Introduction to Programming
Master Program in Data Science
and Advanced Analytics 2018/19

Investigation of factors that provoke the spread of *Aedes* aegypti and *Aedes albopictus*

Margarita Venediktova[1], Rodolfo Saldanha[2], Yu Chun Liu[3] & Shawkatul Islam Aziz[4]

[1]M20180211 - M20180211@novaims.unl.pt
[2]M20180088 - M20180088@novaims.unl.pt
[3]M20180220 - M20180220@novaims.unl.pt
[4]M20181035 - M20181035@novaims.unl.pt

Abstract: Aedes aegypti and Ae. albopictus are the main vectors transmitting dengue, chikungunya, Zika fever virus, Mayaro and Yellow fever viruses. This project aims to identify the factors that cause the expansion of vectors (mosquito population) in the timespan between 1991 and 2014. A dataset with occurrences of Aedes aegypti and Ae. albopictus mosquitoes was used as a base for further analysis. Multiple other variables were identified as potential factors of the increase of the population and were added to the base dataset for further analysis. These variable were incorporated in the model, normalized and used as an input of Self-Organizing map. Later, the initial dataset was denormalized and the results were analyzed and the conclusions were made. The description of data collection and preprocessing, as well as the implementation of self-organizing map technique and the results are discussed in the present paper.

Keywords: Vector-Borne Diseases; Aedes aegypti; Aedes albopictus

Statement of Contribution: Data Collection - All members;

Preprocessing data - Margarita Venediktova , Rodolfo Saldanha, Yu Chun Li;

Exploration of the data – All members;

Results and discussion -Margarita Venediktova , Rodolfo Saldanha, Yu Chun Li;

Report elaboration - Margarita Venediktova , Rodolfo Saldanha, Yu Chun Li.

I. Introduction

The present research focuses on the Aedes aegypti and albopictus vectors' occurrences worldwide in a timespan from 1991 to 2014. The main research goal was to identify the major factor affecting the spread of vectors around the globe. The most comprehensive dataset existing to this day was used in this research. The dataset of occurrences of vectors is used in this research along with other datasets containing information on the climate socioeconomic factors for each country.

The first step of the present research was to identify the factors which are considered critical in the development of Aedes aegypti and Aedes albopictus vectors. The variable commonly interpreted as pivotal were identified through literature review and other reliable sources like World Health Organization [24]. Later these variable were incorporated in the model, normalized and used as an input of Self-Organizing map that is the primary research technique of the paper.

Later, the initial dataset was denormalized and the results were analyzed and the conclusions were made. The most important conclusion of the paper is that based on the result of the SOM, that no factor in isolation can not be inferred as critical for the development of the "outbreak" in the population. However, a combination of several factors can be considered crucial for an increase in the number of occurrences.

II. Problem description and Research goal

First of all, Aedes Aegypti and Albopictus vectors, originated in Africa, are the mosquitos that transmit viruses [1] such as Dengue fever virus, Chikungunya virus, Zika fever virus, Mayaro and Yellow fever viruses, causing a significant fraction of the global infectious disease burden. The expansion of these diseases for instance, dengue [19] is expected to increase due to factors such as the modern dynamics of climate change, globalization, travel, trade, socioeconomics, settlement and also viral evolution, from minor tropical illnesses to diseases of worldwide importance is a demonstration of the importance of dealing with this threat. As the Aedes Aegypti and Albopictus vectors continue to spread worldwide, it is crucial to detect trends and identify the factors that may affect the surveillance of the mosquitos in order to proactively prevent vector establishment until the new technologies or the vaccine being developed [20].

Ecological and human factors appear to play a role in determining the increased incidence of vector-borne diseases. Despite the fact that increasing availability of tests and better awareness of clinicians contribute to more frequent recognition, the increase in the number of epidemics it still can be observed . In 1950, only 9 countries reported cases of dengue; the average annual number of cases reported to WHO varied from 908 in 1950–1959 to 514 139 in 1990–1999 spreaded to 128 countries [21].

Climate change [22], urbanization (in particular with degraded urban environments), human behaviors, mass gathering events, migration of humans and animals, development of air transport and extensive agriculture have all been suggested to have contributed to the rapid worldwide spread of vector-borne diseases. Moreover, epidemiological studies have shown that temperature is a factor in dengue transmission in urban areas. However, climate is only one of many factors affecting vector distribution, such as habitat destruction, land use, pesticide application, and population density [1].

Taken into consideration statements above, the main goal of this research is to collect and analyze data, explore multiple factors aiming to find meaningful patterns among them that may affect directly the increase of the number of occurrences of vectors in distinct countries in the timeframe of 1991 to 2014.

III. Data

a. Data Collection and Description

This section describes the relevance of each dataset to the project and decision-making process of choosing particular variables for further analysis.

Since the research goal was to identify the factors that provoke the sharp increase of vectors that transmit severe diseases, the research started by obtaining the dataset containing information on the number of vectors' occurrences. The most comprehensive dataset containing Aedes aegypti and Aedes albopictus vectors' occurrences was found and became the basis for further analysis.

Moreover, based on the literature review [19] and analysis of the problem in the section above some crucial variables were identified that can be represented in the following groups :

1. **Climate variables**: Monthly average temperature in Celsius and monthly average precipitation (mm) for all 151 countries during 1991-2015. Later the transformation of the climate data from the monthly to the yearly basis were be carried out.
2. **Social and economic variables:** The Human Development Index (HDI) [6] is a composite statistic of life expectancy, education, and income per capita indicators. which are used to rank countries into four tiers of human development, scored from 0-1. Although the HDI index already incorporates level of education as a parameter, it was decided to investigate it at education level by itself and to see if it could have a more direct correlation to the number of occurrences of the vectors. For education index, it takes mean year of schooling and expected year of schooling as parameters. For HDI, it is the geometric mean of the education level, and other two normalized indices.
3. **Population variables**: Annual total population size and population density (per square kilometer) for each country.
4. **Urbanization variables**: Percentage of urban and rural population by year per country to inspect the change of population density in urban and rural area. Definitions of an urban settlement vary widely across countries. However, 2500 to 10000 inhabitants is the amount usually adopted to define an urban area [23]. The data implemented for a given country is its nationally-defined minimum threshold.

b. Data sources and extraction

Originally, the dataset of Aedes aegypti and albopictus vectors' known occurrences was obtained, which consisted of the number of occurrences, the country, the geographical parameters (longitude and latitude) worldwide, and containing 42067 cases of occurrences in total [1]. Starting from there, it was obtained the data regarding to factors that could be useful to the overall analysis of the problem. Weather factors related were looked for and the monthly temperature average [2] and the monthly precipitation [2] were obtained.

Moreover, it was appended the main data information regarding to the population size [4], population density [7], education level [6], urbanization/ rural rate[5] and human development index [6]. By including data from a variety of sources, it was concluded that the created dataset had enough information to a further analysis of outbreak of occurrences.

c. Preprocessing the data

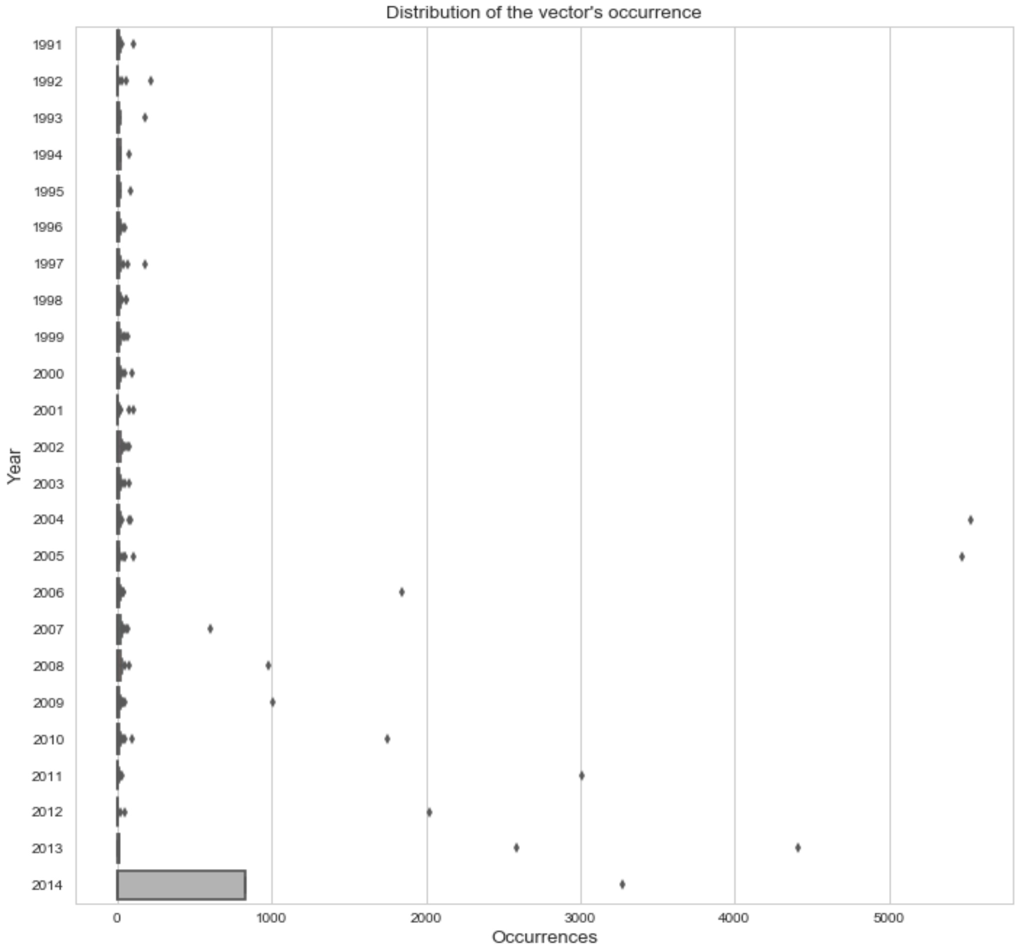
For the purpose of the project, it was necessary to transform the data into one dataframe containing all the desired variables. In order to do this, it was performed preprocessing of several original datasets. Some sources had *csv* extension and others excel extensions. The preprocessing included the following steps:

1. By examining the data regarding the occurrences of the vectors (main dataframe), it was possible to conclude that the key to all datasets was either the country name or country name.
2. One table containing the country names (Country Code) and country codes was found in order to make the aggregations and merges easier.
3. Each dataset required to be normalized in terms of adjusting some non-matching rows. In most of the cases, the table Country Code was used to match the name of the countries with its code (abbreviation), however, due to variety of sources we had to adjust some names of the countries manually by using ***pandas rename method.***
4. Based on assumptions regarding the relevance of certain information for further analysis, some columns were eliminated due to unnecessary information from each of the initial datasets. For this purposes, it was used ***pandas drop method.***
5. After reshaping some data frames in order to match them with the main dataset and, after that, it was possible to merge the multiple datasets into the main dataset, by using *pandas merge method* with the country code mostly as the key label.

d. Exploring the data and modelling

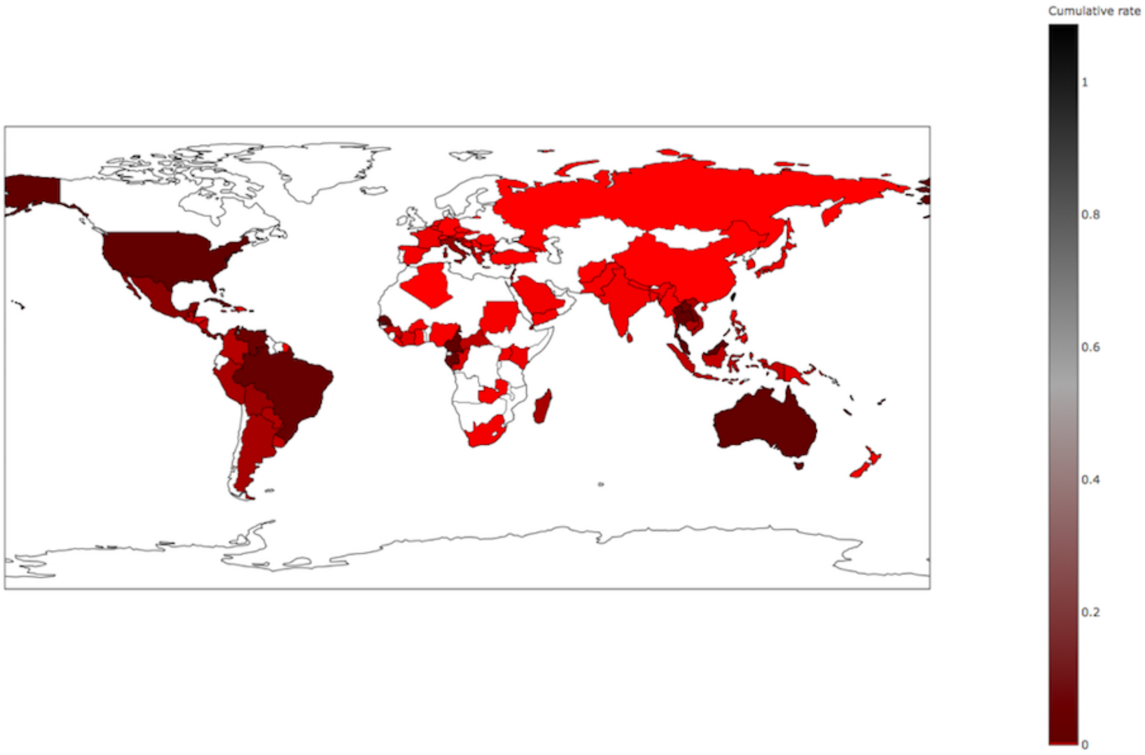
In the graph below, the boxplot of distribution of yearly occurrences is presented. There are some extreme values starting from year 2004. These values belong to observations of occurrences in Taiwan. It was considered more appropriate to eliminate this data from our analysis since it would be more valuable to study this particular situation in isolation.

Figure 1. Distribution of the vector’s occurrence



In order to represent the geographical spread of the mosquito’s population a map of occurrences of vectors was created using python packages ***folium*** and ***plotly.*** In order to highlight the riskiest areas around the globe the number of occurrences was divided by the population size.

Figure 2. Heatmap of the number of vectors per person between 1991-2014



For this research, self-organizing map (SOM) method of clustering was implemented. This method is used primarily for better representation of multidimensional data and finding meaningful relationships in this data. SOM [18], introduced by Prof. Teuvo Kohonen, is an artificial neural network based technique that generates a two-dimensional representation of the data distribution. The output of the method is a grid where each point in original dataset has its position within the grid and located near points with similar characteristics. The order on the grid reflects the neighborhood within the data, such that data distribution features can be read directly from the emerging landscape on the grid.

In this particular case the goal of SOM is to identify clusters of points that preceded the outbreaks of vector occurrences around the globe, in other words, it is being investigated if there some particularities that data points share in the point of time and which one year later caused a sharp increase of the population of mosquitos in a particular area.

The input variables included original attributes, namely : average yearly temperature, average yearly precipitation level, population density, population size, education level, urbanization level and human development index. Apart from mentioned variables, absolute differences, percentage changes and moving average of the same variables were included in the model, since it is supposed that the fluctuation of the all the other variables can be considered an additional factor affecting the target variable.

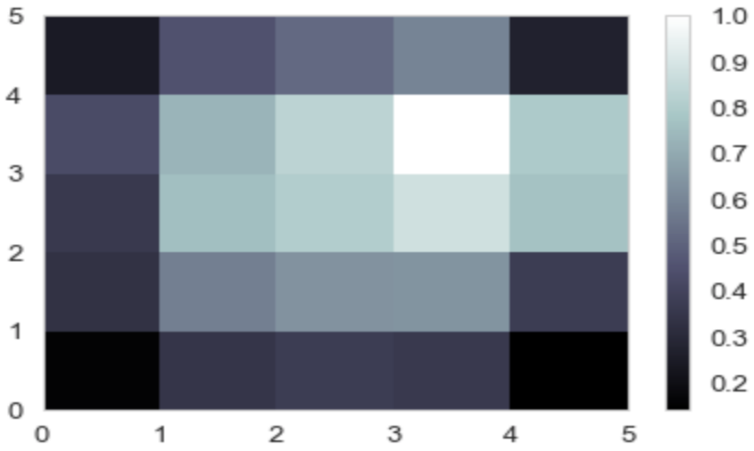
Absolute differences were calculated by using ***numpy diff method ,*** new column was created for the outputs. Growth rates were calculated by using ***pandas pct_change method*** to find out the percentage change, and moving averages with interval equal to one year were calculated by using ***pandas rolling.mean method.***

Afterwards, in order to implement SOM method, all the necessary data was merged into one table and normalized using min-max normalization technique using the function that was created from the scratch**.** The final dataset contained 24 variables and 539 rows . Each row in of the dataset contains the information about the year, the country, the number of occurrences, and other variables that supposedly have an affect on the number of mosquito population. Using this data, the training of SOM was performed, with the following parameters: learning rate equals 0.5, sigma equal 1, and grid size 5*5, meaning in total 25 nodes, number of iterations = 100. For the purpose of the project, in-built Python library MiniSOM was used. It is important to notice here that since the code used to implement SOM is a built-in library, it was not possible to fix the randomness of initial weights of the artificial neural network and, therefore, here it is presented the results of a specific run, which will be different from the subsequent runs of the same algorithm.

IV. Results and Discussion

The following representation (Figure 3) is the result of the implementation of a SOM. In MiniSOM, each region of the map is different in terms of color based on the density of the points which each of the regions contain.

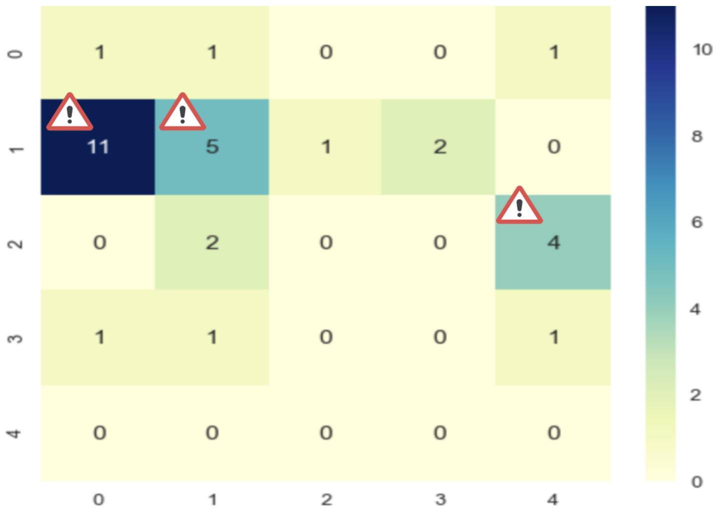
Figure 3. The resulting SOM



In order to investigate which factors provoke the unusual increase of occurrences in a given year, density map was created with data points containing information regarding the year that preceded the "outbreaks". It is important to notice here that for the purpose of the present research, the data point is considered an "outbreak" when the number of occurrences per person in this particular year exceeds the average number of occurrences per person worldwide compared to previous year.

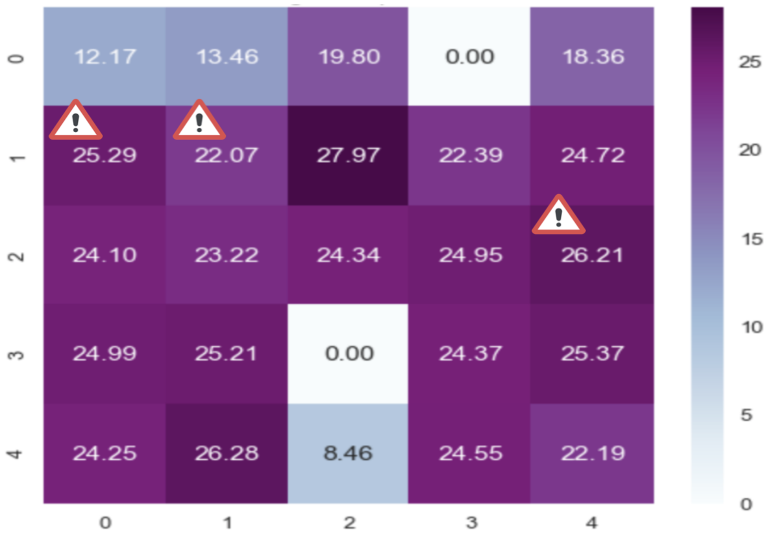
It can be noticed in the heatmap presented in Figure 4 that there are regions that are more densely populated. Therefore, it was inferred that points in each particular region might share factors that provoked the "outbreaks" of the number of vectors in the region.

Figure 4. Points that describe the years that preceded the "outbreak"



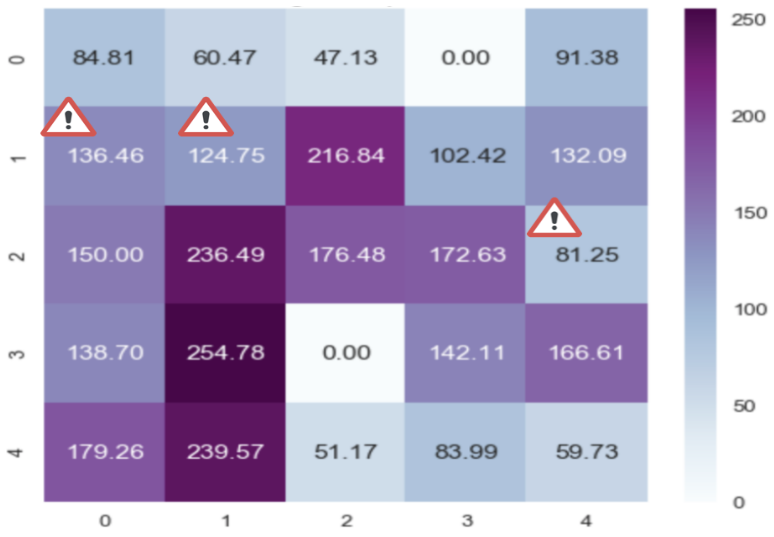
The most dense regions are the region with coordinates (1,0) , (1,1), and (3,4). The average temperature of all the regions is presented on Figure 5:

Figure 5. Average temperature of each region



It can noticed the temperature in the "risky" region is slightly different from the worldwide average (23.01 degrees). In can be inferred that in the short run, temperature on its own does not provoke the growth of mosquito population, since regions with extreme temperatures are not considered "risky".

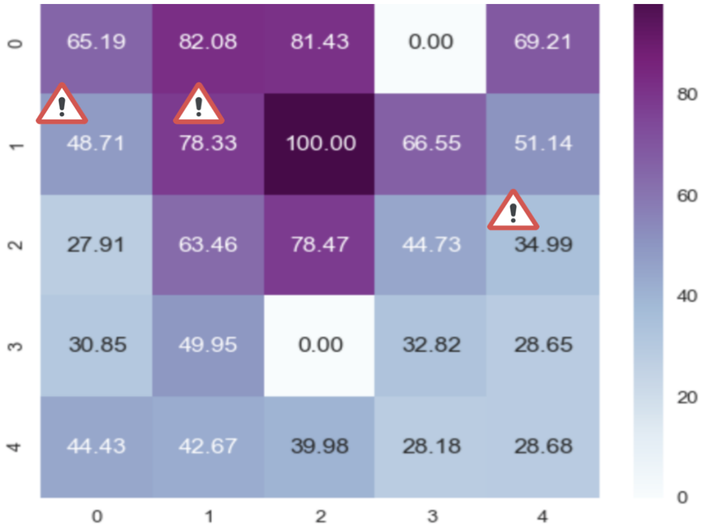
Figure 6. Average precipitation of each region



	0	1	2	3	4
0	84.81	60.47	47.13	0.00	91.38
1	⚠ 136.46	⚠ 124.75	216.84	102.42	132.09
2	150.00	236.49	176.48	172.63	⚠ 81.25
3	138.70	254.78	0.00	142.11	166.61
4	179.26	239.57	51.17	83.99	59.73

The precipitation in the “risky” region (Figure 6) is slightly different form the worldwide average (120 mm) as well. Although, it is known that precipitation level might affect the population of mosquitoes in the short run precipitation only by itself does not provoke the growth of mosquito population.

Figure 7. Average urbanization of each region



	0	1	2	3	4
0	65.19	82.08	81.43	0.00	69.21
1	⚠ 48.71	⚠ 78.33	100.00	66.55	51.14
2	27.91	63.46	78.47	44.73	⚠ 34.99
3	30.85	49.95	0.00	32.82	28.65
4	44.43	42.67	39.98	28.18	28.68

Analyzing Figure 7, area (1,1) has urbanization level (78%) significantly higher than the worldwide average (58%) and area (2,4) has urbanization level (35%) significantly lower than the worldwide average. The results of the

SOM in this case are inconsistent and, therefore, It can be concluded that the impact of urbanization on the vectors expansion needs to be examined more thoroughly.

In the Figure 8 and Figure 9, the average HDI and the average Education index of each region is presented. Worldwide averages are the following: Average worldwide HDI is 0.68 and Average Education index is 0.59. The area (2,4) shares lower levels of the above mentioned metrics. However, area (0,1) and area (1,1) have results closer to the average level.

Figure 8. Average HDI of each region

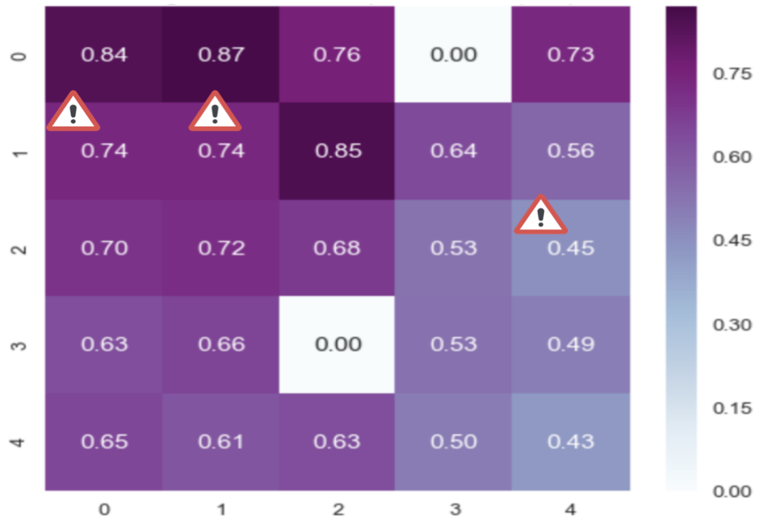
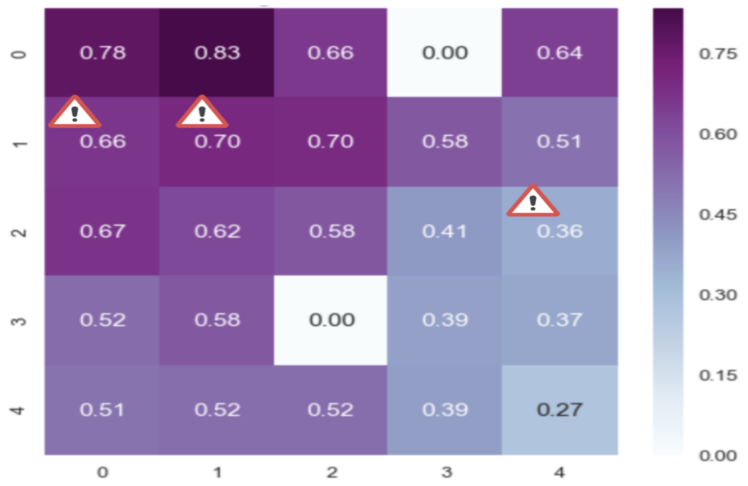
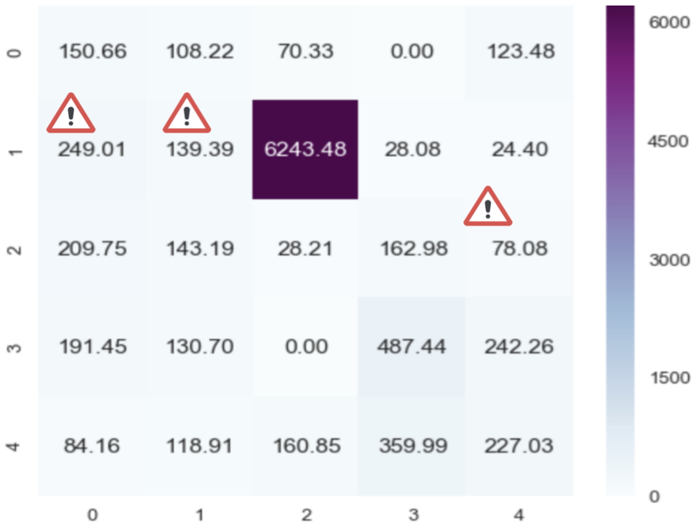


Figure 9. Average Education level of each region



In Figure 10, the average Population density of each region is presented. Worldwide averages is 223. Area (2,4) shares lower levels of the above mentioned metric. However, area (0,1) and area (1,1) have results closer to the average level. Since the results are inconsistent it can be concluded that population density is a factor in isolation can not determine the changes in vector's population.

Figure 10. Average population density of each region



	0	1	2	3	4
0	150.66	108.22	70.33	0.00	123.48
1	⚠ 249.01	⚠ 139.39	6243.48	28.08	24.40
2	209.75	143.19	28.21	162.98	⚠ 78.08
3	191.45	130.70	0.00	487.44	242.26
4	84.16	118.91	160.85	359.99	227.03

Based on the result of the SOM, it can be inferred that no factor in isolation is critical for the development of the "outbreak" in the population. However, a combination of several factors can be considered crucial for an increase in the number of occurrences. For instance, if we consider area (2,4) it can be inferred that low level of education, overall low level of development (social and economic) , low level of urbanization, low population density, temperatures above the global average can provoke the outbreak of the number of mosquitoes.

Basically, there are two dangerous regions in the map (area (0,1) and area (1,1)) that share the following characteristics: levels of precipitation and temperature are similar to worldwide averages, levels of education and overall economic development are significantly higher than worldwide average and population density is close to worldwide average.

V. Conclusion

After the exploration of the collected data, it is clear that the occurrences located in subtropical and tropical areas are more than in other areas as the climate factor is taken into account. *A priori,* based on studies in this field of knowledge [20] [21], it was expected that there is a close relationship between climate factors and the occurrences of the vectors. For instance, high precipitation levels were expected to be a crucial in provoking the "outbreaks". However, such a relationship was not clearly identified.

Moreover, initially high urbanization level in countries with low economic development were considered as "risky zone", however, the results do not support such a relation. In fact, low level of education, overall low level of development (social and economic), low level of urbanization, low population density, temperatures above the global average can provoke the "outbreak" of the number of mosquitos. Region of high economic level, high temperatures, average precipitation level and low urbanization also can be potentially "dangerous" zones.

VI. Acknowledgements

We would like to express our gratitude to professor Jorge Antunes for his great support and guidance throughout the project.

VII. References

1. "The Global Compendium of Aedes Aegypti and Ae. Albopictus Occurrence." Dryad. https://datadryad.org/resource/doi:10.5061/dryad.47v3c.2?fbclid=IwAR1vaQ6nvXiJv8s0hruMEzrNSdYVb_rPAS2ZeHoKYt35vHihv7J75VjaFTk. Ae. aegypti and Ae. albopictus occurrences
2. "Climate Change Knowledge Portal." Climate Change Knowledge Portal. http://sdwebx.worldbank.org/climateportal/index.cfm?page=downscaled_data_download&menu=historical. Temperature and Rainfall data
3. "Disease Outbreaks by Year." World Health Organization. http://www.who.int/csr/don/archive/year/en/?fbclid=IwAR0L_T2HvAuUALKIqaTZBFI55wLK1G-S4NyN_0jFgD1cZY30zY_IseuKEUI. Disease outbreaks data
4. "Population." The World Bank. https://data.worldbank.org/indicator/SP.POP.TOTL?end=2017&start=1980&year_low_desc=false. Population data
5. "Urbanization." Our World in Data. https://ourworldindata.org/urbanization. Urbanization data
6. "Human Development Reports." Human Development Data (1990-2015) | Human Development Reports. http://hdr.undp.org/en/data?fbclid=IwAR0mDXvSTnfp_0YruHvlxZVPkd8fUu-gXoclt4L59Ud30hpKjH5_Z141BGY. HDI data.
7. "World Population Prospects - Population Division." United Nations. Accessed December 06, 2018. https://population.un.org/wpp/Download/Standard/Population/.
8. "R.O.C - Composite Index and Related Indicators." https://eng.stat.gov.tw/ct.asp?xItem=25280&ctNode=6032&mp=5. Taiwan HDI Data
9. "Taiwan National Infectious Disease Statistics System." Dengue Fever, Nationalwide. https://nidss.cdc.gov.tw/en/SingleDisease.aspx?dc=1&dt=2&disease=061. Taiwan Dengue data
10. "Vector-Borne Diseases: Understanding the Environmental, Human Health, and Ecological Connections, Workshop Summary." NCBI. https://www.ncbi.nlm.nih.gov/books/NBK52939/.
11. "Vector-borne Diseases." EFSA. https://www.efsa.europa.eu/en/topics/topic/vector-borne-diseases.
12. "Dengue and Severe Dengue." World Health Organization. http://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue.
13. "Infection Rates by Dengue Virus in Mosquitoes and the Influence of Temperature May Be Related to Different Endemicity Patterns in Three Colombian Cities." NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4962275/
14. "Dengue Transmission." Scitable. https://www.nature.com/scitable/topicpage/dengue-transmission-22399758.
15. “Zika, chikungunya and dengue: the causes and threats of new and re-emerging arboviral diseases”. https://gh.bmj.com/content/3/Suppl_1/e000530

16. “Vector-borne diseases”. https://ecdc.europa.eu/en/climate-change/climate-change-europe/vector-borne-diseases
17. "Human Development Reports." Human Development Data (1990-2015) | Human Development Reports. http://hdr.undp.org/en/content/human-development-index-hdi?fbclid=IwAR0tjaBYm2fi1hX_evEVQa8ZJmzvhWyO_Dxc8utYHKEaO6r7gZCq0FcGzI0.
18. Kohonen Teuvo, Teuvo. *Self-Organizing Maps*. Springer Science & Business Media, 2001.
19. "Epidemiology of Dengue: Past, Present and Future Prospects." NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3753061/.
20. "The Global Distribution of the Arbovirus Vectors Aedes Aegypti and Ae. Albopictus." NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4493616/.
21. "Zika, Chikungunya and Dengue: The Causes and Threats of New and Re-emerging Arboviral Diseases." NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5759716/.
22. "Climate Change Influences on Global Distributions of Dengue and Chikungunya Virus Vectors." NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4342968/.
23. “Urban area”. National Geographic. https://www.nationalgeographic.org/encyclopedia/urban-area/
24. World Health Organization. https://www.who.int