CIAT

Centro Internacional de Agricultura Tropical
International Center for Tropical Agriculture

GIS and Integrated Pest Management: Case Study for predicting BGMV occurrence in Guatemala, Nicaragua and Honduras

Justine Klass, Grégoire Leclerc, Francisco Morales and Pamela Anderson

Octubre 16-18, 1998

CGIAR Consultative Group on International Agricultural Research

Title: GIS and Integrated Pest Management: Case Study for predicting BGMV occurrence in Guatemala, Nicaragua and Honduras

Authors: J. Klass, G. Leclerc, F. Morales and P. Anderson

Affiliations: CIAT (Centro Internacional de Agricultura Tropical), A.A. 6713, Cali, Colombia, S America

Presentation at the **First International Health Geographics Conference**, Baltimore, Maryland, Oct 16-18, 1998

Abstract

The paper demonstrates the use of some GIS and remote sensing techniques that can be used to predict the occurrence of a virus based upon the presence and absence of a virus. The analysis performed here is an attempt to evaluate the fuzzy-approach methods that can be used to predict an accurate virus outbreak using semi-quantitative field data. Some of the techniques included; multi-dimensional logistic regression, geographic weighted logistic regression analysis and fuzzy clustering. The results are presented and evaluated.

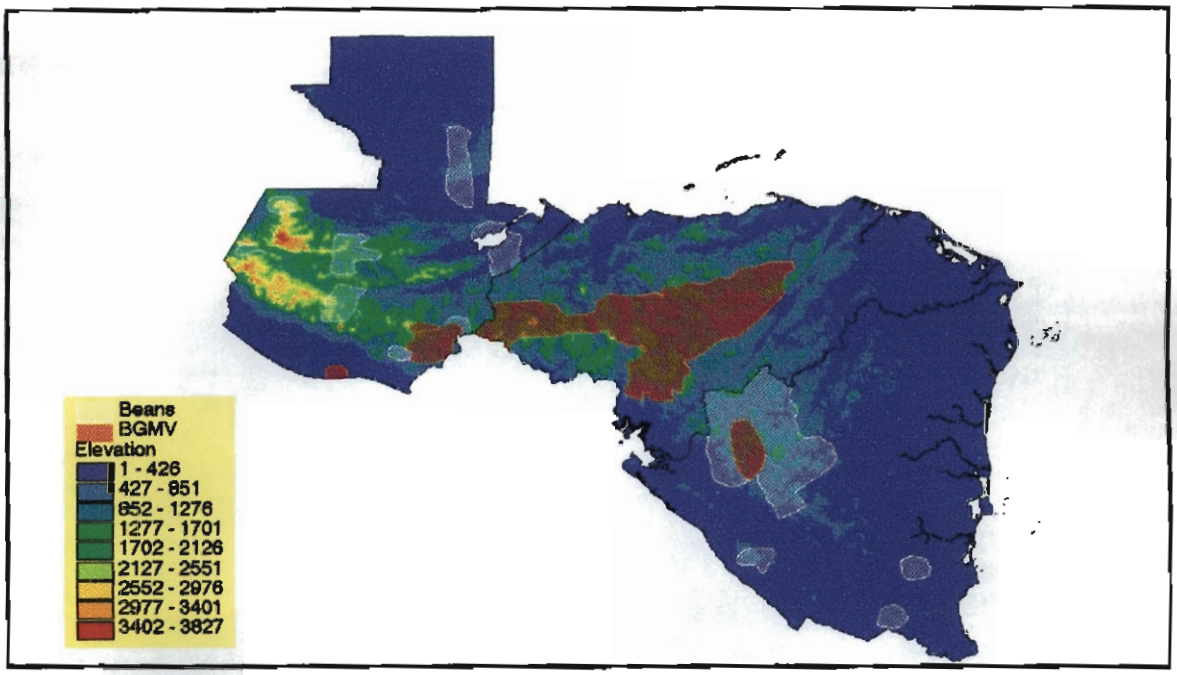
Introduction

Geographical information systems (GIS) and remote sensing are increasingly being used as a tool for investigating and monitoring pests, disease and virus occurrence of humans, plants and animals. For the purpose of this paper, we are interested in a plant disease, transmitted by a vector and that result in significant economic losses.

When studying an outbreak of a virus it is important to study the temporal and spatial changes that are caused by the pathogen. Thus it is important to ask where the epidemic is occurring, when can it be expected to occur and what factors contribute to the development of the disease. We emphasize the importance of analyzing the spatial relationship of the epidemic with its surrounding environmental factors.

A GIS can aid in describing the disease spatially. With additional descriptive information and spatial analysis techniques, the environments of the virus can be modeled to derive differing areas of potential risk within bean growing areas.

For the purpose of this paper, an analysis was conducted in Guatemala, Nicaragua and Honduras based upon the presence and absence of the virus in a single crop, beans. The two main objectives assess production risk areas and evaluate the methods used to predict the occurrence of the virus at a given location. The methods include a multi-dimensional logistic regression, geographically weighted logistic regression analysis, and fuzzy clustering. The results of each technique were compared and contrasted to determine which method provides the most accurate results.



Map 1: Study area illustrating the presence (red) and absence (white) of the virus.

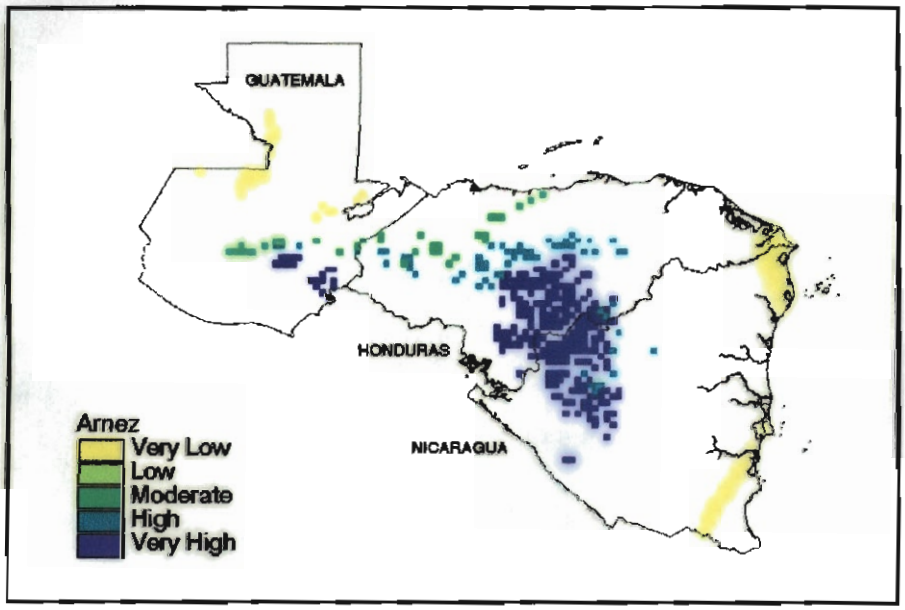
Environmental Factors for the vector, ***Bemisia tabaci***

	Physical Characteristics	Unit	Elevation Restrictions				
Incidence of the virus occurrence	Frequency of Whitefly		Very low (1)	Low (2)	Moderate (3)	High (4)	Very high (5)
Climate							
Humidity for ***Bemisia tabaci***	Precipitation	mm	>3000	2500-3000	2000-2500	1500-2000	<1500
	Consecutive dry months	Month	<1	1-2	2-3	3-4	>4
Temperature requirements	Temperature	C	<19 >30	26-30	---------	19-22	22-26
	Elevation	Msnm	<100 >1500	1500-2000	1000-1500	-------- -	500-1000

Table 1: ***Bemisia tabaci*** environmental factors based on a study in Costa Rica

Source: Arnez, Juan Eliseo Vallejos, 1997, **Sistema Experto para la Evaluacion del Impacto del Complejo *Bemisia tabaci* -Geminivirus, en Frijol, Tomate y Chile Dulce con Fines de Planificacion**, Centro Agronomico Tropical de Investigacion y Ensenanza, Turrialba, Costa Rica. pg 41

According to the table above, the occurrence of ***bemisia tabaci***, is related to elevation, temperature, humidity and the number of consecutive dry months. Thus, based on this study in Costa Rica, from the table it can be assumed that ***bemisia tabaci*** can be found at an elevation ranging from 0 - 1500 meters with a temperature range from 19 - 30 C. The environmental factors as described in table 1 are limited data and the elevation and physical characteristics for the vector are still being verified. However, the main purpose of this analysis is to show the different techniques that can be used with a GIS.



Map 2: Areas of ***bemisia tabaci*** occurrence based upon Arnez's environmental factors.

Methodology

The data used to complete this analysis was based upon the known location of bean growing areas within Nicaragua, Honduras and Guatemala. Of these areas, the critically affected bean golden mosaic virus (BGMV) areas were identified. The information for Nicaragua was based on local expert knowledge both here at CIAT and from Nicaragua (12). While for Honduras and Guatemala the data was obtained from a 1994 publication, **El Mosaico Dorado del Frijol: Avances de Investigacion**, (10)(11).

We constructed a grid representing the presence (value = 1) and absence (value = 0) of the virus for all bean growing areas. This was used as a mask to try to identify all the high-risk virus areas within the bean growing areas.

For the purpose of this paper, three methods were used and compared to estimate the location and probability of BGMV occurrence. These include:

a) the multi-dimensional logistic regression analysis (ESRI, Arc/INFO)
b) a geographically weighted logistic regression analysis (WGR program, Arc/INFO)
c) fuzzy clustering (PCI)

The climate surfaces were created by Corbett, J.D. and O'Brien, R.F (1997) at theTexas Agricultural Experiment Station, Texas A&M University, Blackland Research Center. The potential evapotranspiration surfaces, climate coefficients and surfaces used were created using ANUSPLIN (CRES, ANU, Canberra). These were created at a surface resolution of 3 arc-minutes.

The monthly surfaces were averaged and the mean annual surfaces were used. In this case, the following climate variables were used;

Mean annual minimum temperature
Mean annual maximum temperature
Annual rainfall
Mean annual evapotranspiration
Number of dry months with rainfall < 60 mm per month

From the means used, we characterized the sites in question.

a) Multi-dimensional logistic regression analysis

Multiple regression allows for the examination of the relationship between the known value of a dependent variable with the known values of a set of independent variables. Since the data being analyzed is represented by the presence or absence of the virus within bean growing regions, a logistic regression was used. In a logistic regression, the magnitude of occurrence of the phenomenon being modeled by the dependent variable is unknown. Instead, the known values of the dependent variable are represented by the phenomenon in question at the sample locations.

As in the one-dimensional case, the logistic regression can be used to predict that the probability that the virus will occur at an unsampled location based on the values of the independent variables. The sample was run with the logistic regression and the following results were obtained, table 2.

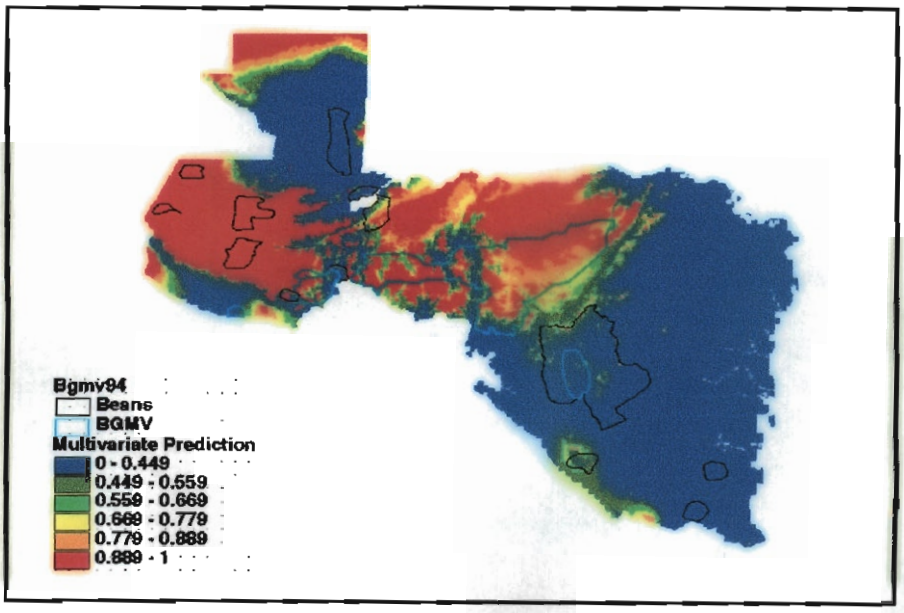
Coef #	Coeff	Coef #	Coeff
0	-25.723	***Mean Rainfall***	-0.041
Drymnth	-0.145	***Min Temperature***	-2.331
Povpe	50.897	***Max Temperature***	2.779
Elevation	-0.002		
RMS	0.345	***Chi-Square***	25.966

Table 2: Coefficient results of the logistic regression

The coefficients represent the transformation that maintains the linear relationship between the dependent and independent variables. These were substituted into the following equation for $P(x)$ and the results are illustrated in map 4.

$$P(x) = \frac{1}{1+\exp(-\sum a_1 \cdot x_1)}$$

where $\mathbf{a_l}$ is the regression coefficient and $\mathbf{x_l}$ are the independent variables



Map 4: Predicted areas of virus occurrence

b) Geographically Weighted Regression (GWR) Analysis

As with the logistic regression analysis, the geographically weighted regression also involves estimating the relationship between one variable and a set of independent variables for a set of zones. As discussed by Fotheringham, Charlton and Brundson, when applying a linear regression to a geographical data, it is assumed that the observations are independent of one another, which is not normally the case, and that there are no local variations in the parameter estimates for a study area. Thus, a GWR permits a local variation of parameter estimates by including a weighting scheme, as illustrated below.

$$P(g) = \beta_0(g) + \beta_1(g)x_1 + \beta_2(g)x_2 + e$$

Where g indicates the location of vector g whose parameters are to be estimated. The weights are chosen such that those observations near the point in space, where the parameter estimates desired have more influence on the result than observations further away. The weight calculations us a Gaussian Scheme where the weight for the ith observation is:

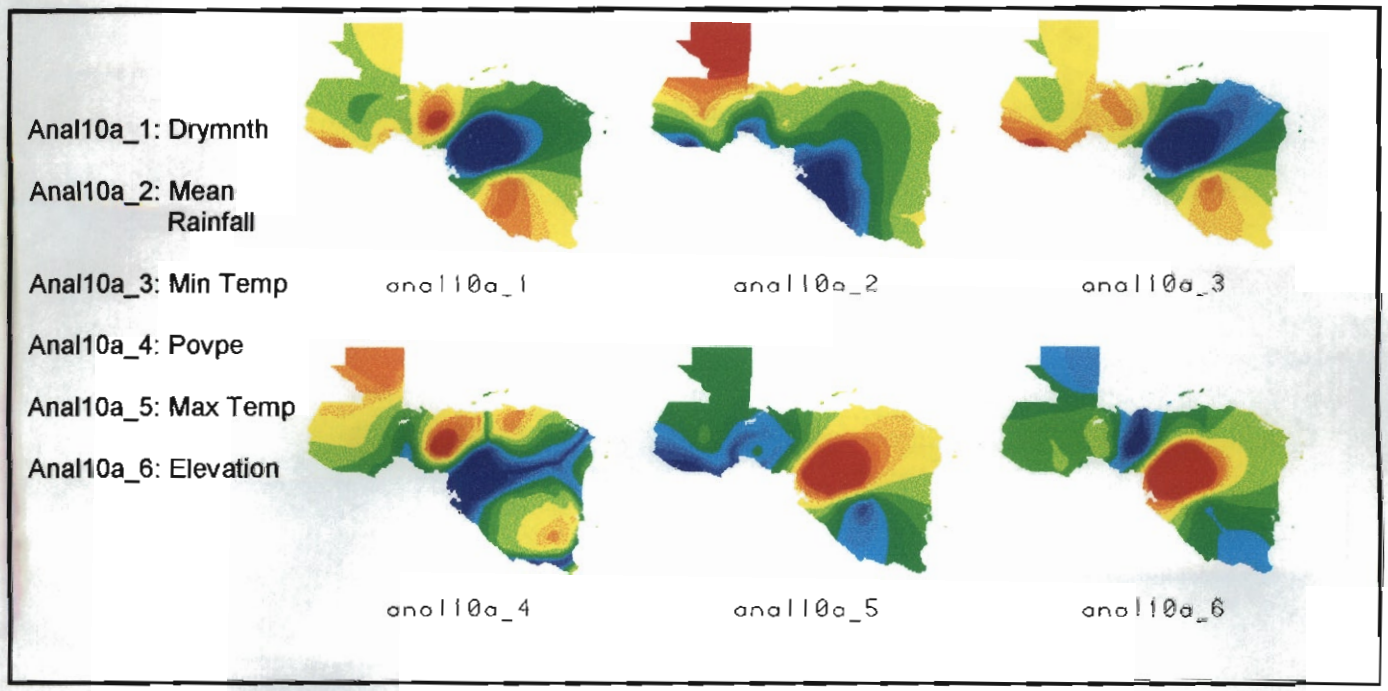
$$W_i(g) = e^{-(d/h)^{\wedge}2}$$

Where d is the Euclidean distance between the location of observation i and location g, and h is a quantity known as the bandwidth. The bandwidth was calculated by a crossvalidation technique (Stewart Fotheringham et al). The following coefficients were obtained using the geographically weighted regression, table 3.

Coef #	Coeff	Coef #	Coeff
Intercept	-2.452110005329	***Mean Rainfall***	-0.005382797206
Drymnth	-0.014944042145	***Min Temperature***	-0.259165555186
Povpe	6.787574445397	***Max Temperature***	0.317158800377
Elevation	-0.000249325821		
RMS Error	0.387393044813	***Error Sum of Squares***	31.665481316819

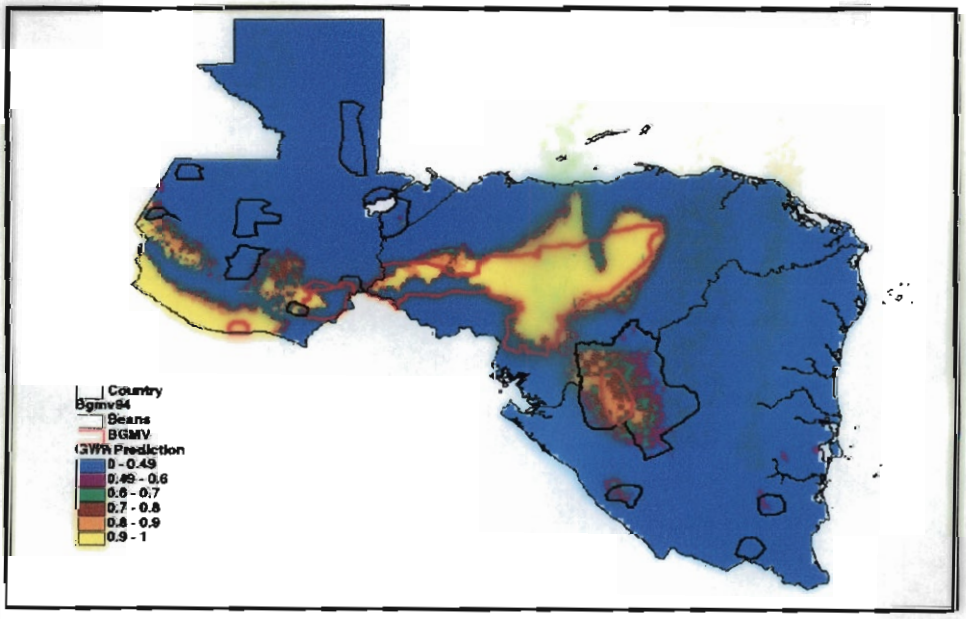
Table 3: Coefficient results from GWR

Using the sample, each theme was weighted accordingly, as illustrated in map 5.



Map 5: Weighting for each independent theme

The resulting parameters were combined with the original independent variable surfaces and mapped with the following results illustrated in Map 6.



Map 6: Geographic Weighted Regression results with a bandwidth of 0.67844

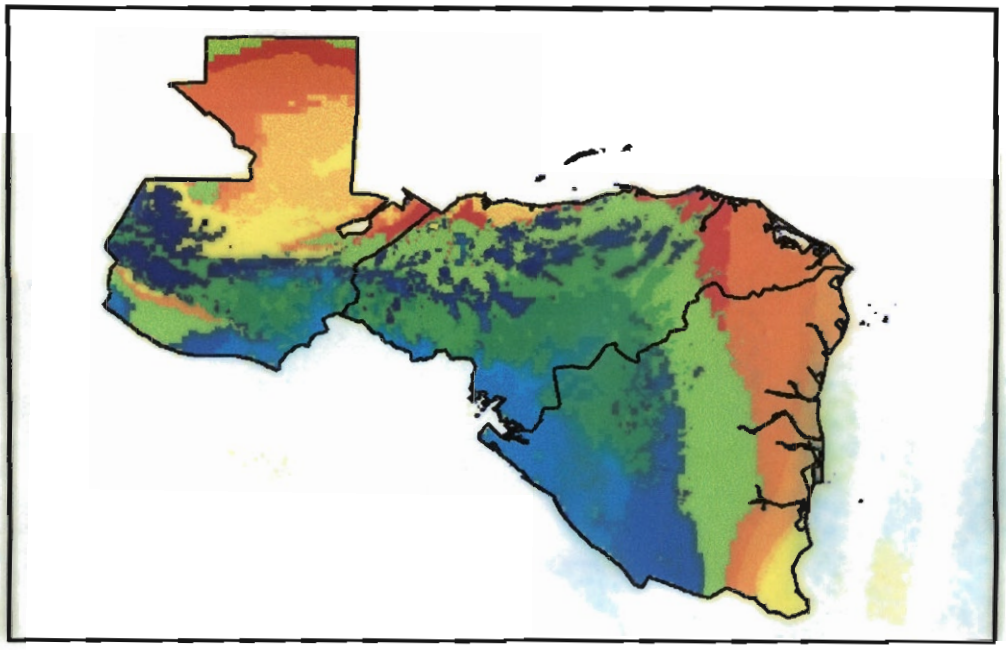
c) Fuzzy Clustering

Fuzzy clustering is a technique used for classifying remotely sensed data with very little intervention. A number of clusters are defined, representing the regions that are to be identified and an algorithm looks for natural groupings in the data and assigns cluster memberships using a maximum likelihood decision rule.

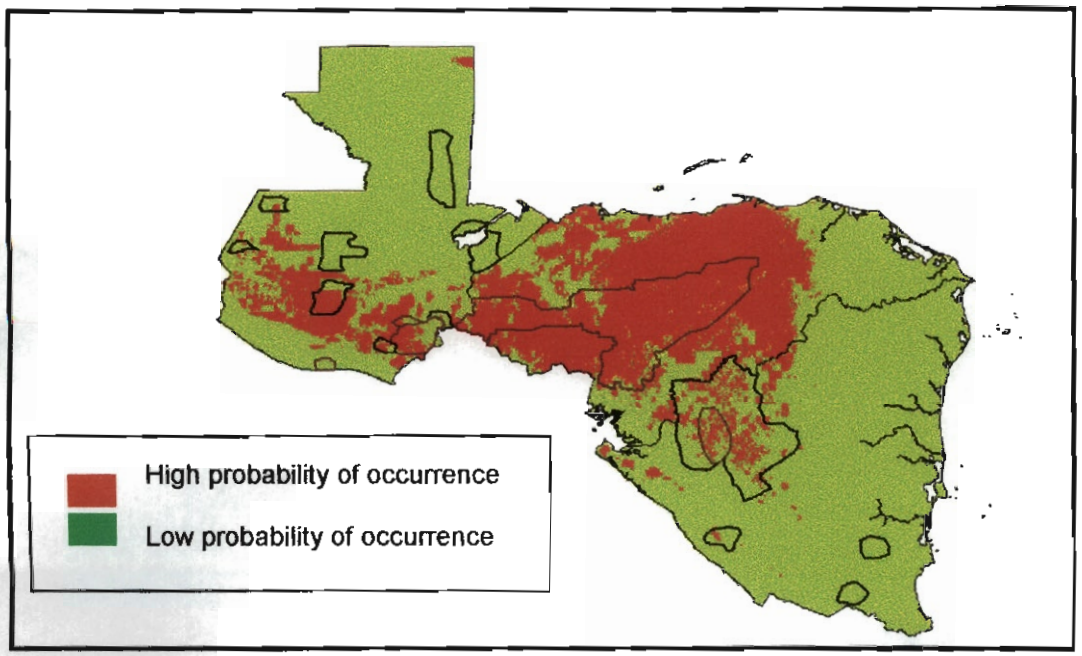
The maximum likelihood classifier method is considered to be superior since it considers the variance and covariance of the spectral signatures when classifying unknown pixels (Lillesand and Kiefer, 1994). To do so, it assumes that each signature has a population of DN values that are normally distributed. Next, with a probability density function, it determines the probability that a certain pixel belongs to a signature. The density function is used to generate equal probability contours around each signature defined. Pixels are then assigned to a class based on the highest probability. This is based on the lowest number of standard deviations from the mean.

A fuzzy clustering was used to partition the space defined by the five climate variables. Once the classifications were completed, we visually identified the clusters that were contained within the BGMV infected areas and viewed the additional potentially affected areas.

For the purpose of this analysis, the climate surfaces were classed into 32 unsupervised fuzzy classifications, map 7. The results are illustrated in map 8 and map 9.

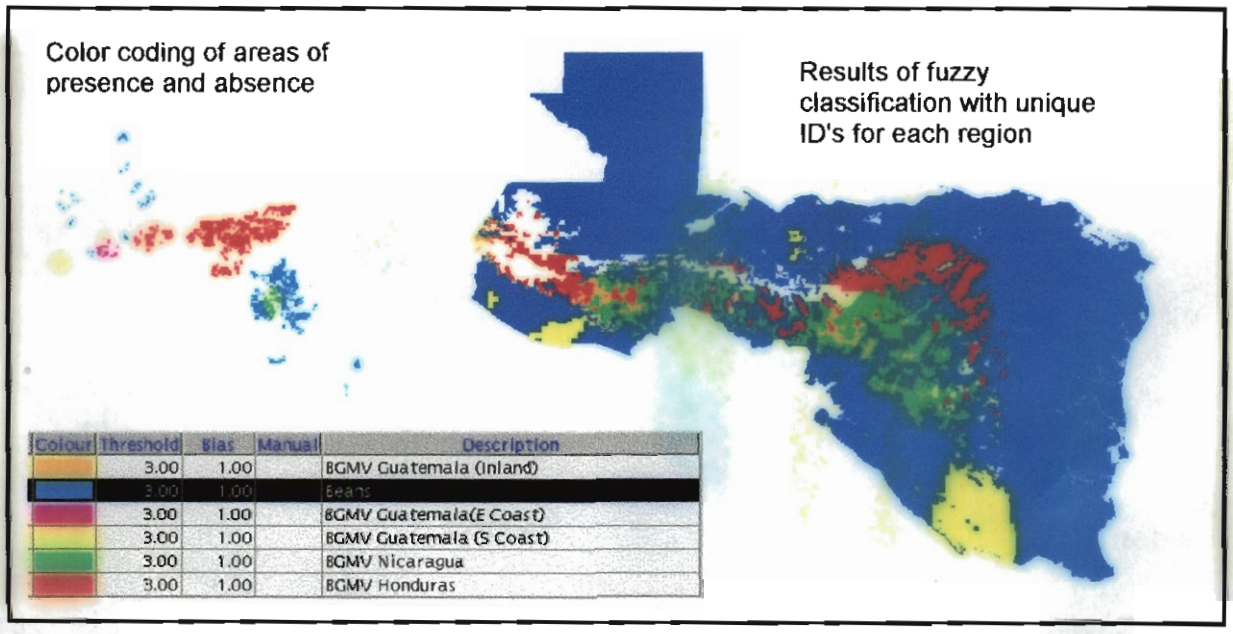


Map 7: Independent themes classed into 32 class fuzzy classification in PCI



Map 8: Occurrence of BGMV using a 32 class fuzzy classification in PCI with two training classification (presence (2) and absence (1))

Since the area being analyzed extends over a wide geographic range, each area will contain a wide range of influencing factors. Thus, each class was given a unique id classified again. The results are illustrated in Map 8.



Map 9: Occurrence of BGMV using a 32 class fuzzy classification in PCI with unique training classification for presence and absence of the virus.

Limitations:

Due to the scale and generalization of the data, accurate predictions of the virus are difficult. For this analysis five climate variables were used with the presence and absence of a virus occurring at a location. Limitations prevail with each of these variables including the virus information.

The climate surfaces were created using climate coefficients and by interpolating between climate stations. (Corbett, J.D and O'Brien, R.F) The main constraints of interpolated climate surfaces include the resolution of the DEM, the source data and the limitations of the interpolated method used.

Secondly, the virus information. The data where the virus occurs is a generalization and may include samples that do not represent the environmental factors required for the virus to exist. Since there is some generalization involved, the defined training areas may have induced some error and included pixels that do not represent the site in question. This was reduced for the virus areas by eliminating areas with an elevation greater than 1500 metres. In addition, cultivated areas were extrapolated using the USGS land use image.

Discussion:

The study area being analyzed covers a large geographic extent that includes varied topography, climate and ecological environments. From the data being used, Honduras has the largest sample size of the virus being present with smaller regions occurring in Nicaragua and Guatemala.

The data used to represent the virus is an approximation of the areas affected and not affected by the virus. Thus, errors can result due to a misrepresentation of the areas in question. This can result by including areas that do not represent the virus as well as omitting virus and non-virus areas. Due to the semi-quantitative field data, fuzzy approaches were more appropriate than simple-linear regressions.

The multi-dimensional logistic regression analysis is a technique suggested by ESRI and is commonly used in GIS for predicting the occurrence of an event based upon presence and absence data. The results are illustrated in Map 4. From the results it can be seen that all the virus occurring areas have been classified positive and include some non-virus occurring areas. Since Honduras contains the largest sample points of the virus occurring, a bias of the environmental factors required for the occurrence of the virus to exist may result. To omit such a bias, the geographically weighted regression can be used.

The geographically weighted regression allows for local variations by including a weighting scheme. From map 6, the areas containing the virus exist within the delineated areas with a strong occurrence in Honduras and near the coast of Guatemala. The general pattern

of occurrence follows a similar trend of the environmental factors required for the vector to exist (map2).

Lastly, a thirty-two class fuzzy classification was used. The results from the fuzzy clustering illustrate a strong occurrence in Honduras and Guatemala with smaller occurrences in Nicaragua. Again, a similar pattern of occurrence can be seen occurring with the environmental factors required for the vector to exist (map2). However, due to the classification of the training areas, the pixels representing Honduras have a greater influence than the other regions containing the virus. Thus, if each polygon is given a unique value and classified, the results differ, as illustrated in map 9. This also shows the areas with similar characteristics for a particular training site.

Conclusion:

The paper demonstrates the use of some GIS techniques that can be used to predict the occurrence of a virus based upon the presence and absence of a virus. The analysis performed here is an attempt to try to evaluate the fuzzy-approach methods that can be used to predict an accurate virus outbreak semi-quantitative field data. This analysis is by no way complete and still requires ground-truthing and a greater understanding of the environmental factors required for the vector to exist.

From the different methods that have been analyzed the classifications that have the closest results to the environmental factors described by Arnez are the geographically weighted regression analysis and the fuzzy clustering. From the coefficient results (table 2 and table 3) it can be noted that the humidity (povpe) and temperature are important contributing factors to the occurrence of the virus.

With the results obtained it is necessary to verify the outputs both in the field and with the experts. In addition, the environmental factors for the occurrence of the vector need to be evaluated since the results illustrate that the virus can occur in areas that are not included in the criteria delineated in table 1.

Future work:

- Verify the results with experts and in the field.
- Improve the sampling data and perform a more in-depth analysis.
- Improve upon the environmental factors required for the vector to exist at a location.

References

1. Anthony C Gatrell, Trevor C Bailey, Peter J Diggle and Barry S Rowlingson, 1996, **Spatial Point Pattern Analysis and its Application in Geographical Epidemiology**, Royal Geographical Society, England

2. A. Stewart Fortheringham, M E Charlton, and C Brundson, 1997, **Two Techniques for Exploring Non-stationarity in Geographical Data**, Geographical Systems. 4: 59-82

3. A. Stewart Fortheringham, Martin Charlton, and Christopher Brundson, 1997, **Geographically Weighted Regression**, http://www.ncl.ac.uk/~ngeog/GWR/

4. Arnez, Juan Eliseo Vallejos, 1997, **Sistema Experto para la Evaluacion del Impacto del Complejo *Bemisia tabaci* -Geminivirus, en Frijol, Tomate y Chile Dulce con Fines de Planificacion**, Msc Thesis, Centro Agronomico Tropical de Investigacion y Ensenanza, Turrialba, Costa Rica.

5. Butler et Al, 1983, **Bemisia tabaci (Homoptera: aleyrodidae): Development, Oviposition, and Longevity in Relation to Temperature**, Annals of the Entomological Society of America, 76, 2: pp 310-313

6. CIAT (Centro Internacional de Agricultura Tropical), 1996, **Sustainable Integrated Management of Whiteflies as Pests and Vectors of Plant Viruses in the Tropics**, CIAT, Cali, Colombia

7. Corbett, J.D, and O'Brien R.F., 1997, **The Spatial Characterization Tool**, Texas Agricultural Experiment Station, Texas A&M University, Blackland Research Center Report No. 97-03, Documentation and CDROM.

8. Lillesand and Kiefer, 1994, **Remote Sensing and Image Interpretation**, John Wiley & sons Inc. New York

9. Morales, F., **El Mosaico Dorado del Frijol: Avances de Investigacion**, 1994, edited and translated by Francisco J Morales, CIAT, Cali, Colombia, pp1 - 17

10. Rodriguez F., Diaz O., Escoto N.D., 1994, Honduras, **In El Mosaico Dorado del Frijol: Avances de Investigacion**, edited and translated by Francisco J Morales, CIAT, Cali, Colombia, pp 45-50

11. Rodriguez R., 1994, Guatemala, **In El Mosaico Dorado del Frijol: Avances de Investigacion**, edited and translated by Francisco J Morales, CIAT, Cali, Colombia, pp 34-39

12. Rojas A., Anderson, P., Morales, F. J., 1994, Nicaragua, **In El Mosaico Dorado del Frijol: Avances de Investigacion**, edited and translated by Francisco J Morales, CIAT, Cali, Colombia, pp 51-61

The data was collected by CIAT and the analysis was performed using the following software: ARC/INFO, MATHCAD and PCI.

The interpolated climate surfaces were obtained from J Corbett and R. F O'Brien of the Texas Agricultural Experiment Station, Texas A&M University.