Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Model for the characterization of epidemies generated by the vector Aedes aegyypti. Case study: Norte de Santander Department, Colombia

To cite this article: M C Bernal *et al* 2018 *J. Phys.: Conf. Ser.* **1126** 012073

Model for the characterization of epidemies generated by the vector Aedes aegyypti. Case study: Norte de Santander Department, Colombia

M C Bernal[1,3], M E Sotelo[2], D C Nimo[3], M M Molina[3], O G Pérez[4], A A Sánchez[3] and J G Toro[3]

[1] Grupo de investigación IngeBioCaribe, Universidad Simón Bolívar, San José de Cúcuta, Colombia

[2] Grupo de investigación ALEF, Universidad Simón Bolívar, San José de Cúcuta, Colombia

[3] Grupo de investigación LIDI, Universidad Nacional Experimental del Táchira, San Cristóbal, Venezuela

[4] Universidad de Pamplona, San José de Cúcuta, Colombia

E-mail: m.bernal@unisimonbolivar.edu.co, m.sotelo@unisimonbolivar.edu.co

Abstract. Applied technologies to health systems have been noticed as a required tool to reduce the reappearance rate of certain illnesses, to identify groups of risk or to involve the patient in the care of his own health. Taking in consideration the fact that in the North Santander Department the occurrence of the outbreak of illnesses transmitted by the vector Aedes aegypti shows intense increase since its re-emergency in 2014, with the presence of several manifestations, one of the most likely solutions to prevent the illness, is the epidemiologic surveillance alongside the strategic prevention/ control of the vector. In this context, this project was developed with the purpose of characterize patients by the building of a knowledge database for digital health that allows to create an architecture to integrate, analyze and structure the historical data according to the detected cases, in order to collect the information from the direct field, store it and process it, to present reports and knowledge models with the right indicators that allow to look out often the symptoms and to control the increase of such epidemies.

1. Introduction

Information and communication technologies applied to health systems are being shown as a necessary tool to reduce the rate of reinsertion in certain diseases, to identify groups of risk or to involve the patient in the care of their own health [1].

The medical area has benefited thanks to the incorporation of technologies in the analysis and monitoring of patients. Within and outside the country, many organizations and institutions decide to rely on the objectivity of the data and are guided by the visibility provided by data processing to improve the patient's experience and to plan the health services, in order to achieve proper control and growth.

In the specific case of the diseases transmitted by the Aedes aegypti vector, their importance lies in the fact that they constitute an important public health problem throughout the world, mainly in the tropical and subtropical regions. Likewise, the vertiginous growth of registered cases and complications derived from these diseases means that the response mechanisms from the health sector need timely and adequate support to guarantee correct decision-making.

This article constitutes the creation of a platform that allows the analysis of medical case records of diseases such as Dengue and Chikungunya, based on historical information in the Department of Norte de Santander, Colombia, which allows for characterization of models to describe the behavior of the presence and evolution of the virus using data mining techniques in this region for its control and monitoring.

2. Epidemiology

Tropical diseases are infectious diseases that predominate in hot and humid climates. They are considered high risk for the population due to the cost of their treatment and the possibility that they become endemic and epidemic [2]. Dengue is a tropical disease, dengue is transmitted from human to human by the bite of hematophagous females of the genus Aedes (A. aegypti and A. albopictus), while infection by Chikungunya, is produced by the Chikungunya virus (gender Alphavirus, Togaviridae family) [3].

The fastest growing vector disease in the world is dengue, transmitted by the Aedes aegypti vector, whose incidence has multiplied by 30 in the last 50 years. Dengue is a viral infection that, when it infects the patient, causes symptoms like an influenza that can evolve to severe dengue known as dengue hemorrhagic fever; It is found in more than 100 countries, placing at risk a population of more than 2.500 million people in tropical and subtropical regions, with an annual incidence of 50 to 100 million cases.

In America, dengue began to be considered a major public health problem only until a few decades ago. In 1950, a program was launched with the objective of eradicating the Aedes aegypti mosquito, the main vector of the DENV, and although this program, coordinated by the Pan American Health Organization (PAHO) was successful in many countries, the mosquito could not be completely eradicated from the region. At the end of the 1970s, as a result of the decrease in economic support for mosquito surveillance and control, the population of the vector increased again in most countries of the tropical region of the Americas [4].

In urban and peri-urban areas the presentation of these diseases is associated with cultural, social and environmental conditions such as overcrowding, poor hygiene both individual and community, inadequate health services and poverty, which generates greater risk for the population suffering from the worst living conditions, here is the importance of their control [5].

In Colombia, dengue outbreaks are cyclical and are reported mainly in the departments of Norte de Santander, Santander, Huila, Tolima, Valle del Cauca and Antioquia; dengue mortality is avoidable in 98% of cases. Nevertheless, the aspects found to be critical in patient care have frequently been the lack of knowledge of the population, which prevents timely consultation, in addition, failures in the diagnosis and barriers to access to health services.

3. Fundamental concepts

For the development of this paper, concepts of knowledge discovery from databases and data mining are fundamentals.

3.1. Knowledge discovery in databases

The Discovery of Knowledge in Databases is the process of identifying valid, new, potentially useful and understandable information in the data. Data mining goes a step further in the knowledge discovery process and basically consists of algorithms that look for patterns or models in the data. Data mining techniques are not new and were improved thanks to a long process of research and development. This evolution began in the 70s when business data was first stored on a computer, and continued with improvements in access to data, and more recently with technologies generated to allow users to browse through the data in real time [6].

3.2. Data mining

According to Hernández and others [6] data mining is responsible for integrating various data analysis techniques for the description of trends and regularities, prediction of behaviors and, in general, the extraction of useful knowledge from large volumes of information, allowing organizations and companies to understand and modeling in a more efficient way the context in which decisions must be made.

Data Mining is an information extraction activity, objective is to discover hidden facts contained in databases through a combination of machine learning, statistical analysis, modeling techniques and database technology, looking for relationships in the data and deduces rules that allow the prediction of future results [6].

3.3. Statistical analysis of categorical data

In social investigations, data sets that reflect some quality or category are involved. These data are known as categorical data. Such data may contain a mixture of different types of variables, many of which are measured in ordered or disordered categories. Variables such as the seasons of the year, the types of a certain product in the market, or the fact that a student approves or not an exam, are examples of variables with disordered categories. Variables such as the level of education or the frequency with which a certain activity takes place (little, regular or much) are examples of variables with ordered categories. Discrete variables can be considered categorical variables, with each category or quality coinciding with its value [7]. The analysis and processing of this type of data is fundamental for the development of research of this type.

4. Methods

Knowledge Discovery in Databases (KDD) refers to the process of discovering useful knowledge in data, where with the application of specific algorithms of data mining data patterns are extracted [8].

The steps used for the development of the research were: selection of the objective, preprocessing of data, transformation, mining of data and interpretation of the results.

Figure 1 shows the stages of the process adapted from Orallo y Ferri [6], with the products obtained as a result of the application of the necessary techniques to analyze the case study.

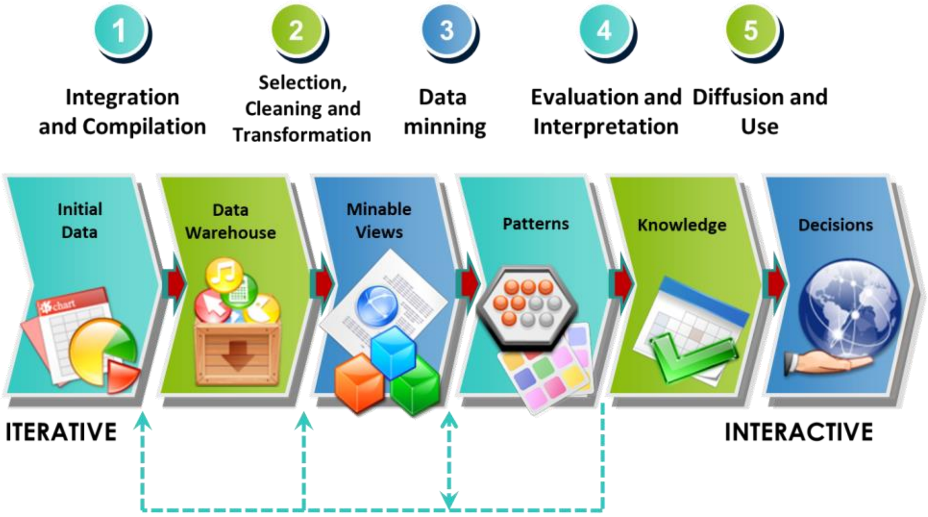

Figure 1. Approach used for the process of Knowledge Discovery in Database (KDD).

4.1. Understanding the business

This first phase focused on understanding the business objectives and requirements from the perspective of the study. The objective of the selection of the objective was to study the problem and decide on the objective of the project. Once the problem was defined, the internal or external data sources were identified and the subset of data necessary for the application of the techniques was selected.

4.2. Understanding the data

The data comprehension phase began with the initial data collection and continued with the activities aimed at familiarizing the data, identifying quality problems and detecting interesting relationships between them that allowed generating characteristics about hidden information, following these steps:

- Extraction of the Data: At this point a study of the data available in the database was done, the structure of the available tables was reviewed, and a representative sample was extracted from it for subsequent studies.
- Filtering of the Data: All the data extracted in the previous point was taken to proceed to eliminate invalid, incorrect, empty and unknown data, additionally the possible values for each available variable were analyzed, in order to be adjusted for use in the later studies.

According to Orallo and Ferri, functional architecture was defined [6]. The functional architecture described in Figure 2, establishes the mechanism for accessing data from its primary collection sources, which facilitates its integration and debugging, making it ready for review and analysis. This architecture allows, through the generation of cubes, to generate timely reports and multiple views for different types of users that allow adequate control and monitoring.

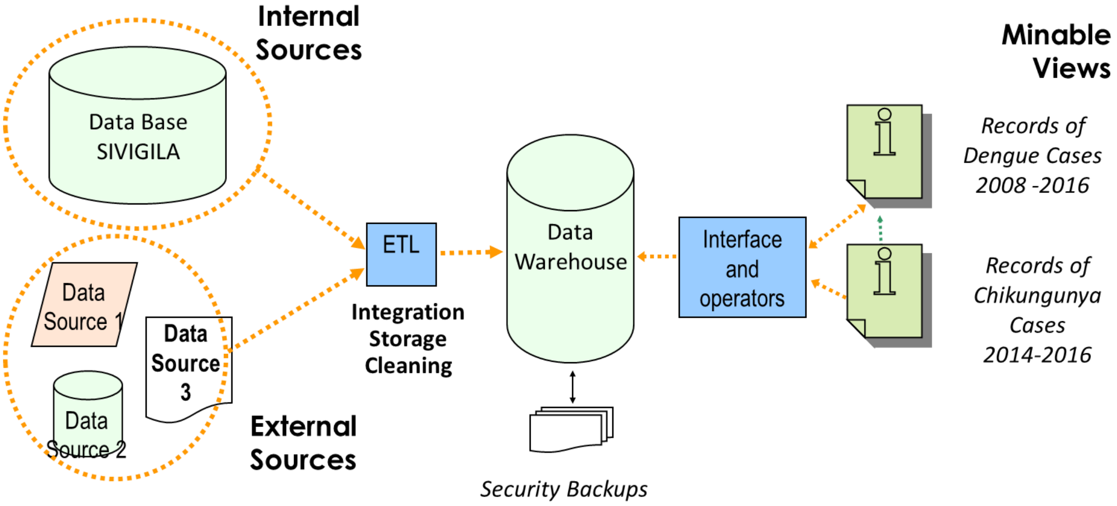

Figure 2. Functional architecture of data integration.

4.3. Preparation of data

This phase contemplates a set of activities focused at the construction of the dataset based on the initial data. During this phase, multiple tasks were developed that included the selection, cleaning and transformation of tables, registers and attributes so that they could be entered into the modeling tool for characterization. For this purpose, the steps are:

- Selection of study variables: After filtering the available data, the variables that best contributed to the description of the study area and those that had the necessary quality criteria for their inclusion were selected. Additionally, new variables not directly available from the initial data were created, these new variables were generated from calculations made from the variables that were available.

Because it is a descriptive study of epidemiological variables, it was analyzed the health problems according to the variables of people, place and time; with the purpose of inferring causality. The variables used for the study are shown in Table 1: of people (age and sex) of time (date of diagnosis and development of the disease), of place (area where he lives) and from diagnosis (ones that have to do specifically with the symptoms presented by the patient described through a set of presence or absence of conditions).

Table 1. Selected variables.

Variable Types			
Person	Time	Place	Diagnosis
Age, gender	date of attendance, date of occurrence of symptoms, days with symptoms.	Department, municipality, locations.	Fever, headache, myalgia, arthralgia, rash, abdominal pain, vomiting, diarrhea, drowsiness, hypotension, hypothermia, platelet fall, fluid retention, bruising.

- Search of exploratory models: The exploratory models were generated in order to purify the selected variables and identify their behavior. This activity was developed as part of the modeling phase.

4.4. Modeling

In this phase, modeling techniques were selected and applied, as well as the values of the parameters and calibration variables were determined. For this task various techniques were used that perform the same function. Some techniques had specific requirements regarding the conformation of the data, which made us return to the preparation phase of the data to make adjustments. This phase contributed to the generation of descriptive models that characterize the presence of the disease according to the behavior of the cases studied. Association rules were extracted as a descriptive model using the a priori algorithm which, once all the nominal variables were configured, obtained the best quality evaluation.

The obtained rules contain a left side with conditionals that must comply, and a right side with the consequences of fulfilling these conditions. They are also using a covering procedure. However, on the right side of the rules, the appearance of any pair or pairs attribute-value was contemplated, for which each possible combination of attribute-value pairs on the right side was considered, later they were pruned using coverage (number of instances predicted correctly) and precision (proportion of number of instances to which the rule applies) [9].

In terms of probabilities, support and confidence are given by the following formulas:

$$support(A \Longrightarrow B) = P(A \cup B) \tag{1}$$

$$confidence\ (A \Rightarrow B) = P(B \,|\, A) = (support(A \cup B))/(support(A)) \tag{2}$$

The support formula shown in Equation 1 is given by the union of two probabilities, that is, the support of the rule (A ∪ B) It is equivalent to the probability that A and B are simultaneously fulfilled. On the other hand, the previous confidence formula shown in Equation 2 is given in terms of conditional probabilities, the probability of occurrence B, given that A occurs. In fact, the interest is centered on rules that have a lot of support, so it looks (regardless of which side they appear), attribute-value pairs that cover many instances. These are called item-sets and each pair attribute-value item. With this, what is sought is to find existing associations in the study data that allow to find relevant patterns. Based on the above, the process performed was as follows [9]:

1. Generate all item sets with an element. Use these to generate those of two elements, and so on. All possible pairs that comply with the minimum support measures are taken. This allows eliminating possible combinations since not all have to be considered.
2. Generate the rules checking that they meet the minimum criterion of confidence.

Figure 3 shows the rules tree; the first number refers to the nominal variable under study and the one in parentheses the times it occurs. Once you have the item sets, the rules are generated.

- For each set of items, generate all its subsets.

- For each set $s \sqsubset l$, generates a rule that complies with Equation 3:

$$s \Rightarrow (l - s) \text{ if:}$$

$$soporte(l)/soporte(s) \geq confidence\_level \tag{3}$$

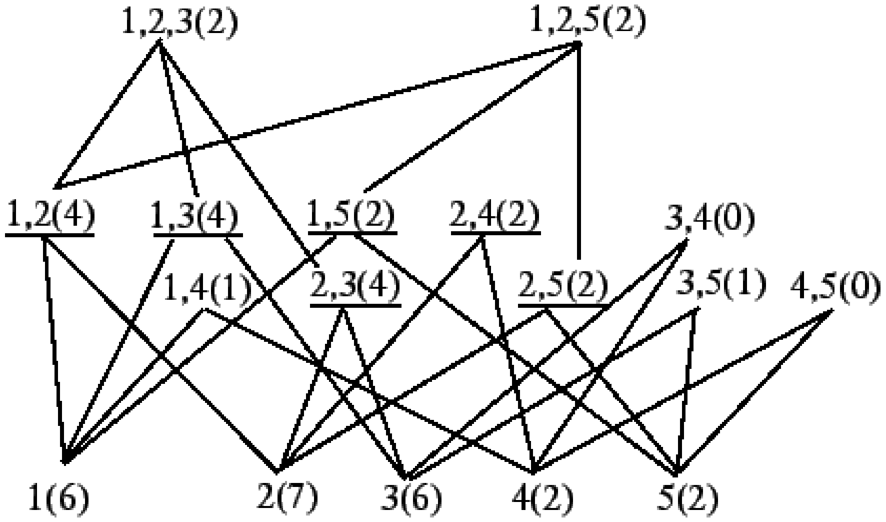

Figure 3. Generation of itemset with a subsample of the study data.

The minimum support (minimum support) is the one that indicates the value established as minimum support for the acceptance of the rule; any rule that is below that support will not be considered.

The confidence metric is the range of error in which the rules to be taken in the model are wanted. This value must be previously established, in this case 90% accuracy was given [10].

4.5. Evaluation

In this phase it was verified that the results obtained in the descriptive modeling phase were valid, novel, useful and understandable from the perspective of data analysis. The modeling results were evaluated, and the steps taken to build the model were reviewed, verifying that these are appropriate in terms of the objectives set and that they also encourage the generation of predictive models in more advanced stages of the project [11].

5. Conclusions

Given the seriousness of epidemiological diseases, it is essential to implement easy methods to reduce their spread, which allows for intelligent monitoring and management. This project establishes an integration architecture that offers an analysis and monitoring mechanism to support the capture, transformation and timely decision making for the intelligent control of vector-borne diseases. The wealth of information transmitted on the vectors of the disease allows us to apply technologies and tools to identify the methods, describe and anticipate their behavior.

Data mining techniques can be used to extract useful patterns. However, it is important to have an appropriate functional architecture that allows quickly and easily, the integration of data from various sources and its transformation for further analysis and exploitation.

Being the Department of Norte de Santander in Colombia an area where there is greater recurrence of these diseases, tools that help monitor and control them will allow to prepare, optimize response times and resources in the proper care and control of epidemics in view to establish routes towards an intelligent health management.

References

[1] Ortiz F, Mendez J F, Ritchie J and Rosado F J 1995 Las organizaciones inteligentes en la toma de decisiones en salud: el caso dengue *Salud Pública de México* **37** 78

[2] Organización Mundial de la Salud 2018 *Organización mundial de la salud* Consulted on: http://www.who.int/topics/tropical_diseases/es/

[3] Nelson M J 1986 *Aedes aegypti: Biology and ecology* (Washington: Panamerican Health Organization)

[4] Castrillón J C, Castaño J C and Urcuqui S 2015 Dengue in Colombia: Ten years of database records *Revista Chilena de Infectologia* **32** 22

[5] Rodríguez A 2015 Chikungunya virus infection: Ecoepidemiological considerations of a new threat for Latin America *One Health Newsletter* **8** 7

[6] Hernández J, Ramírez M and Ferri R 2004 *Introduction to data mining* (España: Pearson)

[7] Hair J, Anderson R, Tatham R and Black W 2007 *Multivariate analysis* (España: Prentice Hall)

[8] Piatetsky-Shapiro G, Frawley W and Matheus C 1991 Knowledge discovery in databases: An overview *AI Magazine* **13** 58

[9] Witten I, Frank E, Hall M and Pal C 2017 *Data mining: Practical machine learning tools and techniques* (United States of America: Morgan kaufmann series in data management systems)

[10] Cabena P, Hadjinian P, Stadler R, Verhees J, Zanasi A 1998 *Discovering data mining: From concept to implementation* (Nueva Jersey: Prentice Hall)

[11] Alpaydin E 2014 *Introduction to machine learning* (United States of America: The MIT Press)

[12] Seewald A and Scuse D 2009 *WEKA manual* (Nueva Zelanda: University of WAIKATO)