World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:7, No:12, 2013

Comparative Study - Three Artificial Intelligence Techniques for Rain Domain in Precipitation Forecast

Nabilah Filzah Mohd Radzuan, Andi Putra, Zalinda Othman, Azuraliza Abu Bakar, Abdul Razak Hamdan

***Abstract*—**Precipitation forecast is important in avoid incident of natural disaster which can cause loss in involved area. This review paper involves three techniques from artificial intelligence namely logistic regression, decisions tree, and random forest which used in making precipitation forecast. These combination techniques through VAR model in finding advantages and strength for every technique in forecast process. Data contains variables from rain domain. Adaptation of artificial intelligence techniques involved on rain domain enables the process to be easier and systematic for precipitation forecast.

***Keywords*—**Logistic regression, decisions tree, random forest, VAR model.

I. Introduction

PRECIPITATION is a process where water release from cluster cloud in form of rain, sleet, sleet ice, snow, or hail. Precipitation is not going down in quantity that same in all place in earth surface. Forecasting precipitation for rain scale cluster generally challenges for meteorologists due to major sources that difficult to be predicted directly. There are two source complexity including formation prediction which requires conversion humidity that enable mistake happen during forecast especially in full domain, and error in application on model involves cause inaccurate result retrieval [1]. Malaysian Meteorological Department finds out especially flow of weather variation rain scale cluster and typhoon increase important in annual precipitation notice at East Malaysia [2].

Precipitation forecast can be done through artificial intelligence techniques. There are various techniques in artificial intelligence including logistic regression, decision trees, random forests, neural network, data mining, and others. This paper presents techniques in artificial intelligence which help in precipitation forecast process, i.e. logistic regression, decision trees, and random forests. These selected techniques are believed to be helping meteorologists meet model variables and parameters that can produce better forecast precipitation [1].

M. R. F. Nabilah and P. Andi are with Center for Artificial Intelligence Technology, Faculty of Technology and Information Science, Universiti Kebangsaan Malaysia, 46300 Bangi, Selangor, Malaysia (phone: 603-8921-6182; fax: 603-8921-6184; e-mail: {nabilah.filzah, drandiputra}@ gmail.com).

O. Zalinda and A. B. Azuraliza are with Data Mining and Optimization, Center for Artificial Intelligence Technology, Faculty of Technology and Information Science, Universiti Kebangsaan Malaysia, 46300 Bangi, Selangor, Malaysia (e-mail: {zalinda, aab}@ftsm.ukm.my).

II. Related Work

The rain cluster distribution data in a same field period of collection time are collected by central of meteorology for research proposes[3]. These data represents a variable that is useful to help in the forecasting process. The connection of relationship of forecast variables not accumulated directly for inducing orographically rain as in Fig. 1 [4].

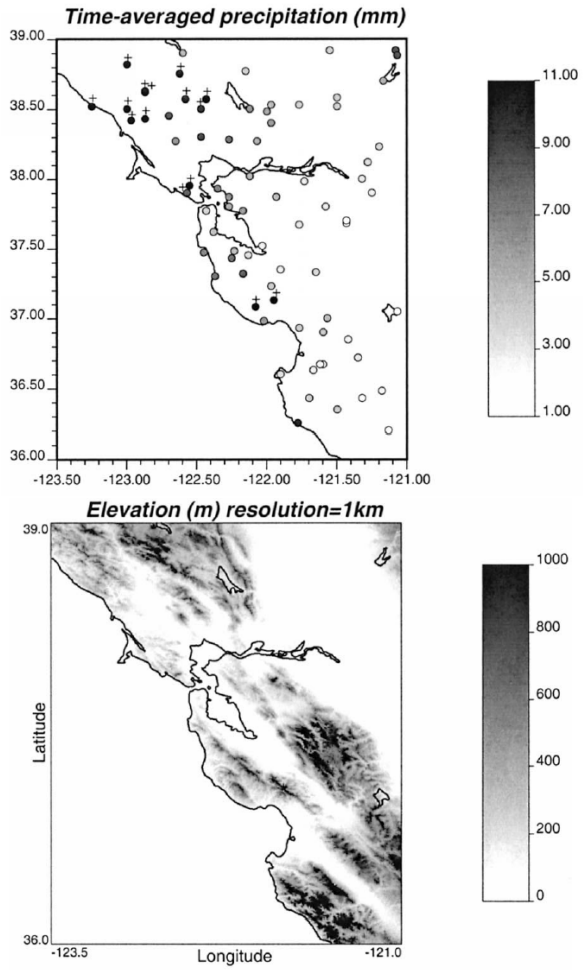

Fig. 1 Example of variable relation in forecast

This rainfall data distribution is used to make a forecast, based on the particular study's objective. The techniques presented in this paper are logistic regression, decision tree and random forest.

In general, logistic regression is a traditional classification method that has been used extensively in various applications including classification document, vision computer and others.

Seldom, logistic regression owns selection features in classification framework uses $\ell 1$ confirmation norm and produces results that attract many applications involving high dimensional data[5]. Two norms have been widely used especially regression model for logistic flat in order to distinguish the optimization problem of unconstrained [5]. An optimization algorithm likes Newton method and conjugate gradient method is involved formulation [6]. Logistic regression formulation involving

$$\text{Prob}(b|a) = 1/\ (1+\exp(-b(w_r a + c)))[5]$$

where, $\text{Prob}(b|a)$ is conditional probability for label b, given sample a, $w \sum Rn$ was heavy vector, and $c \sum R$ was block $wra+c=0$ interpreted as hyperplane in space feature, of which $\text{Prob}(b|a) = 0.5$. The conditional probability $\text{Prob}(b|a)$ larger from 0.5 when $wra+c$ have equivalent such as b, and on the other hand less from 0.5[5].

Combination algorithms work on high dimensional problem data or domain data that are large such as rain domain; in logistic regression. The advantages of logistic regression law are the probable result can be given clearly and regression coefficient can be defined easily [7]. Several studies [1], [5], [7], [8] attracted to use logistic regression law to solve related forecasting problem.

Logistic regression is used to generate a landslide susceptibility map. Logistic regression usage captures not only fact approach with ease and tight assumptions that are needed by multivariable of other statistical method. But, it also demonstrates the possibility of the existence of combination technique to facilitate interpretation model for forecasting of precipitation [9]. Precipitation usually happens on land geology influence and rainfall distribution. This is because lithological and structural often variation lead to a difference in strength and humidity stone and land in the area involved. Fig.2 shows a landslide inventory map in a particular area. This landslide map useful in implementing a precipitation forecast to the area involved.

Logistic regression link predicts variables in incident precipitation through geography cells. The relationship result in a map which shows the probability incident, perhaps can repeat further in the future [10]. Deposition caused loss of millions of Ringgits in property damage during the last decade. Water, land and rain were factors involved in precipitation [10]. Amount of moisture antecedent result in a large rate pore pressures that are increasing in earth [11].

Precipitation in harmful context map landslide is prepared by using ratio model frequency and logistic regression. Defection right is important deposition location that forecast for the occurrence of probability analysis landslide [8]. Increase precipitation occurring on winter and hot. Increase deposition is due to frequency increase precipitation and intensity on effect number that eased o every rainy day[12].

Forecast of precipitation can also be done through a decision tree technique. The decision tree is one of a classification technique based on particular law. Rules produced during learning process use to classify the hidden data that are hidden and afford display relationship between qualities in the diagram. Decision trees are used in dataset that is large. This application is used in building classification trees with a careless set parameter which enables probability estimates from this tree through the frequency branch [13]. The relation of decision tree with learning depression analysis is used to check the result effects. Models based on tree can approach parallel class probability axis with arbitrarily close border [14].

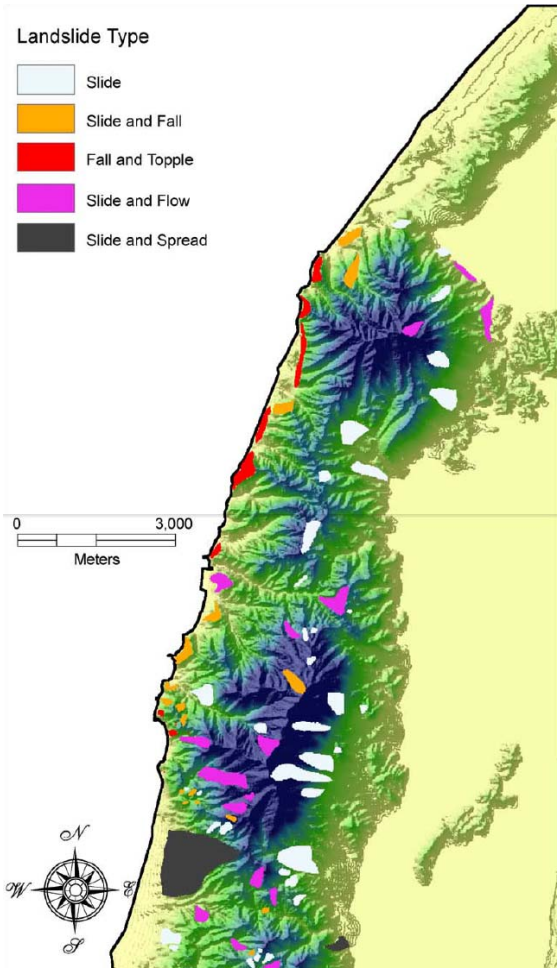

Fig. 2 Study landslide area in inventory map. Stone fall and topple dominant during main highway which carried out parallel to Japan Sea[9]

Decision trees are used in weather forecast through picture radar. A quantitative application based on radar has been confined because of different resources and uncertainty in rain estimate process. Therefore, decision trees help in partitioning radar use data. The partitioning processes include data collection classification, semantic level, conditional labeling, and algorithm cluster unsupervised [15]. Algorithm announces to build decision tree by creating one modus root and determines to all data training for her, chooses the best nature fragmentation, adding a branch as per node root value

separation, and split the data into subsets that mutually exclusive for its bifurcation.

Radar shows the force related to rainfall intensity and flashes hit layer melt for assessment precipitation. On the other hands, amplification in reflectance occurs due to ice shushes and leads this for assessment precipitation [15].

Classification process perpetrated upon the radar picture as practiced by a decision tree technique. This process produces network tree for each data that involved as in Fig. 3.

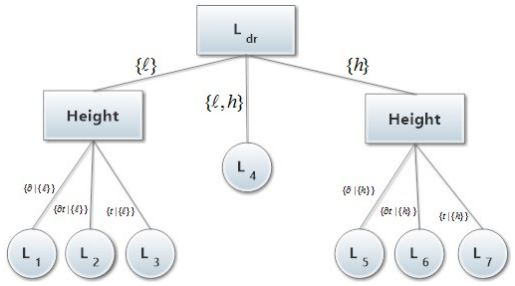

Fig. 3 Example of network produced from technique decision trees [15]

Besides, process technique intelligence is applied to retrieval information until produced decision towards objective of the study involved as in Fig. 4.

Algorithm	Accuracy			
	Rain	Snow	Bright Band	Average
Naive Bayes	75.3%	68.5%	98.6%	80.8%
SVM (SMO)	79.0%	75.1%	98.9%	84.3%
Neural Network	85.2%	75.8%	99.0%	86.7%
KNN	94.4%	94.2%	99.3%	96.0%
ID3 (Uniform)	87.0%	84.0%	97.3%	89.4%
LID3 (Uniform)	89.4%	85.2%	99.3%	91.3%
LID3 (Entropy-Min)	92.3%	87.5%	99.4%	93.1%
LID3 (K-Means)	93.5%	89.0%	99.4%	94.0%

Fig. 4 Example of decision produced by the intelligence technique. Comparison of result with LID3 and Machine Learning Algorithms from Weka [15]

Analysis decision trees develop a basis of regulation that is used for classification of image [16]. Image segmentation combination in various resolution and tree analysis bring result in facilitating selection by including variables and help in deciding suitable scale image analysis.

Apart from that, relationship precipitation with community mammal helps in predicting precipitation from a set ecomorphological mammal features in a community, that is large [17]. The precipitation affects the mammal habitat. Process decision tree helps determine the habitat mammal based on altogether pattern rainfall distribution where it helps to control and preserve this habitat. This situation shows rainfall distribution influence precipitation in the areas involved.

Rain is one domain to predict soil characteristics through decision tree. A method based on decision tree through a tool modernization Cubist were received from Australian Soil Resources Information System (ASRIR), shows display soil characteristics that are continuous for process forecasting. An environment that is changing uses Cubist in selection after considering utility through several measures of intelligent [18]. Power forecast for variable is investigated through correlative study with interest variables involved.

Apart from logistic regression and decision trees, random forest is also one of the techniques which is used to carry out forecast of precipitation. Random forest is closely related to logistic regression and decision tree. Random forest is a combination tree forecaster whereby it depend on random vector values with the same distribution per tree [19]. The random forest technique gives opportunity that is more often exploitation problem in making forecasts deposition [1]. Random forest is a new entry in field data port and designed in result predictions that right [19]. This random forest builds multiple trees, with every tree brought up or developed with a random subset [20].

Random forest uses to help identify regimes that can represent various types of convection, geography, location or condition synoptic [21]. Existence of random forest; a union that trees correlated weak used to place interest by forecaster and provide a benchmark for algorithmic performance that is potential [21]. This technique mark the best way for forecaster to start and distinguish based on day, hour, and location of different weather features.

This technique helps to estimate interest for every variable based on the degradation in classification performance when variable values in process at random. Fig.5 shows an example of technique in random forest that chooses based on a variable. Random forest is one of the techniques used by meteorologist from the meteorological station to predict area that is frequently burnt [22]. This shows the technique used in finding factor that is most important between that lists facilitate at one go forecaster meteorology carrying out the work.

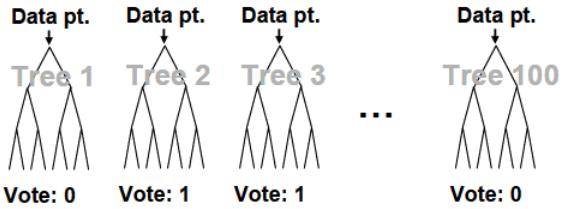

Fig. 5 Diagram concept a random forest weak, weak a group correlated decision tree "vote" in classification for every data point [21]

III. Comparative Analysis

The analysis does compare all three techniques involved namely logistic regression, decision tree, and random forest in rain domain for precipitation forecast.

Logistic regression performances reach multiple analysis techniques regression to research situations involved where result variable is obsolete. In practice, a situation which involves absolute result is fairly common. The forecast can be

done for success that dichotomy result or failure, or increase or decline.

TABLE I
DIFFERENCE ALL THREE MODEL

Technique	Method	Strength	Limitation
Logistic regression	Predict presence or absence a feature or result based on values a set predictor variables	Can estimate models use block entry variables or from methods including forward conditional, forward LR, forward Wald, backward conditional, backward LR.	Need size data that is large to be adapted to get number that is sufficient in both category c elebration variable.
Decision tree	Build models that could forecast value of each target variable based on a few variables included.	Easy for understand and be defined, can use data that is small as preparation, able to handle both categori cal data and numerical, and display as good as in large data within minutes.	Create trees complex that ex cess that not nicely generaliz e data. It cited as "overfitting".
Random forest	Create multi bo ot strapped regr ession tree with out pruning and average yield product except on every tree which expanded with random in forecaster.	Grow or achieve a large amount of stand that do not adapt over fit in data, and at random selection forecaster l ower tendency and provide model that is better for forecast.	More method "black box", that very demanding in form of time and computer resources.

By the basis, regression model logistics consist of:

i. A variable to every 30-50 line data.
ii. Variable applies as it is core model and be defined by a model.
iii. Logistic regression should be having minimum variable, and
iv. Schedule classification as variable that does not depend

Therefore, through this process logistic regression can achieve performance that is positive. Domain rain which has many features or variables can be applied in this technique. Variable in domain rain is dependent on variable namely dummy variable where a record 0 (do not vote) and 1 (vote). Distribution possibility this forces case logistics estimated located between 0 and 1. Visible distribution forecast graph in Fig. 6 which shows that a variable that is selected is easy and short to being interpreted for decision making. Usually, accuracy in decision making for logistic regression seen in value whole range threshold values including thresholds which is resulted in number become very large negative or false positive.

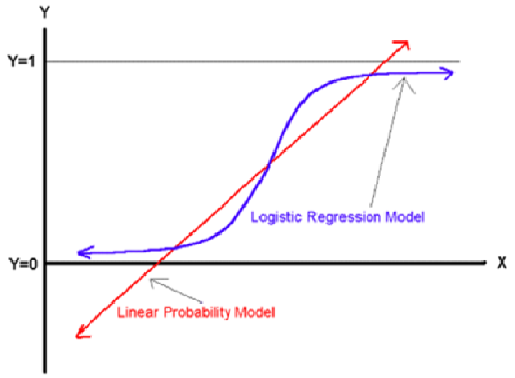

Fig. 6 Regression distribution model diagram logistics

Performance decision tree link visible for domain models rain through analysis achievement data. Decision tree involving three (3) nodes namely root node, internal node, and leaf node. Fraction node from node root to internal node and to leaf node show division variable that is easy to get results analysis for domain rain as in Fig. 7.

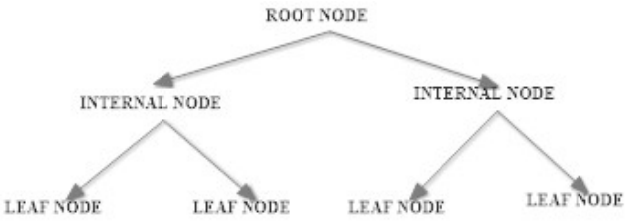

Fig. 7 Node fraction diagram in decision tree

Random forest is a tool that is effective in making forecasts. This technique not ever fit due to the law of large (Law Large Number). Accurate injection in random enable this technique makes classification right and regresses. Apart from that, the framework in the strength question forecast and relations variable giving understanding into capacity random forest prediction. Forecast using bag estimate (out of beg) is real unless theoretical values strength and correlation.

A variable that is selected or involved in this process will process every random branch to every variable so that all included in the forecast. Despite this technique a bit hard but no exemptions occurred against variable as in Fig. 4.

Every model possesses strength and respective weakness. Difference from method aspect, strength, and limit for all three regression model logistics, decision tree, and visible random forest are shown in Table I.

Every technique owns result or variable forecast product that is different. A method model must be used on all these three technique so that it produces forecast precipitation that right namely without error. This combination carried out by way of adapting techniques involved in vector integration (VAR) model.

Estimate model Keynes develops a large into models simultaneously on a large scale with hundred variables in several cases. Criticism on program Keynesian theory and empirical reasons originates in year 1970 [23]. Model VAR

perceived as an alternative approach to approach simultaneous equation.

Therefore, vector autoregression (VAR) is used in multivariate time series analysis. VAR is the model that is most successful, flexible and easy to be applied in model analysis multivariate time series [23]. Forecast from fairly flexible models VAR because it made conditional in potential future lanes to determine variables in the model.

Model VAR involving a system that has various equations where all variables are treated as endogenous. There is per an equation variable namely dependent variable. Every equation leaves values all variables involved because it's dependent variable [24].

Basically, the basis for the system VAR own procedure as follows [23]:

i. Model VAR was a form reduced. There are no variables a contemporary or that same entered on the right side.
ii. All variables involved or include treated or categorized as endogen. Every dependent variable between one another.
iii. ModelVAR is a theoretical model where one form reduced does not show interaction interconnected within involved.
iv. Structural unobserved shock e- shock composed. Lack forms in this application contain structural shock from all similarity of structure. When VAR estimated with actual data, found shock estimates join, and this state cited as we distinguish between incidents from structure represented by ∑. Impulse celebration analysis observes shock may be over, yet have no economic interpretation.
v. A VAR can be used for forecasting, but not for assessment structural analysis and policy.

All three models could be adapted in two means namely algorithmic traditional forecasts, and simulation based forecasting that occur in VAR model.

Algorithmic traditional forecast consider ago problem for future forecasting Yt when parameter ∏ VAR (p) process was assumed to be known and not conditions deterministic or variable sexogenous. Best linear forecaster, in matter error set up is minimum (MSE), where Yt+1 or 1 move forecast based on information that there is in time T [23];

$$Y_{T+1/T} = {}_C + \prod_1 Y_T + \ldots + \prod_p Y_{T-p+1}$$

Resulted displays graph forecast for data that involved like in Fig. 8. This graph can compare all three model involved.

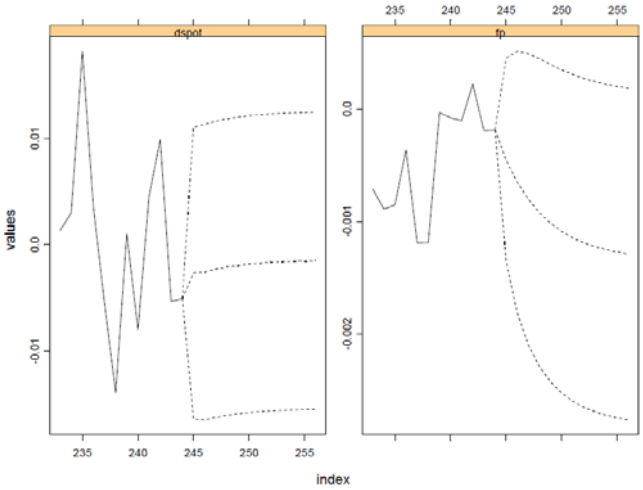

Fig. 8 Algorithmic traditional forecast example of product diagram [23]

This procedure repeated several times to obtain the best yield based on simulation predictions and distance believe variable. However, in move that larger, period for stable forecast values. System VAR is used with set error which is nil. If this result system VAR same to the technique involved, so result retrieval from the process will produce variables that are same and also coefficient that is same.

IV. Discussion

Comparison between all three models involved namely logistic regression, decision tree, and random forest; show every model owns advantage and weakness in making forecasts to rain domain.

The logistic regression advantage is the technique that chooses significant variable and loading in logistic curve for forecasting. A variable that is selected should be easy and short to be interpreted but do not allow for exploration or interpretation that is deep for relationship variable. This technique helps in rain domain variable significant for variables involved that are irregular or random in from the original source.

Advantage decision tree on the other hand is that this technique produces a visual model which brought prediction and can analyze the significant variable and also the relationship though the location tree. Therefore, this technique helps in displaying visual model to every variable involved in domain starting from a basis (most important) until leaf (trifling).

Advantage random forest is a random collection tree which offers a better method from one plant supply. This technique helps to display variable trees in more details so every variable in the domain not exempted from being analyzed.

Despite this advantage technique facilitate process forecast, but it also owns lack or limitation. Logistic regression just picked significant variable for rain domain without other variables. Technique decision tree on the other hand give visual on the whole on a variable until difficult to process to

all variable that occur in the domain. While, technique random forest arranges specification until it not easy to be analyzed.

Advantage and weakness to every technique involved unified in a system or model VAR for turnover that is better to get forecast precipitation for rain domain. Model VAR is seen as able combine to all technique to be doing together forecast with zero error.

Adaptation techniques with domain help in prophesied that altogether prepared for an incident. Precipitation could result in a disaster for the area involved if unheard on factor and when this incident will happen.

V. Conclusion and Suggestion

Precipitation is release water from cluster cloud that inconsistent in area that is different. This scale cluster made variable for rain domain to gauge frequency level for occur once of forecast precipitation. This process forecast can be done by adapt selected a few technique namely logistic regression, decision tree, and random forest. These techniques can help in conducting process forecasting. Variables involved fully-utilized in this all three technique through VAR model. VAR model help in distinguish between them for every techniques and it also help in mutually connection technique to own forecast result that is accurate without error.

References

[1] D. J. Gagne, A. McGovern, and M. Xue, "Machine learning enhancement of storm scale ensemble precipitation forecasts," in *Proceedings of the 2011 workshop on Knowledge discovery, modeling and simulation - KDMS'11*, 2011, p. 45.

[2] D. J. Yik, S. Moten, M. Ariffin, and S. S. Govindan, "Trends in Intensity and Frequency of Precipitation Extremes in Malaysia from 1951-2009," 2009.

[3] S. Moten and M. Ariffin, "Impact of Tropical Cyclones in the West North Pacific and South China Sea on the Asian Monsoon Rainfall during the Pre-monsoon, Monsoon and Post-monsoon Seasons Subramaniam Moten and Munirah Ariffin," 2011.

[4] Phaedon c. Kyriakidis, J. Kim, and norman l. Miller, "Geostatistical Mapping of Precipitation from Rain Gauge Data Using Atmospheric and Terrain Characteristics," *American Meteorology Society*, pp. 1855–1877, 2001.

[5] J. Liu, J. Chen, and J. Ye, "Large-scale sparse logistic regression," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD'09*, 2009, p. 547.

[6] S. Boyd, *Convex Optimization*. 2004, pp. 1–730.

[7] L. Shen and E. C. Tan, "Dimension reduction-based penalized logistic regression for cancer classification using microarray data.," *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 2, no. 2, pp. 166–75, 2005.

[8] S. Lee and B. Pradhan, "Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models," *Landslides*, vol. 4, no. 1, pp. 33–41, Jul. 2006.

[9] L. Ayalew and H. Yamagishi, "The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan," *Geomorphology*, vol. 65, no. 1–2, pp. 15–31, Feb. 2005.

[10] G. C. Ohlmacher and J. C. Davis, "Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA," *Engineering Geology*, vol. 69, no. 3–4, pp. 331–343, Jun. 2003.

[11] W. c. Haneberg and A. onde. Gokce, "Rapid Water-Level Fluctuations in a Thin Colluvium Landslide West of Cincinnati, Ohio," *U.S.Geological Survey Bulletin, Professional Paper 2059-C*, p. 16, 1994.

[12] P. Zhai, X. Zhang, H. Wan, and X. Pan, "Trends in Total Precipitation and Frequency of Daily Precipitation Extremes over China," *journal of climate*, vol. 18, pp. 1096–1108, 2005.

[13] C. Perlich, J. S. Simonoff, and F. Provost, "Tree Induction vs . Logistic Regression: A Learning-Curve Analysis," *journal of machine learning research*, vol. 4, pp. 211–255, 2003.

[14] F. Provost and P. Domings, "Tree Induction for Probability-based Ranking," *Kluwer Academic Publishers*, vol. 18, no. 4, pp. 1–22, 2002.

[15] D. R. McCulloch, J. Lawry, M. Rico-Ramirez, and I. Cluckie, "Classification of Weather Radar Images using Linguistic Decision Trees with Conditional Labelling," in *2007 IEEE International Fuzzy Systems Conference*, 2007, vol. 3, pp. 1–6.

[16] A. S. Laliberte, E. L. Fredrickson, and A. Rango, "Combining Decision Trees with Hierarchical Object-oriented Image Analysis for Mapping Arid Rangelands," *Photogrammetric Engineering and Remote Sensing*, vol. 73, no. February, pp. 197–207, 2007.

[17] J. T. Eronen, K. Puolamäki, L. Liu, K. Lintulaakso, J. Damuth, C. Janis, and M. Fortelius, "Precipitation and large herbivorous mammals I: estimates from present-day communities," *Evolutionary Ecology Research*, vol. 12, pp. 217–233, 2010.

[18] B. L. Henderson, E. N. Bui, C. J. Moran, and D. a. P. Simon, "Australia-wide predictions of soil properties using decision trees," *Geoderma*, vol. 124, no. 3–4, pp. 383–398, Feb. 2005.

[19] L. E. O. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[20] A. Liaw and M. Wiener, "Classification and Regression by random Forest," vol. 2, no. December, 2002, pp. 18–22.

[21] J. K. Williams, D. A. Ahijevych, C. J. Kessinger, T. R. Saxen, M. Steiner, and S. Dettling, "A Machine Learning Approach to Finding Weather Regimes and Skillful Predictor Combinations for Short-Term Storm Forecasting," *National Center for Atmospheric Research*, pp. 1–6, 2008.

[22] P. Cortez and A. Morais, "A Data Mining Approach to Predict Forest Fires using Meteorological Data," *Department of Information System*, 2007.

[23] M. T. Series, "Vector Autoregressive Models for Multivariate Time Series," in *Time series analysis*, 1994, pp. 383–427.

[24] J. G. De Gooijer and R. J. Hyndman, "25 Years of Time Series Forecasting," *International Journal of Forecasting*, vol. 22, no. 3, pp. 443–473, 2006.