Publication pour la présente candidature LAFPT

Titre	Decision Making Environment on Rift Valley Fever in Ferlo (Senegal)
Auteurs	Fanta Bouba, Alassane Bah, Christophe Cambier, **Samba Ndiaye**, Jacques-André Ndione; Maguelonne Teisseire.
Référence	Acta Biotheoretica
Editeur	Springer
Pages	405 – 415
Année	2014
DOI	https://doi.org/10.1007/s10441-014-9235-7
URL	**https://link.springer.com/article/10.1007%2Fs10441-014-9235-7**
Index	https://www.scopus.com/sourceid/12456
ISBN	1572-8358
Encadreur	Oui
Extrait d'une thèse	Oui

Source details

Acta Biotheoretica

Scopus coverage years: from 1935 to 1943, from 1947 to 1949, from 1951 to 1953, from 1955 to 1965, from 1967 to 1988, from 1990 to 2018

Publisher: Springer Nature

ISSN: 0001-5342 E-ISSN: 1572-8358

Subject area: Arts and Humanities: Philosophy

Agricultural and Biological Sciences: General Agricultural and Biological Sciences View all

View all documents > Set document alert Journal Homepage

CiteScoreTracker 2018

Last updated on *11 February, 2019*
Updated monthly

0.90 = Citation Count 2018 / Documents 2015 - 2017 = 62 Citations to date > / 69 Documents to date >

Springer Link

Acta Biotheoretica
September 2014, Volume 62, Issue 3, pp 405–415 | Cite as

Decision Making Environment on Rift Valley Fever in Ferlo (Senegal)

Authors Authors and affiliations

Fanta Bouba, Alassane Bah, Christophe Cambier, Samba Ndiaye, Jacques-André Ndione, Maguelonne Teisseire

Regular Article
First Online: 09 August 2014

136
Downloads

Abstract

The Rift Valley fever (RVF), which first appeared in Kenya in 1912, is an anthropozoonosis widespread in tropical areas. In Senegal, it is particularly felt in the Ferlo area where a strong presence of ponds shared by humans, cattle and vectors is noted. As part of the studies carried out on the environmental factors which favour its start and propagation, the focus of this paper is put on the decision making process to evaluate the impacts, the interactions and to make RVF monitoring easier. The present paper proposes a model based on data mining techniques and dedicated to trade experts. This model integrates all the involved data and the results of the analyses made on the characteristics of the surrounding ponds. This approach presents some advantage in revealing the relationship between environmental factors and RVF transmission vectors for space–time epidemiology monitoring purpose.

Keywords

Data mining Decision-making system Multidimensional modelling Rift Valley fever Spatio-temporal patterns

Copyright information

© Springer Science+Business Media Dordrecht 2014

About this article

Check for updates

Cite this article as:
Bouba, F., Bah, A., Cambier, C. et al. Acta Biotheor (2014) 62: 405. https://doi.org/10.1007/s10441-014-9235-7

Received	Accepted	First Online
22 December 2013	11 July 2014	09 August 2014

DOI	Publisher Name	Print ISSN
https://doi.org/10.1007/s10441-014-9235-7	Springer Netherlands	0001-5342

Online ISSN
1572-8358

About this journal

Reprints and Permissions

Acta Biotheor
DOI 10.1007/s10441-014-9235-7

REGULAR ARTICLE

Decision Making Environment on Rift Valley Fever in Ferlo (Senegal)

Fanta Bouba · Alassane Bah · Christophe Cambier · Samba Ndiaye · Jacques-André Ndione · Maguelonne Teisseire

Received: 22 December 2013 / Accepted: 11 July 2014
© Springer Science+Business Media Dordrecht 2014

Abstract The Rift Valley fever (RVF), which first appeared in Kenya in 1912, is an anthropozoonosis widespread in tropical areas. In Senegal, it is particularly felt in the Ferlo area where a strong presence of ponds shared by humans, cattle and vectors is noted. As part of the studies carried out on the environmental factors which favour its start and propagation, the focus of this paper is put on the decision making process to evaluate the impacts, the interactions and to make RVF monitoring easier. The present paper proposes a model based on data mining techniques and dedicated to trade experts. This model integrates all the involved data and the results of the analyses

F. Bouba (✉) · A. Bah · S. Ndiaye
UMI 209, UMMISCO-UCAD, Dakar, Senegal
e-mail: boubafanta@gmail.com

A. Bah
e-mail: alassane.bah@gmail.com

S. Ndiaye
e-mail: samba.ndiaye@ucad.edu.sn

F. Bouba · C. Cambier
UPMC, Paris, France
e-mail: christophe.cambier@upmc.fr

A. Bah
ESP/UCAD, Dakar, Senegal

C. Cambier
UMI 209, UMMISCO-IRD, Bondy, France

J.-A. Ndione
CSE and LPA/UCAD, Dakar, Senegal
e-mail: jacques-andre.ndione@cse.sn

M. Teisseire
TETIS - Irstea, Montpellier, France
e-mail: maguelonne.teisseire@teledetection.fr

Published online: 09 August 2014

made on the characteristics of the surrounding ponds. This approach presents some advantage in revealing the relationship between environmental factors and RVF transmission vectors for space–time epidemiology monitoring purpose.

Keywords Data mining · Decision-making system · Multidimensional modelling · Rift Valley fever · Spatio-temporal patterns

1 Introduction

The Rift Valley fever (RVF) is an arbovirosis which is transmitted by vectors. Despite a low death rate, this disease, seen as an "economic and biological weapon, has a great power of diffusion and is particularly serious" (Diagne 1992). As a matter of fact, its virus diffusion can be made by aerosol and the vectors eggs are resistant to a long duration period and dryness (Flick and Bouloy 2005). That is why it appears in the common list of several species diseases of the International Organization for Animal Health (Information 2013). In the current approach, the understanding of emergence and spreading process offers a compartmental cellular view. However, significant works (Ndione et al. 2003; Préhaud and Bouloy 1997), have demonstrated the impact of factors such as climate, livestock movement, hydrology on the emergence and persistence of RVF. The complexity of this analysis requires the involvement of all the actors concerned by multidisciplinary dynamics. This approach, based on the multi-criteria and multidimensional analysis, makes it possible to propose health risk management measures by controlling the impacts of the environmental factors and their interactions. It is therefore proposed that a decision making approach be adopted in order to provide trade experts and decision makers with analysis tools, related to the dynamics of animal and human populations. So, to find out new and relevant information, and in order to understand the interactions between the different objects and make predictions, it is necessary to resort to data mining techniques for classification and prediction purposes. These techniques are applied to assess the relevance of the parameters retained for the RVF space–time monitoring. Further in this paper, Sect. 2 is devoted to the description of the studied disease and to the data search applied in epidemiology. Section 3 describes the manipulated data and the experimented data mining algorithms. Section 4 presents (1) the data model gathering the environment and health parameters integrating the time-space dimensions and (2) extracts of correlated analysis on the characterization of temporary pools in the studied area. Section 5 concludes this approach by discussing the choices made and by proposing study tracks for the continuation of this work.

2 Thematic and Scientific Context

2.1 Rift Valley Fever

RVF is an infectious viral disease which affects animals and humans (OMS 2010). This disease, conveyed by arthropod vectors (Diallo et al. 2000), is caused by a

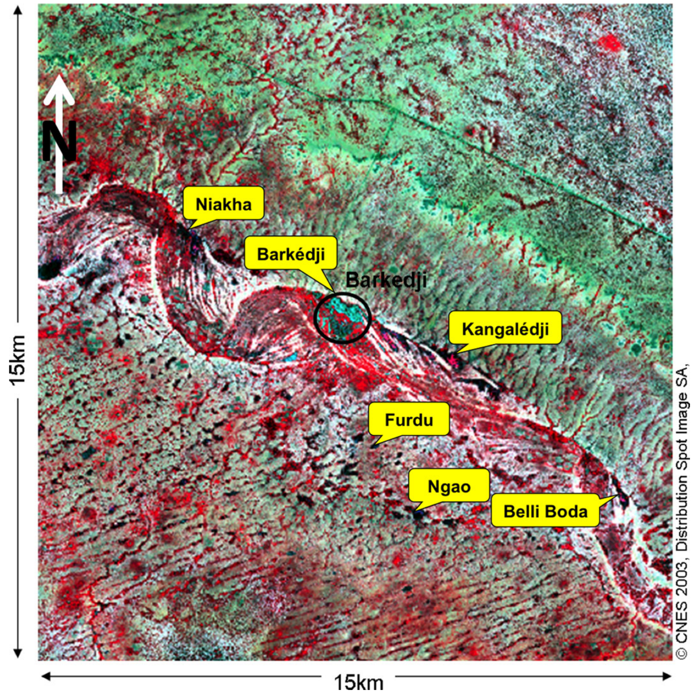

Fig. 1 Ponds of the study area, adapted from Ndione et al. (2009) and Tourre et al. (2008)

virus of the Bunyaviridae family Phlebovirus kind. The RVF can be transmitted by two main modes: the vectorial transmission mode by some blood-sucking insects and the direct transmission mode by personal contact with an infected host (Greboval 2003). In this article, virological data issued from vectors are used. In Senegal, this disease holds an important place in Ferlo. According to Diallo (1995), RVF spread identified in the Ferlo area (Fig. 1) is linked to the mosquito life cycle and ponds evolutionary cycle.

Indeed, Ferlo's ponds constitute both the water supply source for the population and the livestock while being the larva living place of the vectors. In this area, the rainfall intra-seasonal variability, the vegetation dynamics and the turbidity of temporary ponds (the size of which is comparatively small) are the main factors explaining the strong concentration of mosquitoes (Ndione et al. 2009). To understand the vectorial diseases process, it is necessary to take into account the multiple and varied phenomena in the same environment, but also at different scales (time, space and organization).

The partners[1] involved in the QWeCI project (http://www.liv.ac.uk/qweci) in Senegal (Fig. 2) provide data from field investigations (data observed or generated by measurement equipment) and laboratory tests (analysis data).

[1] UCAD (Université Cheikh Anta Diop)—IPD (Institut Pasteur de Dakar)—CSE (Centre de Suivi Ecologique)—DSV (Direction des Services Vétérinaires)—PNLP (Programme National de Lutte contre le Paludisme).

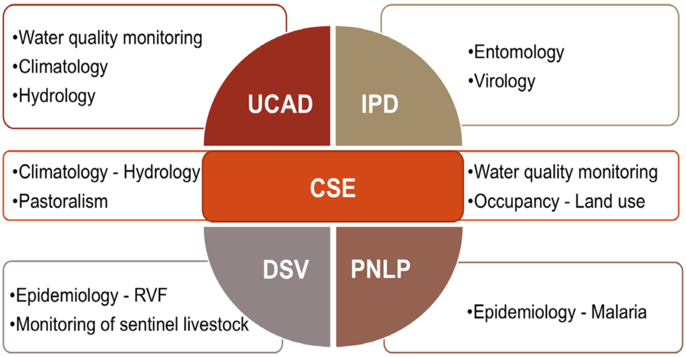

Fig. 2 Organizational complexity

The challenge to be taken up by the research teams is the impact control of the environmental factors on the triggering and spreading of RVF. It is therefore fundamental to take into account the whole set of the different disciplines data in the proposed data model.

2.2 Decision Making Environment and Data Mining in Epidemiology

Data mining is an outstanding stage in the decision-making process allowing to review information, not a priori obvious, in a high data volume. As part of this work, the analysis method of the space time data is defined by Agrawal and Srikant (1995) and Alatrista et al. (2011). The authors of this method focus on including the spatialization notion in the knowledge extraction process. This approach is based on pre-processing the data to include space characteristics into time data. The space sequences thus obtained are then used as inputs for the data mining stage in order to extract spatially frequent sequences. This is achieved with the help of a classical algorithm (Mortazavi-Asl et al. 2000) to extract the sequential motives representing the spatially frequent time evolutions of the studied area.

3 Materials and Methods

This section presents data identified during the analysis phase (field surveys or laboratory results) and the formalisms used to process these data.

3.1 Data Description

3.1.1 Space Data

The present study is carried out with the data of Barkédji rural community (Fig. 1). Thus, a distinction can be made between ponds and encampment (Fig. 3). For possible generalizations, the present paper has chosen to integrate a larger

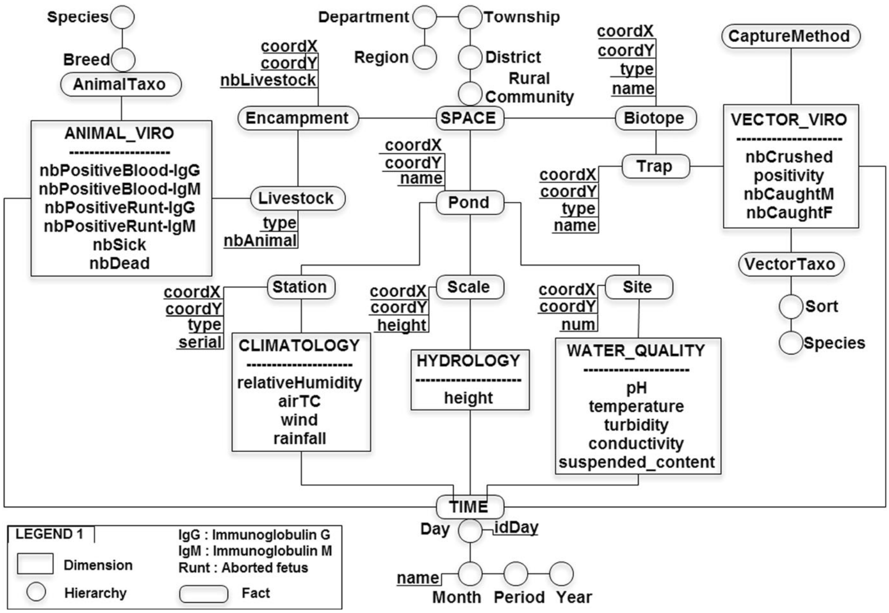

Fig. 3 Data model

dimension by starting from the Senegal administrative division. The pond, however, is the key element of the present study as it is the junction place between vectors and hosts.

3.1.2 Climate, Hydrology and Water Quality Data

Ponds of this geographical area are linked to several elements:

- rainfall stations and automatized meteo stations are installed around ponds: measurements such as wind speed, temperature, rainfall are automatically generated and periodically collected;
- gauging scales are set on water ponds to record the water level;
- water samples are identified by their collection point coordinates: they make it possible to measure pH, temperature and matters in suspension.

3.1.3 Entomological and Virological Data

We also use data from field surveys and laboratory health teams. During the field campaigns, mosquitoes are captured by using bait (sheep, chicken) and CO_2 traps. Then, individual identification follows to determine age, sex, etc. Besides, blood samples are taken up among sentry flocks. The mosquito's virology is assessed based on a crushed sample. On the other hand, animal virology is tested individually on the sampling.

3.2 Methodology

3.2.1 Multidimensional Modelling

The research work on the environment impact on health and more particularly on vector diseases (Tran et al. 2005) made it possible to distinguish two big classes of models: (1) mathematical models (geographical and epidemiological) which are based on a formal representation to provide quantitative descriptions and (2) conceptual models (relational, object and multidimensional) based on the need of the final users. In the present context, the multidimensional approach is best adapted as it provides some formalism dedicated to decision making systems. Multidimensional modelling is described as a representation approach, which considers the analysed subject as a point (the fact) in a several dimension space. Dimensions can be decomposed following a hierarchy linked to the required granularity level Wehrle (2009). In their article on the OLAP algebra Ravat et al. (2010), Ravat et al. propose the following definitions:

- a dimension is defined by its attributes, hierarchies and the set of instances;
- a hierarchy is defined by an ordered set describing the attributes hierarchy (each attributes, called parameter, corresponds to an analysis granularity level);
- a fact is defined by a set of measurements, a set of instances and a function associating each instance to the instances of the dimensions linked to the fact.

3.2.2 Spatial and Temporal Patterns

The spatial dimension is used to define the location of an object in a geographical space and the temporal dimension relates to the time step (date, time, etc.). To better understand this data processing technique, we present the main concepts used, based on the work of Fabrègue et al. (2012):

- item: an item I is a literal value of dimension D_i, $I \in \mathrm{dom}(D_i)$;
- itemset: an itemset, $IS = \langle I_1 I_2 \ldots I_n \rangle$ is a non-empty set of items. All items of an itemset are associated with different dimensions of analysis;
- itemsets sequence: a sequence S is a non empty ordered list of itemsets noted $\langle s_1 s_2 \ldots s_p \rangle$ where s_j is a itemset;
- absolute support: the absolute support of a sequence S is the number of data warehouse areas which satisfy S;
- relative support: the relative support is defined as the ratio between absolute support and the number of areas;
- pattern: let S be the sequence and θ the minimal support defined, if the relative support $\geq \theta$ then S is a spatio-temporal pattern.

4 Results

In this section, we introduce the representation model and the first data mining results obtained with the decision-making environment described by Marakas (2003).

Based on the formalism of multidimensional modelling (Sect. 3.2.1), we built the data model (Sect. 4.1), which is one of the activities of the first step (data management) of this methodology. This model allowed us to implement our data warehouse in which we integrate all handled data (Sect. 3.1).

For the second phase (knowledge management), we use a decision algorithm to identify the species of vectors more present in the different localities of the study area (Sect. 4.2). Then, we apply a spatio-temporal algorithm (Sect. 3.2.2) to get similar parameters of these localities ponds (Sect. 4.3). This work is a technique we proposed to identify and classify risk areas.

4.1 Data Model

In accordance with the multidimensional formalism, based on Golfarelli's proposal (Golfarelli et al. 1998), a constellation model (Fig. 3) has been built, including two (2) fact tables for animal and vector virology and three (3) fact tables for environmental data (climatology, hydrology and water quality). Dimensions have been identified in function of different analysis criteria:

- the vector (mosquitoes) dimensions and the livestock;
- the space dimension hierarchized following the administrative division;
- the time dimension taking into account study periods;
- environment analysis dimensions: scale (measurement ruler used for gauging), station (meteorology station, rainfall station) and site (pond water collection point).

The implemented model is used as a data source for selected algorithms and tools.[2] Yet, it is worth noting that several problems have been encountered:

- irregular temporality: investigations for data production are irregularly carried out, while collection missions are not jointly led between the different teams;
- the data volume: the available data sets are too small and do not guarantee right interpretation;
- the data format: following the project adopted protocols, data are not collected data with the same format. Thus the low quality of this data constitutes a strong constraint in the choice of the algorithms.

[2] This data model is experienced through the implementation of a PostgreSQL data warehouse fed from different sources using Talend ETL (Extract Transform Load). ETL is used to retrieve data from a source, to process those data (cleaning, formatting or structural change) and to load data in another file or database.

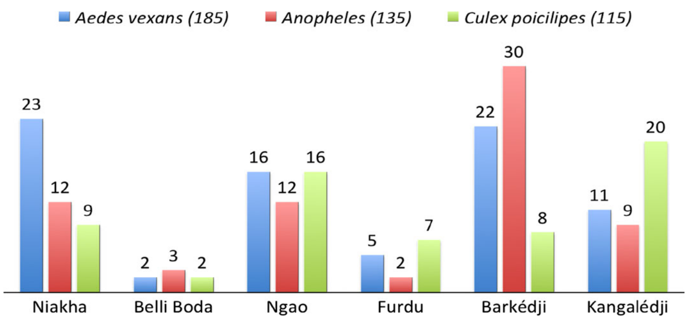

Fig. 4 Distribution of vectors in studied area

The reporting tests made, with the Pentaho tools, on our data warehouse confirm the completeness of the model.

4.2 Decision Tree

The "W-Random Tree" algorithm[3] is applied to describe the vector diversity of the locality related to the studied area ponds.[4] The focus has particularly been put on the potential vectors of RVF (*Aedes vexans*—blue bar, *Anopheles*—red bar, *Culex poicilipes*—green bar). This analysis (Fig. 4)[5] of the vector diversity by locality essentially confirms:

- the high presence of *Aedes vexans* in Niakha, Ngao and Barkédji;
- the high presence of *Culex poicilipes* in Ngao and Kangalédji;
- very few *Anopheles* in Belli Boda and Furdu.
  It is therefore worth identifying the similarity points among these geographical areas. Various studies have shown that the Sahelian climate characterizes all these localities. Then, it became interesting to focus on their water parameters using the space–time patterns.

4.3 Spatio-Temporal Patterns

The data warehouse is transformed in a database of sequences with the algorithm proposed by Alatrista et al. (2011).

This method first retrieves the list of all frequent occurrences (or frequent item) in the database based on the minimum support. A frequent item means that a pattern of greater length was found. For our experimentation, we use data sets for climate elements (temperature, wind, rainfall…) and water quality parameters of the ponds

[3] Tests were performed with RapidMiner to which we integrated the Weka algorithms.

[4] This graphic format was chosen to facilitate interpretation by trade experts.

[5] The number in parenthesis represents the total number of captured vectors in the studied area.

Table 1 Patterns "Ponds"

Patterns	Support
(EC:[42.45;51.30])(TDS:[22.65;27.45])(Temp:>29.05)(pH:[5.36;5.54])	1 (5/5)
(pH:[5.54;5.66])(Temp:[28.35;29.05])	1 (5/5)
(pH:[5.36;5.54])(TDS:[22.65;27.45])(EC:[42.45;51.30])(Temp:[28.35;29.05])	1 (5/5)

(acidity content, dissolved solids, temperature…).[6] Two analyses have been carried out: pond gathered data and site gathered data. For the first data set (ponds), the time dimension is a great deal varied, with a 0.9 minimal support (at least 90 % of the ponds meet the supplied sequences requirements), 21,131 frequent sequences are obtained (Table 1).

These obtained patterns make it possible to certify that the water quality of the five (5) study ponds of the Ferlo zone (Fig. 1) is very similar. Indeed, these patterns have a 100 % support. This means that all the ponds analysed during similar periods have exactly the same behaviour. It is so equally deemed more relevant to confront the variants of the ponds with those of the meteorology-climate environment to control the impact of the latter on the pond water quality. Thus, studying the same data on the basis of the rainfall stations and posts, we obtained very diversified supports, varying between 44 % (temperature, pH, dissolved solids) and 94 % (temperature; Table 2). This data set includes more geographical areas (localities) but fewer dates. The results consist of 67,912 for a 0.4 minimal support (at least 40 % of the localities meet the sequence requirements).

These latest results provide no coherent interpretation. Indeed, this analysis brings no answer based on possible correlations among the pond water characteristics and those supplied by meteorological automatized stations. The low time density could justify this. The meteorological data confirm that the rural community climate in these study areas is very similar. It should then be fit to focus on other parameters such as soil quality to be able to justify the study area vector diversity.

5 Conclusion and Perspectives

As a reminder, the present research work aims at proposing a decision environment describing the interactions between the environment indicators and the spreading of the RVF. This paper has presented the data model and a few extracts of the patterns obtained on the pond water quality. The proposed data model makes it possible to (1) identify the different analysis axes; (2) correlate the different quality measurement indicators (Trujillo et al. 2003); (3) propose views corresponding to the expectations of the different involved disciplines (Golfarelli et al. 2002). Moreover, the data warehouse is also used as a data source for data mining tools and the algorithm of spatio-temporal patterns. These experimentations have made it

[6] TDS: Total Dissolved Solids, EC: Electrical Conductivity, Temp: Temperature, pH : potential Hydrogen.

Table 2 Patterns "Sites"

Patterns	Support
(Temp:[28.35;29.05])	0.94 (15/16)
(pH:≤5.45)(EC:[42.45;51.30])(TDS:[22.65;27.45])	0.75 (12/16)
(pH:[5.45;5.62])(Temp:[28.35;29.05])(TDS:≤22.65)	0.44 (7/16)

possible to confirm that the Ferlo area ponds present the same water quality characteristics. In spite of the weak data set, the realized experiments confirm this integration approach of the spatio-temporal attributes to understand and have control over the RVF health risk. For the continuation of this research work, it is important to be able to confront the achieved results with virological data. In addition, we are working on identifying neighbourhoods of geo-referenced objects following polar and/or Euclidean approach. Thus, the problem will be to identify all correlations between the geographical environmental parameters and the entomological and health data taking into account basic characters such as time and space. Furthermore, observing the livestock transhumance should provide additional elements for space–time projections of the RVF appearance and spreading. In data mining, data make up the information raw material upon which the decisions should be made. So, issues related to the data quality are the cause of the main failures in setting up a decision system. To go thoroughly into the present approach, it would be appropriate to propose a method for rebuilding the missing data, taking into account their proportion and their type.

Acknowledgments Hugo Alatrista Salas, Lilia Beharrou, Centre National d'Etudes Spatiales (CNES), Institut Pasteur de Dakar (IPD), Centre de Suivi Ecologique (CSE).

References

Agrawal R, Srikant R (1995) Mining sequential patterns. In: Yu PS, Chen ALP (eds) Proceedings of the eleventh international conference on data engineering (ICDE), Taipei, Taiwan, pp 3–14. IEEE Computer Society

Alatrista H, Cernesson F, Bringay S, Azé J, Flouvat F, Semaloui N, Teisseire M (2011) Recherche de séquences spatio-temporelles peu contredites dans des données hydrologiques. Revue des Nouvelles Technologies de l'Information (RNTI), numéro spécial Qualité des Données et des Connaissances/Evaluation des Méthodes d'Extraction des Connaissances dans les Données, vol RNTI-E-22, pp 165–188, ISBN: 978-2-70568-286-6

Diagne FF (1992) In: Etude de la Fièvre de la Vallée du Rift chez les ruminants domestiques au Sénégal: enquêtes sérologiques dans la vallée du Fleuve. le Ferlo et la Casamance, PhD, FMPOS/UCAD, pp 29–34

Diallo M (1995) Dynamique comparée des populations de Culicidae à Kédougou (zone soudano-guinéenne) et à Barkédji (zone de savane sahélienne): conséquences dans la transmission des arbovirus. Université Cheikh Anta Diop de Dakar, DEA en Biologie Animale

Diallo M, Lochouarn L, Ba K, Sall A, Mondo M, Girault L, Mathiot C (2000) First isolation of the Rift Valley fever virus from *Culex poicilipes* (Diptera: Culicidae) in nature. Am J Trop Med Hyg 6(62):702–704

Fabrègue M, Braud A, Bringay S, Le Ber F, Teisseire M (2012) Extraction de motifs spatio-temporels à différentes échelles avec gestion de relations spatiales qualitatives, Infosid 2012, Montpellier

Flick R, Bouloy M (2005) Rift valley fever virus. Curr Mol Med 5:827–834

Golfarelli M, Maio D, Rizzi S (1998) The dimensional fact model: a conceptual model for data warehouses. Int J Coop Inf Syst 7(2–3):215–247

Golfarelli M, Rizzi S, Saltarelli E (2002) WAND: a case tool for workload based design of a data mart. 10th National convention on systems evolution for data bases, pp 422–426

Greboval M (2003) Facteurs environnementaux influençant la dynamique des vecteurs du virus de la Fièvre de la Vallée du Rift: conséquences pour la modélisation de la maladie. Thèse de Doctorat Vétérinaire, ENVL - Ecole Nationale Vétérinaire de Lyon

Information on http://www.oie.int/fr/sante-animale-dans-le-monde/maladies-de-la-liste-de-loie-2013 (2013)

Marakas GM (2003) Decision support systems in the 21st century. Prentice Hall, Upper Saddle River

Mortazavi-Asl B, Pinto H, Dayal U (2000) PrefixSpan: mining sequential patterns efficiently by prefix-projected pattern growth. In: Proceedings 17th international conference on data engineering, pp 215–224

Ndione JA, Besancenot J, Lacaux J, Sabatier P (2003) Environnement et épidémiologie de la Fièvre de la Vallée du Rift (FVR) dans le bassin inférieur du fleuve Sénégal. Environnement, Risques et Santé 3(2):176–182

Ndione JA, Lacaux JP, Tourre Y, Vignolles C, Fontanaz D, Lafaye M (2009) In: Mares temporaires et risques sanitaires au Ferlo : contribution de la télédétection pour l'étude de la fièvre de la vallée du Rift entre août 2003 et janvier 2004. Secheresse 20(1):153-160

OMS, La fièvre de la Vallée du Rift. Aide-Mémoire 207, Nations Unies (2010). http://www.who.int/mediacentre/factsheets/fs207/fr/index.html

Préhaud C, Bouloy M (1997) La fièvre de la vallée du rift, un modèle d'étude des fièvres hémorragiques virales. Technical report, Institut Pasteur

Ravat F, Teste O, Zurfluh G (2010) Algèbre OLAP et langage graphique. CoRR abs/1005.0213

Tourre YM, Lacaux JP, Vignolles C, Ndione JA, Lafaye M (2008) Mapping of zones potentially occupied by Aedes vexans and Culex poicilipes mosquitoes, the main vectors of Rift Valley fever in Senegal. Geospatial Health 3(1):69–79

Tran A, Biteau-Coroller F, Guis H, Roger F (2005) Modélisation des maladies vectorielles. Epidémiol et santé anim 47:35–51

Trujillo JC, Luján-Mora S, et Song I (2003) Applying UML for designing multidimensional databases and OLAP applications. In: Siau K (ed) Advanced topics in database research, vol 2. Idea Group Publishing, Hershey, PA, pp 13–36

Wehrle P (2009) Modèle multidimensionnel et OLAP sur architecture de grille. PhD thesis, Institut national des sciences appliquées de Lyon