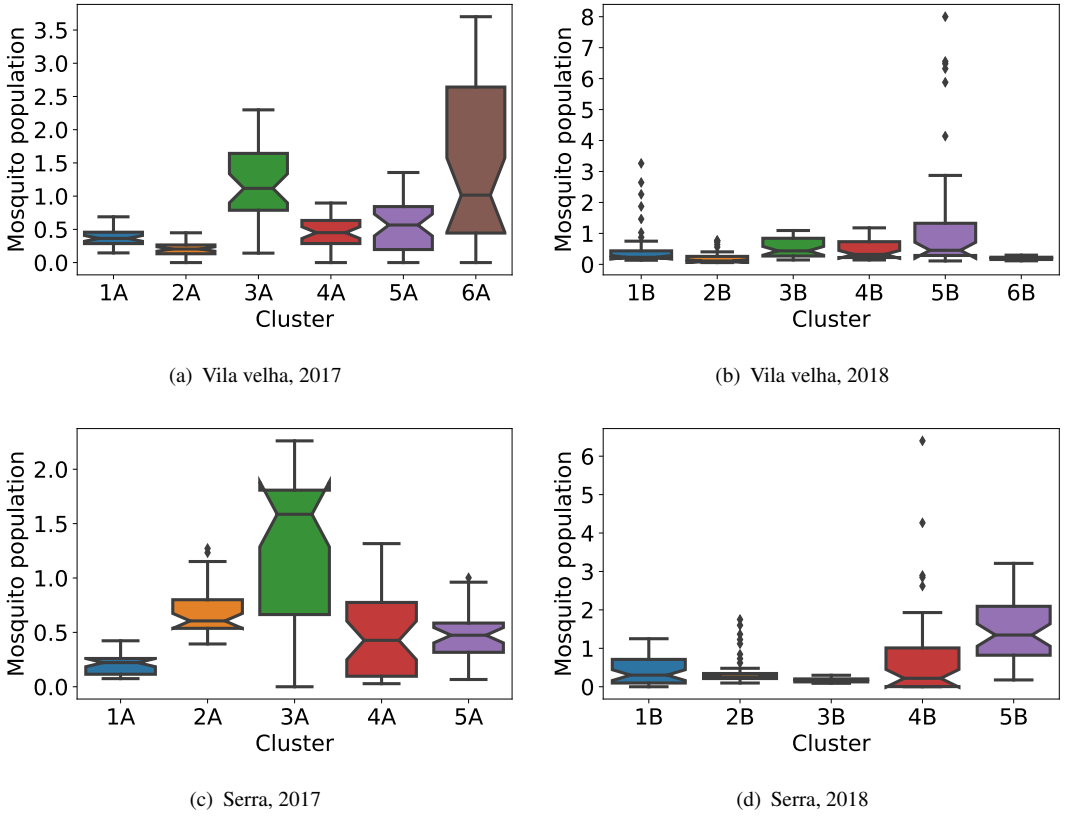
(d) Serra, 2018

FIGURE 5.4: Mean temporal distribution for clusters obtained in both years.

The similarities– at municipality and regional levels– only exist in pockets of time duration, as shown in Figure 5.4. This result reveals the strengths and weaknesses of municipality-level modeling like the one obtained in [54] and [53], and especially for the studies presented in Chapters 5 and 4 which both use the same data in Vila Velha as in this study. While such municipality-level models can capture general trends that are common to most clusters, they do not provide detailed information across different clusters. In areas where the trends in all clusters are similar, then, perhaps, a municipality-level model can be sufficient. Otherwise, there is the need for a disaggregated approach like the one presented in this study for better inference at neighborhood-level.

Figure 5.5 presents boxplots of the vector population series for the derived clusters in both test locations. In the epidemiological sense, minima, maxima, and inter-quartile ranges (IQR) provide a risk profile summary of the component points of each the cluster. Higher maxima mean higher risk exposure at peak periods, while the minima are the lower bounds of the risk exposure in the locations considered. The IQRs describe the pattern variability of the risk exposure in the considered time. In Vila Velha, Cluster 6A has the highest maximum and variability in 2017, which Clusters 1A and 2A come from low risk locations. In 2018, cluster 5B has the highest maximum and variability. In Serra, Cluster 3A has the highest maximum in 2017, while cluster

5B has the largest IQR maximum in 2018. Also, Cluster 2A has the highest minimum in this same location in 2017.



(a) Vila velha, 2017

(b) Vila velha, 2018

(c) Serra, 2017

(d) Serra, 2018

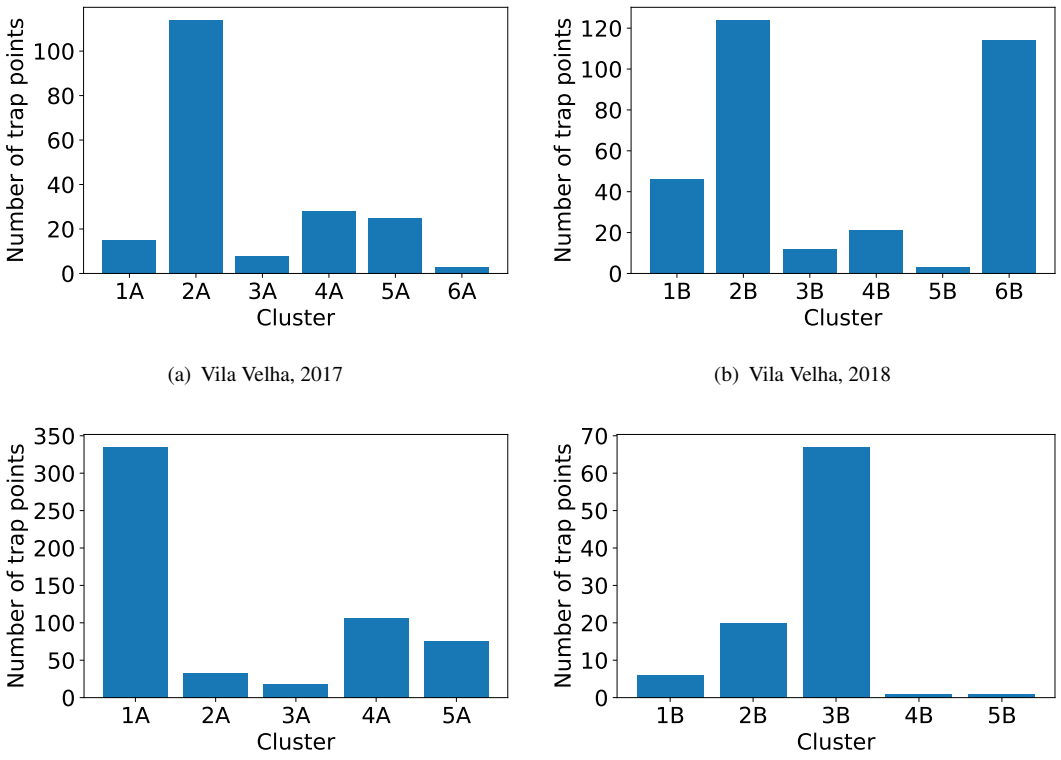
FIGURE 5.5: Descriptive statistics of the resulting female *Ae. aegypti* traps data cluster means.

These inter-cluster differences in environmental variables and vector population suggest that the clustering process has achieved the useful aim of finding homogeneous trap locations: separating the trap points into clusters of different temporal patterns and disease risk profiles.

Figure 5.6 presents bar plots showing the number of traps in all the derived clusters in both study location. Looking at this figure alongside Figure 5.5, it can be seen that most of the trap points are in low risk locations, i.e places where the variability and maximum value reached by the mosquito population are not high. For instance, in Vila Velha, Cluster 6A which has the highest maximum and variability in 2017 has the lowest number of points while Clusters 2A which is the lowerst risk cluster has the highest number of points. This same trend can be seen in the other clusters obtained for both locations.

Figure 5.7 presents the location of the points along with color indicators showing the clusters they belong to. It can be seen that trap points in the same cluster are not necessarily geographically collocated, as is also the case in the results presented in [149].

We examined cluster membership of points common to both years to understand the spatial relationship between clusters obtained in different years in the same location. We used the overlap



(a) Vila Velha, 2017

(b) Vila Velha, 2018

(c) Serra, 2017

(d) Serra, 2018

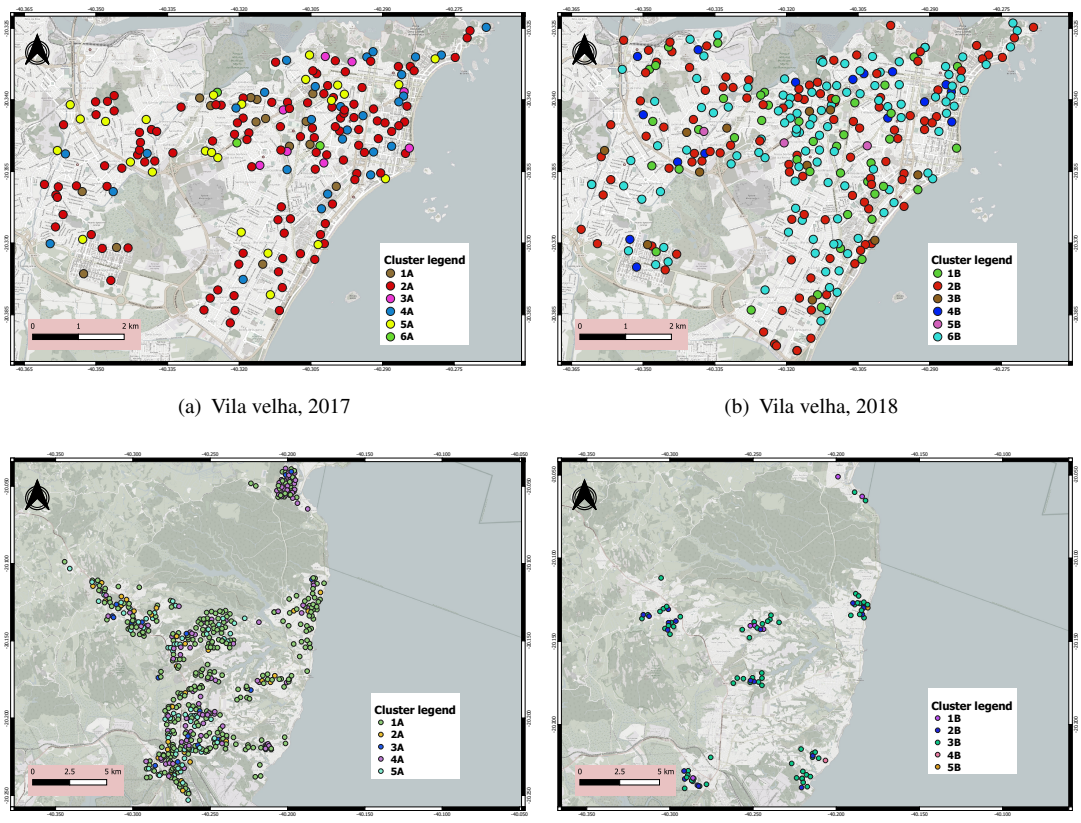
FIGURE 5.6: Bar plots of number of traps in each cluster for each year.

coefficient (OC) [157] to measure spatio-temporal similarities, i.e, the number of common traps contained between every possible cluster pair across both years in the same location:

$$\mathrm{OC}(C_A^i, C_B^j) = \frac{|C_A^i \cap C_B^j|}{\min(|C_A^i|, |C_B^j|)}, \tag{5.9}$$

where C_A^i and C_B^j are the sets of mosquito trap points in the i-th and j-th clusters in 2017 and 2018, respectively, in the same location, after filtering all clusters to retain only traps that exist in both years. In Vila Velha, among the 193 and 325 trap points analyzed in 2017 and 2018, respectively, there are 128 traps which are common to both years. In Serra, among the 567 and 95 trap points analysed in 2017 and 2018, respectively, there are 59 traps which are common to both years.

The spatio-temporal similarities are presented in form of similarity matrices in Figure 5.8. The results for Vila Velha (c.f. Figure 5.8(a)) are discussed in the rest of this paragraph. The highest risk clusters in both years in this location – Clusters 6A and 5B – have an OC of 0.64, which is the highest similarity value obtained in the matrix. This is evidence that there is high correlation between the set of traps with high risk in both years. Also, as shown by their zero OC values with four out of the remaining five clusters, the annual highest risk clusters are decoupled from the lower risk clusters. From an epidemiological standpoint, this is evidence of continuity in risk



(a) Vila velha, 2017

(b) Vila velha, 2018

(c) Serra, 2017

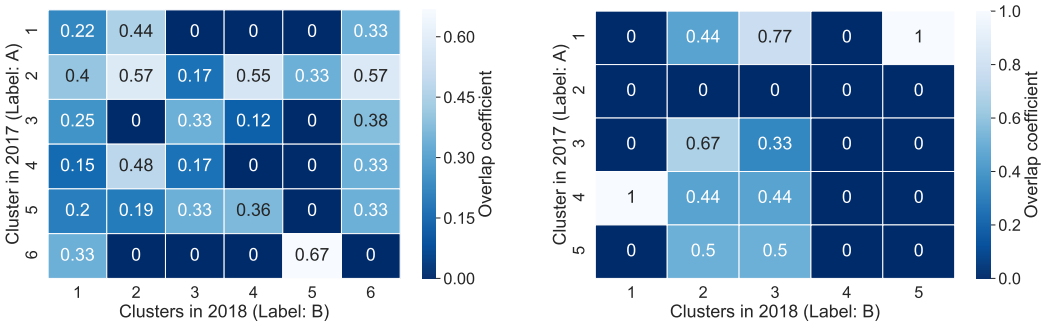
(d) Serra, 2018

FIGURE 5.7: Mosquito trap points color-labelled according to the clusters they have been assigned into. In the background is OpenStreetMap™ view of the study areas: Vila velha and Serra.

level across different control regimes, also showing that the micro-climatic effects that drive the local vector population at these high risk points exhibit some robustness to the control measures that have been applied. In addition, Cluster 2A – the one with lowest risk – has its highest OC of 0.57 with both Clusters 2B and 6B, and its lowest OC of 0.17 with Cluster 3B. As presented in Figure 5.5(b), these clusters (Clusters 2B and 6B) are always low risk throughout the observation period, indicating that that they contain significant amount of the low risk points from Cluster 2A. Since Cluster 3B in Vila Velha is the second highest risk cluster considering the IQR, its low intersection with Cluster 2A (Vila Velha's lowest risk cluster in 2017) is in line. Clusters 1A and 2B, both of relatively low risk in both years, have an OC of 0.44.

In Serra, there are only 59 common traps points in both years. The similarity matrix is subsequently sparse (see Figure 5.8(b)). A significant amount of the sparse relationships in the matrix involve Clusters 5B and 6B, which contain only one trap each and are, thus, unreliable for the kind of analysis conducted here. In spite of the sparseness of the matrix, we still see that Clusters 1A and 3B – the lowest risk clusters in both years – have an OC of 0.77. Also, Clusters 3A and 2B have an overlap coefficient of 0.67. Since Cluster 2B is a high risk cluster in 2018, if we ignore 4B and 5B which contain single trap points each, athere is again a coupling between

high risk points across years and control regimes. These results further point to evidences of continuity in the risk level of the trap points even across different control regimes. Key actors in vector surveillance and control can use the information provided by this similarity matrix for neighborhood-level understanding of control activities effects.



(a) Vila velha

(b) Serra

FIGURE 5.8: Matrix of similarity in set of trap points contained in cluster pairs (each from different years). The similarity is measured by overlap coefficient (OC).

5.5.4 Model results

Table 5.1 presents the quality of the models resulting from the grid search for the optimal temporal window size T. Based on these data, we chose $T = 3$ as the optimal temporal window in both locations. This result is supported by Refs. [158, 159] which shows that the development cycle of *Ae. aegypti* from egg to adult ranges between one-and-a-half to three weeks. The environmental conditions during this development period determine the transition rate of the eggs to adult. The annual best models in both locations are used in all further experiments.

TABLE 5.1: Comparison between mean absolute error (MAE) loss for all models with respect to the temporal window size with a constant learned representation vector size; v = 16 is the learned representation size, while T is the temporal window size considered for each prediction.

		Vila Velha			**Serra**		
	$T \Rightarrow$	3	6	9	3	6	9
2017	Training	0.3117	0.3392	0.4926	0.2254	0.1509	0.2451
	Validation	0.4627	0.1810	0.3745	0.1985	0.3275	0.2729
	Test	**0.6120**	0.6450	0.7565	**0.4048**	0.4703	0.5889
2018	Training	0.1407	0.2067	0.2762	0.2738	0.7999	0.2642
	Validation	0.1998	0.4516	0.3395	0.2407	0.8574	0.3151
	Test	**0.3600**	0.4624	0.4602	**0.4418**	0.9028	0.5372

Table 5.2 presents the quality of models resulting from the grid search for optimal learned representation size with T set to 3. This table shows that in Vila Velha, the learned representation size $v = 128$ produces the best quality on the test data in 2017, while $v = 16$ produces the best quality in 2018. In Serra, $v = 64$ produces the best quality on the test data in 2017, while $v = 32$ produces the best quality in 2018.

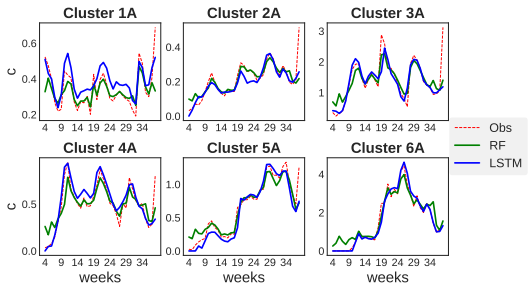
The search for learned representation size is a standard practice with fitting encoder-decoder neural network, and its result does not have a direct epidemiological bearing. However, it can be seen that the best models obtained in 2017 in both locations require higher values of v compared to their 2018 counterparts. This is because the 2017 field mosquito data contain patterns with more variability than 2018 due to an improvement of control activities (see Figures 5.4 and 5.5).

TABLE 5.2: Comparison of MAE for models with respect to varying learned representation vector size for $T = 3$. The learned representation is the encoder output; v: learned representation size

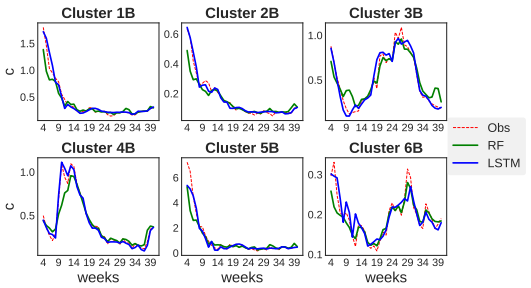
		Vila Velha		**Serra**	
v	Year ⇒	2017	2018	2017	2018
16	Training	0.3117	0.1407	0.2254	0.2738
	Validation	0.4627	0.1998	0.1985	0.2407
	Test	0.6120	**0.3600**	0.4048	0.4418
32	Training	0.2637	0.1501	0.1880	0.2652
	Validation	0.4774	0.1975	0.2456	0.1817
	Test	0.6274	0.4126	0.4231	**0.3986**
64	Training	0.2841	0.1781	0.1630	0.2459
	Validation	0.4765	0.2231	0.1472	0.3632
	Test	0.6203	0.3816	**0.3984**	0.4329
128	Training	0.2802	0.1808	0.2266	0.6732
	Validation	0.3880	0.3189	0.2244	0.8454
	Test	**0.5767**	0.4038	0.4440	0.5794

Figure 5.9 presents the line plots comparing the best LSTM models with the benchmark RF model on training (validation inclusive) and test data in both years. Figure 5.10 present scatterplots comparing both models with respect to their fitness to test data. Spikes in the line plots are indicative of increasing rate of vector population. From an epidemiological standpoint, these spikes are proxies to increasing risk of diseases occurrence in neighborhoods around the cluster component trap points. Hence, forecasting such spikes will serve well as disease outbreak early warning signals. Dips in the line plots, contrarily, are indicative of low rates of vector population. The ability to forecast dips correctly in all clusters is also important since it may lead to better resource allocation through the redeployment of control resources from areas with predicted dips to areas with predicted spikes.

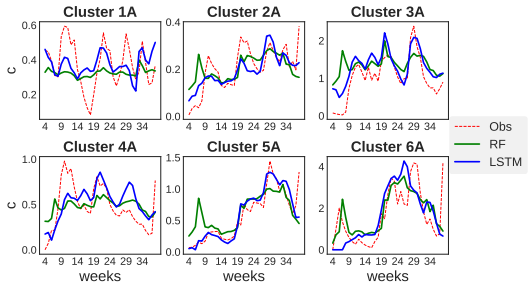
In further experiments, the resulting best LSTM models from Table 5.2 were compared with their corresponding baseline RF models; Table 5.3 shows the results. In Vila Velha, LSTM performs approximately 13 % better than RF on test data in both years. In Serra, LSTM produces an improvements of approximately 17 % and 15 % in 2017 and 2018, respectively. It is worth recalling at this point that LSTMs leverage the sequential ordering input data in the learning process. This is especially useful for learning lagged contributions of predictor features along time. Our results here show significant quantitative evidence of the need for this property of LSTMs for the specific use case addressed in this study.



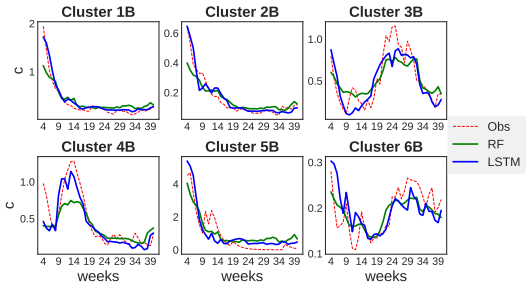
(a) Vila velha, training, 2017



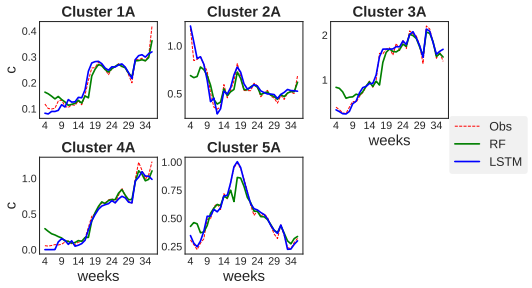
(b) Vila velha, training, 2018



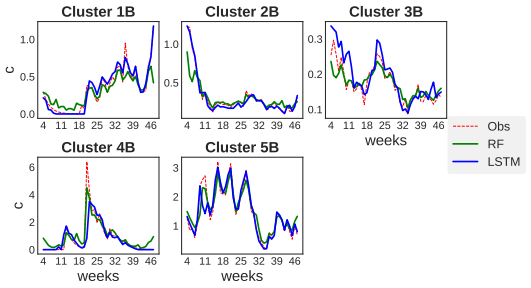
(c) Vila velha, testing, 2017



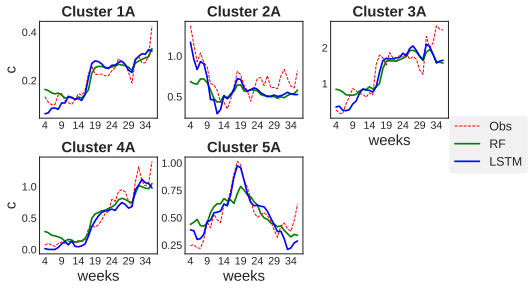
(d) Vila velha, testing, 2017



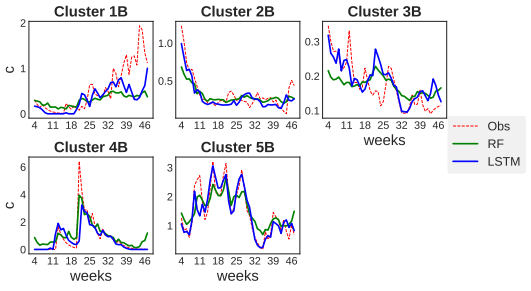
(e) Serra, training, 2017



(f) Serra, training, 2018



(g) Serra, testing, 2017



(h) Serra, testing, 2018

FIGURE 5.9: Line plots comparing observed and predicted values for LSTM and RF models in 2017 and 2018. Validation data points are inserted into their time positions among the training data. Obs: Observed.

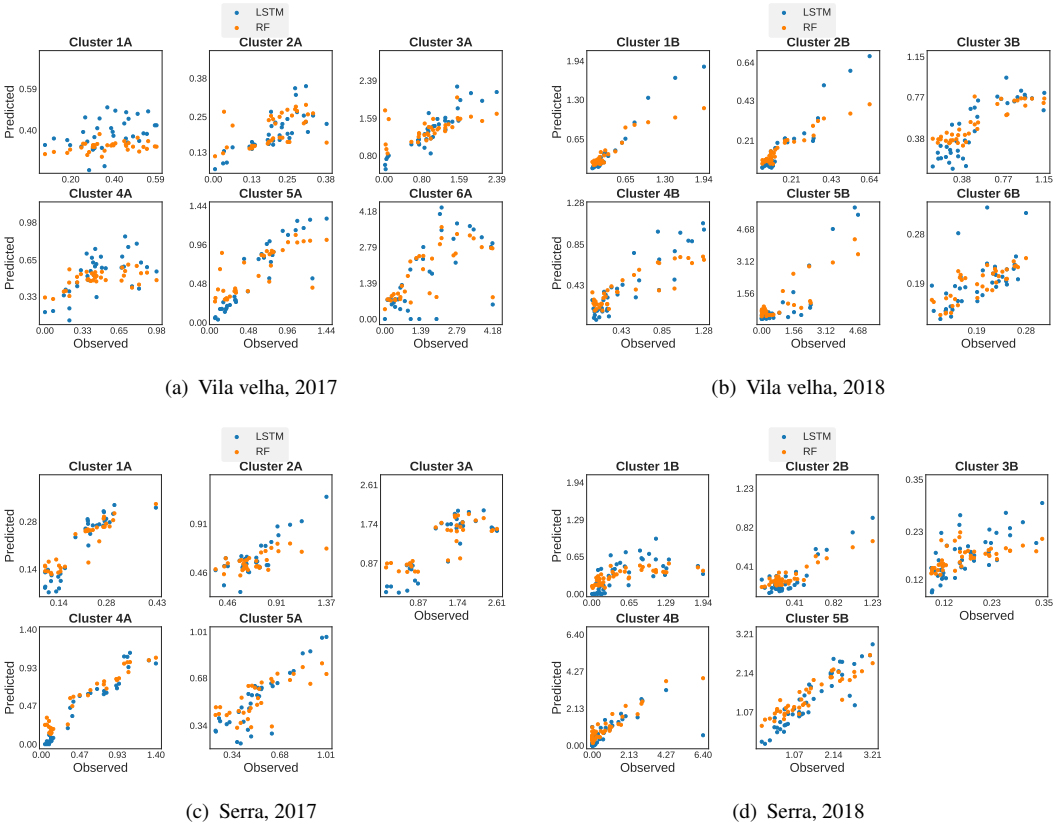
TABLE 5.3: Comparison of MAE for best LSTM and RF in both considered years.

		Vila Velha		**Serra**	
Model	Year ⇒	2017	2018	2017	2018
LSTM	Training	0.2802	0.1407	0.1630	0.2652
	Validation	0.3880	0.1998	0.1472	0.1817
	Test	**0.5767**	**0.3600**	**0.3984**	**0.3986**
RF	Training	0.3203	0.1978	0.2236	0.2168
	Validation	0.1800	0.2511	0.1644	0.3644
	Test	0.6599	0.4128	0.4636	0.4808

Figures 5.9(a) and 5.9(b) present the line plots of both LSTM and RF models on training data in Vila Velha across the two observed years. LSTM overestimates the observed training data in observation weeks 9–34 and in Cluster 1A (Figure 5.9(a)), but still follows the observed trend. RF does not show such overestimation. RF underestimates the observed values in the neighborhood of observation weeks 4–9 in that same cluster, and overestimates around these same weeks in the remaining clusters. LSTM, however, fits the data well in that period. Also, both LSTM and RF fail to reach the observed data value in week 36 in Clusters 1A–5A, but LSTM significantly performs better in that week in Cluster 1. In 2018 (Figure 5.4(b)), RF underestimates the observed data around observation weeks 4–6 in all clusters, except in Cluster 4B. LSTM, on the other hand, fits well the data in these weeks in all the clusters, except in Cluster 5. Also, RF underestimates the observed training data around weeks 9–14 in Cluster 4B, and overestimates its prediction in this same period in Cluster 3B.

Still on Vila Velha, with regards to test data performance, in 2017, as presented in Figure 5.9(c), both LSTM and RF do fit the observed test data well in Cluster 1 compared to the other clusters. This can be attributed to the lower purity of this cluster, as can be inferred from the differences between the training and test data temporal distribution (Compare Cluster 1A training and test patterns in Figures 5.9(a) and 5.9(c), respectively). Nevertheless, for this same cluster, LSTM still follows the trend (spikes and dips) of the vector population in weeks 9–14, 19–24, 31–36, which is a total of 18 out of 36 weeks. RF, on the other hand, remains quasi-invariant in temporal pattern all through the observation weeks.

These results show the robustness of the LSTM to clustering quality variations, which is a major component of the framework proposed in this study. It is worth mentioning, however, that the test results can be improved by improving the clustering procedure. In the other clusters, around weeks 4–9, RF wrongly predicts a spike in Clusters 2A–5A, while LSTM performs better in that period in the mentioned clusters. In epidemiological terms, overestimation (e.g wrongly forecasting a spike) of vector population, as exhibited here by RF, can result in false outbreak alarms.



(a) Vila velha, 2017

(b) Vila velha, 2018

(c) Serra, 2017

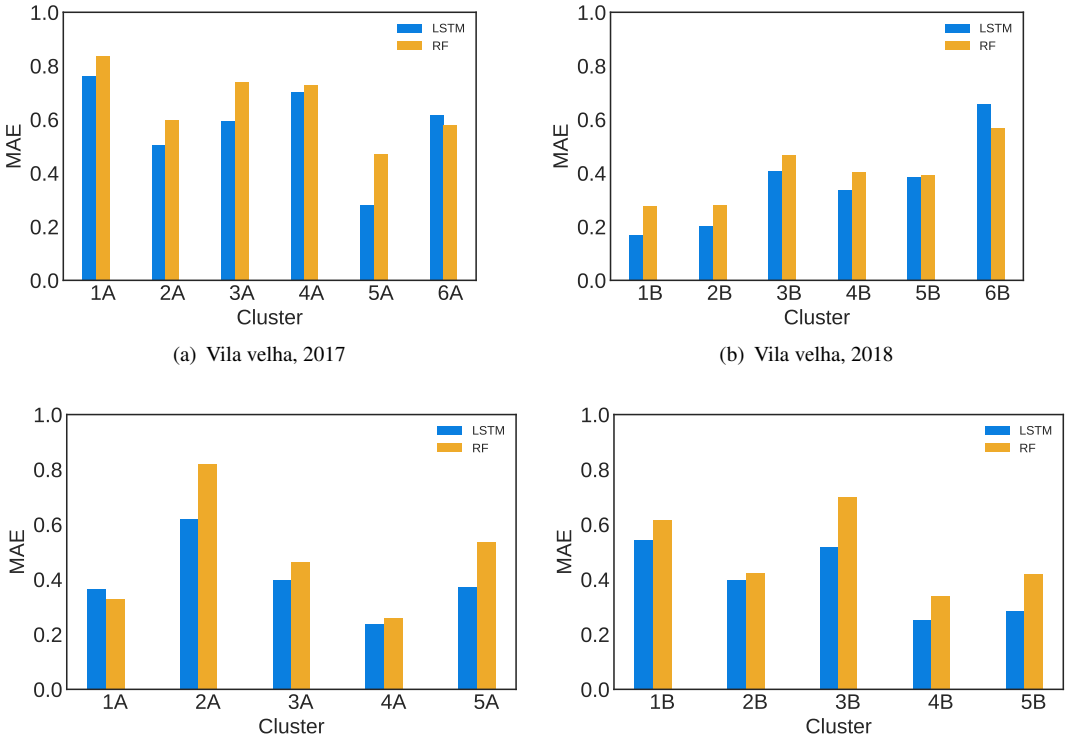
(d) Serra, 2018

FIGURE 5.10: Scatterplots comparing observed and predicted values for LSTM and RF models on test data.

Considering the test data performance in Vila Velha for year 2018 as presented in Figure 5.9(d), RF underestimates the observed data around weeks 4–9 in all clusters. LSTM, however, fits the observed data in all but Cluster 5B during these weeks. Another significant discrepancy between RF and LSTM is around weeks 9–14 in Cluster 4B, in which RF significantly underestimates the observed data, while LSTM produces a good fit.

Figures 5.10(a) and 5.10(b) compare the predicted data by RF and LSTM to the observed test data with a scatterplot visualisation. Here, it is seen that LSTM follows the highest observed values better than RF for all clusters in both years. Again, this is indicative for better capability to forecast possible disease outbreaks. LSTM also follows the lowest observed values better in Clusters 2A–5A in 2017. In 2018, LSTM follows the lowest observed values better in Clusters 3B and 4B.

In Serra, first we discuss the how the models perform on the training data in both years as presented in Figures 5.9(e) and 5.9(f). In 2017, on the training data (Figure 5.9(e)), RF overestimates the observed data in weeks 4–9 for all clusters except Cluster 2A where it underestimates the observed data. Also, RF underestimates the peak reached around observation week 19 in



(a) Vila velha, 2017

(b) Vila velha, 2018

(c) Serra, 2017

(d) Serra, 2018

FIGURE 5.11: Cluster-level comparison of mean absolute error (MAE) for LSTM and RF models. Lower MAE is desirable.

Cluster 5A. On the contrary, LSTM performs relatively well in all these periods. In 2018 (Figure 5.9(f)), RF underestimates the observed data around the observation weeks 4–9 in clusters 2B and 3B.

With regards to the test data performance of the models in Serra, the same disparities from the observed training data shown by RF in 2017 are also reproduced on the test data (See weeks 4–9 in Figure 5.9(g)). In 2018 (Figure 5.9(h)), notable disparities between observed and fitted data by both test models are seen in Cluster 1B starting from observation week 39. This is another case of high intra-cluster variance which has led to different patterns in training and test data of the same cluster. Regardless, for this cluster and in this period, LSTM still attempts to capture some of the temporal variations in this period while RF remains relatively invariant during this period.

From the visualization offered by the scatterplots presented in Figures 5.10(a) and 5.10(b), we observe that, just like in Vila Velha, LSTM follows the lowers observed values better in Serra for all clusters in both years. Also, LSTM follows the highest observed values better in Clusters 2A, 4A, 5A, 1B, 2B, 3B and 5B.

Overall, LSTM qualitatively outperforms RF in generalizing to the test data in both test locations. By following the highest and lower observed data better, LSTM provides the most reliable model for an outbreak early warning system.

Finally, the errors produced by the LSTM models in each location are examined at cluster level in comparison to their RF counterparts. This analysis is presented in Figure 5.11. Here, it is shown that in Vila Velha, LSTM produced less MAE than RF in five out of six clusters in both years. In Serra, LSTM produces better results in four out of the five clusters in 2017 and in all five clusters in 2018.

5.6 Discussion

According to Figure 5.4, the clustering approach applied in this study succeeds in finding common patterns in the trap points series. In this way, the problem of forecasting the many underlying series obtained from the traps is simplified to that of forecasting fewer series. The pattern similarity among traps series in the same cluster is captured by the similarity in the observed training and test data which have been obtained as averages of randomly selected series as shown in Figure 5.9. By this method, we have reduced the forecasting task significantly. In Vila Velha, starting from 195 and 325 trap points series in 2017 and 2018, respectively, we obtain 6 cluster for each year that describe the underlying mosquito vector activity of interest during the time observed. In Serra, starting from 567 and 95 traps in 2017 and 2018, respectively, we are able to summarize them into five clusters in each year. It is noteworthy that the optimal number of underlying clusters obtained is the same for each location in the two observation years. This shows that, in spite of different control conditions and non-matching climate seasons in both years, the underlying pattern mechanisms of the female *Ae. aegypti* in these locations are continuous.

The results show that freely accessible satellite image products which have formed the basis of recent studies [53, 54] in *Ae. aegypti* population dynamics modeling are available at spatial resolutions that make them informative for neighborhood-level temporal modeling. This is useful for muncipality, regional or national monitoring, where a larval survey approach is used to plan preventive and recovery actions. This approach requires that designated field inspectors visit all traps weekly at specified times to inspect and collect the data, resulting sometimes in missing data due to insufficient manpower. This issue is observable in the data of this research because, as mentioned above, among 791 traps in Vila Velha, only 193 traps had significant records for the whole year 2017. Since the cost of collecting in situ data is very high, the financial inefficiency resulting from of large amounts of missing data is also very high. As a result, the framework proposed in this work can serve not only for forecasting purposes, but also for spatio-temporal gap filling, especially when a trap location with missing data had previously be classified into a cluster.

Another point worth discussing is the importance of the lagging effects on some variables. Indeed, many studies have reported varying associations lagged effects between environmental conditions and dengue virus spread dynamics. In the studies presented in Chapters 3 and 4, two weeks of lag was chosen to represent non-synchronous environmental effects. [54] and [53], on the other hand, chose three weeks for same take but in a different study location from Vila Velha. These studies base their choice of lag window on *a priori* entomological knowledge of mosquito development life cycle. However, as reported in [106], this prior based lagged effects knowledge does not generalize globally, and is not necessarily the same for every considered environmental condition. For example, increase in dengue risk has been associated with increasing minimum and maximum temperature by 1–2 two month lags in Mexico, French west indies, and Brazil. Countries closer to the equator, e.g Singapore and Indonesia, report shorter lag effects (2–4 weeks) of temperature on the dengue cases. The study in [160] presents the temporal analysis of the relationship between dengue virus (not vector) and climatic variables in Rio de Janeiro, Brazil between the years 2001–2009. The best result in that study was obtained by considering four weeks ($T = 4$) lag effects of both precipitation and temperature variables. In line with all these works, the results in this research show that $T = 3$ is the most significant choice, in accordance with empirical evidence.

In this study, an experimental approach towards choosing the right temporal window not only in terms of size, but maintaining the sequential ordering of the considered lagged series, has been considered. Indeed, a major advantage of RNNs is that they can, within a specified time window, automatically learn the right lag dependencies differently for each considered environmental variable feature. As shown by the results in Section 5.5.3, learning the sequential dependency in lagged temporal windows improves the quality of our model. This improvement in quality generalizes across multiple vector control regimes and in two different locations.

Finally, as seen in Figure 5.7, the mosquito trap point clusters which were derived in this study are not concentrated around the same geographical zones. Further studies could explore the use of high resolution data to explore micro-climatic effects that drive the intra and inter-cluster environmental differences.

5.7 Chapter conclusions

While in Chapters 3 and 4 RF and GLM approaches were exploited for municipality-level 'now-casting' of *Ae. aegypti* population counts, in this chapter the same satellite image features were used to design a neighborhood-level forecasting framework. To this aim, autoregressive (past vector dynamics) and exogeneous (environmental effects) components were both included in the proposed model, and RNNs were used to learn the model parameters and sequential dependency, especially considering lagged effects.

Eventually, this study results in the following contributions:

- a general RNN-based algorithm for neighborhood-level time series female *Ae. aegypti* population one-week-ahead forecasting using EO products has been proposed and validated;
- forecasting values with better accuracy than those by a state-or-the-art multi-output variant of random forest (RF) has been obtained;
- by applying our modeling pipeline to data from different time periods, the proposed approach has been proved as robust and with generalization capabilities to different conditions;
- finally, the reported results prove that, by using the resulting models in two time periods, the proposed method improves existing vector surveillance techniques in terms of cost, time, and man-power efficiency.

Chapter 6

Conclusions

This chapter provides a comprehensive review of the results of the researches presented in the previous chapters and concludes the thesis by summarizing the novel achievements and highlighting possible future research paths.

6.1 Contributions by this thesis

This thesis has presented novel contributions with regards to the use of optical (multispectral) EO data for epidemiological modeling in urban areas. This domain has gained more attention in recent years for two reasons: (i) urbanisation, whose efficient management is one of the most urgent issue for humanity, and (ii) availability of more EO data from which better details of urban area land covers and of environmental variable can be extracted. Optical EO data provide globally consistent information on vegetation condition, humidity, land surface temperature and precipitation, which are proxy drivers in the propagation of certain disease vectors.

The main topic of this thesis is the prediction of *Ae aegypti* counts because this vector transmits widespread diseases, such as Zika, Dengue, Chikungunya, and Yellow fever. Previous studies which have made contributions in this domain still leave a few gaps. For the case of spatial modeling where land use (specifically vegetation) maps are required as input into the modeling procedure, studies conducted at high resolution ($\approx 6\,\mathrm{m}$) use commercial EO data which are expensive to access. As a result, most other studies rely on Landsat data ($30\,\mathrm{m}$ spatial resolution). In addition, all these studies do not obtain the needed land use map in a robust way. In response, this thesis envisages the use of Sentinel-2 data sets, which are free and provide unprecedented spectral and spatial resolution. For this reason, a robust vegetation mapping technique has been presented.

In addition, temporal *Ae aegypti* population models are generally obtained at municipality level. Ideally, such models should be both accurate and explainable in other to provide detailed information for control activities planning. This calls for trade-offs between model quality, selection, and explanability.

The specific methodologies presented in this thesis correspond to the following list:

- **A high resolution urban vegetation mapping procedure with Sentinel-2 data**. As already mentioned, quality vegetation maps can support the development of better spatial epidemiological models. The mapping methodology presented in this thesis exploits the high temporal resolution of Sentinel-2 to create seasonally aggregated inputs. Also, NDSV features were compared with spectral features, and an RF classifier with SVM and CART classification models. The results show that that seasonally aggregated inputs show better performances than annual greenest pixel for the task at hand, while NDSV improves the separation between the classes "Trees" and "Grass".

- **An accurate and explainable machine learning approach for EO-based temporal population modeling of *Ae aegypti* population at municipality level**. In this regard, this thesis has presented a procedure for explainable modeling of *Ae. aegypti* using EO data estimated environmental factors. Random forest (RF) regression was chosen for modeling, while its wrapped-in quantitative measure of the variable importance (MDI) was used to extract and rank the most informative environmental features. To prove the robustness of RF for the task, other machine learning models including SVR, ANN, KNN and DTR, as well as statistical models, such as LM and GLM, were fitted as baselines. The results show that the RF-based approach is capable of better mapping the complex relationship among the EO variables and vector population. Furthermore, the features selected thanks to the MDI value ranking can be empirically interpreted using the relation curves, and provide hints about the relationships among vector population and these environmental conditions from an operational point of view.

- **A robust statistical regression approach for *Ae aegypti* vector population modeling using EO data**. This thesis has proposed a Weighted Poisson GLM approach, which is able to achieve machine learning (ML) quality results, while also providing the capability to explicitly interpret the causality in the model. To this aim, model selection was performed with the Akaike Information Criterion (AIC), an improved approach with regards to linear correlation which has been used for same task in similar studies [53, 54]. The difference between GLM and RF approach is that GLMs provide model equations which are intuitive to interpret and also provide directions for each EO variable impact on vector population. Public health managers will have the GLM equation(s) at hand, so they will not need weekly complete mosquito data to make predictions. Hence, the models can be used to plan control activities without full mosquito data.

- **A spatio-temporal forecast epidemiological modeling methodology based on Recurrent Neural Network (RNN)**. This thesis has finally introduced a first attempt at neighborhood level one-week-ahead forecasting of *Ae. aegypti* vector population in urban areas. In many cases, due to micro-climatic effects, vector population and disease risk are not uniformly distributed in a city. As a result, municipality level models might fall short. From the experiments presented in Chapter 5, in comparison to RFs which were selected as baseline, the proposed RNN-based Nonlinear Autoregressive Exogenous (NARX) modeling technique is able to leverage estimates of environmental effects obtained from MODIS and GPM remote sensing data to achieve spatio-temporal forecast.

From a geographical point of view, the areas of the world majorly exposed to risks of arbovirosis carried by *Ae. aegypti* mosquito species include Latin America, Central Africa, and South-East Asia, with major disease spread outbreaks already recorded in Brazil, Argentina, Colombia, and Venezuela. Consequently, all experiments reported in this thesis have been performed using data in Argentina and Brazil. Similarly, the discussion about the results may be used to support existing control actions efforts in these countries, and may also be generalised to surrounding countries, or others ones with similar climate profiles.

6.2 Future directions

First of all, apart from the the diseases caused by *Ae. aegypti* mosquito vectors, the methodologies and EO data layers proposed in this study may be useful to model the distribution in urban areas of other kinds of climate-depended pathogens, but more studies should be performed to validate this proposition.

Furthermore, while this thesis has introduced a framework to map different kinds of vegetation in urban areas from Sentinel-2 data, the resulting maps have not been applied for the task of spatial epidemiological modeling. Many studies that focus on spatial epidemiological modeling with free multispectral EO data produce maps that are results of annual aggregates of seasonal data. This is due to limitations like clouds and other distortions affecting optical EO data. Therefore, future studies can consider using the method proposed in this thesis to obtain urban vegetation maps at $10\,\mathrm{m}$ resolution. The resulting maps may then be used for further epidemiological modeling procedures. In addition, thanks to the high temporal resolution of Sentinel-2, the vegetation mapping procedure proposed in this study may be applied to obtain seasonal vegetation maps, which in turn may be exploited to derive seasonal spatial diseases risk models.

With regards to the temporal modeling of *Ae. aegypti* population at municipality level (cf. Chapters 3 and 4), future studies may consider including autoregressive components. Indeed, in the

NARX formulation introduced for spatio-temporal modeling in Chapter 5, the resulting model does not degrade even in case of different (more stringent) control actions put in place by the municipality in the second yea of the study. This is because the vector control effects are implicitly captured by the autoregressive component of the model. As a result, and as more control efforts are put in place, temporal models may also include these effects to ensure the EO-based models still maintain informative quality and explainability. Also, further studies should be performed to validate the ”how” and ”if” referring to the generalization of the proposed methods to neighboring areas, as well as to other parts of the world with different climate conditions.

With regards to spatio-temporal forecasting, the thesis presented an EO-based approach for one-step-ahead forecast of *Ae. aegypti*. Further studies can consider generalising this into multi-step-ahead forecasting. Such method will further support operative actions to better anticipate outbreaks and manage resources. In addition, methods to explain deep neural network predictions can be applied to understand the space-time effects of the EO variables used to obtain the risk models.

Bibliography

[1] United Nations, "2018 revision of world urbanization prospects," 2018.

[2] R. Godfrey and M. Julien, "Urbanisation and health," *Clinical Medicine*, vol. 5, no. 2, p. 137, 2005.

[3] R. Ahas, A. Aasa, Y. Yuan, M. Raubal, Z. Smoreda, Y. Liu, C. Ziemlicki, M. Tiru, and M. Zook, "Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in harbin, paris, and tallinn," *International Journal of Geographical Information Science*, vol. 29, no. 11, pp. 2017–2039, 2015.

[4] R. Goldblatt, W. You, G. Hanson, and A. K. Khandelwal, "Detecting the boundaries of urban areas in india: A dataset for pixel-based image classification in google earth engine," *Remote Sensing*, vol. 8, p. 634, 2016.

[5] N. Joshi, M. Baumann, A. Ehammer, R. Fensholt, K. Grogan, P. Hostert, M. Jepsen, T. Kuemmerle, P. Meyfroidt, E. Mitchard, J. Reiche, C. Ryan, and B. Waske, "A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring," *Remote Sensing*, vol. 8, no. 1, p. 70, Jan 2016.

[6] G. Trianni, G. Lisini, E. Angiuli, E. A. Moreno, P. Dondi, A. Gaggia, and P. Gamba, "Scaling up to national/regional urban extent mapping using Landsat data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 7, pp. 3710–3719, Jul 2015.

[7] W. Grey and A. Luckman, "Mapping urban extent using satellite radar interferometry," *Photogrammetric Engineering & Remote Sensing*, vol. 69, no. 9, pp. 957–961, Sep 2003.

[8] Y. Ban, A. Jacob, and P. Gamba, "Spaceborne SAR data for global urban mapping at 30m resolution using a robust urban extractor," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 103, pp. 28–37, May 2015.

[9] J. Haas and Y. Ban, "Sentinel-1a SAR and sentinel-2a MSI data fusion for urban ecosystem service mapping," *Remote Sensing Applications: Society and Environment*, vol. 8, pp. 41–53, Nov 2017.

[10] M. L. Imhoff, P. Zhang, R. E. Wolfe, and L. Bounoua, "Remote sensing of the urban heat island effect across biomes in the continental USA," *Remote Sensing of Environment*, vol. 114, no. 3, pp. 504–513, Mar 2010.

[11] P. Gamba, B. Houshmand, and M. Saccani, "Detection and extraction of buildings from interferometric SAR data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 1, pp. 611–617, 2000.

[12] R. Sharma, B. Kumar, N. Desai, and V. Gujraty, "SAR for disaster management," *IEEE Aerospace and Electronic Systems Magazine*, vol. 23, no. 6, pp. 4–9, Jun 2008.

[13] F. Zhang, X. Zhu, and D. Liu, "Blending MODIS and Landsat images for urban flood mapping," *International Journal of Remote Sensing*, vol. 35, no. 9, pp. 3237–3253, Apr 2014.

[14] H. Taubenböck, M. Wegmann, A. Roth, H. Mehl, and S. Dech, "Urbanization in India – spatiotemporal analysis using remote sensing data," *Computers, Environment and Urban Systems*, vol. 33, no. 3, pp. 179–188, May 2009.

[15] A. M. Dewan and Y. Yamaguchi, "Land use and land cover change in greater Dhaka, bangladesh: Using remote sensing to promote sustainable urbanization," *Applied Geography*, vol. 29, no. 3, pp. 390–401, Jul 2009.

[16] Y. Ban and O. A. Yousif, "Multitemporal spaceborne SAR data for urban change detection in China," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 4, pp. 1087–1094, Aug 2012.

[17] United Nations Department of Social Affairs, "Sustainable development - the 17 goals," 2020, Accessed: 27.08.2020. [Online]. Available: https://sdgs.un.org/goals

[18] Y. Zhang and B. Guindon, "Using satellite remote sensing to survey transport-related urban sustainability," *International Journal of Applied Earth Observation and Geoinformation*, vol. 8, no. 3, pp. 149–164, Sep 2006.

[19] L. Beck, "Remote sensing and human health: New sensors and new opportunities," *Emerging Infectious Diseases*, vol. 6, no. 3, pp. 217–227, Jun 2000.

[20] J. C. Duque, J. E. Patino, L. A. Ruiz, and J. E. Pardo-Pascual, "Measuring intra-urban poverty using land cover and texture metrics derived from remote sensing data," *Landscape and Urban Planning*, vol. 135, pp. 11–21, Mar 2015.

[21] E. L. Moreno, *Slums of the World: The Face of Urban Poverty in the New Millennium?: Monitoring the Millennium Development Goal, Target 11–world-wide Slum Dweller Estimation*. UN-Habitat, 2003.

[22] P. Gupta, S. A. Christopher, J. Wang, R. Gehrig, Y. Lee, and N. Kumar, "Satellite remote sensing of particulate matter and air quality assessment over global cities," *Atmospheric Environment*, vol. 40, no. 30, pp. 5880–5892, Sep 2006.

[23] C. Lo and D. A. Quattrochi, "Land-use and land-cover change, urban heat island phenomenon, and health implications," *Photogrammetric Engineering & Remote Sensing*, vol. 69, no. 9, pp. 1053–1063, Sep 2003.

[24] C. Neiderud, "How urbanization affects the epidemiology of emerging infectious diseases," *Infection Ecology & Epidemiology*, vol. 5, no. 1, p. 27060, Jan 2015.

[25] S. G. Young, J. A. Tullis, and J. Cothren, "A remote sensing and GIS-assisted landscape epidemiology approach to West Nile virus," *Applied Geography*, vol. 45, pp. 241–249, Dec 2013.

[26] S. Goetz, S. Prince, and J. Small, "Advances in satellite remote sensing of environmental variables for epidemiological applications," in *Remote Sensing and Geographical Information Systems in Epidemiology*. Elsevier, 2000, pp. 289–307.

[27] Centre for Remote Imaging, Sensing, and Processing, "Optical remote sensing," 2001, Accessed: 27.08.2020. [Online]. Available: https://crisp.nus.edu.sg/~research/tutorial/optical.htm

[28] A. Huete, K. Didan, T. Miura, E. Rodriguez, X. Gao, and L. Ferreira, "Overview of the radiometric and biophysical performance of the MODIS vegetation indices," *Remote Sensing of Environment*, vol. 83, no. 1-2, pp. 195–213, Nov 2002.

[29] B.-C. Gao, "NDWI — a normalized difference water index for remote sensing of vegetation liquid water from space," *Remote sensing of environment*, vol. 58, no. 3, pp. 257–266, 1996.

[30] M. L. García and V. Caselles, "Mapping burns and natural reforestation using Thematic Mapper data," *Geocarto International*, vol. 6, no. 1, pp. 31–37, Mar 1991.

[31] J. A. Sobrino, J. C. Jiménez-Muñoz, and L. Paolini, "Land surface temperature retrieval from LANDSAT TM 5," *Remote Sensing of Environment*, vol. 90, no. 4, pp. 434–440, Apr 2004.

[32] V. Herbreteau, G. Salem, M. Souris, J.-P. Hugot, and J.-P. Gonzalez, "Thirty years of use and improvement of remote sensing, applied to epidemiology: From early promises to lasting frustration," *Health & Place*, vol. 13, no. 2, pp. 400–403, Jun 2007.

[33] C. E. Woodcock, R. Allen, M. Anderson, A. Belward, R. Bindschadler, W. Cohen, F. Gao, S. N. Goward, D. Helder, E. Helmer *et al.*, "Free access to Landsat imagery." *SCIENCE VOL 320: 1011*, 2008.

[34] A. S. Belward and J. O. Skøien, "Who launched what, when and why trends in global landcover observation capacity from civilian earth observation satellites," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 103, pp. 115–128, May 2015.

[35] J. S. Miguel-Ayanz, "Comparison of single-stage and multi-stage classification approaches for cover type mapping with TM and SPOT data," *Remote Sensing of Environment*, vol. 59, no. 1, pp. 92–104, Jan 1997.

[36] D. F. Thompson, J. B. Malone, M. Harb, R. Faris, O. K. Huh, A. A. Buck, and B. L. Cline, "Bancroftian filariasis distribution and diurnal temperature differences in the southern Nile delta." *Emerging infectious diseases*, vol. 2, no. 3, p. 234, 1996.

[37] S. I. Hay, R. W. Snow, and D. J. Rogers, "Predicting malaria seasons in Kenya using multitemporal meteorological satellite sensor data," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 92, no. 1, pp. 12–20, Jan 1998.

[38] K. O. Pope, E. Rejmankova, H. M. Savage, J. I. Arredondo-Jimenez, M. H. Rodriguez, and D. R. Roberts, "Remote sensing of tropical wetlands for Malaria control in Chiapas, Mexico," *Ecological Applications*, vol. 4, no. 1, pp. 81–90, Feb 1994.

[39] C. Justice, J. Townshend, E. Vermote, E. Masuoka, R. Wolfe, N. Saleous, D. Roy, and J. Morisette, "An overview of MODIS land data processing and product status," *Remote Sensing of Environment*, vol. 83, no. 1-2, pp. 3–15, Nov 2002.

[40] Z. Wan, Y. Zhang, Q. Zhang, and Z.-L. Li, "Quality assessment and validation of the MODIS global land surface temperature," *International Journal of Remote Sensing*, vol. 25, no. 1, pp. 261–274, jan 2004.

[41] Google, "Google Scholar," 2021, Accessed: 01.03.2021. [Online]. Available: https://scholar.google.com

[42] J. P. Messina, M. U. Kraemer, O. J. Brady, D. M. Pigott, F. M. Shearer, D. J. Weiss, N. Golding, C. W. Ruktanonchai, P. W. Gething, E. Cohn, J. S. Brownstein, K. Khan, A. J. Tatem, T. Jaenisch, C. J. Murray, F. Marinho, T. W. Scott, and S. I. Hay, "Mapping global environmental suitability for Zika virus," *eLife*, vol. 5, Apr 2016.

[43] S. R. Christophers, *Aedes aegypti: the yellow fever mosquito*. CUP Archive, 1960.

[44] J. F. Obenauer, T. A. Joyner, and J. B. Harris, "The importance of human population characteristics in modeling Aedes aegypti distributions and assessing risk of mosquito-borne infectious diseases," *Tropical Medicine and Health*, vol. 45, no. 1, Nov 2017.

[45] C. W. Morin, A. C. Comrie, and K. Ernst, "Climate and dengue transmission: evidence and implications," *Environmental health perspectives*, vol. 121, no. 11-12, pp. 1264–1272, 2013.

[46] L. Lambrechts, K. P. Paaijmans, T. Fansiri, L. B. Carrington, L. D. Kramer, M. B. Thomas, and T. W. Scott, "Impact of daily temperature fluctuations on dengue virus transmission by Aedes aegypti," *Proceedings of the National Academy of Sciences*, vol. 108, no. 18, pp. 7460–7465, Apr 2011.

[47] D. Rogers and S. Randolph, "Climate change and vector-borne diseases," in *Advances in Parasitology*. Elsevier, 2006, pp. 345–381.

[48] G. Chowell and F. Sanchez, "Climate-based descriptive models of dengue fever: the 2002 epidemic in Colima, Mexico." *Journal of environmental health*, vol. 68, no. 10, 2006.

[49] E. P. Astuti, P. W. Dhewantara, H. Prasetyowati, M. Ipa, C. Herawati, and K. Hendrayana, "Paediatric dengue infection in Cirebon, Indonesia: a temporal and spatial analysis of notified dengue incidence to inform surveillance," *Parasites & Vectors*, vol. 12, no. 1, Apr 2019.

[50] O. J. Brady, M. A. Johansson, C. A. Guerra, S. Bhatt, N. Golding, D. M. Pigott, H. Delatte, M. G. Grech, P. T. Leisnham, R. M. de Freitas, L. M. Styer, D. L. Smith, T. W. Scott, P. W. Gething, and S. I. Hay, "Modelling adult Aedes aegypti and Aedes albopictus survival at different temperatures in laboratory and field settings," *Parasites & Vectors*, vol. 6, no. 1, Dec 2013.

[51] D. O. Fuller, A. Troyo, and J. C. Beier, "El niño southern oscillation and vegetation dynamics as predictors of dengue fever cases in Costa Rica," *Environmental Research Letters*, vol. 4, no. 1, p. 014011, Jan 2009.

[52] F. J. Colón-González, G. Bentham, and I. R. Lake, "Climate variability and Dengue fever in warm and humid mexico," *The American Journal of Tropical Medicine and Hygiene*, vol. 84, no. 5, pp. 757–763, May 2011.

[53] A. German, M. Espinosa, M. Abril, and C. Scavuzzo, "Exploring satellite based temporal forecast modelling of Aedes aegypti oviposition from an operational perspective," *Remote Sensing Applications: Society and Environment*, vol. 11, pp. 231–240, Aug 2018.

[54] J. M. Scavuzzo, F. Trucco, M. Espinosa, C. B. Tauro, M. Abril, C. M. Scavuzzo, and A. C. Frery, "Modeling Dengue vector population using remotely sensed data and machine learning," *Acta Tropica*, vol. 185, pp. 167–175, Sep 2018.

[55] M. Espinosa, D. Weinberg, C. H. Rotela, F. Polop, M. Abril, and C. M. Scavuzzo, "Temporal dynamics and spatial patterns of Aedes aegypti breeding sites, in the context of a dengue control program in Tartagal (Salta Province, Argentina)," *PLOS Neglected Tropical Diseases*, vol. 10, no. 5, p. e0004621, May 2016.

[56] S. J. Phillips, R. P. Anderson, and R. E. Schapire, "Maximum entropy modeling of species geographic distributions," *Ecological Modelling*, vol. 190, no. 3-4, pp. 231–259, Jan 2006.

[57] C. Rotela, F. Fouque, M. Lamfri, P. Sabatier, V. Introini, M. Zaidenberg, and C. Scavuzzo, "Space–time analysis of the dengue spreading dynamics in the 2004 Tartagal outbreak, Northern Argentina," *Acta Tropica*, vol. 103, no. 1, pp. 1–13, Jul 2007.

[58] E. P. Crist and R. C. Cicone, "A physically-based transformation of Thematic Mapper data—the TM tasseled cap," *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-22, no. 3, pp. 256–263, May 1984.

[59] G. De'Ath, "Boosted trees for ecological modeling and prediction," *Ecology*, vol. 88, no. 1, pp. 243–251, 2007.

[60] R. Hijmans, S. Cameron, J. Parra, P. Jones, and A. Jarvis, "The worldclim interpolated global terrestrial climate surfaces. version 1.3," 2004.

[61] Q. Lin, "Enhanced vegetation index using Moderate Resolution Imaging Spectroradiometers," in *2012 5th International Congress on Image and Signal Processing*. IEEE, Oct 2012.

[62] A. Schneider, M. A. Friedl, and D. Potere, "A new map of global urban extent from MODIS satellite data," *Environmental Research Letters*, vol. 4, no. 4, p. 044003, Oct 2009.

[63] j. . P. y. . . . v. . . United Nations, title = World urbanization prospects: The 2014 revision, highlights. department of economic and social affairs.

[64] A. Ziemann, G. Fairchild, J. Conrad, C. Manore, N. Parikh, S. D. Valle, and N. Generous, "Predicting Dengue incidence in Brazil using broad-scale spectral remote sensing imagery," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, Jul 2018.

[65] A. Gasparrini, B. Armstrong, and M. G. Kenward, "Distributed lag non-linear models," *Statistics in Medicine*, vol. 29, no. 21, pp. 2224–2234, Aug 2010.

[66] E. L. Estallo, E. M. Benitez, M. A. Lanfri, C. M. Scavuzzo, and W. R. Almiron, "MODIS environmental data to assess Chikungunya, Dengue, and Zika diseases through Aedes (stegomia) aegypti oviposition activity estimation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 12, pp. 5461–5466, Dec 2016.

[67] M. Espinosa, E. M. A. D. Fino, M. Abril, M. Lanfri, M. V. Periago, and C. M. Scavuzzo, "Operational satellite-based temporal modelling of Aedes population in Argentina," *Geospatial Health*, vol. 13, no. 2, Nov 2018.

[68] C. Kummerow, W. Barnes, T. Kozu, J. Shiue, and J. Simpson, "The tropical rainfall measuring mission (TRMM) sensor package," *Journal of atmospheric and oceanic technology*, vol. 15, no. 3, pp. 809–817, 1998.

[69] G. Skofronick-Jackson, W. A. Petersen, W. Berg, C. Kidd, E. F. Stocker, D. B. Kirschbaum, R. Kakar, S. A. Braun, G. J. Huffman, T. Iguchi, P. E. Kirstetter, C. Kummerow, R. Meneghini, R. Oki, W. S. Olson, Y. N. Takayabu, K. Furukawa, and T. Wilheit, "The Global Precipitation Measurement (GPM) mission for science and society," *Bulletin of the American Meteorological Society*, vol. 98, no. 8, pp. 1679–1695, Aug 2017.

[70] C. H. Rotela, L. I. Spinsanti, M. A. Lamfri, M. S. Contigiani, W. R. Almirón, and C. M. Scavuzzo, "Mapping environmental susceptibility to Saint Louis encephalitis virus, based on a decision tree model of remotely sensed data," *Geospatial health*, vol. 6, no. 1, p. 85, Nov 2011.

[71] M. U. Kraemer, M. E. Sinka, K. A. Duda, A. Q. Mylne, F. M. Shearer, C. M. Barker, C. G. Moore, R. G. Carvalho, G. E. Coelho, W. V. Bortel, G. Hendrickx, F. Schaffner, I. R. Elyazar, H.-J. Teng, O. J. Brady, J. P. Messina, D. M. Pigott, T. W. Scott, D. L. Smith, G. W. Wint, N. Golding, and S. I. Hay, "The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus," *eLife*, vol. 4, Jun 2015.

[72] H. L. Cleckner, T. R. Allen, and A. S. Bellows, "Remote sensing and modeling of mosquito abundance and habitats in coastal Virginia, USA," *Remote Sensing*, vol. 3, no. 12, pp. 2663–2681, Dec 2011.

[73] M. Alberti, "The effects of urban patterns on ecosystem function," *International Regional Science Review*, vol. 28, no. 2, pp. 168–192, Apr 2005.

[74] T. V. de Voorde, J. Vlaeminck, and F. Canters, "Comparing Different Approaches for Mapping Urban Vegetation Cover from Landsat ETM+ Data: A Case Study on Brussels," *Sensors*, vol. 8, no. 6, pp. 3880–3902, Jun 2008.

[75] F. van der Meer, "Image classification through spectral unmixing," in *Spatial Statistics for Remote Sensing*. Springer, 1999, pp. 185–193.

[76] A. J. Elmore, J. F. Mustard, S. J. Manning, and D. B. Lobell, "Quantifying vegetation change in semiarid environments," *Remote Sensing of Environment*, vol. 73, no. 1, pp. 87–102, Jul 2000.

[77] J. Yan, W. Zhou, L. Han, and Y. Qian, "Mapping vegetation functional types in urban areas with WorldView-2 imagery: Integrating object-based classification with phenology," *Urban Forestry & Urban Greening*, vol. 31, pp. 230–240, Apr 2018.

[78] C. P. of the European Commission. Sentinel satellites. Accessed: 2017-08-22. [Online]. Available: https://sentinel.esa.int/web/sentinel/missions/sentinel-2

[79] The European Space Agency, "For Corpenicus," 2020, Accessed: 11.09.2020. [Online]. Available: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2/Downloads

[80] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote sensing of Environment*, vol. 202, pp. 18–27, 2017.

[81] E. Angiuli and G. Trianni, "Urban mapping in Landsat images based on normalized difference spectral vector," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 3, pp. 661–665, Mar 2014.

[82] N. N. Patel, E. Angiuli, P. Gamba, A. Gaughan, G. Lisini, F. R. Stevens, A. J. Tatem, and G. Trianni, "Multitemporal settlement and population mapping from landsat using google earth engine," *International Journal of Applied Earth Observation and Geoinformation*, vol. 35, pp. 199–208, 2015.

[83] W.-Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[84] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. j. Rajabi, "Advantage and drawback of support vector machine functionality," in *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, Sep 2014, pp. 63–65.

[85] M. Pal, "Ensemble learning with decision tree for remote sensing classification," *World Academy of Science, Engineering and Technology*, vol. 36, pp. 258–260, 2007.

[86] ——, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, Jan 2005.

[87] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, Apr 2016.

[88] O. Arino, "GlobCover 2009," 2010.

[89] C. Rotela, L. Lopez, M. F. Céspedes, G. Barbas, A. Lighezzolo, X. Porcasi, M. A. Lanfri, C. M. Scavuzzo, and D. E. Gorla, "Analytical report of the 2016 dengue outbreak in Córdoba city, Argentina," *Geospatial Health*, Nov 2017.

[90] L. L. Janssen and F. J. Vanderwel, "Accuracy assessment of satellite derived land-cover data: A review," *Photogrammetric Engineering and Remote Sensing; (United States)*, vol. 60:4, Apr 1994.

[91] R. Ostfeld, G. Glass, and F. Keesing, "Spatial epidemiology: an emerging (or re-emerging) discipline," *Trends in Ecology & Evolution*, vol. 20, no. 6, pp. 328–336, jun 2005.

[92] T. L. Johnson, U. Haque, A. J. Monaghan, L. Eisen, M. B. Hahn, M. H. Hayden, H. M. Savage, J. McAllister, J.-P. Mutebi, and R. J. Eisen, "Modeling the environmental suitability for Aedes (stegomyia) aegypti and Aedes (stegomyia) albopictus (diptera: Culicidae) in the contiguous United States," *Journal of Medical Entomology*, vol. 54, no. 6, pp. 1605–1614, sep 2017.

[93] F. J. Antonio, A. S. Itami, S. de Picoli, J. J. V. Teixeira, and R. dos Santos Mendes, "Spatial patterns of dengue cases in Brazil," *PLOS ONE*, vol. 12, no. 7, p. e0180715, jul 2017.

[94] Z. Liu, Z. Zhang, Z. Lai, T. Zhou, Z. Jia, J. Gu, K. Wu, and X.-G. Chen, "Temperature increase enhances aedes albopictus competence to transmit dengue virus," *Frontiers in Microbiology*, vol. 8, Dec 2017.

[95] A. E. Eiras, M. C. Resende, J. L. Acebal, and K. S. Paixão, "New cost-benefit of Brazilian technology for vector surveillance using trapping system," in *Malaria*. IntechOpen, dec 2019.

[96] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[97] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, Apr 2016.

[98] P. McCullagh, *Generalized linear models*. Routledge, 2018.

[99] A. R. dos Santos, F. S. de Oliveira, A. G. da Silva, J. M. Gleriani, W. Gonçalves, G. L. Moreira, F. G. Silva, E. R. F. Branco, M. M. Moura, R. G. da Silva, R. S. Juvanhol, K. B. de Souza, C. A. A. S. Ribeiro, V. T. de Queiroz, A. V. Costa, A. S. Lorenzon, G. F. Domingues, G. E. Marcatti, N. L. M. de Castro, R. T. Resende, D. E. Gonzales, L. A. de Almeida Telles, T. R. Teixeira, G. M. A. D. A. dos Santos, and P. H. S. Mota, "Spatial and temporal distribution of urban heat islands," *Science of The Total Environment*, vol. 605-606, pp. 946–956, Dec 2017.

[100] Brasil. Ministério da Saúde, "Boletim Epidemiológico v 51," 2020, Accessed: 10.05.2020. [Online]. Available: https://antigo.saude.gov.br/images/pdf/2020/janeiro/20/Boletim-epidemiologico-SVS-02-1-.pdf

[101] ——, "Programa Nacional de Controle da Dengue (PNCD)," 2002, Accessed: 19.05.2020. [Online]. Available: https://bvsms.saude.gov.br/bvs/publicacoes/pncd_2002.pdf

[102] H. Araújo, D. Carvalho, R. Ioshino, A. C. da Silva, and M. Capurro, "Aedes aegypti control strategies in Brazil: Incorporation of new technologies to overcome the persistence of dengue epidemics," *Insects*, vol. 6, no. 2, pp. 576–594, Jun 2015.

[103] D. P. O. de Melo, L. R. Scherrer, and Á. E. Eiras, "Dengue fever occurrence and vector detection by larval survey, ovitrap and MosquiTRAP: A space-time clusters analysis," *PLoS ONE*, vol. 7, no. 7, p. e42125, Jul 2012.

[104] Á. E. Eiras and M. C. Resende, "Preliminary evaluation of the "dengue-MI" technology for aedes aegyptimonitoring and control," *Cadernos de Saúde Pública*, vol. 25, no. suppl 1, pp. S45–S58, 2009.

[105] E. L. Estallo, M. A. Lamfri, C. M. Scavuzzo, F. F. L. Almeida, M. V. Introini, M. Zaidenberg, and W. R. Almirón, "Models for predicting aedes aegypti larval indices based on satellite images and climatic variables," *Journal of the American Mosquito Control Association*, vol. 24, no. 3, pp. 368–376, Sep 2008.

[106] Y. Cheong, K. Burkart, P. Leitão, and T. Lakes, "Assessing weather effects on Dengue disease in Malaysia," *International Journal of Environmental Research and Public Health*, vol. 10, no. 12, pp. 6319–6334, Nov 2013.

[107] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[108] C. Sammut and G. I. Webb, *Encyclopedia of machine learning and data mining*. Springer, 2017.

[109] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2013.

[110] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in neural information processing systems*, 2013, pp. 431–439.

[111] A. Karpatne, I. Ebert-Uphoff, S. Ravela, H. A. Babaie, and V. Kumar, "Machine learning for the geosciences: Challenges and opportunities," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 8, pp. 1544–1554, Aug 2019.

[112] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing," *Geoscience Frontiers*, vol. 7, no. 1, pp. 3–10, Jan 2016.

[113] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, Jun 2016.

[114] D. John, "Artificial intelligence in geoscience and remote sensing," in *Geoscience and Remote Sensing New Achievements*. InTech, Feb 2010.

[115] M. Riedmiller and I. Rprop, "Rprop - description and implementation details," 1994.

[116] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[117] M. Krzywinski and N. Altman, "Classification and regression trees," *Nature Methods*, vol. 14, no. 8, pp. 757–758, Aug 2017.

[118] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric statistical methods*. John Wiley & Sons, 2013, vol. 751.

[119] S.-C. Chen, C.-M. Liao, C.-P. Chio, H.-H. Chou, S.-H. You, and Y.-H. Cheng, "Lagged temperature effect with mosquito transmission potential explains Dengue variability in southern taiwan: Insights from a statistical analysis," *Science of The Total Environment*, vol. 408, no. 19, pp. 4069–4075, Sep 2010.

[120] L. Barsante, K. Paixão, K. Laass, R. Cardoso, A. Eiras, and J. Acebal, "A model to predict the population size of the Dengue fever vector based on rainfall data," *arXiv preprint arXiv:1409.7942*, 2014.

[121] C. C. Jansen and N. W. Beebe, "The dengue vector Aedes aegypti: what comes next," *Microbes and Infection*, vol. 12, no. 4, pp. 272–279, Apr 2010.

[122] K. Kamimura, I. T. Matsuse, H. Takahashi, J. Komukai, T. Fukuda, K. Suzuki, M. Aratani, Y. Shirai, and M. Mogi, "Effect of temperature on the development of Aedes aegypti and Aedes albopictus," *Medical Entomology and Zoology*, vol. 53, no. 1, pp. 53–58, 2002.

[123] N. B. Tjaden, S. M. Thomas, D. Fischer, and C. Beierkuhnlein, "Extrinsic incubation period of dengue: Knowledge, backlog, and applications of temperature dependence," *PLoS Neglected Tropical Diseases*, vol. 7, no. 6, p. e2207, Jun 2013.

[124] M. Trpiš, "Dry season survival of Aedes aegypti eggs in various breeding sites in the Dar es Salaam area, Tanzania," *Bulletin of the World Health Organization*, vol. 47, no. 3, p. 433, 1972.

[125] L. Walsh, "A short review of model selection techniques for radiation epidemiology," *Radiation and Environmental Biophysics*, vol. 46, no. 3, pp. 205–213, Apr 2007.

[126] W. McKinney, *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

[127] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[128] J. K. Lindsey, *Applying generalized linear models*. Springer Science & Business Media, 2000.

[129] A. Lovett, C. Bentham, and R. Flowerdew, "Analysing geographic variations in mortality using poisson regression: The example of ischaemic heart disease in England and Wales 1969–1973," *Social Science & Medicine*, vol. 23, no. 10, pp. 935–943, Jan 1986.

[130] H. Bozdogan, "Akaike's information criterion and recent developments in information complexity," *Journal of mathematical psychology*, vol. 44, no. 1, pp. 62–91, 2000.

[131] H. Akaike, *Information Theory and an Extension of the Maximum Likelihood Principle*. New York, NY: Springer New York, 1998, pp. 199–213.

[132] J.-S. Hwang and T.-H. Hu, "A stepwise regression algorithm for high-dimensional variable selection," *Journal of Statistical Computation and Simulation*, vol. 85, no. 9, pp. 1793–1806, Apr 2014.

[133] L. Valdez, G. Sibona, and C. Condat, "Impact of rainfall on Aedes aegypti populations," *Ecological Modelling*, vol. 385, pp. 96–105, Oct 2018.

[134] E. A. P. de Almeida Costa, E. M. de Mendonça Santos, J. C. Correia, and C. M. R. de Albuquerque, "Impact of small variations in temperature and humidity on the reproductive activity and survival of Aedes aegypti (diptera, culicidae)," *Revista Brasileira de Entomologia*, vol. 54, no. 3, pp. 488–493, 2010.

[135] A. Troyo, D. O. Fuller, O. Calderón-Arguedas, M. E. Solano, and J. C. Beier, "Urban structure and Dengue incidence in puntarenas, costa rica," *Singapore Journal of Tropical Geography*, vol. 30, no. 2, pp. 265–282, Jul 2009.

[136] D. A. da Cruz Ferreira, C. M. Degener, C. de Almeida Marques-Toledo, M. M. Bendati, L. O. Fetzer, C. P. Teixeira, and Á. E. Eiras, "Meteorological variables and mosquito monitoring are good predictors for infestation trends of Aedes aegypti, the vector of Dengue, Chikungunya and Zika," *Parasites & Vectors*, vol. 10, no. 1, Feb 2017.

[137] M. C. de Resende, I. M. Silva, B. R. Ellis, and A. E. Eiras, "A comparison of larval, ovitrap and MosquiTRAP surveillance for Aedes (stegomyia) aegypti," *Memórias do Instituto Oswaldo Cruz*, vol. 108, no. 8, pp. 1024–1030, Nov 2013.

[138] J. D. Hamilton, *Time series analysis*. Princeton New Jersey, 1994, vol. 2.

[139] A. Chattopadhyay, E. Nabizadeh, and P. Hassanzadeh, "Analog forecasting of extreme-causing weather patterns using deep learning," *Journal of Advances in Modeling Earth Systems*, 2020.

[140] N. Wu, B. Green, X. Ben, and S. O'Banion, "Deep transformer models for time series forecasting: The influenza prevalence case," *arXiv preprint arXiv:2001.08317*, 2020.

[141] Y. Qin, D. Song, H. Cheng, W. Cheng, G. Jiang, and G. W. Cottrell, "A dual-stage attention-based Recurrent Neural Network for time series prediction," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, ser. IJCAI'17. AAAI Press, 2017, p. 2627–2633.

[142] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[143] J. L. Elman, "Distributed representations, simple recurrent networks, and grammatical structure," *Machine learning*, vol. 7, no. 2-3, pp. 195–225, 1991.

[144] E. Diaconescu, "The use of NARX neural networks to predict chaotic time series," *Wseas Transactions on computer research*, vol. 3, no. 3, pp. 182–191, 2008.

[145] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," in *Advances in neural information processing systems*, 1997, pp. 473–479.

[146] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[147] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[148] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[149] V. Andreo, X. Porcasi, C. Rodriguez, L. Lopez, C. Guzman, and C. M. Scavuzzo, "Time series clustering applied to eco-epidemiology: the case of Aedes aegypti in Córdoba, Argentina," in *2019 XVIII Workshop on Information Processing and Control (RPIC)*. IEEE, 2019, pp. 93–98.

[150] S. H. Park, B. Kim, C. M. Kang, C. C. Chung, and J. W. Choi, "Sequence-to-sequence prediction of vehicle trajectory via LSTM encoder-decoder architecture," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, Jun 2018.

[151] X. Jin, C. Xu, J. Feng, Y. Wei, J. Xiong, and S. Yan, "Deep learning with s-shaped rectified linear activation units," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[152] A. Kassambara, *Practical guide to cluster analysis in R: Unsupervised machine learning*. STHDA, 2017, vol. 1.

[153] J. Paparrizos and L. Gravano, "k-shape: Efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 2015, pp. 1855–1870.

[154] A. Ng, "Clustering with the k-means algorithm," *Machine Learning*, 2012.

[155] H. Hou and H. Andrews, "Cubic splines for image interpolation and digital filtering," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 26, no. 6, pp. 508–517, 1978.

[156] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[157] M. Vijaymeena and K. Kavitha, "A survey on similarity measures in text mining," *Machine Learning and Applications: An International Journal*, vol. 3, no. 2, pp. 19–28, 2016.

[158] V. S. H. Rao and R. Durvasula, *Dynamic models of infectious diseases*. Springer, 2013, vol. 1.

[159] A. Hussain, F. Ali, O. B. Latiwesh, and S. Hussain, "A comprehensive review of the manifestations and pathogenesis of Zika virus in neonates and adults," *Cureus*, vol. 10, no. 9, 2018.

[160] A. F. Gomes, A. A. Nobre, and O. G. Cruz, "Temporal analysis of the relationship between Dengue and meteorological variables in the city of Rio de Janeiro, Brazil, 2001-2009," *Cadernos de Saúde Pública*, vol. 28, no. 11, pp. 2189–2197, 2012.