IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue, 4, Feb 2019 E-ISSN: 2347-2693

Data Mining Approaches to Predict the Factors that Affect the Agriculture Growth using Stochastic Model

P. Rajesh[1]*, M. Karthikeyan[2]

[1]PG Department of Computer Science, Government Arts College, C.Mutlur,Chidambaram, India
[2]Division of Computer and Information Science, Annamalai University, Annamalainagar, India

Corresponding Author: rajeshdatamining@gmail.com

Available online at: www.ijcseonline.org

***Abstract*—** In the recent times, there has been an increasing demand for efficient strategies in the data mining in agriculture prediction. Data mining is equipment to predict effectively by stochastic model sensing concept. This paper proposes an efficient factor that affects the agriculture growth using different data like rainfall, groundwater and temperature by adopting stochastic modeling and data mining approaches. Firstly, the novel model is proposed to predict the factors affecting the growth of agriculture using stochastic model and numerical illustrations are done and the various expected estimation the sternness of the proposed approach.

***Keywords*—** Data Mining, Agriculture productions, Rainfall, Groundwater, Temperature and Stochastic model

I. INTRODUCTION

Data mining is a familiar tool for finding useful information using pre existing datasets. It is used to retrieves different hidden estimations from the dataset. Data mining include various techniques and algorithms to solve different problems [1]. Furthermore, in this paper a proposed new stochastic model using data mining approach. It is used to retrieve various hidden valuable information using rainfall, temperature, groundwater, enterprises, etc.

In the region of 70 percent of the state's populations are mixed up in agricultural activities as this is one of the major revenue of livelihood in Tamilnadu. Tamilnadu has engaged an area of 1.3 lakh sq. km with an overall area of around 63 L.Ha for plantation. Agriculture, with its related sectors, is the major source of livelihoods in growing country like India. 70 percent of its rural households at a halt depend for the most part on agriculture for their livelihood, with 82 percent of farmers being small and marginal [2].

Tamil Nadu, a farmer-friendly state has set one of the greatest platforms for agricultural enlargement by introducing novel agricultural technologies to re-invent Green Revolution for the second time in the state. Further, the Government has formulated and implemented policies and schemes to achieve a consistent and rapid growth at an accelerated pace.

A stochastic model is one in which the epistemic uncertainties in the variables are taken into account. The uncertainties are those due to natural variation in the process being modeled. The variables in a stochastic model are described with probability distributions. They are commonly used in finance, project management and engineering. There is infinity of possible applications for stochastic modeling - any problem that can be analyzed deterministically can also be analyzed stochastically. Stochastic modelings of corrosion in an offshore facility that may help you understand how stochastic models are developed and analyzed [3].

Statistical model was used to estimate the groundwater dataset using precipitation as well as to predict the groundwater dataset with easily measurable climate date and also using regression model was established regression analysis were conducted to understand the relationship [4]. In any forming system taking to considering in to rainfall, water sources, region and production of crops. The authors to discover the level of concentration in paddy improvement using stochastic model [5]. Water table depth is an important map in environmental models' assessments. To develop an early water table prediction model for North Sinai, Egypt, different approaches consider namely GIS, remote sensing, simulation and stochastic methods. Stochastic (using time-series) modeling used to characterize the water table dynamics in terms of risk [6].

The author discuss with different labours and various types of agriculture workers and other important applicable details as input dataset. Observation of different years of government organization data is to be declared most of the agriculture related labours percentages are decreased year by year. Predicting the data and how to increase the agriculture labours involvement in future events using stochastic model approach [7]. Different stochastic models techniques and definitions explained related to agriculture and other social impact area has been clearly explained [8, 9].

A groundwater table monitoring under the study area and also the time series water table observations collected during the period from 1999 to 2006 used for experiments. The results that observed data of groundwater level. Box and Jenkins univariate stochastic model called as ARIMA model are applied to simulate the groundwater table [10]. Modeling and forecasting of the groundwater datasets for major component of effective management of water resources. One way to predict the groundwater level is analysis using a non-deterministic model [11].

II. METHODOLOGY

A stochastic model is one in which the epistemic uncertainties in the variables are taken into account. The uncertainties are those due to natural variation in the process being modeled. The variable is a quantity whose value changes in time series datasets. A discrete random variable is a variable whose values are obtained by counting. A continuous random variable is a variable which is used to whose values is obtained by measuring. A random variable is an important variable whose value is a numerical outcome of a number.

The proposed model using discrete random variable X has a countable number of possible values denoting the primary fields at the ith decision epoch, i = 1, 2, … n. Y is a another discrete random variable using secondary fields. 'W' denoting the continuous random variable.

In probability theory and statistics, the cumulative distribution function (CDF) of a real-valued random variable X, or just distribution function of X, evaluated at x, is the probability that X will take a value less than or equal to x. L(.) is a cumulative distribution function of continuous random variable of W. vk(T) denoted as probability that there are exactly k decision making epochs in (0, T]. Probability density function (PDF) is a statistical expression f(.) that defines a probability distribution for a continuous random variable as opposed to a discrete random variable. When the PDF is graphically portrayed, the area under the curve will indicate the interval in which the variable will fall. The total area in this interval of the graph equals the probability of a continuous random variable occurring.

The Laplace transform L(.) is invertible on a large class of functions. The inverse Laplace transform takes a function of a complex variable s (often frequency) and yields a function of a real variable time t. Given a simple mathematical or functional description of an input or output to a system, the Laplace transform provides an alternative functional description that often simplifies the process of analyzing the behavior of the system, or in synthesizing a new system based on a set of specifications [12].

Laplace transformation from the time domain to the frequency domain transforms differential equations into algebraic equations and convolution into multiplication. It has many applications in the sciences and technology [13].

In mathematics convolution fk (.) is a mathematical operation on two functions (f and g) to produce a third function that expresses how the shape of one is modified by the other. The term convolution refers to both the result function and to the process of computing it. Convolution is similar to cross-correlation [12].

For continuous functions, the cross-correlation operator is the adjoint of the convolution operator. Convolution has applications that include probability, statistics, computer vision, natural language processing, image and signal processing, engineering, and differential equations [14].

A convolution is an integral that expresses the amount of overlap of one function g as it is shifted over another function f. It therefore "blends" one function with another. For example, in synthesis imaging, the measured dirty map is a convolution of the "true" CLEAN map with the dirty beamistribution). The convolution is sometimes also known by its German name, faltung ("folding") [15].

A three-parameter Weibull distribution can be obtained from a two-parameter Weibull distribution by introducing the location or threshold parameter θ. Therefore, for θ>0, β>0 and ∝<α<-∝, a three-parameter GR distribution has the CDF is

$$F(x;\theta,\beta,\alpha)=\left[1-e^{-((x-\theta)/\beta)^{\alpha}}\right];x>\theta,\beta,\alpha>0 \quad (1)$$

Here satisfy θ>0 and then β>0 be the form and range parameters, in that order. The equivalent probability density function (PDF) is

$$F(x;\alpha,\lambda,\mu)=\alpha\beta^{-\alpha}(x-\theta)^{\alpha-1}e^{-((x-\theta)/\beta)^{\alpha}} \quad (2)$$

The corresponding survival function is

$$\overline{H}(X)=1-\left[1-e-((x-\theta)/\beta)^{\alpha}\right]$$
$$=\left[1-e^{-((x-\theta)/\beta)^{\alpha}}\right] \quad (3)$$

Assume that randomly in time in accordance with a three parameter Weibull distribution. Taking the shape parameter as θ = 1

$$P(x_i<y)=\int_0^{\infty}g_k(x)\overline{H}(X)\,dx \quad (4)$$
$$=\int_0^{\infty}g_k(x)\left[e^{-((x-\theta)/\beta)^{\alpha}}\right]dx$$
$$=\left[g^{*}\left(\theta(1/\beta)^{\alpha}\right)\right]$$

The survival function which gives the probability that the cumulative threshold will fail only after time t.

s(t) = P(T > t) Probability that the total damage survives beyond t

$$= \sum_{k=0}^{\infty} P\{\textit{there are extractly } k \textit{ contacts } (0,t)\} * P\{\textit{the total cumulative threshold } (0,t)\}$$

It is also known from renewal process that

P (exactly k policy decisions in (0, t)) = Fk(t) - Fk+1(t) with F0(t)=1

$$P(T > t) = \sum_{k=0}^{\infty} V_k(t) P(x_i < y)$$

$$= \sum_{k=0}^{\infty} [F_k(t) - F_{k+1}(t)] \left[g^*(\theta(1/\beta))^{\alpha} \right]^k \quad (5)$$

Now, the life time is given by P(T<t) = L(t) = the distribution of life time (T)

$$L(t) = 1 - S(t)$$

$$= 1 - \sum_{k=0}^{\infty} [F_k(t) - F_{k+1}(t)] \left[g^*(\theta(1/\beta))^{\alpha} \right]^k \quad (6)$$

Taking Laplace transformation L(t) we get

$$l^*(s) = \frac{1 - \left[g^*(\theta(1/\beta))^{\alpha} \right]^k f(s)}{1 - g^*(\theta(1/\beta))^{\alpha} f(s)} \quad (7)$$

Let the random variable denoting inter arrival which follows exponential with parameter.

Now $f^*(s) = \frac{\lambda}{\lambda + s}$ substituting in the above equation (7) we get

$$= \frac{1 - \left[g^*(\theta(1/\beta))^{\alpha} \right]^k \left(\frac{\lambda}{\lambda + s} \right)}{1 - g^*(\theta(1/\beta))^{\alpha} \left(\frac{\lambda}{\lambda + s} \right)}$$

$$l^*(s) = \frac{\lambda [1 - \left[g^*(\theta(1/\beta))^{\alpha} \right]^k]}{[\lambda + s - g^*(\theta(1/\beta))^{\alpha} \lambda]} \quad (8)$$

Taking the first order derivatives in the equation (8), we get

$$\left. \frac{d^* l(s)}{ds} \right|_{s=0}$$

Given s = 0

$$E(AG) = \frac{1}{\lambda [1 - g^*(\theta(1/\beta))^{\alpha})]}$$

$$= \frac{1}{\lambda [1 - g^*(\theta)^{\alpha} g^*(1/\beta))^{\alpha})]} \quad (9)$$

$$g^*(\theta) \sim \exp(\theta), g^*(\theta) = \frac{\beta}{\beta + \theta} \quad g^*(1/\beta) = \frac{\beta}{\beta + (1/\theta)}$$

On simplification we get

$$E(AG) = \frac{1}{\lambda \left[1 - \left(\frac{\beta}{\beta + \theta} + \frac{\beta}{\beta + (1/\theta)} \right)^{\alpha} \right]} \quad (10)$$

On simplification we get

$$E(AG) = \frac{1}{\lambda - \left(\frac{\beta}{\beta + \theta} \right)^{\alpha} \lambda + \left(\frac{\beta\theta}{1 + \beta\theta} \right)^{\alpha} \lambda} \quad (11)$$

III. NUMERICAL ILLUSTRATIONS

The following table taken from Department of Economic and Statistics, Department of Agriculture, ENVIS Centre, Tamilnadu State Council for Science, Ministry of Environment and Forests and Climate Change, India Meteorological Department (IMD), Government of India. The dataset display time series data from 2010 to 2016, which is include food grains (L MT), total cereals productivity (Kg./Hec.), total pulses productivity (Kg./Hec.), rainfall (MM), temperature (Celsius) and groundwater level (M). In table 1, include different measurements of data then these type datasets not possible to apply to the proposed stochastic equations (11).

Table 2, shows, the normalization using feature scaling equation 12. This movement is basic when dealing with the parameters using different units and sizes of data. Highlight scaling is a strategy used to institutionalize the scope of independent variables or feature of data. In information handling, it is otherwise called data normalization and is generally performed the information preprocessing step is accustomed to carry all values into the range [0, 1]. This is additionally called unity-based standardization. This can be summed up to limit the scope of values in the dataset between any self-assertive point 'a' and 'b' and also assign (0.1, 0.9) respectively.

Table 1: Actual time series data include food grains (L MT), total cereals productivity (Kg./Hec.), total pulses productivity (Kg./Hec.), rainfall (MM), groundwater level (M) and temperature (Celsius)

Year	Agri. Productions Food Grains (L MT)	Total Cereals (Kg./Hec.)	Total Pulses (Kg./Hec.)	Rainfall (MM)	Ground Water Level (M)	Temperature (Celsius)
2010	126.67	2611	312	937.80	13.20	34.2
2011	124.75	2922	381	1165.10	11.70	32.6
2012	120.78	2897	385	937.00	11.50	33.6
2013	125.04	3918	531	743.10	13.00	32.5

2014	124.30	2526	415	790.60	23.60	32.3
2015	123.22	3907	752	987.90	24.35	33.1
2016	128.03	4419	868	1138.80	21.80	33.4

$$X' = a + \frac{(X - X_{\min})(b-a)}{X_{\max} - X_{\min}} \quad (12)$$

The following pseudo code is used to proposed stochastic equations (3) and normalization equations (4). In this approach, the get required inputs based on primary time series datasets (table 1) and processes of those data using the following pseudo code, finally the code deliver various expected factors for affecting agriculture growth estimation in table 3 to table 5 and figure 2 to figure 4.

A. Pseudo for Normalization and Stochastic Model

BEGIN
Initialize the model parameters
SET $n \leftarrow 5$
SET $\beta \leftarrow$ rainfall [937.80, 1165.10, 937.00, 743.10, 790.60, 987.90, 1138.80]
SET $\theta \leftarrow$ groundwater [13.20, 11.70, 11.50, 13.00, 23.60, 24.35, 21.80]
SET $\alpha \leftarrow$ temperature [34.2, 32.6, 33.6, 32.5, 32.3, 33.1, 33.4]
SET $\lambda \leftarrow$ foodgrain [126.67, 124.75, 120.78, 125.04, 124.30, 123.22, 128.03]
INPUT: Proposed stochastic model parameters *(β, θ, α, λ)*
OUTPUT: Expected factors that affect the agriculture growth
Generate control sequence using equation *(12)*
for $i \leftarrow 1$ to n **do**
SET $a \leftarrow 0.1$ and **SET** $b \leftarrow 0.9$
for $j \leftarrow 1$ to n **do**
$beta[j] \leftarrow a + ((\beta[j] - Min(\beta))\ (b-a)) / (Max(\beta) - Min(\beta))$
RETURN *beta*
end for
for $j \leftarrow 1$ to n **do**
$delta[j] \leftarrow a + ((\delta[j] - Min(\delta))\ (b-a)) / (Max(\delta) - Min(\delta))$
RETURN *delta*
end for
for $j \leftarrow 1$ to n **do**
$theta[j] \leftarrow a + ((\theta[j] - Min(\theta))\ (b-a)) / (Max(\theta) - Min(\theta))$
RETURN *theta*
end for
for $j \leftarrow 1$ to n **do**
alpha [j]
$\leftarrow a + ((\alpha[j] - Min(\alpha))\ (b-a)) / (Max(\alpha) - Min(\alpha))$
RETURN *alpha*
end for
for $j \leftarrow 1$ to n **do**
lambda [j]
$\leftarrow a + ((\lambda[j] - Min(\lambda))\ (b-a)) / (Max(\lambda) - Min(\lambda))$
RETURN *lambda*
end for
Generate initial sequences using Equation (11)
SET *count* $\leftarrow 1$
SET $x \leftarrow 1$
while *count* < 6 **do**
PART1[x] $\leftarrow$ *(beta[x]/ (beta[x] +theta[x])) ^alpha[x])*lambda[x]*
PART2[x] $\leftarrow$ *((beta[x]* theta[x])/ (1+beta[x]*theta[x]))^alpha[x]) *lambda[x]*
EAG[x] $\leftarrow$ *(1/ (lambda[x] - PART1+PART2))*
RETURN *EAG*
count+1 and x+1
End while
END

Table 2 shows the normalized form of data using the proposed stochastic equation and normalization equation. The normalization and stochastic model equation also entirely solved using the above pseudo code techniques.

Table 2: Normalized data include Agri. Productions, Rainfall, Groundwater and Temperature

Year	Agri. Productions Food Grains (L MT)	Total Cereals Productivity (Kg./Hec.)	Total Pulses Productivity (Kg./Hec.)	Rainfall (MM)	Ground Water Level (M)	Temp. (Celsius)
	λ_1	λ_2	λ_3	β	θ	α
2010	0.7499	0.1359	0.1000	0.4691	0.2058	0.9000
2011	0.5381	0.2674	0.1993	0.9000	0.1125	0.2263
2012	0.1000	0.2568	0.2050	0.4676	0.1000	0.6474
2013	0.5701	0.6883	0.4151	0.1000	0.1934	0.1842
2014	0.4884	0.1000	0.2482	0.1900	0.8533	0.1000
2015	0.3692	0.6836	0.7331	0.5641	0.9000	0.4368
2016	0.9000	0.9000	0.9000	0.8501	0.7412	0.5632

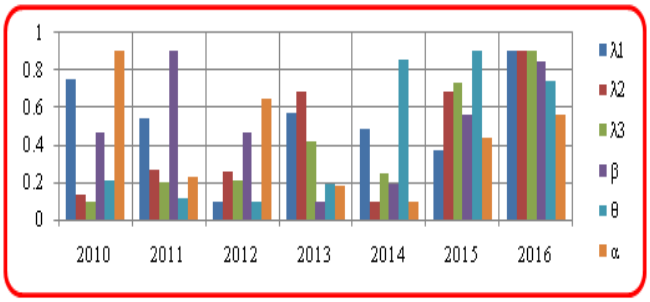

Fig. 1: Comparison of Normalized data include Agri. Productions, Rainfall, Groundwater and Temperature

Table 3: Expected Prediction using β, θ, α (decrease) and λ (fixed)

Rainfall (β)	Groundwater (θ)	Temperature (α)	Agri. Productions (λ)	Expected Prediction
0.5	0.5	0.5	0.2	0.9000
0.4	0.4	0.4	0.2	0.7563
0.3	0.3	0.3	0.2	0.5642
0.2	0.2	0.2	0.2	0.3425
0.1	0.1	0.1	0.2	0.1000

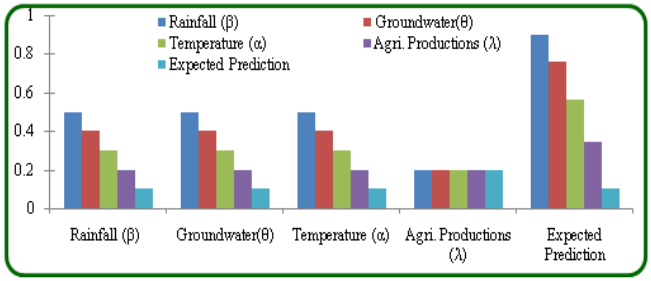

Fig. 2: Expected Prediction using β, θ, α (decrease) and λ (fixed)

Table 4: Expected Prediction using θ, α, λ (fixed normal) and β (decrease)

Rainfall (β)	Groundwater (θ)	Temperature (α)	Agri. Productions (λ)	Expected Prediction
0.5	0.3	0.3	0.3	0.9000
0.4	0.3	0.3	0.3	0.6425
0.3	0.3	0.3	0.3	0.4166
0.2	0.3	0.3	0.3	0.2270
0.1	0.3	0.3	0.3	0.1000

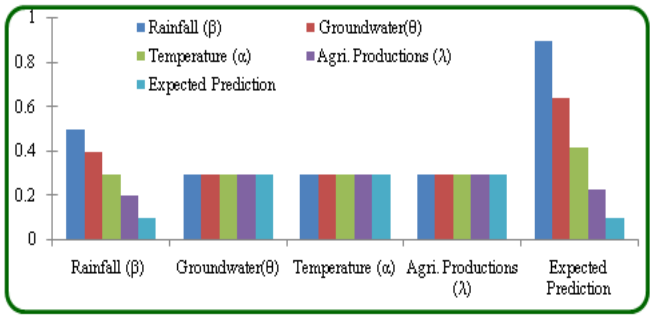

Fig. 3: Expected Prediction using θ, α, λ (fixed normal) and β (decrease)

Table 5: Expected Prediction using β, θ, λ (fixed normal) and α (increase)

Rainfall (β)	Ground Water (θ)	Temp. (α)	Agri. Productions (λ)	Expected Prediction
0.3	0.3	0.1	0.3	0.1000
0.3	0.3	0.2	0.3	0.2889
0.3	0.3	0.3	0.3	0.4867
0.3	0.3	0.4	0.3	0.6909
0.3	0.3	0.5	0.3	0.9000

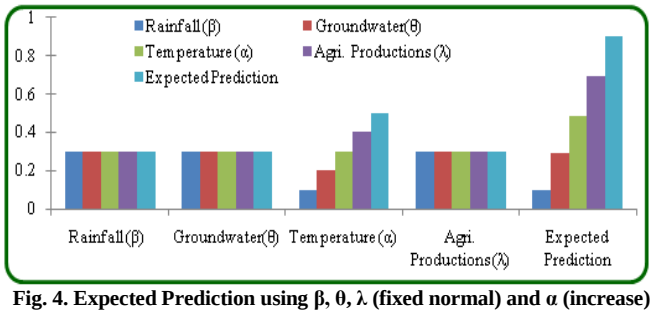

Fig. 4. Expected Prediction using β, θ, λ (fixed normal) and α (increase)

IV. RESULTS AND DISCUSSION

Table 1 shows the actual time series data from 2010 to 2016. The dataset includes different primary and secondary fields, which is include agriculture productions (in L MT), rainfall (in MM), groundwater (in M) and temperature (in Celsius). The dataset contains various scales of measurements of the data, it is not suitable for apply the proposed pseudo code using stochastic model equation.

In table 2 shows the normalized or uniform format of dataset using the pseudo code. The primary field agriculture productions values are assigned as 'λ' symbol, the rainfall values are assigned as 'β'. Similarly, other secondary parameters ground water levels are assigned as 'θ' and temperature level is named as 'α' and fertilizer named as 'θ'. In this assumption is very useful for applying numerical values easily to the proposed model.

Numerical illustration of table 2, in 2016 occurs to archive a maximum agriculture production are increases at the same time rainfall, groundwater and temperature also increased normally. In 2012 has low production growth in agriculture, which is reflected using the parameters of rainfall, groundwater and temperature also marginal, in this result depicted in figure 1.

Further more in table 3, which is used to predict the estimation using the proposed pseudo code. In this case, if the rainfall, groundwater and temperature namely β, θ, α values are decreased and the value of λ is fixed normally, the proposed system delivered the result of expected agriculture growth also decreased and the results show in fig. 2.

The result and discussion of table 4 and figure 3, the proposed system delivered a prediction of expected agriculture growth is decreased. In this regard the value of rainfall 'β' is decreased and other parameters like groundwater, temperature and agriculture productions namely θ, α, λ are assigned a normal fixed value equally. In this case, the expected prediction of agriculture growth also decreased.

Further more in table 5 and figure 4, if the β, θ and α namely rainfall, groundwater and agriculture productions values are kept fixed normally and other parameters namely λ decreased year by year; in this nature, the expected agriculture growth also increased using the proposed system.

V. CONCLUSION AND FUTURE SCOPE

The novelty of proposed system is used to predict for three factors that affected the agriculture growth strongly and also proved using the proposed system. In future the proposed model deeply analysis and how to use that increasing the affected factors like rainfall, groundwater and temperature. The three factors are safely considered to increase then the increasing the agriculture growth are one of the sustainable developments in nations like India. The pseudo code converts into the GUI tools to predict the factors that

affecting the agriculture growth in future. This model is not only in the field of agriculture area and additionally in a more extensive setting to utilize other social effect zones.

REFERENCES

[1] Rajesh, P. and Karthikeyan, M., "A comparative study of data mining algorithms for decision tree approaches using WEKA tool", Advances in Natural and Applied Sciences, vol. 11(9), 2017, pp. 230-243.

[2] https://en.wikipedia.org/wiki/Economy_of_India.

[3] https://www.quora.com/An-example-of-stochastic-model

[4] Yan, S., Yu, S., Wu, Y., Pan, D., Dong, J., "Understanding groundwater table using a statistical model", Water Science and Engineering, vol. 11(1), 2018, pp. 1-7.

[5] Rajesh, P. and M. Karthikeyan, "Prediction of Agriculture Growth and Level of Concentration in Paddy - A Stochastic Data Mining Approach", Advances in Intelligent Systems and Computing, 2018, pp. 127-139.

[6] EI-Sayed Omran. E., "A stochastic simulation model to early predict susceptible areas to water table level fluctuations in North Sinai, Egypt", The Egyptian Journal of Remote Sensing and Space Science, vol. 19(2), 2016, pp. 235-257.

[7] Rajesh, P. and Karthikeyan, M., "Predication of Labour Demand in Agriculture Based On Comparative Study of Different Data Using Data Mining and Stochastic Approach", International Journal of Engineering Science Invention, vol. 2, 2018, pp. 86-89.

[8] Bartholomew, D. J., "The Stochastic model for social processes",. 3nd ed., John Wiley and Sons, New York, 1982.

[9] Mucherino, A., Papajorgji, P.J., and Pardalos, P.M, "Data mining in agriculture", Springer Science & Business Media., 2009.

[10] Adhikary, SK., Mahidur Rahman, Md., and Gupta, AD., "A Stochastic Modelling Technique for Predicting Groundwater Table Fluctuations with Time Series Analysis", International Journal of Applied Sciences and Engineering Research, vol. 1(2), 2012, pp. 238-249.

[11] Mohammad Mirzavand, Seyed Javad Sadatinejad, Hoda Ghasemieh, Rasool Imani and Mehdi Soleymani Motlagh, "Prediction of Ground Water Level in Arid Environment Using a Non-Deterministic Model", Journal of Water Resource and Protection, vol. 6, 2014, pp. 669-676.

[12] Korn, G. A. and Korn, T. M., "Mathematical Handbook for Scientists and Engineers", 2nd ed., McGraw-Hill Companies, 2016.

[13] https://en.wikipedia.org/wiki/Laplace_transform.

[14] https://en.wikipedia.org/wiki/Convolution.

[15] http://mathworld.wolfram.com/Convolution.

