



HAL
open science

Data mining techniques on satellite images for discovery of risk areas

Boukaye Boubacar Traore, Bernard Kamsu-Foguem, Fana Tangara

► **To cite this version:**

Boukaye Boubacar Traore, Bernard Kamsu-Foguem, Fana Tangara. Data mining techniques on satellite images for discovery of risk areas. *Expert Systems with Applications*, 2017, vol. 72, pp. 443-456. 10.1016/j.eswa.2016.10.010 . hal-01527238

HAL Id: hal-01527238

<https://hal.science/hal-01527238>

Submitted on 24 May 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Open Archive Toulouse Archive Ouverte (OATAO)

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in: <http://oatao.univ-toulouse.fr/>
Eprints ID: 17667

To link to this article: DOI:10.1016/j.eswa.2016.10.010

<https://doi.org/10.1016/j.eswa.2016.10.010>

To cite this version:

Kamsu-Foguem, Bernard and Traore, Boukaye Boubacar and Tangara, Fana *Data mining techniques on satellite images for discovery of risk areas*. (2017) *Expert Systems with Applications*, vol. 72. pp. 443-456.
ISSN 0957-4174

Any correspondence concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr

Data mining techniques on satellite images for discovery of risk areas

Boukaye Boubacar Traore ^{a,b}, Bernard Kamsu-Foguem ^{a,*}, Fana Tangara ^b

^a Université de Toulouse, Ecole Nationale d'ingénieurs de Tarbes (ENIT), Laboratoire de Génie de Production (LGP), EA 1905, 47 Avenue d'Azereix, BP 1629, 65016 Tarbes Cedex, France

^b Université des Sciences, des Techniques et des Technologies de Bamako (USTTB), Faculté des Sciences et Techniques (FST), Centre de Calcul, de Modélisation et de Simulation, Colline de Badalabougou, ancien Lycée Bada/a B.P. E28 11 – FAST 223 Bamako, Mali

A B S T R A C T

The high rates of choiera epidemic mortality in less developed countries is a challenge for health facilities to which it is necessary to equip itself with the epidemiological surveillance. To strengthen the capacity of epidemiological surveillance, this paper focuses on remote sensing satellite data processing using data mining methods to discover risk areas of the epidemic disease by connecting the environment, climate and health. These satellite data are combined with field data collected during the same set of periods in order to explain and deduct the causes of the epidemic evolution from one period to another in relation to the environment. The existing technical (algorithms) for processing satellite images are mature and efficient, so the challenge today is to provide the most suitable means allowing the best interpretation of obtained results. For that, we focus on supervised classification algorithm to process a set of satellite images from the same area but on different periods. A novel research methodology (describing pre-treatment, data mining, and post-treatment) is proposed to ensure suitable means for transforming data, generating information and extracting knowledge. This methodology consists of six phases: (1.A) Acquisition of information from the field about epidemic, (1.B) Satellite data acquisition, (2) Selection and transformation of data (Data derived from images), (3) Remote sensing measurements, (4) Discretization of data, (5) Data treatment, and (6) Interpretation of results. The main contributions of the paper are: to establish the nature of links between the environment and the epidemic, and to highlight those risky environments when the public awareness of the problem and the prevention policies are absolutely necessary for mitigation of the propagation and emergence of the epidemic. This will allow national governments, local authorities and the public health officials to effective management according to risk areas. The case study concerns the knowledge discovery in databases related to risk areas of the choiera epidemic in Mopti region, Mali (West Africa). The results generated from data mining associations indicate that the level of the Niger River in the wintering periods and some societal factors have an impact on the variation of choiera epidemic rate in Mopti town. More the river level is high, at 66% the rate of contamination is high.

Keywords:

Data mining
Discretization
Remote sensing
Risk identification
Knowledge

1. Introduction

Choiera remains a major public health problem in developing countries. More than fifty countries in the world suffer choiera outbreaks with an average of 200,000 cases and lethality rate of 4% (Piarroux, 2002). There is no effective vaccine included in the WHO vaccination programs (Fournier, 1996). Mali located in the heart of West Africa is no exception to the rule. Noting that the geographical factor plays a major role in this epidemic, because in 78% of cases people were the bathing takes place in the lakes, streams or rivers according to a study by Dao et al. (2009),

particularly in wintering periods. Indeed, all outbreak areas are located along the Niger River crosses Mali on its length South to the North. The progression of the epidemic has occurred in the flow direction of the West River Country to the North.

The infectious diseases are gaining more ground every year thus increasing the risk of epidemics. They have in common small organisms, often arthropods (virus, mosquitoes, flies, ticks, lice or fleas), which are called vectors because they transmit the infectious agent from one host to another. The distribution area of these vectors is progressively expanding, with climate change opening up some possibilities for the reappearance of disease in previously protected populations. Sometimes new vulnerable regions are thus seriously affected.

To overcome the limitations of traditional techniques to fight against epidemics, it is necessary to enjoy the benefits of space

* Corresponding author.

E-mail addresses: boukaye@gmail.com (B.B. Traore), Bernard.Kamsu-Foguem@enit.fr (B. Kamsu-Foguem), fbtangara@yahoo.fr (F. Tangara).

technology in the fight and the stalking of infectious diseases. The potential advantages of the space technology are numerous:

- Economies: satellites cover a large area without long travel of the field teams.
- Security: the satellites cover dangerous or inaccessible areas,
- Dynamics: decryption of the spatiotemporal dynamics of diseases,
- Monitoring: provision of risk maps and epidemiological surveillance platform for better monitoring and anticipation of epidemics.

Epidemiology turns to space using remote sensing satellite technology through Tele-epidemiology. Tele-epidemiology associates satellite observation data (geographical, climatological, environmental, land use, etc.) with data sets from multidisciplinary areas (health data, temperature or humidity, socioeconomic data, biological readings, etc.) (Marechal, Ribeiro, Lafaye, & Guell, 2008). That allows to analyse the mechanisms of emergence, propagation and transmission of epidemic diseases, firstly, by investigating the tripartite relationships Climate-Environment-Health, and secondly, by updating the links that exist between the epidemic diseases and the environment in which they develop, using remote sensing satellite technology (Lacaux, Tourre, Vignolles, Ndione, & Lafaye, 2007).

Statistical and non-statistical products have been developed to facilitate the work of epidemiologists: quantification of the cases of disease, identification of the factors characteristic of the affected areas, identification of the high risk areas, development of forecasting models, management of prevention programmes, etc. Remote sensing has a rising significance in the current information and communication society.

The purpose of this work is to demonstrate the link between the geographical factor of the Niger River and the spread of the cholera epidemic in affected communities using satellite images and field data processed by data mining techniques. Thus this work contributes to strengthening epidemiological surveillance policy that focuses on a particular area of Mali and generally in the countries concerned about the production of images representing maps of risk areas of epidemics.

The processing of remote sensing of satellite data by data mining techniques presents a major challenge in the health sector, especially in epidemiology. Data mining is a process for automatic information discovery in large databases (Tan, Steinbach, & Kumar, 2005). Data mining technique allow to extract previously unknown patterns that can generate interesting knowledge (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), such as risk areas (through risk maps) from huge databases of satellite images. These patterns can be seen as a kind of summary of the input data, and will be used in further analysis to predict other cases, for instance expand the production of risk maps of epidemic on the whole national territory.

The three main phases of data mining are pre-treatment, treatment, and post-treatment. Most of the work of data mining lies in the post and pre-processing data: 80% post and pre-processing data, and 20% data mining techniques (application of algorithms) (Qin, Zhang, Zhu, Zhang, & Zhang, 2009; Zhang, Chengqi, & Qiang, 2003). Most data mining algorithms do not know how to handle numerical data but rather categorical data. To solve this problem, it is necessary to discretize numerical data. The discretization, is the process to convert numerical values into categorical values and it is positioned in the pre-processing data phase. The discretization is one of the current problems of data mining because it has much more impact on the results and therefore easily leads to misinterpretations.

The structure of the rest of the paper is as follow. We present theoretical background in Section 2, with an additional litera-

ture review in Section 3. Our novel methodology and techniques adopted are presented in Section 4. Section 5 presents experimental results and Section 6 gives some discussions and finally Section 7 provides concluding remarks.

2. Theoretical background

Remote sensing is a method to observe the earth surface or the atmosphere from space using satellite systems recording images which can be interpreted to yield useful information. Remote sensing supports the mapping of various territories (terrestrial, marine and coastal ecosystems) and it has many applications in mapping land use and cover, agriculture, soils mapping, military observation, deforestation, urban growth, etc. Satellite remote sensing imagery provides an effective way to collect critical information. Nowadays, the challenge is to develop technical solutions based on remote sensing satellite data to identify, predict and study risk areas that can benefit to all sectors particularly the health sector.

There are many satellites positioned around the globe: some are geostationary above a region, and some are orbiting around the Earth, but all are sending a non-stop stream of data to the surface. There is a considerable growth of stored data in databases that hide important knowledge and deserves to find and use in order to help decision-makers (Alatrística-Salas et al., 2015). In this article, we present an approach to determine risk areas of epidemic crisis using data mining techniques (search for patterns or regularities) in satellite data.

Satellite data consists of data of Earth or other planets collected by satellites. We focus on satellite data on earth from the Landsat program that is the longest running enterprise for acquisition of satellite imagery of Earth and because their images are free and of good quality. On July 23, 1972, the Earth Resources Technology Satellite was launched. This was eventually renamed to Landsat. The most recent, Landsat 8, was launched on February 11, 2013 (Short, 1982). The instruments on the Landsat satellites have acquired millions of images and can be viewed through the United States Geological Survey (USGS: Earth Explorer, Landsat Look Viewer and GloVis) website. Data mining techniques are applied to a data base made up of satellite data for the identification and study of epidemic risk areas.

The data mining process usually consists of three phases or steps: (1) pre-processing or data preparation; (2) modelling and validation; and (3) post-processing or deployment (Compieta, Di Martino, Bertolotto, Ferrucci, & Kechadi, 2007). During the first phase, the data may need some cleaning and transformation according to some constraints imposed by some tools, algorithms, or users. One has to make sure that the data are free of noise and some transformations are needed for visualizing very large data sets. The second phase consists of choosing or building a model that better reflects the application behaviour. In other words, once a model is chosen or developed, it should be evaluated in terms of its efficiency and accuracy of its predictive results. Finally, the third step consists of using the model, evaluated and validated in the second phase, to effectively study the application behaviour. Usually, the model output requires some "post-processing" in order to exploit it.

In this work, we focus on advanced databases such as satellite images databases which can be constituted by the three most common types of satellite images: Visible imagery, Infrared imagery, and Water vapor imagery (Fugazzi & Spokane NWS, 2008). Therefore, it is a question to produce risk maps and spread of epidemics, for that we propose a methodology which combines data on health systems and data from the processing of satellite images. This proposed methodology is focused on the discretization phase, supervised classification and finally good interpretations by associating field data.

3. State of the art

Satellite remote sensing in epidemiology (Tele-epidemiology) is an interesting perspective in the mapping and modelling of epidemic and epidemic risk factors (Nnadi, Nimzing, Mark, & Onyedibe, 2011). We use remote sensing data, to understand cholera epidemic and its factors namely water resources and field data to discover the risk areas of cholera epidemic. Often epidemic study areas are larger than a single satellite scene or may be discontinuous, particularly in flooding or tornado hazard events. Liu and Hodgson (2016) proposed a spatial optimization model to solve the large area satellite image acquisition planning problem in the context of hazard emergency response.

From critical review of research literature (Mabaso & Ndlovu, 2012), one of actual challenge about epidemics diseases in general, was the development of operational early warning systems taking into account the impacts of climate-driven threats. Therefore, many applications were developed, Kansakar and Hossain (2016) presented a general review of applications based on remote sensing image through many themes such as land cover, land use mapping, carbon biomass assessment, food security, disaster management, water resources, ocean management and health and air quality. To use an abundance of environmental data for human health, Liu et al. (2015) developed the Epidemiological Applications of Spatial Technologies (EASTWeb) for mosquito-borne diseases (e.g. Malaria, Rift Valley fever) based on data like temperature, precipitation, spectral indices, and evapotranspiration. This application does not process really satellite images but it is more focused on field data.

In the purpose to contrai, prevent and eradication the arthropod vectors, depending on environmental factors, Jamison et al. (2015) use multidisciplinary research incorporating climate, geographic information systems, remote sensing, and ecology. The tripartite relationships Environment-Climates-health are central to the mapping of risk areas of epidemic. Ecological factors with Rift Valley were associated to produce Rift Valley fever (RVF) risk map in Kenya (Mosomtai et al., 2016) based on remote sensing data. The main ecological variables are: Wet soil fluxes measured, Normalized Difference Vegetation Index (NOVI) and Evapotranspiration (ET). The map generated seems more to a Geographic Information System (GIS) application with a superposition of layers: the individual maps of three most significant variables are produced separately, and the final risk mapping result are derived from amalgamating these maps. Another GIS application was developed from the strong climate-fascioliasis outbreaks relationship. Halimi, Farajzadeh, Delavari, and Arbabi (2015) produce risk map of fascioliasis outbreaks in Iran using Ollerenshaw's fascioliasis risk index integrating a GIS. They find four fascioliasis outbreak risk classes. This study focuses on the relationships between climate factor and fascioliasis epidemic, but do not take into account satellite remote sensing image process.

In the same domain of télé-epidemiology, Xu et al. (2014) establish the impact of temperature on childhood pneumonia in Brisbane, Australia by using remote sensing data combined to data from emergency department visits (EDVs). Certainly, it may be easy to use the quasi-Poisson generalized linear model combined with a distributed lag non-linear model, but rather suitable for discrete variables, there is a need to have a dependent variable with a Poisson distribution and it is difficult to check the almost systematic dispersion problem (Takahashi & Kurosawa, 2016).

To predict cholera epidemics in the Bengal Delta region using satellite remote sensing, Jutta, Akanda, and Islam (2012) studied the variability on space-time of chlorophyll. The bacterium *Vibrio cholerae* cannot be measured from space directly. The cholera factor maintained is the Chlorophyll, which is estimated from satellite

images by measuring the pigment part (chlorophyll) existing in plankton. This approach is limited because the observation of pigment part in all plankton does not mean at 100% the presence of chlorophyll. It should be specified the period and how many percent the chlorophyll is involved in all cases of cholera epidemic in region. It would be interesting to integrate field data to improve the precision and accuracy risk areas of arthropod epidemic. What is certain is that Cholera is an acute diarrhoeal infection caused by ingestion of water or food contaminated with the bacterium *Vibrio cholerae*. In the most countries contaminated, Cholera is a seasonal epidemic usually during the rainy season, which is the case of Mopti (Mali).

The accessibility of the massive quantity of health data (unstructured and multi-structured data) presents considerable opportunities and challenges for the real-time tracking of diseases, predicting disease outbreaks, and detecting pathogens and causes of emergence (Asokan & Asokan, 2015). Health human resources have much to gain and a great deal to understand from the deployment of geospatial analysis and exploitation of generally obtainable spatial data sources for epidemiological modelling and sustainable health monitoring system (Malone John, Yang, Leonardo, & Zhou, 2010). There is a strong need of new means for interdisciplinary data analysis, modelling and selection for multidimensional, complex settings and this may facilitate innovative insights and responses to complex environmental problems (Lausch, Schmidt, & Tischendorf, 2015). Earth observation data value addition process for intelligent geo-information products implies four phases: (i) Earth observation data acquisition, (ii) data processing (geo-information data mining and products), (iii) knowledge transformation for specific application, (iv) decision support and policy formulation.

The developed methodologies can contribute to the processing of land mapping over large areas, with data mining of remotely sensed imagery (Vieira et al., 2012). In order to improve the prediction accuracy of identification of landslide-related factors using a geographic information system, data mining models can be used in conjunction with other models (e.g. artificial neural network and Fuzzy logic) (Song et al., 2012). In such circumstances, the data mining techniques can be engaged to examine topographic and vegetative features for the verification of landslides implied by specific climatic conditions (e.g. heavy rainfall) on the regions of a country (Tsai, Lai, Chen, & Lin, 2013). Regardless of the technique employed, there are some open research issues in spatial data mining techniques such as validity testing, the selection of relevant characteristics, interpretation of models, and treatment of multi-dimensional data (Arentze, 2009). Furthermore, various theories (e.g. Probability theory, Dempster-Shafer theory, Possibility theory and Rough set theory) can be employed for dealing with uncertainties (e.g. vague/ambiguous or imprecise/inaccurate/incomplete information) in spatial data analysis and geographic information systems (Tavana, Liu, Elmore, Petry, & Bourgeois, 2016). Particularly, the extraction of complex models from multiresolution remote sensing images has to be enhanced by considering hierarchical strategies (e.g. involving segmentation and clustering) to facilitate the possible discovery of more complex patterns with a higher semantic level (Kurtz, Passat, Gançarski, & Puissant, 2012).

We are focusing on the pre-treatment, and one of the challenges of pre-treatment is the discretization of the input data (Pitarch et al., 2015). Bad achievement of this step necessarily impacts the final results. Misunderstanding or misuse of the discretization in pre-treatment has the potential to interfere with data mining and post-treatment. Our discretization approach occurs on spatial data specifically on satellite images. After a better process of discretization on satellite images, we apply data mining algorithms to data base made up of satellite images.

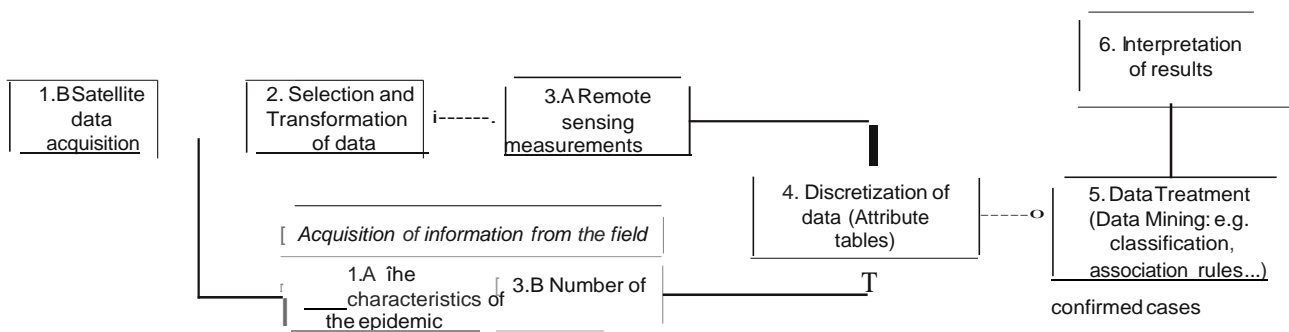


Fig. 1. Process to discover risk areas of epidemic.

4. Methodology and approaches adopted

The methodology consists of six phases as described by the figure below (Fig. 1): the first phase starts with Acquisition of information from the field about epidemic and satellite data acquisition, the second phase concerns the selection and transformation of the useful data, the third phase describes remote sensing measurements (e.g. the Normalized Difference Vegetation Index (NOVI), a function measuring the reflectance of the planet in both red and near-infrared bands) and the collection of the number of confirmed cases from the field, the fourth phase concerns the preparation of ail data by discretization techniques to making them available for the fifth phase concerning data treatment with data mining algorithms and finally the sixth phase is related to the interpretation of the results obtained by producing epidemic risk maps.

4.1. A acquisition of information [rom the field about epidemic

The acquisition of field information is done through the health facilities and field specialists.

There are two major types of information.

- **Number of confirmed cases (Epidemiological data of the field)**: it is question to get the number of confirmed cases stored in databases from one location to another and from one period to another from health structures (health data). These data allow seeing the evolution of the epidemic in different places and at different times.
- **The characteristics of the epidemic**: epidemiological diseases such as the epidemic of Ebola virus disease, Malaria, Choiera, Lyme disease ... they ail have their special characteristics but they have in common certain transmission vectors. The vector is the agent carrying viruses and bacteria from a given epidemic risk source to infect the population: birds, animals, ticks, mosquitoes, flies, worms.... . So, it needs to determine the links between vectors and supportive environments for the emergence of these vectors; these environments are constantly expanding due to climate change, harmful human activities (e.g. inappropriate land use practices). The target sensitive environments (water, vegetation , etc.) include some potential risk areas that will be detected by satellites.

It is important to analyze the emergence, propagation, and transmission of infectious diseases and establish the relations that exist between the infectious diseases and the environment, ail using space technology. Next step is the acquisition of satellite data.

4.1. B satellite data acquisition

There are many satellites positioned around the globe: some are geostationary above a region, and some are orbiting around



Fig. 2. LANDSAT 8 Satellite Sensor (15 m).

the Earth, but ail are sending a non-stop stream of data to the surface: Landsat (Fig. 2), SPOT, Ikonos, ASTER, AVIRIS, AVHRR, Quickbird, RadarSat, TMS, DTED or WorldView

The images of some satellites are free, this is the case of Landsat (Fig. 2), which have the advantage of being available free of charge by the United States Geological Survey (USGS). Once downloaded to the USGS website <http://earthexplorer.usgs.gov/>, the images are provided in the form of several tiff files, each file describes one band. Landsat 7 images have 8 channels or bands, Landsat 8 have 11 (Wulder et al., 2012). Each band is composed of two attributes: the first one describes the division name's abbreviation of the infrared band and the second one represents the wavelength (Table 1).

This information related to bands helps us make a better choice to select the useful bands to the calculation of remote sensing indices. Subsequently, we describe the selection process and data transformation.

4.2. Selection and transformation of data (data derived from images)

The purpose of this step is to identify and select relevant and important data involved in the research of environmental factors conducive to the epidemic. To make the selection of useful bands, it is necessary to have:

- information on the origin and types of environments conducive to the emergence and spread of the epidemic. This information is provided by specialists.
- satellite data of the study field.

From the viewpoint of rangeland management, the aim of a remote sensing method is to extract information that is directly related to a management question or to create other data layers that are co-related to what you are eventually interested in. Some approaches have been proposed to analyze and make information

Table 1
OLI and TIRS spectral bands (LANDSAT 8) Thermal InfraRed (TIR).

Band	Wavelength	Useful for mapping
Band 1-Coastal aerosol	0.43-0.45	coastal and aerosol studies
Band 2-Blue	0.45-0.51	Bathymetric mapping, distinguishing soil from vegetation and deciduous from coniferous vegetation
Band 3-Green	0.53-0.59	Emphasizes peak vegetation, which is useful for assessing plant vigor
Band 4-Red	0.64-0.67	Discriminates vegetation slopes
Band 5-Near Infrared (NIR)	0.85-0.88	Emphasizes biomass content and shorelines
Band 6-SWIR 1	1.57-1.65	Discriminates moisture content of soil and vegetation; penetrates thin clouds
Band 7-SWIR 2	2.11-2.29	Improved moisture content of soil and vegetation and thin cloud penetration
Band 8-Panchromatic	0.50-0.68	15 meter resolution, sharper image definition
Band 9-Cirrus	1.36-1.38	Improved detection of cirrus cloud contamination
Band 10-Thermal Infrared (TIRS) 1	10.60-11.19	100 meter resolution, thermal mapping and estimated soil moisture
Band 11-Thermal Infrared (TIRS) 2	11.50-12.51	100 meter resolution, Improved thermal mapping and estimated soil moisture

Table 2
Definition of spectral indexes. Bands are designated in the formulas as R (Red), B (Blues), G (Green), NIR (Near-Infrared), and RE (Red-Edge).

Index	Name	Citation	Formula
NOVI	Normalized Difference Vegetation Index	(Tucker, 1979)	$(\text{NIR}-\text{R})/(\text{NIR}+\text{R})$
NGRDI	Normalized Green Red Difference Index	(Tucker, 1979)	$(\text{G}-\text{R})/(\text{G}+\text{R})$
GNDVI	Green Normalized Difference Vegetation Index	(Mages et al., 2004)	$(\text{NIR}-\text{G})/(\text{NIR}+\text{G})$
SR	Simple Ratio	(Tucker & Sellers, 1986)	NIR/R
SAVI	Soil-Adjusted Vegetation Index (L=0.5)	(Huete, 1988)	$[(\text{NIR}-\text{R})/(\text{NIR}+\text{R}+\text{L})]\{1+\text{L}\}$
EVI	Enhanced Vegetation Index	(Huete et al., 2002)	$2.5\{(\text{NIR}-\text{R})/(\text{NIR}+6\text{R}-7.5\text{B}+1)\}$
TVI	Triangular Vegetation Index	(Broge & Leblanc, 2000)	$0.5[120(\text{NIR}-\text{G})-200\{\text{R}-\text{G}\}]$
VAR!	Visible Atmospherically Resistant Index	(Gitelson, Kaufman, Stark, & Rundquist, 2002)	$(\text{G}-\text{R})/(\text{G}+\text{R}-\text{B})$
NDREI	Normalized Difference Red Edge Index	(Gitelson & Merzlyak, 1994)	$(\text{RE}-\text{R})/(\text{RE}+\text{R})$

using satellite images such as Vegetation Indices, Spatial Prediction Techniques (Hengl, Heuvelink, & Stein, 2004), Thermal Remote Sensing Products (Tahakata et al., 2009), Biophysical Parameter Estimates, Classification Techniques (Alrababah et al., 2006) ... (the most known is Vegetation Indices) (Anderson, Hanson, & Haas, 1993).

Several indices have been developed specifically for the analysis of remote sensing data but in literature, the vegetation indices are the most used. Based on information provided by experts in the field, the choice of the indices is made and thereafter, followed by the selection data involved in the calculations of these indices.

These vegetation indices use the principle of differentiation between areas of vegetation with their sub-zones (e.g. forest and arid zones) and water zones (e.g. Lake, river, and pond). In the literature, different vegetation indices (Vis) have been developed (Table 2, Prabhakara, Hively, & McCarty, 2015): Rouse, Haas, Schell, and Oeering (1994) proposed the Normalized Difference Vegetation Index (NOVI) to estimate vigour of plants, Huete presented a Soil-Adjusted Vegetation Index (SAVI) and, Roujean and Breon developed a Renormalized Difference Vegetation Index (ROVI). A Triangular Vegetation Index (TVI) was presented by Broge & Leblanc, the Three-band Gradient Difference Vegetation Index (TGOVI) was established by Tan et al. (2005) Normalized Difference Pond Index (NOPI) and the Normalized Difference Turbidity Index (NOTI) are presented by Lacaux et al. (2007) and Bicout, Vautrin, Vignolles, and Sabatier (2015).

Among vegetation indices approach, the NOVI is widely used (Hobbs, 1995; Anderson et al., 1993). NOVI is constructed from of the Red (R) and Near Infrared (NIR) and highlights the difference between the red visible band and the near infrared visible band by using the standard equation: $[(\text{NIR} - \text{RED})/(\text{NIR} + \text{RED})]$ (Tucker, 1979). This index is sensitive to the strength and quantity of vegetation and the values range from -1 to + 1. Negative values correspond to surfaces other than vegetation zones such as snow, water or clouds for which the reflectance in the red is greater than the near infrared. For bare soil, the reflectance is roughly the same magnitude in the Red and Near Infrared, the NOVI values are close to 0. Vegetation zones have positive NOVI values generally between 0.1 and 0.7. The values beyond 0.7 correspond to dense

vegetation. When a time series of images is observed for the same area, we can deduce the stress state of vegetation for different values of NOVI: in wintering or during droughts, values will be lower than in summer or in times of rainy seasons.

To calculate NOVI from Landsat image, for example, you combine the first three bands; you get a conventional optical image (RGB) and after join the band of near infrared. The NOVI calculation generated an image file or layer, and this could be realized through some software image processing programs (e.g., ENVI, IDRISI, Matlab, R, EROAS Imagine) or Geographic Information System (GIS) programs that can manipulate raster calculations (e.g., ESRI ArcGIS, GRASS).

4.3. Remote sensing measurements

Having opted for a remote sensing measurement (e.g. NOVI), we get the corresponding images to remote sensing measurements. There is a need to extract the corresponding attribute tables from composite data. For this, there are several methods (Gonzalez et al., 2015) including the use of satellite image processing software (e.g. Envi software).

The attribute table of each remote sensing data are continuous values, for each band values are between 0 and 65,535. However, the majority of data mining algorithms have difficulty in processing continuous values. Therefore, the continuous attributes must be divided into intervals to accommodate the decision rules. The attribute data of the satellite images and the evolving of epidemic databases on the field must be prepared by discretization techniques for the next step.

4.4. Discretization of data

In this step, data is transformed or consolidated into forms appropriate for data mining process. We have two types of data to discretize: epidemical databases of a number of confirmed cases from Health structures and remote sensing measurements data (e.g. NOVI image file or layer). When an image is created through remotely sensed data, it needs to go through some form of valida-

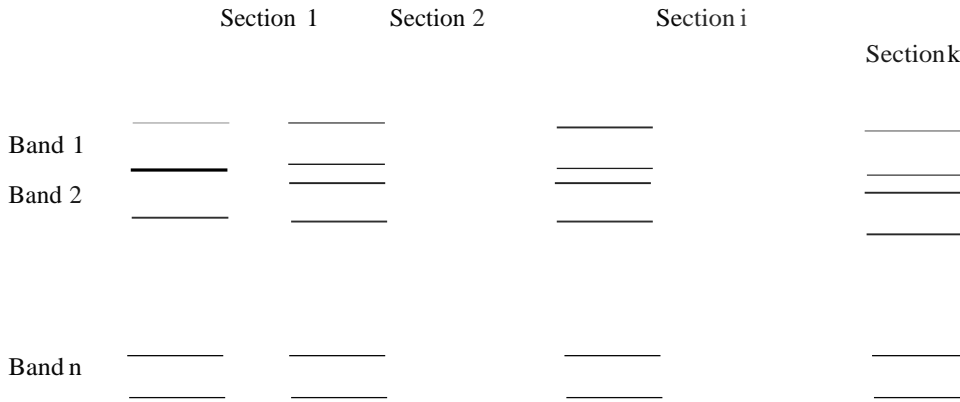


Fig. 3. The outcome of discretization on a sample (Xie et al., 2016).

tion procedure using observational and or sampling techniques to increase the confidence in the final image.

Discretization is the process of converting numerical values into categorical values: the continuous attributes must be divided into multiple intervals to adapt to the data mining algorithm. To discretize remote sensing measurements data, the discretization algorithm based on information entropy proposed by Xie et al. (2016) is adopted. They define three steps to discretize sensing remote data: (i) Making a Decision table, (ii) Breakpoint and equivalence class, (iii) finally run the discretization algorithm (KNN) and implemented in Matlab software. Discretisation algorithm allows subdividing each band into several intervals (sections) (Fig. 3). It is important to resize images obtained with the Nearest Neighbour resampling method. This is the ideal technique if the new image must undergo a classification since it then consists of the original pixel brightness and the simple rearranged in a position to give correct image geometry (Richards, 1999). The process assigns a value to each corrected pixel from the nearest uncorrected pixel. The gains of nearest neighbour include simplicity and the ability to preserve original values in the unaltered scene.

After obtaining discretized data, next step consists in establishing the correlation between the discretized data from remote sensing images and epidemical database of a number of confirmed cases to discovered risk areas of epidemics.

4.5. Data treatment

The method of data treatment is supervised classification. Classification models predict categorical class labels. The classification is one of the main tasks of data mining which is an inductive, iterative and interactive process to discover in large databases of valid data models, new, useful and understandable (Han, Kamber, & Pei, 2011).

The data for classification are divided into two sets: a data set for training sites and a data set for test sites to consolidate the pattern found. We will use this pattern for classifying new areas into two classes: epidemic risk or not an epidemic risk. Supervised classification methods are used to generate a map with each pixel assigned to epidemic risk class or not epidemic risk class based on its multispectral composition. The classes are determined based on the spectral composition of training sites defined by the user. In this work, Digital image classification and analysis was performed using Envi image processing software.

Description of the classification algorithm used: Maximum Likelihood (Abkar, Sharji, & Mulder, 2000) (Richards, 1999) (Ahmad & Quegan, 2012)

We used the most common supervised classification method used with remote sensing image data Maximum likelihood classification based on general approach Bayes classification, which states that a posteriori distribution $P(i|w)$, i.e., the probability that

a pixel with feature vector w belongs to class i , is given by:

$$P(i|w) = \frac{P(w|i)P(i)}{P(w)} \quad (1)$$

where $P(w|i)$ is the likelihood function, $P(i)$ is the a priori information, i.e., the probability that class i occurs in the study area and $P(w)$ is the probability that w is observed, which can be written as:

$$P(w) = \sum_{i=1}^M P(w|i)P(i) \quad (2)$$

where M is the number of classes. $P(w)$ is often treated as a normalisation constant to ensure $\sum_{i=1}^M P(i|w)$ sums to 1. Pixel x is assigned to class i by the rule:

$$x \in i \text{ if } P(i|w) > P(j|w) \text{ for all } j \neq i \quad (3)$$

ML often assumes that the distribution of the data within a given class i obeys a multivariate Gaussian distribution. It is then convenient to define the log likelihood (or discriminant function):

$$g_i(w) = \ln P(i|w) = - \frac{1}{2} (w - \mu_i)' \Sigma_i^{-1} (w - \mu_i) - \frac{1}{2} \ln |\Sigma_i| \quad (4)$$

Since log is a monotonic function, Eq. (3) is equivalent to:

$$x \in i \text{ if } g_i(w) > g_j(w) \text{ for all } j \neq i \quad (5)$$

Each pixel is assigned to the class with the highest likelihood or labelled as unclassified if the probability values are all below a threshold set by the user. The general procedures in ML are as follows:

1. The number of land cover types within the study area is determined.
2. The training pixels for each of the desired classes are chosen using land cover information for the study area. For this purpose, the Jeffries-Matusita (JM) distance can be used to measure class separability of the chosen training pixels. For normally distributed classes, the JM separability measure for two classes, J_{ij} , is defined as follows:

$$J_{ij} = \frac{1}{2} \left(1 - \exp(-2a) \right) \quad (6)$$

where a is the Bhattacharyya distance and is given by:

$$a = \frac{1}{2} (\mu_i - \mu_j)' \Sigma^{-1} (\mu_i - \mu_j) + \frac{1}{2} \ln \left(\frac{|\Sigma_i| + |\Sigma_j|}{2} \right) \quad (7)$$

J_{ij} ranges from 0 to 2.0, where $J_{ij} > 1.9$ indicates good separability of classes, moderate separability for $1.0 < J_{ij} < 1.9$ and poor separability for $J_{ij} < 1.0$.



Fig. 4. Study area: Mopti region (Mopti department in red band).

3. The training pixels are then used to estimate the mean vector and covariance matrix of each class.
4. Finally, every pixel in the image is classified into one of the desired land cover types or labelled as unknown.

In ML classification, each class is enclosed in a region in multispectral space where its discriminant function is larger than that of all other classes. These class regions are separated by decision boundaries, where, the decision boundary between class i and j occurs when:

$$g_i(\mathbf{w}) = g_j(\mathbf{w}) \quad (8)$$

For multivariate normal distributions, this becomes:

$$-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{w} - \boldsymbol{\mu}_i) - \ln(2\pi r_i) - \ln(|\mathbf{C}_d|) - \frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{w} - \boldsymbol{\mu}_j) - \ln(2\pi r_j) - \ln(|\mathbf{C}_d|) = 0 \quad (9)$$

which can be written as:

$$-\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{w} - \boldsymbol{\mu}_i) - \ln(|\mathbf{C}_d|) + \frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{w} - \boldsymbol{\mu}_j) - \ln(|\mathbf{C}_d|) = 0 \quad (10)$$

This is a quadratic function in N dimensions. Hence, if we consider only two classes, the decision boundaries are conic sections (i.e. parabolas, circles, ellipses or hyperbolas).

4.6. Interpretation of results

Finally, the last step is to interpret maps epidemic risk which must be made available to health structures or to the proper authorities to better use.

After processing remote sensing data by the supervised classification method, we can obtain detailed accuracy by class on the generated maps. These measurements of accuracy give the success rate in percentage of the applied classification. Indeed, it is a set of maps from different periods (monthly or annual data) that are produced. During the same period of considered satellite images, the field data (in tabular form) of collected various cases of contamination of the epidemic (monthly or annual data) are also available.

From the results provided by classification and the field data, it can be useful to make a comparative analysis allowing to:

- Better explain the observed trends from field indicators,
- Specify the nature of the link between the environment and the epidemic,
- Clarify the impact of environmental changes on the epidemic.

5. Case study: choiera epidemic in mopti (Mali)

The case study is the region of Mopti in Mali (Fig. 4), on the basis of indications provided by the National Health Directorate

(NHD) of Mali. The epidemiological disease considered is choiera and Mopti is the most affected region in Mali according to the epidemiologic data of NHD.

5.1.1. The characteristics of the epidemic: choiera

Choiera is a contagious epidemic enteric-borne infection caused by the bacterium *Vibrio choiera*, or *Comma bacillus* discovered by Pacini in 1854 and rediscovered by Robert Koch in 1884 (Harris, LaRocque, Qadri, Ryan, & Calderwood, 2012). Strictly limited to the human species, it is characterized by sudden and abundant diarrhoea (gastroenteritis) leading to severe dehydrations. Surveillance and prompt reporting allow for containing choiera epidemics rapidly (WHO, 2008).

Factors of propagation of the choiera epidemic (Dao et al., 2009):

- 78% of contamination cases was bathing in rivers,
- The most affected regions (Mopti, Segou, and Koulikoro) are located along the Niger River and constitute the core (starting point) of epidemics of choiera in the last three decades,
- The most favourable period in Mali is the rainy season Uune, July, August).

Finally, the geographical (environment) factor retained in the circle of Mopti is the Niger River. It is a question to determine the impact of the Niger River on the choiera epidemic during the rainy season.

5.1.2. Satellite data acquisition

Study Area:

From 1995 to 2004, Mopti region was the only one to be affected by all the choiera epidemics in Mali, located on the Niger River (Dao et al., 2009). The study area concerns Mopti town, 7262 km² of the area. Mopti is located in fair Mid Mali, so it is a culminating point between North and South, with a population over than 368 512 people (INSAT, 2009) made up of several ethnic groups around the Niger River to carry on activities trade, transport, fisheries, livestock and agriculture. The Niger River crosses the region and is joined by the Bani, an important tributary, at the city of Mopti. We focus on Mopti town with Path = 197 and Row = 50 or Lat = 14.5 and Long = 4.3 (Fig. 5).

Satellite data:

We opted for the Landsat archive satellite images of the study area, and they are available for download on <http://glovis.usgs.gov/> or <http://earthexplorer.usgs.gov/>. It is able to download images covering 30m², after specifying the criteria: address/Place (Path and Row) and date.

Although, the Landsat program cycle is 16 days in the same locality (scenes) some twenty images are produced each year

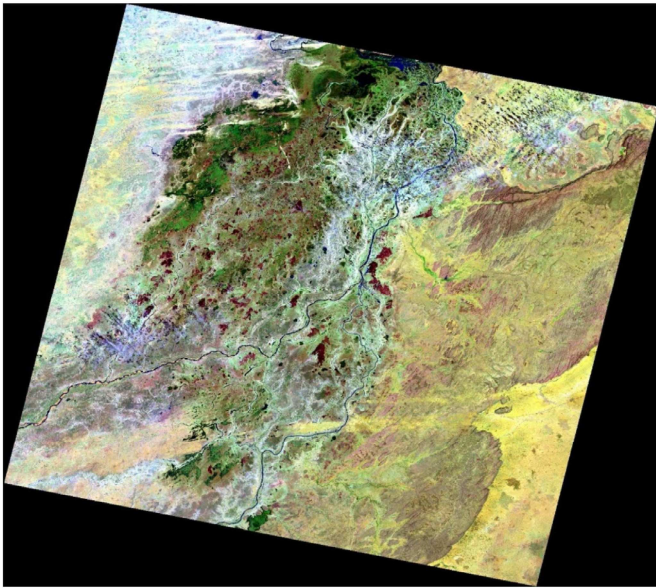


Fig. 5. Satellite image of Mopti (Path=I97 and Row=50).

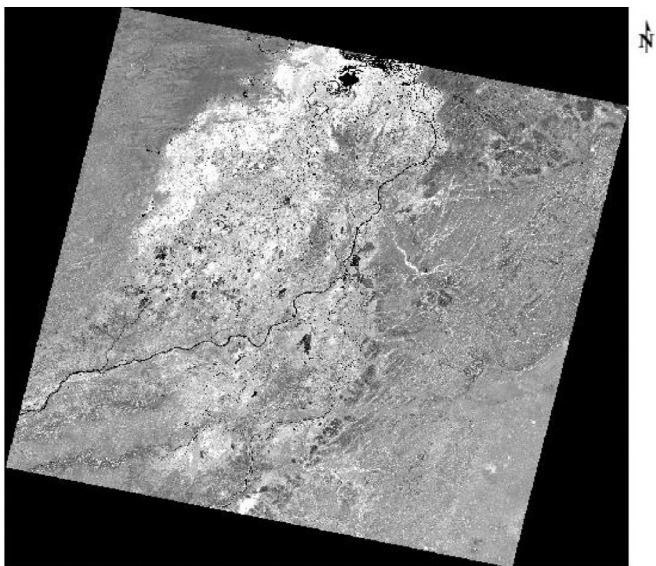


Fig. 6. NDVI file.

by scene (Fu & Weng, 2016). However, in our research, we took into account the domain expert knowledge who stipulates that the rainy season is the conducive period to the choiera epidemic particularly in July and August. Therefore, we used one satellite image per year while respecting the indicated periods.

5.2. Selection and transformation of data

After the download satellite data, the bands selected to display images are SWIR2 (Band 7) Band, NIR (Band 5) and BLUE (Band 2) with Stretch (0.5% Clip) and Gamma (0.8). It can be applied some corrections on satellite images such as atmospheric rectifications for better images processing.

5.3.1. Remote sensing measurements

At this step, we generate the NOVI file because NOVI is the best index to make the difference between vegetation and water area (Fig. 6) by using the NOVI formula $(NIR - R)/(NIR + R)$ (Tucker, 1979).

5.3.2. Number of confirmed cases

The Table 3 below describes the confirmed cases from Health Ministry structures (Dao et al., 2009). During the decade "1995-2004", choiera has hit ail regions of Mali with variations across regions and years (Table 3 and Fig. 7). Indeed, the regions of Mopti (3412 cases) and Segou (1500 cases) were involved in almost ail of the 2004-1995 choiera epidemics with a lull in 2002 in Segou (no case).

Kayes and Sikasso regions experienced fewer choiera outbreaks between 1995 and 2004. During this period, Kayes region experienced a single outbreak with 3497 cases and Sikasso region two outbreaks with 50 cases.

Finally, the case study concerns **Mopti** region, because it is the most affected region of the choiera epidemic in Mali from 1995 to 2004.

5.4. Discretization of data

Prior to the images classification into two groups (risk areas and not risk areas), it is necessary to prepare satellite images by a discretization technique and after resize data with the Nearest Neighbour resampling method (Richards, 1999) that are applied to images obtained (Software : Matlab, Envi). For our study, the Mopti town is considered, which means that ail satellite images processed (described in Table 4) have the same Map Coordinates:

Upper Left Coordinate:

Lat: 14 31 52.16

Lon:-4 142.64

Lower Right Coordinate:

Lat: 14 28 38.42

Lon:-4 1041.67

5.5. Data treatment: classification

The Table 5 gives different results after supervised classification (Maximum Likelihood) (Abkar et al., 2000).

5.6. Interpretation of results

We present the results (Table 6, Fig. 8) from classification process step and make them available by a codification system (discretization) (Tables 7 and 8) for data mining technique that allows good interpretation (Table 9).

Given a database (Table 8), the problem is to generate ail association rules that have support and confidence greater than the user-specified minimum support (called minsup) and minimum confidence (called minconf) respectively. An association rule $X \Rightarrow Y$ is an observed regularities induced a form of implicative relationship between two sets of items (i.e. binary attributes) X and Y such that X and Y are disjoint, and that X is unordered and Y is unordered. The support of a rule $X \Rightarrow Y$ (i.e. the frequency in the database) is the number of sets of items (i.e. itemsets) that contains XuY divided by the number of itemsets in the database. The confidence of a rule (i.e. the number of times a rule has been found to be true) is the number of itemsets that contains XuY, divided by the number of itemsets that contains X.

Many algorithms exist for discovering association rules (Kamsu-Foguem, Rigal, & Mauget, 2013), the best known being the Apriori algorithm. Unlike the other algorithms, Apriori finds ail association rules between frequent itemsets by adding to large sets, and pruning small sets (Luna, Cano, Sakalauskas, & Ventura, 2016). This is consistent with our objective to extract ail possible types of rules for later evaluation of the results by the user.

For the general problem of mining the association rules, m items potentially lead to 2^m frequent itemsets. To address this

Table 3
Distribution of choiera cases in Mali from 1995 to 2004 by region. (From 1997 to 2000, no cases of choiera in Mali.)

Regions	Number of confirmed cases						Total
	1995	1996	2001	2002	2003	2004	
Kayes	0	3497 (63.1%)	0	0	0	0	3497 (28.7%)
Koulikoro	57 (2.5%)	33 (0.6%)	0	0	99 (6.9%)	1 (0.0%)	190 (1.6%)
Sikasso	0	0	0	0	45 (3.1%)	5 (0.2%)	50 (0.4%)
Segou	166 (7.4%)	453 (8.2%)	16 (23.5%)	0	645 (44.9%)	220 (7%)	1500 (12.3%)
Mopti	987 (43.8%)	735 (13.3%)	52 (76.5)	16 (100%)	597 (46%)	1025 (35.8%)	3422 (28%)
Tombouctou	776 (34.4%)	616 (11.1%)	0	0	36 (2.5%)	1258 (44%)	2686 (22.1%)
Gao	234 (10.4%)	205 (3.7%)	0	0	0	351 (12.3%)	790 (6.5%)
Bamako	34 (1.5%)	2 (0.1%)	0	0	15 (1%)	0	51 (0.4%)
Total	2254 (100%)	5541 (100%)	68 (100%)	16 (100%)	1437 (100%)	2860 (100%)	12,176 (100%)

• District.

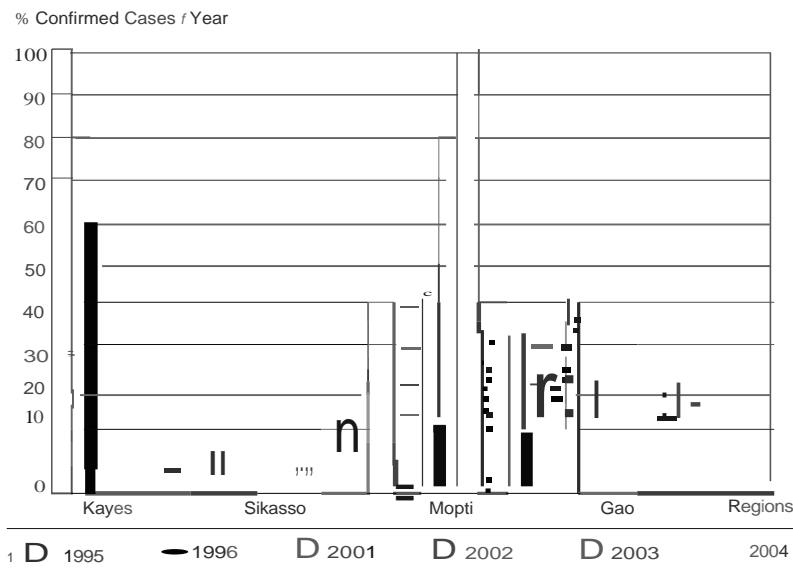


Fig.7. Distribution of choiera cases in Mali from 1995 to 2004 by region.

problem, the Apriori algorithm uses an estimation procedure in order to determine the itemsets that should be measured at each iteration. Thus, an itemset X of length k is frequent if and only if every subset of X , having length $k - 1$, is also frequent; i.e., if an itemset of size k is a frequent itemset, then all the itemsets below $(k - 1)$ size must also be frequent itemsets. This consideration permits a significant reduction of the search space, and allows rule discovery in a computationally reasonable time (Potes Ruiz, Kamsu Foguem, & Grabat, 2014).

Let k -itemset be an itemset having k items, L_k be a set of large itemsets and C_k a set of candidate k -itemsets. The Apriori algorithm is summarized in (Fig. 9). As described in (Agrawal & Srikant, 1994), the first iteration of the algorithm counts item occurrences to determine the large 1-itemsets. During the following iterations, the large itemsets L_{k-1} found in the $(k - 1)$ th iteration are used to generate the candidate itemsets C_k , using the Apriori-gen function (Fig. 9), which includes two phases taking as argument L_{k-1} : union and pruning. In the union phase, all k -itemsets candidates are generated. Then, in the pruning phase, all candidates generated in the union phase with some non-frequent $(k - 1)$ -itemset are removed. Nevertheless, because of its popularity and good performance, we have chosen the classical Apriori algorithm for extracting association rules in our application example.

The Association rules generated from Apriori Association Rules Algorithm (Agrawal & Srikant, 1994) of SPMF software (Fournier-Viger et al., 2014) are presented in Table 9. According to the results produced by processing satellite image and comparing them to the

field data, we can conclude that the level of the Niger River has an impact on the variation of choiera epidemic rate in Mopti town.

More the river level is high, at 66% the rate of contamination is high, and more the confirmed cases is low, then the risk area is also low with a full confidence. More the confirmed cases are high then the risk area is also high with a full confidence. The underlying interpretation is the societal factors that occur during the wintering period: throw garbage and faeces in Niger River causes bacterium *Vibrio*, the use (drinking and bathing) of this water can cause the choiera epidemic.

Mopti, crowded city, is located right in the heart of Mali, it is a culminating point between North and South, made up of several ethnic groups around the Niger River to carry on several activities: trade, transport, fisheries, livestock and agriculture. This implies to continue the awareness and education of the riverside communities and improve their living conditions to avoid the use of contaminated water and food.

The satellite remote sensing in choiera epidemic in Mopti is an interesting perspective to understand choiera epidemic and its propagation factors namely dynamic changes occurring in the Niger River, combining to field data to discover the risk areas of choiera.

6 Discussions

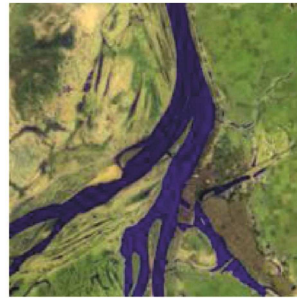
The high rates of choiera epidemic mortality in Mali in general and particularly Mopti is a challenge for health facilities to which it is necessary to equip itself with the epidemic surveillance tech-

Table 4
Satellite images processed.



date: 1995/08/05
Landsat 4 Present

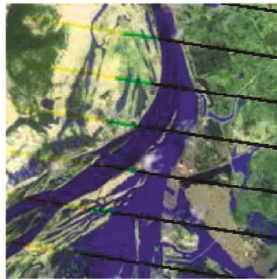
NO IMAGES
AVAILABLE IN
LANDSAT ARCHIVES
AT THIS AREA
FROM
1995/08/05-1998/02/18



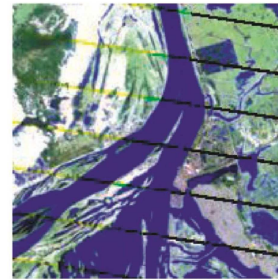
date: 2001/07/28
Landsat 7 SLC-on (1999-2003)



date: 2002/07/31
Landsat 7 SLC-on (1999-2003)



date: 2003/08/19
Landsat 7 SLC-off (2003->)



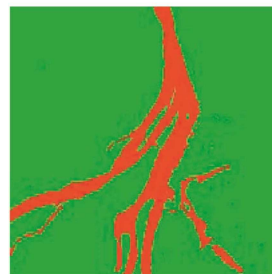
date: 2004/08/21
Landsat 7 SLC-off (2003->)

Table 5
Results after classification.

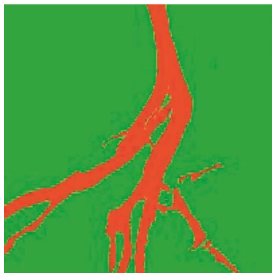


Date: 1995/08/05
Ground Truth (Percent)
R.risk areas (Red): 17.24%
Not risk areas (Green): 82.76 %

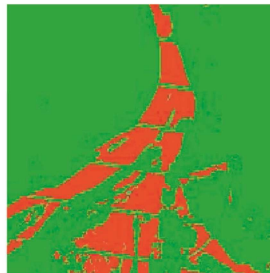
NO IMAGE AVAILABLE
INLANDSAT ARCHIVES
AT THIS AREA
FROM
1995/08/05-1998/02/18



Date: 2001/07/28
Ground Truth (Percent)
R.risk areas (Red): 16.75%
Not risk areas (Green): 83.25 %



date: 2002/07/31
Ground Truth (Percent)
R.risk areas (Red): 14.73%
Not risk areas (Green): 85.27%



date: 2003/08/19
Ground Truth (Percent)
R.risk areas (Red): 19.69%
Not risk areas (Green): 80.31 %



date: 2004/08/21
Ground Truth (Percent)
R.risk areas (Red): 23.93%
Not risk areas (Green): 76.07%

Table 6
Data base.

Years	Confirmed cases	Risk areas (%'100)
1995	987	1724
2001	52	1675
2002	16	1473
2003	597	1969
2004	1025	2393

Table 7
Codification system.

Confirmed cases			Risk areas		
Code	Interval	Description	Code	Interval	Description
2	[0-350]	Low	11	[09%-16%]	Low
3	[351-700]	Medium	22	[17%-24%]	High
	[701-1050]	High			

Table 8
Database codified.

Confirmed cases	Risk areas (%'100)
3	22
	11
1	11
2	22
3	22

niques. Among the epidemic surveillance techniques, we process satellite data to prevent and facilitate the management of an epidemic crisis by earth observations. Earth observation is the gathering of information about planet Earth's physical, chemical and biological systems via remote sensing technologies supplemented by earth surveying techniques, encompassing the collection, analysis and presentation of data (Unninayar & Olsen, 2015). There is the need to minimize negative impacts of epidemics along with the opportunities generated by earth observation (via satellite data) to improve social and economic well-being.

In the same logic to discovery risk areas, (Kazansky, Wood, and Sutherlun (2016) studied the current and potential role of satellite remote sensing in the campaign against malaria. The purpose is to establish the malaria risk map to facilitate the management to the government and public health. The proposed methodology is based on indicators from satellites and other sources, allowing establish which regions are most likely to be at risk of malaria. The main inconvenient of this method is that it is difficult (impossible) to implement a system to predicting risk areas without integrating data mining techniques to predict effective and efficient risk areas with rates accuracy.

In the purpose to provide the best interpretation, Punia, joshi, and Porwal (2011) identified 13 classes by a decision tree classification, for the free seasons, to explored the potential of multi-temporal IRS P6 (Resourcesat) Advanced Wide Field Sensor (AWiFS) data of land use land caver for Delhi, India. The main difference is that they do not consider field data from domain expert knowledge to validate the process in the presented methodology.

Hudak and Brockett (2004) use unstandardized Principal Components Transformation (PCT) to reconstruct the fire history of

Table 9
Interpretation of the results obtained.

Association rules generated	Interpretation
1=> 11 #SUP: 2 #CONF: 1	If confirmed cases is low ([0-350]) Then risk area is also low ([09%-16%]) with a full confidence (100%)
22=> 3 #SUP: 2 #CONF: 0,66667	If the risk area is high ([17%-24%]) Then confirmed cases is high ([701-1050]) with a good confidence (66%)
3=> 22 #SUP: 2 #CONF: 1	If confirmed cases is high ([701-1050]) Then risk area is also high ([17%-24%]) with a full confidence (100%)

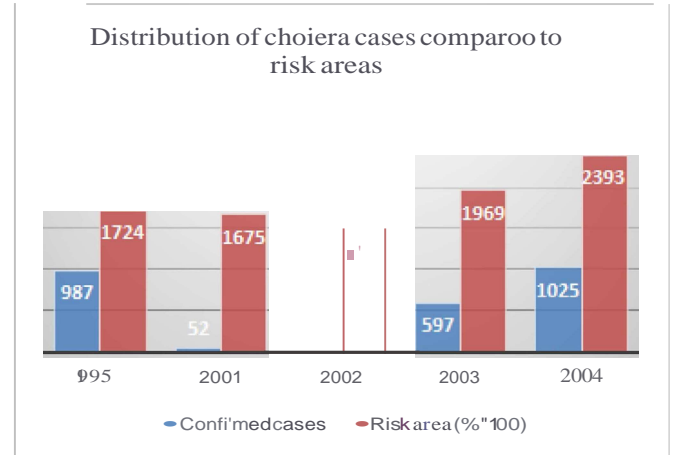


Fig. 8. Data Base corresponding graph.

Input: Database D

```

1)  $L_1 = \{ \text{large 1-itemsets} \};$ 
2) for ( $k = 2; Lk-1 \neq \emptyset; k++$ ) do begin
3)    $ek = \text{apriori-gen}(Lk-1);$  // New candidates
4)   for all transactions  $t \in D$  do begin
5)      $et = \text{subset}(ek, t);$  // Candidates contained in  $t$ 
6)     for all candidates  $c \in et$  do
7)        $c.\text{count}++;$ 
8)   end
9)    $Lk = \{ c \in edc.\text{count} \geq \text{minsup} \}$ 
10) end
11) Answer =  $\cup_k Lk;$ 

```

apriori-gen ($Lk-1$)

```

12) for all itemsets  $c \in ek$  do begin
13)   for all (k-1)-subsets  $s$  of  $c$  do begin
14)     if ( $s \notin Lk-1$ ) then
15)       delete  $c$  from  $ek$ 
16)   endfor
17) endfor

```

Fig. 9. The Apriori algorithm.

a semi-arid savannah landscape in southern Africa. They also use a supervised classification to process Lansat satellite data to differentiated burned and unburned areas. However, the only uses of satellite images are not sufficient for best accuracy, efficiency, objectivity, and consistency because fire severity maps are more dependent on strong field validation (White, Ryan, Key, & Running, 1996) than maps of fire existence or not.

The approach proposed by Pitarch et al. (2015) combines a satellite image processing and field data for looking for patterns and mapping of agricultural lands in order to improve food security early warning systems (Vintrou et al., 2012). However, this approach has limitations in its discretization phase and because the chosen discretization intervals were too large to capture texture temporal variations and moreover recommended finer discretization of these intervals. The advantage of our approach

compared to those in [Bicout et al. \(2015\)](#) is that we treat several satellite images of the same study area combined with field data collected over the same periods and this is what allows us to have a better interpretation.

[Contiu and Groza \(2016\)](#) proposed a hybrid system for crop classification from satellite images through a vote-based method for conflict resolution in ensemble learning. We have the same advantage by associate the domain expert knowledge with mines extracting from data bases to increase the classification accuracy. We consider that the existing methods for processing satellite images are mature and efficient, so the challenge today is to provide the most suitable means allowing the best interpretation of obtained results. For that, we focus on supervised classification algorithm (Maximum Likelihood) to process a set of satellite images from the same area but on different periods. Maximum Likelihood algorithm is the most common supervised classification method used with remote sensing image data ([Richards, 1999](#)), based on the probability that a pixel belongs to a particular class. However, there are other image classification methods with certain characteristics. The Minimum Distance algorithm is also attractive but it does not use covariance data. Another variant of Minimum Distance is the Mahalanobis distance, the difference is the use of the covariance matrix. The parallelepiped algorithm method is a simple classifier for machine-learning tasks, however there can be significant gaps between the parallelepipeds; pixels in those regions will not be classified. A deep learning for image classification based on neural networks have become new revolution in artificial intelligence and relevant for several domains: the audible or visual signal analysis, facial recognition, disaster recognition, voice recognition, computer vision, automated language processing ([Hinton, Osindero, & Teh, 2006](#)). Deep learning is a set of algorithms which aims to model high-level abstractions in data by using a deep graph with multiple processing layers ([Schmidhuber, 2015](#)). Deep learning classification seems more suited to pattern recognition, but they do not seem capable of logical reasoning ([Guo et al., 2016](#); [Haohan & Bhiksha, 2015](#)). The latest developments in deep learning have shown promise for modelling static data, such as computer vision, applying them to complex temporal data (e.g. the risk areas of the images) is gaining increasing attention ([Langkvist, Karlsson, & Loutfi, 2014](#)).

Therefore, one of the biggest contributions of the proposed technique is to allow the best interpretation of obtained results taking into account field data. This study also demonstrated that the Niger River has impact on the evolution of the choiera epidemic in Mopti region, all through the techniques of remote sensing from space for more precisely processing of satellite images.

7. Conclusion and related works

Humanity will always be faced with epidemics and the emergence of new diseases. Global monitoring and efforts to address potential disruptions effects of infectious diseases whose emergence with multifactorial aspects are challenging. For monitoring of infectious diseases, the use of remote sensing is very beneficial in the tele-epidemiology that consists to collect data in the field and satellite data to face and prevent infection disease related to the environment discoverable by satellites.

We focus on the use of data mining technique to discover risk or not risk areas of epidemic crisis from satellite images. For our analysis, we used supervised classification method to process a set of satellite images from the same area but on different periods. A novel method of six phases (describing pre-treatment, data mining, and post-treatment) is described to ensure suitable means for transforming data, generating information and extracting knowledge. The proposed method was successfully evaluated

with the implementation in Mopti region, Mali (West Africa), to discover choiera epidemic.

The main contributions of the paper are: discretization in the pre-processing phase improving the quality of the obtained results, establishment of the link between the environment and the epidemic, and identification of most risky areas for the propagation and emergence of the epidemic. This will allow national governments, local authorities and the public health officials to effective management according to risk areas.

As discuss research limitations, it would be desirable to have only good quality images, but from 2003 to 2004, only Landsat 7 ETM+SLC-off (2003-present) supplied images of Landsat program. On May 31, 2003, Scan Line Corrector (SLC) in the ETM+ instrument failed, causing stripes on the images (2003/08/19 and 2004/08/21) hence its name SLC-off ([Wulder et al., 2012](#)). The stripes (zigzag) have less impact on the results. The green stripes above the River classified Not risk areas (green) are false positives and should therefore be classified into Risk areas (Red), thus the relative growth in percentage of risk areas providing a reasonable argumentation and justifies the outcomes. In a medium-term perspective, there is a need to improve the processing of satellite images namely the images with less quality to achieve the desired results. In this work, we use the same data format to collect images from satellite systems. Therefore, as future research, we intend to take into account images from different satellites in future works. Processing satellite images from heterogeneous sources is also a considerable challenge for the scientific communities.

Acknowledgements

This work is supported by a funding within the grant of "Programme de Formation des Formateurs des Universités de Bamako et de Ségou", delivered by the Malian government. We want to thank the two anonymous reviewers whose significant comments have contributed to improve the content of the paper. We congratulate Shester GUEUWOU for his linguistic ability of proofreading as a native English speaker.

References

- Abkar, A.-A., Sharifi, M. A., & Mulder, N.J. (2000). Likelihood-based image segmentation and classification: A framework for the integration of expert knowledge in image classification procedures. *International journal of Applied Earth Observation and Geoinformation*, 2(2), 104-119.
- Agrawal, R., & Srikant, R. (1994, June). *Fast algorithms for mining association rules in large databases*. San Jose, California: IBM Almaden Research Center Research Report RJ 9839.
- Ahmad, A., & Quegan, S. (2012). Analysis of maximum likelihood classification on multispectral data. *Applied Mathematical Sciences*, 6(129), 6425-6436.
- Alatrasta-Salas, H., Azé, J., Bringay, S., Cernesson, F., Selmaoui-Folcher, N., & Teisseire, M. (2015). A knowledge discovery process for spatiotemporal data: Application to river water quality monitoring. *Ecological Informatics*, 26(March), 127-139 Part 2.
- Alrababah, M. A., & Alhamad, M. N. (2006). Land use/cover classification of arid and semi-arid mediterranean landscapes using landsat ETM. *International journal of Remote Sensing*, 27, 2703-2718.
- Anderson, G. L., Hanson, J. D., & Haas, R. H. (1993). Evaluating landsat thematic mapper derived vegetation indices for estimating above-ground biomass on semiarid rangelands. *Remote Sensing of the Environment*, 45(2), 165-175.
- Arentze, T. A. (2009). Spatial data mining, cluster and pattern recognition. *International Encyclopedia of Human Geography*, 325-331.
- Asokan, G. V., & Asokan, V. (2015). Leveraging "big data" to enhance the effectiveness of "one health" in an era of health informatics. *Journal of Epidemiology and Global Health*, 5(December (4)), 311-314.
- Bicout, D.J., Vautrin, M., Vignolles, C., & Sabatier, P. (2015). Modeling the dynamics of mosquito breeding sites vs rainfall in barkedji area. Senegal. *Ecological Modelling*, 317(December (10)), 41-49.
- Broge, N. H., & Leblanc, E. (2000). Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. *Remote Sensing and Environment*, 76, 156-172.
- Compieta, P., Di Martino, S., Bertolotto, M., Ferrucci, F., & Kechadi, T. (2007). Exploratory spatio-temporal data mining and visualization. *Journal of Visual Languages and Computing*, 18, 255-279.

- Contiu, S. & Groza, A. (2016). Improving remote sensing crop classification by argumentation-based conflict resolution in ensemble learning. *Expert Systems with Applications*, 64(December (1)), 269-286.
- Dao, S., Konaté, I., Oumar, A. A., Sacko, M., Maiga, I., Toure, K., et al. (2009). Les épidémies de choléra au Mali de 1995 à 2004. *Santé Publique*, 3, 21, 263-269.
- Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. *Advances in knowledge discovery and data mining*, 1, 1-34.
- Fournier-Viger, P., Gomariz, G., T., A., Soltani, A., Wu, C., & Tseng, V.S. (2014). SPMF: A java open-source pattern mining library. *Journal of Machine Learning Research*, 15, 3389-3393.
- Fournier, J. M. (1996). Medico-Surgical Encyclopaedia. *Infectious Diseases*, 8026 F 1986; 11 :6D.
- Fu, P., & Weng, Q. (2016). Consistent land surface temperature data generation from irregularly spaced landsat imagery. *Remote Sensing of Environment*, 184(October), 175-187.
- Fugazzi, M., & Spokane NWS, W. A. (2008). Satellite imagery overview for pilots. *NOAA's National Weather Service*, 8(2), 1-4.
- Gitelson, A. A., Kaufman, Y. J., Stark, R., & Rundquist, D. (2002). Novel algorithms for remote estimation of vegetation fraction. *Remote Sensing of Environment*, 80, 76-87.
- Gitelson, A. A., & Merzlyak, M. N. (1994). Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves. *Journal of Photochemistry and Photobiology*, 22, 247-252.
- González, M. P., Bonaccorso, E., & Pape, M. (2015). Applications of geographic information systems and remote sensing techniques to conservation of amphibians in northwestern Ecuador. *Global Ecology and Conservation*, 3(Uanuary), 562-574.
- Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew Michael, S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187(April (26)), 27-48.
- Halimi, M., Farajzadeh, M., Delavari, M., & Arbabi, M. (2015). Developing a climate-based risk map of fascioliasis outbreaks in Iran. *Journal of Infection and Public Health*, 8(September-October (5)), 481-486.
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (p. 744). Hardcover/USA: Publisher: Morgan Kaufmann; San Francisco. July 6.
- Haohan, W., & Bhiksha, R. (2015). A survey: time travel in deep learning space: an introduction to deep learning models and how deep learning models evolved from the initial ideas. arXiv:1510.04781v2 [cs.LG].
- Harris, J. B., LaRocque, R. C., Qadri, F., Ryan, E. T., & Calderwood, S. B. (2012). "Cholera" *Lancet*, 379(30 June (9835)), 2466-2476.
- Hengl, T., Heuvelink, G. B. M., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120, 75-93.
- Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7), 1527-1554.
- Hobbs, T. J. (1995). The use of NOAA-AVHRR NOVI data to assess herbage production in the arid rangelands of Central Australia. *International Journal of Remote Sensing*, 16(7), 1289-1302.
- Hudak, A. T., & Brockett, B. H. (2004). Mapping fire scars in a southern African savannah using Landsat imagery. *International Journal of Remote Sensing*, 25(20 August (16)), 3231-3243.
- Huete, A. R. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, 25, 295-309.
- Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83, 195-213.
- INSTAT - Institut national de la statistique. (2016). *Quatrième recensement général de la population et de l'habitat 2009 du mali* http://www.instat-mali.org/content/rghp/ramig09_rghp.pdf last access 10 September.
- Jamison, A., Tuttle, E., Jensen, R., Bierly, G., & Ganser, R. (2015). Spatial ecology, landscapes, and the geography of vector-borne disease: A multi-disciplinary review. *Applied Geography*, 63(September), 418-426.
- Jutla, A. S., Akanda, A. S., & Islam, S. (2012). Satellite remote sensing of space-time plankton variability in the Bay of Bengal: Connections to cholera outbreaks. *Remote Sensing of Environment*, 123(August), 196-206.
- Kansakar, P., & Hossain, F. (2016). A review of applications of satellite earth observation data for global societal benefit and stewardship of planet earth. *Space Policy*, 36(May), 46-54.
- Kazansky, Y., Wood, D., & Sutherland, J. (2016). The current and potential role of satellite remote sensing in the campaign against malaria. *Acta Astronautica*, 121, 292-305.
- Kamsu-Foguem, B., Rigal, F., & Mauget, F. (2013). Mining association rules for the quality improvement of the production process. *Expert Systems With Applications*, 40(4), 1034-1045 March.
- Kurtz, C., Passat, N., Gaçarski, P., & Puissant, A. (2012). Extraction of complex patterns from multiresolution remote sensing images: A hierarchical top-down methodology. *Pattern Recognition*, 45(February (2)), 685-706.
- Lacaux, J. P., Tourre, Y. M., Vignolles, C., Ndione, J. A., & Lafaye, M. (2007). Classification of ponds from high-spatial resolution remote sensing: Application to Rift Valley Fever epidemics in Senegal. *Remote Sensing of Environment*, 106(15 January (1)), 66-74.
- Langkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42(Quene (1)), 11-24.
- Lausch, A., Schmidt, A., & Tischendorf, L. (2015). Data mining and linked open data - New perspectives for data analysis in environmental research. *Ecological Modelling*, 295(Qanuary (10)), 5-17.
- Liu, S., & Hodgson, M. E. (2016). Satellite image collection modeling for large area hazard emergency response. *ISPRS Journal of Photogrammetry and Remote Sensing*, 118(August), 13-21.
- Liu, Y., Hu, J., Snell-Feikema, I., VanBemmel, M. S., Lamsal, A., & Wimberly, M. C. (2015). Software to facilitate remote sensing data access for disease early warning systems. *Environmental Modelling & Software*, 74(December), 247-257.
- Luna, J. M., Cano, A., Sakalauskas, V., & Ventura, S. (2016). Discovering useful patterns from multiple instance data. *Information Sciences*, 357(August (20)), 23-38.
- Mabaso, M. L. H., & Ndlovu, N. C. (2012). Critical review of research literature on climate-driven malaria epidemics in sub-Saharan Africa. *Public Health*, 126(November (11)), 909-919.
- Malone John, B., Yang, G.-j., Leonardo, L., & Zhou, X.-N. (2010). *Chapter 4 - implementing a geospatial health data infrastructure for control of usian schistosomiasis in the People's Republic of China and the Philippines*. *Advances in Parasitology*, 73, 71-100.
- Marechal, F., Ribeiro, N., Lafaye, M., & Guell, A. (2008). Satellite imaging and vector-borne diseases: The approach of the French National Space Agency (CNES). *Geospatial Health*, 3(1), 1-5.
- Moges, S. M., Raun, W. R., Mullen, R. W., Freeman, K. W., Johnson, G. V., & Salie, J. B. (2004). Evaluation of green, red and near infrared bands for predicting winter wheat biomass, nitrogen uptake, and final grain yield. *Journal of Plant Nutrition*, 27(8), 1431-1441.
- Mosomtai, G., Evander, M., Sandström, P., Ahlm, C., Sang, R., Hassan, O. A., et al. (2016). Association of ecological factors with Rift Valley fever occurrence and mapping of risk zones in Kenya. *International Journal of Infectious Diseases*, 46(May), 49-55.
- Nnadi, N. E., Nizing, L., Mark, O. O., & Onyedibe, K. I. (2011). Landscape epidemiology: An emerging perspective in the mapping and modelling of disease and disease risk factors. *Asian Pacific Journal of Tropical Disease*, 1(September (3)), 247-250.
- Piarroux, R. (2002). Cholera: Epidemiology and transmission. Experience from a number of humanitarian interventions in Africa, the Indian Ocean and Central America. *Bulletin de la Société de Pathologie Exotique*, 95(5), 345-350.
- Pitarch, Y., Ienco, D., Vintrou, E., Bégué, A., Laurent, A., Poncelet, P., et al. (2015). Spatio-temporal data classification through multidimensional sequential patterns: Application to crop mapping in complex landscape. *Engineering Applications of Artificial Intelligence*, 37(Uanuary), 91-102.
- Potes Ruiz, P., Kamsu Foguem, B., & Crabot, B. (2014). Generating knowledge in maintenance from experience feedback. *Knowledge-Based Systems*, 68(Septembe), 4-20.
- Prabhakara, K., Hively, W., & McCarty, G. W. (2015). *Evaluating the relationship between biomass, percent groundcover and remote sensing indices across six winter cover crop fields in Maryland, United States*. *International Journal of Applied Earth Observation and Geoinformation*, 39(Uuly), 88-102.
- Punia, M., Joshi, P. K., & Porwal, M. C. (2011). Decision tree classification of land use land cover for Delhi, India using IRS-P6 AWiFS data. *Expert Systems with Applications*, 38(May (5)), 5577-5583.
- Qin, Y., Zhang, S., Zhu, X., Zhang, J., & Zhang, C. (2009). POP algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases. *Expert Systems with Applications*, 36(March (2)), 2794-2804 Part 2.
- Richards, J. A. (1999). *Remote sensing digital image analysis* (p. 240). Berlin: Springer-Verlag.
- Rouse, J. W., Jr., Haas, R. H., Schell, J. A., & Deering, D. W. (1994). *Third earth resources technology satellite-1 symposium- volume 1: Technical presentations*. NASA SP-351. compiled and edited by S.C. Freden, M. Enrico P., & B. Margaret A. pages, published by NASA, Washington, D.C., (1974), p.309.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61(Uanuary), 85-117.
- Short, N. M. (1982). *The landsat tutorial/ workbook, basics of satellite remote sensing (NASA reference publication 1078)*. Washington DC, USA: National Aeronautics & Space Administration.
- Song, K.-Y., Oh, H.-j., Choi, J., Park, I., Lee, C., & Lee, S. (2012). Prediction of landslides using ASTER imagery and data mining models. *Advances in Space Research*, 49(1 March (5)), 978-993.
- Tahakata, C., Amin, R., Sarma, P., Banerjee, C., Oliver, W., & Fa, J. E. (2009). Remotely-sensed active fire data for protected area management: Eight-year patterns in the Manas National Park, India. *Environmental Management*, 45(2), 414-423 February 2010.
- Takahashi, A., & Kurosawa, T. (2016). Regression correlation coefficient for a Poisson regression model. *Computational Statistics & Data Analysis*, 98(Uune), 71-78.
- Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining* (1st ed.). Essex, England: Pearson Education May 12 ISBN-13 978-0321321367.
- Tavana, M., Liu, W., Elmore, P., Petry, F. E., & Bourgeois, B. S. (2016). A practical taxonomy of methods and literature for managing uncertain spatial data in geographic information systems. *Measurement*, 81(March), 123-162.
- Tsai, F., Lai, J.-S., Chen, W. W., & Lin, T.-H. (2013). Analysis of topographic and vegetative factors with data mining for landslide verification. *Ecological Engineering*, 61(Part C, December), 669-677.
- Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8, 127-150.
- Tucker, C. J., & Sellers, P. J. (1986). Satellite remote sensing of primary production. *International Journal of Remote Sensing*, 7(11), 1395-1416.
- Unninayar, S., & Olsen, L. M. (2015). *Monitoring, observations, and remote sensing - global dimensions*. Colorado, USA: Reference Module in Earth Systems and Environmental Sciences.

- Vieira, M. A., Formaggio, A. R., Renno, C. D., Atzberger, C., Aguiar, D. A., & Mello, M. P. (2012). Object based image analysis and data mining applied to a remotely sensed landsat time-series to map sugarcane over large areas. *Remote Sensing of Environment*, 123(August), 553-562.
- Vintrou, E., Desbrosse, A., Bégué, A., Traoré, S., Baron, C., & Lo, S. D. (2012). Crop area mapping in West Africa using landscape stratification of MODIS time series and comparison with existing global land products. *International journal of Applied Earth Observation and Geoinformation*, 14(February (1)), 83-93.
- White, J. D., Ryan, K. C., Key, C. C., & Running, S. W. (1996). Remote sensing of forest fire severity and vegetation recovery. *International journal of Wildland Fire*, 6, 125-136.
- WHO (World Health Organization). (2008). *WHO Cholera: Prevention and control*. Health topics: WHO Retrieved 2008-12-08.
- Wulder, M. A., Masek, J. G., Cohen, W. B., Loveland, T. R., & Woodcock, C. E. (2012). Opening the archive: How free data has enabled the science and monitoring promise of Landsat. *Remote Sensing of Environment*, 122(July), 2-10.
- Xie, L., Li, G., Xiao, M., & Peng, L. (2016). Novel classification method for remote sensing images based on information entropy discretization algorithm and vector space model. *Computers & Geosciences*, 89, 252-259.
- Xu, Z., Liu, Y., Ma, Z., Li, S., Hu, W., & Tong, S. (2014, July). Impact of temperature on childhood pneumonia estimated from satellite remote sensing. *Environmental Research*, 132, 334-341.
- Zhang, S., Chengqi, Z., & Qiang, Y. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17, 375-381.