Knowledge Discovery and Leaf Spot Dynamics of Groundnut Crop through Wireless Sensor Network and Data Mining Techniques

A. K. Tripathy
Centre of Studies in Resources Engineering,
Indian Institute of Technology Bombay, Powai, Mumbai, Maharashtra, India. 400076
email: aktripathy@iitb.ac.in

J. Adinarayana
Centre of Studies in Resources Engineering,
Indian Institute of Technology Bombay, Powai, Mumbai, Maharashtra, India. 400076
email: adi@iitb.ac.in

K. Vijayalakshmi
Department of Entomology,
ANGR Agricultural University, Rajendranagar, Hyderabad, India. 500030
email: vijayalkorlipara@yahoo.com

S. N. Merchant
Department of Electrical Engineering,
Indian Institute of Technology Bombay, Powai, Mumbai, Maharashtra, India. 400076
email: merchant@ee.iitb.ac.in

U. B. Desai
Indian Institute of Technology Hyderabad, Yeddumailaram, Andhra Pradesh, India. 502205
email: ubdesai@iith.ac.in

S. Ninomiya
Institute for Sustainable Agro-ecosystem Services (ISAS),
The University of Tokyo, Nishitokyo, Tokyo, Japan. 188-0002
email: snino@isas.a.u-tokyo.ac.jp

M. Hirafuji
Hokkaido Agricultural Research Center
National Agriculture and Food Research Organization & University of Tsukuba
Memuro-cho, Hokkaido, Japan 082-0071
email: hirafuji@affrc.go.jp

T. Kiura
Agroinformatics Division, Agricultural Research Center
National Agriculture and Food Research Organization (NARO),
3-1-1, Kannondai, Tsukuba, Ibaraki, Japan 305-866
email: kiura@affrc.go.jp

Abstract

Data driven precision agriculture aspects, particularly the dynamic disease management, require dynamic crop-weather-environment data at micro level. An experiment was conducted during four consecutive seasons (2009 Kharif, 2009-10 Rabi, 2010 Kharif and 2010-11 Rabi) in a semi-arid tropic region of India to understand the crop-weather-environment-disease relations using wireless sensory and field-level surveillance data on the groundnut crop for Leaf Spot (LS) disease, which is economically important yet more prone in the semi-arid tropic. Tailor-made various data mining techniques (Naïve Bayes classification with Gaussian distribution, rapid association rule mining and multivariate regression mining) were developed and applied to turn the data into useful information/knowledge/relations/trends and correlation to understand crop-weather-environment-disease continuum. These dynamics obtained from the data mining techniques and

trained through proposed multivariate regression (MVR) mathematical models were validated with ground level surveillance data as well as ARI model (derived from ongoing long term weather-based experiment with diversely pooled data experimented from 10 seasons in semi-arid and arid zones). It was found that LS disease infection is strongly influenced by minimum temperature (18-20⁰C), prolonged duration of leaf wetness (7-10 hours), high humidity (75% or more) and age of the crop. These findings have been used for development of prediction models (One week and cumulative predictions), which can assist the user community to take respective ameliorative measures and it has been found that cumulative prediction model has performed better than ARI model with respect to ground level observations in all 16 diverse dates of sowing experiments spanning for two model-years.

Keywords: Data Mining, Knowledge Discovery, Wireless Sensor Network, Precision Farming and Pest/Disease Management at micro-scale level.

1. INTRODUCTION

Crop losses, particularly oilseeds, due to pests and diseases are quite substantial, particularly in the semi-arid conditions [33]. Among the oilseed crops, groundnut extent its large space despite its prone to pests and diseases. Significant crop losses by pests/diseases have been reported from many countries [3, 16, 26]. Disease is a product of the interaction between host and pathogens in favourable environment overtime. Even though susceptible host and aggressive pathogen are presence, the absence of congenial weather at canopy level, generally referred to as microclimate, restricts the disease pathogen development.

The major worldwide foliar diseases of groundnut, particularly in Semi-arid tropic region, caused by fungi are Rust (*Puccinia arachidia*), late Leaf Spot (LS) (*Cercosporidium personatum*), and early LS (*Cercospora arachidicola*), are the most important foliar diseases of groundnut worldwide, and can cause yield losses as high as 50% [32, 34]. Early LS develops small necrotic flecks that usually have light to dark-brown centers, and a yellow halo. The spots may range from 1-10 mm in diameter and sporulation appears on the abaxial (upper) surface of leaflets. Late LS develops small necrotic flecks that enlarge and become light to dark brown. The yellow halo is either absent or less conspicuous in late LS with sporulation on the abaxial (lower) surface of leaves. LS cause premature defoliation and are commonly a major constraint to both seed yield and good quality haulm production [6].

Weather and climate are very important in determining the precise epidemiology of outbreak of the disease. Critical threshold of the meteorological elements for the incidence, spread and intensification of disease determined in the laboratory condition have little relevance to the field condition. Therefore, they have to be determined and monitored under

83 field conditions through simultaneous observation of micrometeorological parameters and the pertinent data [23].
84 Temperature, humidity, rainfall, leaf wetness are the foremost weather parameters influencing the LS disease incidence [4,
85 6, 11, 18, 34]. It is rather difficult to establish a direct cause and effect relationship between any single climatic factor and
86 disease activity as the effect of these weather elements on the disease is usually confounded [32]. The combined effects of
87 temperature and wetness duration on a variety of diseases have been described by polynomial equations comprising
88 numerous parameters that lacked a clear epidemiologic significance [22]. Duthie [10] and Wu et al. [34] described a
89 nonlinear model for evaluating the combined effects of temperature and wetness duration. While parameters of this model
90 can be interpreted to provide information on the mechanisms involved in the disease response, reports on fitting empirical
91 data and interpreting the response have not been established.

92
93 With the ever-increasing amount of information about the farms, farmers are not only harvesting in terms of agriculture
94 output but also a large amount of data. In order to extract only the useful information, these data need to be analyzed and
95 used for optimization [13, 31]. Such data can be used in productive decision making if appropriate data mining (DM)
96 techniques are applied. DM is a process of extracting the most important and useful information from large set of data and to
97 uncover previously unknown patterns and hidden relationships within the data [2, 22, 31] that may be relevant to current
98 dynamic agricultural problems. In fact, applications of DM in agriculture emerged just recently, and therefore, there are not
99 yet any specific DM techniques specifically tailored to solve agricultural-related problems [22]. However, different data
100 mining techniques/algorithms can be developed and tailor-made to turn the data into useful information/
101 knowledge/relation/ trends and obtain correlation to understand crop-weather-environment-disease continuum.

102
103 In the practices of agricultural information, the agricultural intelligence monitoring system is an important and
104 indispensable link [35]. The advanced sensor technology and the intelligence information processing technology are
105 important methods to guarantee correctly and quantitatively to gather real time data on different parameters pertaining to
106 weather, crop, soil and environment, which in turn help in developing open solutions for majority of the agricultural
107 processes. With the advent of Wireless Sensor Network (WSN), information acquisition, dissemination and the processing
108 technology has seeped gradually into agricultural domain by the characteristics such as low-cost, low-power consumption,
109 small devices equipped with limited sensing, data processing and wireless communication capabilities, which perfectly
110 suites for precision agriculture where decisions are to be made at micro-climatic level at right time/place/input [19, 27].

111
112 Weather based pest/disease forewarning models have been developed to certain extent [4, 6, 11, 17, 18, 32, 34].
113 However, a functionally viable model for disease forecast considered to be one of the important components in the
114 integrated pest/disease management strategy. With the current scenario of disease resistance to fungicides/pesticides and

their high costs, it is still more important to develop an early warning model to provide caution to the farmers regarding the occurrence of disease, their peak activity and migration to develop effective and efficient management system. Hence, the present investigation has been carried out to discover the correlation (direct/hidden) and dynamics of disease with micro-weather parameters obtained from WSN in field conditions, which lead to developing a forewarning model with.

2. MATERIALS AND METHODS

In order to study the crop-weather-disease interactions, a test bed for WSN experiment was chosen at Agriculture Research Institute (ARI) of Acharya N G Ranga Agricultural University, Hyderabad, which falls in semi-arid tropic region. The test bed, where long term weather-based experiments are being carried out on groundnut crop, had provided a platform for validation (including comparisons) of the proposed model with existing pest/disease model. This work is a part of Indo-Japan initiative to develop a real time decision support system, called GeoSense [29, 3], integrating Geographical Information Communication Technology (Geo-ICT) and WSN for Precision Agriculture.

2.1 Standard Experimental Setup

Studies on crop-weather-disease interaction were carried out consecutively in two Kharif (monsoon) 2009 and 2010, two Rabi (post-monsoon) (2009-10 and 2010-11) agriculture seasons. A standard field experiment design was laid out in the test bed (Figure 1 and 2). Where P_1 is unprotected plot (a normal situation in farmers field) and P_2 is weather based protection plots. Four different dates of sowing (D_1 , D_2 , D_3 and D_4) were taken in to consideration. These different dates determine the impact of disease incidence in order to observe dynamics in pre (D_1), normal (D_2 and D_3) and post normal weeks of sowing (D_4).

Figure 1.

Figure 2.

Apart from this, to have uniform and unbiased observation, surveillance data has been collected from each plot in randomly selected one square meter area locations of the plot (S_1D_1 , S_1D_2 , S_1D_3 , S_1D_4 , S_2D_1 , S_2D_2 , S_2D_3 , S_2D_4 ,.....) during flowering to harvest crop phenological stages (Figure 2) in three replicated plots (R_1 , R_2 and R_3).

2.2 Surveillance Data Collection

LS population dynamics (surveillance data) were obtained at every week from flowering to reproductive phenological stages, where majority of disease incidences occur, at various locations in the experimental site. The surveillance data has been collected weekly instead of daily as there won't be any significant visible changes in disease incidences. A total of 48 observations (12 weeks X 4 different dates of sowing) were made with respect to different dates of sowing in each season. Along with this, Groundnut crop age (that is at which stage of the crop the disease attack takes place and their dynamics trends) also recorded week- wise to understand the infection dynamics of LS. The weeks in a year are mapped into integer values by considering the first week of January as first standard week.

Table 1.

Five sticks were placed in each one square meter and numbered 1, 2, 3, 4 and 5 (Figure 2). Subsequently, data has been collected manually from the plants adjacent to each stick for all four seasons. The surveillance data includes number of leaf-lets that LS have infected, how many plants infected in the one square meter area, date of flowering, date of recording the surveillance data, etc. The plants have been numbered in the form of ID like $R_1P_1D_1-1$, $R_1P_1D_1-2$, $R_1P_1D_2-1$, etc. so as to retrieve their respective data easily from the database (e.g. $R_1P_1D_1-1$ indicates Plant Number 1 of replica R_1 from P_1 plot with date of Sowing D_1 from one square meter area, $R_2P_1D_2-2$ indicates Plant Number 2 of replica R_2 from P_1 plot with date of sowing D_2). This is the one of the standard design practices that has been used in the long term experiment in the test bed. Per cent disease intensity (PDI) was recorded on at least 30 plants each plot from in the test bed on the basis of a '0 – 5' scale shown in Table 1.

2.3 Sensory Data collection

AgiSens [27], an end to end WSN system was deployed in the test bed to monitor micro level weather parameters. AgiSens, a set of battery-powered nodes embedded with wireless sensors (temperature sensor, humidity sensor, leaf wetness sensor, soil moisture sensor), was used to continuously monitor sensory soil-crop-weather parameters. The sensory data was routed through multi-hop communication to an embedded gateway (*stargate*) where data stored in a structured database. The embedded gateway is connected to internet through general packet radio service (GPRS) interface. It was facilitated to smooth transfer of data to the server (remotely located) for data storing. The data collected, thus, was stored in an Open Source database PostgreSQL for further analysis and mining.

Figure 3.

Figure 3 shows the schematics of wireless sensor network with agricultural / environment sensors deployed in the field. Each node was able to transmit/receive data packets to/from other nodes every 15 minute over a transmission range of 25 meter. Data collected by the sensors were wirelessly transferred in a multi-hop manner to a base station node (*stargate*) connected with embedded gateway for data logging. In a WSN, when the transmission range of a sensor node is not sufficient, it uses multi-hop communication to reach the destination node or sink node. This data forwarding mechanism continues till it reaches the sink node.

The base station has a GPRS connectivity through which it routes data to the remote GeoSense server setup at Agro-Informatics Lab at CSRE, IIT Bombay and collect all the sensory information. The sensory data coming to the server through GPRS is raw and has been converted in to a usable format in real time through appropriate conversion formula (by

using open-source server-side scripting language PHP) in the server end. Both raw and the real data have been stored in different database for further analysis and mining. Other related weather data (sunshine hours SH, wind speed WS, rainfall RF and evapotranspiration ET) were obtained from the weather station with in the vicinity of the test bed (Figure 4). Sensory data was collected during the Kharif seasons of 2009 and 2010 as well as Rabi season 2009-10 and 2010-11 with multiple sensor nodes, i.e. M_2 , M_3 , M_4 and M_5 for four different dates of sowing (D_1 , D_2 , D_3 and D_4) in four different test plots, in groundnut field (Figure 1 and Figure 4). M_1 sensor node was deployed in P_2 plot of D_1 date of sowing to collect sensory data in weather based protected environment. These sensory data have been used for comparative pilot study of protected and unprotected plots.

Figure 4.

2.4 Proposed MVR Mining/Prediction Model

Various data mining (DM) techniques and algorithms were tailor made to comprehend the pest dynamics and the general processing flow followed is depicted in Figure 5. Raw sensory data obtained from the experimental field is not uniform in its collection. Owing to the climatic conditions or non-function of field sensors or due to network errors, there have been a few breaks in collecting continuous sensory data that may lead to biased outcomes while developing the model. Expectation–Maximization (EM) algorithm was used to deal with missing data [15]. Relative data from the nearest sensor node was used to fill the missing data with EM algorithm. The data set is provided in daily and weekly means wherever is required. Quality data is accomplished by performing satisfactory data pre-processing methodology such as data selection, data reduction and elimination of null values or other noise values. Though the real-time sensory data were collected at 15 minute interval, all such data were not used for the current experiment. For example, the temperature was taken by computing maximum and minimum values of the respective day. Relative humidity was recorded as RH1 (7: 30 AM) and RH2 (3: 00 PM), which is a standard practice in Indian agriculture system (AgroMet-Cell, 2009). Leaf Wetness (LW) data has been used in the scale of 1 to 15 (0 being the completely dry to 15 being saturated) as Leaf Wetness Index, and the value above 5 has been taken as wet leaf and computed for wetness period [27], which is generally taken as the value that attract pest/disease.

Figure 5.

2.4.1 Tailor made Naïve Bayes with Gaussian distribution

This Data Mining (DM) technique was tailor made to understand LS dynamics from sensory and weather station based meteorological and other surveillance parameters in Groundnut crop. As Groundnut crop was infested with multiple pests

and diseases, multi-level classification modules were developed in the model, which classifies crop pests and diseases based on the severity. Naive Bayes classification with Gaussian distribution [12, 21] was developed and applied in the experiment. Bayesian network principle was used to model uncertainty by combining experimental knowledge and observational evidences. Gaussian Naive Bayes (NB) classifier, which is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence, was used to assume the presence (or absence) of a particular feature of a class unrelated to the presence (or absence) of any other feature [21]. For example, a disease symptom may be considered to be LS owing to the shape and place of the spot on the leaflet, color, texture, damaged leaflets, etc. Even though these features depend on the existence of the other features, NB considers all of these properties independently which contribute the probability of occurrence of LS (early/late LS).

Hypothetically, the probability model for a classifier is a conditional model $p(C|V_1, \dots, V_n)$ upon a dependent class variable C with a small number of results, conditional on some feature variables V_1 through V_n . However, if the number of features n is large, then constructing such a model on probability base is infeasible. From Bayes' theorem,

$$p(C|V_1, \dots, V_n) = \frac{p(C)p(V_1, \dots, V_n|C)}{p(V_1, \dots, V_n)}.$$

The numerator is equal to the joint probability model $p(C, V_1, \dots, V_n)$ which can be redrafted using recurrent applications of the description of conditional probability:

$$\begin{aligned} &= p(C)p(V_1, \dots, V_n|C) = p(C)p(V_1|C)p(V_2, \dots, V_n|C, V_1) \\ &= p(C)p(V_1|C)p(V_2|C, V_1)p(V_3|C, V_1, V_2) \dots p(V_n|C, V_1, V_2, V_3, \dots, V_{n-1}). \end{aligned}$$

The "Naïve" conditional independence assumptions come into force: assume that each feature V_i is conditionally independent of every other feature V_j for $j \neq i$.

That is $p(V_i|C, V_j) = p(V_i|C)$ for $i \neq j$, and the joint model can be articulated as

$$= p(C) \prod_{i=1}^n p(V_i|C).$$

All model parameters (*i.e.*, class priors and feature probability distributions) can be estimated with relative frequencies from the training set. A class' prior can be calculated by assuming equiprobable classes (*i.e.*, priors = 1/number of classes), or by calculating an estimate for the class probability from the training set (*i.e.*, prior for a given class = number of samples in the class / total number of samples) [12]. If one is dealing with continuous data, a typical assumption is the Gaussian

distribution with model parameters of mean and variance. The mean, μ , is calculated by $\mu = \frac{1}{N} \sum_{i=1}^N X_i$ where, N is the

number of samples and x_i is the value of a given sample. The variance, σ^2 , is calculated by $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$. if a given class and feature value never occurs together in the training set then the frequency-based probability estimate will be zero, which will wipe out all information in the other probabilities when they are multiplied. It is, therefore, often desirable to incorporate a small-sample correction in all probability estimates such that no probability is ever set to be exactly zero.

2.4.2 Tailor-made Association Rule Mining

Rapid association rule mining [1, 8] was developed/tailor-made in association with Naïve Bayes-Gaussian Distribution (NB-GD) to find out the correlation of multiple weather parameters with respect to LS. This phenomenon is to identify signature patterns as well to discover LS presence with other weather parameters. This algorithm helps in discovering effects of disease over the other weather parameters. The outcomes are in the form of correlation index values ranging from -1 to 1.

Association Rule can be defined as when,

Let $I = \{i_1, i_2, i_3, \dots, i_n\}$ be a set of objects, D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with a unique identifier TID (Transaction ID). A transaction T is said to contain X , a set of items in I , if $X \subseteq T$. An association rule is an implication of the form $X \rightarrow Y$, where $X \subseteq I$; $Y \subseteq I$, and $X \cap Y = \emptyset$. The rule $X \rightarrow Y$ has support s in the transaction set D if $s\%$ of the transactions in D contain $X \cup Y$. In other words, the support of the rule is the probability that X and Y hold together among all the possible presented cases. It is said that the rule $X \rightarrow Y$ holds in the transaction set D with confidence c . If $c\%$ of transactions in D contain X as well as Y . In other words, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the predecessor X . The problem of discovering all association rules from a set of transactions D consists of generating the rules that have a support and confidence greater than given thresholds. These rules can be named strong correlation.

A plant having disease could be always a combination of symptoms with stress or due to change in influential weather parameters. The task is to find the strength between the symptoms or diseases how frequently they are associated with weather or other host/disease.

Let $I = \{\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_m\}$ be a feature set. The association of features can be defined as follows:

$$Y(I) = \sum_{L' \subseteq I, |L'| \geq 2} \frac{s(I' \rightarrow I'')}{\text{NumberOfObservations}} \frac{|I'|}{|I|} \quad (1)$$

257 where, $I' \subseteq I$ and $I'' = I - I'$

$$258 \quad Y = \begin{cases} 0 & \text{no-correlation} \\ < \alpha & \text{Moderate-correlation} \\ \geq \alpha & \text{Strong-correlation} \end{cases}$$

259

260 Input (TD-transaction Database, Ms-Minimum Support, Mc-Minimum Confidence, Mcn-Minimum
261 correlation/association) used to Developed or tailor-made algorithm from [2, 8] is as fallow:

262

263 1. C_1 = Candidate l-itemsets

264 2. L_1 =Frequent l-itemsets

265 3. for ($K=2$; $L_{K-1} \neq \varnothing$; $K++$)

266 4. {

267 5. $C_K = L_{K-1} \bowtie L_{K-1}$

268 6. for each $c \in C_K$

269 7. If any subset of $c \notin L_{K-1}$

270 8. then $C_K = C_K - \{c\}$

271 9. for each $c \in C_K$

272 10. { If $\text{support}(c) \geq Ms$ then $L_k = L_K \cup \{c\}$ }

273 11. for each $c \in L_k$

274 12. {

275 13. If $Y(c) \geq Mcn$

276 14. {

277 15. for each $c=(x \cup y)$

278 16. If $\text{confidence}(x \rightarrow y) \geq Mc$

279 17. then $LS=LS \cup \{x \rightarrow y\}$

280 18. }

281 19. }

282 20. }

283 While developing regression function for modeling experimental data with more than one independent argument,
284 multivariate regression has been used. A multivariate regression is the use of more than one input variable for the pertinent
285 of complex models. When multiple predictors are used, the regression line cannot be visualized in two-dimensional space.

286 However, the line can be computed by expanding the equation for single-predictor linear regression to include the
 287 parameters for each of the predictors [25, 30].

288 There is a continuous random variable called the dependent variable, Y (e.g. LS), and a number of independent
 289 variables, $x_1, x_2, x_3, \dots, x_p$ (e.g. Leaf Wetness, Temperature, Rainfall, etc). The purpose is to predict the value of the
 290 dependent variable (also referred to as the response variable) using a linear function of the independent variables. The
 291 values of the independent variables (also referred to as predictor variables, regressors or covariates) are known quantities
 292 for purposes of prediction, the model is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon \quad (2)$$

293 Where, ε is “noise” variable and normally a distributed random variable with mean equal to zero and standard deviation
 294 σ whose value to be known. Also, the values of the coefficients $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ will be found out subsequently. All these
 295 $(p + 2)$ unknown values from the available data are subsequently estimated. The data consist of n rows of observations also
 296 called cases, which give us values $y_i, x_{i1}, x_{i2}, \dots, x_{ip}; i = 1, 2, 3, \dots, n$. The estimates for the β coefficients are computed
 297 so as to minimize the sum of squares of differences between the fitted (predicted) values at the observed values in the data.
 298 The sum of squared differences is given by

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_p x_{ip})^2 \quad (3)$$

299 Let's denote the values of the coefficients that minimize this expression by $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$. These are the
 300 estimates for the unknown values and are called OLS (ordinary least squares) estimates in the literature [18]. Once it

301 computes the estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$, one can calculate an unbiased estimate $\hat{\sigma}^2$ for σ^2 using the formula:

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip})^2 \quad (4)$$

302 Later, plug in the values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ in the linear regression model equation (1) to predict the value of the
 303 dependent value from known values of the in-dependent values, $x_1, x_2, x_3, \dots, x_p$. The predicted value, $\hat{Y}$, is computed by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p \quad (5)$$

304 Predictions based on equation No. 4 are the most probable predictions possible as they will be unbiased (equal to the true
 305 values on the average) and will have the smallest expected squared error compared to any unbiased estimates if one makes
 306 the following assumptions:

307 a) Linearity: The expected value of the dependent variable is a linear function of the independent variables, i.e.,

$$E(Y | x_1, x_2, \dots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (6)$$

b) Independence: The “noise” random variables ε_i are independent between all the rows. Here ε_i is the “noise” random variable in observation i for $i = 1, \dots, n$.

c) Unbiasness: The noise random variable ε_i has zero mean, i.e., $E(\varepsilon_i) = 0$ where $i = 1, 2, \dots, n$.

d) Homoskedasticity: The standard deviation of ε_i equals the same (un-known) value, σ , where $i = 1, 2, \dots, n$.

e) Normality: The “noise” random variables, ε_i , are *Normally Distributed*.

A notable fact is that, even if one can drop the assumption of normality and allows the noise variables to follow arbitrary distributions, these estimates are very good for prediction. One can show that predictions based on these estimates are the best linear predictions in that they minimize the expected squared error. The Normal distribution assumption was required to derive confidence intervals for predictions. In data mining applications, the training data set and the validation data set are two distinct sets of data that are both representative of the relationship between the dependent and independent variables. The training data is used to estimate the regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$. The validation data set constitutes a “hold-out” sample and is not used in computing the coefficient estimates. This enables to estimate the error in the predictions without having to assume that the noise variables follow the Normal distribution. The training data has been used to fit the model and to estimate the coefficients. These coefficient estimates are used to make predictions for each case in the validation data. The prediction for each case is then compared to value of the dependent variable that was actually observed in the validation data. The average of the square of this error enables us to compare different models and to assess the accuracy of the model in making predictions.

The regression algorithm estimates the value of the target as a function of the predictors for each case of the test data. The relationships between predictors (e.g. $T_{\max}$, ET, RH₁, etc.) and target (LS) is summarized in a multivariate regression equation, which could then be applied to a different data set in which the target values are unknown. The above technique was used in generating the prediction model. In addition, multiple correlation coefficient (R^2) is also used as an indicator of the reliability of a relationship identified by the regression analysis.

Following is a multivariate regression models developed/tailor-made for LS through XLminer software tool:

Leaf Spot (Linear)-

$$Y_{LS} = -552.16 - 18.65 * T_{\max} + 8.03 * T_{\min} + 2.74 * RH_1 - 0.185 * RH_2 + 175.56 * RF - 0.58 * RD + 3.49 * LW + 82.06 * SH + 26.85WS + 0.54 * ET - 2.56 * AC \quad (7)$$

335 Leaf Spot (Non Linear)-

$$Y_{LS} = -251.82 - 1.034 * T_{\max} + 4.36 * RH_1 - 1.92 * RH_2 + 3.05 * RF - 0.24 * ET - 7.12 * T_{\min} + 0.17 * SH + 0.65 * AC + 0.11 * (T_{\min})^2 + 2.70 * (RH_1)^2 - 6.03 * (RH_2)^2 + 0.34 * (RF)^2 + 0.58 * (LW)^2 + 1.087 * (RD)^2 - 4.22 * (AC)^2 \quad (8)$$

336 Where

Tmax =Maximum Temperature, AC= Age of the Crop, ET = Evapotranspiration,
 RH = Relative Humidity, WS=Wind Speed, RD=Rain Days,
 Tmin = Minimum Temperature, LW = Leaf Wetness, SH = Sunshine Hours
 RF = Rain Fall, RH₁ = Relative Humidity at 7:30 am RH₂ = Relative Humidity at 3pm

337

338 Multivariate prediction model, thus developed, has been used by taking historical data and was used for one week
 339 prediction. For example, one week prediction of Kharif 2010 season LS incidence level has been predicted for one week
 340 prediction by using Kharif 2009 data and similarly for Rabi 2010-11 seasons one week prediction was carried out with
 341 2009-10 Rabi historical parameters.

342 In addition, Complex Polynomial Cumulative model [1] was also developed/tailor-made by including various aspects
 343 viz. maximum pest population, time of first appearance, time of maximum pest population/disease severity as well as life
 344 cycle, season, weather parameter, stage of crop, incidence at flowering stage, growing degree days, previous year record,
 345 Correlation with other pest/disease, previous season crop for pest/disease forecasting:

$$Y = a_0 + \sum_{i=1}^p \sum_{j=0}^1 a_{ij} Z_{ij} + \sum_{i=1}^p \sum_{j=0}^1 b_{ii'j} Z_{ii'j} + e \quad (9)$$

$$Z_{ij} = \sum_{w=n1}^{n2} r_{iw}^j X_{iw} \text{ and } Z_{ii'j} = \sum_{w=n1}^{n2} r_{ii'w}^j X_{iw} X_{i'w}$$

346 Where, Z_{ii} 's and Z_{ij} 's the independent variables which are functions of the basic weather variables like maximum
 347 temperature, relative humidity, leaf wetness, etc. Y variable to forecast LS indices, X_{iw} = Value of i^{th} weather variable
 348 in w^{th} week, r_{iw} = Correlation coefficient between Y and i^{th} weather variable in w^{th} week, $r_{ii'w}$ = Correlation coefficient
 349 between Y and product of X_i and $X_{i'}$ in w^{th} week, $n1$ = initial incidence, $n2$ = first Peak population week. The
 350 developed multivariate regression model has been tested with respect to ground level studies in both Kharif 2009 and Rabi
 351 season 2009-10. The same is validated in the subsequent Kharif 2010 and Rabi 2010-11 seasons along with ground level
 352 studies for forecasting of LS incidences and compared with long term ARI experiment model. However, it is required to
 353 quantify the relations.

354 **3. RESULTS AND DISCUSSIONS**

355 Raw sensory data have been converted to real data at server end as when data comes by an automated tool developed at
356 Agro-Informatics, CSRE, IIT Bombay. Different life cycles of pathogens and their interactions with host plant produce a
357 wide range of responses to environmental and climatic drivers. Correlation studies of disease–weather–environment–crop
358 interaction were carried out and their hidden correlations were discovered and quantified with tailor-made NB-GD and
359 Rapid Association Rule mining techniques in the test bed. Subsequently, the developed multivariate regression mining
360 techniques were applied for disease modeling and predictions. The correlation values (both positive and negative) of
361 predictor (e.g. Tmax, RH₁, ET, etc.) versus target (LS) infection index were obtained from various datasets (sensory,
362 weather-station and surveillance) during flowering to harvesting stages and are depicted in Figure 6 for all the four
363 seasons. A correlation Index matrix was obtained by using Rapid Association Rule mining algorithm, in which correlation
364 index of disease prediction parameters (LS/weather/environment/crop age) were quantified in the range of -1 to +1 (Table
365 2), with negligible, moderate and strong level drawn from all four seasons.

Table 2.

Figure 6.

366

367 Based on concept of infection index with respect to pest/disease risk model [9, 25], correlation index greater than 0.5 in
368 the scale of -1 to +1 has been considered as strong positive, whereas the value of -0.5 or more considered as strong
369 negative. It was found that RH₁, LW and AC have strong positive correlation and Tmin has strong negative correlation
370 with LS. There is an increase in disease intensity with the period of high humidity as the latter favours abundant
371 sporulation and germination of the spores of the pathogen, and in agreement with several researchers [18, 34, 6, 32]. Since
372 the optimum temperature for LS ranges from 15⁰C to 25⁰C [7], naturally occurring temperature during the rainy season are
373 not likely to limit the process. The infection only takes place when the leaf surface is wet [6], this also supports the
374 assumption that temperature would rarely limit infection. Moderate +ve correlations were found with rainfall, rainy days
375 and wind Speed, while SH shown is moderate –ve correlation. Tmax, RH₂ and ET found to have negligible correlation
376 with LS, and is agreement with [32].

377

378 **3.1 Leaf Spot dynamics**

379 Leaf Spot (LS) incidence in groundnut has been observed round the year with prevailing weather conditions. By means of
380 the developed MVR model, with standard ground level surveillance observation and weather parameters from sensory as
381 well as nearby weather station, disease incidence analysis for all four seasons were quantified. The MVR model findings

(solid line with triangle) have been evaluated compared with ARI model (broken line) [32] resulted from long term experiments/analysis as well as with ground level surveillance data (GLS) (solid line with oval) data (Figures 7).

Leaf Spot dynamics in Kharif 2009, D1 sowing plot (Figure 7) shows the trend is similar with all the three observation. However, it has been observed that the developed MVR model has attained almost close to the ground level studies (GLS). The initial peak of 13.44% from ARI model and 10.47% MVR model as compared to 3.22% at GLS has been found on 18 September 2009. A maximum peak 71.63% (ARI) and 77.84 00 % (MVR) to 84.49 % (GLS) incidence was observed during first and second week of October (i.e., during 08 -14 October 2009). Significant increase of LS incidence was observed after initial two weeks of mild infection. This is due to the longer duration leaf wetness due to 40 to 50 mm rainfall with more rainy days propelled by high humidity during September second and third weeks, which favours abundant sporulation and germination of the spores of pathogen. [18, 34] also recorded similar observations in other climatic conditions. Similarly, other dates of sowing (D2, D3 and D4) and other seasons (Rabi 2009-10, Kharif 2010 and Rabi 2010-11) have been witnessed similar trends and patterns. Figure 8 shows the supporting photograph/evidence for the presence of LS in the test bed and depicts that almost all the leaf lets are infected.

Figure 7.

Figure 8.

The most prevailing weather parameters are quantified, including the crop age. Apart from this, typical periods of weather that can have a significant effect on the disease development in groundnut crop are also developed. Notably, from this experiment it has been observed that leaf wetness of 7-10 hours per day could be adequate for high levels of infection from sources of moisture other than rainfall, fog, dew, sprinkler irrigation, etc. Consequently, with direct leaf wetness measures (using leaf wetness sensor) has improved the prediction of favorable periods for infection. However, the increased precision probably may come at high price in terms of the purchase and maintenance of the sensor equipment. Temperature during the leaf wetness period also has a significant effect on the ability of LS pathogens to infect groundnut, and it was found that maximum infection occurred during 18⁰ C to 22⁰ C. It was, thus discovered from the four seasons experiment that prolonged hours (7-10) of leaf wetness with low temperature in high humidity of more than 75% (monsoon season) conditions favored the disease development. This discovery with MVR model enables one to predict on sa short term basis for precision protection.

3.2 Predictions

The weather based modeling for early warning of disease infestation may provide appropriate tool for investigating disease status and making strategic contingent measures. With correlation studies revealing the crop-weather-environment-

disease relationships/ interactions, there is a possibility of developing an early warning model (both cumulative and non-cumulative) on disease infestations. One week prediction model was carried out with previous year (all four different dates of sowing of 2009 Kharif and 2009-10 Rabi season) data of the same week, same date of sowing and same season. These computations have been carried out for near to equal to the peak period only as rest of the prediction has a less disease management significance if the crop is already infected during these peak stages. Two color schemes were followed as warnings (for LS management at the designated week) marked as light grey strip and dark gray indicating prediction of disease severity and to take respective control measures so as to reduce from disease.

Table 3.

Table 4.

Perusal of Tables (3, 4 and 5) reveal that both in Kharif 2010 and Rabi 2010-11, that one week (1wk) prediction is very closer to the proposed regression/empirical model (MVR), whereas the cumulative (CWK) method is probably a preferable prediction strategy as it is closer to the ground level data. However, in case of D2 sowing all the prediction models found close to each other (with a variation of 2 to 5% during peak incidence). These interactions indicate that if LS disease is controlled on 17/8/2010 (roughly during very early pod stage R3), there is a possibility to counter this disease multiplicity and the yield losses as far as D2 is concerned (Table 3). Similarly, in case of Rabi 2010-11 season, the first control measure is advised during R3 (pod) stage (12/12/2011 marked in dark gray colour in Table 4) failing which the LS incidence increases rapidly and the control after this period (after 2/1/2011 i.e. after two weeks marked dark gray color in Table 9) will have no significance as the crop has already severely infected and will not save the crop.

In general for Kharif 2010 season, the CWK value was observed to be in the range of 2 to 4% increased value with respect to ground level data. However, a 4 to 6% increase in prediction value was observed in case of 1wk prediction approach as compared to the ground level data and 5 to 10% increase in prediction observed in ARI prediction model. However, in case of D2, ARI prediction found to be deviated to the trend of 18% as compared to GLS. Thus, the CWK model found to be better accurate prediction method. This could be the fact that micro-level weather data has been used in MVR model as compared to the general weather station data has been used by ARI prediction model. Similar trends were also observed during 2010-11 Rabi season and care should be taken after the appearance of first peak incidence of LS. It has been observed that there were minimum deviations between actual and predicted (cumulative prediction) values of LS population, indicating the feasibility of predicting population occurrence using the prevailing weather parameters. Like this the prediction model has been tested and validated (with ARI model and ground level surveillance data) for both Kharif 2010 and Rabi 2010-11 seasons with four different date of sowings

Table 5.

441

Fungicide application usually is required for LS control and fungicides may be applied on a set 14-day calendar schedule, according to a weather-based LS advisory. In well-crop rotated fields, the first fungicide spray should be applied to susceptible cultivars at the very early pod stage (R3). After the first spray, apply fungicides every 14 days and continue applications until full seed stage R6 [5]. In all the four dates of sowing both in Kharif 2010 and Rabi 2010-11, the first control measure and second control measure have shown two weeks; however, in D4 (Table 5) rapid growth of disease was not found due dry weather with less hours of leaf wetness, hence the incident found to be less and the second control measure can be delayed by four weeks (i.e. 13/3/2011).

Both one week as well as cumulative predictions have been carried out for all the dates of sowing with the help of derived MVR model and compared with existing long-term experiment-based ARI model. It was found that the proposed MVR forecast model is very close to the ground level surveillance data as compared to ARI model in all most all 16 experiments. The MVR model developed with diverse climatic condition with two (wet and dry) consecutive model-years along with different dates of sowing experimented with two groundnut varieties (TMV-2 and NARAYANI) gave confidence levels in applying it with diverse fragile agro climatic regions. Though this model has been designed for precision protection with on-farm experiments, considering diverse experiments after validation with long term ARI prediction model, the MVR prediction model can be used in the prevailing agro-climatic region with inclement climatic conditions. The developed MVR model is able to forecast the probable disease incidence level at least one or two weeks in advance including its disease severity. This will help the stake-holders to take judicious ameliorate decisions. The model is able to predict the first disease peak level where the immediate care could be taken and at the same time it also predicts the disease growth rate so as to take the counter measures. With tested and validated forecasting ability with respect to different dates of sowing, the prediction models can be used for giving out a warning to growers in a timely manner ahead of the established response measures traditionally taken in agriculture. At present, disease prevention and precaution measures advices are given from the experts from the local academic/research bodies with the help of results obtained from the long term experiments and nearby and sparse weather stations.

Although the prediction models have been developed in association with weather-environmental parameters for LS disease, with reasonably accuracy for one - two weeks ahead, predictions of short-term weather pattern from the LS disease growth rate/incidence rate is questionable/debatable.

4. CONCLUSIONS

An attempt has been made to understand the hidden relationships between most prevailed LS disease that causes economic losses, and weather/environmental parameters in Groundnut crop. Real-time micro-climatic parameters such as

temperature, humidity and leaf wetness were obtained through Zigbee based wireless sensor network, where as the crop-weather-environment disease dynamics and hidden relations were quantified through tailor-made integrated DM techniques. The most prevailing weather parameters (Temperature, Leaf Wetness, Humidity, etc.) were quantified including the crop age. Apart from this the typical periods of weather that can have a significant effect on the disease development in a crop are also discovered. It has been observed that warm temperatures (18-22⁰C) and long periods of high humidity (75% or more) with prolonged hours (7- 10 hours) of leaf wetness are favorable condition for LS disease. Wet periods of sufficient duration to support infection usually consist of dew periods at night or extended rainy periods. When adequate moisture is present, LS infections may occur in a relatively short period when temperatures are warm, but a longer wet period is required when temperatures are cool. Hence, potential for damage from LS is greater where levels of humidity and rainfall are high. Frequent irrigation with small amounts of water can also create prolonged periods of high humidity and leaf wetness favorable for infection.

A MVR model has been developed and tested with four agriculture seasons (two dry seasons and two wet seasons) with micro-climatic sensory and related weather/environmental data from the nearby weather station and has been validated with ongoing long term weather-based experiment model (ARI model with 10 seasons diversely pooled data). Apart from this, the proposed model has been validated and compared with four different dates of sowing for each season. It is observed that the proposed MVR forecast model is very close to the ground level surveillance data as compared to ARI model in all 16 diverse dates of sowing experiments spanning for two model-years. Though this model has been designed for precision protection with on-farm experiments, considering diverse experiments, the model can be used in the agro climatic region with inclement climatic conditions.

In addition, one week and cumulative prediction models were developed. Out of two developed prediction models as well as ARI model, cumulative prediction model appears to be more accurate with respect to ground level surveillance data. Thus, it appears that cumulative prediction model, which issues the forecast for population buildup, initiation and severity of LS diseases will help the stake holders to take judicious ameliorative decisions so as to save the crop from LS infection and its subsequent effects on environment and crop yield.

ACKNOWLEDGMENT

This research work is a part of the Indo-Japan multi-disciplinary ICT initiative ‘Geo-ICT and Sensor Network-based Decision Support Systems for Agriculture and Environment Assessment’, sponsored by the DST and JST (project No: INT/JP/JST/P-07/2007). The authors wish to thank Dr. D. Raji Reddy and Dr. G. Sreenivas of Agromet-Cell, ANGR

505 Agricultural University, Rajendranagar, Hyderabad, India for their valuable collaboration in the field research. A. K.
506 Tripathy would like to thank Don Bosco Institute of Technology, Mumbai, India for sponsoring to carry out Ph.D. research
507 at IIT Bombay.

508 REFERENCES

- 509 1. Agrawal, R. and Mehta S.C. (2007). Weather Based Forecasting of Crop Yields, Pests and Diseases - IASRI Models,
510 Journal of Indian Society of Agricultural Statistics, 61(2): pp. 255-263.
- 511 2. Agrawal, R., Imielinski, T. and Swami, A.N. (1993). Mining Association Rules between Sets of Items in Large
512 Databases. SIGMOD. June 1993, 22(2): pp. 207-16.
- 513 3. AgroMet-Cell (2009). Entomology Work 2008-09, AgroMet-Cell, Agriculture Research Institute, ANGRAU,
514 Hyderabad, Annual Report, pp. 57-69.
- 515 4. Bailey, J. E, Johnson, G.L. and Toth, S. J. Jr. (1993). Evolution of a weather-based peanut leaf spot sprays advisory
516 in North Carolina. Plant Dis 78: 530-535.
- 517 5. Boote, K. J. (1982). Growth Stages of Peanut (*Arachis hypogaea* L.), Florida Agricultural Experiment Stations
518 Journal Series No. 2846. Peanut Science: January 1982, Vol. 9, No. 1, pp. 35-40.
- 519 6. Butler, D. R, Wadia, K. D. R., Reddy, R. K., Das, N. D., Johnson, B., Meena, K., Krishnamurty, K., Sreenivas,
520 B. and Srivastava N. N. (2000). A weather-based scheme to advise on limited chemical control of groundnut leaf
521 spot diseases in India. Experimental Agriculture, 36 (4). pp. 469-478.
- 522 7. Butler, D. R., Wadia, K. D. R., Jadav, D. R. (1994). Effects of leaf wetness and temperature on late leaf spot infection
523 of groundnut. Plant pathology 43:112-120.
- 524 8. Das, A., Ng, W. K. and Woon, Y. K. (2001). Rapid association rule mining, CIKM '01, Proceedings of the tenth
525 international conference on Information and knowledge management (CKIM'01), November 5-10, 2001, Atlanta,
526 GA, USA, pp. 474-481.
- 527 9. Das, I, Naveen, C. P. R. G., Yadav, S. S., Kodilkar, A. A., Shah, N.G., Merchant, S. N. and Desai, U. B. (2009).
528 WSN Monitoring of Weather and Crop Parameters for Possible Disease Risk Evaluation for Grape Farms - Sula
529 Vineyards, A Case Study, Proceedings of Conference on Geospatial Technologies and Applications, Geomatrix-
530 2009, IIT Bombay, Mumbai, 27-29, 2009
- 531 10. Duthie, J. A. (1997). Models of the response of foliar parasites to the combined effects of temperature and duration of
532 wetness. Phytopathology 87:1088-1095.
- 533 11. Davis, D. P. Jacobi, J. C. and Backman, P. A. (1993). Twenty-four-hour rainfall, a simple environmental variable for
534 predicting peanut leaf spot epidemics. Plant Dis.77:722-725
- 535 12. George, H. J. and Pat, L. (1995). Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the
536 Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Mateo. pp. 338-345.

13. Georg, R., Rudolf, K., Martin S., Peter, W. (2008). Data Mining with Neural Networks for Wheat Yield Prediction. P. Perner (ed.), Springer LNAI, Vol. 5077, pp. 47-56.
14. Greene, William H. (2002). Econometric analysis (5th ed.). New Jersey: Prentice Hall. ISBN 0-13-066189-9. Retrieved on 15 July 2009.
[http://bib.tiera.ru/DVD-\[4\] 010/Greene_W.H._Econometric_analysis_\(2002\)\(5th_ed.\)\(en\)\(983s\).pdf](http://bib.tiera.ru/DVD-[4] 010/Greene_W.H._Econometric_analysis_(2002)(5th_ed.)(en)(983s).pdf).
15. Gupta, M. R. and Chen, Y. (2010). Theory and Use of the EM Algorithm, Foundations and Trends in Signal Processing, Vol. 4, No. 3 (2010) 223–296
16. Izge, A. U., Mohammed, Z. H. and Goni, A. (2007). Levels of variability in groundnut (*Arachis hypogaea* L.) to cercospora leaf spot disease – implication for selection, African Journal of Agricultural Research Vol. 2 (4), pp. 182-186, April 2007.
17. Jacobi, J. C., Davis, D. P. and Brannen, P. M. (1995). AU-Peanuts Advisory I: Modification of rule-based system for scheduling peanut leaf spot fungicide application. Plant Dis 79: 666-671.
18. Jhorar, O. P., Mavi, H.S. and Dhiman, J. S. (1987). A graphical Technique for short-range Forecast of Early and Late Leaf Spot disease of groundnut. Journal Res. Punjab Agriculture Univ. 24(4): 607-12, December 1987.
19. Liu, H., Liao, G., Yang F. (2008). Application of wireless sensor network in agriculture producing. Journal of Agricultural Network Information, 2008, (11):16 – 18.
20. Matese, A., Gennaro, S. F. Di, Zaldei, A., Genesio, L., Vaccari, F. P. (2009). A wireless sensor network for precision viticulture: The NAV system, Computers and Electronics in Agriculture, Volume 69, Issue 1, November 2009, Pages 51-58.
21. Mitchell, T. M. (2010). Generative and Discriminative Classifier: Naive Bayes and Logistic Regression, In Machine Learning, McGraw Hill (Draft of January 19, 2010), pp. 1-17.
22. Monroe, J. S., Santini, J. B. and Latin, R. (1997). A model defining the relationship between temperature and leaf wetness duration, and infection of watermelon by *Colletotrichum orbiculare*. Plant Dis. 81:739-742.
23. Prabhakar, T. V., Jamadagni, H. S., Sahu, A, Prasad, R. V. (2010). "Lessons from the Sparse Sensor Network Deployment in Rural India", 11th International Conference on Distributed Computing and Networking (ICDCN' 10).Springer LNCS, Vol- 5935, pp. 104-115.
24. Reddy, D. D. R. and Ravindra, B. (2008). Groundnut Insect Pests, Diseases, Nutritional Disorders, Acharya N.G. Ranga Agricultural University, Information Bulletin No.4, pp. 20-36. (2008)
25. Royle, D.J. and Burr, D. J. (2008). The Place of Multiple Regression Analysis in Modern Approaches to Disease Control, EPPO Bulletin, 9(3), 155-163.
26. Singh, F. and Oswalt, D.L.,(1992). Major Diseases of Groundnut, ICRISAT-skill development series No.6, pp. 50-36.

27. SPANN Lab (2009). SPANN Lab. WSN project report, Electrical Engineering Department, IIT Bombay, Available at, [Http://www.ee.iitb.ac.in/spann/agrisens/](http://www.ee.iitb.ac.in/spann/agrisens/) (Accessed on 23 March 2009).
28. Subrahmanyam, P., Reddy, D. V. R., Gibbonsa, R. W., Rao, V. R., Garrenab, K. H.(1990). Peanut Production, Disease and Harvesting in india, International Journal of Pest Management, Volume 25, Issue 1 March 1990 , pages 25 - 29
29. Sudharsan, D., Adinarayana, J., Tripathy, A. K., Desai, U. B., Merchant, S. N., Ninomiya S., Hirafuji, M., Kiura, T. and Tanaka, K. (2011). GeoSense: Geo-ICT and Wireless Sensor Network based Dynamic Real-time System for Precision Agriculture (2011); Proceeding of CIGR, Conference CIGR2011, September 19-23, 2011, Tokyo, Japan
30. Tan, P. N., Steinbach, M., Kumar, V.(2006). Data Mining and Regression, Pearson Addison Wesley (edn.), pp. 729-733 (2006)
31. Tripathy, A.K., Adinarayana, J., Sudharsan, D., Merchant, S.N., Desai, U.B., Vijayalakshmi, K., Reddy, D.R., Sreenivas, G., Ninomiya, S., Hirafuji, M., Kiura, T., Tanaka, K.(2011). Data Mining and Wireless Sensor Network for Agriculture Pest/Disease Prediction; 2011 World Congress on Information and Communication Technologies (WICT 2011), 11-14 December 2011, Mumbai, India. IEEE Catalog: CFP1168R-CDR, ISBN: 978-1-4673-0125-1, Pp: 1233-1238.
32. Vijayalakshmi, K., Reddy, Raji D., Varma, N.R.G. and Pranuthi, G. (2009). Weather based pest and disease forewarning models in groundnut in the context of climate change, Workshop Proceedings: Impact of Climate Change on Agriculture, ISPRS Archives XXXVIII-8/W3: pp. 48-50.
33. Wang, Y., Huang, L., Wu, J. and Xu, H. (2007). Wireless Sensor Networks for Intensive Irrigated Agriculture, Consumer Communications and Networking Conference, 2007. CCNC 2007. 4th IEEE , vol., no., pp.197-201, Jan. 2007
34. Wu, L., Damicone, J. P., Duthie, J. A. and Melouk, H. A. (1999). Effects of temperature and wetness duration on infection of peanut cultivars by *Cercospora arachidicola*, *Phytopathology* 89: 653-659.
35. Yuanguai, L., Haixia, X., Wei, Y. and Xingwu, Z. (2009). Application of Zigbee wireless sensor network in Precision agriculture [J]. *Journal of Qiongzhou University*, 2009, (5): 32-34, 37.

601 **Table Captions**

602 Table 1. Groundnut LS on PDI scale (Modified from Subrahamanyam et al.[28])

603 Table 2. Overall interpretation of correlation index values

604 Table 3. LS prediction Kharif 2010, D2

605 Table 4. LS prediction Rabi 2010-11, D2

606 Table 5. LS prediction Rabi 2010-11, D4

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

List of Tables

Table 1. Groundnut LS on PDI scale

Score	Description	Disease severity (%)
0	No disease.	0
1	Lesions largely on lower leaves; very few lesions on middle leaves; defoliation of some leaflets evident on lower leaves.	1-10
2	Lesions on all lower and middle leaves; 50% defoliation of lower leaves.	11-40
3	Lesions on all leaves but less severe on top leaves; defoliation of all lower and some middle leaves.	41-60
4	Defoliation of all lower and middle leaves; lesions severe on top leaves and some defoliation of top leaves evident.	61-80
5	Defoliation of almost all leaves leaving bare stems; some leaflets may be present, but with severe LSs.	81-100

Table 2. Overall interpretation of correlation index values

Correlation value (- 1 to +1)	Correlation Levels (Such as Leaf Spot with RH ₁)	Variables (Leaf Spot)
0.0 to 0.1 1.0 (+ve)	Negligible or	Tmax ET RH ₂
0.0 to - 0.1 (-ve)	No correlation	
> 0.15 to 0.5	Moderate (+ve)	RF RD, WS
< - 0.15 to - 0.5	Moderate (-ve)	SH
> 0.5 to 1 (+ve)	Strong (Good Correlation)	AC LW RH ₁
< -0.5 to - 1 (-ve)	Strong (Good Correlation)	Tmin

Table 3. LS prediction Kharif 2010, D2

Date	GLS	MVR model	1WK	CWK	ARI
24/7/2010	0.00	0.00	0.00	0.00	0.00
30/7/2010	0.00	3.55	3.55	3.55	2.41
11/8/2010	0.00	0.70	3.11	2.57	3.84
17/8/2010	3.90	5.80	5.09	4.84	6.42
26/8/2010	27.50	30.50	31.57	29.07	35.04
4/9/2010	37.60	38.80	40.25	39.47	43.89
8/9/2010	63.60	70.20	73.28	68.65	73.13
16/9/2010	79.90	81.40	83.45	80.17	85.08
23/9/2010	60.90	65.20			
30/9/2010	64.00	60.00			

Table 4. LS prediction Rabi 2010-11, D2

Date	GLS	MVR model	1WK	CWK	ARI
19/11/2010	0.58	3.78	3.78	3.78	5.79
26/11/2010	0.97	5.31	4.89	4.53	7.12
5/12/2010	2.34	6.89	6.11	5.29	8.16
12/12/2010	5.88	8.29	7.76	6.81	12.89
24/12/2010	42.07	45.21	45.81	44.21	53.57
2/1/2011	50.46	51.87	53.06	48.44	68.11
11/1/2011	57.64	61.28	63.78	59.64	70.24
23/1/2011	55.19	51.69			
30/1/2011	43.48	47.31			
9/2/2011	42.12	45.26			
17/2/2011	40.78	43.02			
25/2/2011	38.45	41.59			

648
649
650

651
652
653

654
655
656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

Table 10. LS prediction Rabi 2010-11, D3

Date	GLS	MVR model	1WK	CWK	ARI
24/12/2010	6.44	8.15	7.09	5.12	7.56
13/01/2011	4.51	7.54	7.88	5.89	10.89
24/01/2011	6.71	9.23	10.31	6.77	15.37
30/01/2011	14.85	16.78	18.12	13.19	23.08
12/2/2011	31.23	38.89	37.06	35.09	42.73
20/2/2011	34.81	47.11	45.28	37.43	52.24
27/2/2011	41.93	52.35	55.24	45.88	58.09
5/3/2011	52.41	60.14	61.28	54.21	65.95
13/3/2011	67.02	71.58			
20/3/2011	79.18	77.62			
28/3/2011	69.02	61.28			

Table 5. LS prediction Rabi 2010-11, D4

Date	GLS	MVR model	1WK	CWK	ARI
13/1/2011	1.84	3.78	4.19	2.03	4.38
25/1/2011	4.30	5.89	7.66	4.55	8.04
30/1/2011	10.67	15.25	17.25	12.45	18.69
12/2/2011	26.12	30.81	28.04	26.68	36.14
20/2/2011	25.83	31.01	29.73	26.86	37.55
27/2/2011	25.22	29.14	30.81	27.39	38.95
5/3/2011	26.32	32.78	35.12	30.45	41.27
13/3/2011	46.32	50.27	53.04	47.23	57.11
21/3/2011	48.24	52.41	54.07	50.17	61.02
29/3/2011	55.09	56.34			
8/4/2011	56.71	54.89			
11/4/2011	51.63	48.11			

List of Figure Captions

- Figure 1. Standard experimental layout
- Figure 2. Surveillance data collection from one square meter area
- Figure 3. WSN architecture in the experimental site with AgriSens
- Figure 4. AgriSens sensor network deployment in the field
- Figure 5. DM Processing flow for pest/disease dynamics
- Figure 6. Correlation index values for LS with weather and crop age
- Figure 7. LS in Kharif 2009 in D1
- Figure 8. Leaf Spot incidence as seen in the test bed

Figure 1. Standard experimental layout

Figure 2. Surveillance data collection from one square meter area

Figure 3. WSN architecture in the experimental site with AgriSens

Figure 4. AgriSens sensor network deployment in the field

Figure 5. DM Processing flow for pest/disease dynamics

Figure 6. Correlation index values for LS with weather and crop age

Figure 7. LS in Kharif 2009 in D1

736
737

738
739
740

Figure 8. Leaf Spot incidence as seen in the test bed