IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MULTI-DIMENSIONAL DATA PREPARATION: A PROCESS TO SUPPORT VULNERABILITY ANALYSIS AND CLIMATE CHANGE ADAPTATION

[1]Wrushabh S. Sirsat
[1]P.G. Student
[1]Department of Computer Science & Engineering, Sipna C.O.E.T, Amravati, India

ABSTRACT:-

Agriculture is the backbone of a country's economic system, considering that it not only provides food and raw materials but also employment opportunities for a large percentage of the population. In this way, determining the degree of agricultural vulnerability represents a guide for sustainability and adaptability focused on changing future conditions. In many cases, vulnerability analysis data is restricted to use by authorized personnel only, leaving open data policies aside. Furthermore, data in its native format (raw data) by nature tend to be diverse in structure, storage formats, and access protocols. In addition, having a large amount of open data is important (though not sufficient) to obtain accurate results in data-driven analysis. These data require a strict preparation process and having guides that facilitate this process is becoming increasingly necessary. In this study, we present the step by step processing of several open data sources in order to obtain quality information for feedback on different agricultural vulnerability analysis. The data preparation process is applied to a case study corresponding to the upper Cauca river basin in Colombia. All data sources in this study are public, official and are available from different web platforms where they were collected. In the same way, a ranking with the importance of variables for each dataset was obtained through automatic methods and validated through expert knowledge. Experimental validation showed an acceptable agreement between the ranking of automatic methods and the ranking of raters.

INTRODUCTION:-

Agriculture is one of the activities most affected by climatic factors. It not only provides food and raw materials but also employment opportunities for a large percentage of the population. Although agriculture contributes approximately 5% to 7% of GDP in modern economies, as this percentage increases, the economic system becomes more vulnerable. The effects of variability and climate change on food production are now a reality. These phenomena have begun to affect the production of the ten main crops (barley, cassava, corn, palm oil, rapeseed, rice, sorghum, soybeans, sugarcane, and wheat), which represent key food sources for human beings. These food sources represent 83% of all calories produced on arable land, and for this reason, understanding how much can be affected has become an urgent task for researchers around the world. In this way, determining the degree of agricultural vulnerability represents a guide for sustainability and adaptability focused on changing future conditions.

Agricultural vulnerability has become a fundamental basis for analyzing the risks of climate variability. In recent decades, several studies have focused on analyzing and measuring this type of risk. Fig. 1 presents the approaches around *Vulnerability Analysis* and *Climate Change Adaptation* published since 2010.

FIGURE 1. Research works around Vulnerability Analysis and Climate Change Adaptation applied to several domains since 2010.

This systematic mapping was developed using the methodology proposed by Petersen, where the databases of Scopus, Google Scholar, and Science Direct were consulted. 80 related works were found and classified into three topics: environmental, soil and crops, and water supply. We highlight those related to agriculture and food security below. In this sense, RHoMIS (Rural Household of Multiple Indicators Survey) is a methodology composed of surveys and databases to monitor the agricultural sector through food systems. Likewise, the International Model for Policy Analysis of Agricultural Commodities and Trade (IMPACT) is a network of economic, water, and related crop models that simulates national and international agricultural markets. Following this line of research but at a regional scale, several approaches integrate physical, agro ecological and socioeconomic indicators. These indicators were grouped into the components of exposure, sensitivity, and adaptability using a composite index method. Finally, Agriculture, Vulnerability, and Adaptability (AVA) is a methodology for calculating the vulnerability of productive systems in the upper Cauca River basin in Colombia through multiple key indicators.

These types of approaches collect and generate valuable information in workshops organized between different stakeholders. However, to obtain acceptable results there are

Some limitations that increase the complexity of the entire process. The first corresponds to the type of analysis (qualitative or quantitative) developed in these studies. Using a qualitative approach had several challenges, thus, most approaches are predominantly quantitative. The second refers to the difficulty in reaching an agreement between participants. Sometimes the panels of experts become unpleasant experiences by not reaching a consensus among the stakeholders. This leads to a third limitation which lies in the time required to implement the analyses. The enormous amount of non-trivial work represents a high time of analysis.

The above implies having sufficient and relevant data sources for such analyses. However, data in its native format (raw data) by nature tend to be diverse in structure, storage formats, and access protocols. There are often intrinsic spatial-temporal relationships between different data sources, which may offer relevant knowledge for a given information query [9]. This need is often unsatisfied due to data inconsistencies. If an adequate cleaning process is not applied, subsequent analyses will not be accurate enough. In other words, a great deal of information and knowledge will be lost when entering erroneous data ("garbage in, garbage out").

1.1 DOMAIN INTRODUCTION:-

Weather forecasting is the application of science and technology to predict the state of the atmosphere for a given location. Ancient weather forecasting methods usually relied on observed patterns of events, also termed pattern recognition. For example, it might be observed that if the sunset was particularly red, the following day often brought fair weather. However, not all of these predictions prove reliable. Here this system will predict weather based on parameters such as temperature, humidity and wind. This system is a web application with effective graphical user interface. User will login to the system using his user ID and password. User will enter current temperature; humidity and wind, System will take this parameter and will predict weather from previous data in database. The role of the admin is to add previous weather data in database, so that system will calculate weather based on these data.

Weather forecasting system takes parameters such as temperature, humidity, and wind and will forecast weather based on previous record therefore this prediction will prove reliable. This system can be used in Air Traffic, Marine, Agriculture, Forestry, Military, and Navy etc. Forecasting the temperature and rain on a particular day and date is the main aim of this paper. In the paper we forecast rain and temperature for Europe; year up to 2051 and also we forecast temperature of world; year up to 2100.Our paper is aimed to provide real time weather forecast service at finest granularity level with recommendations. We grab user's location (longitude, latitude) using GPS data service whenever user requests for our services. Our system will process the users query and will mine the data from our repository to draw appropriate results. Users will be provided with recommendations also and that is the key facility

of our service. Personalized forecast is generated for each individual user based on their location.

LITERATURE REVIEW:-

Weather forecasting has been one of the most challenging difficulties around the world because of both its practical value in popular scope for scientific study and meteorology. Weather is a continuous, dynamic, multidimensional chaotic process, and data-intensive and these properties make weather forecasting a stimulating challenge. It is one of the most imperious and demanding operational responsibilities that must be carried out by many meteorological services all over the globe. Various organizations / workers in India and abroad have done demonstrating using supported time series data manipulation. The various methodologies viz. statistic decomposition models, Exponential smoothing models, ARIMA models and their dissimilarities like seasonal ARIMA models, vector ARIMA models using flexible time series, ARMAX models i.e. ARIMA with following informative variables etc., which has been used for forecasting purposes. Many trainings have taken place within the analysis of pattern and circulation of rainfall in many regions of the world. Totally altered time series methods with different purposes are used to investigate weather information in many different literatures. Accurate and timely weather forecasting is a major challenge for the scientific research. Weather prediction modelling involves a combination of many computer models, observations and acquaintance of trends and designs. Using these methods, practically accurate forecasts can be made up. Regression is a statistical experimental technique and it must be widely used in many business, the behavioural sciences, social and climate recasting and many other areas.

In the year.1980, Agrawal et al. explained the phenomena for time series regression models for forecasting the yield of rice in Raipur district on weekly data using weather parameters.
In the year 1999 Palmer, N.T said that ensemble forecasts as input to a simple decision-model analysis, it is shown that probability forecasts of weather and climate have greater potential economic value than corresponding single deterministic forecasts with uncertain accuracy.

In the year 2000 Maier & Dandy stated that ANNs are being used increasingly for the prediction and forecasting of a number of water resources variables, including rainfall, flow, water level and various water quality parameters. But the modelling of ANN is poorly described. Takagi.H (2000) introduces the two-new patent using NN+FS the Soft computing techniques. In the year 2001 Sivapragasam et al, found that SVM has higher prediction accuracy of hydrologic variables than that of the non-linear prediction (NLP) method.SSA-SVM results in a significant improvement in the case study on Singapore rainfall prediction with a correlation coefficient of 0.70 as opposed to 0.51 obtained by NLP. Zhang, (2001) propose that the linear ARIMA model and the nonlinear ANN model are used jointly, aiming to capture different forms of relationship in the time series data. It has been proposed that the combination method can be an e5ective way to improve forecasting performance. In the year 2002 Taylor & Baize (2002) concluded that there is strong potential for the use of weather collective predictions in NN load forecasting. Meek et al, (2002)

discussed the one of the most important aspect of modeling time series is handling seasonality in data. Seasonality can be handled using ART models by explicitly allowing or including relevant repressor variables in the linear regressions at the leaves. In the year 2003 Kokkinos et al, concluded that nonlinear models of the speech production system have been presented, that are constructed on the reconstructed attractor of speech signals. Temeyer et al, (2003) stated that although the nonlinear models usually performed best, for some parameters at some times, the linear models were better.

In the year 2006 SomvanshiK et al, predicted that the ANN is more appropriate ten AIRMA model in long term prediction. Leng et al, (2006) evident that the GA-based pruning method, as a global search tool, is superior to the OBS-based pruning method to identify the significance of the existing EBF neurons. Gooijer et al. (2006) reviewed the progress on time series forecasting

In the year Ni.X (2008) found that neural network is suitable for solving data mining problem and it will improve the efficiency of data mining methods. Ingsrisawang et al, found that machine learning techniques are suitable for prediction of rainfall in same day period. Choudhary & Garg. (2008) proposed that a hybrid GA-SVM system for predicting the future direction of stock prices. Qi & Zhang, (2008) found that the most effective way to model and forecast trend time series with NNs, a recent popular nonlinear modeling tool. Mutlu et al, (2008) found the comparison the ANN models to forecast daily flows at multiple gauging stations in the Eucha watershed in north-west

In the year 2016 new research worked emerged, Hirani & Mishra reports a detailed survey on rainfall predictions using different rainfall prediction methods extensively used over last 20 years. Liu et al, found the weak generalization ability for the FNN by using BP neural network,

TECHNOLOGY

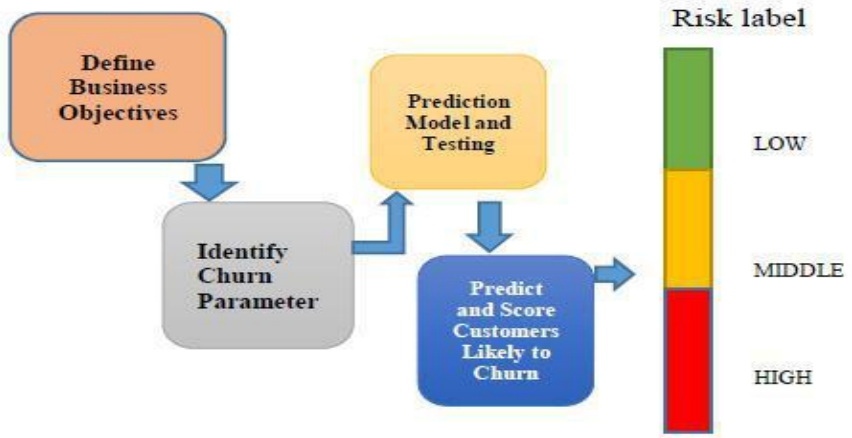


Fig no.2.2: Technology Diagram

This section explains how the Ensemble based Classifiers and well-known Base Classifiers were used in the Churn Prediction Model.

A. Decision Tree

Decision Tree was developed to overcome the drawbacks of ID3 algorithm. C4.5 utilizes the benefits of greedy approach and uses a series of rules for classification. Although this approach gives a high classification accuracy rate it fails to respond to noisy data. Gain is the main metric used in the decision tree to decide the root node attribute.

B. Naive Bayes

Naïve Bayes is a brute-force method for training the model. The underlying principle behind Naïve Bayesian classifier is Bayes Theorem. For the classification problem, each predictor attribute was consider separately with class label for model construction using training dataset. Predictor attribute includes the area, service calls, evening calls night calls etc. Apply the conditional probability for each attribute belongs to all the predictor attributes given
Class label represents churn. The disadvantage of this methods is, it is not suited for the large dataset.

C. Support Vector Machine

SVM algorithm was proposed by **Boser, Guyon, and Vapnik.** It was very well used for both classification and regression problem. SVM maps all the data points to a higher dimensional plane to make the data points linear separable. The plane which divides data points is known as hyper plane. It can be used for small dataset to give an optimal solution. SVM cannot be more effective for noisy data. SVM model tries to find out the churn and non-churn customer. In order to divide the dataset into churner and non-churner group, first it will take all the data points in n*dimensional* plane and divide the data points into churner and non-churner group based on maximum marginal hyper plane.

Based on the maximum marginal hyper plane it will divide the data points into churner and non-churner group. Here *n* represents the number of predictor variable associated with the dataset.

D. Bagging

Bagging (or Bootstrap aggregation) is one of the Ensemble based Classifiers which consist of bag of similar type or dissimilar type base classifiers. Bagging algorithm
Helps to reduce the variance of the classifier used for the Churn Prediction Model in order to increase the performance. The steps in designing of Churn Prediction Model using Bagging algorithm are as follows, first, it is required to divide the input dataset into *k* subset with replacement, then it requires to train the model by using the *(k-1)* subset and test the model using the dataset which has not been used for training model. The experimental results showed
That, bagging is effective, because it predicts the test instances using the classifier which has more accuracy from the bag of classifier, Bagging requires heavier computational resource for the Model construction.

E. Boosting

Boosting Ensemble technique is designed in such a way that it will maintain a weight for each training tuple. After a classifier is learned from the training tuple, weights are updated for the subsequent classifier. The final Boosted Classifier combines the vote of each individual classifier for prediction to improve the performance of the classifier. In similar to SVM, this model is also not suited for noisy data. The key idea for the customer Churn Prediction using Boosting algorithm is to train a series of classifier simultaneously and keep updating the model accuracy for improving the performance of the classifier.

F. Random Forest

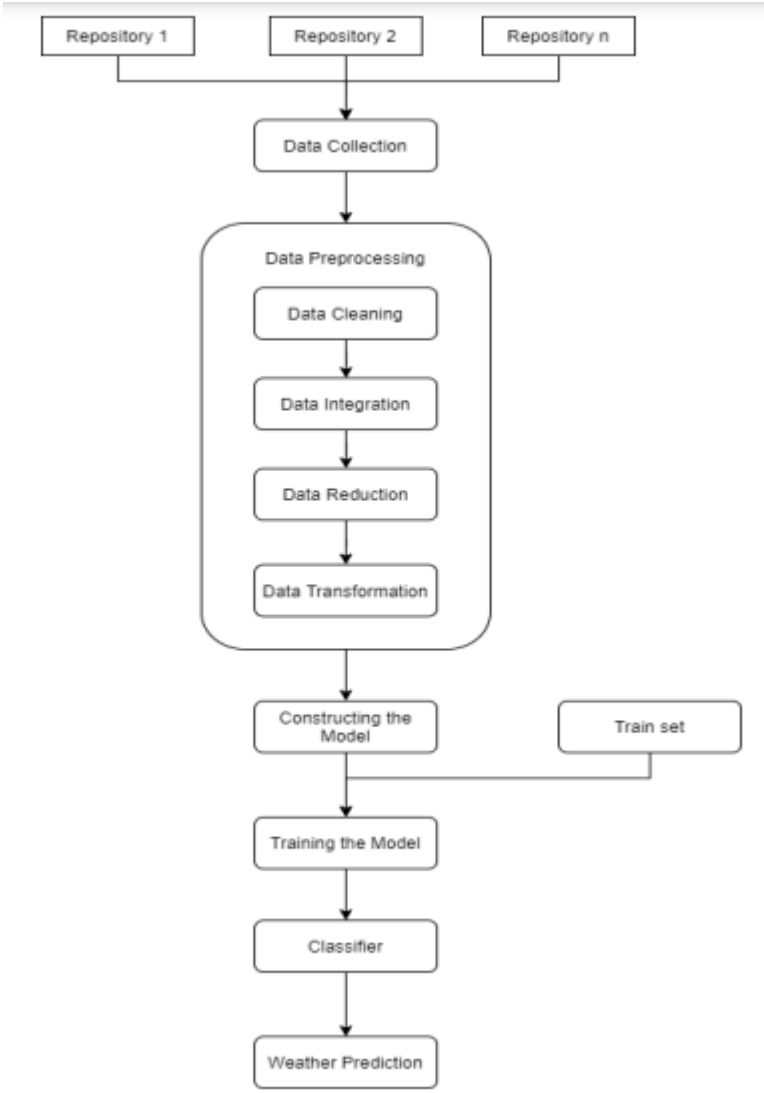
Random forest (**Breiman 2001)** works based on the random subspace method. The designed strategy used in Random Forest is divide and conquer. It forms number of Decision Trees and each Decision Tree is trained by selecting any random subset of attribute from the whole predictor attribute set. Each tree will grow up to maximum extent based on the attribute present in the subset. Then after, based on average or

weighted average method, the final Decision Tree will be constructed for the prediction of the test dataset. Random forest runs efficiently in large dataset. It can handle thousands of input variables without variable deletion. It also handle the missing values inside the dataset for training the model. It is difficult to handle the unbalanced dataset by using Random Forest.

SYSTEM DESIGN:-

The system is built on windows 2007 operating system. The system uses advanced java technology along with machine learning concepts. MySQL is used for storing data. This system uses three-tier architecture. The web service layer provides the android user to rate movies, view similar recommendations given by the system and comment on it. The proposed system is a better system than any other existing systems. This system has added the positive features of existing systems and has overcome the drawbacks of existing systems. The system uses all the existing algorithms i.e. content based, context based and collaborative based algorithms. All these algorithms are combined to give more precise result. The following modules are developed as: A. Admin The system admin will add movie in a database, view movies and update it. B. Recommendation Engine This recommendation engine will calculate the similarities between the different users. On the basis of that similarities calculated, this engine will recommend movie to a user. C. Movie Web Service This will allow user to rate movies, comments on movies. This service will also show the movie recommendation to the users. D. Android User The android user can rate a movie, can comment on any movie, and can see similar movies recommended by other users who are similar to this user.

The performance evaluation of each classifier is discussed in a tabular manner and the classifier with best accuracy discovered. All the weather prediction techniques are demonstrated by applying the classifiers on a dataset namely weatherdata.csv in the Weka tool. The dataset consists of 32686 Values with 10 attributes naming temperature, heat index, humidity, pressure, wind direction, wind speed, precipitation, gust speed, sea level pressure, conditions. The conditions attribute acts as a class label. There are 28 attributes in the dataset. Those classes are Smoke, Clear, Haze, Overcast, Scattered Clouds, Shallow Fog, Mostly Cloudy, Fog, Partly Cloudy, Fog Patches, Thunderstorms with Rain, Rain, Light Rain, Light Drizzle, Drizzle, Mist, Volcanic Ash, Thunderstorm, Light Thunderstorms with Rain, Light Thunderstorm, Squalls, Heavy Rain, Light Haze, Sandstorm, Widespread Dust, Funnel Cloud, Heavy Thunderstorms with Rain, Heavy Thunderstorms with Hail, Light Rain Showers. The steps followed in the weather prediction system are depicted as a Flow Chart in the Figure 1.



Weather Prediction can be done by using various classifiers in data mining techniques. Some of the techniques that assist in Weather Prediction are discussed in the following. Each technique has its own advantages and disadvantages and each of the classification technique becomes handy depending upon the requirement and the conditions. Naïve Bayes: This Naïve Bayes classifier depends on easiest Bayesian system models. This classifier works on Bayes theorem. It predicts the probabilities for each record to have membership in a class. This classifier is exceptionally versatile requiring various parameters in an issue. It is based on conditional probability and the attributes however independent with each other. The class with highest probability is known as Maximum a Posteriori (MAP).

IMPLEMENTATION

FUNCTIONALITY

We can divide our process in two modules namely:

1) Weather Mining
2) Recommendation

B. Weather Mining

Data collection: We have collected weather data from WORLD DATA CENTER for climate, Hamburg. We have decided to use NWS API for data collection in future. Data formatting and cleaning: We have converted our data from .NC (netcdf) format to .CSV (comma-separated values) format because WEKA supports .CSV format.

Clustering: Using WEKA, we have performed clustering on weather data to draw inferences.

Recommendation: We have planned to use recommendation algorithm as user to location collaborative algorithm similar to user to item collaborative algorithm. This algorithm uses user location (N*M) metrics.

Visualization: To generate visualization for user, we have used NOAA weather and climate tool kit.

For the part of the implementation, on which your project focused most, which algorithms you implemented or used and if any modifications were needed to those algorithms or if you did some initial pre-processing, discuss here for the other phases of data mining, discuss brief. E.g., if you focused most on visualization, you can talk about: which data (Example: downloaded from some website put the URL here; did some survey, then talk about how you did the survey etc) collection approach was used in the project?

C. Recommendation: Extract the location of the user. Extract the destination of the user and then recommend the best path according to the conditions.

CONCLUSION AND FUTURE SCOPE

CONCLUSION

Traditionally, weather forecasting has always been performed by physically simulating the atmosphere as a fluid and then the current state of the atmosphere would be sampled. In the previous system the future state of the atmosphere is computed by solving numerical equations of thermodynamics. But this model is sometimes unstable under disturbances and uncertainties while measuring the initial conditions of the atmosphere. This leads to an incomplete understanding of the atmospheric processes, so it restricts weather prediction.

Our proposed solution of using Machine learning for weather predicting is relatively robust to most atmospheric disturbances when compared to traditional methods. Another advantage of using machine learning is that it is not dependent on the physical laws of atmospheric processes. In the long run weather prediction using Machine Learning has a lot of advantages and thus it should be used globally.

FUTURE SCOPE:

The project mainly focuses on forecasting weather conditions using historical data. This can be done by extracting knowledge from this given data by using techniques such as association, pattern recognition, nearest neighbor etc.

Disaster Mitigation: Predicting storms, floods, droughts
Helping those sectors which are most dependent on weather such as agriculture, aviation also depends on weather conditions.

REFERENCES

1) A. Payne and P. Frow, "A strategic framework for customer relationship management," *J. Marketing*, vol. 69, no. 4, pp. 167–176, Oct. 2005.

2) L. D. Xu, "Enterprise systems: State-of-the-art and future trends," *IEEE Trans. Ind. Inf.*, vol. 7, no. 4, pp. 630–640, Nov. 2011.

3) A. Berson, K. Thearling, and S. Smith, *Building Data Mining Applications for CRM*. New York: McGraw-Hill, 1999.

4) K. Coussement and D. V. Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 313–327, Jan. 2008.

5) W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *Eur. J. Oper. Res.*, vol. 218, no. 1, pp. 211–229, Apr. 2012.

6) W. J. Reinartz and V. Kumar, "The impact of customer relationship characteristics on profitable lifetime duration," *J. Marketing*, vol. 67, no. 1, pp. 77–99, Jan. 2003.

7) P. Datta, B. Masand, D. R. Mani, and B. Li, "Automated cellular modeling and prediction on a large scale," *Artif. Intell. Rev.*, vol. 14, no. 6, pp. 485–502, Dec. 2000.

8) D. Popović and B.D. Bašić, "Churn prediction model in retail banking using fuzzy C-means algorithm," *Informatica*, vol. 33, no. 2, pp. 235–239, May 2009.

9) P. Gut, D. Ackerknecht, and S. K. für A. T. am ILE, *Climate Responsive Building: Appropriate Building Construction in Tropical and Subtropical Regions*. SKAT, 1993

10) D. K. Ray, P. C. West, M. Clark, J. S. Gerber, A. V. Prishchepov, and S. Chatterjee, "Climate change has likely already affected global food production," *PLOS ONE*, vol. 14, no. 5, pp. 1–18, 2019.

11) C. H. Ramírez, J. B. Valencia, and C. F. O. Paniagua, "Modelos de Vulnerabilidad Agrícola ante los efectos del cambio climático," *CIMEXUS*, vol. 9, no. 2, pp. 31-48–48, Jan. 2015.

12) K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic Mapping Studies in Software Engineering," in *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, Swinton, UK, UK, 2008, pp. 68–77

13) CGIAR, "Encuesta rural sobre intervenciones climaticas inteligentes," 2016. [Online]. Available:
14) http://cac.foodsecurityportal.org/regional-sub-portal-blog-entry/latinamerica/886/riesgo-y-resiliencia. [Accessed: 17-Oct-2018].

15) R. P. Jose *et al.*, "Assessing the Vulnerability of Agricultural Crops to Riverine Floods in Kalibo, Philippines using Composite Index Method," in *GISTAM*, 2017.

16) CDKN, "Agricultura, Vulnerabilidad y Adaptación: metodología para medir la vulnerabilidad Del sector agrícola - Climate and Development Knowledge Network," 2011. [Online]. Available: http://cdkn.org/project/agricultura-vulnerabilidad-adaptacioncuenca- alta-cauca/. [Accessed: 10-Jul-2018].