9

Data Warehouse and the Deployment of Data Mining Process to Make Decision for Leishmaniasis in Marrakech City

Habiba Mejhed, Samia Boussaa and Nour el houda Mejhed
[1]Université Cadi Ayyad, Marrakech Département Génie Informatique,
Laboratoire Micro Informatique Systèmes Embarques Systèmes sur Puce.
Ecole Nationale des sciences appliquees,
[2]Université CadiAyyad,
Marrakech Laboratoire d'Ecologie et Environnement,
Faculté des Sciences Semlalia,
[3]&Université Louis Pasteur-Strasbourg I,
Laboratoire de Parasitologie, Faculté de Pharmacie,
[4]Universite Sidi Mohammed Ben Abdallah,
Fes Département Génie Informatique,
Ecole Nationale des sciences appliquées de Fes.
[1,2,4]Morocco
[3]France

1. Introduction

In the last decade, the epidemiology applied more and more tools to help to make decision.
The aim is to translate epidemic data using the modelling concepts of health information systems for the 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 rapid and complete information needed for the 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 and policy framework 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 clinical decision and public health.
The clinical decision support provided epidemiologist technologies necessary to facilitate this difficult task (Degoulet & Fieschi, 1998), (Gilbert, 2004), (Teh, 2009).
The data warehouse remains a valuable tool for storage and data accessibility, 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.

If the data warehouse used to store historical data, with the finality analysis, the data mining is defined as a process of exploration and modelling data in order to discover new correlations, to find trends and stable patterns in the data. It proposes a number of tools from different disciplines, in particular, to decision making in epidemiology [Daniel, 2005), [René & Gilles, 2001), (Pascal et all, 2007), (Stéphane, 2007), (Egmont et all, 2002), (Hang & Xiubin, 2008). 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.

In Morocco, leishmaniasis remains a severe public health problem. Many foci were described in rural areas of Ouarzazate (Rioux et all 1986), Essaouira (Pratlong et all, 1991), Azilal (Rhajaoui et all, 2004), Chichaoua (Guernaoui et all 2005) and Al Haouz (Boussaa et all, 2009), (Rioux et all, 1986) but also in suburban areas, in Taza (Guessous et all, 1997) and Fez (Rhajaoui et all, 2004) Marrakesh is an interesting study site because it lies close to the focus of cutaneous leishmaniasis in the south of Morocco (Ouarzazate, Chichaoua and Al Haouz), and current studies have classified the area of Marrakesh as being at risk of cutaneous leishmaniasis (Boussaa et all, 2005), (Boussaa et all, 2007).

As the best choice of a vector-control strategy is dictated by sandfly ecology, we try to simplify this complex of diseases and quantify the climatic factors which can determine the distribution and activity of sandflies vectors in Marrakesh city. According to the WHO (2005), the activity of sandfly fauna is affected by many climatic factors as temperature, humidity and wind, besides seasons and according to sandfly species.

In this chapter we present a simple contribution to the fight against leishmaniasis in Morocco. The idea is to propose to epidemiologists an application based on tools of data warehouse and data mining to help them to make 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 develop a data warehouse to store and extrapolate data collected in Marrakech city, and we used a data mining tools for Leishmaniasis data analysis to get a better decision-making.

We studied three forms of leishmaniasis according to sandfly vectors collected in Marrakech city by (Boussaa et all, 2005), (Boussaa et all, 2007): Phlebotomus papatasi proven vector of zoonotic cutaneous leishmaniasis with Rodents as a reservoir; P. sergenti proven vector of anthroponotic cutaneous leishmaniasis and P. longicusis potential vector of visceral leishmaniasis with canine as reservoir hosts.

2. Material and methods

A. Sandfly collections

Specimens were collected in Marrakech city (31º36'N, 8º02'W, 471m a.s.l.) between October 2002 and September 2003 as described by (Boussaa et all, 2005). The specimens caught were preserved in 70% ethanol. They were cleared in potash 20% and Marc-Andre solution, then dehydrated and mounted in Canada balsam (Abonnenc, 1972). The identification was made by examining the morphology of male genitalia, female spermathecae and pharynges. For Larroussius species, we revised our specimens according to results of [Boussaa et all, 2008).

B. Data analysis

We have a data file.xls containing all the information on the activity of sandflies P.Papatasi, P. Sergenti and P. Longicuspis based on climate change (date, temperature and density). We can present data from Excel files in the following diagrams:

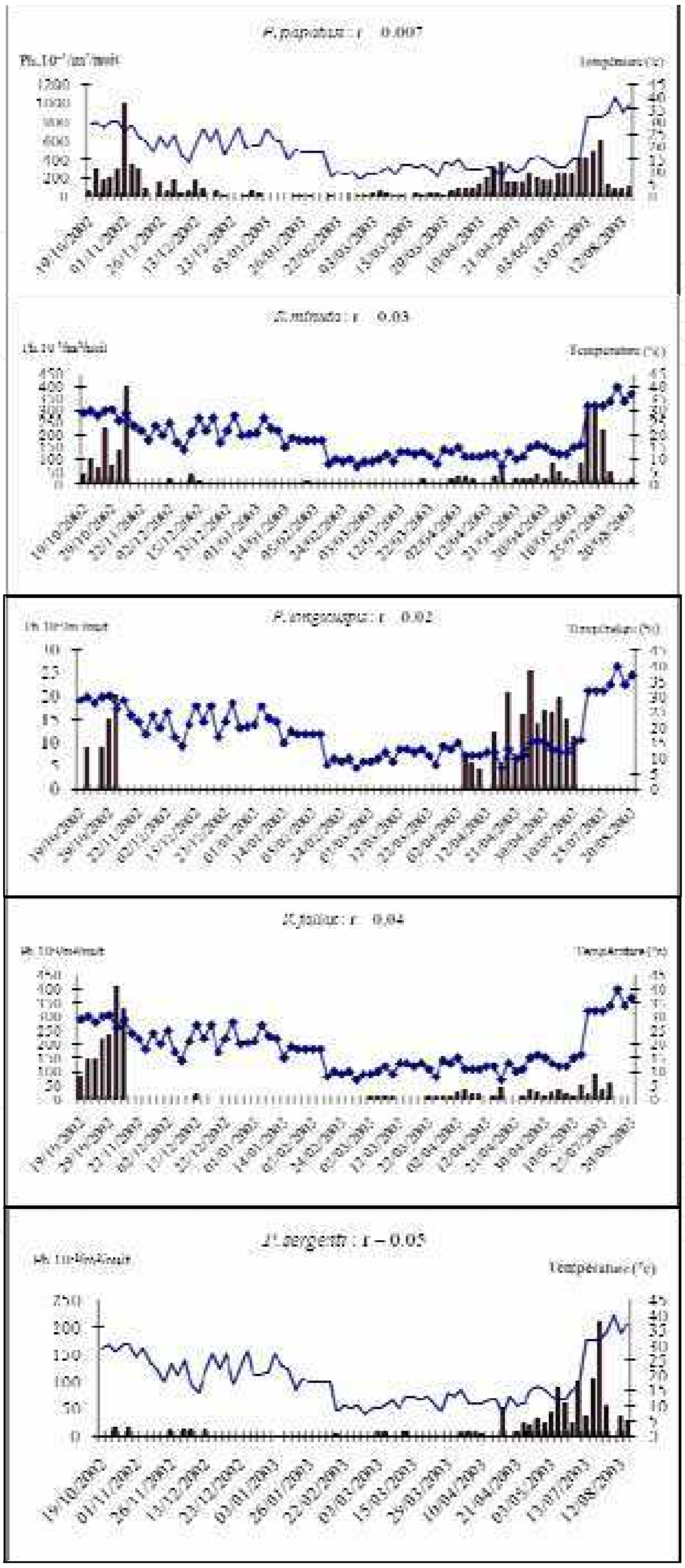


Fig. 1. The activity of sandfly population in Marrakech area

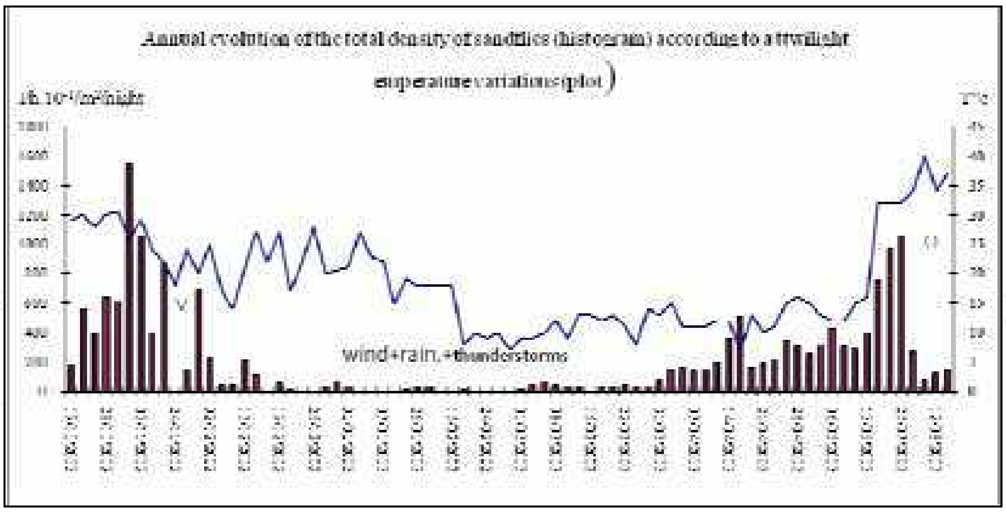


Fig. 2. Annuel evolution of the global density of the sandflies vector according to the time and the climatic change

3. Development of a data warehouse Leishmaniasis

In this section we proceed to the conception of our data scheme. Our investigations on the sandflies collections were carried out in Marrakech city. The goal is to design a database, to develop DW and to load this database, and then the relationships in the cube are built automatically to give the answer to question posed in this case.

3.1 Background

The creation of a data warehouse involves several steps:

The conception: the implementation of a data warehouse usually begins by framing the project, define the needs and goals expressed by policymakers, modeling 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.

a. The static extraction will be performed when the DW must be loaded for the first time and is conceptually a copy of operational data.
b. 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.

3.2 Needs specification

We have developed the analysis tools concerning, the population of sandflies, according to the various species listed in Marrakech city and their density, we considered the human population which may become affected after bites of infected sandflies.

These tools allow knowing the following information: (a) The density of each species of sandflies listed, (b) The period at risk for the spread of the disease, (c) The rate of infection of humans by infected sandflies, (d) The rate per unit time, which a man loses her immunity and becomes susceptible, (e) The rate of infected and susceptible in humans.

To meet the needs of decision makers, we implemented the DW with respect to the architecture described in the following section.

3.3 Data warehouse architecture

Schematic below shows the architecture of the data warehouse applied to Leishmaniasis data.

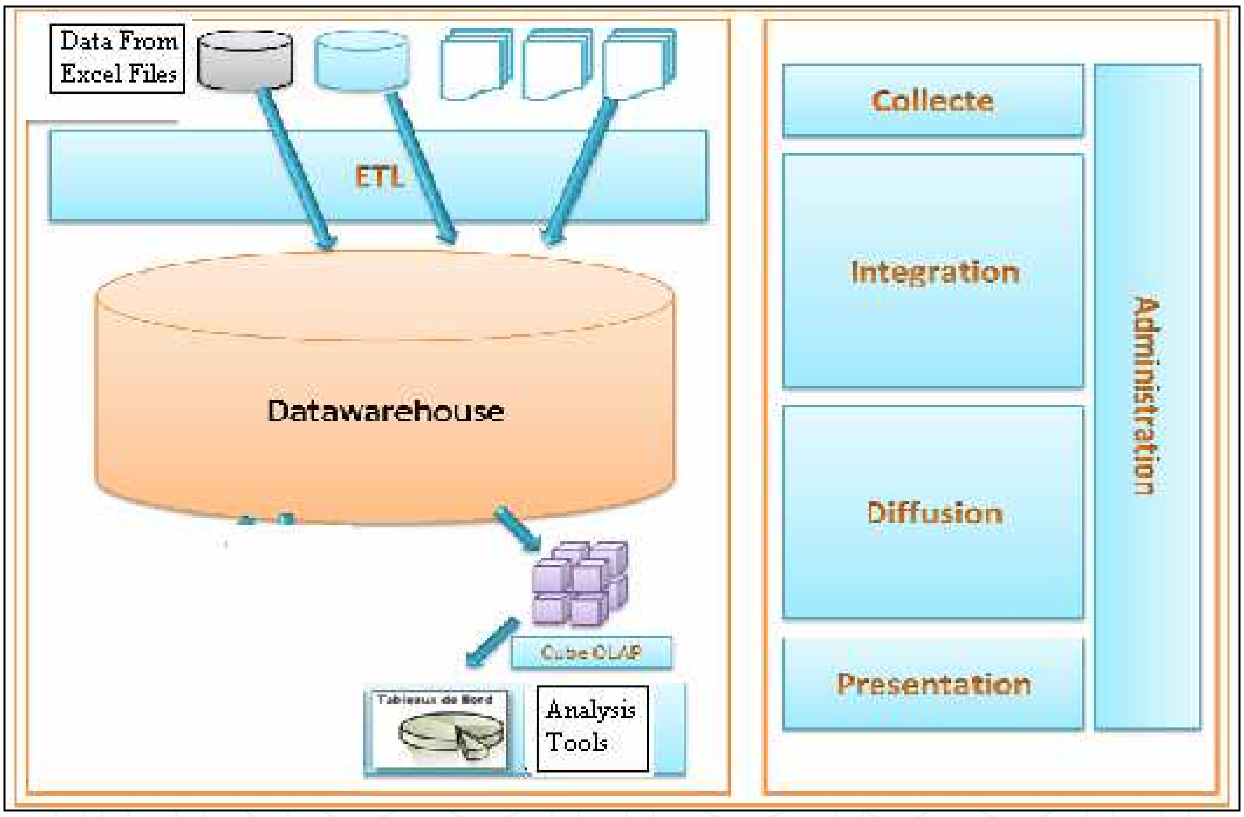


Fig. 3. Data warehouse Architecture

To ensure a robust, flexible and portable solution, we adopted a software architecture divided into several parts:

Collection of raw data files: the job of this module is periodically connected to all servers, to check the generation of new data files

Conversion of the Files: The application of this module is written in Java. This module convert the data files collected by the module above, and filter the information contained therein. It permits to leave only the information that will be used in the future treatments.

Loading data into the data warehouse: ETL process allows Extraction, Transformation and Loading of data from various sources (databases, files) into DW. ETL process is the most

important module to design a Data warehouse with respecting two constraints: data sources and data types (data quality).
Building of multidimensionnel cubes :

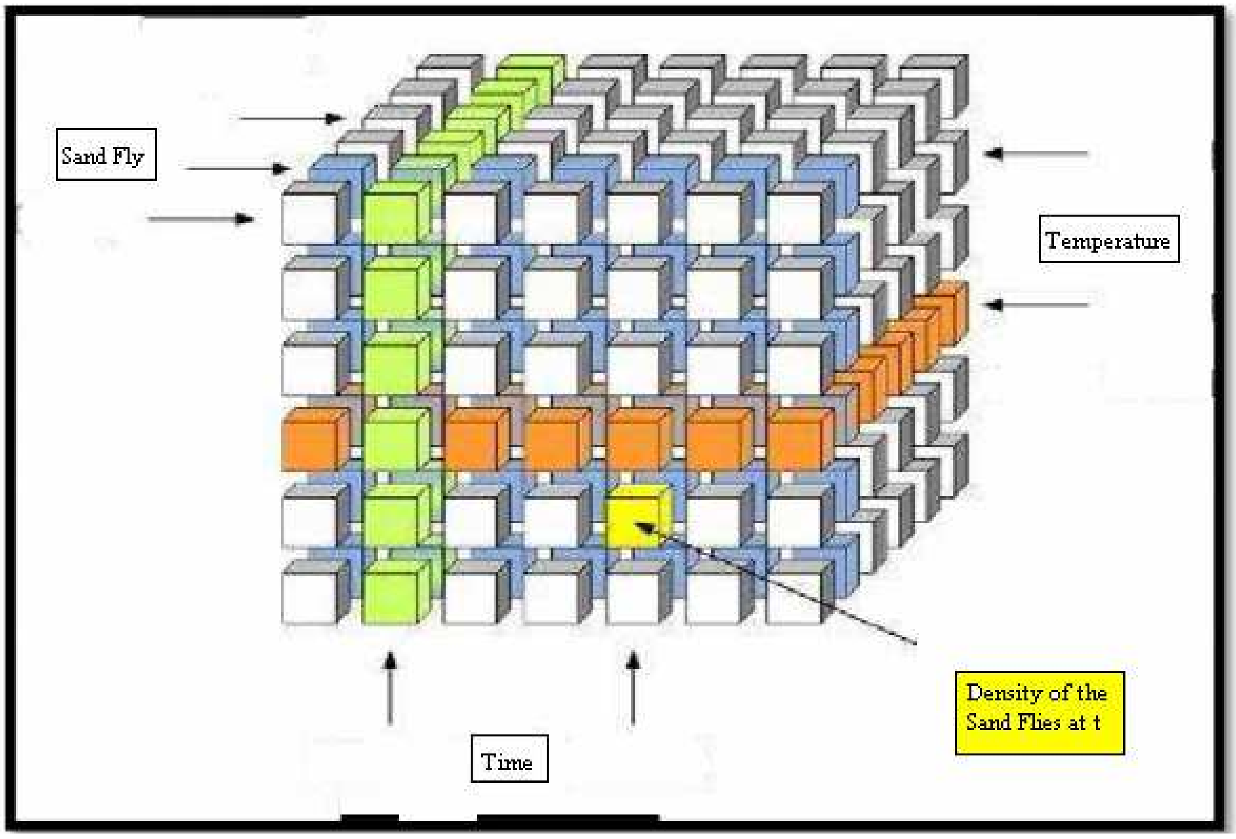


Fig. 4. Example of multidimensional view (Cube) of the Leishmaniasis data.

Creation of the graphical interfaces, graph and report: used for operation, querying cubes and creating reports.

3.4 Data model

This section deals with the transmission of the Leishmaniasis disease from the sandflies vector to human. There are four actors in this case, sandflies species, climatic change and Time and Human. The data dictionary given in the following table:

data	Description of data
Human	Sexe Age statut
sandflies	specie Name
Temperature	Degree
time	wear Month Day

The data layer architecture of Leishmaniasis is illustrated schematically by:

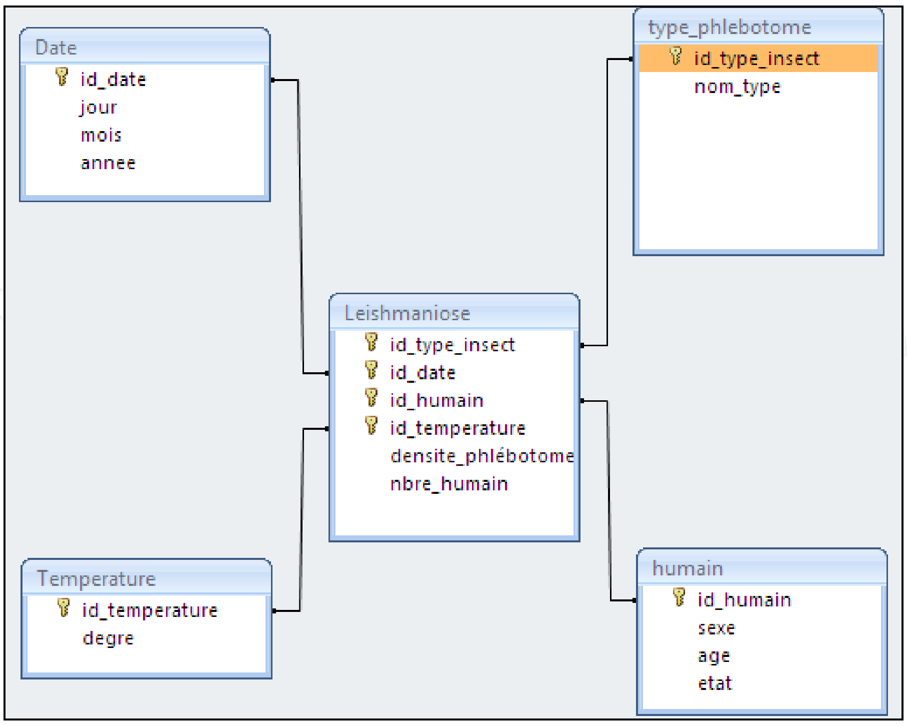


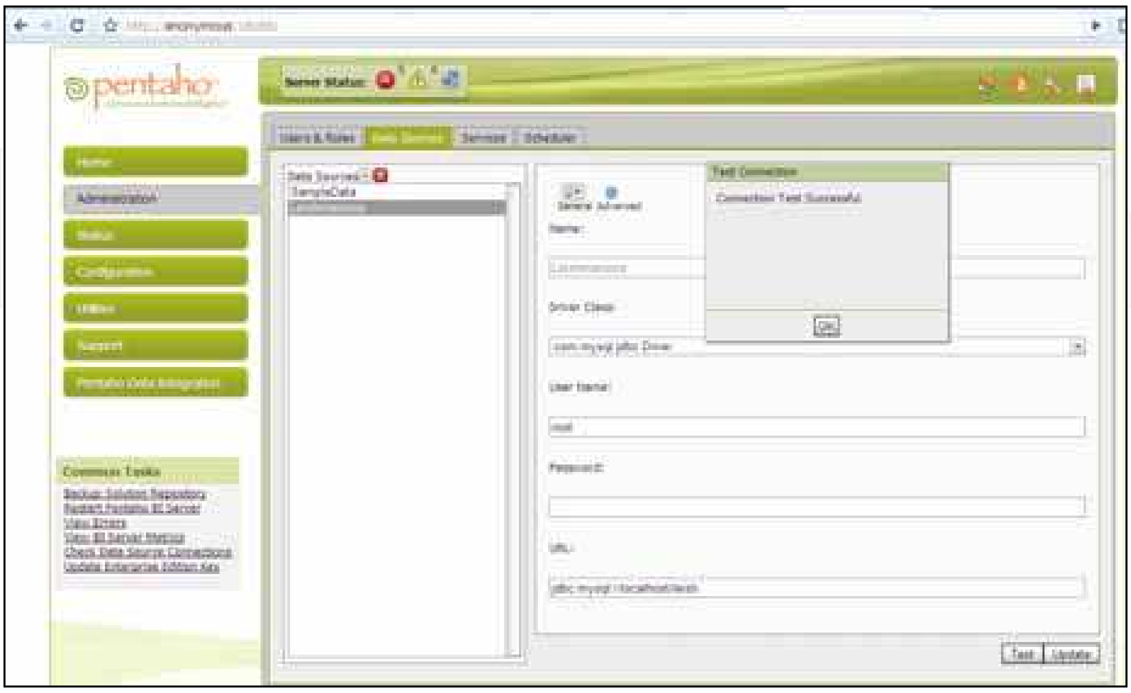
Fig. 5. The model data warehouse diagram

The model chosen must comply with the requirements and needs of use, in our case, we opted for a star pattern respecting the nature of the information we have.

Dimension Tables: Date, Temperature, Sandflies (P. Sergenti, S. Minuta, S. Fallax, P. Logicuspis, P. Papatasi), Human.

Fact Table: Leishmaniose.

Data is extracted from Excel files using java code. The program consists of two classes, one for extraction and one for export useful data to the database.



Connection Test Successful

Name:
Leishmaniose

Driver Class:
com.mysql.jdbc.Driver

User Name:
root

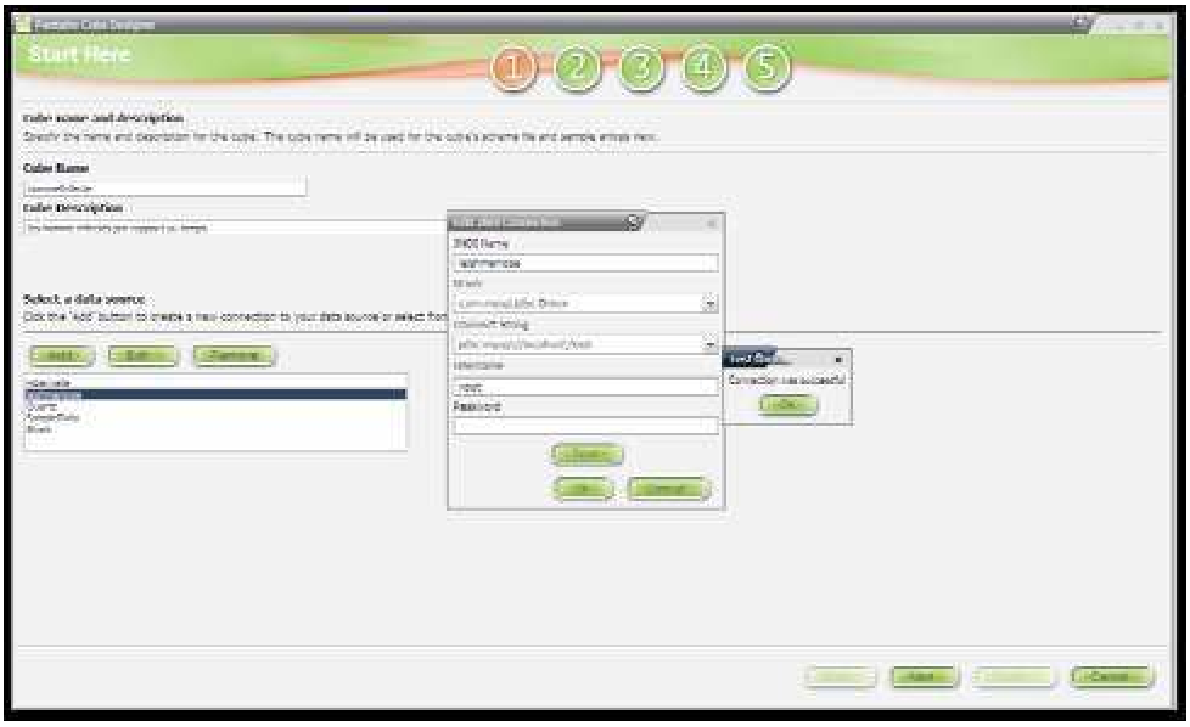
Password:

URL:
jdbc:mysql://localhost/leish

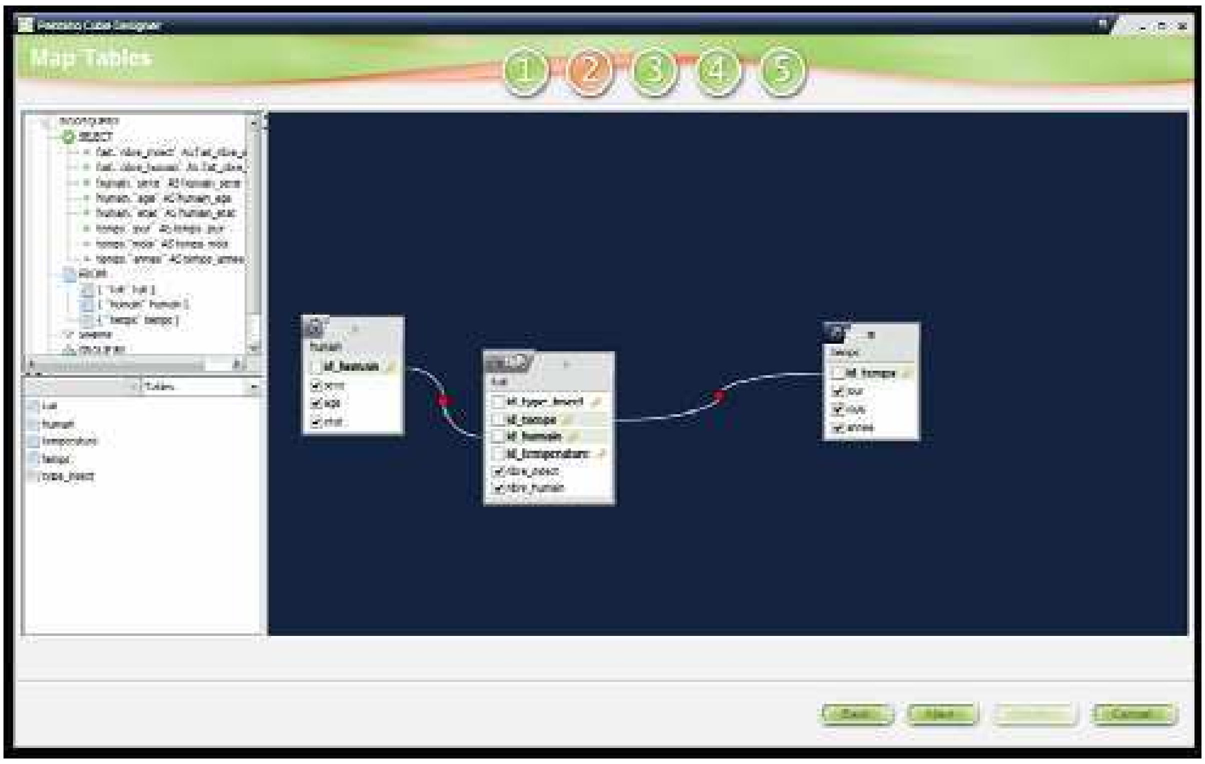
3.5 Dimensional cube

Pentaho integr a CubeDesign tool, it allows to have the cube in XML format and to publish it in the User Pentaho Console. Cube creation through 5 steps as shown in the below:

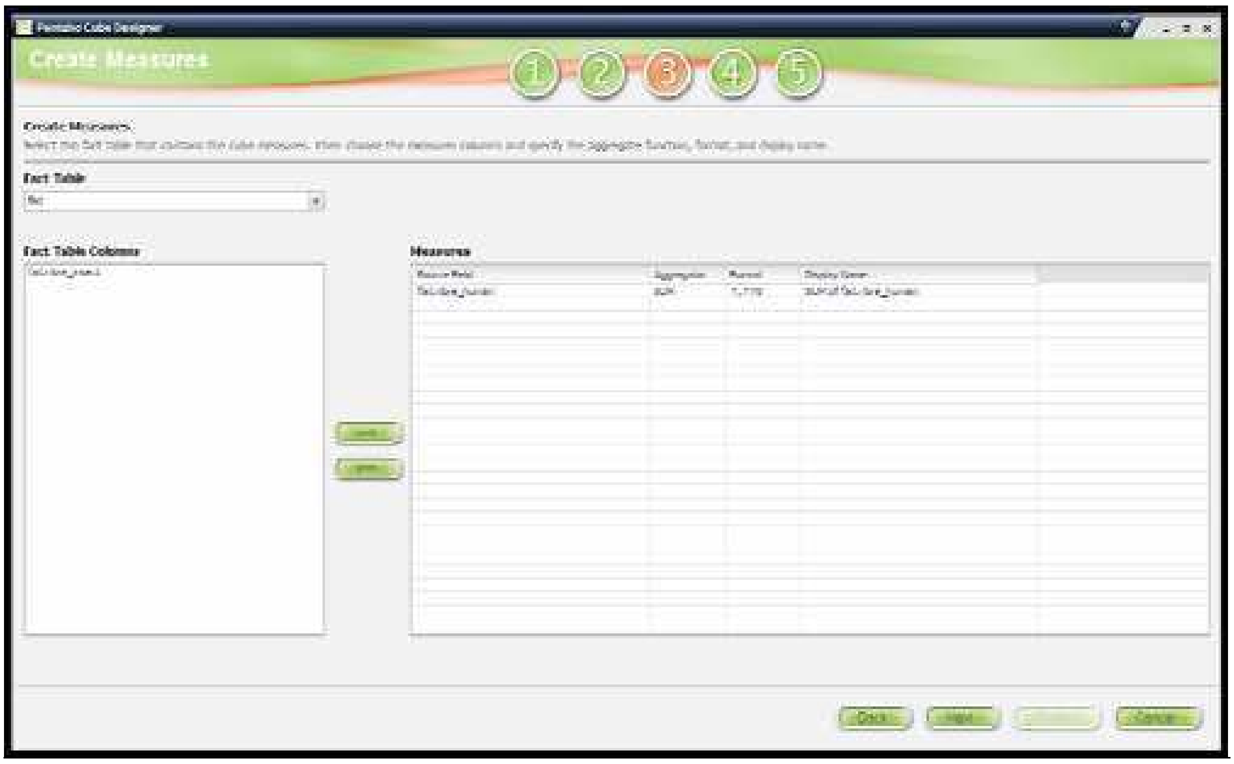
Step 1. With the CubeDesigner we establish the connection to database Leishmaniose.



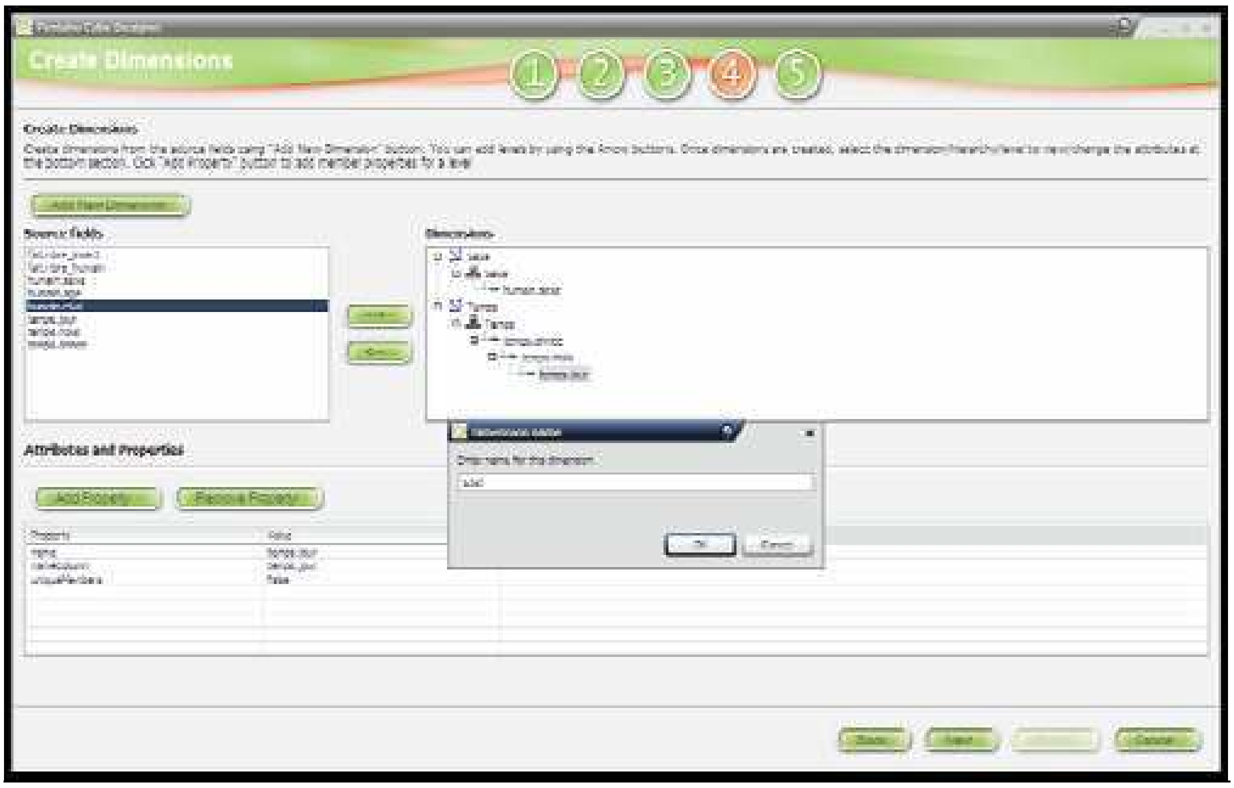
Step 2. Select tables and views useful to visualise the Cube. For example, to calculate the density of the P. Sergenti and the infected human in a fixed interval time.



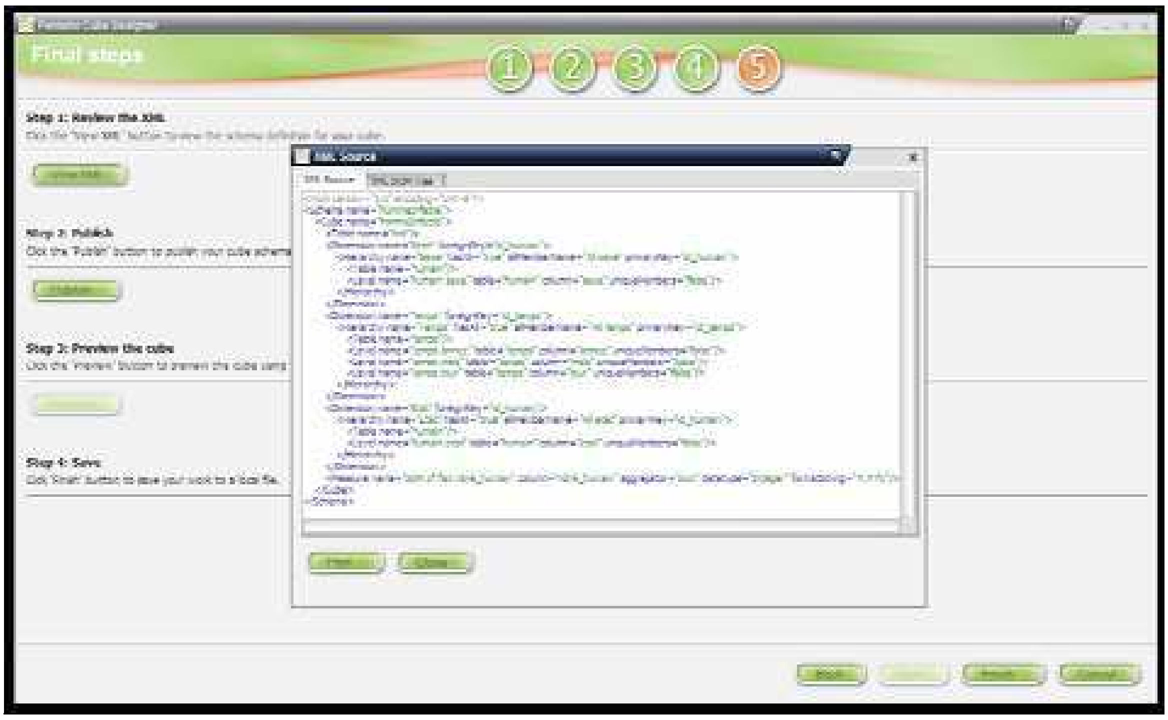
Step 3. Choose of the measures: Infected human



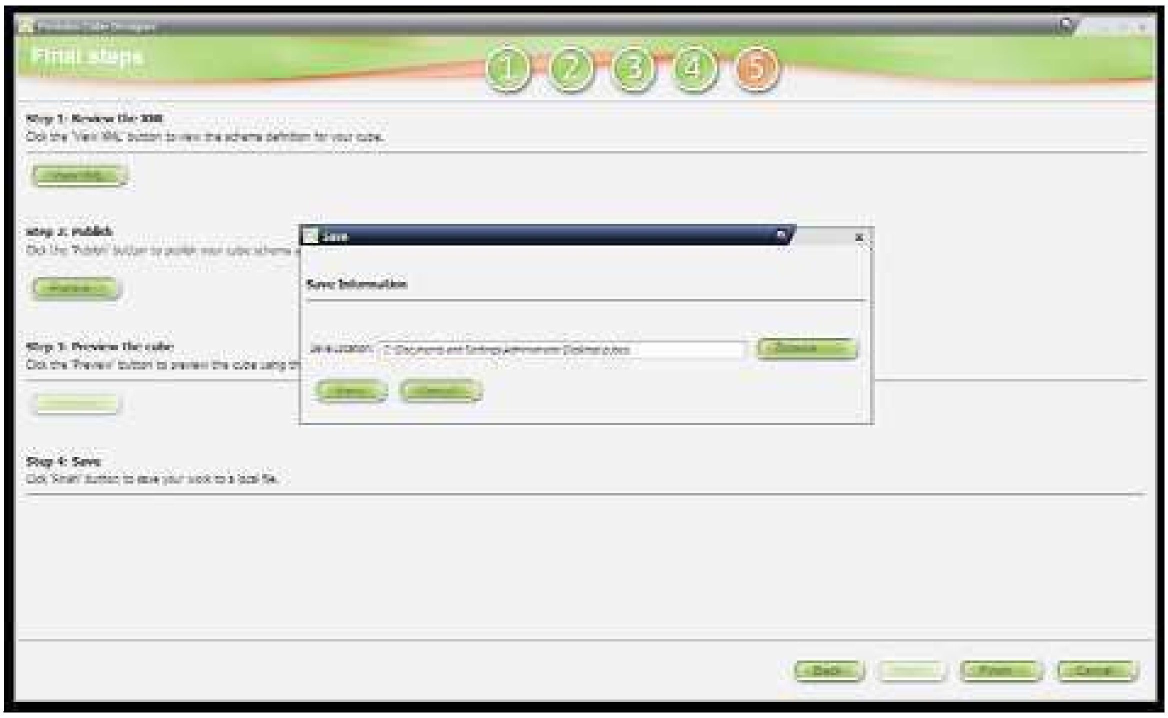
Step 4. All the dimensions and their aggregation will be specified via the interface below.



Step 5. The interface below allows us to view the file for the model; it's possible to update it as needed.



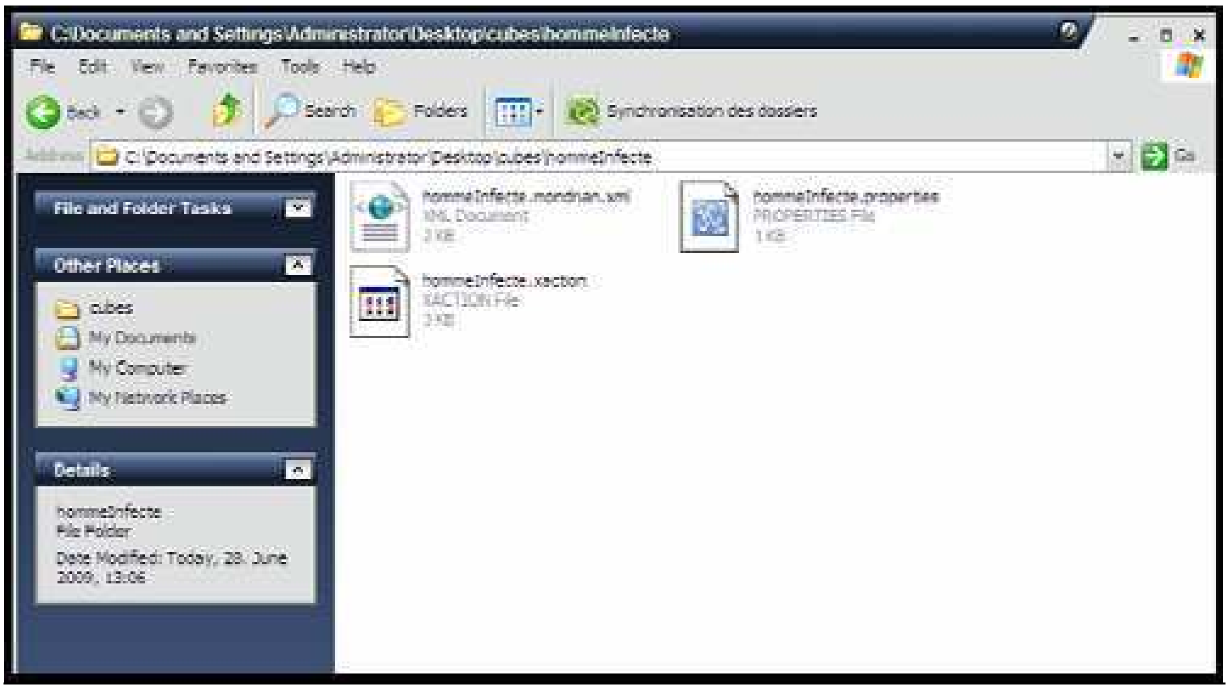
Finally, the pattern is saved



Click the "View XML" button to view the schema definition for your cube.

Click "Finish" button to save your work to a local file.

Tree files are generated : (a) Xml file to save the pattern produced by CubeDesigner. (b) Properties file : for the allocation of the database. (c) Xaction file : presents a set of all protocols to data access.


C:\Documents and Settings\Administrator\Desktop\cubes\hommeInfecte
File Edit View Favorites Tools Help
Back Search Folders Synchronisation des dossiers
Address C:\Documents and Settings\Administrator\Desktop\cubes\hommeInfecte
Go
File and Folder Tasks
Other Places
cubes
My Documents
My Computer
My Network Places
Details

hommeInfecte.mondrian.xml
XML Document
3 KB

hommeInfecte.properties
PROPERTIES File
1 KB

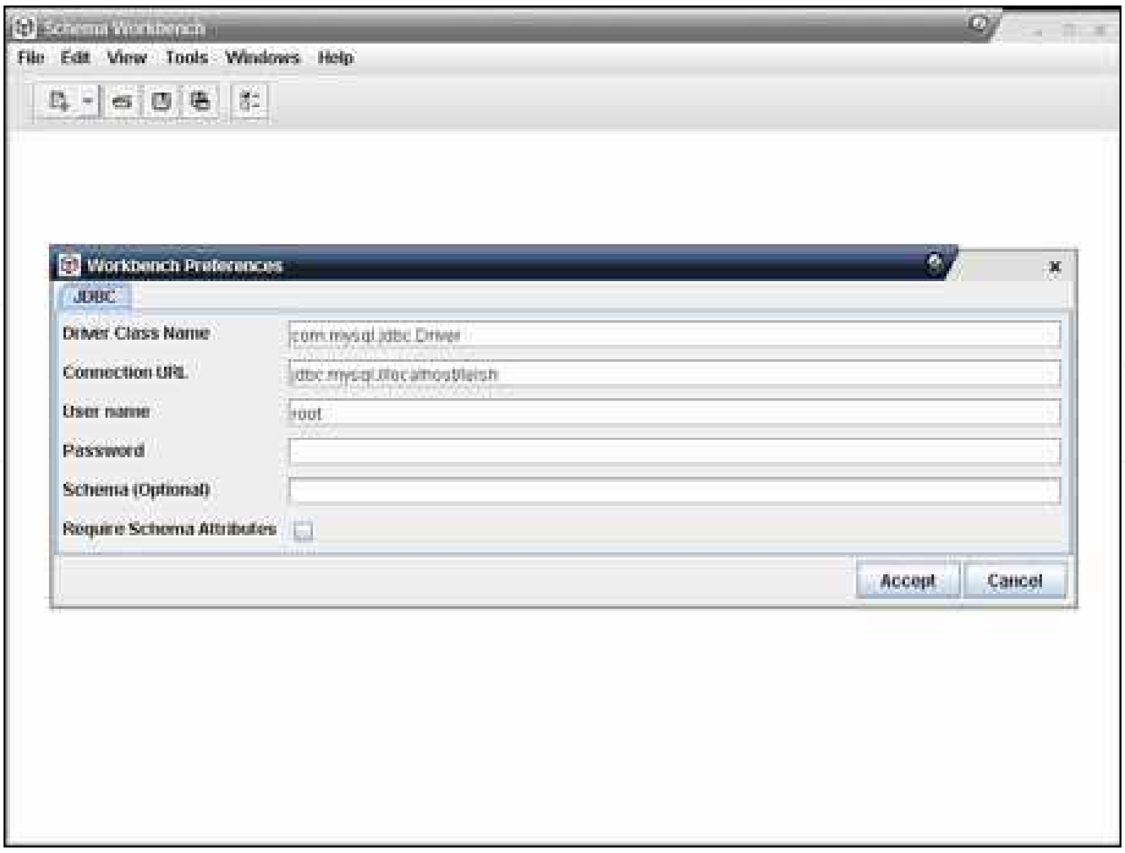
hommeInfecte.xaction
XACTION File
3 KB

hommeInfecte
File Folder
Date Modified: Today, 28. June 2009, 13:06

3.6 Pattern publication with workbench

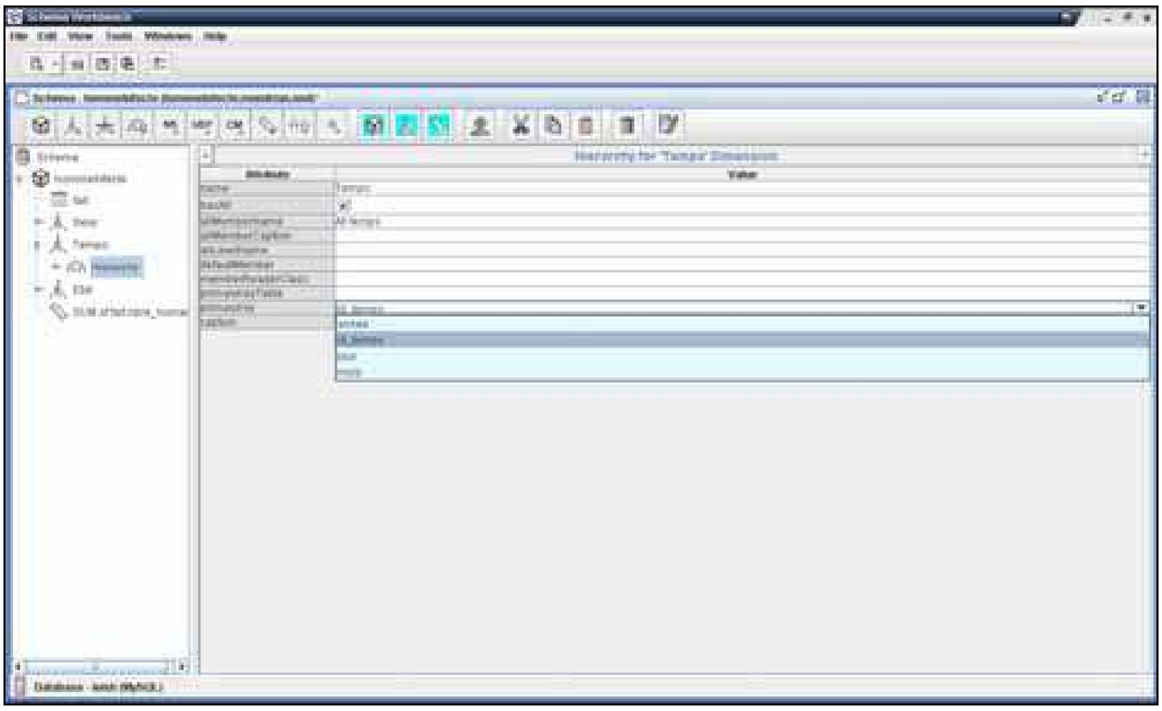
Workben is a tool to create diagrams, for our case it is just used to refine and publish the pattern designed by CubeDesigner.

Step 1. Configuration of the Workbench references to establish connection to database.

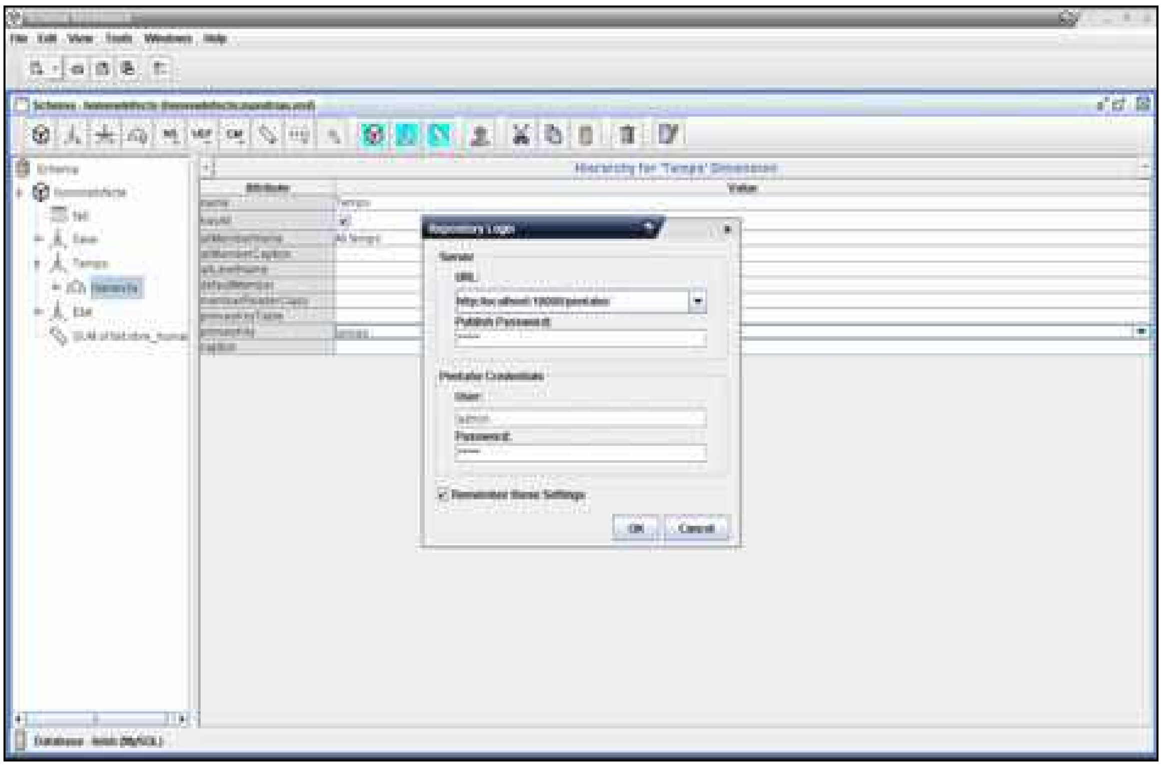


Driver Class Name: com.mysql.jdbc.Driver
Connection URL: jdbc:mysql://localhost/leish
User name: root
Password:
Schema (Optional):
Require Schema Attributes ☐

Step 2. Pattern raffinement.


Schema Workbench
File Edit View Tools Windows Help
Hierarchy for 'Tempo' Dimension
Attribute
Value
Schema
Tempo
Hierarchy
Database - loisir (MySQL)

Step 3. Now, it is important to have URL for the host user, The password for the publisher and The login and password of the user server.


File Edit View Tools Windows Help
Hierarchy for 'Tempo' Dimension
Attribute
Value
Repository Login
Server
Pentaho Credentials
Remember these Settings
OK
Cancel
Database - loisir (MySQL)

URL: http://localhost:18080/pentaho

Publish Password: *****

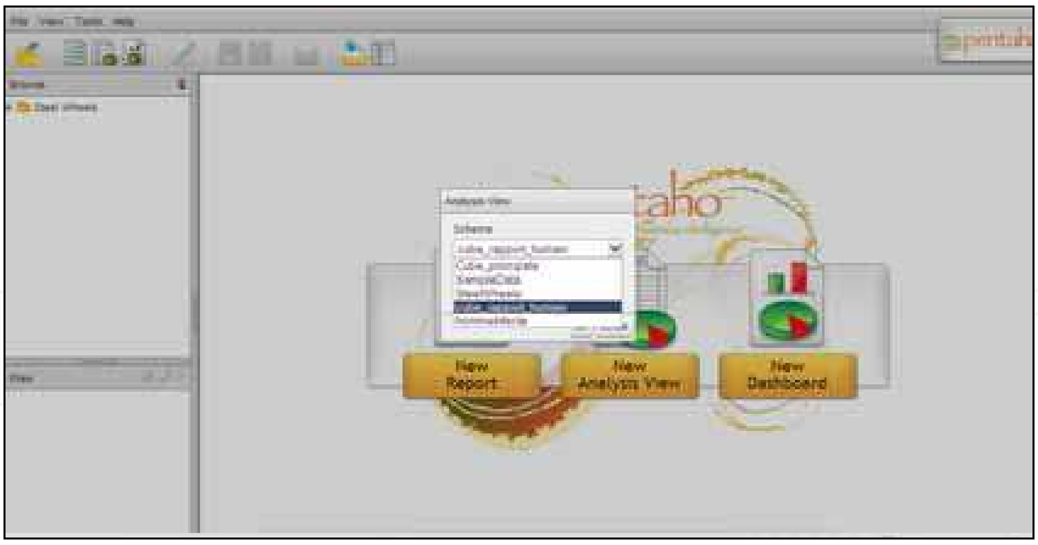
User: admin

Password: *****

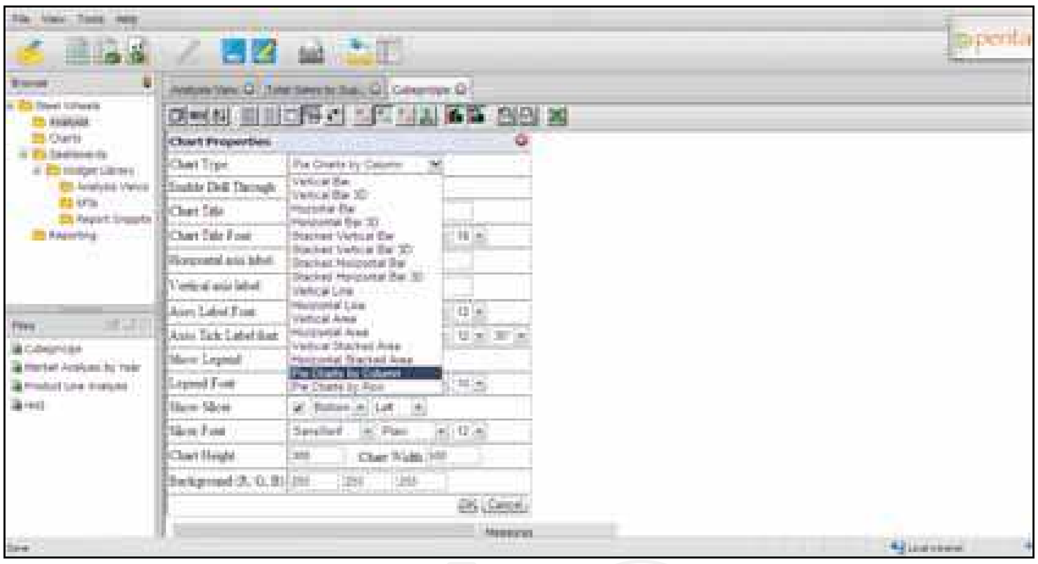
☑ Remember these Settings

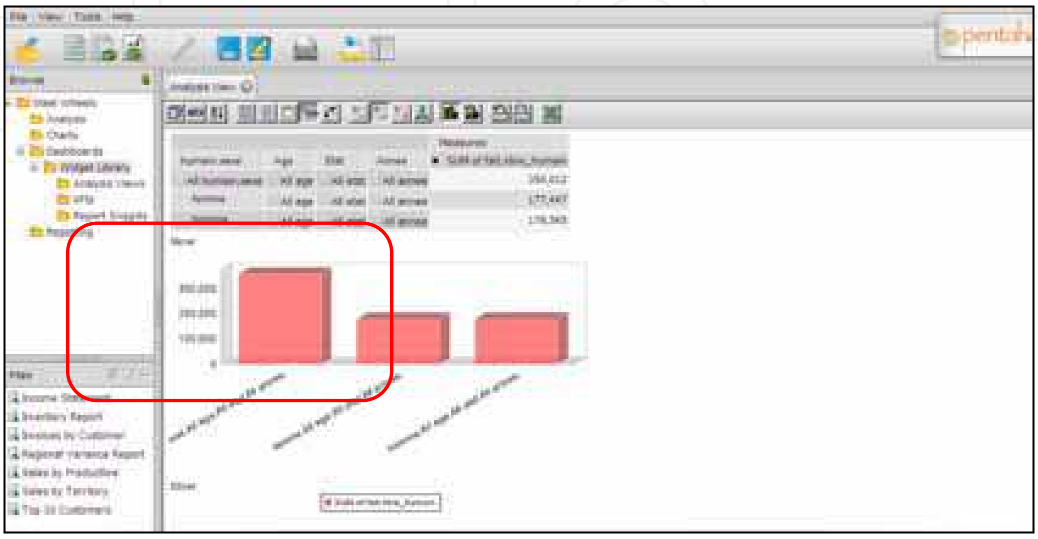
3.7 Visualization phase

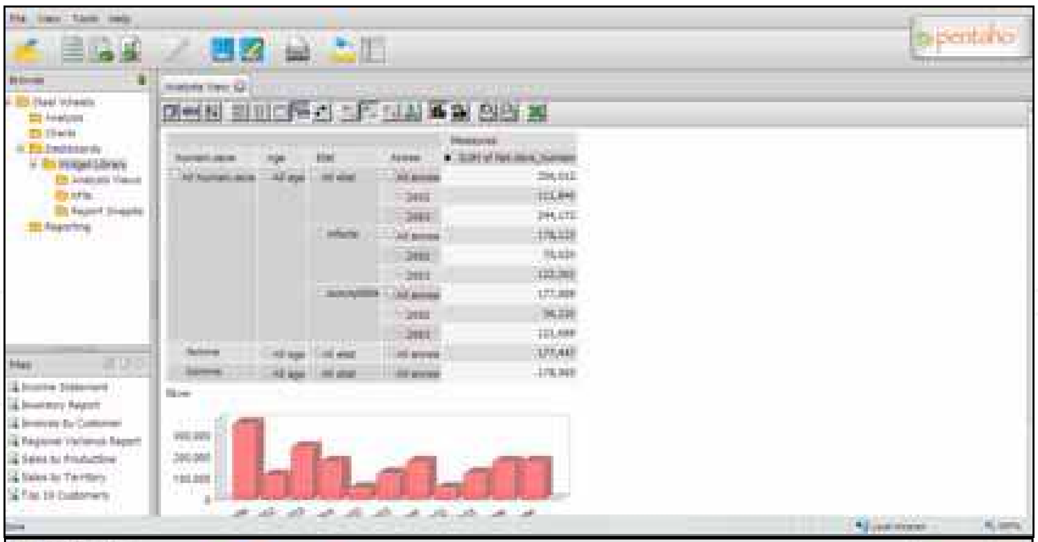
'User Console' tool gives different views of the pattern previously published.

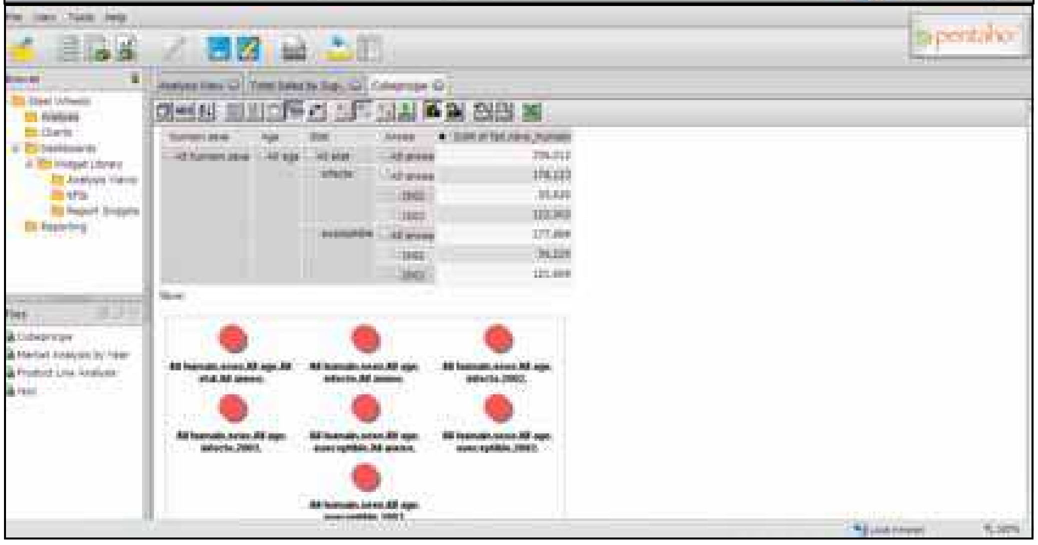


It’s possible to have several modes to visualize data from database. As illustrated in the following figures.

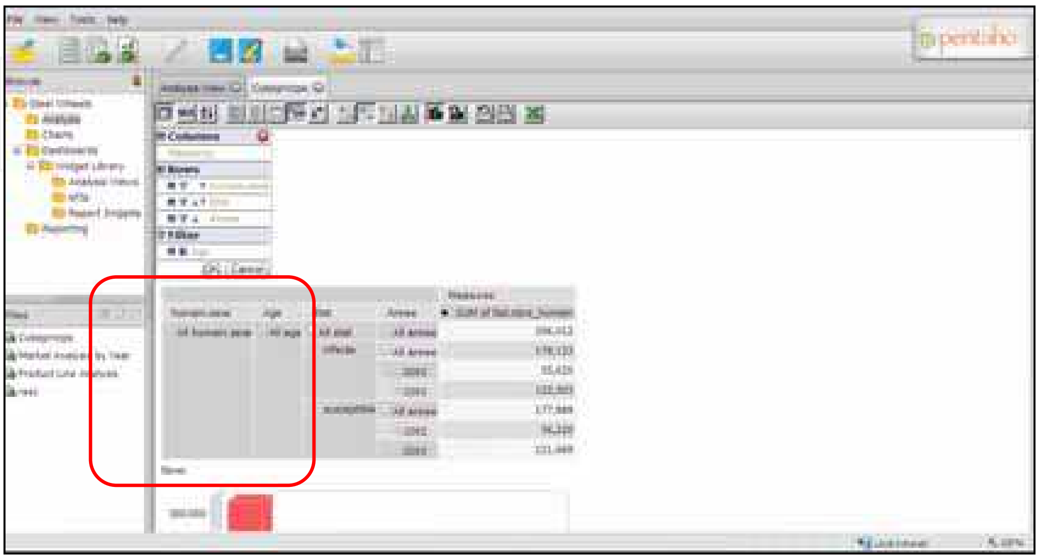








The hearlth professional may use OLAP concept to fialter and visualize other type of information.

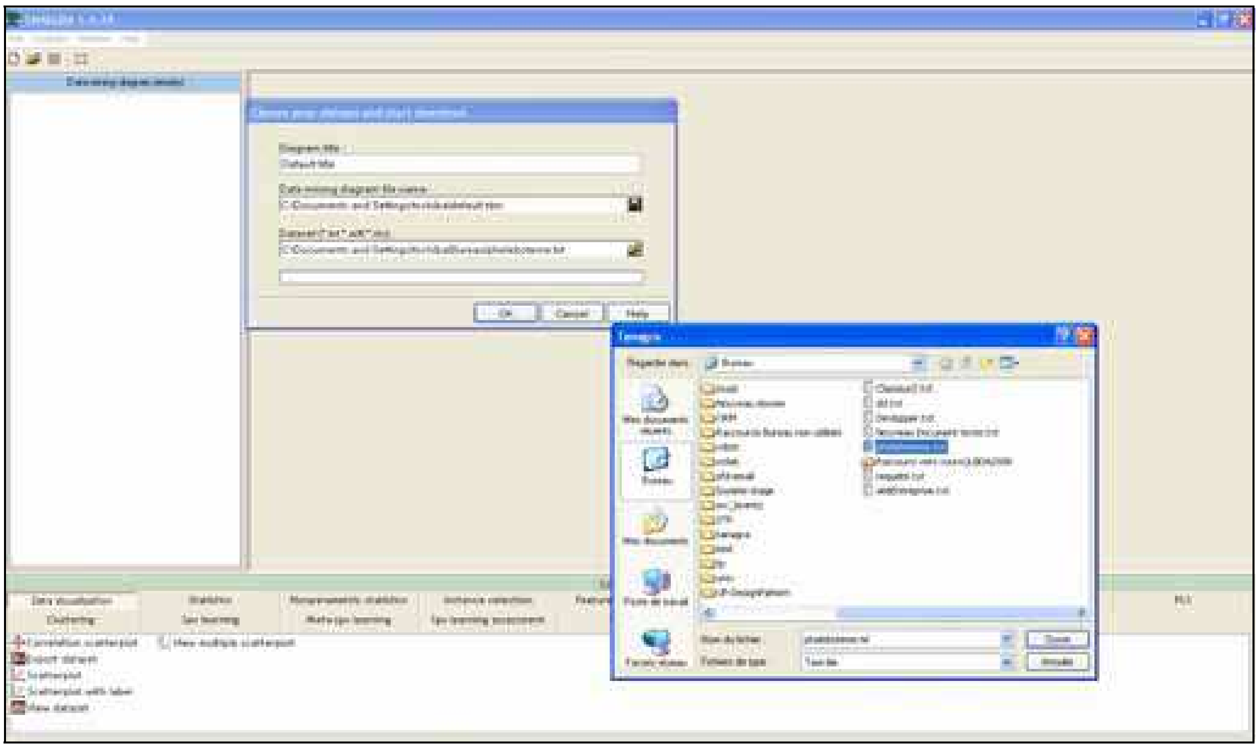


4. Data mining: application to the Leishmaniasis

Given the seriousness of leishmaniasis in Morocco, it was essential to deploy easy methods to reduce its exploitable spread if not eradicate it completely. Our proposal aims to exploit tools of data mining process on this infectiou disease. By definition, the data mining attempts to extract knowledge from vast volumes of data.
The wealth of information transmitted on vectors of disease allows us to apply these tools to identify methods and anticipate behaviour, therefore, make a good decision.

4.1 K-means and deployment

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



4.2 Descriptive statistics

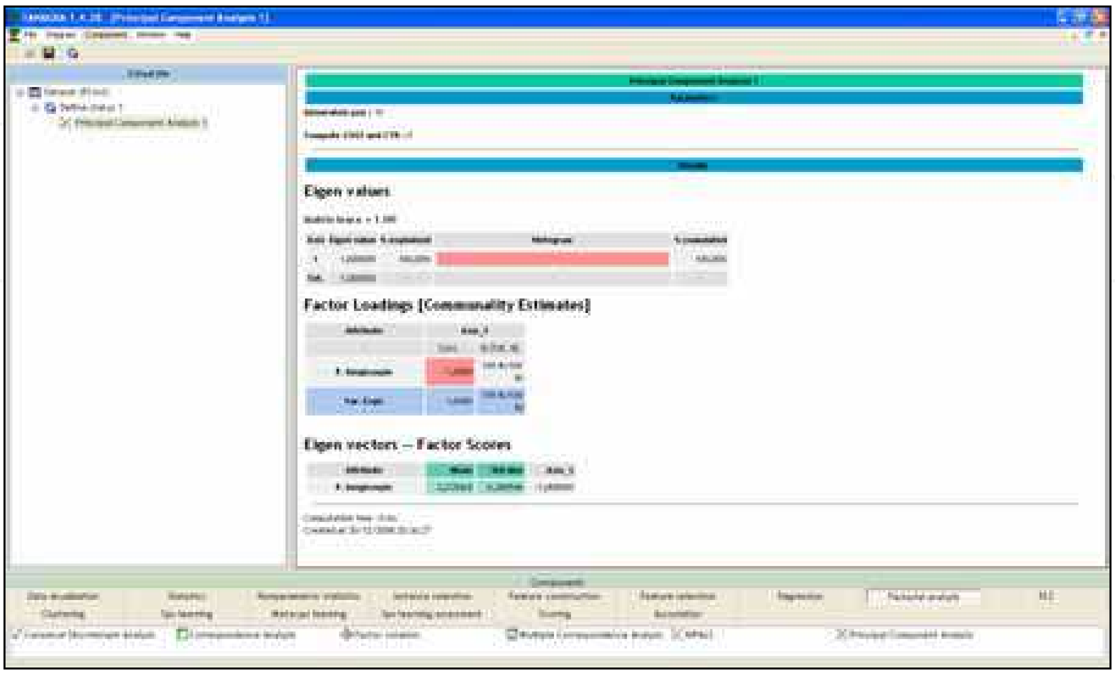
We can do descriptive statistics to variables. We calculate the frequency histograms on all columns to count the number of active and additional comments.

4.3 Method principal component analysis (ACP)

They are three categories of data mining 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.

4.4 Visualization of our data

To implement the ACP method, we can see, for example, date and temperature data concerning P. longicuspis. After we define an analysis of the variables studied. The result is given in the following figure:



Eigen values

Factor Loadings [Communality Estimates]

Eigen vectors – Factor Scores

To better assess the relative positions to sandflies in the first factorial design, 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 point cloud :

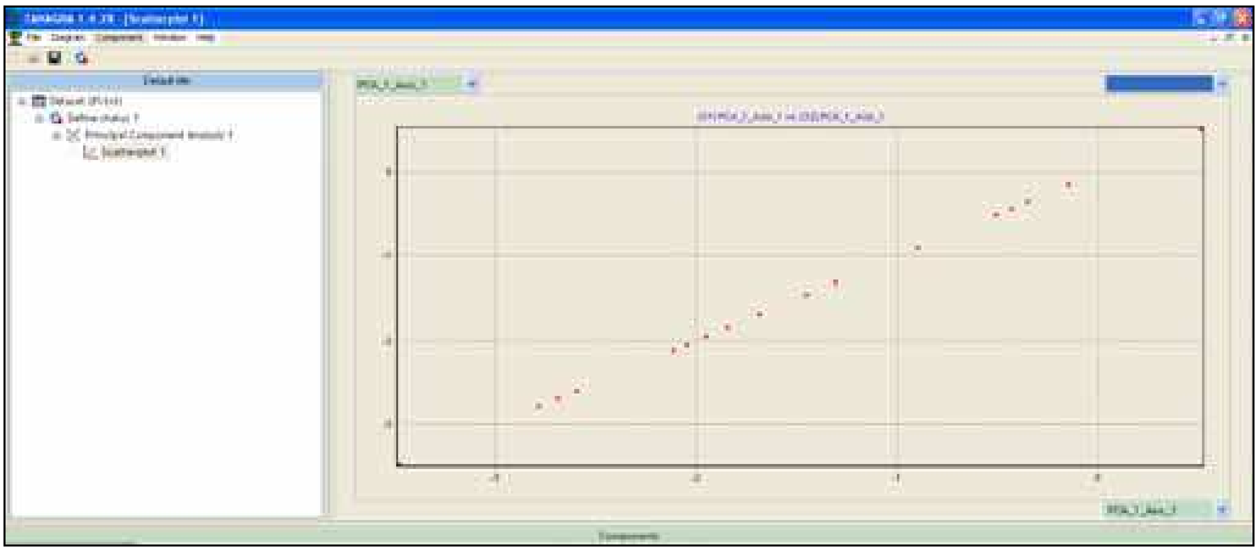


Fig. 6. Point cloud data on (date, temperature, for P. longicuspis)

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

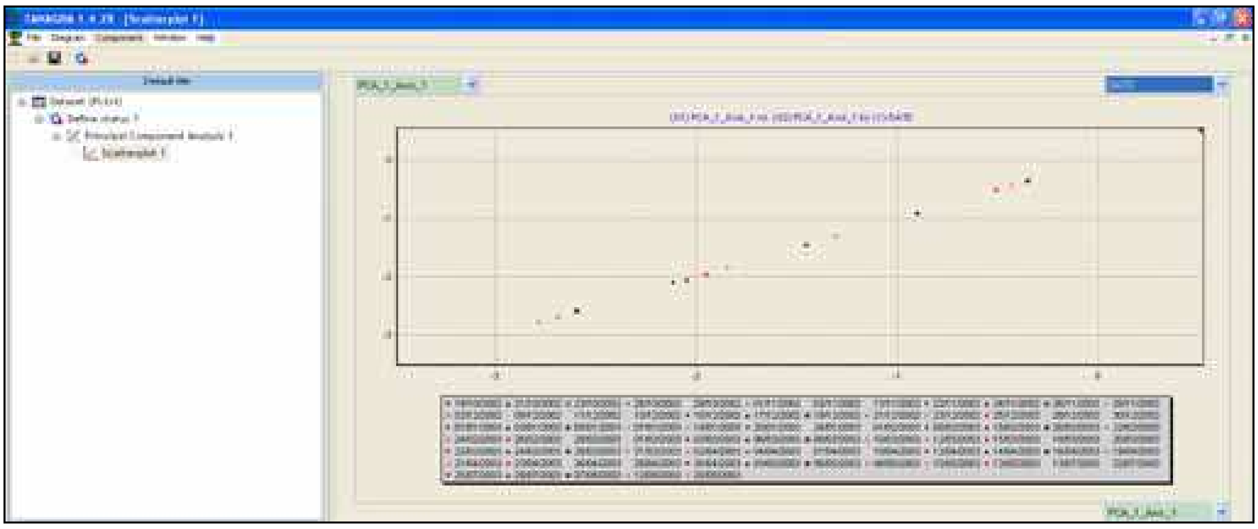


Fig. 7. Using the date_format as illustrative variable

Now we apply the ACP method on all data:

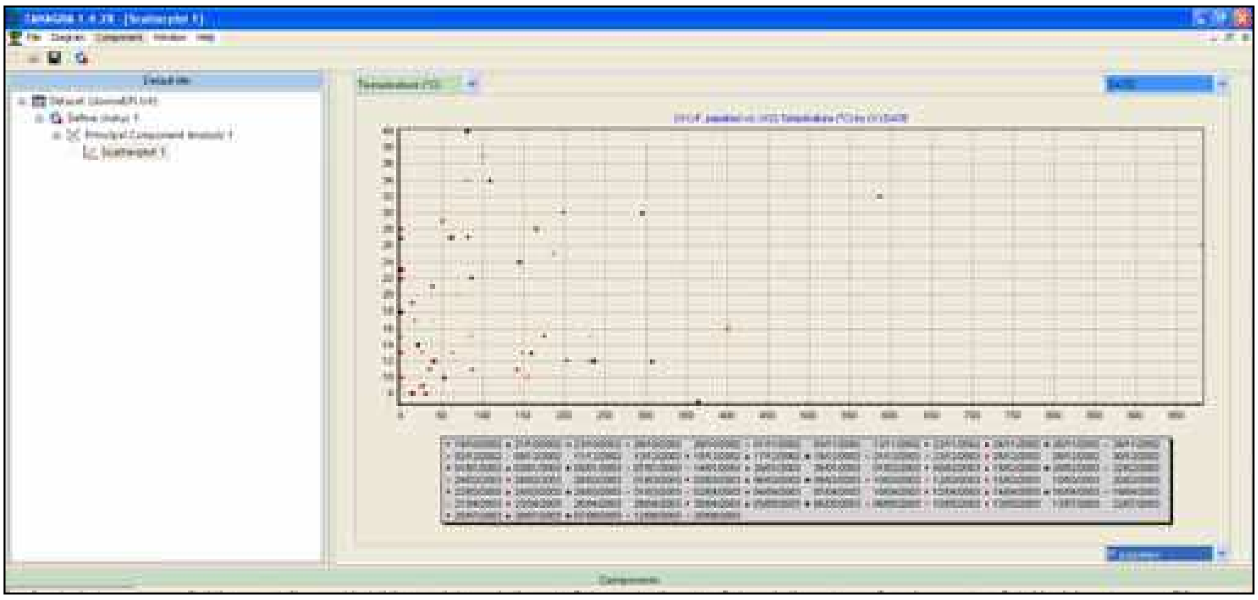


Fig. 8. Implementation of the ACP method on sandflies data

The most difficult decision in any project is to determine what method should be implemented. In our case, we prefer use the open source tools:

a. ETL « Talend Open Studio » Talend Open Studio offer: A complete range of components, Traces and statistics of treatment in real time, the addition of specific code and the integration on the decisional open source.
b. MySQL to create the database and data warehouse for the persistent storage area.
c. Pentaho BI Suite: cover many areas of Business intelligence through various software (owned by Pentaho or integrated into).
d. Pentaho Analysis (Mondrian + JPivot), Cube Designe, schema Workbench. For analysis OLAP
e. Open source data mining software Tanagra.

5. Discussion

Our investigations were conducted in Akioud, an urban district in Marrakech city, during one-year-study. This site was selected, considering the presence, of all the sandfly species inventoried in the urban area of Marrakech (Boussaa et all, 2007).

According to the correlation between the weekly density of the three vectors (P. papatasi, P. sergenti and P. longicuspis) and the factor R0, we can prevent the risk of leishmaniasis in this area.

a. For P. papatasi population, R0 factor is superior to 1 during two periods of the year: November and May-June-July, which correspond to the periods of risk of zoonotic cutanous leishmaniais caused by L. major in this area.
b. For P. sergenti population, R0 is superior to 1 during the period of July–August and inferior to 1 in the rest of the year. So, this period corresponds to the phase of risk of anthroponotic cutanous leishmaniais caused by L. tropica in this area. We observe that R0 reaches its peak during the period of August when the temperature is very high.
c. For P. longicuspis, the results have shown two periods of risk of visceral leishmaniasis in this area: October and May-June.

(Boussaa et all, 2005) classified Marrakech area as being at risk of cutaneous leishmaniasis because of the high density of P. papatasi throughout the year, its position close to the cutaneous leishmaniasis foci in the arid region (Rioux et all, 1986) and the omnipresence of Meriones shawi, main L. major reservoir host in Morocco. In this work, we approve the conclusion of [Boussaa et all, 2005) and we exploit these data to assist in the decision on the issue of the fight against leishmaniasis in Morocco.

Indeed, these are spatio-temporal models which permit to follow and control the evolution of leishmaniasis in terms of time and region, and to find the appropriate threshold of the population of sandflies for stopping the multiplication of the disease. In Chichaoua (70 Km from Marrakech city), focus of anthroponotic cutanous leishmaniais, (Bacaër and Guernaoui, 2006) suggest that the epidemic could be stopped if the vector population were reduced by a factor (R0)2 = 3.76.

6. Conclusion

To fight against the spread of the cutaneous and visceral leishmaniasis and to address the need of the responsible for population health to set the policies that determine the nature of health care provided, and most importantly, for fast evaluation of the severity of the multiplication of the disease, we proposed to develop a model of data warehouse to storage appropriate data of the sandflies seasonality and the information about the susceptible human (patient with suspected Leishmania infection). Then, structural key and functional parameters will be measured to ensure the advanced treatment. By employing data mining process we have the ability to extract knowledge and to generate summaries for better health decision-making. Through this complete and operational platform, it will be easy to control the sandflies seasonality and the rate of the transmission disease.

7. References

P. Degoulet et M. Fieschi (1998). *Informatique et santé Collection, Paris, Springer-Verlag*, Volume 10,

Gilbert Saporta (2004). Data mining ou fouille de données, RST « Epidémiologie » *Data Mining*.

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

René Lefébure et Gilles Venturi (2001). Data Mining : Gestion de la relation client, personnalisations de site web*, Eyrolles*, mars 2001.

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

Stéphane Tufféry (2007). Data Mining et Statistique Décisionnelle, Technip, nouvelle édition revue et enrichie, juin 2007.

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

Boussaa, S., Guernaoui, S., Pesson, B., Boumezzough, A. (2005). Seasonal fluctuations of phlebotomine sand fly populations (Diptera: Psychodidae) in the urban area of Marrakech, Morocco. pp. 86–91, *Acta Trop.* 95.

Boussaa, S., Pesson, B., Boumezzough, A. (2007). Phlebotomine sandflies (Diptera: Psychodidae) of Marrakech city, Morocco. pp. 715-724, *Ann. Trop. Med. Parasitol.* 101.

WHO (2005). Lutte contre les leishmanioses. *Série de Rapports Techniques*.

Rhajaoui, M., Fellah, H., Pratlong, F., Dedet, J.P., Lyagoubi, M. (2004). Leishmaniasis due to Leishmania tropica MON-102 in a new Moroccan focus. *Trans. R. Soc. Trop. Med. Hyg.* 98, pp. 299–301.

Teh and All (2009). Development of a Data warehouse for Lymphoma Cancer Diagnosis and Treatment Decision Support. *Proceedings of the 10th WSEAS International Conference on MATHEMATICS and COMPUTERS in BIOLOGY and CHEMISTRY,* Pp. 15 _24. ISSN: 1790-5125, ISBN: 978-960-474-062-8.

Abonnenc E. (1972). Les phlébotomes de la région éthiopienne (Diptera: Phlebotomidae). *Mémoire de l'ORSTOM.* 55, pp. 1–289.

Bacaër, N., Guernaoui, S. (2006). The epidemic threshold of a simple seasonal model of cutaneous leishmaniasis. *J Math. Biol.* 53, pp. 421–436.

Boussaa, S., Boumezzough, A., Remy, P. E., Glasser, N., Pesson, B. (2008). Morphological and isoenzymatic differentiation of Phlebotomus perniciosus and Phlebotomus longicuspis (Diptera: Psychodidae) in Southern Morocco. *Acta Trop.* 106, pp. 184–189.

Boussaa, S., Pesson, B., Boumezzough, A. (2009). Faunistic study of the sandflies (Diptera: Psychodidae) in an emerging focus of cutaneous leishmaniasis in Al Haouz province, Morocco. *Ann. Trop. Med. Parasitol.* 103, pp. 73-83.

Guernaoui, S., Boumezzough, A., Pesson, B., Pichon, G. (2005). Entomological investigations in Chichaoua: an emerging epidemic focus of cutaneous leishmaniasis in Morocco. J. *Med. Entomol.* 42, pp. 697–701.

Guessous-Idrissi, N., Chiheb, S., Hamdani, A., Riyad, M., Bichichi, M., Hamdani, S., Krimech, A. (1997). Cutaneous leishmaniasis: an emerging epidemic focus of Leishmania tropica in north Morocco. *Trans. R. Soc. Trop. Med. Hyg.* 91, pp. 660–663.

Pratlong, F., Rioux, J.A., Dereure, J, Mahjour, J, Gallego, M., Guilvard, E., Lanotte, G., Périères, J, Martini, A., Saddiki, A. (1991). Leishmania tropica au Maroc. IV. *Diversité isozymique intrafocale. Ann. Parasitol. Hum. Comp. 66*, pp. 100–104.

Ramaoui, K., Guernaoui, S., Boumezzough, A. (2008). Entomological and epidemiological study of a new focus of cutaneous leishmaniasis in Morocco. *Parasitol.* Res. 103, pp. 859-863.

Rioux, J.A., et all (1986). Les leishmanioses cutanées du bassin méditerranéen occidental: de l'identification enzymatique à l'analyse éco-épidemiologique, l'exemple de trois 'foyers', tunisien, marocain et français. Montpellier, France: *Institut Méditerranéen d'Etudes Épidémiologiques et Ecologiques*. pp. pp.365–395.

Hang Xaio, Xiubin Zhang (2008). Comparison Studies on Classification for Remote Sensing Image Based on Data Mining Method*, WSEAS TRANSACTIONS on COMPUTERS.* Volume 7, ISSN: 1109-2750, pp. 552 558.