IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

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

[1]Wrushabh S. Sirsat
[1]Student
[1]Sant Gadge Baba Amravati University

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 [1]. 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 [2]. 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 [3]. Fig. 1 presents the approaches around *Vulnerability Analysis* and *Climate Change Adaptation* published since 2010. This systematic mapping was developed using the methodology proposed by Petersen [4], 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 [5]. 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 [6]. 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 [7]. Finally, Agriculture, Vulnerability, and Adaptability (AVA) [8] 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 analyzes . The enormous amount of non-trivial work represents a high time of analysis.

The above implies having sufficient and relevant data sources for such analyzes. 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 analyzes 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”) [10].

Literature Review

In the paper presented by A H M Jakaria et.al [1], a study was made by collecting the weather data of Nashville in Tennessee and data of surrounding cities. The training data set consisted of two months worth of weather data of July and August 2018. Many machine learning models were implemented such as Extra Tree Regression, Random Forest Regression, Support Vector Regression and Ridge Regression. They have found Random Forest Regressor to be a better regressor as it ensembles multiple decision trees while making decision. Their evaluation results have shown that machine learning models can give accurate results comparable to the traditional models.

In a study conducted by Man Galih Salman et al [2], a comparison was made among different deep learning models such as Convolutional Networks, Conditional Restricted Boltzmann Machine, Recurrent Neural Network. The dataset for training and testing the models have been collected from the indonesian Meteorological department. Rainfall is the feature being predicted while keeping the other variables independent. RNN method was found to give adequate accuracy when compared to the other models temperature, dewpoints wind speed, precipitation and many other weather parameters were used to train the neural network model. The data required for training the model has been obtained from National Climate Data Center from the year 2007 to 2017. After using the data to train the model using LSTM algorithm, it was found that LSTM algorithm gave substantial results accuracy wise, among other weather prediction techniques.

In the paper presented by E B Abhrahamsen et al. [4] , two different neural network models were studied. One is an Artificial Neural Network and the other is an Artificial Neural Network with Exogenous input. Four different models were built each with a prediction period of 1, 3,6 and 12 hours. All the ANN models use the ReLu as the activation function in the hidden layer and a linear activation function in the output layer. IN the ANN models only temperature was used as the input to the network. Where as in ARX models another feature, precipitation was introduced as another input and improved the prediction accuracy. So they have concluded that with the introduction of more inputs accuracy of these models can be improved.

Holmstrom et al. proposed a technique to forecast the maximum and minimum temperature of the next seven days, given the data of past two days [6]. They utilized a linear regression model, as well as a variation of a functional linear regression model. They showed that both the models were outperformed by professional weather forecasting services for the prediction of up to seven days. However, their model performs better in forecasting later days or longer time scales. A hybrid model that used neural networks to model the physics behind weather forecasting was proposed by Krasnopolsky and Rabinivitz [7]. Support vector machines was utilized for weather prediction as a classification problem by Radhika et al. [9]. A data mining based predictive model to identify the fluctuating patterns of weather conditions was proposed in [11]. The patterns from historical data is used to approximate the upcoming weather conditions. The proposed data model uses Hidden Markov Model for prediction and k-means clustering for extracting weather condition observations. Grover et al. studied weather prediction via a hybrid approach, which combines discriminatively trained predictive models with deep neural networks that models the joint statistics of a set of weather-related variables [5].

Montori et al. used the concept of crowdsensing, where participating users share their smart phone data to environmental phenomenons [8]. They introduced an architecture named SenSquare, which handles data from IoT sources and crowdsensing platforms, and display the data unifiedly to subscribers. This data is used in smart city environment monitoring. However, none of these works use the idea of combining data from neighboring places.

Climate change model construction

Global climate models incorporate the latest scientific understanding of the physical processes at work in the atmosphere, oceans, and Earth's surface and how they are all interconnected. A global climate model can produce projections of precipitation, temperature, pressure, cloud cover, humidity, and a host of other climate variables for a day, a month, or a year. Models were selected for use in the research based on a rigorous set of criteria, including the model's effectiveness in reproducing past and current climate within our region. If a model can replicate known historical conditions in the Asia Pacific region, we have higher confidence when projecting the future climate. While there are many global climate models available, the analyses described on this paper used a core set of models that performed best.

Comprehensive climate models are constructed using expert judgments to satisfy many constraints and requirements. Overarching considerations are the accurate simulation of the most important climate features and the scientific understanding of the processes that control these features. Typically, the basic requirement is that models should simulate important features, particularly

surface variables such as temperature, precipitation, windiness, and storminess. This is a less-straightforward requirement than it seems because a physically based climate model also must simulate all complex interactions in the coupled atmosphere– ocean–land surface–ice system manifested as ultimate variables of interest. The models should also be capable of simulating changes in statistics caused by relatively small changes in the Earth’s energy budget that result from natural and human actions. Climate processes operate on time scales ranging from several hours to millennia and on spatial scales ranging from a few centimeters to thousands of kilometers. Principles of scale analysis, fluid dynamical filtering, and numerical analysis are used for intelligent compromises and approximations to make possible the formulation of mathematical representations of processes and their interactions. These mathematical models are then translated into computer codes executed on some of the most powerful computers in the world. Available computer power helps determine the types of approximations required.

As a general rule, growth of computational resources allows modellers to formulate algorithms less dependent on approximations known to have limitations, thereby producing simulations more solidly founded on established physical principles. These approximations are most often found in "closure" or "parameterization" schemes that take into account unresolved motions and processes and are always required because climate simulations must be designed so they can be completed and analyzed by scientists in a timely manner, even if run on the most powerful computers.

The increase confidence in attribution of global scale climate change to human induced greenhouse emissions, and the expectation that such changes will increase in future, has lead to an increased demand in predictions of regional climate change to guide adaptation. Although there is some confidence in the large scale patterns of changes in some parameters, the skill in regional prediction is much more limited and indeed difficult to assess, given that we do not have data for a selection of different climates against which to test models. Much research is being done to improve model predictions, but progress is likely to be slow. Despite their limitations, climate models provide the most promising tool of providing information on climate change. This will include assessments of the ability of the models used to predict current climate, and the range of predictions from as large a number of different models as possible.

Climate models have shown steady improvement over time as computer power has increased, the understanding of physical processes of climatic relevance has grown, datasets useful for model evaluation have been developed, and the computational algorithms have improved. Model ranking according to individual members of this basket of indicators varies greatly, so this aggregate ranking depends on how different indicators are weighted in relative importance. Nevertheless, the conclusion that models have improved over time is not dependent on the relative weighting factors, as nearly all models have improved in most respects. The construction of metrics for evaluating climate models is itself a subject of intensive research.

Model selection criteria

The ability of the model to simulate the present climate conditions is an important consideration taken into account in the selection of GCM to be used in this study. The performance of individual GCMs may differ for individual climate variables as well as for different regions of the world. Typically, GCMs are validated for their ability to reproduce spatial patterns (McKendry et al. 1995, Huth 1997) of selected variables and their annual cycles (Nemesova & Kalvova 1997, Nemesova et al. 1999). GCMs are run by a number of research centres. Some differences exist among the models, which result in various climate sensitivities in a range likely between 1.0°C and 3.5°C (Figure 1) with a best estimate value of 2.0°C over a 80 year simulation period. However, selecting appropriate models is difficult especially when many models are available with different projection results.

Vintage. Recent models are likely to be more reliable as they incorporated the latest knowledge in their construction.

Resolution. Recent models tend to have finer resolution than older models. Higher resolution models contain more spatial details (eg complex topography, better-defined land–sea boundaries etc) and some key processes of climate variability such as ENSO events are better represented.
However, finer resolution does not necessarily guarantee better model performance.

Validity. Selection of models is based on how well they simulate present day climate. The validity of a model is assessed by comparing observed data with simulated data. The easy method is by 'upscaling' the observed data to a GCM grid size to compare it with the GCM simulated data. Statistical measures such as standard deviation, standard error, etc are useful for this analysis (Giorgi and Mearns 1991; Murphy et al. 2004).

Representativeness of results. Preferably, results from more than one GCM are to be applied in an impact assessment. Selecting some representative GCM results helps in illustrating a range of changes in a key climate variable in the study region. For example, if a number of models show less annual precipitation, no change in precipitation and more annual precipitation, users can choose one for each
cluster of the simulation results for illustrating future potential impacts of their studies.

Other criteria that should be met by climate scenarios if they are to be useful for impact researchers and policy makers are as follows:
□□**Applicability in impact assessments**. They should describe changes in a sufficient number of variables on a spatial and temporal scale that allows for impact assessment. For example, impact models may require input data on variables such as precipitation, solar radiation,
temperature, humidity and wind speed at spatial scales ranging from global to site and at temporal scales ranging from annual means to daily or hourly values.
□□**Accessibility**. They should be straightforward to obtain, interpret and apply for impact assessment. Many impact assessment projects include a separate scenario development component which specifically aims to address this last point.

Hindcasting to gauge accuracy of the model

Hindcasting is a way of testing a mathematical model. Known or closely estimated inputs for past events are entered into the model to see how well the output matches the known results. An example of hindcasting would be entering climate forcings (events that force change) into a climate model. If the hindcast accurately showed weather events that are known to have occurred, the model would be considered successful.

The various IPCC model outputs are evaluated and rated, based on the reliability of their hindcasts in terms of replicating observed conditions. The errors in these hindcasts are computed for individual parameters and for specific regions. Multiple criteria are considered for each parameter, including replication of the means, as well as modelled vs. observed variances and potentially additional measures, such as trends and seasonality. The errors are then used to construct “distances” between the model and observations for each model simulation.

Twenty-three different climate model simulations were undertaken for the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. These model simulations were acquired from the program for Climate Model Diagnosis and Intercomparison and were used to generate climate change projections for the Asia/Pacific region. These model simulations were individually evaluated with respect to their abilities to faithfully reproduce observed seasonal patterns of mean sea-level pressure, temperature, and rainfall over the Asia/Pacific (60–180ºE, 55ºN–25ºS) region for a 30-year period (1961-1990).

Data availability and cost

WORLDCLIM data set was chosen in favor of the data from the IPCC Distribution center because of the higher resolution data set available in WORLDCLIM that can be readily applied for impact studies. WORLDCLIM provides a number of outputs that have been downscaled from Global Circulation Models at no cost and the database has 400 times higher spatial resolution than previously available surfaces (New et al., 2002). Data is based on interpolated climate surfaces for global land areas in four different spatial resolutions; 30 seconds (about 0.86 km2 at the equator), 2.5 minutes, 5 minutes, and 10 minutes (about 344 km2 at the equator) and can easily be exported to GIS.

IPCC Data Distribution Centre

The following information has been extracted from IPCC-TGICA (2007).IPCC Data Distribution Centre (DDC) was established in 1998, following a TGICA recommendation to facilitate the timely distribution of a consistent set of up-to-date scenarios of changes in climate and related environmental and socio-economic factors for use in climate impact and adaptation assessment.

The DDC is a shared operation between the British Atmospheric Data Centre in the UK, the Max-Plank Institute for Meteorology in Germany, and the Center for International Earth Science Information Network at Columbia University, New York, USA. It provides three main types of data and guidance, which meet certain criteria established by the TGICA. They are: socio-economic data and scenarios that follow the assumptions used for the construction of SRES emission scenarios (see Box 2); climate observations and scenarios; and data and scenarios for other environmental changes.

The climate observation data set contains 0.5° latitude/longitude gridded monthly global land surface of 11 climate variables for the period 1901–2000 supplied by the Climatic Research Unit. The data set can be used to examine climate variability over the 20th century, to evaluate the simulations of various GCMs over the period 1961-1990, and to combine observed data with GCM projections. The variables are precipitation and wet-day frequency; mean, maximum and minimum temperatures; vapour pressure and relative humidity; sunshine percent and cloud cover; frost frequency; and wind speed. The monthly averaged results from climate change simulations by a number of climate modelling centres are also available. The results have been extracted from transient AOGCM simulations which include greenhouse gas only and combined greenhouse gas and sulphate aerosol forcings. Ensembles and time-slice experiments are also being provided. Main variables that are available are cloud cover, diurnal temperature range, precipitation, solar radiation, mean temperature, minimum temperature, vapour pressure, and wind speed.

IPCC TGICA applied some criteria for GCM experiment results could be placed at the DDC, which follows criteria set by Parry (2002). The climate models should

be full 3D coupled ocean-atmospheric GCMs,

be documented in the peer-reviewed literature,

have performed a multi-century control run (for stability reasons) and

have participated in the Second Coupled Model Inter-comparison Project (CMIP2).

In addition, the models should

have performed a 2×CO2 mixed layer run,

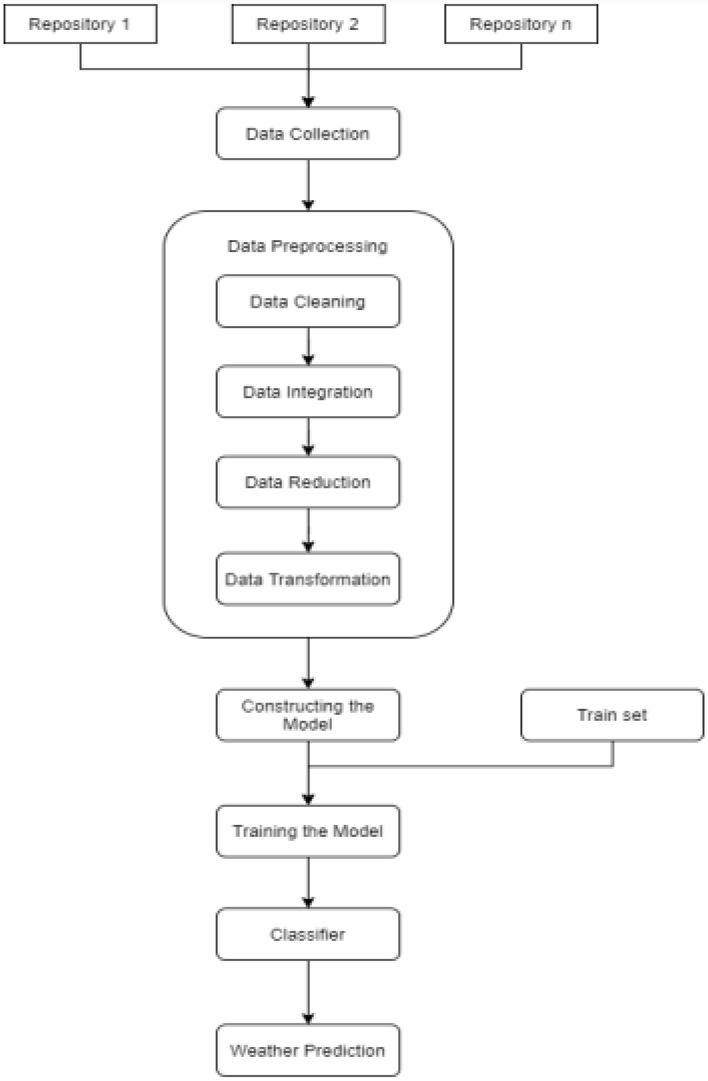
have participated in the Atmospheric Model Inter-comparison Project (AMIP),

have a resolution of at least T40, R30 or 3° × 3° latitude/longitude and

consider explicit greenhouse gases (eg CO2, CH4 etc).

Proposed Work

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).

Objective

1. Do develop and implement a mechanism for weather forecasting
2. To implement old data performer using machine learning
3. To implement proposed mechanism for next 10 years of weather prediction
4. To study and implement various prediction algorithm for weather forecasting

References

[1] 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

[2] 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.

[3] 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.

[4] 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

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

[6] IFPRI, "The international model for policy analysis of agricultural commodities and trade (impact): Model description for version 3," 2015. [Online]. Available: http://www.ifpri.org/publication/international-model-policyanalysis-agricultural-commodities-and-trade-impact-model-0.
[Accessed: 18-Oct-2018].

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

[8] 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].

[9] P. Kathiravelu and A. Sharma, "A dynamic data warehousing platform for creating and accessing biomedical data lakes," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10186 LNCS, pp. 101–120, 2017.

[10] E. Rahm and H. H. Do, "Data Cleaning: Problems and Current Approaches," *IEEE Data Engineering Bulletin*, vol. 23, p. 2000,2000.

[11] S. Sadiq and M. Indulska, "Open data: Quality over quantity," *International Journal of Information Management*, vol. 37, no. 3, pp. 150–154, Jun. 2017.