Title: Using GIS techniques to aid in predicting a plant virus in beans

Authors: J. Klass[1], G. Leclerc[1], F. Morales[2] and J Wellens[3]

[1] Landuse and GIS Department
CIAT (Centro Internacional de Agricultura Tropical),
A.A. 6713, Cali, Colombia, S America
Tel: 1-650-833-6625(USA/Direct);Fax: 1-650-833-6626
E-mail: j.klass@cgiar.org, g.leclerc@cgiar.org

[2] Department of Virology
CIAT (Centro Internacional de Agricultura Tropical),
A.A. 6713, Cali, Colombia, S America
Tel: 1-650-833-6625(USA/Direct);Fax: 1-650-833-6626
E-mail: f.morales@cgiar.org

[3] Department of Geography
University of Leicester
Leicester, UK, LE1 7RH
Tel: +44 (0)116 252 3846;Fax: +44 (0)116 252 3854
E-mail: jw27@leicester.ac.uk

Abstract

Geographical information systems (GIS) assist us in mapping and analyzing outbreaks of diseases in plants, animals and humans. This paper describes how GIS are being used to model the intensity of the outbreak of a plant virus, bean golden mosaic virus (BGMV) in Guatemala, Honduras and El Salvador. BGMV is a geminivirus affecting beans *(Phaseolus vulgaris)* and is transmitted by a vector, the sweet potato whitefly (*Bemisia tabaci*). Once a plant is infected by the virus yield losses, at varying locations, can range from 40% to 100%. Plant pathologists can improve upon integrated pest management strategies to monitor virus movement and outbreaks by estimating the likelihood of risk in a cropping systems. For the purpose of this analysis three techniques were selected (multivariate logistic regression, Fourier transform with principle components analysis and a multi-process boolean analysis) to predict the spatial occurrence of BGMV in beans. The methods selected are based on the location of the virus (presence/absence) and the environmental factors determining the distribution of the vector. The process involves predicting the distribution of the vector by modeling and mapping the probability of occurrence using environmental variables, such as minimum and maximum temperature ranges, elevation, rainfall and number of dry months. The results of the methods are compared, evaluated and discussed.

Key Words: climate, *Bemisia tabaci,* model, GIS, disease, pest, *Phaseolus vulgaris*, bean

1. Introduction

In Guatemala, Honduras and El Salvador, bean (*Phaseolus vulgaris)* production is affected by biotic constraints during the dry season. Of these constraints, bean golden mosaic virus (BGMV) is a major problem (Morales, 1994). The virus is transmitted by the sweet-potato whitefly, *Bemisia tabaci*, to a wide range of crops, including including subsistence crops such as beans and non-traditional export crops such as melons, tomatoes and broccoli.

To minimize crop loss, farmers overuse pesticides causing crop and environmental contamination and farmer health problems. Pesticides are transported with crop shippers both locally and internationally. Preventative measures can be taken to minimize pesticide use by developing a simple model that can identify areas at risk.

Mapping of disease has had a long history and is becoming more prevalent with the increased use and availability of geographic information systems (GIS). GIS enables researchers to spatially view the location of viruses at various scales ranging from farm or country to region or continent. The scale of analysis is dependent upon the availability of the data and hardware and software capabilities of the computer.

The main objective was to create potential BGMV risk maps using techniques most suitable to the data available. For this study the logistic regression, fourier transform with principle components analysis, and a multi-process boolean analysis were evaluated. In order to run the analysis it was necessary to identify the bean-growing regions with and without BGMV and the environmental factors required for predicting the occurrence of the vector, *Bemisia tabaci.*

The data on the presence or absence of bean and BGMV information was collected from CIAT and the countries investigated. The information was digitized and analyzed over eleven months using the GIS software, Arc/Info and ArcView3.1.

2. Methods and Materials

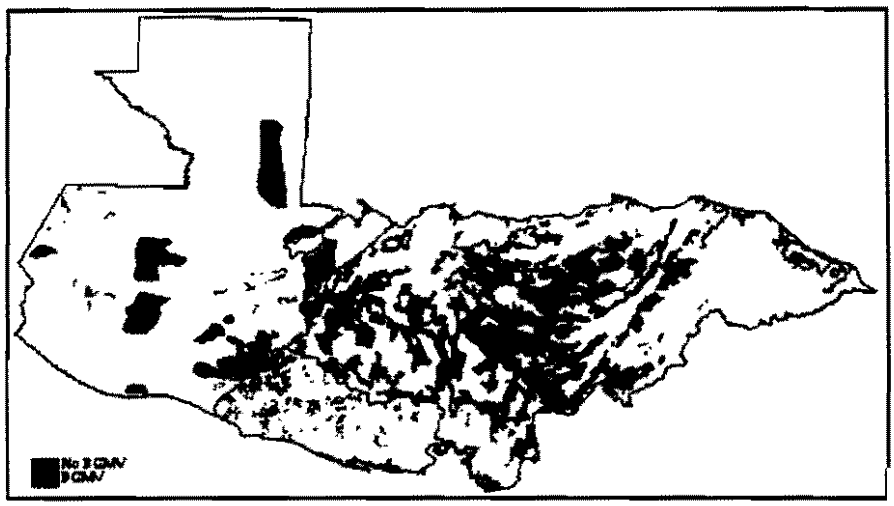
2.1 Bean data

Bean-growing areas were identified using agricultural census data, elevation and temperature ranges best suited for cultivating beans, interviews of local bean breeders, publications (ICTA, 1990; Cabrera *et al.* 1997, Morales, 1994), delineation of bean growing areas by local experts, and existing bean distribution maps.

The information was converted into digital data at the country level by extracting data by elevation or digitizing (large scale maps) or scanning and digitizing of smaller paper maps. Once each country's bean surface was created, all were combined to create the final bean surface used in the analysis. The final surface

(figure 1) indicates the location of BGMV and non-BGMV areas within the bean-growing regions.

Figure 1: Observed location of BGMV and non-BGMV areas within bean-growing region



Each country contains one or more bean growing seasons with one season most susceptible to BGMV outbreaks. For the purpose of this analysis, the bean-growing season with the worst BGMV problems were selected. For both Honduras and Guatemala major BGMV problems have been reported to occur between September/October to December/January while in El Salvador the season is from November to January.

2.1.2 BGMV data

The data used to create the areas containing the virus was obtained from three sources: (1) a publication *Mosaico Dorado del Frijol: Avances de Investigacion* (Morales, 1994); (2) expert local knowledge drawn on 1:75,000 scale maps, and (3) publications obtained various from bean experts (Cabrera *et al.* 1997, ICTA, 1990)

Maps published (with the permission) in *Mosaico Dorado del Frijol: Avances de Investigacion* (1994), locating the virus in each country were scanned, registered to existing administrative boundary coverage (CIAT) and digitized. The data sets was created using Arc/Info. Once the data were rasterized, areas located at elevations greater than 1000 metres above sea level were eliminated, since the virus, to date, has been reported to occur at elevations of less than 1000 metres (Rojas *et al.* 1994, Rodriguez *et al.* 1994).

2.1.3 Climate data information

Climate surfaces were developed at CIAT based on 30 year climate averages from about 10,000 stations in Latin America (Jones *et al.* 1999). The surfaces were interpolated using "the inverse square of the distance between the five nearest stations and the interpolated point"(Jones *et al.* 1999:). The

temperature surfaces were "standardized to the elevation of the pixel in the DEM using a lapse rate model" (Jones *et al.* 1999: pp). Included in the analysis were mean minimum monthly temperature (Jan to Dec), mean maximum monthly temperature (Jan to Dec), monthly rainfall (Jan to Dec) and consecutive dry months. Dry months are defined as less than 60 mm of rainfall in a month.

2.1.4 Environmental factors determining whitefly distribution

The distribution of the vector, *Bemisia tabaci* is influenced by temperature and host plant (Byrne *et al.*1991; Cohen *et al.*1991). Generally, whitefly populations increase during dry seasons with moderate to high temperatures. A detailed risk classification of whitefly occurrence is described in Table 1. These classes were based upon a classification developed by Arnez (1997) and modified by Morales (1999).

Table 1: Environmental factors influencing the distribution of *Bemisia tabaci*

Factors	**Characteristics**	**Unit**	**Risk Classification (Incidence of virus occurrence)**				
			Very low (0)	Low (1)	Moderate (2)	High (3)	Very High (4)
Environ-mental factors	Elevation	Msnm	>1500	1200-1500	1000-1200	500-1000	0-500
	Precipitation	mm	>3000	2000-3000	1500-2000	1000-1500	<1000
	Consecutive dry months	Month	0	0-1	1-2	2-3	3-6
	Temperature	C	<12 >32	12-16	17-21	22-26	27-31
Cropping Systems	Risk Assessment	Crop type	Cereals	Vegetables not in 3-5	Bean, potato, peppers (sweet and chili)	Tobacco, tomato, eggplant, cucumber	Cotton, soybean, melons, squash

Source: Morales (1999) based on Arnez(1997)

The incidence of the virus is ranked from very low (0) to very high (4) based on climatic factors favorable to whiteflies. Whitefly-geminivirus transmitting populations are highest (rank = 4) between elevations of 500-1000 with less than 1500 mm of rainfall, more than 4 consecutive dry months and a temperature range of 22°C to 26° C (Morales, 1999).

2.1.5 Other Geographical Information Data

CIAT's administrative boundary coverage was created in the GIS lab at CIAT in 1996. Boundaries were digitized country by country and adjusted to the DCW (Digital Chart of the World. Boundary information was obtained from maps of various scales (Barona, 1997).

The DEM was developed by the U.S. Geological Survey's EROS Data Center, Sioux Falls, South Dakota, 1996. The elevations are regularly spaced at approximately 1 kilometer (USGS, 1997) and were derived from eight data sources, both vector and raster. For Central America, the main source used was

digital terrain elevation data (DTED) set with enhancements made by the DCW data (USGS, 1999b; USGS, 1997).

3. Methods

Three methods were selected for this analysis and are the logistic regression, fourier transform with principal components analysis and a multi-process boolean analysis.

3.1 Logistic Regression

Multiple regression examines the relationship between the known values of a dependent variable, BGMV, with the known values of a set of independent variables, the number of dry months, elevation, rainfall, and minimum and maximum temperature. The data being analyzed are represented by the presence or absence of the virus within bean growing regions, where the dependent variable is discrete and dichotomous, with values 1 and 0. The areas where the virus is present are coded with a value 1; and bean producing regions without the virus are represented by 0. Since the data for the virus was available in a boolean format, a logistic regression was used.

The analysis was performed using the GRID component of the GIS software Arc/INFO. The logistic regression uses a maximum likelihood estimate to perform the regression and can be represented as:

$$P(x) = 1/(1+\exp(-(a_0 + a_1(mintemp) + a_2(maxtemp) + a_3(rain) + a_4(drymonth) + a_5(elevation))))$$

By substituting the coefficients into the above equation, a probability surface is created expressing the areas of virus occurrence ranging from 0 to 1, where 1 is the highest probability of occurrence. For example, grid cells with the value .95, have a 95% probability of containing the virus, while grid cells closer to 0 have a zero to no probability of BGMV occurrence.

The logistic regression used values at each sample point for BGMV and extrapolated values occurring in each independent layer occurring at the point location. The analysis was run on one bean-growing season because for the months in which the virus does not occur, results imply virus occurrence.

3.2 Fourier Transform and Principle Components Analysis

The season selected for the analysis was based on the planting seasons reported by experts within each country and varies from September/November to December/January. To compare between season, the seasonal timings need to be synchronized to a standard time (Jones *et al.* 1999). This can be accomplished using the Fourier transform (Jones *et al.* 1999). By synchronizing the season, a probability model (Principal Components Analysis (PCA)) can be used to locate the areas of probability of occurrence of an event such as a virus (Jones *et al.* 1997).

3.3 Multiprocess Boolean Analysis

Normally when performing a boolean analysis the criteria are either true (1) or false (0). By using a multi-process boolean analysis it is possible to use a multicriteria evaluation method for assessing and aggregating many criteria. First, it is necessary to rescale the criteria values into a standard numerical range. By reclassifying each criteria it is easier to relate and compare the various layers of different data information. Once each layer is reclassified, all layers can be combined and the results analyzed.

3.3.1 Creating the themes based on the risk classifications

For each month, the climate surfaces were classified from 0 – 4 using the risk classifications defined in table 1. This was accomplished using a boolean process where the factor for a particular risk group when true is assigned the value of that risk group, or else it is assigned a NODATA value.

3.3.2 Potential Risk using all themes

Using the reclassified climate surfaces, risk maps were created for each month by combining the themes by month. The potential risk was calculated using the formula

Riskmap1 = (elevation + temperature + rain + dry_months) / 16

The final risk map contains fractional values ranging from 0 to 1, where 0 is the least favourable and 1 is the most favourable. The new values allows the original risk factors (0 to 4) to be assigned a proportion of risk depending on the numeric combinations occurring at a particular location.
For example;

Table 2: Fractional Risk Calculation using the Risk Classification Values (0-4) assigned to each Environmental Theme

	Mor_ elev	mor_ temp	mor_ rain	mor_ dry	Sum	Risk Value
Site A	4	4	4	4	16/16	1
Site B	3	2	3	4	12/16	0.6

The results can be compared to the observed BGMV results, as discussed later. If the results of the boolean analysis indicate that there is little to no correlation between the occurrence of the vector and the environmental factors, the criteria for each risk classification, as defined in Table 1, will need to be re-assessed.

3.4 Methods used to verify the results

The results obtained were verified using three methods; (1) visually comparing the observed (Figure 1) with the expected, (2) calculating the goodness of fit for each cell and (3) calculating the percent accuracy of prediction.

4. Results

In this section the prediction results obtained for each of the methods are discussed.

4.1 Multivariate Logistic regression

The monthly regression surfaces illustrate the months that would be predicted to have the most severe whitefly problems. In Honduras, the worst problems occurred in December; while El Salvador was subject to whitefly problems for the entire bean-growing season (Nov – Jan). Guatemala showed the fewest problems occurring during September and October in the south of country, with increases in the virus as vectors migrated from north in November and December. A risk map of the most likely virus occurrence locations can be viewed by combining the regression surfaces.

Coefficients obtained from the logistic regression reflect the influence of each factor. For Honduras the min temp and max temp were the most important factors for the months of November, December and January. In El Salvador, dry months followed by the max and min temp showed the greatest influence for the months of December and January. Lastly, in Guatemala, the most influential factor was max temp followed by dry months and min temp. For all locations, rainfall and elevation were calculated to have marginal affects. According to the environmental information (Table 1), however, elevation is considered an important factor since the virus has not been reported to occur above 1000 meters (Rojas *et al.* 1994; Rodriguez *et al.* 1994).

The results obtained by logistic regression (Figure 2) were verified visually (Figure 2 vs Figure 1) and using the goodness of fit at each location. The visual comparison showed a good fit for Guatemala and El Salvador where the expected results were similar to the observed. Honduras, on the other hand, was not as good since the high-risk areas where BGMV has been reported was not visible in Figure 2.

The goodness of fit (figure 3) showed the best results for the logistic regression were in Guatemala and El Salvador, while the results obtained in Honduras were only good in the central west.

Figure 2: Logistic Regression Results; Probability of BGMV occurrence within the bean-growing region.

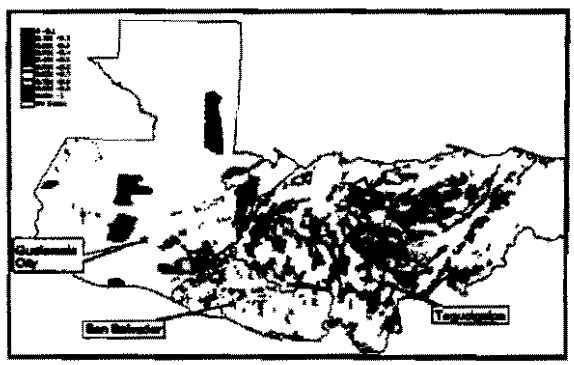
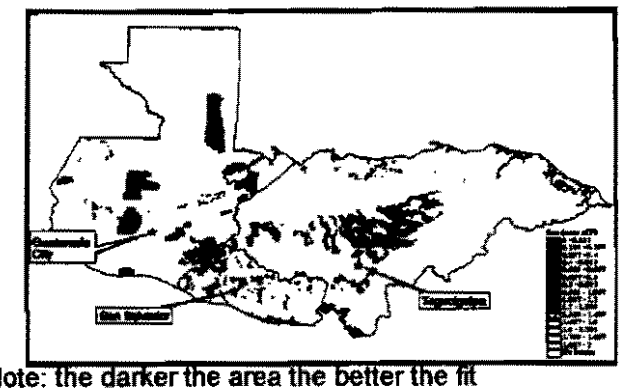


Figure 3: Logistic Regression Results: Goodness of Fit



Note: the darker the area the better the fit

4.2 Fourier transform with Principle Components Analysis (PCA)

The Fourier transform synchronizes the start of the different seasons and then the PCA is run on the data points representing the location of the virus. The analysis included all months since the program cannot differentiate between the different bean growing seasons in each country.

The result obtained (Figure 4) is coarse due to the resolution of the climate surface and should be used to obtain possible locations of the virus at a regional or continental level. Based on the results obtained from calculating the goodness of fit (Figure 5), areas in which the virus was present in Honduras were quite accurately predicted.

Figure 4: Fourier transform and Principle Components Analysis Results

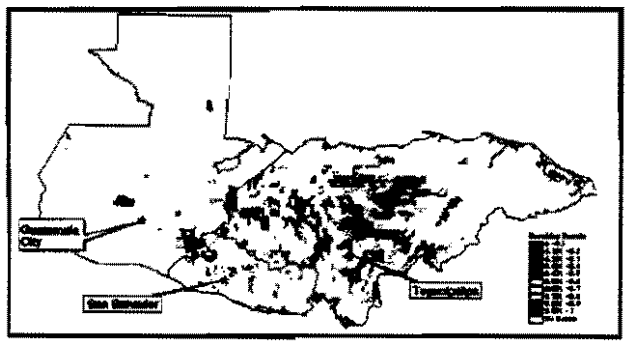
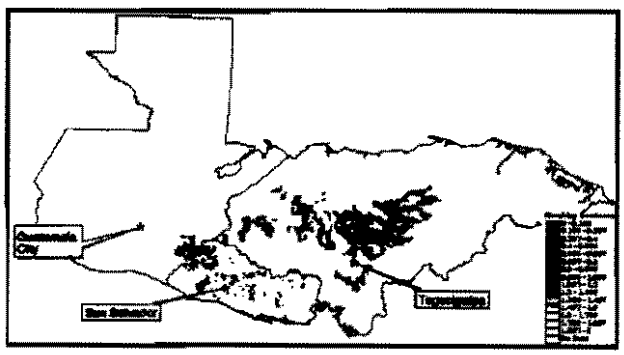


Figure 5: Fourier transform and PCA: Goodness of Fit



4.3 Multi-process Boolean Analysis

The boolean analysis was based on the risk classifications described in Table 1. Each layer was classified from 0 to 4 and combined to obtain the probability of occurrence at a particular location.

From Figure 6, most of the bean areas are highly susceptible to BGMV outbreaks with minor problems predicted in the north of Honduras and the highlands of Guatemala. The boolean method over predicted the occurrence of the virus in both BGMV and non-BGMV areas (Figure 6). As illustrated by the goodness of fit (Figure7), the boolean failed to accurately predict the occurrence of no-BGMV in non-BGMV areas.

Figure 6: Multi-process boolean Analysis

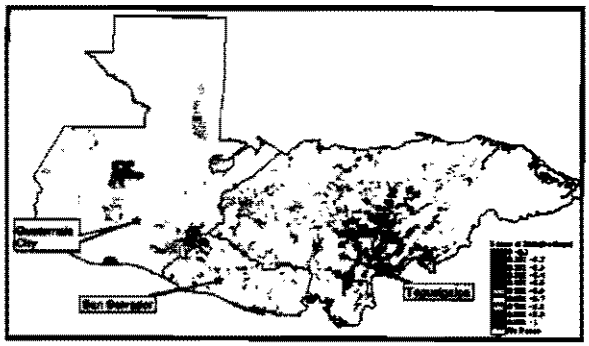
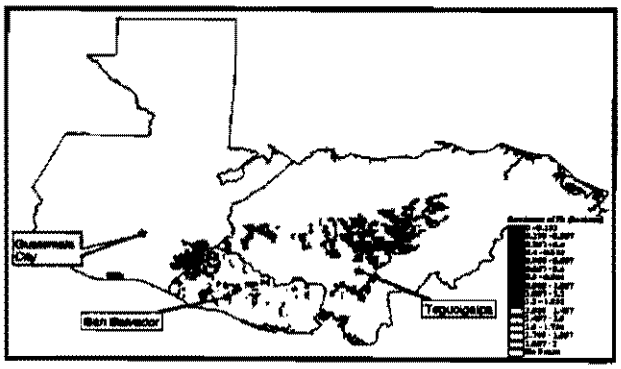


Figure 7: Multi-process boolean Analysis; Goodness of Fit



4.4 Comparison of the different methods

Based on the results illustrated in Figures 3 – 7, the method with the best goodness of fit for predicting BGMV in BGMV areas was the Boolean. The boolean method, however was terrible at accurately predicting no-BGMV in non-BGMV areas. The logistic regression was good at predicting the occurrence of BGMV and no-BGMV in Guatemala and El Salvador, but with less accurate results for Honduras. Lastly, the Fourier transform with PCA was the better than the logistic and boolean at predicting the occurrence of BGMV in BGMV areas for Honduras.

A simple method to compare the accuracy of each method is to calculate the percent of correctly and incorrectly predicted virus occurrence in both the BGMV and non-BGMV areas between the observed and expected results.

The accuracy of the results, for each BGMV and non-BGMV area was evaluated using a threshold for the probability of occurrence (p> 0.49). The accuracy for each method was determined and summarized in Table 3. For each method the error of misclassification was calculate. The logistic had the least error (21%) followed by the Boolean (64%) and lastly the Fourier transform (68%).

Table 3: Summary of the prediction of BGMV and non-BGMV vs the observed BGMV and non-BGMV results for the three methods.

	Observed BGMV (%)			***Observed no-BGMV (%)***		
	Logistic	*Fourier*	*Boolean*	*Logistic*	*Fourier*	*Boolean*
Predict BGMV	11	16	28	4	25	64
Predict No BGMV	17	9	0	68	16	8
No Data	0	3	0	0	31	0.0
Total	28			72		

The logistic regression was the most accurate, with 11% of the virus predicted to occur in BGMV areas and 68% no-BGMV predicted to occur in non-BGMV areas. The logistic predicted 4% BGMV in non-BGMV areas while predicting 17% no-BGMV in BGMV areas.

Second best was the Boolean method that predicted a 28% BGMV occurrence within BGMV areas and 64% occurring in non-BGMV areas. The boolean method tended to over-predict the occurrence of the virus and tended to predict a low occurrence of no-BGMV, 8% in total.

Lastly, the Fourier transform method predicted 16% BGMV occurrence in BGMV areas with 25% occurring in non-BGMV areas. A total 25% no-BGMV was predicted, with 16% occurring in non-BGMV and a 9% mis-classification in BGMV areas.

Of the two presence/absence methods, the logistic was better at predicting BGMV areas for probabilities greater than 0.49. The Fourier transform with PCA was good at predicting an overview of the problem areas and should be run with a climate surfaces at finer resolution. Lastly, the boolean method, although the least accurate, illustrates the need to re-asses the risk classifications required for the virus to occur (Table 2).

5. Conclusion

The techniques used in this study were compared and contrasted to see which would be the best to use with the data available. The logistic regression can be used in conjunction with the environmental factors required for the vector. By

comparing and contrasting these, it is possible to improve on the definition of parameters which predict whitefly occurence.

Having completed the initial evaluations, it is possible to predict the likelihood of the plant virus outbreak. The techniques need to be improved to develop a climate-based model for assessing areas of risk within a particular cropping system. Once an accurate modeling technique has been identified this can aid scientists and policy-makers in determining critical areas where preventative measures are required.

In the future it will be necessary to include specific information for BGMV such as bean varieties and their resistance to the virus, intensity of outbreaks at each location and the control practices implemented to reduce further crop losses.

6. Acknowledgements

Thanks to the collaboration of the following institutes who made it possible to obtain the information required for the analysis. These include;
Colombia: *CIAT (Centro Internacional de Agricultura Tropical)*; Guatemala: *ICTA; Profrijol*; Honduras: *CIAT-Honduras Office; El Zamorano, Escuela Agricola Panamericana*; El Salvador: *University of El Salvador; CENTA; FAO-CENTA*
Many thanks to all the GIS staff at CIAT and all my colleagues for their support through the brainstorming and kept me on track. Thanks to the Canadian International Development Agency (CIDA) for providing funding through the Innovative Research Award for students (1998). This enabled me to visit the study areas and collect the information for the analysis. A special thanks to Dr S. Fujisaka for taking the time to give constructive feedback, edit the paper and make me question what I was doing at every step.

7. References

Arnez, Juan E. V. (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 (CATIE).* Turrialba, Costa Rica.

Barona, E. (1997) Coberatura de America Latina – Division Administrativa, unpublished working paper, CIAT, Cali, Colombia

Brunt, A.A., Crabtree, K., Dallwitz, M.J., Gibbs, A.J., Watson, L. and Zurcher, E.J., (eds.)(1996 onwards) Plant Viruses Online: Descriptions and Lists from the VIDE Database. Version: 16th January 1997. *URL http://biology.anu.edu.au/Groups/MES/Vide*

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

Byrne, D.N., Bellows, T.S. JR., and Parilla, M.P. (1991) Whiteflies in Agricultural Systems. Pp 227 –259. In: Gerling, D. (ed.). *Whiteflies: Their Bionomics, Pest Status and Management.* Intercept Ltd., United Kingdom

Byrne, D.N., Rathman, R.J., Orum, T.V., Palumbo, J.C. (1996) Localized migration and dispersal by the sweet potato whitefly, *Bemisia tabaci.* In *Oecolgia.* **105**. 320-328.

Cabrera, Carlos Atilio Pérez (1997) Situación Actual del Cultivo de Frijol en El Salvador, Documento presentado a la Misión de Evalución Externa de PROFRIJOL, CENTA, El Salvador

Cabrera, Carlos Atilio Pérez, Edgardo Mendoza Puquirre, (1999) Importancia del Frijol en El Salvador, CENTA, El Salvador

Cliff, A. D. and Haggett, P. (1996) The Impact of GIS on Epidemiological Mapping and Modelling. Pp 321-343. In: Spatial Analysis: Modelling in a GIS Environment. Longley, P. and Batty, M. (eds.). GeoInformation International, Cambridge, United Kingdom.

Cohen, S. (1991) Epidemiology of Whitefly-Transmitted Viruses. Pp 211 –225. In: Gerling, D. (ed.). Whiteflies: Their Bionomics, Pest Status and Management. 1991. Intercept Ltd., United Kingdom

Gerling, D. (ed.) (1991) Whiteflies: Their Bionomics, Pest Status and Management. Intercept Ltd., United Kingdom

Hilje, L and O. Arboleda (1993) Las moscas blancas (Homópetera: Aleyrodidae) en Centro América y el Caribe. *Memoria del Taller sobre moscas blancas.* CATIE, Turrialba, Costa Rica. 78p

ICTA (1990) Recomendaciones Tecnicas Agropecuarias, Instituto de Ceincia y Tecnologia Agricolas (ICTA), Guatemala

Jones, P. and Gladkov, A. (1999) FloraMap. Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia

Jones, P., Beebe, S.E., Tohme, J. and Galwey, N.W. (1997) The use of geographical information systems in biodiversity exploration and conservation. In *Biodiversity and Conservation.* **6**. 947-958

Morales, F. J. (ed). (1994) El Mosaico Dorado del Frijol: Avances de Investigacion. CIAT, Cali, Colombia, 193pp.

Morales, F. J. (1999) Personal Communication, CIAT, Cali, Colombia

Digital Data Bibliography:

- Administrative Boundary Coverage for Central America, 1997, CIAT, Cali, Colombia
- Bean golden mosaic virus surfaces, 1998, digitized from the Bean golden mosaic virus, Advances in research, 1994, CIAT, Cali, Colombia
- Climate Surfaces: P Jones, 1991, CIAT, Cali, Colombia
  Mean minimum monthly temperature (Jan to Dec) and Mean maximum monthly temperature (Jan to Dec) were corrected to elevation of the pixel by a lapse rate model for the mean tropical atmosphere from night soundings in the Caribbean. Data from Rhiel, H. (1979) Climate and weather in the tropicas. Academic Press London. p 62.
- USGS

USGS (1995), Digital Chart of the World(DCW), Defence Mapping Agency, *http://164.214.2.54/guides/dtf/dcw.htm*

USGS, (1998), GTOPO30 Documentation.
http://edcwww.cr.usgs.gov/landdaac/gtopo30/gtopo30.html

USGS (1999b), GTOPO30 Source Data.
http://edcwww.cr.usgs.gov/landdaac/gtopo30/gifs/gt30src.gif