Applied Statistical Model and Remote Sensing for Decision Management System for Soybean

Mohammad Taheri
Department of Electrical Engineering and Computer Science, South Dakota State University
Brookings, SD, 57007, USA
mohammad.taheri@jacks.sdstate.edu

Dheeman Saha
Department of Electrical Engineering and Computer Science, South Dakota State University
Brookings, SD, 57007, USA
Dheeman.Saha@sdstate.edu

Gary Hatfield
Department of Mathematics and Statistics, South Dakota State University
Brookings, SD, 57007, USA
gary.hatfield@sdstate.edu

Emmanuel Byamukama
Department of Agronomy, Horticulture and Plant Science, South Dakota State University
Brookings, SD, 57007, USA
emmanuel.byamukama@sdstate.edu

Sung Y. Shin
Department of Electrical Engineering and Computer Science, South Dakota State University
Brookings, SD, 57007, USA
sung.shin@sdstate.edu

ABSTRACT

This paper proposes a Decision Management System to identify the white mold regions from the soybean fields using Autologistic Statistical Model (ASM) and Remote Sensing (RS) data analysis with commercially available Big Data sets as input data. In order to develop an identification model, numerous types of data need to be considered. In this study, the data that was used is satellite image pixel values, and data gathered from the field such as precipitation, yield, elevation, humidity, wind speed, wind direction and geospatial locations. The model evaluated the outcome using this information as input parameters and provided an overall estimation of the white mold region in the soybean fields. Based on the evaluation of the result, the accuracy rate of the proposed methods 84% which is a promising result due to the fact that each pixel of the satellite image is 30 by 30 meters.

CCS Concepts

•Mathematics of computing → Markov networks; •Computing methodologies → Image segmentation; •Computing methodologies → Feature selection; Support vector machines;

Keywords

Decision Management System; Satellite Imagery; Remote Sensing (RS); Autologistics Statistical Model (ASM); Support Vector Machine (SVM); Markov Random Model; Big Data

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

RACS'17, September 20–23, 2017, Krakow, Poland
© 2017 Association for Computing Machinery.

ACM ISBN 978-1-4503-5027-3/17/09…$15.00

https://doi.org/10.1145/3129676.3129710

1. INTRODUCTION

Precision Agriculture (PA) is one of the key factors of the agricultural revolution. In PA, it is essential to monitor the growth of crop production at different growth stages of the crop and measuring the changes such as crop yield, environmental parameters such as humidity, temperature, wind direction, and wind speed. Over the past few decades, PA is being developed with Geospatial technologies including Geographic Information Systems (GIS), the Global Positioning System (GPS), and Remote Sensing (RS) [1]. One of the best ways to improve efficiency in any agricultural activities is to make use of statistical models and RS data as a result of using these both tools, the developed model will give an estimated idea of infected regions in cultivated fields and will aid the farmers to make better decisions in the upcoming seasons.

PA is another area in which statistical models and big data analysis along with RS can be applied for multiple purposes. It can track the growth rate of a specific crop, distribution of different crops in a region, monitoring the crop disease over a region, the crop's yield and much more. Landsat-8 images are used as the source for RS. Landsat is a NASA-led enterprize for the acquisition of the earth landscape. Since its introduction used extensively in various fields from the agriculture to the education-purpose. The Landsat series of satellites provides the longest temporal record of moderate resolution multispectral data of the Earth's surface on a global basis for more than 40 years. The Landsat record stayed unbroken, proving a unique resource to assist a broad range of specialists in managing the world's food, water, forests, and other natural resources for a growing world population. It is a record unmatched in quality, detail, coverage, and value [2].

In this study, different data sources and data analysis algorithms were used to make a decision management system that identifies the white mold regions in soybean fields, such as remote sensing and satellite imageries, machine learning techniques, Autologistic Model which is a Markov Random Model where the outcome is randomized, and data gathered from soybean fields. Our Region of Interest (ROI) is the southern part of Brookings, SD and the northern part of Sioux Falls, SD, as of 2017, the principal crops production in this region are corn and soybean.

Support Vector Machine (SVM) is a supervised machine learning technique used in this study to classify the pixels in Landsat images into two predefined classes; white mold and healthy soybean. To identify the White-Mold in the selected region/field, an adequate number of Landsat images along with machine learning techniques can be used for a better classification of the Landsat pixels. Machine learning techniques are based on the study of patterns, in this case, patterns among Landsat-8 pixels, so they can make predictions based on the information given to them and information they have gathered from the data them by themselves [3]. There are many approaches to machine learning, some commonly used approaches are Support Vector Machine (SVM), similarity measurement, Artificial Neural Network (ANN), genetic algorithms and much more. ANN is one of the common supervised learning techniques, which can be also used for classification, but greater computational burden, proneness to overfitting, and not having enough training input data are some of its downsides [4]. In this paper SVM was used to classify the pixels into two classes and extract white mold pixels, the result is cross checked with an expert for validation and later used as one of the input data for the statistical model.

In this paper, a decision management system was developed based on techniques such as; remote sensing, statistical models, machine learning and, data sources such as satellite imagery and data gathered from the soybean fields. The developed model shows promising result based on the evaluation of the generated result by the system with 84% accuracy. An overview of the system is shown in Figure 1.

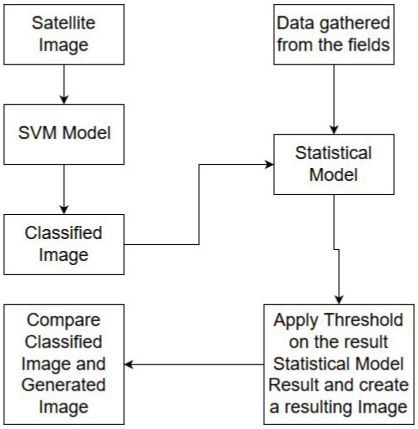
Satellite Image
Data gathered from the fields
SVM Model
Statistical Model
Classified Image
Compare Classified Image and Generated Image
Apply Threshold on the result Statistical Model Result and create a resulting Image

Figure 1. Overview of the proposed method

This paper is organized as follows; Section 2 describes input data used in this study. Section 3 explain the overall process and the methodology used. Experimental result in Section 4. Then in Section 5 conclusion is discussed.

2. Input Data

In addition to satellite imageries, data was gathered from the selected soybean fields. We have found several factors, primarily using WolframAlpha [5]; the computational engine then relays the user's input to relevant nodes such as the nearest weather station, for instance, which could be found in the input. However, the result relying on this methodology varies city by city, since not every city is equipped with a weather station. Hence, if the gathered information turned out to be insufficient, we have also used the National Weather Service [6] as additional sources.

Regarding the yield data, we've used the USDA's annual report. Although we cannot obtain the exact data, field by field, we can nonetheless obtain the annual estimate on data such as acres harvested, planted, production and so on, county by county.

We've tried to use the LiDAR [7] for additional data, but the USGS doesn't provide the LiDAR dataset to South Dakota, as of 2017, its service appears to be limited to urban areas.

3. METHOD

3.1 Overview of the Statistical Model

The proposed model for the evaluation of the agricultural fields consists of 3 major input images. The data is generated from different platforms. The three types of data that will be considered for the model are; statistical data, pixel values extracted from the selected soybean images and data gathered from the fields. The statistical data is generated by analyzing the clusters of pixels. Where the field's data is generated by extracting the information from the images and the features are also extracted from the image which consists of the information of the pixels. The last set of data consists of the yield, elevation, humidity, and climate. These set of data are generated externally which is available online [5].

3.2 Statistical Model pre-processing

Using the algorithm defined for classification of white mold and healthy soybean pixels in the input images were classified either as white mold or healthy soybean first by using SVM as a supervised machine learning technique and then by cross validation with the filed expert. As it can be seen in Figure 2 a) some pixels have a different color than the others. It not easy to say that one specific color is representing white mold. Thus, the classification and extraction algorithm was used to extract the white mold pixels. In addition, it is important to only work with pixels representing the soybean field and not the pixels outside of the field, thus 2 pixels are removed from each side of the image. An example of the input test data and the pre-processed one with 2 pixels removed from each side is shown in Figure 2.

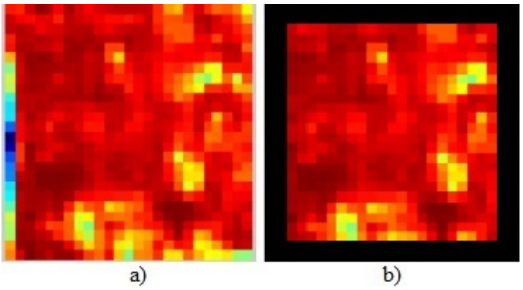

Figure 2. An example of the a) input test image and b) the corresponding result generated by the extraction algorithm

After pre-processing step, the classification and extraction algorithm is applied to extract the white mold pixels, which is shown in Figure 3.

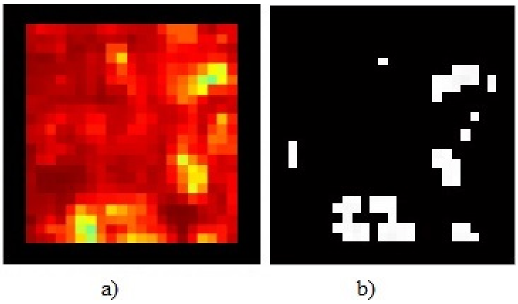

Figure 3. An example of the a) pre-processed test image and b) the corresponding result generated by the classification and extraction algorithm

When classification and extraction step is done, a label is assigned to each pixel in the image, 1 if it is white mold and 0 if it is not (healthy soybean). Then a table is generated which holds the coordinates of each pixel and the label assigned to it called label table. Later, the data collected from the fields merged to this label table. The new table was used as input data for the statistical model which explained in Section 3.3.

3.3 Statistical Model

The primary idea of the statistical model makes use of the Autologistic Model [8] which is a Markov Random Model [9] which is a decision-making framework where the outcome is randomized. The Autologistic Model is primary suitable for white mold detection from the soybean fields where the binary outcome of the decision determines the presence of infected regions from a given field. The inference of the outcome is more stable if the idea of the Autologistic Model is combined with the Pseudolikelihood Inference [8] which is an approximation of a joint probability distribution of the random variable. Therefore, combining both the statistical ideas the decision of the binary outcome indicates the presence or absence of white mold from the input images.

The Autologistic Model was reparametrized [10] by centering the autocovariate with the resulting conditional log odds given by:

$$\log\frac{\mathrm{P}\left(Z_i = 1|\{Z_j:(i,j)\}\in E\right)}{\mathrm{P}\left(Z_i = 0|\{Z_j:(i,j)\}\in E\right)} = x_i'\beta + \eta\sum_{j:(i,j)\in E}\left(Z_j - \mu_j\right)$$

Where $\mu_j = \{1+\exp(-x_j'\beta)\}^{-1}$the independence expectation of is Z_j, x_i is a p-vector of spatial predictors associated with the ith area unit. $\beta$is the p-vector of the spatial regression coefficient, the η is a spatial dependence parameter and Z_j is the autocovariate. The coefficient for 3 test fields are provided in the below Table 1.

Table 1. Coefficient of the Model

Field	β_X	β_Y	η	Pixel Grid Size
1	-0.1447	-0.3300	2.5030	23x24
2	-0.05267	0.00932	2.19900	18x11
3	-1.2720	-0.8881	1.1850	7x14

The coefficient of the model of each field are interpreted by the centered autocovariate value. The centered autocovariate is a signed value measured locally against the large-scale structure. This captures the spatial dependence and the η represents the "reactivity" of an observation to its neighboring pixel values, conditioning on the large-scale structure represented by the regression component of the model β_X.

The outcome of the model is the "Fitted Value" which is defined as the probability of the pixel being white-mold. If the outcome of the joint probability is 1 then presence of white-mold is there in that pixel location or vice versa. Since the model is prepared to have an overall assumption of the white-mold detection the value is selected to its mean value which is 0.5.

But optimal threshold value can be obtained based on the predicted probabilities of a specific region or through determining a global fitted value. We can even look at the Receiver Operating Characteristic (ROC) curve and other measures to determine the optimal classification.

3.4 Result of statistical modeling

Since the outcome of the model is to determine the regions of the white-mold the inference for estimated model coefficients in Table 1 is available for the autologistic function as an option for confidence intervals obtained using bootstrapping. However, spatial bootstrapping can be time consuming. Examining the empirical distributions from a Monte Carlo study provides insight into possible theoretical distributions that can be used to make inference about the underlying population.

3.5 Monte Carlo Simulation

As there is a certain level of uncertainty associated with the outcome of the model. So, a parametric Monte Carlo Simulation is used in three different grid layouts to examine the distributions of η and proportion of white mold for a model with parameters similar to estimated values from Field 3. The function autologistic in R package ngspatial [11] was used to simulate 1000 draws from the centered autologisitc model with parameters $\beta_X = -1, \beta_Y = -1,$ and $\eta = 1.5$ for grid sizes 10x10, 30x30, and 50x50. Run times were 17 seconds for the 10x10 grid, 37.7 minutes for the 30x30 grid, and 7.49 hours for the 50x50 grid on a desktop computer with a 64-bit Operating System, an Intel Core i7 CPU at 2.93 GHz, and 12 GB RAM.

An example of the histograms of η for the grid sizes is given in Figure 4 and numerical summaries are in Table 2.

Table 2. Numerical Summary of Empirical Distributions of η

Grid Size	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Normality p-value
10x10	0.270	1.339	1.524	1.529	1.707	3.374	<0.001
30x30	1.204	1.420	1.481	1.493	1.559	1.862	<0.001
50x50	1.300	1.450	1.489	1.496	1.536	1.759	<0.001

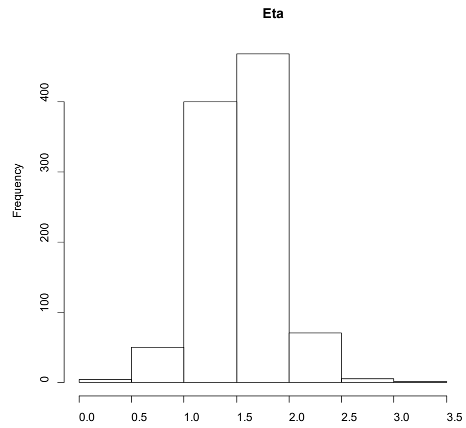

Figure 4. Proportion of White-Mold in Grid Layout 10x10

The p-value tests the hypothesis where Null: The data is from a normal distribution and Alternative: The data is not from a normal distribution. A p-value must be reported to indicate which hypothesis to conclude. Either fail to reject the null hypothesis (p-value greater than or equal to 0.01) or reject the null hypothesis (p-value less than 0.01).

The empirical distributions of η are unimodal and slightly skewed to the right. An example of the histograms of the proportion of white mold is given in Figure 5 and numerical summaries are in Table 3.

Table 3. Numerical Summary of Empirical Distributions of Proportion of White Mold

Grid Size	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
10x10	0.02	0.38	0.50	0.50	0.62	0.97
30x30	0.328	0.472	0.502	0.501	0.531	0.658
50x50	0.400	0.482	0.498	0.499	0.517	0.588

It is seen in Table 2 that the range for a 10x10 grid is 0.95, for a 30x30 grid is 0.330, and for a 50x50 grid is 0.188. Thus, as the number of pixels increases, the range of values for proportion of white mold decreases indicating an increase in precision.

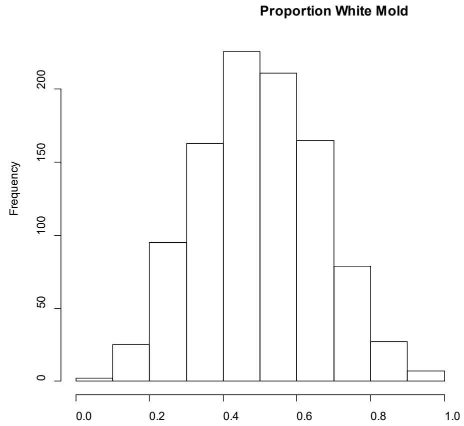

Figure 5. Proportion of White-Mold in Grid Layout 10x10

The distributions of the proportion of white mold are unimodal and symmetric. Based on these results, we recommend a grid size of at least 30x30 pixels.

4. RESULTS

In the post-processing, the result is generated based on the result of the statistical model. After applying the statistical model on the data, the result need to be mapped back to an image using a threshold on the fitted probability value generated by the statistical model. Then the result will be compared with the result of the algorithm for classifying White mold and healthy soybean pixels. Thus, we can evaluate the result. An example of the generated image based on the result of statistical model is shown in Figure 6.

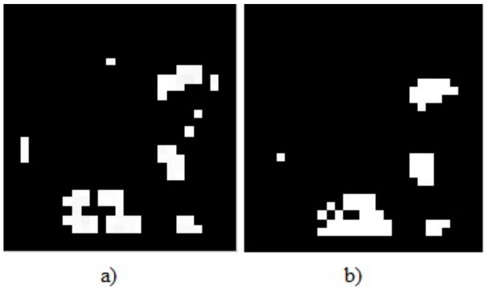

Figure 6. a) Result of algorithm designed for classifying white mold and healthy soybean pixels, b) the result of mapping back the result of statistical data into an image.

After comparing and matching pixels in images a) and b) the matching percentage is 84% on average for three test images.

5. Conclusion

Satellite imagery is wieldy used in different areas such as precision agriculture which can provide useful information about the fields. Thus, it can be used as one the sources to make a decision management system for a specific crop. Using Remote Sensing along with Statistical Models we can enhance the accuracy of the system. The centered autologistic model was used as our statistical model which provides coefficients that are easily interpreted and have a high level of accuracy for predicting the occurrence of white mold. Covariates can be more statistically significant if amount of information for the three soybean fields is increased and more data sources are used. The overall accuracy for three soybean fields is 84% which looks promising as each pixel in the Landsat satellite imagery is 30 by 30 meters. But more data sets are needed to determine the optimal threshold value for classifying predicted pixel probabilities as white mold.

6. ACKNOWLEDGMENTS

This work was supported by SD GOED Research Team Development (RTD) grant by the South Dakota Board of Regents.

7. REFERENCES

[1] J. C. B. Hofmann-Wellenhof , H. Lichtenegger, *Global Positioning System: Theory and Practice*, 5th Editio. 2001.

[2] “About Landsat,” 2017. [Online]. Available: https://landsat.usgs.gov/about-landsat.

[3] E. Alpaydin, *Introduction to Machine Learning*, Second Edi. MIT Press, 2004.

[4] J. V Tu, “Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes.,” *J. Clin. Epidemiol.*, vol. 49, no. 11, pp. 1225–31, Nov. 1996.

[5] “WolfarmAlpha, computational knowledge engine.” [Online]. Available: https://www.wolframalpha.com/.

[6] “National Weather Service,” 2017. [Online]. Available: http://www.weather.gov/.

[7] “LiDar data on earthexplorer,” 2017.

[8] J. Hughes, M. Haran, and P. C. Caragea, “Autologistic models for binary data on a lattice,” *Environmetrics*, vol. 22, no. 7, pp. 857–871, Nov. 2011.

[9] A. R. Cassandra and A. R. Cassandra, “EXACT AND APPROXIMATE ALGORITHMS FOR PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES,” 1998.

[10] P. C. Caragea and M. S. Kaiser, “Autologistic models with interpretable parameters,” *J. Agric. Biol. Environ. Stat.*, vol. 14, no. 3, pp. 281–300, Sep. 2009.

[11] J. Hughes, “ngspatial: A Package for Fitting the Centered Autologistic and Sparse Spatial Generalized Linear Mixed Models for Areal Data,” *R J.*, vol. 6, no. December, pp. 81–95, 2013.