Data warehouse and Data mining tools for Epidemic Data analysis to get a better decision-making: The case of leishmaniasis in Morocco

HABIBA MEJHED
Département Génie Informatique, Ecole Nationale des sciences appliquées
Université Cadi Ayyad, BP 575, Boulevard Abdelkarim Khattabi, Guéliz ,
Marrakech, Maroc
mejhed90@gmail.com
SAMIA BOUSSAA
Laboratoire d'Ecologie et Environnement, Faculté des Sciences Semlalia, Université Cadi Ayyad, Marrakech, Morocco & Laboratoire de Parasitologie, Faculté de Pharmacie, Université Louis Pasteur-Strasbourg I, France.
samiaboussaa@yahoo.fr
NOUR EL HOUDA MEJHED
Département Génie Informatique, Ecole Nationale des sciences appliquées
Université Sidi Mohamed Ben Abdallah, Fès, Maroc
noramej@hotmail.com

Abstract: In this paper, we introduced the tools of datawarehouse and datamining to improve epidemiological knowledge of leishmaniasis in Morocco. With the aim to contributes in the planning for the prevention and control against leishmaniasis in Morocco, our investigations were care out in Marrakech area for sandfly collections. We develop an information system to facilitate decision-making, access to information and data storage in a datawarehouse. In addition, we use the concepts of Datamining in order to select, explore and transform our epidemic data to prediction index.

Keywords: Epidemiology, Leishmaniasis, Morocco, Information system, database, Decision aid, Datawarehouse, Datamining.

1 Introduction

In the last decade, the epidemiology used more and more the techniques of help to decision, the aim is to translate epidemic data using the modelling concepts of information systems decision-making. The information is a value-increasing necessary to plan and control the activities of an organism with effectively. It is the raw material that will be transformed by information systems. Often, the availability of data makes it very difficult, if not impossible, to extrapolate the information that really matter. It is essential to have quick and complete information needed for decision-making process: the strategic indicators are extrapolated mainly operational data in a database, through a selection process or synthetic gradually. The widespread use of data analysis techniques has made the information system a strategic element for achieving the business. Therefore, the decision-making systems have emerged in the 80s (decision support system) and offer techniques and means to extract information from a set of memorized data. As a result, the volume of information collected during an epidemiological case study enables the development of new observing systems to analyze and extract some indicators as appropriate medical decision and public health. The decision support provided epidemiologists technologies necessary to facilitate this difficult task [1], [2].

The datawarehouse remains a valuable tool, used for storage and accessibility of the data, it is defined as a collection of information that integrates and reorganize the data from a variety of sources and make them available for analysis and assessment to scheduling and

decision-making. There are three approaches to create a data warehouse and represent information, namely:

- Relational OLAP (ROLAP) : The data is stored in a relational database, OLAP server simulates multidimensional approach.
- Multidimensional OLAP (MOLAP) : Structure storage as a cube with direct access to data in the cube
- Hybrid OLAP (HOLAP) : it is the hybrid approach of the two previous approaches

If the data warehouse used to store historical data, with the finality analysis, the data mining [3, 4, 5, 6, 7] is defined as a process of exploration and modelling data in order to discover a new correlations, trends or patterns in the data. It proposes a number of tools from different disciplines, in particular, to decision makers in epidemiology. Data Mining combines between various sciences domains (Databases, Statistics, Artificial Intelligence) to construct models from the data, and under the criteria fixed in advance and make a maximum of knowledge useful to make decision.

To put it in the context of the problem that we address in this paper, we recall that since 1970s, Moroccan leishmaniasis has been the subject of a Franco-Moroccan programme to study different foci. On results, three species *L. major*, *L. tropica* and *L. infantum* were identified as responsible parasite in these foci [12].

Currently, the leishmaniases is diseases to obligatory declaration and remains a true problem of public health [10]. Tow clinical entities observed, visceral and cutaneous, they are spilled extensively in the whole territory. These are diseases caused by three species with genetic diversity and different epidemiological dynamics. The bibliography is rich in research works carried out in Morocco, especially in south of Morocco, to study the phenomenon vectors bionomics as an initial step in the development of effective strategies for the control of leishmaniasis [8-12]. We note that the situation is becoming increasingly alarming.

This paper is a simple contribution to the fight against leishmaniasis in Morocco. The idea is to propose to epidemiologists an application based on tools of datawarhouse and datamining to help them in the decision. In the first time, we conceived and modelled our information system to establish the pattern of the database on which we are going to work. Then we construct a datawarehouse to store and extrapolate data collected in Marrakech city, the information concerning the three species: *Phlebotomus papatasi* vector of zoonotic cutaneous leishmaniasis, *P. sergenti* vector of anthroponotic cutaneous leishmaniasis and *P. longicuspis* vector of visceral leishmaniasis.

2 Construction of a data warehouse Leishmaniasis

2.1 Background

The creation of a data warehouse involves several steps:

The conception: the implementation of a datawarehouse usually begins by framing the project, define the needs and goals expressed by policymakers, and modelling and designing a data structure. There are two data models, the star pattern, in this model, we must define one (or more) table (s) made with one or several measures (values of indicators). Both must have multiple dimension tables whose primary keys form the primary key tables done. Warning: The dimension tables are not linked.

Then the model snowflake which is derived from the star schema where the tables are standard size (of the table remains unchanged). With this scheme, each dimension is divided according to his (or her) hierarchy (s).

- ***The acquisition of data:*** The data will be extracted from the sources.
  - The static extraction will be performed when the DW must be loaded for the first time and is conceptually a copy of operational data.
  - The incremental extraction, is used for the periodic updating of the DW, and captures just the changes in data sources at the last extraction.

The choice of extracting data is based mainly on their quality, selection of data from the database is not a simple task to do.

Data cleaning: This phase will improve the quality of duplicate data, inconsistencies between the values logically related, missing data, unexpected use of a field, impossible value or wrong ...

Loading DW: The loading of data in the DW is the process is to load the data cleaned and prepared in the DW.

2.2 Description of our database

In this section we proceed to the conception of the scheme of our data. We have a data file containing all the information on the activity of sandflies P. papatasi, Siminula, P. Sergentine, S.fallax, P. longicuspis based on climate change (date, temperature, density). The goal is to design a database for this case, the conception DataWarehouse and loading the database, then build a cube specific answer to the question of the case study.

2.2.1 Construction of the model:

The tables of the database Leishmaniose are as follows:

Tables of dimensions:

- Date_dim
- Psergenti
- Temperature_dim
- Siminula
- S.fallax
- P.longicuspis
- P.paptasi

Table of facts:

- Fait

The model chosen must comply with the requirements and needs of the use, in our case, we opted for a star pattern in the nature of the information we have. The next capture represents the logical schema of the database:

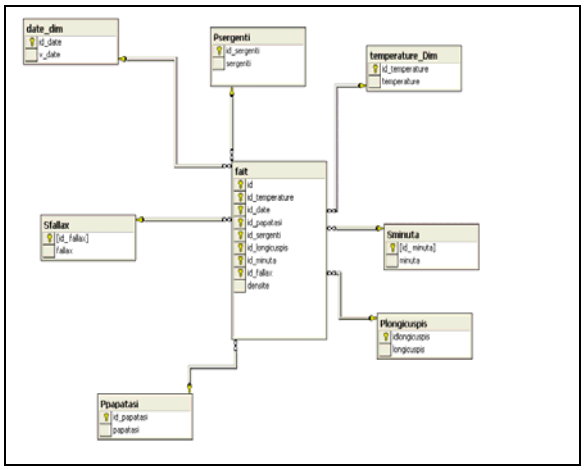
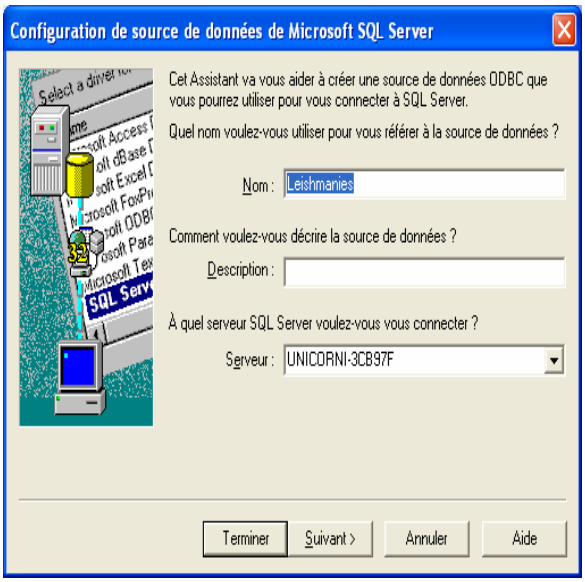


Figure 1: The star model data of Leishmaniosis

2.2.2 Definition of connection to a data source system

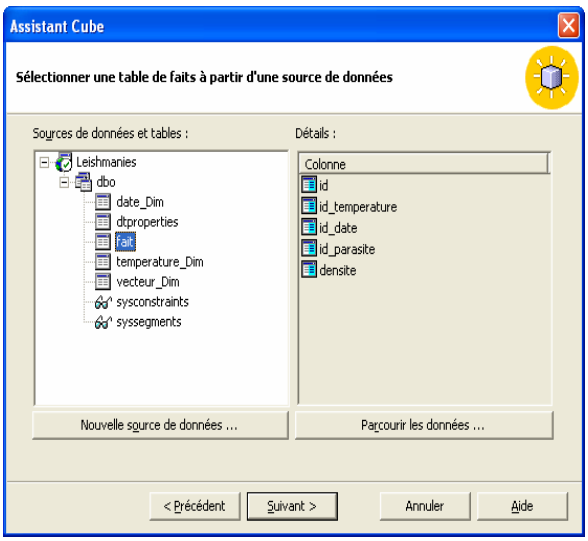
A data source contains the information needed to access the data source object. After that the database has been registered with ODBC to simplify access and avoid connections management system databases used.



2.2.3 Establishment of the cube to extract information

Modeling data structure facilitates multidimensional data analysis and query performance. They exist the tools to transform data stored in relational databases of information relevant and easy to operate, through the creation of a data cube. A cube is a multidimensional data structure. The cubes are defined by a set of dimensions and measures.

We proceed to the creation of the cube wizard mode, you must choose the fact table from the data source:

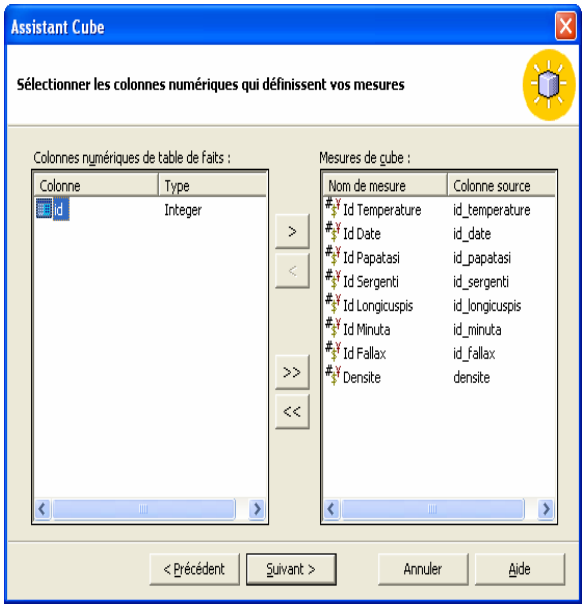


Then, we can browse the data:

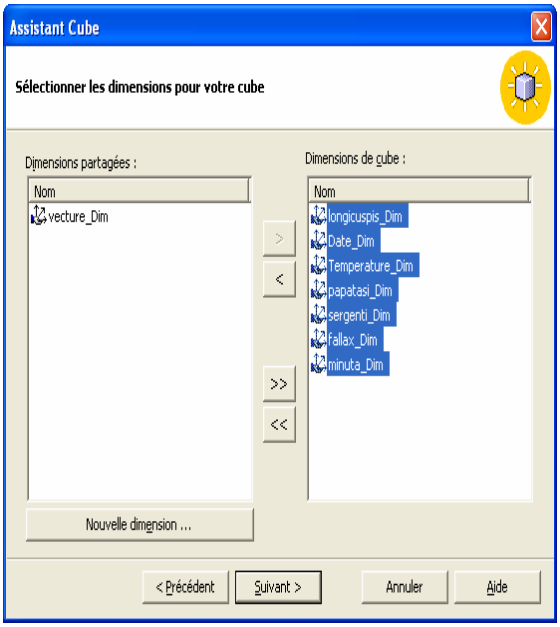
Parcourir les données : "dbo"."fait" (1000 premières lignes)

	id	id_temperature	id_date	id_papatasi	id_sergenti	id_longicuspis	id_minuta
1	78	1	1	1	1	1	1
2	79	2	2	2	2	2	2
3	80	3	3	3	3	3	3
4	81	4	4	4	4	4	4
5	82	5	5	5	5	5	5
6	83	6	6	6	6	6	6
7	84	7	7	7	7	7	7
8	85	8	8	8	8	8	8
9	86	9	9	9	9	9	9
10	87	10	10	10	10	10	10
11	88	11	11	11	11	11	11
12	89	12	12	12	12	12	12
13	90	13	13	13	13	13	13
14	91	14	14	14	14	14	14
15	92	15	15	15	15	15	15
16	93	16	16	16	16	16	16
17	94	17	17	17	17	17	17
18	95	18	18	18	18	18	18
19	96	19	19	19	19	19	19
20	97	20	20	20	20	20	20
21	98	21	21	21	21	21	21

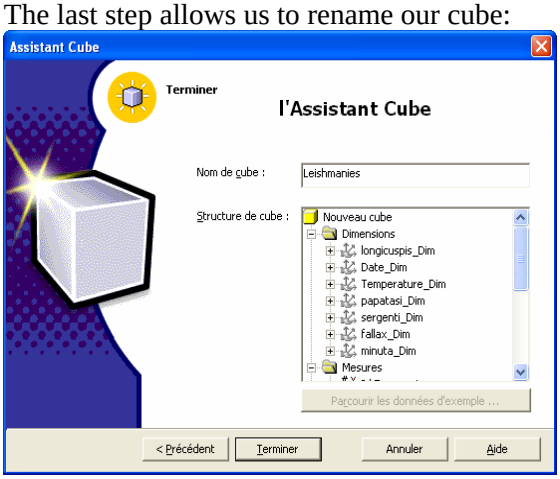
After, we select the columns that define our measures:



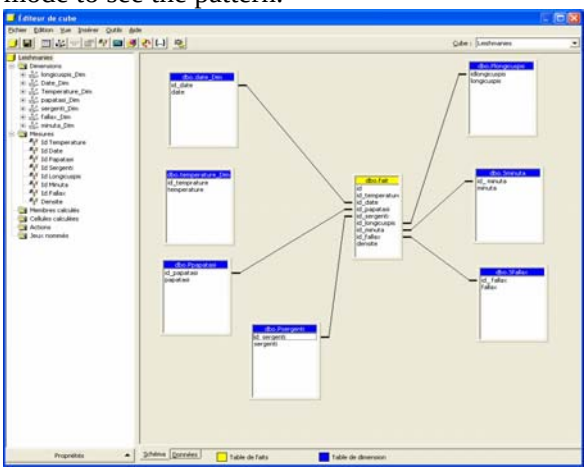
Now, we must move to the creation of the size, we select the size of our Leishmaniasis cube as is illustrated by the following figure:



The last step allows us to rename our cube:



Once created, the cube can be opened in edit mode to see the pattern.



2.3 Storage and treatment of a cube

We can specify the storage options for data and aggregations of our cube. Before using the data from our cubes, we must treat them. We have three modes of storage MOLAP, ROLAP and HOLAP

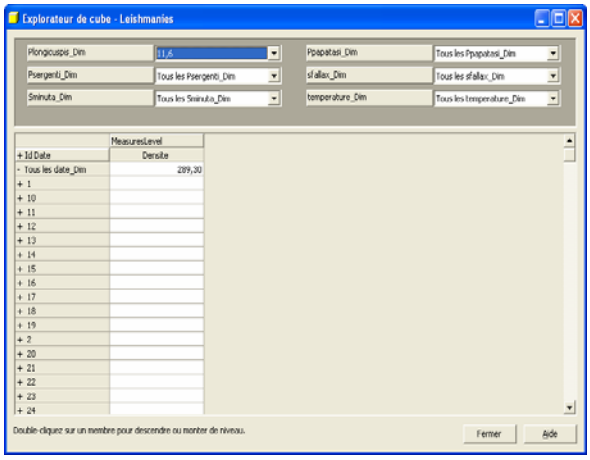
When we treat a cube, aggregations created for it are calculated and the cube is loaded with the calculated aggregations and data.

Now that we've created the structure of the Leishmaniasis cube, we must choose the means of storing it and specify the amount of the precalculeted values to store them. Once you've done this, we must make data in the cube. In our case, we selected MOLAP as a storage mode, we created the concept of aggregation (performance) for the Leishmaniasis cube and we made the processing of the cube. The treatment of the cube load data from the ODBC source and calculates the partial values as defined in the conception of the aggregation.

2.4 Exploration data leishmaniasis cube

Using the Explorer cube, we can view the data in different means and according to criteria fixed from the beginning, this allows as to filter the amount of data visible dimension, extract down to view more details or to extract up to display less.

The Leishmaniasis cube is treated, which is to provide political analysis and aggregation and retrieve data, we can then browse the data. As it's shown in the example of the capture follows:



3 Data mining: application to the Leishmaniasis

Given the seriousness of leishmaniasis in Morocco, it was essential to deploy easy and exploitable ways to reduce its spread if not eradicate it completely. Our proposal aims to support and benefit studies on leishmaniasis of of technological development tools. The datamining, by definition, is a technic for extracting knowledge from a large volume of data.

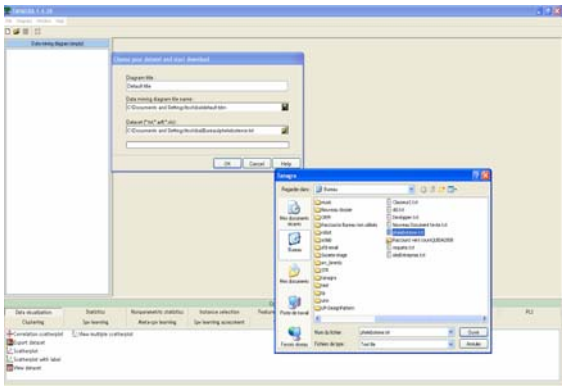
The wealth of information on transmitted vectors of disease allows us to apply these tools to identify methods and anticipate behavior, therefore, make better decisions.

3.1 K-Means and deployment

The technologies that are on the market, we offer complete platforms and integrated data analysis to meet all requests of indicators developed in the industry. We are able to have any type of data stored in our database, to implement operations to analyze the data and present results in a need predefined by the user.

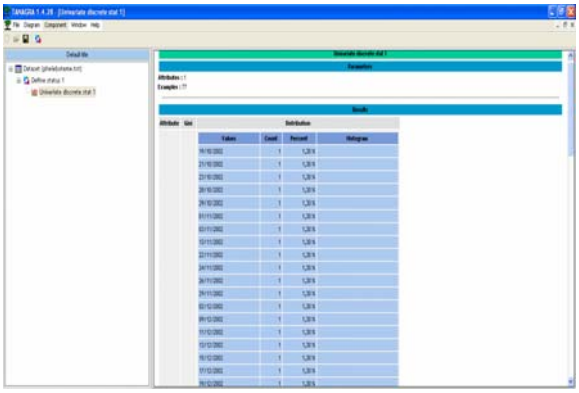
The software that we use to develop our application offers a wide range of approaches ranging from methods of descriptive statistical analysis to predictive modeling methods.

The first step is to create a new diagram and import the data as shown in the screenshot below.



3.2 Descriptive statistics

We can do descriptive statistics to variables. We calculate the frequency histograms on all columns to count the number of active and additional comments.



3.3 Method principal component analysis (ACP)

They are three families datamining algorithms: supervised methods, unsupervised methods and methods of data reduction. Each category is based on a number of techniques.

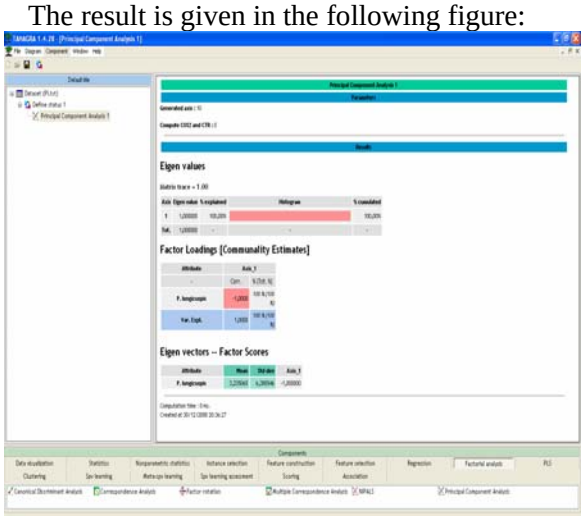
In this section, we chose the third type using the method of principal component analysis (ACP).

Given a set of observations described by variables exclusively digital (x1, x2, ..., xp), the APC aims to describe the same data set with new variables in reduced numbers. These new variables will be linear combinations of original variables. Principal component analysis can therefore be seen as a technique to reduce dimensionality.

3.3.1 Visualization of our data

To implement the ACP method, we can see, for example, date and temperature data concerning the kind P.longicuspis. After we define an analysis of the variables studied.

The result is given in the following figure:



To better assess the positions relatives to sandflies in the first factorial model, we add the component display. We put abscissa variable representing the first axis, calculated using the ACP, and ordered the second axis. We get the cloud of point after

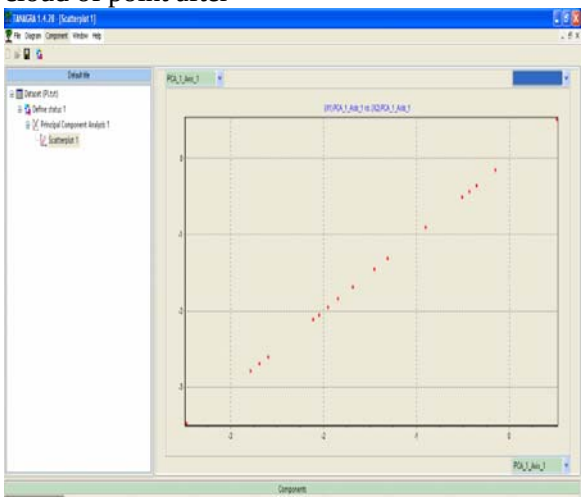


Figure 2: Cloud data on (date, temperature, P. longicuspis)

Among the variables are, we want to check the effect of the date variable that can distort our results. We will color the points according to this variable, we select as a variable component in this illustrative visualization

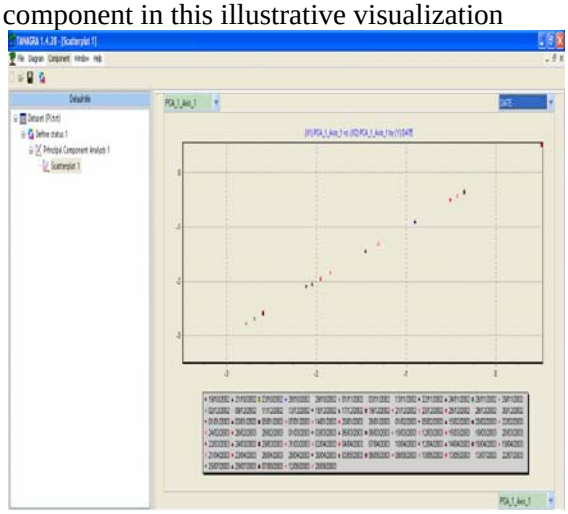


Figure 3: Using the date_format as illustrative variable

Now we apply the ACP method on all data:

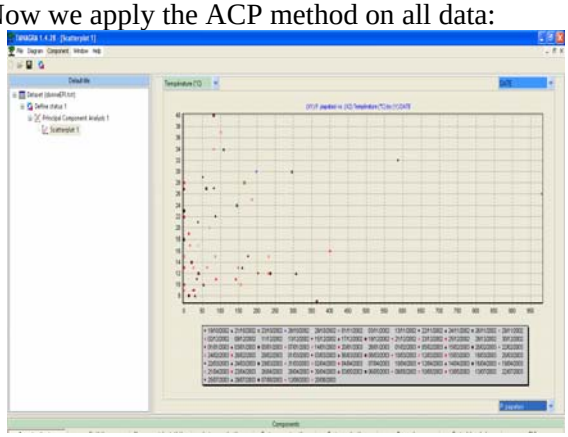
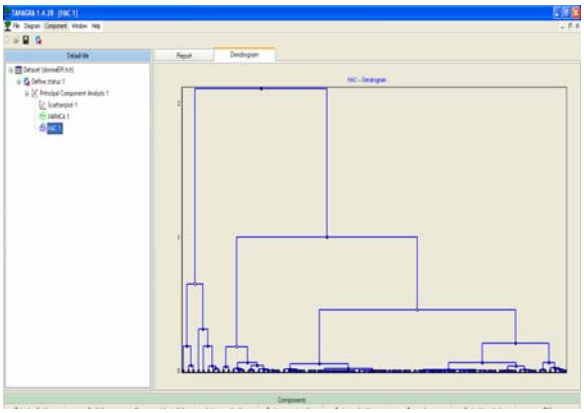


Figure 4: Implementation of the ACP method on all sandflies data



4 Conclusion

In conclusion, we proposed a model of data warehouse and application of data mining methods to meet the needs of knowledge in the epidemiology considering the case of

Leishmania in Morocco. This will help ensure the advanced treatment and to generate summaries for better health decision-making.

References :

[1] Informatique et santé Collection dirigée par P. Degoulet et M. Fieschi, Paris, Springer-Verlag France A. Albert, F.H. Roger-France, P. Degoulet et M. Fieschi Volume 10 - 1998

[2] « DATA MINING » ou FOUILLE DE DONNÉES Gilbert Saporta RST « Epidémiologie », 4.2 Data Mining 5/12/04

[3] Daniel T. Larose (adaptation française T. Vallaud): Des données à la connaissance : Une introduction au data-mining (1Cédérom), Vuibert, 2005 .

[4] René Lefébure et Gilles Venturi, Data Mining : Gestion de la relation client, personnalisations de site web, Eyrolles, mars 2001

[5] Pascal Poncelet, Florent Masseglia and Maguelonne Teisseire (Editors). Data Mining Patterns: New Methods and Applications, Information Science Reference, ISBN: 978-1599041629, (October 2007).

[6] Stéphane Tufféry, Data Mining et Statistique Décisionnelle, Technip, nouvelle édition revue et enrichie, juin 2007

[7] Egmont-Petersen, M., de Ridder, D., Handels, H. (2002). "Image processing with neural networks - a review". Pattern Recognition 35: 2279–2301. doi: 10.1016/S0031-3203(01)00178-9.

[8] Boussaa et al. 2005. Acta Trop., 95, 86–91.

[9] Boussaa et al. 2007. Ann. Trop. Med. Parasitol. 101: 715-724.

[10] OMS. 1990. Lutte contre les leishmanioses. Série de Rapports Techniques.

[11] Rhajaoui et al. 2004. Trans. R. Soc. Trop. Med. Hyg., 98, 299–301.

[12] Rioux et al. 1986. Leishmania. Taxonomie et Phylogenèse. ed. Rioux, J. A. pp. 365–395.